PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu nr 309

RESEARCH PAPERS
of Wrocław University of Economics No. 309

Społeczno-gospodarcze aspekty statystyki

Redaktorzy naukowi

Zofia Rusnak
Edyta Mazurek

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Joanna Szynal

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
ISBN 978-83-7695-398-4

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp 9

Tadeusz Bednarski: Rola Jerzego Spławy-Neymana w kształtowaniu metod statystycznej analizy przyczynowości 11

Filip Borowicz: Ocena możliwości uzupełnienia danych BAEL informacjami ze źródeł administracyjnych w celu dokładniejszej analizy danych o bezrobociu 19

Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Przydatność testów nieparametrycznych Kruskala-Wallisa i mediany w długoterminowej ocenie parametrów kruszyw melafirowych 27

Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Karty kontrolne w ocenie jakości kruszyw dla budownictwa drogowego 42

Czesław Domański: Uwagi o procedurach weryfikacji hipotez z brakującą informacją 54

Stanisław Heilpern: Zależne procesy ryzyka 62

Artur Lipieta, Barbara Pawełek, Jadwiga Kostrzewska: Badanie struktury wydatków w ramach wspólnej polityki UE z wykorzystaniem analizy korespondencji 78

Agnieszka Marciniuk: Dwa sposoby modelowania stopy procentowej w ubezpieczeniach życiowych 90

Beata Bieszk-Stolorz, Iwona Markowicz: Model nieproporcjonalnej intensywności Coxa w analizie bezrobocia 114

Edyta Mazurek: Statystyczna analiza podatku dochodowego od osób fizycznych 127

Katarzyna Ostasiewicz: Awersja do nierówności w modelowaniu użytkowania dóbr wspólnych 159

Piotr Peternek: Porównanie kart kontrolnych indywidualnych pomiarów uzyskanych z wykorzystaniem uogólnionego rozkładu lambda oraz krzywych Johnsona 179

Małgorzata Podogrodzka: Starzenie się ludności a płodność w Polsce w latach 1991-2010 – ujęcie regionalne 192

Renata Rasińska, Iwona Nowakowska: Jakość życia studentów w aspekcie znajomości wskaźników zrównoważonego rozwoju 203

Maria Rosienkiewicz, Jerzy Detyna: Analiza efektywności metod wyboru zmiennych objaśniających do budowy modelu regresyjnego 214
Jerzy Śleszyński: National Welfare Index – ocena nowego miernika rozwoju trwałego i zrównoważonego 236
Maria Szmuksta-Zawadzka, Jan Zawadzki: Wykorzystanie oszczędnych modeli harmonicznych w prognozowaniu na podstawie szeregów czasowych o wysokiej częstotliwości w warunkach braku pełnej informacji..... 261
Anna Zięba: O możliwościach wykorzystania metod statystycznych w badaniach nad stresem 278

Summaries

Tadeusz Bednarski: Role of Jerzy Spława-Neyman in statistical inference for causality 18
Filip Borowicz: Assessing the possibility of supplementing the Polish LFS data with register records for more detailed unemployment data analysis. 26
Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Usefulness of nonparametric Kruskal-Wallis and median tests in long-term parameters assessment of melaphyre crushed rocks 41
Mariusz Donocik, Bogdan Kisiała, Mirosław Mróz, Beata Detyna, Jerzy Detyna: Control charts in the assessment of aggregates quality for road construction..... 53
Czesław Domański: Some remarks on the procedures of the verification of hypotheses under incomplete information..... 61
Stanisław Heilpern: Dependent risk processes 77
Artur Lipieta, Barbara Pawełek, Jadwiga Kostrzewska: Study of the structure of expenditure under the EU's common policy using correspondence analysis 89
Agnieszka Marciniuk: Two ways of stochastic modelling of interest rate in life insurances 113
Beata Bieszk-Stolorz, Iwona Markowicz: The Cox non-proportional hazards model in the analysis of unemployment..... 126
Edyta Mazurek: Statistical assessment of Personal Income Tax 158
Katarzyna Ostasiewicz: Inequality aversion in modeling the use of common pool resources 178
Piotr Peternek: Comparison of control charts of individual measurements based on general Lambda distribution and Johnson curves..... 191
Małgorzata Podogrodzka: The ageing of the population and fertility in Poland in the years 1991-2010 by voivodeships..... 202
Renata Rasińska, Iwona Nowakowska: Students' life quality in terms of knowledge of sustainable development indicators 213

