ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

Krystyna Pruska[*]

SENSITIVITY OF THE CHI-SQUARE FIT TEST TO THE DIVISION OF HYPOTHETICAL SET OF INVESTIGATED VARIABLE VALUES[1]

ABSTRACT. In the paper there is considered an influence of division of hypothetical set of investigated variable values on making a decision either to reject or not to reject the null hypothesis in the goodness-of-fit chi-square test. The procedures of determination of distinguished categories and the number of categories are very important.

Some results of simulation experiments are presented. Two groups of experiments are considered: one with the same expected frequencies and the other with different expected frequencies for particular categories.

Key words: chi-square fit test, Monte Carlo experiments.

I. INTRODUCTION

The chi-square fit test is one of oldest nonparametric tests. It allows to check if the investigated random variable follows a distribution from a certain family of distributions. For this purpose the set of random variable values given by the null hypothesis is divided into some separate classes and the test statistic takes into account the differences between the numbers of sample observed values which belong to specific classes and the numbers of values which should belong to these classes when the null hypothesis is true. Due to this, the decision taken in the null hypothesis verification process depends on the differences between the sample outcomes and theoretical expectation dealing with the phenomenon investigated. However, this decision is also depended on the number of classes predetermined in the set of hypothetical values of the random variable and on the way in which they were determined.

[*] Ph.D., Associate Professor, Chair of Statistical Methods, University of Łódź.

[1] The paper was presented at the Seminar "Testy zgodności ich zastosowania" which had taken place on 21.03.2001 in Łódź.

In the paper the results of the simulation analysis of the chi-square fit test sensitivity to the division of hypothetical set of random variable values into classes applied are presented in the case of simple samples and continuous population distribution.

II. THE CHI-SQUARE FIT TEST

Let X denote the statistical attribute with respect to which the population is investigated, F_X – cumulative distribution function (cdf) of random variable which is a model of attribute X. Let us assume that a simple sample $X_1,...,X_n$ was drawn from the population. The chi-square fit test (see e.g. Fisz (1969), Domański i Pruska (2000)) allows to verify the following hypothesis:

$$H_0 : F_X \in \mathcal{F}_0$$

against the alternate:

$$H_1 : F_X \notin \mathcal{F}_0,$$

where $\mathcal{F}_0$ is known cdf family. If the family $\mathcal{F}_0$ consists of one element i.e. $\mathcal{F}_0 = \{F_0\}$, then the null hypothesis states that the investigated variable follows the distribution with cdf F_0.

The procedure of H_0 verification starts from dividing the set of the $\mathcal{F}_0$ random variable values into r separate classes. Then, we find the value of the test statistic i.e.:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}, \tag{1}$$

where

n_i – number of the i-th class observations (i = 1, ..., r),

p_i – probability of appearing in the sample value from the i-th class (i= 1 ,..., r).

Probabilities p_i (i = 1 ,..., n) are called theoretical probabilities.

If the H_0 hypothesis is true the χ^2 statistic follows the asymptotic chi-square distribution with $r-l-1$ degrees of freedom, where l is the number of the hypothetical distribution parameters estimated from the sample by likelihood method.

In the test we apply the following region of rejection: $\langle \chi^2_\alpha; +\infty)$, where $P(\chi^2 \geq \chi^2_\alpha) = \alpha$ and α is the level of significance. If the value of the χ^2 statistic belongs to the interval $\langle \chi^2_\alpha; +\infty)$, we reject H_0. Otherwise, there is no ground for rejecting H_0.

Statistic (1) is used in the chi-square fit test for simple samples. Some modifications of this test for nonsimple samples are also known (see. e.g. Bracha (1996), Domański and Pruska (2000)).

