ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 175, 2004

*Krystyna Pruska**

TESTING THE IDENTITY OF DISTRIBUTIONS OF TWO DISCRETE RANDOM VARIABLES

Abstract. The comparing distributions of two discrete random variables appears often in statistical research. In many cases we can apply the test for two means for it. If the means are equal and we do not know the set of values of investigated variables, it is possible to use the properties of sample proportions for testing the identity of two distributions.

In this paper testing the identity of distributions for two univariate and two bivariate random variables is considered. The power of proposed tests is also analysed.

Key words: homogeneity test, test for proportions.

1. INTRODUCTION

The homogeneity χ^2-test is known procedure for verifying hypothesis about the identity of some distributions (see for example: C. Cramer (1958), C. Bracha (1996), C. Domański and K. Pruska (2000), J. Koronacki and J. Mielniczuk (2001)).

In this paper alternative tests to the homogeneity χ^2-test are considered. The results of Monte Carlo experiments concerning the power of these tests are presented.

2. HOMOGENEITY TESTS FOR DISTRIBUTIONS OF k POPULATIONS

We consider k populations with regard to variables $X_1, ..., X_k$ respectlively. We draw independently $n_{i.}$ elements from i-th population where

$$H_0: \text{Distributions of } X_1, ..., X_k \text{ are identical}$$

* Prof., Chair of Statistical Methods, University of Łódź.

against

$$H_1: \text{Distributions of } X_1, ..., X_k \text{ are not identical.}$$

We assume that set of values of variables $X_1, ..., X_k$ is classified into l categories: $K_1, ..., K_l$.

Let

$$p_{ij} = P(X_i \in K_j) \quad \text{for} \quad i = 1, ..., k \text{ and } j = 1, ..., l \tag{1}$$

The expression $P(X_i \in K_j)$ denotes then the value of variable X_i belongs to category K_j.

If hypothesis H_0 is true that the hypothesis:

$$H_0^*: p_{1j} = p_{2j} = ... = p_{kj} \quad \text{for} \quad j = 1, ..., l$$

is true too.

Let

$$n = \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}. \tag{2}$$

and

$$n_{.j} = \sum_{i=1}^{l} n_{ij} \tag{3}$$

where n_{ij} is a number of sample elements which belong to i-th population and j-th category.

If hypothesis H_0^* is true we can assume that

$$p_{1j} = p_{2j} = ... = p_{kj} = p_j \quad \text{for} \quad j = 1, ..., l \tag{4}$$

and an estimator of p_j has the form:

$$p_j^* = n_{.j}/n \tag{5}$$

In classical homogeneity test for verification of hypothesis H_0 we apply the test statistic:

$$CHI = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} \tag{6}$$

The statistic CHI has asymptotic distribution $\chi^2_{(k-1)(l-1)}$ when hypothesis H_0 is true. In the test we apply right-side region of rejection.

It is possible to propose a different method for verification of hypothesis H_0.

We consider two populations with regard to variables X_1, X_2 respectively and we want to verify hipothesis:

H_0: Distributions of variables X_1 and X_2 are identical

against

H_1: Distributions of variables X_1 and X_2 are not identical.

We draw independently a sample of size n from the first population and a sample of size m from the second population.

Firstly, we consider univariate case, which means that variables X_1, X_2 are univarite.

We can take the following statistic as test statistic:

$$CHI1 = \sum_{j=1}^{l-1} \frac{(n_j/n - m_j/m)^2}{p_{0j}q_{0j}/H} \tag{7}$$

where n_j is a number of elements in the first sample which belong to j-th cathegory, m_j is a number of elements in the second sample which belong to j-th cathegory and: $H = nm/(n+m)$, $p_{0j} = (n_j + m_j)/(n+m)$, $q_{0j} = 1 - p_{0j}$.

Statistic $CHI1$ has asymptotic distribution χ^2_{l-1}. In the test we apply right-side region of rejection.

