UDC 37.02:519.85

INSPECTION OF HYPOTHESES OF THE LAW DISTRIBUTIONS USING THE CHI-SQUARE TEST OF EXPERIMENTAL RESEARCH RESULTS IN PEDAGOGY

IVAN SARAHAVETS, NASTASSIA MATELENAK

Polotsk State University, Belarus

The paper proposes one of the ways of testing the hypothesis of distribution law of a sample aggregate. It consists of the following: intuitively choosing the distribution law of a random variable (CB) and estimating its parameters for constructing the statistics of a chi-square test, dividing the definition set of the chosen law into intervals with the same probability of falling into the values of the selected CB. The boundaries of the intervals can be easily found using the distribution function of the selected CB. It is easy to determine the theoretical frequencies of hits in the intervals as they are equal and do not need to be adjusted (here the number of intervals, the sample size) and the statistics of the criterion

Usually, when constructing the distribution law of a random variable (CB) from a selective set, the following elements are present:

1) By the geometric characteristics of the sample and by the relationships between the numerical characteristics, the form of the distribution law is chosen. Sometimes the form of the distribution law is established from intuitive considerations;

2) the parameters of the chosen distribution law are estimated;

3) the consistency of the constructed distribution law with the sample data is checked.

The traditional method of implementation is as follows. We break up a set of selective values on k equal intervals, where $k \approx 1 + 3{,}22 \cdot \lg n$, n – sample size, m_i – number of sample values, of i interspace, $i = 1, 2, ..., k$. Further, we find the theoretical probabilities $p_i^/$ and theoretical frequencies $m_i^/ = n \cdot p_i^/$, The lower limit of the first interval is assumed to be equal to the minimum value of the domain of definition of the chosen law, and the upper limit of the latter to the maximum. Theoretical frequencies must satisfy the condition $m_i^/ \geq 5$. If for some intervals this condition is not met, then they are combined with neighboring ones. After this, the statistics of the chi-square test are calculated (χ^2) and the result is compared with the critical value.

In this paper, in order to perform step three, it is proposed to split the set of values of the chosen distribution law into k intervals with the same probability $1/k$ hit in each interval. Boundary Intervals x_i are determined using the distribution function $F(x)$ chosen law as a result of solving the equations $F(x_i) = i/k$ $(i = 1, 2, ..., k-1)$. For all practically important distributions, it is possible to specify the boundaries in advance when dividing into l intervals (l =3, 4,…, k).

Example 1. Construct the law of distribution of SV for the sample:

Table 1.

i	1	2	3	4	5	6	7	8	9	10	11	12	13
X_i	0,07	0,09	0,16	0,23	0,53	0,75	0,9	1,24	1,41	1,68	2,1	2,2	2,21
i	14	15	16	17	18	19	20	21	22	23	24	25	
X_i	2,32	2,41	2,62	2,78	3,14	4,03	4,54	9,95	10,8	11,47	12,06	15,91	

Solution. 1st method. In this case $n = 25, 1 + 3{,}22 \cdot \lg 25 \approx 5{,}5, \Rightarrow k = 5$. As $X_{\min}$ choose $X_0 = 0$, as $X_{\max}$ choose 16. Then the length of each interval $\Delta = 16/5 = 3{,}2$ and it is easy to calculate the number of values (frequencies) at intervals. Such a frequency distribution is possible for an exponential distribution with a distribution function $F(x) = 1 - \exp(-\lambda \cdot x), \quad x > 0$, that is, a hypothesis is advanced H_0: the sample is made from the general population with the exponential distribution law. Numerical characteristics of the sample $\overline{X} = 3{,}824; D^* = 19{,}788; S^* = \sqrt{D^*} = 4{,}448$. We estimate the parameter $\lambda = \frac{1}{\overline{X}} = \frac{1}{3{,}824}$ and calculate the theoretical probabilities $p_i^/$ and theoretical frequencies $m_i^/$. As a result, we get the calculation table:

Table 2

Intervals	(0; 3,2)	(3,2; 6,4)	(6,4; 9,6)	(9,6; 12,8)	(12,8; 16)
m_i	18	2	0	4	1
$p_i^/$	0,5669	0,2455	0,1063	0,0461	0,0352
$m_i^/$	14,1725	6,1375	2,6575	1,1525	0,88

The condition is to satisfy $m_i^/ \geq 5$ 2, 3, 4 and 5 intervals should be combined, and for calculating the statistics of the criterion χ^2 we obtain only two intervals (the minimum number of intervals is 3), that is, the hypothesis H_0 deviates.

2nd method. Hypothesis H_0 has already been put forward and for the distribution function, the expression is $F(x)=1-\exp(-\frac{x}{\overline{X}}),\ \ x>0$. We break the sample set into intervals with the same probability $p^/ = 1/5 = 0{,}2$ hit in them the values of the investigated CB. The boundaries of the intervals, as already noted, we find from the equation $F(x_i)=i/5 \Rightarrow 1-\exp(-x_i/\overline{X})=i/5$. Solving this equation for x_i, we get $x_i=-3{,}824\cdot\ln(1-i/5)$. It is finding x_i and counting the number of values found in the intervals, we obtain.

