DOI: https://doi.org/10.21323/2414-438X-2022-7-1-42-57

Available online at https://www.meatjournal.ru/jour
Review article
Open Access

Received 15.12.2021
Accepted in revised 22.02.2022
Accepted for publication 25.03.2022

METHODS FOR NONPARAMETRIC STATISTICS IN SCIENTIFIC RESEARCH. OVERVIEW. PART 2

Marina A. Nikitina*, Irina M. Chernukha
V. M. Gorbatov Federal Research Center for Food Systems of Russian Academy of Sciences, Moscow, Russia

Keywords: *nonparametric statistics, null and alternative hypotheses, type I error, type II error, goodness-of-fit tests, tests for homogeneity*

Abstract
The use of nonparametric methods in scientific research provides a number of advantages. The most important of these advantages are versatility and a wide range of such methods. There are no strong assumptions associated with nonparametric tests, which means that there is little chance of assumptions being violated, i. e. the result is reliable and valid. Nonparametric tests are widely used because they may be applied to experiments for which it is not possible to obtain quantitative indicators (descriptive studies) and to small samples. The second part of the article describes nonparametric goodness-of-fit tests, i. e. Pearson's test, Kolmogorov test, as well as tests for homogeneity, i. e. chi-squared test and Kolmogorov-Smirnov test. Chi-squared test is based on a comparison between the empirical (experimental) frequencies of the indicator under study and the theoretical frequencies of the normal distribution. Kolmogorov-Smirnov test is based on the same principle as Pearson's chi-squared test, but involves comparing the accumulated frequencies of the experimental and theoretical distributions. Pearson's chi-squared test and Kolmogorov test may also be used to compare two empirical distributions for the significance of differences between them. Kolmogorov test based on the accumulation of empirical frequencies is more sensitive to differences and captures those subtle nuances that are not available in Pearson's chi-squared test. Typical errors in the application of these tests are analyzed. Examples are given, and step-by-step application of each test is described. With nonparametric methods, researcher receives a working tool for statistical analysis of the results.

For citation: Nikitina, M.A., Chernukha, I.M. (2022). Methods for nonparametric statistics in scientific research. Overview. Pert 2. *Theory and practice of meat processing,* 7(1), 42-57. https://doi.org/10.21323/2414-438X-2022-7-1-42-57

Funding:
The research was supported by state assignment of V. M. Gorbatov Federal Research Centre for Food Systems of RAS, scientific research No. FNEN–2019–0008.

Introduction

German philosopher, psychologist and teacher Johann Friedrich Herbart at the beginning of the 19th century wrote: "Any theory trying to be consistent with experience, first of all, must be continued until it accepts quantitative determinations that arise in experience or lie in its foundation. If not, it hangs in the air, exposed to every wind of doubt and being unable to contact with other, already strengthened opinions".

Thus, the researcher, having received data during the experiment, must process them correctly using mathematical methods in order to draw a correct and reasonable conclusion.

As a rule, researchers use methods of parametric statistics, which is not always correct. Many parametric methods have direct analogues in nonparametric statistics. For example, Student test and analysis of variance determine the significance of differences in mean values for two or more groups; and Mann-Whitney U-test determines the significance of differences in the average rank for two groups; Pearson's correlation coefficient allows determining the linear relationship between two numerical indicators; and Spearman rank correlation coefficient allows determining linear relationship between the ranks of two indicators. In some cases, there is no direct analogy with nonparametric method.

Nonparametric methods of mathematical statistics do not require knowledge of the functional form for the theoretical distribution. The name "nonparametric methods" itself emphasizes their difference from classical (parametric) methods, in which it is assumed that the unknown theoretical distribution belongs to some family that depends on a finite number of parameters (for example, the family of normal distributions), and which allow estimating unknown values of these parameters based on the results of observations and testing certain hypotheses regarding their values [1].

Common characteristics for most nonparametric methods [2,3] are: 1) fewer assumptions about the type of distribution; 2) the sample size is less strict; 3) the measurement may be nominal or ordinal; 4) independence of randomly selected observations, except for paired ones; 5) the focus is on the ranking order or data frequency; 6) hypotheses are expressed regarding the ranks, medians or data frequency.

Based on the practice of statistical data analysis, there are three main spheres of nonparametric statistics [4]:

- sphere at the junction of parametric and nonparametric methods;
- rank statistical methods;
- nonparametric estimates for functions, primarily distribution density, regression dependence, as well as statistics used in classification theory.

In the first part of the article [5], a review of simple nonparametric methods is given. Two groups of nonparametric tests are considered: 1) to identify differences in the indicator distribution (Rosenbaum Q-test, Mann-Whitney U-test); 2) estimates of the significance for shift in the values of the studied indicator (sign G-test, Wilcoxon T-test).

In the second part of the article, nonparametric tests for testing hypotheses of distribution type (Pearson's chi-squared test, Kolmogorov test) and nonparametric tests for testing hypotheses of homogeneity (Pearson's chi-squared test for homogeneity, Kolmogorov-Smirnov test) are considered.

The purpose of the article is to give a working tool for solving specific research and applied problems using methods of nonparametric statistics.

Materials and methods

The materials of the study are recent publications in the statistical analysis of which methods of nonparametric statistics are used, i. e. goodness-of-fit tests (Kolmogorov test, Kolmogorov-Smirnov test, Pearson's chi-squared test).

Goodness-of-fit tests

It is known that one of the most important tasks for mathematical statistics is the establishment of a theoretical law of distribution for a random variable characterizing the studied indicator, based on empirical distribution. The solution of this problem allows: 1) choosing the right method of statistical data processing; 2) determining the type of model that describes the relationship between the analyzed indicators.

Goodness-of-fit tests are used to check the agreement between the experimental data and the theoretical model. So, goodness-of-fit test is a test for testing a hypothesis about an assumed distribution law [6].

The researcher states two hypotheses: null hypothesis (H_0) and alternative hypothesis (H_1). Next, the hypotheses are tested using various tests.

H_0: The resulting empirical indicator distribution does not differ from the theoretical distribution (normal, uniform, exponential, etc.).

H_1: The resulting empirical distribution of the indicator differs from the theoretical distribution.

To test the null hypothesis H_0, some random variable U is chosen, which characterizes the disagreement between the theoretical and empirical distributions, the distribution law for which is known, for sufficiently large n, and almost does not depend on the distribution law for the random variable X.

When knowing the distribution law of the random variable U, a critical value U_α can be found, at which the null hypothesis H_0 is true, as well as the probability that the random variable U assumes a value greater than U_α, i. e. the function $P(U > U_\alpha) = \alpha$ is small, where α is the test significance level.

If the value observed in the experiment $U_i = U > U_\alpha$, i. e. it falls into the critical region, this means that such large U values are practically impossible and contradict the hypothesis H_0. In this case, the hypothesis H_0 is rejected.

If $U_i = U \leq U_\alpha$, then the difference between the empirical and theoretical distributions is insignificant, and the hypothesis H_0 may be considered as not contradicting the experimental data.

In this case, the researcher can make two types of errors when testing hypotheses: type I error and type II error [6].

Type I error. If we reject the null hypothesis H_0 (i. e., we consider the null hypothesis H_0 is false), while in fact the null hypothesis H_0 is true, then the researcher makes an error consisting in the incorrect rejection of the null hypothesis.

Type II error. If we accept the null hypothesis H_0 (i. e., we do not agree with the alternative hypothesis H_1), while in fact the null hypothesis H_0 is false, then the researcher makes an error consisting in incorrect acceptance of the null hypothesis.

It is worth noting that the probability of making a type I error is established quite easily, because it is equal to α, while for type II errors, it must be specially calculated.

Pearson's goodness-of-fit test or Pearson's chi-squared test

Pearson's goodness-of-fit test (or Pearson's chi-squared test) is the most commonly used to test the hypothesis that a certain sample belongs to a theoretical distribution law [7,8].

Given data for the problem: let there be a sample of values for a random variable X with size n: $x_1, x_2, \ldots, x_k$ and a set of corresponding frequencies $m_1, m_2, \ldots, m_k$ (k is the number of partition intervals). As a measure of difference between the empirical and theoretical distributions, the value χ^2 is taken, which is equal to the sum of the squared deviations of the relative frequencies $\frac{m_i}{n}$ from the probabilities p_i calculated from the assumed distribution and taken with a certain coefficient c_i:

$$\chi^2 = \sum_{i=1}^{k} c_i \left(\frac{m_i}{n} - p_i\right)^2 \quad (1)$$

The coefficient c_i is chosen in such a way that for the same deviations $\left(\frac{m_i}{n} - p_i\right)^2$, the deviations at which p_i is small have more weight, and the deviations at which p_i is large have less weight. Therefore, $\frac{n}{p_i}$ ratio is taken as c_i. We obtain the measure of difference of the following form:

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i}\left(\frac{m_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{{m_i}^2}{np_i} - n =$$
$$= \sum_{i=1}^{k} \frac{\left(m_i - m_i^{theor}\right)^2}{m_i^{theor}} = \sum_{i=1}^{k} \frac{m_i^2}{m_i^{theor}} - n,$$

so that, with $n \to \infty$, the sample distribution of χ^2 tends to the limit distribution of χ^2 with the number of degrees of freedom $v = k - r - 1$, where r is the number of parameters of the hypothetical probability distribution estimated from the sample data. Numbers m_i and m_i^{theor} are called empirical and theoretical frequencies, respectively.

Application of Pearson's chi-squared test

1. The measure of difference between empirical and theoretical frequencies is determined by the formula (2) and the experimental value of the test is calculated.
2. For the chosen significance level α, using the table of χ^2 distributions, the critical value χ^2_{cr} is found with the number of degrees of freedom $v = k - r - 1$.
3. If the experimental value χ^2_{exp} is greater than the critical value, i.e. $\chi^2_{exp} \geq \chi^2_{cr}$, then the null hypothesis H_0 is rejected; and if $\chi^2_{exp} < \chi^2_{cr}$, the null hypothesis H_0 does not contradict the experimental data.

Limitations of Pearson's chi-squared test

1. Sample size must be large enough: $n \geq 30$.
2. The theoretical frequency for each cell should not be less than 5.
3. The selected ranks should cover the entire range of the indicator's variability. Classification into ranks should be the same in all compared distributions.
4. Ranks should be non-overlapping.