Maria Rosienkiewicz, Jerzy Detyna: Efficiency analysis of chosen methods of explanatory variables selection within the scope of regression model construction 235
Jerzy Śleszyński: *National Welfare Index* – assessment of a new measure of sustainable development 260
Maria Szmuksta-Zawadzka, Jan Zawadzki: The application of harmonic models in forecasting based on high frequency time series in condition of lack of full information 277
Anna Zięba: About statistical methods in the study on stress 284

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 309 • 2013
Społeczno-gospodarcze aspekty statystyki ISSN 1899-3192

Czesław Domański
Uniwersytet Łódzki

UWAGI O PROCEDURACH WERYFIKACJI HIPOTEZ Z BRAKUJĄCĄ INFORMACJĄ*

Streszczenie: Ze względu na trudno dostępne dane występuje konieczność weryfikacji hipotez opartych na próbach z brakującą informacją. Pociąga to za sobą analizę metody pomiaru ilości informacji zawartej w posiadanych obserwacjach, służących konkretnemu testowi. Wiążą się z tym zagadnienia wyboru najlepszej procedury testowej w przypadku posiadania pełnej bądź niepełnej informacji. Przykładem takich badań niekompletnych prób są nauki genetyczne. W genetyce pomiar względnej informacji wynika z potrzeby planowania eksperymentów, porównywania technologii, interpretacji danych i zrozumienia własności metod wnioskowania. W artykule zasygnalizowane zostały problemy dotyczące wnioskowania statystycznego przy niepełnej informacji.

Słowa kluczowe: miary względnej informacji, testy statystyczne, weryfikacja hipotez przy niepełnej informacji.

1. Wstęp

Wiele dyscyplin naukowych o charakterze aplikacyjnym bazuje na procedurach testowania hipotez opartych na próbach z brakującą informacją. Pociąga to za sobą analizę metody pomiaru ilości informacji zawartej w posiadanych obserwacjach, służących konkretnemu testowi. Pomiar zawartości informacji w obserwacjach powinien być odniesiony do pełnej informacji, którą moglibyśmy posiadać, gdyby dane były kompletne. Wiążą się z tym zagadnienia wyboru najlepszej procedury testowej w przypadku posiadania pełnej bądź niepełnej informacji. Staramy się odnieść do coraz częściej występującego w praktyce problemu, gdy badacz dokonał wyboru procedury testowej i chce poznać wpływ brakującej informacji na test w kategoriach relatywnej utraty informacji.

Przykładem takiej dyscypliny są nauki genetyczne. W genetyce pomiar względnej informacji wynika z potrzeby planowania eksperymentów, porównywania technologii, interpretacji danych i zrozumienia własności metod wnioskowania.

* Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2011/01/B/HS4/02746.

W odróżnieniu od podobnego zagadnienia dotyczącego estymacji, gdzie pojęcie udziału brakującej informacji jest dobrze zbadane i opisane w literaturze (por. np. [Dempster, Laird, Rubin 1977; Meng, Rubin 1991]), w przypadku weryfikacji hipotez należy rozważyć, którą z hipotez uznajemy za zerową, a którą za alternatywną.

Procedury testowania hipotez, zwłaszcza o charakterze nieparametrycznym lub quasi-parametrycznym, są zwykle konstruowane w odniesieniu do szczególnego (parametrycznego) modelu. Jednakże bez konkretnego modelu łączącego cechy nieobserwowalne z obserwowalnymi danymi pomiar ilości brakującej informacji nie jest możliwy w przypadku ogólnym. Niektóre odporne procedury estymacji lub testowania mogą dostarczyć bardziej użytecznych wniosków przy mniejszej ilości informacji (por. [Chernoff 1979; Meng 2001]).

