METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XVII/3, 2016, s. 33 – 42

TEST ZGODNOŚCI χ^2 OPARTY NA PRÓBACH NIEPROSTYCH

Czesław Domański

Katedra Metod Statystycznych, Uniwersytet Łódzki
e-mail: czedoman@uni.lodz.pl

Streszczenie: Klasyczna teoria wnioskowania statystycznego dostarcza nam metod estymacji nieznanych parametrów rozkładu, szacowania postaci funkcji określającej ten rozkład oraz weryfikację hipotez na podstawie prób prostych, tzn. takich, w których obserwacje są stochastycznie niezależne i mają ten sam rozkład prawdopodobieństwa.
Problemy związane z estymacją, w szczególności metody adaptacji centralnego twierdzenia granicznego dla prób nieprostych oraz weryfikację hipotez o zgodności rozkładów dla prób nieprostych za pomocą testu χ^2 będą przedmiotem tego artykułu.

Słowa kluczowe: estymacja, weryfikacja hipotez, centralne twierdzenie graniczne

UWAGI WSTĘPNE

Klasyczna teoria wnioskowania statystycznego dostarcza nam metod estymacji nieznanych parametrów rozkładu, szacowania postaci funkcji określającej ten rozkład oraz weryfikację hipotez na podstawie prób prostych, tzn. takich, w których obserwacje są stochastycznie niezależne i mają ten sam rozkład prawdopodobieństwa. Na ogół jednak ze względu na koszty i efektywność badania posługujemy się próbami nieprostymi (złożonymi). Wyniki obserwacji w tych próbach są realizacjami stochastycznie zależnych zmiennych losowych o różnych rozkładach. W badaniach reprezentacyjnych wyróżniamy m.in. następujące schematy:

- losowanie zależne (bez zwracania) z różnymi prawdopodobieństwami wyboru,

- warstwowe,
- zespołowe,
- zespołowe i wielostopniowe.

W szczególności, losowanie bez zwracania eliminuje stochastyczną niezależność obserwacji, proces warstwowania powoduje zróżnicowanie prawdopodobieństw wyboru elementów próby, natomiast losowanie wielostopniowe wpływa na różnorodność rozkładów.

Jednym z najczęściej stosowanych w GUS schematów losowania jest schemat losowania dwustopniowego, w którym jednostki pierwszego stopnia są warstwowane i w warstwach wybierane z prawdopodobieństwami proporcjonalnymi do wartości pewnej cechy uznanej za miarę ich wielkości, natomiast jednostki losowania drugiego stopnia są lub nie warstwowe i losowane zgodnie z schematem losowania prostego bez zwracania.

Problemy związane z estymacją, w szczególności metody adaptacji centralnego twierdzenia granicznego dla prób nieprostych oraz weryfikację hipotez o zgodności rozkładów dla prób nieprostych za pomocą testu χ^2 będą przedmiotem tego artykułu.

TWIERDZENIA GRANICZNE DLA POPULACJI SKOŃCZONYCH

Metoda reprezentacyjna zajmuje się sposobami losowania prób z populacji skończonych i szacowania na podstawie otrzymanych prób nieznanych parametrów w tych populacjach. Ponieważ populacje są skończone, to i próby też muszą być skończone. Co więcej, jeżeli N oznacza liczebność populacji generalnej, zaś n - liczebność próby, to rozsądne jest tylko rozpatrywanie sytuacji, w których $n < N$, (przypadki, gdy $n = N$, nie stanowią przedmiotu zainteresowań metody reprezentacyjnej). Względy ekonomiczne i organizacyjne zmuszają statystyków do zastąpienia prób prostych (próba prosta to taka, w której wyniki każdej obserwacji są niezależne mają taki sam rozkład, jak rozkład badanej cechy w populacji) próbami nieprostymi. Ten fakt w połączeniu z założeniem, że $n < N < \infty$ oraz tym, że w wielu sytuacjach n jest zmienną losową uniemożliwia stosowanie znanych z rachunku prawdopodobieństwa twierdzeń granicznych do prób nieprostych.