Testing the hypothesis about the normal distribution of the general population

1. Based on a sample of size n, arrange the interval statistical array by classification of the given data into k ranges $[a_i; a_{i+1})$ with the corresponding frequencies m_i. Rearrange interval statistical array into statistical array by replacing each range $[a_i; a_{i+1})$ with its mean value: $x_i = \frac{a_i + a_{i+1}}{2}$. Now we have Table 1.

Table 1. Interval statistical array

Ranges for observed values of a random variable *X*	$[a_1; a_2)$	$[a_2; a_3)$	...	$[a_i; a_{i+1})$	$[a_k; a_{k+1})$
Frequencies m_i	m_1	m_2	...	m_i	m_k
Mean value x_i	x_1	x_2	...	x_i	x_k

2. Using Table 1, calculate mathematical expectation estimate $\overline{x}$ and sample standard deviation σ_v.
3. Calculate $z_i = \frac{a_i - \overline{x}}{\sigma_в}$, $i = \overline{2,3,\ldots k}$, where a_i is the left end of the i^{th} range. Set value z_1 equal to minus ∞, and value z_{k+1} equal to plus ∞.
4. Assuming a normal distribution of the general population, determine the theoretical frequencies m_1^{theor}, m_2^{theor}, …, m_k^{theor} by the formula:

$$m_i^{theor} = n \cdot p_i,$$

where $p_i = \Phi(z_{i+1}) - \Phi(z_i)$ is the probability of a random variable X to fall within the range $[a_i; a_{i+1})$; $\Phi(x)$ is the cumulative Laplace distribution function.

5. Calculate χ^2_{exp} by the formula:

$$\chi^2_{exp} = \sum_{i=1}^{k} \frac{\left(m_i - m_i^{theor}\right)^2}{m_i^{theor}} \quad (3)$$

or

$$\chi^2_{exp} = \sum_{i=1}^{k} \frac{m_i^2}{m_i^{theor}} - n \quad (4)$$

6. Using the table, calculate χ^2_{cr} [9,10,11,12], considering the given level of significance α and the number of degrees of freedom $v = k - 3$.
7. Compare χ^2_{exp} and χ^2_{cr}.

If $\chi^2_{exp} < \chi^2_{cr}$, there is no reason to reject the hypothesis about the normal distribution of the general population.

If $\chi^2_{exp} \geq \chi^2_{cr}$, the hypothesis about the normal distribution of the general population should be rejected.

Testing the hypothesis about the distribution of a random variable according to a uniform law

1. Group the sample data by arranging them as a sequence k of the ranges $[a_i; a_{i+1})$ and their corresponding frequencies m_i, $i = \overline{1,\ldots,k}$, $a_1 = a$, $a_{k+1} = b$.
2. From a given variational array, calculate the probabilities p_i of X to fall within the range by the formula:

$$p_i = P(a_i < X < a_{i+1}) = \frac{a_{i+1} - a_i}{b - a} \quad (5)$$

3. Calculate theoretical frequencies by the formula:

$$m_i^{theor} = n \cdot p_i,$$

where n is sample size.

4. Calculate χ^2_{exp} by the formula (4).
5. For given significance level α and the number of degrees of freedom $v = k - 1$, calculate χ^2_{cr} using the table [9,10,11,12].
6. Compare χ^2_{exp} and χ^2_{cr}.

If $\chi^2_{exp} < \chi^2_{cr}$, there is no reason to reject the hypothesis of uniform distribution of X within the range [a; b].

If $\chi^2_{exp} \geq \chi^2_{cr}$, then the hypothesis of uniform distribution should be rejected.

Example 1. 48 cows were examined for deviations of the annual milk yield from the average. Grouped data are given in Table 2.

Table 2. Given data for the problem

Annual milk yield, kg	0÷1000	1000÷2000	2000÷3000	3000÷4000	4000÷5000
Number of cows, animals	2	8	23	13	2

Evaluate the hypothesis about the normal distribution of the general population at a significance level α ≤ 0.05 with Pearson's chi-squared test.

Solution. Let's rearrange interval statistical array into statistical array by replacing each range $[a_i; a_{i+1})$ with its mean value $x_i = \frac{a_i + a_{i+1}}{2}$. Now we have Table 3.

Table 3. Statistical array

x_i	500	1500	2500	3500	4500
m_i	2	8	23	13	2

Using Table 3, let's calculate mathematical expectation estimate $\overline{x}$ and sample standard deviation σ_v.

Mathematical expectation:

$$\overline{x} = \frac{1}{48}\sum_{i=1}^{5} x_i m_i = \frac{1}{48} \cdot (500 \cdot 2 \ + \ 1500 \cdot 8 \ + \ + 2500 \cdot 23 \ + \ 3500 \cdot 13 \ + \ 4500 \cdot 2\ ) \ \approx 2604;$$

Sample variance

$$D_v = \frac{1}{48}\sum_{i=1}^{5} x_i^2 m_i - \overline{x}^2 = \frac{1}{48} \cdot (500^2 \cdot 2 \ + \ 1500^2 \cdot 8 \ + \ + 2500^2 \cdot 23 \ + \ 3500^2 \cdot 13 \ + \ 4500^2 \cdot 2) - - 2604^2 \ \approx 759982.64;$$

Sample standard deviation

$$\sigma_v = \sqrt{D_v} = \sqrt{759982.64} \approx 871.77.$$

Let's calculate $p_i = \Phi(z_{i+1}) - \Phi(z_i)$, the probability of a random variable X to fall within the range $[a_i; a_{i+1})$; $\Phi(x)$ is the cumulative Laplace distribution function, $z_i = \frac{a_i - \overline{x}}{\sigma_в}$.

$$p_1 = \Phi\left(\frac{1000 - 2604}{871.77}\right) - \Phi(-\infty) = -0.4671 + 0.5 = = 0.0329;$$

$$p_2 = \Phi\left(\frac{2000 - 2604}{871.77}\right) - \Phi\left(\frac{1000 - 2604}{871.77}\right) = = -0.2549 + 0.4671 = 0.2122;$$

$$p_3 = \Phi\left(\frac{3000 - 2604}{871.77}\right) - \Phi\left(\frac{2000 - 2604}{871.77}\right) = = 0.1736 + 0.2549 = 0.4285;$$

$$p_4 = \Phi\left(\frac{4000 - 2604}{871.77}\right) - \Phi\left(\frac{3000 - 2604}{871.77}\right) = = 0.4452 - 0.1736 = 0.2716;$$

$$p_5 = \Phi(+\infty) - \Phi\left(\frac{4000 - 2604}{871.77}\right) = 0.5 - 0.4452 = 0.0548.$$

Let's calculate m_i^{theor} by the formula $m_i^{theor} = n \cdot p_i$ and complete Table 4.

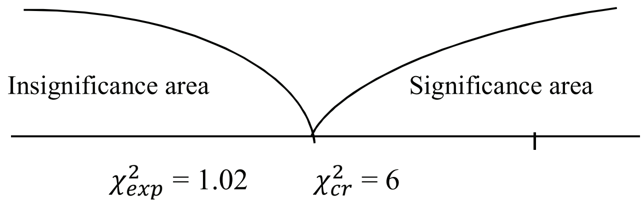
Table 4. Calculation results

№	x_i	m_i	m_i^2	p_i	m_i^{theor}	$\frac{m_i^2}{m_i^{theor}}$
1	500	2	4	0.0329	1.5792	2.532928
2	1500	8	64	0.2122	10.1856	6.28338
3	2500	23	529	0.4285	20.568	25.71956
4	3500	13	169	0.2716	13.0368	12.9633
5	4500	2	4	0.0548	2.6304	1.520681
Σ		48		1	48	49.01986

$\chi^2_{exp} = 49.01986 - 48 = 1.01986 \approx 1.02.$

Using the table [9, 10, 11, 12], for α ≤ 0.05 and $v = k - r = 5 - 3 = 2$ let's determine $\chi^2_{cr} = 6$

Let's plot the axis of significance:



Since 1.02 < 6 ($\chi^2_{exp} < \chi^2_{cr}$), hypothesis about the normal distribution of the general population should be accepted.

Example 2. In some areas, the distribution of cows by live weight was recorded. Grouped data are given in Table 5.

Table 5. Given data for the problem

Live weight, kg	400÷420	420÷440	440÷460	460÷480	480÷500
Livestock, animals	12	39	88	82	86

Evaluate the hypothesis about the normal distribution of the general population at a significance level α ≤ 0.05 with Pearson's chi-squared test.

Solution. Let's rearrange interval statistical array into statistical array by replacing each range $[a_i; a_{i+1})$ with its mean value $x_i = \frac{a_i + a_{i+1}}{2}$. Now we have Table 6.

Table 6. Statistical array

x_i	410	430	450	470	490
m_i	12	39	88	82	86

Using Table 6, let's calculate mathematical expectation estimate $\overline{x}$ and sample standard deviation σ_v.

Mathematical expectation:

$$\overline{x} = \frac{1}{307}\sum_{i=1}^{5} x_i m_i = \frac{1}{307} \cdot (410 \cdot 12 \ + \ 430 \cdot 39 \ + \ + 450 \cdot 88 \ + \ 470 \cdot 82 \ + \ 490 \cdot 86\ ) = 462.443;$$

Sample variance

$$D_v = \frac{1}{307}\sum_{i=1}^{5} x_i^2 m_i - \overline{x}^2 = \frac{1}{307}\cdot(410^2\cdot 12 + 430^2\cdot 39 + 450^2\cdot 88 + 470^2\cdot 82 + 490^2\cdot 86) - 462.443^2 \approx 513.5757;$$

Sample standard deviation

$$\sigma_v = \sqrt{D_v} = \sqrt{513.5757} \approx 22.66.$$

Let's calculate $p_i = \Phi(z_{i+1}) - \Phi(z_i)$, the probability of a random variable X to fall within the range $[a_i; a_{i+1})$; $\Phi(x)$ is the cumulative Laplace distribution function, $z_i = \frac{a_i - \overline{x}}{\sigma_в}$.

$$p_1 = \Phi\left(\frac{420 - 462.443}{22.66}\right) - \Phi(-\infty) = -0.4693 + 0.5 = 0.0307;$$

$$p_2 = \Phi\left(\frac{440 - 462.443}{22.66}\right) - \Phi\left(\frac{420 - 462.443}{22.66}\right) = -0.3389 + 0.4693 = 0.1304;$$

$$p_3 = \Phi\left(\frac{460 - 462.443}{22.66}\right) - \Phi\left(\frac{440 - 462.443}{22.66}\right) = -0.0389 + 0.3389 = 0.2991;$$

$$p_4 = \Phi\left(\frac{480 - 462.443}{22.66}\right) - \Phi\left(\frac{460 - 462.443}{22.66}\right) = 0.2794 + 0.0389 = 0.3192;$$

$$p_5 = \Phi(+\infty) - \Phi\left(\frac{480 - 462.443}{22.66}\right) = 0.5 - 0.2794 = 0.2206.$$

Let's calculate m_i^{theor} by the formula $m_i^{theor} = n \cdot p_i$ and complete Table 7.

