METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XVI/3, 2015, str. 18 – 25

OGRANICZENIA STOSOWANIA TESTÓW STATYSTYCZNYCH

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

Streszczenie: Celem artykułu jest wskazanie ograniczenia stosowania testów statystycznych opartych na teorii Neymana-Pearsona. Najważniejszą cechą każdego testu jest jego moc, którą możemy zbadać, wtedy gdy jednoznacznie mamy sformułowaną zarówno hipotezę zerową, jak i hipotezę alternatywną. W artykule przedstawimy przykład takiego testu, dla którego określimy empiryczną moc i jego rozmiar.

Słowa kluczowe: model statystyczny, idea Neymana-Pearsona, moc i rozmiar testu, test Coxa, testy dla prób nieprostych

WSTĘP

Celem artykułu jest wskazanie pewnych ograniczeń stosowania testów statystycznych opartych na teorii Neymana-Pearsona.

Przystępując do wnioskowania statystycznego, czyli do sformułowania sądów o zbiorowości na podstawie pobranej z niej próby, konieczne jest ustalenie tzw. trójki probabilistycznej: przestrzeni prób Ω, σ ciała $\aleph$ określonego na Ω i miary probabilistycznej P określonej na $\aleph$.

Niech A będzie wyróżnionym σ-ciałem podzbiorów zbioru $\chi \subset R^n$, zaś X jest mierzalnym przekształceniem $(\Omega,\aleph)\rightarrow(\chi,A)$. Rozkład $P^X(A)=P\left(X^{-1}(A)\right)$ jest miarą na przestrzeni (χ,A). W problemach statystycznych zakłada się, że rozkład P^X należy do pewnej określonej klasy rozkładów $\mathcal{P}$ na (χ,A). Znając tę klasę oraz mając dane wyniki obserwacji zmiennej losowej X, chcemy wysnuć poprawne wnioski o nieznanym rozkładzie P^X. Wobec tego matematyczną podstawą badań statystycznych jest przestrzeń mierzalna (χ,A) i rodzina rozkładów $\mathcal{P}$. Przestrzeń probabilistyczna $(\Omega,\aleph,\mathrm{P})$ odgrywa rolę pomocniczą. Sformułowanie: dana jest przestrzeń probabilistyczna $(\Omega,\aleph,\mathrm{P})$, oznacza, że znany jest model probabilistyczny

pewnego zjawiska lub doświadczenia, czyli wiemy, jakie są możliwe wyniki tego doświadczenia, jakie zdarzenia wyróżniamy oraz jakie prawdopodobieństwa tym zdarzeniom przypisujemy. Reasumując, wiedza a priori o przedmiocie badań jest sformułowana w postaci pewnych modeli probabilistycznych. Probabilistyka może wynikać z samego charakteru badanego zjawiska lub też być wprowadzana przez badacza.

Zauważmy, że $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$ jest rodziną rozkładów prawdopodobieństwa na odpowiednim σ-ciele zdarzeń losowych w χ.

Przestrzeń próby wraz z rodziną rozkładów $\mathcal{P}$, tzn. obiekt:

$$(\chi, \aleph \{P_\theta: \theta \in \Theta\}) \tag{1}$$

nazywamy modelem statystycznym (przestrzenią statystyczną), natomiast odwzorowania z χ w R^k – statystykami lub k-wymiarowymi statystykami.

Jeżeli $\mathbf{X} = (X_1, X_2, ..., X_n)^T$, przy czym $X_1, X_2, ..., X_n$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, to będziemy stosować też oznaczenie:

$$(\chi, \{P_\theta: \theta \in \Theta\})^n, \tag{2}$$

w którym χ jest zbiorem wartości zmiennej losowej X (a więc każdej ze zmiennych $X_1, X_2, ..., X_n$) oraz P_θ to rozkład tej zmiennej losowej. Używa się wtedy również terminologii: $X_1, X_2, ..., X_n$ jest próbą z rozkładu P_θ lub próbą z populacji P_θ dla pewnego $\theta \in \Theta$.