Now we consider two bivariate variables: $X_1 = (Y_1, Z_1)$ and $X_2 = (Y_2, Z_2)$. For testing of hypothesis H_0' against hypothesis H_0' we can apply the following statistics:

$$CH1 = \sum_{i=1}^{r-1} \sum_{j=1}^{s-1} \frac{(n_{ij}/n - m_{ij}/m)^2}{p_{0ij}\,q_{0ij}/H} \tag{8}$$

or

$$CH2 = \sum_{i=2}^{r} \sum_{j=2}^{s} \frac{(n_{ij}/n - m_{ij}/m)^2}{p_{0ij}\,q_{0ij}/H} \tag{9}$$

where r is a number of categories which are marked out in set of variables Y_1 and Y_2 (the same categories for both variables), s is a number of categories which are marked out in the set of variable Z_1 and Z_2 (the same categories for both variables), n_{ij} is a number of elements which belong to i-th category with regard to values of variables Y_1, Y_2 and j-th category with regard to values of variables Z_1, Z_2 in the first sample, m_{ij} is a number of elements which belong to i-th category with regard to values of variables Y_1, Y_2 and j-th category with regard to values of variables Z_1, Z_2 in the second sample, $H = nm/(n+m)$, $p_{0ij} = (n_{ij} + m_{ij})/(n+m)$, $q_{0ij} = 1 - p_{0ij}$.

Statistics $CH1$ and $CH2$ have asymptotic distribution $\chi^2_{(r-1)(s-1)}$. In the test we apply right-side rejection area.

The distributions of test statistics CHI, $CHI1$, $CH1$, $CH2$ depend on number of categories which are considered in the set of values of variables X_1, X_2. The categories ought to be nonempty and disconnected, and their union ought to be the whole set of values. In the bivariate case for quantitative variables we can propose the following algorithm for creating categories for sets of observations: $\{(y_{11}, z_{11}), ..., (y_{1n}, z_{1n})\}$ from the first population and $\{(y_{21}, z_{21}), ..., (y_{2m}, z_{2m})\}$ from the second population:

– we determine values:

$$a = \min\{y_{11}, ..., y_{1n}, y_{21}, ..., y_{2m}\} \tag{10}$$

$$b = \max\{y_{11}, ..., y_{1n}, y_{21}, ..., y_{2m}\} \tag{11}$$

– we divide interval $[a; b]$ into r intervals (categories) $A_1, ..., A_r$ which have the same length (r is fixed);

– we determine the observations for which the values of variables Y_1 and Y_2 belong to category A_i, $i = 1, ..., r$; we denote the observations by $(y^{(i)}_{1t_1}, z^{(i)}_{1t_1}), ..., (y^{(i)}_{1t_h}, z^{(i)}_{1t_h})$ and $(y^{(i)}_{2\tau_1}, z^{(i)}_{2\tau_1}), ..., (y^{(i)}_{2\tau_g}, z^{(i)}_{2\tau_g})$ for $i = 1, ..., r$;

– for each category A_i, $i = 1, ..., r$, we determine:

$$c_i = \min\{y^{(i)}_{1t_1}, ..., y^{(i)}_{1t_h}, y^{(i)}_{2\tau_1}, ..., y^{(i)}_{2\tau_g}\} \tag{12}$$

$$d_i = \max\{y^{(i)}_{1t_1}, ..., y^{(i)}_{1t_h}, y^{(i)}_{2\tau_1}, ..., y^{(i)}_{2\tau_g}\} \tag{13}$$

– for each i $(i = 1, ..., r)$ we divide interval $[c_i; d_i]$ into s intervals $B^{(i)}_1, ..., B^{(i)}_s$ (s is fixed);

– we create rs categories:

$$A_1 \times B^{(2)}_1, ..., A_1 \times B^{(2)}_s, A_2 \times B^{(2)}_1, ..., A_2 \times B^{(2)}_s, ..., A_r \times B^{(2)}_1, ..., A_r \times B^{(2)}_s$$

In the test we determine a number of observations which belong to categories: $A_1 \times B^{(1)}_1, ..., A_1 \times B^{(1)}_s, A_2 \times B^{(1)}_1, ..., A_2 \times B^{(1)}_s, ..., A_r \times B^{(1)}_1, ..., A_r \times B^{(1)}_s$.

If we want to apply the presented tests, samples from both population ought to be large.

3. MONTE CARLO ANALYSIS OF POWER OF HOMOGENEITY TESTS

Monte Carlo experiments are carried out in order to compare the power of homogeneity tests. For fixed population distributions and for different size of samples the hypothesis about the identity of distributions of two discrete random variables are verified. For given pair of distributions the experiments are repeated 1000 times and a number of cases of rejection of hypothesis H_0 is determined. In case of univariate random variables six categories are mareked out in all experiments and in case bivariate random variables – thirty six categories. The results of calculations are presented in Tab. 1 for univariate distributions and in Tab. 2 for bivariate distributions. In Tab. 1 symbol P_λ denotes Poisson's distribution with parameter λ and symbol $D_{n,p}$ denotes binomial distribution with parameters n and p.