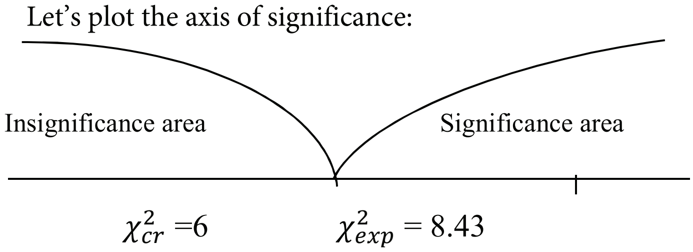
Table 7. Calculation results

№	x_i	m_i	m_i^2	p_i	m_i^{theor}	$\frac{m_i^2}{m_i^{theor}}$
1	410	12	144	0.0307	9.4249	15.2786767
2	430	39	1521	0.1304	40.0328	37.99384505
3	450	88	7744	0.2991	91.8237	84.33552558
4	470	82	6724	0.3192	97.9944	68.61616582
5	490	86	7396	0.2206	67.7242	109.2076392
Σ		307		1	307	315.432

$\chi^2_{exp} = 315.432 - 307 = 8.43185 \approx 8.43.$

Using Table [9, 10, 11, 12], for α ≤ 0.05 and $v = k - r = 5 - 3 = 2$ let's determine $\chi^2_{cr} = 6$

Let's plot the axis of significance:



Since 8.43 > 6 (when accepting the null hypothesis, it should be $\chi^2_{exp} < \chi^2_{cr}$), hypothesis about the normal distribution of the general population should be rejected.

Using the example of research in various fields, we will show the application of Pearson's goodness-of-fit test. The article [15] shows the applicability of target hazard quotient (THQ) estimates for communicating the danger of seafood due to metal contamination. The food recall data set was collected by the Laboratory of government chemists (LGC, UK) between January and November 2007. For example, seafood products originating in only 3 countries were recalled more than 10 times due to metal contamination (Spain, 50 times; France, 11 times; Indonesia, 11 times). Products containing swordfish and sharks have been recalled more than 10 times, mostly due to mercury contamination. Based on the food alert/recall system, the application of THQ risk assessment in cases of seafood contamination with metals is questionable, as THQ implies frequent (or even daily) lifetime exposure. Infrequent recalls due to metal contamination and lack of trend make it highly unlikely that a person would be exposed to repeated significant levels of metal ions in seafood. Pearson's goodness-of-fit chi-squared test, nonparametric correlation (Kendall's tau) and Kruskal-Wallis test were used to confirm the hypothesis and perform statistical processing. The work [16] shows a study of perception, belief and behavior in relation to nutritional and complementary practices in inflammatory bowel disease (IBD). 80 patients with IBD completed a closed-ended 16-item questionnaire that was divided into three subsections: 1) baseline/demographic characteristics; 2) disease characteristics; 3) dietary and complementary beliefs and behaviors. One-sample chi-squared goodness-of-fit tests were used for each question, and two-sided Pearson's chi-squared tests of independence were used for testing differences in response to each question between baseline/demographic variables.

The processing time of 1.0 cm, 1.5 cm and 2.0 cm potato cubes with 0.4%, 0.8% and 1.2% aqueous solutions of sodium carboxymethyl cellulose at flow rates of 453 ml/s, 534 ml/s and 599 ml/s was measured for the performance of vertical scraped surface heat exchanger (VSHE) rotating at 60, 110 and 160 rpm, and the particle flow distribution characteristics for each set of conditions were studied in [17]. Statistical data processing using Pearson's chi-squared test showed that most distributions for the residence time of individual particles in the vertical flow in VSHE may be described by the gamma model, while for the horizontal VSHE, many of the individual distributions correspond to the normal model in addition to the gamma model. VSHE orientation turned out to be an important factor influencing the forces acting on particles during the flow in the VSHE. Interactions of particles with each other, as well as a combination of process parameters, caused a "tail" of some particles, which led to a shift in the distribution to the right. The purpose of the article [18] was to assess the purchasing behavior of consumers and the decision-making process when buying bread and to suggest ways to improve bread positioning in the market. 1601 correctly completed questionnaires were used for the analysis. Results were presented as response rates and statistical tests. The analysis included the evaluation of statistical hypotheses about independence (significance level $\alpha = 0.01$) using goodness-of-fit chi-squared test and Pearson's randomness coefficient. Then the significance level was compared with the p value. For the p value > α, the null hypothesis was not rejected. The most important factors in choosing bread are freshness, appearance and price. Importance of price increases with the age of the respondents and decreases with the income of the surveyed consumers. The importance of a brand, as well as referrals from family and friends, increases slightly as consumer income increases. When making a purchase decision, most respondents do not make a difference between yeast and rye-yeast bread baking technologies. However, it cannot be stated that the preference for rye-yeast bread increases with the age of the respondents to the detriment of yeast bread, or vice versa.

In [19], gender differences were determined in the self-assessment of social functioning in patients with comorbidity of affective disorders and chronic coronary artery disease. The study included 248 cardiac patients (194 men (78.2%) and 54 women (21.8%)) with chronic coronary artery disease and affective disorders. The mean age of patients with chronic disease in men was (57.2 +/− 6.5) years, and in women it was (59.3 +/− 7.1), p = 0.04. Qualitative and quantitative indicators were examined using the Mann-Whitney test, Wilcoxon test and T-test; chi-squared test (Pearson's goodness-of-fit test) was used to estimate frequencies. The purpose of the study in [20] was to reveal the parents' ideas about the main trends and structural features of children's Internet addiction. The study was based on the results of a mass survey. The survey was conducted in 2019 on a multi-stage sample (by gender, age, type of location), consisting of the adult population at the Tyumen region. The authors carried out a detailed socio-statistical analysis of Internet risks for children based on self-assessments of all respondents (with identification of socio-demographic groups), risk assessments for children according to parents. The structure of "Parents" subsample by gender and type of location was proportional to the structure of the main sample. According to the authors, "Children" subsample included respondents' children of minority age. The risk of Internet addiction was included in the structure of 12 Internet risks and examined on the basis of 4 components (behavioral, cognitive, social and affective components). The analysis used Cronbach's alpha consistency ratings, index method, Spearman rank correlation coefficients, Pearson's goodness-of-fit test, F-test for equality of several means, case classification and triangulation method. The study [21] examined the relationship between mean micturition volume and urinary incontinence episodes per 24 hours after adjusting for fixed frequencies in children with overactive bladder. Patients were aged 5 to 12 years with >= 4 episodes of daytime urinary incontinence during the 7-day period prior to study entry. Mean number of episodes of urinary incontinence per 24 hours at the end of the study was the dependent variable. Explanatory variables included treatment, mean number of episodes of urinary incontinence per 24 hours at baseline, and change in mean micturition volume from baseline to the end of the study. Statistical significance and degree of conformity were analyzed using Pearson's chi-squared test. The aim of the study [22] was to evaluate the effectiveness of a pediatric mortality index of 3 in predicting mortality at the intensive care unit. This was an observational study conducted in the intensive care unit from January 2016 to October 2018. All patients aged 1 month to 15 years who were hospitalized to the intensive care unit were included. The authors analyzed the relationship between the pediatric mortality index of 3 and mortality. Indicators of the pediatric mortality index of 3 were assessed by calibration and discrimination. Calibration assessed the pediatric mortality index of 3 at various mortality risks using the standardized mortality rate (SMR) and Pearson's goodness-of-fit test (chi-squared test). The study [23] evaluated the impact of health-related quality of life on the use of health services using four different scoring data models. Health-related quality of life was measured using a brief six-dimensional instrument and a functional assessment of colon cancer therapy, while health service use was measured by the number of monthly clinical consultations and the number of monthly hospitalizations. Goodness-of-fit statistics (Pearson's chi-squared test, Akaike information criterion and Bayesian tests) were used to determine the best model. In [24], a cross-sectional diagnostic study was described. 83 medical records of patients with suspected heart failure admitted to the emergency and internal medicine department of the Ramiro Priale Priale National Hospital were examined. Pearson's chi-squared test was used to analyze categorical variables and ANOVA was used for continuous variables. P-values < 0.05 were considered significant.

Kolmogorov test

Kolmogorov goodness-of-fit test is designed to test the hypothesis that the sample belongs to some distribution law, i.e. to check that the empirical distribution corresponds to the expected model.

In this test, the maximum value of the absolute difference between the empirical distribution function $F_n(x)$ and the corresponding theoretical distribution function $d = max|F_n(x) - F(x)|$ is a measure of difference between theoretical and empirical distributions. This random variable is denoted as $\lambda = D\sqrt{n}$ and is called *Kolmogorov goodness-of-fit λ-test*.

Application of Kolmogorov test

1. Arrange the results of observations in ascending order: $x_1 \leq x_2 \leq \cdots \leq x_n$ or represent them as an interval variational array.

2. Calculate the empirical relative frequencies for each rank by the formula:

$$f_j = \frac{m_j}{n} \quad (6)$$

3. Determine the values of the empirical distribution function $F_n(x)$ by calculating the accumulated empirical relative frequencies by the formula:

$$\sum f_j = \sum f_{j-1} + f_j \quad (7)$$

where $\sum f_i$ is the relative frequency accumulated in the previous ranks; j is the order number of the rank;

The obtained values $\sum f_j$ is empirical distribution function.