III. SIMULATION ANALYSIS OF CHI-SQUARE TEST SENSITIVITY FOR SIMPLE SAMPLES

The values of the chi-square fit test statistic depend on the division into classes of the set of values of the null hypothesis random variable. Following Greń (1987), each class should contain at least 8 observations, and according to Fisz (1969), the number of classes should meet the condition $np_i \geq 10$, where n is the sample size, and p_i are theoretical probabilities given in formula (1). Following Zieliński (1979), if classes are defined so that their theoretical probabilities are equal, it is enough that each class contains at least one observation. In the paper written by Krysicki et al. (1986) the numbers of classes are proposed for the given sample size.

Appropriately arranged simulation experiments allow to observe the influence of the division into classes of the hypothetical set of the investigated random variable values on the test statistic, i.e. on the decision of rejecting H_0.

In this paper the properties of the chi-square fit test are investigated in the case of verifing fit with the normal distribution when normal, t-Student or chi-square populations are considered.

From the population constituted by 50 000 numbers generated from a specified distribution, the samples of sizes n = 100, 300, 500, 700, 900, 1000 were drawn.The hypothetical set of the random variable values was divided into r classes, for r = 10, 15, 20, 25, 30, and for n equal to 900 and 1000 the division into 35, 40, 45 and 50 classes was also considered. Two types of the division were analized:

A) division with equal theoretical probabilities i.e. $p_i = 1/r$ for $i = 1,..., r$;

B) division with unequal theoretical probabilities involving classes with equal lenghts (apart from the first and last class), i.e. with lengths equal to : $(x_{\max} - x_{\min}) / r$, where $x_{\max}$ i $x_{\min}$ stand for the greatest and smallest observed value of the investigated variable; the first class is: $(-\infty\ ;\ x_{\min} + 1/r]$ for the normal and t-Student distribution and $[0\ ;\ x_{\min} + 1/r]$ for the chi-square distribution; the last class is: $(x_{\max} - 1/r\ ; +\infty)$ for all three distributions.

Three groups of experiments were carried out with respect to the population distribution. The following distributions were considered :

– normal distributions: $N(10; 2)$, $N(2; 10)$, $N(3; 1)$;

– t-Student distribution: $S_{10}, S_{20}, S_{30}, S_{40}, S_{50}, S_{100}$;

– chi-square distribution: $\chi^2_{10}, \chi^2_{20}, \chi^2_{30}, \chi^2_{40}, \chi^2_{50}, \chi^2_{100}$.

For all experiments the level of significance $\alpha = 0.05$ was applied.

For the normal population the parameters in the null hypothesis were not specified. For the t-Student population the H_0 was constituted by the normal distribution $N(0;1)$, and for the χ^2_k population the H_0 contained normal distribution $N(k;\sqrt{2k})$.

In the experiments carried out for the normal distrbution and for the equal probabilities division, all decisions, regardless of the sample size and number of classes, were correct. For the divisions with unequal theoretical probabilities the null hypothesis was reject 4 times for each of the normal distributions. These cases were the experiments for which : $n = 400$ and $r = 20$, $n = 400$ and $r = 25$, $n = 600$ and $r = 20$, $n = 700$ and $r = 15$. They contained classes with numbers of observations smaller than 8 but these were not the only such cases among the ones considered in this group.

The experiments results for the chi-square distribution are presented in tables 1–3, with table 1 giving decisions taken for the divisions with unequal theoretical probabilities performed according to the rules described above and for smaller number of classes than originally planned (some classes were connected to make their numbers of observations be at least 8). For 24 cases presented in table 1, in 3 cases different decisions of rejecting H_0 with respect to the division into classes were observed, with 2 cases having the right decision for smaller number of classes (if we assume that the chi-square distribution is close enough to the appropriate normal distribution for $k > 30$). Due to many possibilities of creating classes when there are many of them and they contain small numbers of observations, the remaining experiments with unequal theoretical probabilities pertained only to the division into classes of type B.