PROBLEMY ANALIZY STATYSTYCZNEJ

Zadaniem analizy statystycznej jest wykrycie, który z rozkładów z modelu statystycznego mógł „wyprodukować" (wygenerować) daną próbę. Ważnym problemem jest prawidłowe zdefiniowanie P. Niewłaściwy wybór P może doprowadzić do błędnego wnioskowania, co w statystycznym żargonie określa się nieraz jako błąd trzeciego rodzaju. Podstawowy warunek właściwego wyboru to szczegółowa znajomość procedury rzeczywiście zastosowanej do uzyskania wyników próby.

Obecny etap rozwoju społeczeństw charakteryzuje się gwałtownym rozpowszechnianiem technologii informacyjnych. Prezentowane metody analizy statystycznej obejmują pewne obszary, które z jednej strony oparte są na przyjętym modelu, a z drugiej strony na metodach nieparametrycznych, które uwzględniają dane pod względem najogólniejszych rodzajów wzorców. Często wymagają zastosowania technik iteracyjnych, w których odpowiednie programy komputerowe pozwalają wykonywać wiele symulacji, a szereg oszacowań prowadzi do ostatecznego rozwiązania.

W przypadku badań częściowych jednym z ważniejszych zagadnień, koniecznych do rozważenia, jest spełnienie lub niespełnienie założenia

o niezależności elementów próby, na podstawie której dokonywane jest wnioskowanie statystyczne. Okazuje się bowiem, iż często w zastosowaniach praktycznych założenie o istnieniu próby prostej nie jest spełnione. Szczególnie dotyczy to prób w obszarze nauk ekonomiczno-społecznych, a zwłaszcza w badaniach marketingowych. W takich przypadkach stosowane są złożone schematy losowania próby, inne niż losowanie niezależne, mogą powodować nielosowość prób, co w konsekwencji prowadzi do powstawania błędów (najczęściej niedoszacowania). Wprowadzone wagi układu wpływają bezpośrednio nie tylko na estymatory parametrów populacji ale również na wariancję tych estymatorów. Rozważania teoretyczne nie dostarczają jednak odpowiednich miar dotyczących zmienności próby, szczególnie dla małych prób.

Prezentowane w literaturze przedmiotu procedury estymacji dla prób złożonych dotyczą w większości parametrów populacji takich jak wartość średnia i wartość globalna, natomiast brak jest propozycji estymacji innych parametrów populacji na przykład nieliniowych miar nierówności oraz korelacji, jak również estymacji parametrów pozycyjnych. Analogiczna sytuacja jest obserwowana w przypadku weryfikacji hipotez statystycznych dla prób złożonych. Fakt, że próba nie jest prosta ma duży wpływ na rozmiar testu. Prowadzone badania wykazały [Domański, Pruska 2000], że dla prób złożonych dla testu niezależności χ^2 faktycznie prawdopodobieństwo błędu I rodzaju przekraczało nawet 0,5, przy założonym poziomie istotności 0,05. Modyfikacje statystyk testowych powodują, że faktyczny rozmiar testu waha się pomiędzy 0,04 a 0,06 [por. np. Bracha 1996].

TEORIA NEYMANA – PEARSONA

Głównym odkryciem Jerzego Spławy Neymana było to, że testy istotności nie mają sensu, o ile nie ma przynajmniej dwóch możliwych hipotez. Zatem nie można weryfikować hipotezy, czy dane pasują do rozkładu normalnego, o ile nie jesteśmy przekonani, że mogą one pasować do jakiegoś innego rozkładu lub klasy innych rozkładów. Wybór tej hipotezy alternatywnej określa sposób realizacji testu istotności. Prawdopodobieństwo wykrycia tej alternatywy, o ile jest poprawna, nazywa się „mocą" testu. W matematyce jasność rozumowania osiąga się nadając wyraźne, dobrze zdefiniowane nazwy określonym koncepcjom. W celu odróżnienia hipotezy, którą wykorzystuje się do wyznaczania p-wartości Fishera, od innych możliwych hipotez, Neyman i Pearson nazwali testowaną hipotezę hipotezą zerową, a pozostałe hipotezą alternatywną. W ich podejściu p-wartości oblicza się, aby przetestować hipotezę zerową, ale moc testu określa, jak zachowa się p-wartość, gdy w rzeczywistości prawdziwa jest hipoteza alternatywna.
Doprowadziło to Neymana do dwóch wniosków.