4. Determine the corresponding values of the assumed theoretical distribution function by counting the accumulated theoretical relative frequencies for each rank by the formula:

$$\sum f_j^{theor} = \sum f_j^{theor} + f_j^{theor} \quad (8)$$

where $\sum f_j^{theor}$ is the theoretical relative frequency accumulated in the previous ranks.

5. Calculate the absolute differences between the empirical and theoretical accumulated relative frequencies for each rank. Designate them as d.

6. Determine the largest absolute difference d_{max}.

7. Using the table of Kolmogorov test critical values [9, 10, 11, 12], for a given significance level α and a number of observations n, determine the critical value d_{cr}.

If $n > 100$, then d_{cr} is calculated by the formula:

$$d_{cr} = \begin{cases} \frac{1.36}{\sqrt{n}} & for\ \alpha \leq 0.05 \\ \frac{1.63}{\sqrt{n}} & for\ \alpha \leq 0.01 \end{cases} \quad (9)$$

If $d_{max} \geq d_{cr}$, then the null hypothesis is rejected: differences between distributions are significant.

If $d_{max} < d_{cr}$, then it is considered that there is no reason for rejecting the null hypothesis, i.e. the difference between the empirical and theoretical distribution function is not significant.

Limitations of test

Ranks should be arranged in ascending order.

Example. When weighing the fattened young cattle (103 animals) delivered to the meat processing plant, the following primary (raw) array was obtained according to live weight (kg):

413	454	419	412	427	435	404	430	421	399	414	386
428	441	397	417	418	423	420	416	407	427	428	417
398	424	419	401	424	411	426	380	419	406	410	409
416	410	403	426	407	400	423	425	394	432	409	418
419	388	423	434	402	431	405	436	405	424	405	412
413	444	392	411	428	394	433	395	433	420	430	398
437	422	394	416	424	434	407	443	406	422	414	429
417	406	419	429	406	388	421	415	417	394	431	411
422	410	432	409	439	421	414					

Determine whether the data obtained are normally distributed or not at a significance level $\alpha \leq 0.05$.

Solution. Let's rearrange the primary array into the variational array (Table 8).

Table 8. Variational array by the live weight of young cattle when delivered to a meat processing plant

W	380-389	390-399	400-409	410-419	420-429	430-439	440-449	450-459	Sum
f	4	10	16	30	26	13	3	1	n=103

Let's determine empirical relative frequencies for each rank by the formula:

$$f_j = \frac{m_j}{n},$$

where m_j is the frequency of a given number of points, n is the total number of points appearances.

$$f_1 = \frac{m_1}{n} = \frac{4}{103} = 0.039$$
$$f_2 = \frac{m_2}{n} = \frac{10}{103} = 0.097$$
$$f_3 = \frac{m_3}{n} = \frac{16}{103} = 0.155$$
$$f_4 = \frac{m_4}{n} = \frac{30}{103} = 0.291$$
$$f_5 = \frac{m_5}{n} = \frac{26}{103} = 0.252$$
$$f_6 = \frac{m_6}{n} = \frac{13}{103} = 0.126$$
$$f_7 = \frac{m_7}{n} = \frac{3}{103} = 0.029$$
$$f_8 = \frac{m_8}{n} = \frac{1}{103} = 0.0097$$

Let's determine accumulated empirical relative frequencies by the formula:

$$\sum f_j = \sum f_{j-1} + f_j$$

where $\sum f_i$ is the relative frequency accumulated in the previous ranks; j is the order number of the rank.

$$\sum f_1 = f_1 = 0.039$$

$$\sum f_{1+2} = \sum f_1 + f_2 = 0.039 + 0.097 = 0.136$$

$$\sum f_{1+2+3} = \sum f_{1+2} + f_3 = 0.136 + 0.155 = 0.291$$

$$\sum f_{1+2+3+4} = \sum f_{1+2+3} + f_4 = 0.291 + 0.291 = 0.582$$

$$\sum f_{1+2+3+4+5} = \sum f_{1+2+3+4} + f_5 = 0.582 + 0.252 = \\ = 0.834$$

$$\sum f_{1+2+3+4+5+6} = \sum f_{1+2+3+4+5} + f_6 = \\ = 0.834 + 0.126 = 0.960$$

$$\sum f_{1+2+3+4+5+6+7} = \sum f_{1+2+3+4+5+6} + f_7 = \\ = 0.960 + 0.029 = 0.989$$

$$\sum f_{1+2+3+4+5+6+7+8} = \sum f_{1+2+3+4+5+6+7} + f_8 = \\ = 0.989 + 0.0097 = 0.9987 \approx 1$$

Let's determine theoretical relative frequencies for each rank. For the 1st rank, the theoretical relative frequency is calculated by the formula:

$$f_1^{theor} = \frac{1}{k},$$

where k is the number of ranks ($k = 8$).

$$f_1^{theor} = \frac{1}{k} = \frac{1}{8} = 0.125.$$

This theoretical relative frequency applies to all ranks.

Let's determine accumulated theoretical relative frequencies.

$$\sum f_1^{theor} = f_1^{theor} = 0.125;$$

$$\sum f_{1+2}^{theor} = \sum f_1^{theor} + f_2^{theor} = 0.125 + 0.125 = \\ = 0.250$$

$$\sum f_{1+2+3}^{theor} = \sum f_{1+2}^{theor} + f_3^{theor} = 0.250 + 0.125 = \\ = 0.375$$

$$\sum f_{1+2+3+4}^{theor} = \sum f_{1+2+3}^{theor} + f_4^{theor} = 0.375 + 0.125 = \\ = 0.500$$

$$\sum f_{1+2+3+4+5}^{theor} = \sum f_{1+2+3+4}^{theor} + f_5^{theor} = \\ = 0.500 + 0.125 = 0.625$$

$$\sum f_{1+2+3+4+5+6}^{theor} = \sum f_{1+2+3+4+5}^{theor} + f_6^{theor} = \\ = 0.625 + 0.125 = 0.750$$

$$\sum f_{1+2+3+4+5+6+7}^{theor} = \sum f_{1+2+3+4+5+6}^{theor} + f_7^{theor} = \\ = 0.750 + 0.125 = 0.875$$

$$\sum f_{1+2+3+4+5+6+7+8}^{theor} = \sum f_{1+2+3+4+5+6+7}^{theor} + f_8^{theor} = \\ = 0.875 + 0.125 = 1$$

Calculate the absolute differences between the accumulated empirical and theoretical frequencies:

$$d_1 = \left|\sum f_1 - \sum f_1^{theor}\right| = |0.039 - 0.125| = 0.086;$$

$$d_2 = \left|\sum f_2 - \sum f_2^{theor}\right| = |0.136 - 0.250| = 0.114;$$

$$d_3 = \left|\sum f_3 - \sum f_3^{theor}\right| = |0.291 - 0.375| = 0.084;$$

$$d_4 = \left|\sum f_4 - \sum f_4^{theor}\right| = |0.582 - 0.500| = 0.082;$$

$$d_5 = \left|\sum f_5 - \sum f_5^{theor}\right| = |0.834 - 0.625| = 0.209;$$

$$d_6 = \left|\sum f_6 - \sum f_6^{theor}\right| = |0.960 - 0.750| = 0.210;$$

$$d_7 = \left|\sum f_7 - \sum f_7^{theor}\right| = |0.989 - 0.875| = 0.114;$$

$$d_8 = \left|\sum f_8 - \sum f_8^{theor}\right| = |1 - 1| = 0.$$

The results are shown in Table 9.

Table 9. Calculation results

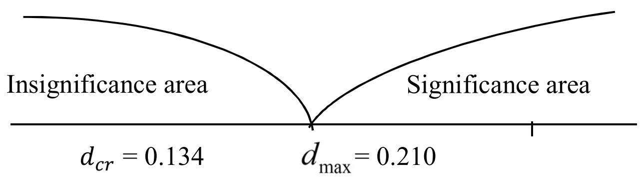
Number of points	Empirical frequency	Empirical relative frequency	Accumulated empirical relative frequency	Accumulated theoretical relative frequency	Difference
1	4	0.039	0.039	0.125	0.086
2	10	0.097	0.136	0.250	0.114
3	16	0.155	0.291	0.375	0.084
4	30	0.291	0.582	0.500	0.082
5	26	0.252	0.834	0.625	0.209
6	13	0.126	0.960	0.750	0.210
7	3	0.029	0.989	0.875	0.114
8	1	0.0097	1	1	0
Sums	103	1			

Let's determine the largest absolute difference d_{max} (yellow color cell).

Since in this problem $n > 100$, then the critical value d_{cr} is calculated by the formula (9) for a significance level $\alpha \leq 0.05$:

$$d_{cr} = \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{103}} = 0.134$$

Let's plot the axis of significance:



Since $d_{max} \geq d_{cr}$, then the null hypothesis is rejected, i.e. the empirical distribution for the live weight of cattle delivered to a meat processing plant differs from the normal (uniform) distribution.

Let's give examples for the use of Kolmogorov test in scientific research. The article [25] analyzed the growth of rice, wheat and common food grains in India for the period from 1950 to 2019. The distribution was assessed using Kolmogorov test. It was found that the availability of rice (70.05 kg/year), wheat (70.73 kg/year) and total grains (182.96 kg/year) will decrease in 2021 compared to this year. The article [26] analyzed a questionnaire survey of 227 respondents regarding purchasing preferences for organic food in Slovakia. To achieve the goal and provide a deeper analysis of the results, 3 assumptions and 5 hypotheses were made. According to the survey results, 65% of respondents buy organic products, of which 39% buy organic products at least once a week. Up to 98% of respondents have already heard about the concept of organic food and know what it means. 37% of respondents buy mostly organic fruits and vegetables; 18% of respondents buy mostly organic meat and meat products, and 13% of respondents prefer organic dairy products. The most preferred place to buy organic products are specialized stores (36%); buying organic products directly from the manufacturer is the most popular way for 29% of respondents; hypermarkets and supermarkets are a favorite place to buy organic products for 19% of respondents; and 12% of respondents buy organic products mainly in farmers' markets. Only 4% of respondents prefer another way to buy organic products. The quality of organic products and the absence of pesticides are the most important criteria for purchasing organic products (36%). The results of the study were evaluated using the goodness-of-fit chi-squared test and Kolmogorov test, and the following conclusion was made: there is a difference in the preferences of the respondents. In Slovakia, there is a relationship between consumer preferences for organic food and traditional food, and there is a strong preference to buy organic food. The aim of the study [27] was to present a correct model for probability distribution based on data obtained from surveys on the temperature of food storage in household refrigerators at home. The temperature in household refrigerators was determined as a risk factor for foodborne disease outbreaks for microbial risk assessment. Temperature was measured by visiting 139 homes directly with a data logger from May to September 2009. The overall average temperature for all refrigerators participating in the survey was 3.53 ± 2.96 °C, with 23.6% of refrigerators having temperatures above 5 °C. Probability distributions were generated from the measured temperature data. Statistical ranking was determined by Kolmogorov goodness-of-fit test or Anderson-Darling test to determine the appropriate probability distribution model. This result showed that the LogLogistic distribution (-10.407, 13.616, 8.6107) was the most appropriate for the microbial risk assessment model.