Table 1

Results of simulation experiments concerning the power homogeneity tests in case of univariate discrete random variables

Compared distributions		Size of sample		Number of cases (among 1000 cases) of rejection of hypothesis H_0 (H_0^*) for test statistic		Number of cases (among 1000 cases) of rejection of hypothesis in test for two means
I population	II population	I population	II population	*CHI*	*CHI*1	
1	2	3	4	5	6	7
P_3	P_3	200	300	39	66	54
		800	800	41	68	49
		900	900	42	56	48
		1 000	1 000	44	67	45
P_5	P_5	200	200	33	62	39
		800	800	44	69	41
		900	900	57	81	49
		1 000	1 000	44	65	51
P_{10}	P_{10}	200	200	47	63	52
		800	800	49	76	51
		900	900	47	69	54
		1 000	1 000	52	72	59

Table 1 (condt.)

1	2	3	4	5	6	7
$D_{20\frac{1}{4}}$	$D_{20\frac{1}{4}}$	400 800 900 1 000	500 800 900 1 000	42 37 50 31	66 63 72 43	38 34 53 49
$D_{30\frac{1}{5}}$	$D_{30\frac{1}{5}}$	400 800 900 1 000	500 800 900 1 000	52 49 45 46	58 69 64 70	53 58 34 51
$D_{36\frac{1}{6}}$	$D_{36\frac{1}{6}}$	400 800 900 1 000	500 800 900 1 000	43 53 50 56	61 67 67 75	50 48 43 53
$D_{42\frac{1}{6}}$	$D_{42\frac{1}{6}}$	400 800 900 1 000	500 800 900 1 000	45 53 66 40	60 65 78 70	46 42 51 57
$D_{49\frac{1}{7}}$	$D_{49\frac{1}{7}}$	400 800 900 1 000	500 800 900 1 000	47 44 49 56	59 64 72 73	43 42 45 38
P_3	P_4	400 800 900 1 000	500 800 900 1 000	1 000 1 000 1 000 1 000	1 000 1 000 1 000 1 000	1 000 1 000 1 000 1 000
P_3	P_{10}	200 800 900 1 000	300 800 900 1 000	1 000 1 000 1 000 1 000	1 000 1 000 1 000 1 000	1 000 1 000 1 000 1 000
P_{10}	P_{11}	200 400 800 900 1 000	300 500 800 900 1 000	686 933 997 1 000 1 000	712 948 997 1 000 1 000	934 997 1 000 1 000 1 000
$D_{30;\frac{1}{5}}$	$D_{30;\frac{1}{5}}$	200 400 800 900 1 000	300 500 800 900 1 000	36 57 68 74 64	64 67 88 96 70	56 45 43 50 4
$D_{42;\frac{1}{6}}$	$D_{49;\frac{1}{7}}$	200 400 800 900 1 000	200 500 800 900 1 000	49 54 54 62 55	62 71 80 83 81	42 53 51 47 39

Table 1 (condt.)

1	2	3	4	5	6	7
P_3	$D_{20;\frac{1}{4}}$	200	300	1 000	1 000	1 000
		800	800	1 000	1 000	1 000
		900	900	1 000	1 000	1 000
		1 000	1 000	1 000	1 000	1 000
P_5	$D_{20;\frac{1}{4}}$	100	200	151	151	39
		200	300	319	296	47
		400	500	505	475	43
		600	500	601	577	40
		800	800	828	778	39
		900	900	869	811	50
		1 000	1 000	903	859	42
P_{10}	$D_{20;\frac{1}{4}}$	200	300	1 000	1 000	1 000
		800	800	1 000	1 000	1 000
		900	900	1 000	1 000	1 000
		1 000	1 000	1 000	1 000	1 000
P_{10}	$D_{20;\frac{1}{2}}$	100	200	819	725	47
		200	300	978	935	49
		800	800	1 000	1 000	34
		900	900	1 000	1 000	58
		1 000	1 000	1 000	1 000	43

Source: author's calculations.

Table 2

Results of simulation experiments concerning with the power homogeneity tests in case of bivariate discrete random variables

Compared distributions		Size of sample		Number of cases (among 1000 cases) of rejection of hypothesis H_0 (H_0^*) for test statistic		
I population	II population	I population	II population	*CHI*	*CH*1	*CH*2
1	2	3	4	5	6	7
$(X, X+Y)$	$(X, X+Y)$	400	300	42	46	40
		400	400	41	41	46
		1 000	1 000	42	51	40
		3 000	3 000	48	56	47
		5 000	5 000	69	61	57
$(Z, U+Z)$	$(Z, U+Z)$	400	300	49	50	55
		400	400	39	45	35
		1 000	1 000	53	59	63
		3 000	3 000	50	57	58
		5 000	5 000	42	53	45

Table 2 (condt.)

