# The method of specification the degree of reliability for “zero hypotheses” about the distribution laws basing on Pearson’s and Kolmogorov’s consent criteria

a. Note: 1) As it’s mentioned in [14] the sign (+) is placed in the first interval if the value of the sign coinciding with the upper interval limit is included into the same interval; 2) the $f_i$ sum of empirical frequencies is equal to the original set Y power: $\sum f_i = | Y | = N = 12$.

Ilya Igushkin
*Engineering Institute*
*Kazan Federal University*
ilyha133@mail.ru
Natalya Verzun
*Department of information systems and technologies*
*St. Petersburg State University of Economics*
verzun.n@unecon.ru

Anatoly Shikhalev
*Engineering Institute*
*Kazan Federal University*
shihalev_48@mail.ru

Irina Akhmetova
*Institute of Management, Economics and Finance*
*Kazan Federal University*
iraahmetova@mail.ru

Dmitry Vorontsov
*Engineering Institute*
*Kazan Federal University*
DPVoroncov@kpfu.ru

Vadim Zubenko
*Engineering Institute*
*Kazan Federal University*
*zubenkovadim93@gmail.com*

VIII International Conference on Information Technology and Nanotechnology (ITNT-2022)

## Abstract

In the recent years the modern socio-economic indicators usually are very unstable. Therefore the using of initial (raw) data even in the so-called "small samples" volume (25-30 values) could be problematic [1], [2]. For the initial data now it is often necessary to use the smaller number of it called the “ultra-small samples” which add some technical restrictions when studying many problems in the mathematical statistics framework, in particular when analyzing the task of the Pearson’s consent criteria χ2 calculating with the recommended significance level Rpredet. ≥ 95% (i.e. “predetermined level”). In this research we propose the method of "(re)conciliation" (or “concordance”) for the determining and calculating the type of distribution laws for the random variable when Rcalc. < Rpredet. (i.e. Rcalculated < Rpredetermined). Moreover, also was created the "discrepancy criterion” (or “non-conformity criterion”) for the "zero hypotheses" concerned with the distribution laws of random variables testing for ultra-small samples cases.

## I. Introduction and problem statement

The initial data (ID) in Table 1 are given about the disc cultivator paws (in pcs.) produced by the OOO “Agromaster” (a limited liability company (LLC) under the laws of Russian Federation) in Kazan city:

Table I. Results of the monthly items

| Lot, № | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | ∑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Items, in pcs. | 80 | 67 | 79 | 77 | 81 | 75 | 76 | 82 | 68 | 90 | 69 | 75 | 919 |
| Defect items, in pcs. | 5 | 2 | 4 | 4 | 5 | 3 | 5 | 6 | 3 | 5 | 5 | 3 | 50 |

The task is to estimate the initial or basic data (BD) for its belonging as a random variable (i.e. the number of monthly defect cases) to the particular distribution law (DL). Firstly a "null hypothesis" is stated about the belonging of the random variable (RV) to some particular distribution law (DL): to the normal distribution law, Poisson’s distribution or to any other law. The validity of the "null hypothesis" is testing with the Pearson’s χ2 consent criterion with predetermined degrees of freedom number df [3] and significance level α = 0.05 [4] which gives the reliability level P = 95%.So, if the requirement (1) is realized:

χ2calc. ≤ χ2table (df; α = 0.05) (1),then the "null hypothesis" will be accepted. In another case this hypothesis will be rejected.

χ2calc. > χ2table (df; α = 0.05) (2)

Then there are some problems: 1) How much is different exactly in percentage points (the "null hypothesis") from the expected reliability Rtable = 95% with the Pearson's parameter χ2calc. from the Rcalc. (i.e. “calculated”) with a tabular value of χ2table (df; α = 0.05); 2) The difference ∆R between it is acceptable or it isn’t? Some answers to these questions are presented in this research.

## II. Test of the "null hypothesis" about the normal (Gauss) distribution law of random variable

According to [8] usually is better to start from the hypergeometric probability distribution which could be easily approximated by the binomial distribution [9] or Poisson’s distribution [10]. And if the binomial law of random variable distribution formally consists of the Bernoulli form and the number of combinations from combinatorics [11] and for its using we should know only the number of parties and the average defects number, then some clarifications are needed for the Poisson’s and normal distribution laws application. So, for the normal distribution law hypothesis testing we need to determine a series of empirical frequencies i = 1, n, where “n” is the number of rows-variants from the variation series (VR). Then it is necessary to realize the random variable data adjustment (creating the theoretical distribution) with the theoretical frequencies fitheor. finding. Let’s propose that we have the original set (third row in the Table 1) with the name Y = {yj}, j = 1, N = 12. Then the set of elements yj i.e. the defect cases as the random variable will look like: Y = {5, 2, 4, 4, 5, 3, 5, 6, 3, 5, 5, 3}. Next, the variation series with regard to (5) will take the form presented in the Table 2.

Table II. Variation series as the complex of variants (rows) X.

| Variants count, i | Variants, $x_i^{start}$ - $x_i^{start}$ | Empirical frequencies, $f_i$ | | Center of variants, $x_i^{av}$ |
|---|---|---|---|---|
| | | *Count* | *Number* | |
| 1 | 2.0 - 3.0 (+) | //// | 4 | 2.5 |
| 2 | 3.0 - 4.0 | // | 2 | 3.5 |
| 3 | 4.0 – 5.0 | ///// | 5 | 4.5 |
| 4 = n | 5.0 – 6.0 | / | 1 | 5.5 |

a. *Note*: 1) As it’s mentioned in [14] the sign (+) is placed in the first interval if the value of the sign coinciding with the upper interval limit is included into the same interval; 2) the $f_i$ sum of empirical frequencies is equal to the original set Y power: $\sum f_i = | Y | = N = 12$.