The aim of the work [28] was to study the strong Markov property for stochastic differential equations controlled by G-Brownian motion (G-SDE). First, the authors extended the conditional G-expectancy of deterministic time to optional points of time. The strong Markov property for the G-SDE was obtained using Kolmogorov tightness criterion. The article [29] considers the process of the defect appearance in the body of a workpiece obtained by casting. The medium with many randomly distributed discontinuities was schematically a regular structure formed by a set of elements in the form of a regular tetrahedron with spherical depressions at the vertices. The proposed technique makes it possible to create a model of a continuous homogeneous medium that is equivalent in its deformation properties to the original discontinuous material. Using this approach, a power approximation of the extension curve for a model medium was obtained. The rupture of the material was fixed using Kolmogorov plastic deformation test. This test was used in the evaluation of the limit state of the valve chamber under operating conditions.

Nonparametric tests for homogeneity

Hypotheses of homogeneity are hypotheses assuming that the samples under study are taken from the same general population.

Let there be two independent samples with sizes n_1 and n_2 obtained from populations with unknown theoretical distribution functions $F_1(x)$ and $F_2(x)$. Hypotheses are stated:

H_0: Empirical distribution 1 does not differ from empirical distribution 2, i.e. $F_1(x) = F_2(x)$.

H_1: Empirical distribution 1 differs from empirical distribution 2, i.e. $F_1(x) \neq F_2(x)$.

Pearson's chi-squared test for homogeneity

Pearson's chi-squared test may be used to evaluate the homogeneity of two or more independent samples, i.e. to test the hypothesis that there are no differences between two and more empirical distributions of the same indicator. Source data should be presented in the form of Table 10:

Table 10. Source data template (cross-tab table or contingency table)

Empirical frequencies		Indicator ranking					Sum
		1	…	*j*	…	*k*	
Ranks of the indicator	1						
	…						
	i						
	…						
	c						
Sum							

Such tables are called cross-tab tables or contingency tables.

The algorithm for calculating Pearson's chi-squared test is the same as for Pearson's goodness-of-fit test (see above), but for each cell of the i^th^ row and j^th^ column, its own theoretical frequency is determined by the formula:

$$m_{ij}^{theor} = \frac{\sum_i m_{ij} \cdot \sum_j m_{ij}}{N}, \quad (10)$$

where N is the sum of frequencies of the entire contingency table; $\sum_i m_{ij}$ is the sum of frequencies in all cells of the i^{th} row; $\sum_j m_{ij}$ is the sum of frequencies in all cells of the j^{th} column.

Pearson's chi-squared test is calculated by the formula:

$$\chi_{exp}^2 = \sum_{i=1}^{c} \sum_{j=1}^{k} \frac{\left(m_{ij} - m_{ij}^{theor}\right)^2}{m_{ij}^{theor}}. \quad (11)$$

The number of degrees of freedom is calculated by the formula:

$$df = (k-1) \cdot (c-1) \quad (12)$$

where c is the number of ranks for the indicator (number of compared distributions).

If the number of degrees of freedom is equal to 1, i.e., if the indicator only takes two values, the adjusting for continuity is needed. The adjusting for continuity is applied under the following conditions:

1) when the empirical distribution is compared to the uniform distribution, and the number of indicator rankings $k = 2$, and the number of degrees of freedom $v = k - 1 = 1$.

2) when two empirical distributions are compared, and k=2, i.e. number of rows and number of columns is both equal to 2 and $v = (k-1) \cdot (c-1) = 1$.

In these cases, it is necessary to reduce the absolute difference $\left|m_{ij} - m_{ij}^{theor}\right|$ by 0.5 prior to squaring. χ_{exp}^2 is calculated by the formula:

$$\chi_{exp}^2 = \sum_{i=1}^{c} \sum_{j=1}^{k} \frac{\left(\left|m_{ij} - m_{ij}^{theor}\right| - 0.5\right)^2}{m_{ij}^{theor}} \quad (13)$$

Example. During the survey, high school students were asked which of the three possible areas of education (mathematics, natural sciences or human sciences) they would prefer in the future. Among the respondents were both young males and young females [30]. The data are summarized in Table 11.

Table 11. Given data for the problem

Empirical frequencies		Indicator ranking		
		Mathe-matics	Natural sciences	Human sciences
Ranks of the indicator	Young males 1	18	10	3
	Young females 2	10	9	15

Such table is called a cross-tab table with size of 2 × 3.

Is it possible to state that at a significance level $\alpha \leq 0.05$ the preference for one or another area of education is somehow related to the gender factor?

Solution. Let's state the hypotheses:

H_0: distribution of preferences for the area of education in young males and young females is not significantly different from the random distribution.

H_1: distribution of preferences for the area of education in young males and young females is significantly different from the random distribution.

In Table 12 sums of frequencies are calculated by rows and columns.

Table 12. Intermediate cross-tab 2 × 3 calculations

Empirical frequencies		Indicator ranking			Sum
		Mathe-matics	Natural sciences	Human sciences	
Ranks of the indicator	Young males 1	18	10	3	31
	Young females 2	10	9	15	34
Sum		28	19	18	65

For each of the cells, a special theoretical frequency related only to this cell should be calculated by the formula:

$$m_{ij}^{theor} = \frac{\sum_i m_{ij} \cdot \sum_j m_{ij}}{N}.$$

There are 65 frequencies in total, of which 28 frequencies correspond to mathematics, 19 frequencies correspond to natural sciences, and 18 frequencies correspond to human sciences. The proportion of each education area is 28/65, 19/65, 18/65, respectively. In all rows, mathematics should be 28/65 of all the answers, natural sciences should be 19/65, and human sciences should be 18/65. Knowing the sums of frequencies for each row, you can calculate the theoretical frequencies for each cell.

$m_{11}^{theor} = \frac{31 \cdot 28}{65} = 13.35;$

$m_{12}^{theor} = \frac{31 \cdot 19}{65} = 9.06;$

$m_{13}^{theor} = \frac{31 \cdot 18}{65} = 8.58;$

$m_{21}^{theor} = \frac{34 \cdot 28}{65} = 14.65;$

$m_{22}^{theor} = \frac{34 \cdot 19}{65} = 9.94;$

$m_{23}^{theor} = \frac{34 \cdot 18}{65} = 9.42.$

Let's complete Table 13.

Table 13. Calculation results

Rank — indicator ranking	m_{ij}	m_{ij}^{theor}	$m_{ij} - m_{ij}^{theor}$	$\left(m_{ij} - m_{ij}^{theor}\right)^2$	$\frac{\left(m_{ij} - m_{ij}^{theor}\right)^2}{m_{ij}^{theor}}$
Young males — mathematics	18	13.35	4.65	21.59	1.62
Young males — natural sciences	10	9.06	0.94	0.88	0.10
Young males — human sciences	3	8.58	-5.58	31.19	3.63
Young females — mathematics	10	14.65	-4.65	21.59	1.47
Young females — natural sciences	9	9.94	-0.94	0.88	0.09
Young females — human sciences	15	9.42	5.58	31.19	3.31

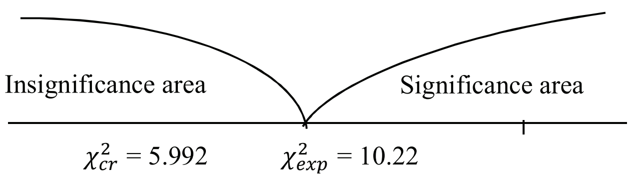
$\chi^2_{exp} = 1.62 + 0.10 + 3.63 + 1.47 + 0.09 + 3.31 = 10.22.$

The number of degrees of freedom is calculated by the formula:

$$v = (k - 1) \cdot (c - 1) = (3 - 1) \cdot (2 - 1) = 2$$

Using the table of critical values [9, 10, 11, 12], χ^2 distributions for $v = 2$ and $\alpha \leq 0.05$ $\chi^2_{cr} = 5.992$.

Let's plot the axis of significance:



Since $\chi^2_{exp} > \chi^2_{cr}$, the null hypothesis should be rejected and the alternative hypothesis should be accepted, i.e. the dependence of preference in choosing a further education on the gender of the respondent was proved.