Table 1

Decisions taken with the chi-square fit test while verifying H_0 hypothesis of fitness of the population distribution $\chi^2(k, \sqrt{2k})$ with the $N(k, \sqrt{2k})$ distribution in the chosen cases of unequal theoretical probabilities and for sample size $n = 1000$ [a)]

Degrees of freedom k	Planned number of classes r_1	Diminished number of classes r_2	Decision for	
			r_1	r_2
1	2	3	4	5
10	35	20	–	–
	40	28	–	–
	45	28	–	–
	50	33	–	–
20	35	27	–	–
	40	28	–	–
	45	30	–	–
	50	30	–	–
30	35	25	–	–
	40	25	–	+
	45	28	–	+
	50	34	–	–
40	35	25	–	–
	40	27	–	–
	45	32	–	–
	50	36	–	–
50	35	25	–	–
	40	27	–	–
	45	32	–	–
	50	29	–	–
100	35	24	+	+
	40	26	+	+
	45	33	+	+
	50	34	+	–

[a)] The „+" sign means no grounds for rejecting H_0; „–" stands for rejecting H_0.

Source: own calculations.

Tables 2–3 present the decisions taken with the chi-square fit test for two types of division into classes: A and B. Table 2 refers to the chi-square population, and table 3 to the t-Student population. We may observe that the type of division into classes affects the inference concerning H_0. This is corroborated by the results in tables 4–5, which contain the numbers of different decisions taken while verifying the same hypothesis for different divisions into classes. In the performed experiments the number of rejections of H_0 for both chi-square and t-Student populations is bigger for the division of type B, i.e. for the division with unequal theoretical probabilities. There are also cases in which, regardless of the numbers of classes, for the fixed sample size the same decisions are taken. For the normal populations they comprised almost all cases. The results for the remaining two distributions are given in tables 4–5.

Table 2

Decisions taken with the chi-square fit test while verifying H_0 hypothesis of fitness of the population distribution $\chi^2(k, \sqrt{2k})$ with the distribution $N(k, \sqrt{2k})$ in the cases of equal and unequal theoretical probabilities [a)]

Degrees of freedom k	Sample size n	Number of classes r	Decision		Degrees of freedom k	Sample size n	Number of classes r	Decision	
			$p_i =$ const	$p_i \neq$ const				$p_i =$ const	$p_i \neq$ const
1	2	3	4	5	6	7	8	9	10
10	100	10	+	–	10	500	10	–	–
		15	+	–			15	–	–
		20	+	–			20	–	–
		25	+	–			25	–	–
		30	+	–			30	–	–
10	300	10	–	–	10	700	10	–	–
		15	–	–			15	–	–
		20	–	–			20	–	–
		25	–	–			25	–	–
		30	–	–			30	–	–
10	900	10	–	–	10	1000	10	–	–
		15	–	–			15	–	–
		20	–	–			20	–	–
		25	–	–			25	–	–
		30	–	–			30	–	–
		35	–	–			35	–	–
		40	–	–			40	–	–
		45	–	–			45	–	–
		50	–	–			50	–	–
20	100	10	+	+	20	500	10	–	–
		15	+	–			15	–	–
		20	+	+			20	+	–
		25	–	–			25	–	–
		30	+	+			30	–	–
20	300	10	–	–	20	700	10	–	+
		15	–	–			15	–	–
		20	+	–			20	–	–
		25	–	–			25	–	–
		30	–	–			30	–	–

Table 2 (cont.)