W przypadku losowania bez zwracania musi być spełniony warunek $n < N$. Tym samym nie możemy korzystać z twierdzeń granicznych znanych z rachunku prawdopodobieństwa, w których zakłada się, że $n \to \infty$. Przytoczymy tutaj twierdzenie Lindeberga Fellera [por. Fisz 1967].

Twierdzenie Lindeberga – Fellera. Niech $\{Y_k\}(k = 1,2,...)$ będzie ciągiem niezależnych zmiennych losowych o skończonych wariancjach oraz $G_k(y)$, μ_k i $\sigma_k \neq 0$ oznaczają odpowiednio dystrybuantę, wartość oczekiwaną i odchylenie standardowe zmiennej Y_k. Niech dalej $C_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2}$, a $F_n(x)$ oznacza dystrybuantę standaryzowanej zmiennej.

$$Z_n = \frac{1}{C_n} \sum_{k=1}^{n} (Y_k - \mu_k)^2,$$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt,$$

a $\Phi(z)$ oznacza dystrybuantę rozkładu N(0,1).

Warunkiem koniecznym i dostatecznym na to, aby

$$\lim_{n\to\infty} \max_{1\le k\le n} \frac{\sigma_k}{C_n} = 0, \qquad \lim_{n\to\infty} F_n(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt \tag{1}$$

jest, by dla dowolnego $\varepsilon > 0$ zachodziła relacja

$$\lim_{n\to\infty} \frac{1}{C_n^2} \sum_{k=1}^{n} \int_{|y-\mu_k|>\varepsilon C_n} (y-\mu_k)^2 dG_k(y) = 0 \tag{2}$$

Zamiast zapisu (2) będziemy stosować następujący:

$$Z_n \xrightarrow{L} N(0,1) \tag{3}$$

W schemacie losowania z prawdopodobieństwami proporcjonalnymi do wartości cechy Y ze zwracaniem, pomimo, że populacja generalna jest skończona można stosować twierdzenie Lindeberga – Fellera. Na podstawie tego twierdzenia można wykazać, że estymator Hansena-Hurwitza [por. Bracha 1998] ma przy $n \to \infty$ rozkład normalny zauważmy, że n może być dowolnie duże (jednostki w próbie mogą być dowolnie duże, a ponadto zmienne Y_i oraz Y_i' $(i \ne i')$ niezależne (losowanie ze zwracaniem).

W przypadku, gdy $n < N$ i zmienne losowe są zależne, to nie mamy podstaw do zastosowania twierdzenia Lindeberga-Fellera do estymacji wartości średniej populacji.

Trudności wynikające z założenia $n < N$ i zależności zmiennych jako pierwszy próbował rozwiązać Madow (1948). Rozpatrywał on, zamiast danej populacji U, ciąg populacji $\{U_\upsilon\}$, który został generowany przez wielokrotne reprodukowanie poszczególnych elementów z populacji U, przy założeniu, że zarówno liczebności tych populacji jak i liczebności prób z nich losowanych rosną nieograniczenie, tzn. przy $\upsilon \to \infty$, $N_\upsilon \to \infty$, $n_\upsilon \to \infty$ oraz $\frac{n_\upsilon}{N_\upsilon} \to q < 1$.

Hajek [1960] przeformułował twierdzenie Madowa, które można przedstawić następująco [por. także Erdös i Réyi 1959].

Twierdzenie Lindeberga-Fellera-Hajka. Dany jest ciąg populacji $\{U_v\}_{v=1}^{\infty}$, gdzie

$$U_v = \{Y_{v1},...,Y_{vN_v}\}, \tag{4}$$

odpowiadający temu ciągowi ciąg parametrów $\{\mathbf{Y}_v\}_{v=1}^{\infty}$, gdzie $\mathbf{Y}_v = (Y_{v1},...,Y_{vN_v})^T$, a także ciąg danych $\{d_v^{'}\}_{v=1}^{\infty}$, gdzie $d_v^{'} = \{y_{v1},...,y_{vn_v}\}$ oraz odpowiadające im ciągi o wyrazach ogólnych