In the studies [31-35], chi-squared test was used. The study [31] examined the association of interleukin-6 (IL-6) (IL-6-174G/C), transforming growth factor-beta 1 (TGF-beta1-29C/T) and calmodulin 1 gene. 16C/T-polymorphism (CALM1-16C/T) was clinically determined in Pakistani patients with osteoarthritis and corresponding control group. The study included 295 subjects, including 100 patients with osteoarthritis, 105 patients with predisposition to osteoarthritis and 90 patients from the control group. The study design was based on biochemical analysis of osteoarthritis using hyaluronic acid serum enzyme-linked immunosorbent assay and genetic analysis based on PCR with an amplification-resistant mutation system. Allele probabilities were statistically estimated using Pearson's chi-squared test. The authors [32] studied the role and interaction of proteins involved in the control and stimulation of neurotransmission in predisposition to migraine. The study was performed on 183 migraineurs (148 women and 35 men) and 265 non-migraine controls (202 women and 63 men). Labeling of single nucleotide polymorphisms of neurexin was carried out to assess the association between neurexin and predisposition to migraine. Chi-squared test was used to compare allele frequencies in test cases and controls, and odds ratios were estimated with 95% confidence intervals. The authors [33] present a retrospective crossover observational study of the epidemiological profile of all dengue cases confirmed and reported to the Minister of Health in Pernambuco between 2015 and 2017. The data include all municipalities of Pernambuco with the exception of Fernando de Noronha. People infected with dengue were classified according to the type of dengue fever (without and with the symptoms or severe dengue), age, sex, ethnicity, and intermediate geographic region of residence (Recife, Caruaru, Serra Talhada, or Petrolina). The distribution of cases by years was estimated using chi-squared test. The aim of the study [34] was to evaluate eating behavior, health-related and nutrition-related problems among students with symptoms of orthorexia nervosa. The participants were 1120 college students from seven universities in Poland studying health-related (n=547) and other specialties (n=573). Students were examined with ORTO-15 test, the health problems scale and the food intake frequencies questionnaire. Then, based on principal component analysis, eight nutrition patterns were derived ("sweets and snacks", "legumes and nuts", "fruits and vegetables", "refined breads and animal fats", "dairy products and eggs", "fish", "meat", "fruit and vegetable juices"). Pearson's correlation, Pearson's chi-squared test, Student t-test and one-sided ANOVA were used for further analysis. In the work [35], the authors studied the potential roles and mechanisms of si-STOML2 (stomatin-like protein 2) in the migration and invasion of human hepatoma LM3 cells. Stomatin-like protein 2 expression levels in tissues and cells were separately analyzed by quantitative real-time PCR (qRT-PCR) and Western blotting. Cell viability, migration and invasion were assessed using the cell count-8 kit, wound healing and transwell assay kit, respectively. mRNA and various protein factors levels were separately measured by qRT-PCR and Western blotting. The correlation analysis between the expression of stomatin-like protein 2 and the clinical/pathological features of liver cancer patients was assessed using the chi-squared test.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov test statistics is the following:

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot max|F_1(x) - F_2(x)|, \qquad (14)$$

where $F_1(x)$ and $F_2(x)$ are empirical distribution functions from two samples with sizes n_1 and n_2. Let's assume that the functions $F_1(x)$ and $F_2(x)$ are continuous.

Application of Kolmogorov-Smirnov test

1. Arrange the results of observations in ascending order: $x_1 \leq x_2 \leq \cdots \leq x_n$ or represent them as an interval variational array.

2. Calculate the empirical relative frequencies for each rank for distribution 1 by the formula:

$$f_{1j} = \frac{m_{1j}}{n_1}$$

where m_{1j} is the empirical frequency in the given rank; n_1 is the number of observations in the sample.

3. Calculate the empirical relative frequencies for each rank for distribution 2 by the formula:

$$f_{2j} = \frac{m_{2j}}{n_2},$$

where m_{2j} is the empirical frequency in the given rank; n_2 is the number of observations in the sample.

4. Calculate the accumulated empirical relative frequencies for distribution 1 by the formula:

$$\sum f_{1j} = \sum f_{1j-1} + f_{1j}$$

where $\sum f_{1j-1}$ is the relative frequency accumulated in the previous ranks; j is the order number of the rank; f_{1j} is the relative frequency of the given rank.

5. Calculate the accumulated empirical relative frequencies for distribution 2 by the same formula.

$$\sum f_{2j} = \sum f_{2j-1} + f_{2j}$$

where $\sum f_{2j-1}$ is the relative frequency accumulated in the previous ranks; f_{2j} is the relative frequency of the given rank.

6. Calculate the absolute differences between the accumulated relative frequencies for each rank. Designate them as d. Determine the largest absolute difference d_{max}.

7. Calculate λ'_{exp} by the formula:

$$\lambda'_{exp} = d\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}_{max} \quad (15)$$

where n_1 is the number of observations in the first sample; n_2 is the number of observations in the second sample.

8. Using the table of critical values [9, 10, 11, 12], for a given significance level α, determine λ_{cr}. If $\lambda'_{exp} \geq \lambda_{cr}$, then the differences between the distributions are significant. If $\lambda'_{exp} < \lambda_{cr}$, then the differences between the distributions are not significant.

Limitations of Kolmogorov-Smirnov test

1. When comparing two empirical distributions, it is necessary that $n_1, n_2 \geq 50$.

2. Ranks must be arranged in ascending or descending order by some indicator. We cannot accumulate frequencies by the ranks that differ only qualitatively and do not represent a scale of order.

Example. To evaluate the effectiveness of a drug, one group of subjects was given a test drug tested on animals, and the other group of subjects was given a placebo (a physiologically inert substance, the positive therapeutic effect of which is associated with the patient's subconscious psychological expectation). Table 14 represents data on the number of occurrences of influenza symptoms over a two-year period in the group taking prophylactic drug at the beginning of the period and in the group taking placebo [12].

Table 14. Given data for the problem

Number of diseases	Number of patients taking the drug	Number of patients taking placebo
	m_{1j}	m_{2j}
0	32	26
1	26	30
2	15	11
3	6	14
4 and more	6	19
Sum	85	100

Can we state that at a significance level $\alpha \leq 0.05$ the effect of the drug is sufficiently greater than of placebo?

Solution. Let's state the hypotheses:

H_0: Empirical distribution 1 differs from empirical distribution 2, i.e. the effect of the drug significantly exceeds the effect of the placebo.

H_1: Empirical distribution 1 does not differ from empirical distribution 2, i.e. the effect of the drug does not significantly exceed the effect of the placebo.

Let's determine empirical relative frequencies for each rank for sample 1 (first test) by the formula:

$$f_{1j} = \frac{m_{1j}}{n_1}$$

$$f_{11} = \frac{m_{11}}{n_1} = \frac{32}{85} = 0.3765;$$

$$f_{12} = \frac{m_{12}}{n_1} = \frac{26}{85} = 0.3059$$

etc.

The results of the calculations are represented in Table 15.

Table 15. Calculation results

Number of diseases	Empirical frequencies		Empirical relative frequencies		Accumulated empirical relative frequencies		Difference $\left\lvert\sum f_{1j} - \sum f_{2j}\right\rvert$
	m_{1j}	m_{2j}	f_{1j}	f_{2j}	$\sum f_{1j}$	$\sum f_{2j}$	
0	32	26	0.3765	0.2600	0.3765	0.2600	0.1165
1	26	30	0.3059	0.3000	0.6824	0.5600	0.1224
2	15	11	0.1765	0.1100	0.8588	0.6700	0.1888
3	6	14	0.0706	0.1400	0.9294	0.8100	0.1194
4 and more	6	19	0.0706	0.1900	1.0000	1.000	0
Sum	85	100	1	1			

Let's determine empirical relative frequencies for each rank for sample 2 (second test) by the formula:

$$f_{2j} = \frac{m_{2j}}{n_2}$$

$$f_{21} = \frac{m_{21}}{n_2} = \frac{26}{100} = 0.2600;$$

$$f_{21} = \frac{m_{22}}{n_2} = \frac{30}{100} = 0.3000$$

etc.

Let's calculate the accumulated empirical relative frequencies for sample 1 by the formula:

$$\sum f_{1j} = \sum f_{1j-1} + f_{1j}.$$

$$\sum f_{11} = f_{11} = 0.3765$$

$$\sum f_{12} = \sum f_{11} + f_{12} = 0.3765 + 0.3059 = 0.6824$$

etc.

Let's calculate the accumulated empirical relative frequencies for sample 2 by the same formula:

$$\sum f_{2j} = \sum f_{2j-1} + f_{2j}.$$

$$\sum f_{21} = f_{21} = 0.2600$$

$$\sum f_{22} = \sum f_{21} + f_{22} = 0.2600 + 0.3000 = 0.5600$$

etc.

Let's determine the absolute differences between the accumulated empirical relative frequencies by the formula:

$$d_j = \left|\sum f_{1j} - \sum f_{2j}\right|.$$

$$d_1 = |\sum f_{11} - \sum f_{21}| = |0.3765 - 0.2600| = 0.1165;$$

$$d_2 = \left|\sum f_{12} - \sum f_{22}\right| = |0.6824 - 0.5600| = 0.1224;$$

etc.

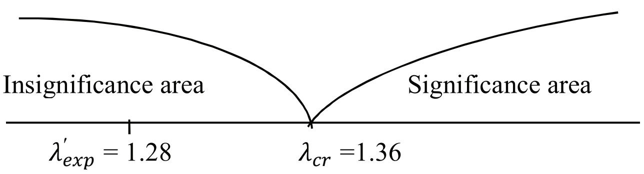
From Table 15, let's determine the largest absolute difference d_{max}. This is $d_{max} = 0.1888$ (highlighted in yellow).

Let's calculate λ'_{exp}:

$$\lambda'_{exp} = d_{max}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} = 0.1888 \cdot \sqrt{\frac{85 \cdot 100}{85 + 100}} \approx 1.28$$

Using the table of critical values [9, 10, 11, 12], for a given significance level $\alpha \leq 0.05$, let's determine $\lambda_{cr} = 1.36$.

Let's plot the axis of significance:



Since $\lambda'_{exp} < \lambda_{cr}$, then the null hypothesis is not rejected, i. e. the effect of the drug significantly exceeds the effect of the placebo.