1	2	3	4	5	6	7	8	9	10
20	900	10	–	–	20	1000	10	–	–
		15	–	–			15	–	–
		20	–	–			20	–	–
		25	–	–			25	–	–
		30	–	–			30	–	–
		35	–	–			35	–	–
		40	–	–			40	–	–
		45	+	–			45	–	–
		50	–	–			50	–	–
30	100	10	+	+	30	500	10	–	–
		15	+	+			15	–	–
		20	+	+			20	+	–
		25	+	+			25	+	–
		30	–	–			30	+	–
30	300	10	+	+	30	700	10	+	–
		15	+	–			15	–	–
		20	+	+			20	+	+
		25	–	–			25	+	–
		30	–	–			30	+	+
30	900	10	–	–	30	1000	10	+	–
		15	–	–			15	+	–
		20	–	–			20	–	–
		25	+	–			25	–	–
		30	–	–			30	–	–
		35	–	–			35	+	–
		40	–	–			40	–	–
		45	–	–			45	+	–
		50	–	–			50	–	–
40	100	10	+	+	40	500	10	–	–
		15	+	+			15	+	+
		20	+	+			20	+	–
		25	+	+			25	+	+
		30	+	+			30	–	–
40	300	10	+	+	40	700	10	+	–
		15	+	+			15	+	–
		20	+	+			20	–	–
		25	+	+			25	+	–
		30	–	–			30	–	–
40	900	10	–	–	40	1000	10	+	–
		15	+	–			15	+	–
		20	–	–			20	–	–
		25	–	–			25	+	–
		30	+	–			30	–	–
		35	–	–			35	+	–
		40	+	+			40	–	–
		45	+	+			45	+	–

Table 2 (cont.)

1	2	3	4	5	6	7	8	9	10
40	900	50	–	–	40	1000	50	–	–
50	100	10	+	+	50	500	10	+	–
		15	+	+			15	–	+
		20	+	–			20	+	–
		25	–	+			25	+	–
		30	+	+			30	+	+
50	300	10	+	+	50	700	10	+	+
		15	+	+			15	–	–
		20	–	–			20	+	–
		25	–	–			25	+	–
		30	+	+			30	+	–
50	900	10	–	–	40	1000	10	+	–
		15	–	–			15	+	–
		20	+	–			20	–	–
		25	+	–			25	+	–
		30	–	–			30	–	–
		35	+	–			35	+	–
		40	+	–			40	–	–
		45	–	–			45	+	–
		50	–	–			50	–	–
100	100	10	+	+	100	500	10	+	+
		15	–	–			15	+	+
		20	+	+			20	+	+
		25	+	+			25	–	–
		30	+	+			30	+	+
100	300	10	+	–	100	700	10	+	+
		15	–	–			15	+	–
		20	+	+			20	+	+
		25	+	+			25	+	+
		30	+	+			30	+	–
100	900	10	+	+	100	1000	10	+	–
		15	+	+			15	–	–
		20	+	–			20	+	–
		25	+	–			25	+	–
		30	+	+			30	+	–
		35	–	–			35	+	+
		40	+	–			40	+	+
		45	+	–			45	+	+
		50	+	+			50	+	–

[a] The „+" means no grounds for rejecting H_0; „–" stands for rejecting H_0.

Source: own calculations.

Table 3

Decisions taken with the chi-square fit test while verifying H_0 hypothesis of fitness of the population distribution S_k with the distribution $N(0;1)$ in the cases of equal and unequal theoretical probabilities $^{a)}$

Degrees of freedom k	Sample size n	Number of classes r	Decision		Degrees of free-dom k	Sample size n	Number of classes r	Decision	
			$p_i =$ const	$p_i \neq$ const				$p_i =$ const	$p_i \neq$ const
1	2	3	4	5	6	7	8	9	10
10	100	10	+	+	10	500	10	+	–
		15	+	–			15	+	–
		20	+	+			20	+	–
		25	+	+			25	+	–
		30	+	–			30	+	–
10	300	10	+	–	10	700	10	+	–
		15	+	–			15	–	–
		20	+	–			20	–	–
		25	+	–			25	+	–
		30	+	–			30	+	–
10	900	10	+	–	10	1000	10	–	–
		15	+	–			15	+	–
		20	–	–			20	+	–
		25	–	–			25	+	–
		30	+	–			30	–	–
		35	+	–			35	+	–
		40	–	–			40	+	–
		45	+	–			45	–	–
		50	+	–			50	–	–
20	100	10	+	+	20	500	10	+	+
		15	+	+			15	+	–
		20	+	+			20	+	–
		25	+	–			25	+	+
		30	+	–			30	+	–
20	300	10	+	+	20	700	10	+	–
		15	+	–			15	+	–
		20	+	–			20	+	+
		25	+	+			25	+	–
		30	+	–			30	+	+
20	900	10	+	+	20	1000	10	+	–
		15	+	–			15	+	–
		20	+	–			20	+	–
		25	+	+			25	+	+
		30	+	–			30	+	+
		35	+	–			35	+	–
		40	+	–			40	+	–
		45	+	–			45	+	–
		50	+	–			50	+	–

Tabel 3 (cont.)