$$\bar{Y}_v = \frac{1}{N_v}\sum_{j=1}^{N_v} Y_{vj}\ , \tag{5}$$

$$S_v^2 = \frac{1}{N_v - 1}\sum_{j=1}^{N_v}(Y_{vj} - \bar{Y}_v)^2\ , \tag{6}$$

$$\bar{y}_v = \frac{1}{n_v}\sum_{j=1}^{n_v} y_{vi}\ . \tag{7}$$

Niech $U_{v\varepsilon}$ będzie dla dowolnego $\varepsilon > 0$ podzbiorem zbioru U_v i niech jego elementy spełniają warunek

$$\left|Y_{vj} - \bar{Y}_v\right| > \varepsilon\sqrt{n_v\left(1 - \frac{n_v}{N_v}\right)}S_v\ , \tag{8}$$

gdy $n_v \to \infty$, $Nv - nv \to \infty$ dla $v \to \infty$.

Warunkiem koniecznym i dostatecznym, aby

$$z_v = \frac{\bar{y}_v - \bar{Y}_v}{\sqrt{\left(\frac{1}{n_v} - \frac{1}{N_v}\right)}S_v} \xrightarrow{L} N(0,1)\ , \tag{9}$$

jest relacja

$$\lim_{v\to\infty}\frac{\sum_{j\in U_{v\varepsilon}}\left(Y_{vj}-\bar{Y}_v\right)^2}{\sum_{j\in U_v}\left(Y_{vj}-\bar{Y}_v\right)^2}=0 \tag{10}$$

Relację (10) nazywa się warunkiem Lindeberga-Hajka.

Scott i Wu [1981] udowadniają dalsze własności estymatorów w przypadku losowania prostego bez zwracania.

Twierdzenie Scotta-Wu. Jeżeli jest spełniony warunek

$$\lim_{v\to\infty}\left(1-\frac{n_v}{N_v}\right)\frac{S_v^2}{n_v}=0, \tag{11}$$

to dla dowolnego $\varepsilon > 0$

$$\lim_{v\to\infty} P\left\{\left|\bar{y}_v-\bar{Y}_v\right|<\varepsilon\right\}=1. \tag{12}$$

Często zamiast równości (12) stosujemy zapis

$$\left(\bar{y}_v-\bar{Y}_v\right)\xrightarrow{P}0.$$

Z twierdzenia Scotta-Wu wynika, że średnia z próby wylosowanej według schematu losowania prostego bez zwracania jest zgodnym z estymatorem średniej $\bar{Y}$.

Hajek [1964] dla schematu losowania z prawdopodobieństwami proporcjonalnymi dla wartości cechy Y bez zwracania, zaproponował procedurę losowania z odrzuceniem (rejective sampling). Procedura ta polega na tym, że dla danych $p_j=\frac{x_i}{X}$, gdzie X oznacza sumę obserwacji zmiennej dodatkowej, określone są wielkościami a_j będące funkcjami p_j i spełniające warunek

$$\sum_{j=1}^{N}a_j=1$$

Następnie jednostki losowane są ze zwracaniem i prawdopodobieństwami wyboru w każdym ciągnieniu proporcjonalnym do a_j. Jeżeli po n ciągnieniach próba zawiera n różnych jednostek, to ją akceptuje. Jeżeli natomiast pewne jednostki powtarzają się, to cała próba jest odrzucona i losowana jest nowa.

Rosen [1972] udowodnił centralne twierdzenie graniczne dla estymatora Horvitza – Thompsona opartego na próbie losowej sekwencyjnej. Prace takie są prowadzone przez wielu badaczy zarówno metodami analitycznymi jak i symulacyjnymi [por. np. Bracha 1990, 1998]. Na podstawie wyników tych badań autorzy sugerują wysoką ostrożność przy wyciąganiu wniosków o zgodności rozkładów rozważanych estymatorów z rozkładem normalnym [por. także monografia Fullera 2009].