In the studies [36–44], Kolmogorov-Smirnov test was used. The study [36] aimed to determine the relationship between the management of household solid waste (HSW) and non-household solid waste (NHSW) (X variable) in Huancavelica County and municipal government (Y variable) in 2016. The population and sample were 12,249 and 140 people, respectively. The collected data were analyzed using Kolmogorov-Smirnov test. The paper [37] represents the results of physicochemical and rheological studies of wet foams obtained from hen egg albumin with the addition of xanthan gum and/or arabic gum using the batch method. Physicochemical analysis included determination of foam density, gas phase volume fraction, overrun, stability and distribution of gas bubbles suspended in liquid. The study of hydrocolloids effect on the distribution of gas bubbles was based on standard descriptive parameter estimation and the use of the nonparametric Kolmogorov-Smirnov test. The study [38] evaluated the expression of basic fibroblast growth factor and the number of osteoblasts during orthodontic tooth movement after administration of *Bifidobacterium bifidum* probiotic in male Wistar rats. Orthodontic tooth movement was carried out using a nickel titanium coil spring with a force of 10 g applied between the first incisor and the maxillary first molar of a Wistar rat. Samples were then removed on days 3, 7 and 14. Maxillary tissue was isolated for immunohistochemical examination and hematoxylin-eosin staining. All data were analyzed using an independent t-test (p<0.05), which was implemented based on Kolmogorov-Smirnov test and Levene test (p>0.05). In the study [39], it was proposed to use a queuing network to simulate the diffusion of molecules in accordance with Fick's law. The proposed model was tested using Kolmogorov-Smirnov test to compare the results obtained from the simulation with the theoretical standard deviations obtained based on Einstein-Smoluchowski test. The article [40] develops two different approaches to simulative diagnostic procedures for models of Markov chains based on bands. The first approach uses a formal test based on Kolmogorov-Smirnov or Cramer-von Mises statistics.

The article [41] shows a study to determine the effect of consumption of roasted soyabeans and textured soy protein on the clinical and metabolic status of older women with borderline metabolic syndrome parameters. A randomized single-blinded controlled clinical trial included 75 women aged over 60 years with a diagnosis of metabolic syndrome based on ATP III. Participants were randomly assigned to three groups of 25 people who consumed for 12 weeks: 1) soyabeans; 2) textured soy protein; and 3) control diet. Fasting blood samples were taken at the beginning and end of the study to compare metabolic responses. Kolmogorov-Smirnov test, ANOVA, ANCOVA, paired t-test, and repeated measurements analysis of the generalized linear model were used to evaluate the study results. As a result of the study, it was found that nutrition and physical activity of the participants in the two groups did not differ significantly. After 12 weeks of intervention, the soyabean-treated participants showed significant reductions in total cholesterol (p <0.001), low-density lipoproteins, and very-low-density lipoproteins. Thus, short-term consumption of roasted soyabeans and textured soy protein improves lipid profile, markers of glucose intolerance and oxidative stress. Although roasted soyabeans were more effective than textured soy protein. Moderate daily intake of roasted soyabeans as a snack or textured soy protein as a food supplement for individuals with borderline metabolic syndrome parameters may be a safe and useful way to avoid disease progression. The work [42] was aimed at analyzing the consumption of sugar (sucrose) by the low-income population of Brazil. A cross-sectional descriptive study was conducted to evaluate typical customers of a popular restaurant (PR) in Brazil (Brazilian food aid program for low-income people). In the final sample, 1232 adult PR clients were interviewed. Exclusion criteria were pregnant women, diabetics, or people on any special sucrose-restricted diet. People were enrolled at lunchtime while they waited in line to pick up food. The invitation to participate were made to the first person in the queue, then to the 15th person, and so on until the sampling was complete. A three-day, 24-hour review was used to estimate sugar intake. Sociodemographic and anthropometric data were collected so that client profiles could be compiled. To characterize the sample, a statistical analysis of descriptive data (frequency, mean value, median, percentage and standard deviation) was carried out. Statistical normality tests (Kolmogorov-Smirnov test) were performed for all analyzes to test the assumptions of the statistical tests. The average total energy value (TEV) for the estimated three-day period was 1980.23 ± 726.75 kcal. A statistically significant difference was found between income groups ($p < 0.01$). The northern and northeastern regions have the lowest median income in Brazil, statistically different from the south ($p < 0.01$) and southeast ($p < 0.01$) regions. The northern region showed the lowest sugar consumption from industrial products, in contrast to the northeast ($p = 0.007$), southeast ($p = 0.010$) and south ($p = 0.043$) regions. The north region also has the lowest consumption of home-cooked foods among other regions ($p < 0.001$). Total sugar (sucrose) intake did not differ with body mass index ($p = 0.321$). There was no significant difference in sugar (sucrose) intake over the three days ($p = 0.078$). The addition of sugar (sucrose) contributed to 36.7% of all sugar (sucrose) intake, and sweetened beverages contributed to 22.53% of all sugar (sucrose) intake. Home-cooked products accounted for 20.06% of sugar (sucrose) consumption and industrial products accounted for 22.53% of sugar (sucrose) consumption. Thus, consumption of free sugar (sucrose) is still the largest contributor to total sugar (sucrose) intake, followed by sweetened beverages, especially on weekends. The average percentage of sugar (sucrose) intake exceeds the World Health Organization's recommendation of consuming less than 5% of total energy from sugars. Because this population group has a high percentage of overweight and obesity, sugar (sucrose) consumption may increase health outcomes by increasing public health costs.

The article [43] presents a study assessing the consumption of meat and products obtained from hunting by the consumer population. To do this, a survey was conducted on the frequency of eating meat from the four most representative species in Spain, two large species: wild boar (Sus scrofa) and red deer (Cervus elaphus), as well as two small species: rabbit (Oryctolagus cuuniculus) and red partridge (Alectis rufa), as well as processed meat products (salami sausages) made from the meat of these animals. The survey was conducted on 337 habitual consumers of these products. The overall average per capita meat consumption in this population group is 6.87 kg of meat per year or 8.57 kg of meat per year if processed meat products are also considered. The consumption of rabbit, red partridge, red deer, and wild boar was 1.85, 0.82, 2.28, and 1.92 kg per person per year, respectively. Using probabilistic methods, distributions of meat consumption frequencies were estimated for each of the studied hunted species. The distribution of consumption frequencies was statistically proven by the chi-squared test and Kolmogorov-Smirnov test.

The aim of the study [44] was to describe the nutritional value of food and non-alcoholic beverages advertised in a lineup for children compared to a general lineup on two national private free-to-air television channels in Colombia. The methods chosen were: a cross-sectional descriptive study. The recording was made in July 2012 for four days randomly selected from 6:00 am to 12:30 pm. Nutrient content has been classified according to the Food Standards Agency nutrition profile criteria for nutrients indicating risk, the Pan-American Health Organization for trans fats, and Colombian Resolution 333 dated 2011, which classifies foods as a source of protective nutrients. Descriptive statistics was used, i. e. Kolmogorov-Smirnov test to establish normality and Pearson's chi-squared test to compare variables. The p value of < 0.05 was taken into account. As a result, the following data were obtained: 1560 advertising clips were shown in 52 hours of recording, of which 23.3% (364) clips advertised food and drinks, of which 56.3% were shown in a lineup for children. In terms of nutritional value, in the lineup for children, a high percentage of foods and non-alcoholic beverages classified as "rich" in sugar, sodium, saturated fats (69.0%, 56.0%, 57.1%) was noted, compared with the general lineup. In contrast, the percentage of foods and non-alcoholic beverages classified as "rich" in total fat content was higher in the general lineup (70.4% vs. 29.6%, respectively). Thus, in the lineup for children, a large impact of food and non-alcoholic beverage advertising was observed characterized by a high content of high-risk nutrients and a low content of foods.

Thus, the possibilities of nonparametric statistics are shown in the analysis of seemingly incomparable results.

Conclusion

The second part discusses nonparametric tests for testing hypotheses of distribution type and nonparametric tests for testing hypotheses of sampling homogeneity. Pearson's chi-squared test, Kolmogorov-Smirnov test, Kolmogorov test were reviewed. Using examples, the use of tests was discussed, and their capabilities and limitations were evaluated. Based on the literature review, brief descriptions of studies in which methods of nonparametric statistics have been successfully applied are given. These tests may be used when comparing descriptive characteristics, which allows statistical processing of the results, for example, tasting evaluation of the product or morphological analysis of the section. Nonparametric methods also allow to compare groups with different unequal number of parameters.