1	2	3	4	5	6	7	8	9	10
30	100	10	+	+	30	500	10	+	+
		15	–	+			15	+	+
		20	+	+			20	+	+
		25	+	+			25	+	+
		30	+	+			30	+	–
30	300	10	+	–	30	700	10	+	+
		15	+	–			15	+	–
		20	+	+			20	+	–
		25	+	+			25	+	–
		30	+	–			30	+	+
30	900	10	+	–	30	1000	10	+	–
		15	+	+			15	+	+
		20	+	+			20	+	+
		25	+	+			25	+	–
		30	+	–			30	+	+
		35	+	–			35	+	+
		40	+	–			40	+	–
		45	+	–			45	+	–
		50	+	+			50	+	–
40	100	10	+	+	40	500	10	+	+
		15	+	+			15	+	+
		20	+	+			20	+	+
		25	+	+			25	+	+
		30	+	+			30	+	+
40	300	10	+	+	40	700	10	+	+
		15	+	–			15	+	–
		20	+	+			20	+	+
		25	+	+			25	+	+
		30	–	–			30	+	+
40	900	10	+	–	40	1000	10	+	+
		15	+	–			15	+	+
		20	+	+			20	+	+
		25	+	–			25	+	+
		30	+	–			30	+	–
		35	+	+			35	+	+
		40	+	+			40	+	–
		45	+	–			45	+	+
		50	+	–			50	+	+
50	100	10	+	+	50	500	10	+	+
		15	+	+			15	+	+
		20	+	+			20	+	+
		25	+	–			25	+	–
		30	+	+			30	+	–

Tabel 3 (cont.)

1	2	3	4	5	6	7	8	9	10
50	300	10	+	+	50	700	10	+	+
		15	+	+			15	+	–
		20	+	+			20	+	+
		25	+	+			25	+	–
		30	+	+			30	+	+
50	900	10	+	+	40	1000	10	+	–
		15	+	–			15	+	+
		20	+	+			20	+	+
		25	+	+			25	+	–
		30	+	+			30	+	+
		35	+	+			35	+	+
		40	+	+			40	+	–
		45	+	+			45	+	+
		50	+	–			50	+	–
100	100	10	+	+	100	500	10	+	+
		15	+	+			15	+	+
		20	+	+			20	+	+
		25	+	+			25	+	+
		30	+	+			30	+	+
100	300	10	+	+	100	700	10	+	+
		15	+	+			15	+	+
		20	+	+			20	+	–
		25	+	+			25	+	–
		30	+	+			30	+	+
100	900	10	+	+	100	1000	10	+	+
		15	+	+			15	+	+
		20	+	+			20	+	+
		25	+	+			25	+	+
		30	+	+			30	+	–
		35	+	+			35	+	+
		40	+	+			40	+	+
		45	+	+			45	+	+
		50	+	+			50	+	+

[a] The „+” means no grounds for rejecting H_0; „–” stands for rejecting H_0.

Source: own calculations.

Table 4

Number of cases in given group of experiments for population distribution $\chi^2(k,\sqrt{2k})$
(number all cases in given group of experiments)

Number of degrees of freedom k	Number different decision dealing with H_0 for division A and B	Number of cases of rejecting H_0		Number of cases of the same decision dealing with H_0 for fixed sample size regardless the number of classes	
		division A	division B	division A	division B
10	6 (58)	51 (58)	57 (58)	6 (6)	6 (6)
20	9 (58)	43 (58)	48 (58)	2 (6)	4 (6)
30	17 (58)	29 (58)	46 (58)	0 (6)	3 (6)
40	19 (58)	17 (58)	36 (58)	1 (6)	3 (6)
50	26 (58)	17 (58)	40 (58)	0 (6)	3 (6)
100	23 (58)	8 (58)	39 (58)	1 (6)	0 (6)

Source: own calculations.