TEST ZGODNOŚCI χ^2 DLA PRÓB NIEPROSTYCH

Niech zmienna losowa Y przyjmuje wartości należące do $k(k \geq 2)$ rozłącznych przedziałów. Oznaczmy przez p_i prawdopodobieństwo tego, że zmienna X przyjmuje wartości z i-tego przedziału, przy czym $p_i > 0$ dla $i = 1,...,k$ oraz $\sum_{i=1}^{n} p_i = 1$. Na podstawie próby prostej należy zweryfikować hipotezę:

$$H_0 : \mathrm{p} = \mathrm{p}_0$$

wobec hipotezy alternatywnej:

$$H_1 : \mathrm{p} \neq \mathrm{p}_0,$$

gdzie: $\mathbf{p} = [p_i]_{i=1,...,k-1}$, p_0 jest $(k-1)$ wymiarowym wektorem hipotecznych prawdopodobieństw związanych z $\mathrm{p}(\mathrm{p}_0 = [p_{0i}]_{i=1,...,k-1})$.

Do weryfikacji hipotezy H_0 proponuje się statystykę macierzową [por. np. Rao 1982].

$$\chi^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)^T \, \mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0), \tag{13}$$

gdzie:

$$\mathbf{P}_0 = diag(p_0) - \mathbf{p}_0\mathbf{p}_0^T, \qquad \hat{\mathbf{p}} = [\hat{p}_i]_{i=1,...,k-1} \tag{14}$$

przy czym $\hat{p}_i$ jest nieobciążonym estymatorem p_i.

Przy założeniu prawdziwości hipotezy H_0 statystyka dana wzorem (13) ma asymptotyczny rozkład χ^2 o $k-1$ stopniach swobody.

Dla prób złożonych Holt i inni [1980] przedstawili modyfikacje statystyki testu zgodności χ^2 następującej postaci:

$$\chi_*^2 = \frac{\chi^2}{\hat{\lambda}} \tag{15}$$

gdzie:

$$\hat{\lambda} = \frac{n}{k-1}\sum_{i=1}^{k}\frac{\hat{D}^2(p_i)}{p_{i0}} \tag{16}$$

przy czym $\hat{D}^2(p_i)$ oznaczają estymatory wariancji badanej cechy odpowiednie dla określonego schematu losowania. Statystyka (15) przy założeniu prawdziwości hipotezy H_0 ma rozkład χ^2 o $(k-1)$ stopniach swobody. Hipotezę H_0 odrzucamy na poziomie istotności α, gdy zachodzi nierówność $\chi_*^2 \geq \chi_\alpha^2$.

W przypadku gdy $k=2$, wówczas weryfikujemy hipotezę $H_0: p=p_0$ wobec hipotezy alternatywnej $H_1: p \neq p_0$ za pomocą statystyki [por. np. Bracha 1998].

$$\chi_*^2 = \frac{(\hat{p}-p_0)^2}{\hat{D}^2(p)} \tag{17}$$

gdzie $\hat{p}$ jest estymatorem p.

Statystyka (17) przy prawdziwości hipotezy H_0 ma dla dużych wartości n rozkład zbliżony do rozkładu χ^2 o jednym stopniu swobody.

Przeprowadziliśmy kilka eksperymentów metodą Monte Carlo dla prób nieprostych badając rozmiary testu χ^2 i jego modyfikacji χ_*^2. W pierwszym eksperymencie porównaliśmy rozmiary rozważanych testów dla prób nieprostych (losowanie bezzwrotne) z populacji skończonych o rozkładzie normalnym o zadanych parametrach dla $N=1000$, 2000, 10000. Na podstawie tak wylosowanych prób weryfikowaliśmy hipotezę prostą H_0, że próba pochodzi z populacji o rozkładzie normalnym za pomocą klasycznego testu χ^2 i zmodyfikowanego χ_*^2 uwzględniającego efekt schematu losowania. Badanie zostało przeprowadzone dla kilkunastu wariantów podziału wyników próby na klasy np. $N=1000$ liczby klas $k=2{,}4{,}6{,}8$ odpowiednio dla różnej wielkości próby spełniającej warunki zbieżności statystyki χ^2 do rozkładu chi-kwadrat. Badanie przeprowadzono dla $q=10000$ repetycji.