## III. An adjustment of the created variation series

This operation proposes the theoretical frequencies fitheor calculating (we don’t know and even propose here any structural changes in production, etc., see [15]).

For this case [4] we will create the calculation Table 3 for which two variables are needed: the average weighted xav.weigh. (6), standard deviation σ (7), its dispersion σ2 and standard deviation σ and also the constant const:

$$x_{av.}^{weigh.} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{(2,5\cdot4+3,5\cdot2+4,5\cdot5+5,5\cdot1)}{4+2+5+1} \approx 3,8 \text{ (pcs.)} \quad (6)$$

Table III. THEORETICAL FREQUENCIES $F_I^{THEOR.}$ VALUES CALCULATION

| i | $x_i^{av}$ | ($x_i^{av}$– $x_{av}^{weigh}$) | t = ($x_i^{av}$ – $x_{av}^{weigh}$)/σ | φ (t) | $f_i^{theor}$= const · φ (t) |
|---|---|---|---|---|---|
| ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6*** |
| 1 | 2.5 | - 1.3 | - 1.29 | 0.1736 | 2.083 ≈ 2 |
| 2 | 3.5 | - 0.3 | - 0.30 | 0.3814 | 4.577 ≈ 5 |
| 3 | 4.5 | 0.7 | - 0.69 | 0.3144 | 3.773 ≈ 4 |
| 4 | 5.5 | 1.7 | 1.68 | 0.0973 | 1.168 ≈ 1 |

## IV. Pearson's consent criterion calculation

Table IV. Pearson's consent criterion calculation

| i | $f_i$ | $f_i^{theor}$ | ($f_i$ -$f_i^{theor}$) | ($f_i$ -$f_i^{theor}$)$^2$ | $\chi^2_{calci}$ = ($f_i$ - $f_i^{theor}$)$^2$ / $f_i^{theor}$ |
|---|---|---|---|---|---|
| 1 | 4 | 2 | 2 | 4 | 2 |
| 2 | 2 | 5 | - 3 | 9 | 1.80 |
| 3 | 5 | 4 | 1 | 1 | 0.25 |
| 4 | 1 | 1 | 0 | 0 | 0 |
| | | | | **Calculated value $\chi^2_{calc.}$:** | 4.05 |

From the Pearson’s coefficient table with the previously fixed significance level α = 0.05 we could find the tabular value χ2table (df = 1; α = 0.05) = 3.84, whereas χ2calc = 4.05. As a result and basing on the (2) condition χ2calc.>χ2.table we could see that the discrepancy between the empirical and theoretical frequencies could not be estimated as the random and previously proposed hypothesis about the normal distribution of random variable is not confirmed with the our determined reliability Rpredet. = 95%. After that the next questions appear: 1) How much the reliability which has been (exactly) achieved in Rcalc. < Rpredet. is less than previously proposed?; 2) Is the discrepancy large? For the receiving an answer to the first question above we will use the alternative Kolmogorov’s consent criterion as the most interesting for us is the calculation of the expression (10) in percent.

## V. Kolmogorov's consent criterion calculation

Table V. “D” value from formula (11) for Kolmogorov’s consent criterion calculation

| i | $f_i^{v}$ | $f_i^{theor}$ | Cumulated frequencies | | \| $q_i$ - $q_i^{theor}$\| |
|---|---|---|---|---|---|
| | | | *Empirical,$q_i$* | *Theoretical,$q_i^{theor}$* | |
| 1 | 4 | 2 | 4 | 2 | 2 - max |
| 2 | 2 | 5 | 6 | 7 | 1 |
| 3 | 5 | 4 | 11 | 11 | 0 |
| 4 | 1 | 1 | 12 | 12 | 0 |

Table VI. Admissible deviations in percentage points from "null hypothesis" for the distribution law of random variable estimation.

| **The nature of the "null hypothesis" merit of fit (see formula (14) for calculating ∆R = $P_{determ.}$ - $R_{calc.}$,inperc. points)** | **Admissible deviation from the "null hypothesis" in perc. points at determ. significance level α on Pearson's consent criterion** |
|---|---|
| Minimum discrepancy | up to 3% |
| Ordinary (usual) discrepancy | 3% - 10% |
| Approximate (or tentative) discrepancy | 10% - 20% |
| Evaluative discrepancy | 20% - 40% |
| Rapid calculation (of) discrepancy | more than 40% |

Although Kolmogorov's test is usually considered as with the "less power" than χ2-Pearson’s criterion but its additional (and not alternative) using could be useful in the cases where the ratio (1) is not realized but the inequality (2) is satisfied. Then Kolmogorov’s criterion of consent could be expressly and particularly used as the estimation of the measure of disagreement between the desired (or predetermined) reliability Rdeterm. = 95% and the calculated reliability Rcalc. Using here the modified (by us) Yadov’s table as a normative scale makes it possible to estimate the nature of the revealed discrepancy in the linguistic scale (see Table 6). Such look-up will not only give the possibility to estimate the significance of the discrepancy but also take it into consideration in the subsequent discussions and calculations.