Pierwszy z nich dotyczy mocy testu, która jest miarą dobroci testu. Z dwóch testów lepiej zastosować ten o większej mocy.

Drugi wniosek związany jest ze zbiorem wariantów hipotezy alternatywnej, który może być zbyt duży. Przeprowadzający analizę nie ma możliwości określenia,

czy dane z próby mają rozkład normalny (hipoteza zerowa), czy dowolny inny możliwy rozkład. To zbyt szeroki zbiór wariantów hipotezy alternatywnej i żaden test nie może mieć wysokiej mocy przeciwko wszystkim możliwym wariantom takiej hipotezy.

OGRANICZENIA WYNIKAJĄCE Z TEORII NEYMANA – PEARSONA

W latach pięćdziesiątych Neyman opracował ideę testu ograniczonych hipotez, gdzie zbiór wariantów hipotezy alternatywnej jest bardzo wąsko określony. Pokazał, że takie testy mają większą moc, niż te, które przyjmują obszerniejsze zbiory hipotezy alternatywnej.

Znane testy zgodności nie zawsze mogą być stosowane w szczególności do weryfikacji np. postaci rozkładów płac i dochodów. Po pierwsze, wiele z tych testów to testy normalności, podczas gdy rozkład normalny dobrze opisuje rozkłady płac jedynie w wyjątkowej sytuacji badania jednorodnych grup pracowników. Po drugie, popularne testy zgodności o szerszym zastosowaniu, jak test χ^2 czy też test λ-Kołmogorowa, nie powinny być stosowane ze względu na założenia, które na ogół są niespełnione. Testy oparte na statystyce Kołmogorowa wymagają, aby parametry rozkładu teoretycznego były znane. Zastosowanie testu χ^2 wydaje się także wątpliwe, gdyż w przypadku bardzo dużych prób (kilka lub nawet kilkudziesięciu tysięcy obserwacji), z jakimi mamy często do czynienia, badając rozkłady płac i dochodów, test ten odrzuca hipotezę zerową nawet wtedy, gdy dopasowanie rozkładów jest niemal pewne.

Przedstawimy testy zgodności Coxa, które uwzględniają postulat Neymana.

Testy zgodności Coxa

Niech H_f oznacza hipotezę zerową, że funkcja gęstości rozkładu populacji generalnej przyjmuje postać $f(y,\theta)$, przy czym θ oznacza wektor nieznanych parametrów rozkładu.

Niech H_g oznacza hipotezę, że funkcja gęstości rozkładu populacji generalnej ma postać $g(y,\eta)$, gdzie η – wektor nieznanych parametrów, przy czym $f(y,\theta)$ i $g(y,\eta)$ są rozłącznymi rodzinami rozkładów. Cox [1961, 1962] wyjaśnia, że pod pojęciem rozłączne rodziny rozkładów (*separate families*) należy rozumieć, że jeden z porównywanych rozkładów nie może być przedstawiony z żądaną dokładnością za pomocą funkcji gęstości drugiego z rozkładów. Rozkład taki nie może być więc szczególnym przypadkiem ani też rozkładem granicznym drugiego. Hipoteza H_g służy wskazaniu rozkładu alternatywnego, dla którego chcemy osiągnąć wysoką moc testu – informacja o tym rozkładzie jest zawarta w sprawdzianie testu (T_f).

Sprawdzianem testu Coxa jest statystyka:

$$T_f = L_f\left(\hat{\theta}\right) - L_g\left(\hat{\eta}\right) - E_{\hat{\theta}}\left[L_f\left(\hat{\theta}\right) - L_g\left(\hat{\eta}\right)\right], \tag{3}$$

gdzie $L_f\left(\hat{\theta}\right), L_g\left(\hat{\eta}\right)$ oznaczają wartości logarytmów funkcji wiarygodności dla oszacowanych metodą największej wiarygodności parametrów rozważanych rozkładów.