W tabeli 1 przedstawiono rozmiary rozważanych testów dla trzech poziomów istotności $\alpha=0{,}10$; 0,05; 0,01 i liczby stopni swobody $(lss=7)$ dla $N=1000$ o liczebności próby n = 40, 50,…, 100, 120, dla N = 2000 o liczebności próby n = 50, 100, 150, 200, 300 oraz dla $lss=14$ i $N=10000$ liczebności próby n = 200, 300, 400, 500. Natomiast w tableli 2 dla $N=1000$ prezentowane są rozmiary rozważanych testów w zależności od liczby stopni swobody $(lss=2)$

dla liczebności próby n = 10, 15, 20, 30, 40, 50, 100; $(lss = 4)$ dla n = 15, 20, 30, 40, 50, 100; $(lss = 6)$ dla n = 20, 30, 40, 50, 100.

WNIOSKI

1. Rozmiar klasycznego testu χ^2 dla N =1000 we wszystkich przypadkach przewyższa założone poziomy istotności, natomiast zmodyfikowany test χ^2_* nie przekracza przyjętych poziomów istotności α = 0.10 i α =0.05, a także na ogół dla α =0.01. Podobne rezultaty uzyskaliśmy dla N =10000 (por. tab. 1).
2. Wraz ze wzrostem liczby swobody (lss) na ogół rozmiar klasycznego testu χ^2 coraz bardziej odbiega od przyjętego poziomu istotności α, natomiast wraz ze wzrostem liczby swobody rozmiar zmodyfikowanego testu χ^2_* coraz bardziej zbliża się do przyjętego poziomu istotności (por. tabela 2).

Reasumując, wyniki badań na tym etapie należy wyraźnie podkreślić, że dla próby nieprostej (losowanie zależne) klasyczny test zgodności χ^2 podaje na ogół w rozważanych przypadkach niewłaściwe wskazanie odnoście do weryfikacji hipotezy. Najczęściej w losowaniu bez zwracania rzeczywisty błąd pierwszego rodzaju znacznie przewyższa przyjęty poziom istotności α.

Z dotychczasowych doświadczeń wynika, że rozważane testy w dalszym ciągu badać należy dla prób prostych jak i prób złożonych. Niektóre postulaty wielu autorów dotyczące zasad stosowania testu zgodności χ^2 powinny być zweryfikowane [por. Domański, Pruska 2000].

BIBLIOGRAFIA

Bracha Cz. (1998) Metoda reprezentacyjna w badaniach opinii publicznej i marketingu. EFEKT, Warszawa.

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

Domański Cz. (2007) Verification of Hypotheses Concerning Parameters of the Regression Model for Complex Samples. Acta Universitatis Lodziensis, Folia Oeconomica, 206, 103-112.

Erdös P., Réyi A. (1959) On the central limit theorem for samples from a finite population. Publications of the Mathematics Institute of Hungarian Academy of Science, 4, 49-57.

Fisz M. (1967) Rachunek prawdopodobieństwa i statystyka matematyczna. PWN, Warszawa.

Fuller W. A. (2009) Sampling statistic. John Wiley & Sons, Hoboken, New Jersey.

Hajek J. (1960) Limiting distribution in sample random sampling from a finite populations. Publications of the Matematics Institute of the Hungarian Academy of Science, 5, 361-374.

Hajek J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. Annals of Mathematical Statistics, 1491-1523.

Holt D., Scott A. J., Evings P. D. (1980) Enings chi-squared Test with survey. Journal of the American Statistical Association, A, 143, 303-320.

Madow W. G. (1948) On the limiting distributions of estimates based on sample from finite populations. Annals of Mathematical Statistics, 19, 535-545.

Rao C. R. (1982) Modele liniowe statystyki matematycznej, PWN, Warszawa.

Rosén B. (1972) Asymptotic theory for successive sampling with varying probabilities without replacement: I and II. AMS, 43, 373-397 and 748-776.

Särndal C. E., Swensson B., Wretman J. (1997) Model Assisted Survey Sampling. Springer-Verlag, New York, Berlin, Heidelberg.