REFERENCES

1. Prokhorov, Yu.V. (2013). Nonparametric methods: Encyclopedia. Moscow: The Great Russian Encyclopedia. 2013. (In Russian)
2. Tomkins-Lane, C.C. (2006). An introduction to non-parametric statistics for health scientists. *University of Alberta Health Sciences Journal*, 3(1), 20–26.
3. Pett, M.A. (1997). Nonparametric statistics for health care research. London, Thousand Oaks, New Delhi: SAGE Publications, 1997.
4. Orlov, A.I. (2015). Current status of nonparametric statistics. *Polythematic Online Scientific Journal of Kuban State Agrarian University,* 106, 239–269. (In Russian)
5. Nikitina, M.A., Chernukha, I.M. (2021). Methods for nonparametric statistics in scientific research. Overview. Part 1. *Theory and Practice of Meat Processing*, 6(2), 151–162. https://doi.org/10.21323/2414-438X-2021-6-2-151-162
6. Kobzar, A.I. (2006). Applied mathematical statistics. Moskow: Fizmatlit, 2006. (In Russian)
7. Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175. https://doi.org/10.1080/14786440009463897
8. Chernoff, H., Lehmann, E. L. (1954). The Use of maximum likelihood estimates in tests for goodness of fit. *The Annals of Mathematical Statistics*, 25(3), 579–586. https://doi.org/10.1214/aoms/1177728726
9. Gubler, E.V., Genkin, A.A. (1973). Application of nonparametric statistical criteria in biomedical research. Leningrad: Medicine, 1973. (In Russian)
10. Rosenbaum, S. (1954). Tables for a nonparametric test of location. *Annals of Mathematical Statistics*, 25 (1), 146–150. https://doi.org/10.1214/aoms/1177728854
11. Stepanov, V.G. (2019). Application of nonparametric statistical methods in agricultural biology and veterinary medicine research. St-Petersburg: Lan. 2019. (In Russian)
12. Edelbaeva, N.A., Lebedinskaya, O.G., Kovanova, E.S., Tenetova, E.P., Timofeev, A.G. (2019). Fundamentals of nonparametric statistics. Moscow: YUNITY-DANA. 2019. (In Russian)
13. Krupin, V.G., Pavlov, A.L., Popov, L.G. (2013). Higher mathematics. Probability theory, mathematical statistics, random processes. Moscow: Publishing House of the Moscow Power Engineering Institute, 2013. (In Russian)
14. Gmurman, V.E. (2004). A guide to solving problems in probability theory and mathematical statistics. Moscow: High School, 2004. (In Russian)
15. Petroczi, A., Naughton, D.P. (2009). Mercury, cadmium and lead contamination in seafood: A comparative study to evaluate the usefulness of Target Hazard Quotients. *Food and Chemical Toxicology*, 47(2), 298–302. https://doi.org/10.1016/j.fct.2008.11.007
16. Sinclair, J., Dillon, S., Bottoms, L. (2022). Perceptions, beliefs and behaviors of nutritional and supplementary practices in inflammatory bowel disease. *Sport Sciences for Health*. https://doi.org/10.1007/s11332-022-00901-8 (unpublished data)
17. Lee, J.H., Singh, R.K. (1993). Residence time distribution characteristics of particle flow in a vertical scraped surface heat exchanger. *Journal of Food Engineering*, 18(4), 413–424. https://doi.org/10.1016/0260-8774(93)90055-O
18. Skořepa, L., Pícha, K. (2016). Factors of purchase of bread — prospect to regain the market share? *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 64(3), 1067–1072. https://doi.org/10.11118/actaun201664031067
19. Lebedeva, E.V., Schastnyy, E.D., Nonka, T.G., Repin, A.N. (2021). Gender differences in self-reported social functioning of patients with chronic coronary artery disease and affective disorders. *Byulleten Sibirskoy Meditsiny*, 20(1), 75–82. https://doi.org/10.20538/1682-0363-2021-1-75-82
20. Romashkina, G.F., Khuziakhmetov, R.R. (2020). The risks of internet addiction: structure and characteristics of perception. Education and Science Journal, 22(8), 108–134. https://doi.org/10.17853/1994-5639-2020-8-108-134 (In Russian)
21. Snijder, R., Bosman, B., Stroosma, O., Agema, M. (2020). Relationship between mean volume voided and incontinence in children with overactive bladder treated with solifenacin: post hoc analysis of a phase 3 randomised clinical trial. *European Journal of Pediatrics*, 179(10), 1523–1528. https://doi.org/10.1007/s00431-020-03635-2
22. Malhotra, D., Nour, N., El Halik, M., Zidan, M. (2020). Performance and analysis of pediatric index of mortality 3 score in a Pediatric ICU in Latifa Hospital, Dubai, UAE. *Dubai Medical Journal*, 3(1), 19–25. https://doi.org/10.1159/000505205
23. Wong, M.Y., Yang, Y.S., Cao, Z.Q., Guo, V.Y.W., Lam, C.L.K., Wong, C.K.H. (2018). Effects of health-related quality of life on health service utilisation in patients with colorectal neoplasms. *European Journal of Cancer Care*, 27(6), Article e12926. https://doi.org/10.1111/ecc.12926
24. Calderon-Gerstein, W., Bruno-Huaman, A., Damian-Mucha, M., Huayllani-Flores, L. (2021). Predictive capacity of the brain natriuretic peptide in the screening of heart failure in a high altitude population. *Respiratory Physiology and Neurobiology*, 289, Article 103654. https://doi.org/10.1016/j.resp.2021.103654
25. Ray, S., Bhattacharyya, B. (2020). Statistical modeling and forecasting of arima and arimax models for food grains production and net availability of India. *Journal of Experimental Biology and Agricultural Sciences*, 8(3), 296–309. https://doi.org/10.18006/2020.8(3).296.309
26. Kádeková, Z., Récky, R., Nagyová, L., Košičiarová, I., Holienčinová, M. (2017). Consumers´ purchasing preferences towards organic food in Slovakia. *Potravinarstvo Slovak Journal of Food Sciences*, 11(1), 731–738. https://doi.org/10.5219/846
27. Bahk, G.-J. (2010). Statistical probability analysis of storage temperatures of domestic refrigerator as a risk factor of foodborne illness outbreak. *Korean Journal of Food Science and Technology*, 42(3), 373–376.
28. Hu, M., Ji, X., Liu, G.M. (2021). On the strong Markov property for stochastic differential equations driven by G-Brownian motion. *Stochastic Processes and their Applications*, 131, 417–453. https://doi.org/10.1016/j.spa.2020.09.015
29. Poroshin, V., Shlishevsky, A., Tsybulya, K. (11–15 September 2017). *Development of a model of a homogeneous continuous medium based on the material with defects in the form of hollows.* International conference on modern trends in manufacturing technologies and equipment (ICMTMTE2017), 129, Article 02017. Sevastopol, Russia. https://doi.org/10.1051/matecconf/201712902017
30. Lupandin, V.I. (2009) Mathematical methods in psychology. Ekaterinburg: Ural University Publishing House, 2009. (In Russian)
31. Badshah, Y., Shabbir, M., Hayat, H., Fatima, Z., Burki, A., Khan, S. et al. (2021). Genetic markers of osteoarthritis: early diagnosis in susceptible Pakistani population. *Journal of Orthopaedic Surgery and Research*, 16(1), Article 124. https://doi.org/10.1186/s13018-021-02230-x
32. Alves-Ferreira, M., Quintas, M., Sequeiros, J., Sousa, A., Pereira-Monteiro, J., Alonso, I. et al. (2021). A genetic interaction of NRXN2 with GABRE, SYT1 and CASK in migraine patients: a case-control study. *Journal of Headache and Pain*, 22(1), Article 57. https://doi.org/10.1186/s10194-021-01266-y
33. Do Nascimento, I.D.S., Pastor, A.F., Lopes, T.R.R., Farias, P.C.S., Goncales, J.P., Do Carmo, R.F. et al. (2020). Retrospective cross-sectional observational study on the epidemiological profile of dengue cases in Pernambuco state, Brazil, between 2015 and 2017. *BMC Public Health*, 20(1), Article 923. https://doi.org/10.1186/s12889-020-09047-z
34. Plichta, M., Jezewska-Zychowicz, M. (2019). Eating behaviors, attitudes toward health and eating, and symptoms of orthorexia nervosa among students. *Appetite*, 137, 114–123. https://doi.org/10.1016/j.appet.2019.02.022
35. Zhu, W.Y., Li, W., Geng, Q., Wang, X.Y., Sun, W., Jiang, H. et al. (2018). Silence of stomatin-like protein 2 represses migration and invasion ability of human liver cancer cells via inhibiting the nuclear factor kappa B (NF-κB) pathway. *Medical Science Monitor,* 24, 7625–7632. https://doi.org/10.12659/MSM.909156
36. Espinoza-Quispe, C.E., Marrero-Saucedo, F.M., Benavides, R.A.H. (2021). Solid Waste Management in the County of Huancavelica, Peru. *Letras Verdes*, 28, 163–177. https://doi.org/10.17141/letrasverdes.28.2020.4269
37. Kruk, J., Ptaszek, P., Kaczmarczyk, K. (2021). Technological aspects of xanthan gum and gum Arabic presence in chicken egg albumin wet foams: Application of nonlinear rheology and nonparametric statistics. *Food Hydrocolloids*, 117, Article 106683. https://doi.org/10.1016/j.foodhyd.2021.106683
38. Triwardhani, A., Anggitia, C., Ardani, I.G.A.W., Nugraha, A.P., Riawan, W. (2021). The increased basic fibroblast growth factor expression and osteoblasts number post Bifidobacterium bifidum probiotic supplementation during orthodontic tooth movement in Wistar rats. *Journal of Pharmacy and Pharmacognosy Research*, 9(4), 446–453.

39. Honary, V., Nitz, M., Wysocki, B.J., Wysocki, T.A. (2019). Modeling 3-D diffusion using queueing networks. *Biosystems,* 179, 17–23. https://doi.org/10.1016/j.biosystems.2018.12.006

40. Huang, X.-W., Emura, T. (2019). Model diagnostic procedures for copula-based Markov chain models for statistical process control. *Communications in Statistics: Simulation and Computation,* 50(8), 2345–2367. https://doi.org/10.1080/03610918.2019.1602647

41. Bakhtiari, A., Hajian-Tilaki, K., Omidvar, S., Nasiri-Amiri, F. (2019). Clinical and metabolic response to soy administration in older women with metabolic syndrome: A randomized controlled trial. *Diabetology and Metabolic Syndrome*, 11, Article 47. https://doi.org/10.1186/s13098-019-0441-y

42. Botelho, R.B.A., De Cássia Akutsu, R., Zandonadi, R.P. (2019). Low-income population sugar (Sucrose) intake: A cross-sectional study among adults assisted by a brazilian food assistance program. *Nutrients,* 11(4), Article 798. https://doi.org/10.3390/nu11040798

43. Sevillano Morales, J., Moreno-Ortega, A., Amaro Lopez, M.A., Arenas Casas A., Cámara-Martos, F., Moreno-Rojas, R. (2018). Game meat consumption by hunters and their relatives: a probabilistic approach. *Food Additives and Contaminants — Part A Chemistry, Analysis, Control, Exposure and Risk Assessment*, 35(9), 1739–1748. https://doi.org/10.1080/19440049.2018.1488183

44. Mejía-Díaz, D.M., Carmona-Garcés, I.C., Giraldo-López, P.A., González-Zapata, L. (2014). Nutritional content of food, and non-alcoholic beverages advertisements broadcasted in children's slot of Colombian national television. *Nutricion Hospitalaria*, 29(4), 858–864. https://doi.org/10.3305/nh.2014.29.4.7214

AUTHOR INFORMATION

Marina A. Nikitina, Candidate of technical sciences, docent, leading scientific worker, the Head of the Direction of Information Technologies of the Center of Economic and Analytical Research and Information Technologies, V. M. Gorbatov Federal Research Center for Food Systems. 26, Talalikhina str., 109316, Moscow, Russia. Tel: +7–495–676–95–11 extension 297, E-mail: m.nikitina@fncps.ru
ORCID: https://orcid.org/0000–0002–8313–4105
* corresponding author

Irina M. Chernukha, Doctor of technical sciences, professor, Academician of the Russian Academy of Sciences, Head of the Department for Coordination of Initiative and International Projects, V. M. Gorbatov Federal Research Center for Food Systems. 26, Talalikhina, 109316, Moscow, Russia. Tel: +7–495–676–95–11 extension 109, E-mail: imcher@inbox.ru
ORCID: https://orcid.org/0000–0003–4298–0927

All authors bear responsibility for the work and presented data.

All authors made an equal contribution to the work.

The authors were equally involved in writing the manuscript and bear the equal responsibility for plagiarism.

The authors declare no conflict of interest.