Table 5

Number of cases in given group of experiments for population distribution S_k
(number all cases in given group of experiments)

Number of degrees of freedom k	Number different decision dealing with H_0 for division A and B	Number of cases of rejecting H_0		Number of cases of the same decision dealing with H_0 for fixed sample size regardless the number of classes	
		division A	division B	division A	division B
10	43 (58)	10 (58)	53 (58)	3 (6)	5 (6)
20	35 (58)	1 (58)	36 (58)	6 (6)	0 (6)
30	22 (58)	2 (58)	22 (58)	5 (6)	1 (6)
40	13 (58)	1 (58)	14 (58)	5 (6)	2 (6)
50	17 (58)	0 (58)	17 (58)	6 (6)	1 (6)
100	6 (58)	1 (58)	7 (58)	6 (6)	4 (6)

Source: own calculations.

The performed experiments confirm the chi-square fit test sensitivity to the way of dividing the hypothetical set of values of the investigated random variable.

IV. FINAL REMARKS

The problem of division into classes of the set of values of the random variable always appears for the chi-square fit test. Especially for continuous variables there are many ways of divisions. It happens that hypothesis verification should be made on the basis of grouped data and there is no possibility of creating classes with equal theoretical probabilities. In such cases only another aggregation of observations is possible and dividing them so that the division would be most appropriate for comparing the empirical distribution with the hypothetical one. The classes arrangement may be performed on the basis of the density function graph or of the probability function of the null distribution and the sample based information. When we possess detailed data for continuous variables we should rather apply classes with equal theoretical probabilities. For such division in none of the performed experiments "empty classes" appeared (i.e. classes containing no observations). If the random variable is descrete the classes are determined by particular variants of values and their aggregation is possible but generally it is not possible to arrive at the classes with equal probabilities.

Apart from the choice of the way of dividing data for the chi-square fit test, there is the problem of determining the number of subsets that should be distinguished (see. e.g. Krysicki et al. (1986)).

It seems that simulation methods and becoming more and more efficient computers allow to make many observations on the properties of the chi-square fit test. The range of the meaning of the term "large sample' changes. Now, it is possible to carry out many experiments which were not possible in the past.

REFERENCES

Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.

Domański Cz., Pruska K. (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.

Fisz M. (1969), *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa.

Greń J. (1987), *Statystyka matematyczna. Podręcznik programowany.* PWN, Warszawa.

Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M. (1986), *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, cz. II,* PWN, Warszawa.

Zieliński R. (1979), *Generatory liczb losowych. Programowanie i testowanie na maszynach cyfrowych.* Wydawnictwa Naukowo-Techniczne, Warszawa.

Krystyna Pruska

WRAŻLIWOŚĆ TESTU ZGODNOŚCI CHI-KWADRAT NA PODZIAŁ HIPOTETYCZNEGO ZBIORU WARTOŚCI BADANEJ ZMIENNEJ

Test zgodności chi-kwadrat jest jednym z najstarszych testów zgodności. Podejmowanie decyzji o odrzuceniu lub nieodrzuceniu sprawdzanej hipotezy uzależnione jest nie tylko od wyników próby losowej, ale także od podziału hipotetycznego zbioru wartości badanej zmiennej na klasy. Liczba klas ma wpływ na rozkład i wartość statystyki testu.

W pracy rozpatrywany jest problem podziału na klasy w przypadku rozkładów ciągłych. Porównywane są wyniki weryfikacji hipotez dla jednakowych i niejednakowych prawdopodobieństw teoretycznych.