Jeżeli przestawimy role H_f i H_g odpowiednio jako hipotezy zerowej i alternatywnej, otrzymujemy statystykę testu postaci:

$$T_g = L_g\left(\hat{\eta}\right) - L_f\left(\hat{\theta}\right) - E_{\hat{\theta}}\left[L_g\left(\hat{\eta}\right) - L_f\left(\hat{\theta}\right)\right]. \tag{4}$$

Statystyki T_f i T_g przy założeniu prawdziwości odpowiednio H_f i H_g mają rozkłady asymptotyczne normalne: $T_f \sim as\ N(0, D(T_f))$, $T_g \sim as\ N(0, D(T_g))$.

W przypadku H_f jako hipotezy zerowej w wyniku przeprowadzenia testu możliwe są następujące decyzje:

- nie ma podstaw do odrzucenia H_f (T_f bliskie zera);
- odrzucamy H_f na korzyść H_g (T_f istotnie różne od zera i ujemne);
- odrzucamy H_f na korzyść innego rozkładu (nie $g(y, \eta)$), gdy T_f jest istotnie różne od zera i dodatnie.

W Tabeli 1 prezentowane są wyniki badania rozmiaru testu Coxa. Zarówno dla rozkładu gamma, jak i dla rozkładu logarytmiczno-normalnego jako rozkładów alternatywnych. Wobec hipotezy zerowej, że rozkład jest Daguma, rozmiar testu Coxa dobrze odzwierciedla przyjęty poziom istotności α. Dla rozkładu gamma dla α = 0,05 rozmiar testu jest zadowalający dla badanych liczebności prób, a dla rozkładu Singha-Maddali dla $n \leq 1000$. (por. Tabela 1).

Wyniki badania mocy testu zgodności Coxa prezentowane są w Tabeli 2, przy czym rozważano przypadki, gdy rzeczywisty rozkład był rozkładem gamma, logarytmiczno-normalnym lub Singha-Maddali, nie zaś przyjętym w hipotezie zerowej rozkładem Daguma. Należy podkreślić, że rozkłady alternatywne dobrane zostały tak, że obejmowały rozkłady o podobnym kształcie i położeniu w stosunku do rozkładu zerowego. Wszystkie są bowiem rozkładami o dodatniej asymetrii a ich parametry oszacowano na podstawie tego samego rozkładu empirycznego. Najbardziej podobnym do rozkładu Daguma jest oczywiście rozkład Singha-Maddali; oba rozkłady należą do systemu rozkładów Burra. Moc testu dla rozkładu gamma jako rozkładu alternatywnego jest bardzo wysoka dla wszystkich liczebności prób i poziomów istotności. Dla rozkładu logarytmiczno-normalnego odsetek właściwych decyzji jest bliski 100% dla $n \geq 2000$, natomiast dla rozkładu Singha-Maddali zadowalające wyniki otrzymano dopiero dla n = 5000. Wynika to z faktu, że dla tej pary rozkładów funkcje gęstości są prawie identyczne. Badanie

przeprowadzono metodą Monte Carlo powtarzając 10000 razy każdy wariant eksperymentu.

Tabela 1. Rozmiar testu Coxa dla wybranych rozkładów alternatywnych

n	Rozkład					
	logarytmiczno-normalny		gamma		Singha-Maddali	
	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
200	0,047	0,011	0,049	0,006	0,051	0,029
300	0,047	0,011	0,049	0,010	0,055	0,031
400	0,053	0,011	0,046	0,014	0,050	0,022
500	0,051	0,012	0,055	0,009	0,043	0,015
600	0,048	0,014	0,049	0,013	0,057	0,023
700	0,053	0,012	0,041	0,013	0,049	0,023
800	0,055	0,014	0,053	0,013	0,055	0,024
1000	0,040	0,014	0,042	0,012	0,053	0,014
2000	0,037	0,015	0,052	0,014	0,062	0,021
5000	0,056	0,015	0,046	0,011	0,061	0,016

Źródło: na podstawie pracy Jędrzejczak [2011]

Tabela 2. Empiryczna moc testu Coxa dla wybranych rozkładów alternatywnych (w %)

n	Rozkład					
	logarytmiczno-normalny		gamma		Singha-Maddali	
	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
200	218	47	951	848	72	30
300	141	24	997	979	72	37
400	523	227	1000	998	97	40
500	643	357	1000	999	139	57
600	716	431	1000	1000	167	80
700	797	543	1000	1000	162	65
800	830	610	1000	1000	208	92
1000	846	594	1000	1000	286	123
2000	991	921	1000	1000	624	289
5000	1000	1000	1000	1000	985	901

Źródło: obliczenia własne

KRYTERIUM WYBORU TESTU

W podejściu Fishera testy istotności pozwoliły uzyskać liczbę, którą nazywał p-wartością. Jest to obliczone prawdopodobieństwo związane z zaobserwowanymi danymi, przy założeniu, że hipoteza zerowa jest prawdziwa.