1	2	3	4	5	6	7
$(W, U+W)$	$(W, U+W)$	400 400 1 000 3 000 5 000	300 400 1 000 3 000 5 000	45 46 46 43 60	45 48 51 55 58	51 32 49 42 54
$(X, X+Y)$	$(Z, U+Z)$	400 400 1 000 3 000 5 000	300 400 1 000 3 000 5 000	1 000 1 000 1 000 1 000 1 000	1 000 1 000 1 000 1 000 1 000	1 000 1 000 1 000 1 000 1 000

Source: author's calculations.

Firstly, we consider univariate case. We can notice that we obtain similar results for two presented tests and the test for two means. When we have two different distributions with the same means than the number of rejection of null hypothesis is a little greater for the test with statistic *CHI* than for the test with statistic *CHI*1 in the case of distributions from different family of distributions. In case of distributions from the same family distributions we observe a little greater number of rejection of null hypothesis for test with statistic *CHI*1, but obtained numbers are not large in comparison with a number of conducted experiments.

For univariate distributions the test power is greater for greater size of sample. For different distributions with the same means the estimates of test power are equal one for given sizes of sample.

We can also notice that the considered homogenous tests are sensitive to differences between means of distributions.

For bivariate case we consider the following variables: $(X, X+Y)$, $(Z, U+Z)$, $(W, U+W)$ where the distributions of variables X, Y, U, W, Z have the form:

$$P(X=1)=0.3,\ P(X=2)=0.2,\ P(X=3)=0.1, \tag{14}$$

$$P(X=4)=0.1,\ P(X=5)=0.2,\ P(X=6)=0.1$$

$$P(Y=1)=0.2,\ P(Y=2)=0.3,\ P(Y=3)=0.1, \tag{15}$$

$$P(Y=4)=0.1,\ P(Y=5)=0.1,\ P(Y=6)=0.2$$

$$P(U=1)=0.15,\ P(U=2)=0.1,\ P(U=3)=0.15,\ P(U=4)=0.2, \quad (16)$$
$$P(U=5)=0.15,\ P(U=6)=0.2,\ P(U=7)=0.05$$

$$P(W=1)=0.05,\ P(W=2)=0.2,\ P(W=3)=0.25,\ P(W=4)=0.2, \quad (17)$$
$$P(W=5)=0.1,\ P(W=6)=0.1,\ P(W=7)=0.1$$

$$P(Z=1)=0.05,\ P(Z=2)=0.15,\ P(Z=3)=0.15,\ P(Z=4)=0.2, \quad (18)$$
$$P(Z=5)=0.15,\ P(Z=6)=0.15,\ P(Z=7)=0.1$$

We assume that variables X, Y, U, W, Z are independent.

We consider three tests with statistics CHI, $CH1$ and $CH2$ for testing of hypothesis H_0. On the basis of Tab. 2 we notice that the results are similar for the tests.

4. FINAL REMARKS

Theoretical considerations and Monte Carlo analysis, which was carried out for homogeneity tests, show that tests with statistic CHI, $CHI1$ can be alternative for univariate random variables and tests with statistics CHI, $CH1$, $CH2$ can be alternative for bivariate random variables.

REFERENCES

Bracha C. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.
Cramer H. (1958), *Metody matematyczne w statystyce*, PWN, Warszawa.
Domański C., Pruska K. (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.
Koronacki J., Mielniczuk J. (2001), *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwo Naukowo-Techniczne, Warszawa.

Krystyna Pruska

WERYFIKACJA HIPOTEZY O ZGODNOŚCI DWÓCH ROZKŁADÓW SKOKOWYCH

Potrzeba badania zgodności rozkładów zmiennych losowych pojawia się przy porównywaniu przebiegu różnych zjawisk. Bardzo często wystarczy zweryfikować hipotezę o równości dwóch średnich, by stwierdzić, że rozkłady nie są jednakowe. Zdarza się jednak, że wartości oczekiwane rozpatrywanych zmiennych są takie same, a jednocześnie nie jest możliwe, na podstawie logicznych przesłanek i wstępnych badań empirycznych, dokładne określenie zbioru wartości rozważanych cech. W takich przypadkach można zaproponować stosowanie testów, wykorzystujących własności wskaźników struktury z próby.

W pracy rozważane są możliwości weryfikacji hipotezy o zgodności dwóch rozkładów skokowych jednowymiarowych i dwuwymiarowych oraz moc rozpatrywanych testów.