Table 3

Intervals	(0; 0,853)	(0,853; 1,953)	(1,953; 3,504)	(3,504; 6,154)	(6,154; ∞)
m_i	6	4	8	2	5

Since for each interval the theoretical frequency is $m_i^/ = 25\cdot 0{,}2 = 5$, we can calculate the statistics of the criterion χ^2 (3 and 4 intervals are combined):

$$\chi^2_{emp}=\sum_i \frac{(m_i-m_i^/)^2}{m_i^/}=\frac{(6-5)^2}{5}+\frac{(4-5)^2}{5}+\frac{(10-10)^2}{10}+\frac{(5-5)^2}{5}=0{,}4.$$

According to the table of quantiles of the chi-square the distribution is $\chi^2_{4-1-1;\,0{,}95}=5{,}99$. As $\chi^2_{emp}<\chi^2_{2;\,0{,}95}$ (0,4 < 5.99), then the hypothesis H_0 is adopted.

Most often we have to test the hypothesis of a normal distribution, since many statistical criteria are based on this distribution. In this case, the need for the item disappears. We establish with an accuracy of 5 digits after the comma of the boundary for the partition of the normal distribution $N(0;1)$ on 4, 5, 6 and 7 intervals with the same theoretical probability $1/k$ $(k=4;5;6;7)$ hit in each interval.

$$\left.\begin{array}{r} k=4:\ \ -\infty;-0{,}67448;0;0{,}67448;+\infty,\\ k=5:\ \ -\infty;-0{,}84162;-0{,}25334;0{,}25334;0{,}84162;+\infty,\\ k=6:\ \ -\infty;-0{,}96739;-0{,}43070;0;0{,}43070;-0{,}96739;+\infty,\\ k=7:\ \ -\infty;-1{,}06757;-0{,}56594;-0{,}18;0{,}18;0{,}56594;1{,}06757;+\infty. \end{array}\right\} \quad (1)$$

Example 2. Table 4 contains a fragment of the results normality verification of a pedagogical experiment conducted at the Department of Higher Mathematics of Polotsk State University.

Explanations to Table 4. The essence of the experiment was to compare the values of the control group (CG) and experimental group (EG). At the initial stage (Stage 0), the homogeneity of the groups taken for the experiment was tested (group 10BB-KG, and group 10 TV-EG). To compare the indicators of the groups selected, the sum of the points obtained in the testing and in the certificate was chosen.
In subsequent stages, the quality of the knowledge the students received by traditional presentation of the material (CG) and in the experimental (EG) was compared (stages 1 and 2 - the results of the 1st and 2nd semesters

were chosen as indicators for comparison - the sum of the scores obtained in control papers and in exams in the first and second semesters).

Table 4. – $UN_i = \left(XN_i - \overline{XN}\right) / SN_x, VN_i = \left(YN_i - \overline{YN}\right) / SN_y (N = 0,1,2)$