Jerzy Spława Neyman znalazł sposób przedstawienia testowania hipotez tak, by prawdopodobieństwa związane z decyzjami podjętymi na podstawie testu można było obliczyć. Musiał powiązać p-wartości testu statystycznego z rzeczywistością.

W podejściu Neymana i Pearsona badacz przyjmuje pewną liczbę, taką jak 0,05 i odrzuca hipotezę zerową, gdy tylko p-wartość testu istotności jest nie większa od 0,05. W ten sposób w długiej perspektywie badacz będzie odrzucał prawdziwą hipotezę zerową dokładnie w 5% przypadków. Idea Neymana i Pearsona testowania hipotez ma charakter częstościowego podejścia do prawdopodobieństwa. Jednocześnie Neyman przedstawił problem potrzeby określenia dobrze zdefiniowanej hipotezy alternatywnej, którą testuje się wobec hipotezy zerowej, co pozwala na określenie mocy wybranego testu.

Lehman [1959] napisał monografię z dziedziny testowania hipotez, który pozostaje najbardziej kompletnym dziełem prezentującym podejście Neymana i Pearsona do testowania hipotez statystycznych.

W przypadku, gdy dysponujemy próbami nieprostymi tzn. uzyskanymi na podstawie losowania np. zależnego, warstwowego, złożonego, wielostopniowego, zastosowanie odpowiednich testów dla prób prostych może prowadzić do zmiany własności statystycznych testowanych o czym wspomniano w rozdziale PROBLEMY ANALIZY STATYSTYCZNEJ.

Kish [1965] wprowadził pojęcie współczynnika, który nazwał efektem strategii (schematu) losowania (design effect).

Uwzględniając ten współczynnik uzyskujemy modyfikację testów dla prób prostych, którą można zastosować do odpowiednich prób złożonych opartych na różnych schematach losowania lub ich kombinacjach.

Wnioskowanie statystyczne oparte na próbach nieprostych gwarantuje poprawne decyzyjne wtedy, gdy posługujemy się zmodyfikowanymi statystykami testowymi. Szerszą charakterystykę tych problemów przedstawił Bracha [1996], [por. także Domański, Pruska 2000].

BIBLIOGRAFIA

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

Cox D. R. (1961) Tests of Separate Families of Hypotheses. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Contributions to the Theory of Statistics, 105-123, University of California Press, Berkeley, Calif.

Cox D. R. (1962) Further Results on Tests of Separate Families of Hypotheses, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 24, No. 2 (1962), pp. 406-424.

Domański C., Pruska K. (2000) Nieklasyczne metody statystyczne, Polskie Wydawnictwo Naukowe, Warszawa.

Domański C., Pekasiewicz D., Baszczyńska A., Witaszczyk A. (2014) Testy statystyczne w procesie podejmowania decyzji, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Fisher R. A. (1942) The Design of Experiments, Edinburh.

Jędrzejczak A. (2011) Metody analizy rozkładów dochodów i ich koncentracji, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Lehmann E.L. (1968) Testowanie hipotez statystycznych, PWN, Warszawa.

Neyman J. (1969) Zasady rachunku prawdopodobieństwa i statystyki matematycznej, PWN, Warszawa.
Rao C.R. (1982) Modele liniowe statystyki matematycznej, PWN, Warszawa.

LIMITATIONS OF APPLICABILTY OF STATISTICAL HYPOTHESIS TESTING

Abstract: The aim of this article is to show limits of applications of statistical tests based on the theory Neyman-Pearson. The most important feature of each test is its power, which we can examine if only we had formulated both the null hypothesis and the alternative hypothesis. In this paper we present an example of a test, for which we defined empirical power and size.

Keywords: a statistical model, the Neyman-Pearson idea, power and size of the test, the Cox test, tests for not simple samples