Tabela 1. Porównanie rozmiaru testu zgodności χ^2 z testem zmodyfikowanym χ^2_* dla prób nieprostych wylosowanych z populacji skończonych o rozkładzie normalnym dla N =1000, 2000, liczby stopni swobody $lss = 7$ dla N=10000, $lss = 14$

Wielkość próby n	Poziom istotności					
	α = 0,10		α = 0,05		α = 0,01	
	χ^2	χ^2_*	χ^2	χ^2_*	χ^2	χ^2_*
$N = 1000 (lss = 7)$						
40	0,136	0,091	0,071	0,046	0,020	0,010
50	0,128	0,086	0,071	0,046	0,023	0,012
60	0,123	0,085	0,068	0,047	0,018	0,008
70	0,111	0,078	0,067	0,040	0,020	0,013
80	0,105	0,081	0,064	0,042	0,017	0,013
90	0,116	0,090	0,057	0,041	0,014	0,009
100	0,106	0,092	0,062	0,053	0,014	0,010
120	0,111	0,090	0,059	0,048	0,016	0,008
$N = 2000 (lss = 7)$						
50	0,128	0,089	0,064	0,042	0,021	0,012
100	0,102	0,071	0,057	0,034	0,015	0,009
150	0,097	0,082	0,054	0,043	0,010	0,007
200	0,076	0,057	0,033	0,025	0,005	0,005
300	0,083	0,087	0,051	0,052	0,009	0,010
$N = 10000 (lss = 14)$						
200	0,134	0,104	0,081	0,062	0,024	0,012
300	0,116	0,088	0,068	0,053	0,021	0,014
400	0,125	0,106	0,075	0,063	0,010	0,007
500	0,103	0,097	0,049	0,046	0,014	0,009

Źródło: obliczenia własne

Tabela 2. Porównanie rozmiaru testu zgodności χ^2 z testem zmodyfikowanym χ^2_* dla prób nieprostych wylosowanych z populacji skończonych o rozkładzie normalnym dla N =1000 w zależności od liczby stopni swobody lss = 2,4,6

Wielkość próby n	Poziom istotności					
	α = 0,10		α = 0,05		α = 0,01	
	χ^2	χ^2_*	χ^2	χ^2_*	χ^2	χ^2_*
$lss = 2$						
10	0,1145	0,0593	0,0670	0,0318	0,0192	0,0093
15	0,1095	0,0490	0,0612	0,0259	0,0160	0,0070
20	0,1026	0,0497	0,0540	0,0221	0,0147	0,0064
30	0,0979	0,0427	0,0502	0,0202	0,0119	0,0046
40	0,0871	0,0375	0,0441	0,0188	0,0104	0,0040
50	0,0857	0,0379	0,0425	0,0178	0,0084	0,0036
100	0,0722	0,0367	0,0319	0,0163	0,0058	0,0027
$lss = 4$						
15	0,1398	0,0831	0,0813	0,0452	0,0263	0,0128
20	0,1272	0,0756	0,0737	0,0406	0,0224	0,0105
30	0,1204	0,0730	0,0699	0,0368	0,0215	0,0096
40	0,1237	0,0768	0,0682	0,0384	0,0208	0,0101
50	0,1163	0,0715	0,0651	0,0408	0,0187	0,0105
100	0,1056	0,0727	0,0533	0,0367	0,0139	0,0082
$lss = 6$						
20	0,1516	0,0982	0,0906	0,0533	0,0331	0,0167
30	0,1405	0,0930	0,0851	0,0506	0,0285	0,0143
40	0,1327	0,0883	0,0779	0,0492	0,0249	0,0138
50	0,1213	0,0857	0,0727	0,0463	0,0222	0,0119
100	0,1161	0,0934	0,0625	0,0492	0,0160	0,0122

Źródło: obliczenia własne

χ^2 GOODNESS OF FIT TEST IN NON-SIMPLE SAMPLING

Abstract: Classical theory of statistical inference provides methods of estimation of unknown population parameters, density estimation and statistical hypothesis testing, based on simple random sampling, that is sampling scheme, in which all individuals are stochastically independent and identically distributed.

Problems with estimation, especially with adaptation of central limit theorem to non-simple sampling and verification of goodness of fit hypothesis with χ^2 test will be subject of this article.

Keywords: estimation, hypothesis verification, central limit theorem