	Stage 0				Stage 1				Stage 2			
	10 BB		10 TB		10 BB		10 TB		10 BB		10 TB	
№	*X0*	*U0*	*Y0*	*V0*	*X1*	*U1*	*Y1*	*V1*	*X2*	*U2*	*Y2*	*V2*
1	17	-1,5362	21	-1,4254	11	-1,8524	12	-2,4493	11	-1,7806	12	-2,5319
2	17	-1,5362	24	-1,2568	12	-1,6852	17	-1,4344	12	-1,6104	17	-1,6155
3	18	-1,4689	25	-1,2006	12	-1,6852	17	-1,4344	13	-1,4402	17	-1,6155
4	19	-1,4016	28	-1,0320	13	-1,518	18	-1,2314	14	-1,27	19	-1,2490
5	19	-1,4016	29	-0,9758	15	-1,1835	19	-1,0284	14	-1,27	22	-0,6991
6	21	-1,2671	29	-0,9758	15	-1,1835	19	-1,0284	14	-1,27	24	-0,3326
7	22	-1,1998	29	-0,9758	15	-1,1835	20	-0,8255	16	-0,9295	24	-0,3326
8	23	-1,1325	30	-0,9196	16	-1,0163	21	-0,6225	16	-0,9295	24	-0,3326
9	23	-1,1325	32	-0,8073	16	-1,0163	21	-0,6225	16	-0,9295	24	-0,3326
10	24	-1,0653	33	-0,7511	18	-0,6818	21	-0,6225	17	-0,7593	24	-0,3326
11	24	-1,0653	35	-0,6387	19	-0,5146	21	-0,6225	17	-0,7593	24	-0,3326
12	25	-0,9980	35	-0,6387	19	-0,5146	23	-0,2165	17	-0,7593	26	0,0339
13	25	-0,9980	37	-0,5263	20	-0,3473	24	-0,0135	18	-0,5891	26	0,0339
14	32	-0,5270	40	-0,3577	20	-0,3473	24	-0,0135	18	-0,5891	26	0,0339
15	33	-0,4597	42	-0,2454	20	-0,3473	24	-0,0135	18	-0,5891	26	0,0339
16	35	-0,3252	46	-0,0206	20	-0,3473	25	0,1894	19	-0,4189	26	0,0339
17	36	-0,2579	48	0,0918	21	-0,1801	25	0,1894	20	-0,2487	27	0,2172
18	37	-0,1906	49	0,1480	21	-0,1801	26	0,3924	20	-0,2487	27	0,2172
19	39	-0,0561	53	0,3727	21	-0,1801	26	0,3924	20	-0,2487	29	0,5838
20	39	-0,0561	54	0,4289	22	-0,0129	26	0,3924	21	-0,0785	29	0,5838
21	39	-0,0561	57	0,5975	22	-0,0129	27	0,5954	23	0,2618	29	0,5838
22	44	0,2803	59	0,7099	23	0,1544	27	0,5954	23	0,2618	29	0,5838
23	44	0,2803	61	0,8222	24	0,3216	27	0,5954	23	0,2618	31	0,9503
24	44	0,2803	62	0,8784	24	0,3216	29	1,0014	24	0,4320	32	1,1336
25	45	0,3476	64	0,9908	24	0,3216	30	1,2043	24	0,4320	32	1,1336
26	47	0,4822	64	0,9908	24	0,3216	30	1,2043	24	0,4320	33	1,3168
27	49	0,6167	65	1,0470	25	0,4888	30	1,2043	25	0,6022	38	2,2332
28	49	0,6167	75	1,6089	26	0,6561	30	1,2043	25	0,6022		
29	49	0,6167	78	1,7775	26	0,6561	31	1,4073	25	0,6022		
30	51	0,7513	87	2,2832	26	0,6561	32	1,6103	27	0,9426		
31	52	0,8186			28	0,9905			27	0,9426		
32	53	0,8858			28	0,9905			27	0,9426		
33	53	0,8858			29	1,1578			27	0,9426		
34	53	0,8858			30	1,325			28	1,1128		
35	54	0,9531			30	1,325			29	1,2830		
36	54	0,9531			31	1,4922			29	1,2830		
37	56	1,0877			31	1,4922			32	1,7937		
38	56	1,0877			32	1,6595			32	1,7937		
39	59	1,2895			32	1,6595			32	1,7937		
40	63	1,5586										
41	63	1,5586										
42	68	1,8950										

As a criterion for comparison, a two-sample t-criterion for comparing averages was selected [1, p. 128]. This criterion is based on the normal distribution and first of all it is necessary to check the normality of the received samples. This fact is verified in this paper with the help of the method of splitting the sample collection into intervals proposed above. First of all, for each sample that is marked $X0, Y0, X1, Y1, X2, Y2$, the mean values were found $\overline{X}, \overline{Y}$, Selective variances and mean square deviations Dx, Dy, Sx, Sy.

$\overline{X0}$	39,833	$\overline{Y0}$	46,367	$\overline{X1}$	22,077	$\overline{Y1}$	24,067	$\overline{X2}$	21,46	$\overline{Y2}$	25,815
D0x	220,92	D0y	316,72	D1x	35,757	D1y	24,271	D2x	34,52	D2y	29,772
S0x	14,864	S0y	17,797	S1x	5,980	S1y	4,9266	S2x	5,875	S2y	5,456

After that, the $\frac{X-\overline{X}}{Sx}, \frac{Y-\overline{Y}}{Sy}$, each of which has a distribution $N(0,1)$ provided that the corresponding sample has a normal distribution. Each sample of CG was divided into 6 intervals, and EG into 5 intervals. Using formulas (1) with $k=6$ and $k=5$ it is easy to get a partition of each sample into intervals with the same theoretical probability. For an example, we give a partition in case $X1$ and calculate for this partition the statistics of the chi-square test: $X1$: $m_i^/ = \frac{39}{6} = 6,5$, $m_1 = 9, m_2 = 3$, $m_3 = 9, m_4 = 5, m_5 = 4, m_6 = 9$ (combine the 1st and 2nd intervals of the 5th and 6th): $\chi^2_{emp} = \frac{(12-13)^2}{13} + \frac{(9-6,5)^2}{6,5} + \frac{(5-6,5)^2}{6,5} + \frac{(13-13)^2}{13} = 1,38 < 3,84 = \chi^2_{1;\,0,95}$.

It is also proved that all the samples satisfy the normality condition.
In conclusion, we note that the proposed method allows to determine the minimum sample size for testing hypotheses by the chi-square test. It is easy to propose the following formula: $n \geq 5 \cdot (r+2)$, where r – number of parameters of the law being checked.

REFERENCES

1. Мюллер, П. Таблицы по математической статистике / П. Мюллер, П. Нойман, Р. Шторм. – М. : Финансы и статистика, 1982. – 272 с.