W przypadku małych prób pomiar informacji wymaga zastosowania innych miar, jak informacja Fishera, która jest uzasadniona w przypadku dużych prób. Problem małych prób często pojawia się w sytuacji niekompletnych danych, ponieważ brakująca informacja obniża efektywny rozmiar.

Ważnym wyzwaniem jest znalezienie miary (por. [Nicolae, Meng, Kong 2008]), która spełnia następujące warunki:

1) jest wiarygodnym indeksem mówiącym o względnej informacji w odniesieniu do konkretnego celu badawczego,

2) zależy od konkretnego zestawu danych,

3) jest odporna w sensie możliwości uniwersalnego stosowania, również do małych prób,

4) nie nastręcza trudności obliczeniowych,

5) pozostaje zgodna z pewnymi ważnymi aksjomatami odnoszącymi się do łączenia wyników badań.

Warunek wiarygodności (1) jest oczywisty.

Warunek (2) jest konieczny, ponieważ z reguły badaczowi zależy na pomiarze względnej informacji zawartej w konkretnych danych, które ma do dyspozycji, a nie na pomiarze jakiegoś „przeciętnego" zestawu danych.

Warunek (3) jest właściwy, ponieważ w typowym badaniu, np. demograficznych badaniach giełdowych czy w genetyce, istnieje potrzeba pracy z dużymi bazami danych, charakteryzującymi się różnorodnością i złożonością struktur (np. rodzina jednoosobowa lub rodziny wielodzietne). Ponadto badanie zwykle podlega ograniczeniom czasowym, w związku z czym nie jest możliwe projektowanie odrębnych miar dopasowanych do poszczególnych struktur w danych.

Warunek (4) jest ważny, gdyż każda metoda cechująca się dużą złożonością i nieefektywnością algorytmu, bez względu na jej teoretyczną wyższość, prawdopodobnie nie znalazłaby zastosowania w rutynowych badaniach społeczno-ekonomicznych.

Warunek (5) zapewnia pożądany stopień zgodności przy łączeniu wyników badań (np. badania dostarczające więcej informacji otrzymują mniejszą wagę w łącznym indeksie).

2. Imputacja przy założeniu spełnienia hipotezy zerowej

W tym punkcie prześledzimy zagadnienie zastosowania imputacji w przypadku brakujących danych przy testowaniu hipotez, przy założeniu spełnienia hipotezy zerowej. W szczególności przypuśćmy, że $x_1,...,x_n$ są niezależnymi realizacjami pochodzącymi z rozkładu Bernoulliego (p), ale tylko $n_0 < n$ z nich jest obserwowalne w rzeczywistości. Przyjmując założenie, że brakujące dane są rozłożone w próbie w sposób losowy [Rubin 1976], oznaczmy dane zaobserwowane przez $x_1,...,x_{n_0}$. Oczywiście prosty test dla dużych prób (przy założeniu dostatecznie dużego n_0) dla hipotezy $H_0 : p = p_0$ ma statystykę testową postaci (gdzie $\overline{x_e}$ oznacza średnią z obserwacji):

$$T_e = \frac{\overline{x_e} - p_0}{\sqrt{p_0(1-p_0)/n_0}}. \tag{1}$$

Statystyka (1) ma rozkład normalny N(0, 1) przy założeniu prawdziwości hipotezy zerowej. Przypuśćmy, że brakujące x są imputowane z użyciem dwóch metod opartych na średniej. Pierwsza metoda polega na imputowaniu każdego brakującego x przez wartość średnią, czyli $\overline{x_e}$. Druga procedura polega na imputowaniu każdego brakującego x przez wartość średnią przy założeniu prawdziwości hipotezy zerowej. Oczywiście przy każdym ze sposobów imputacji, jeżeli traktujemy dane imputowane jak rzeczywiste obserwacje i stosujemy test (1), gdzie $n_0 = n$, wnioski będą fałszywe, jeśli nie dostosujemy rozkładu normalnego N(0, 1) związanego z hipotezą zerową. W przypadku pierwszej metody średnia ze wszystkich danych, zarówno zaobserwowanych, jak i imputowanych, wynosi $x_1^* = \overline{x_e}$. W związku z tym, jeśli błędnie potraktujemy dane imputowane jako rzeczywiste obserwacje, obliczymy statystykę testową postaci:

$$T_1^* = \frac{\overline{x_1^*} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{1}{\sqrt{r}} T_e, \tag{2}$$

gdzie $r = n_0 / n$.

W odróżnieniu od sprawdzania testu (2) druga metoda doprowadzi do statystyki postaci:

$$T_0^* = \frac{\overline{x_0^*} - p_0}{\sqrt{p_0(1-p_0)/n}} = \sqrt{r} T_e, \tag{3}$$

gdyż średnia ze wszystkich danych, zarówno zaobserwowanych, jak i imputowanych, wynosi

$$x_0^* = r\overline{x_0} + (1-r)p_0.$$

Nasuwają się dwa wnioski z powyższych rozważań.

Po pierwsze, w obydwóch przypadkach otrzymana w wyniku imputacji statystyka testowa oparta na „kompletnych danych" jest proporcjonalna do statystyki danej wzorem (1). W konsekwencji imputacja przy założeniu prawdziwości hipotezy zerowej lub jej pominięcie prowadzi do tej samej odpowiedzi, o ile rozkład związany z hipotezą zerową zostanie odpowiednio dopasowany.

Po drugie, równości (2) i (3) dają odpowiednio:

$$r = \left(\frac{T_e}{T_1^*} \right)^2 \quad \text{i} \quad r = \left(\frac{T_0^*}{T_e} \right)^2 . \tag{4}$$

Wyniki (4) są ważne, ponieważ $r = n_0 / n$ mierzy względną wielkość prób, a zatem „względną informację" przy założeniu niezależności i identyczności rozkładów. Stąd wynika, że pomiar względnej informacji utożsamiany jest z pytaniem, jak duże jest prawdopodobieństwo popełnienia błędu pierwszego rodzaju w pierwszym teście, gdy dane imputowane przy założeniu prawdziwości hipotezy alternatywnej traktowane są jako obserwacje, lub jak bardzo konserwatywny jest drugi test, gdy dane imputowane przy założeniu prawdziwości hipotezy zerowej traktowane są jako obserwacje.

3. Metody dla dużych prób

Podstawą rozważań dla dużych prób jest tożsamość wiążąca współczynniki wiarygodności, w której wartość oczekiwana odpowiada warunkowemu rozkładowi brakujących danych przy danym rozkładzie empirycznym obserwacji.

W szczególności niech X_{co} będzie kompletnym zestawem danych, natomiast X_e posiadanymi informacjami. Zauważmy, że X_e jest funkcją X_{co}. Niech $\ell(\theta|D)$ oznacza logarytm wiarygodności θ pod warunkiem uzyskania danych *D*. Wówczas dla każdego θ_1 oraz każdego θ_2.

$$\begin{aligned} \ell(\theta_1|X_{co}) - \ell(\theta_2|X_{co}) = \left[\ell(\theta_1|X_e) - \ell(\theta_2|X_e) \right] + \\ + \left[\log f(X_{co}|X_e, \theta_1) - \log f(X_{co}|X_e, \theta_2) \right] \end{aligned} . \tag{5}$$

Warunkowa wartość oczekiwana funkcji $f(X_{co}|X_e, \theta)$ przyjmuje postać:

$$E\left[lod(\theta_1, \theta_2|X_{co}) | X_e, \theta \right] = lod(\theta_1, \theta_2|X_e) + \left[\log \frac{f(X_{co}|X_e, \theta_1)}{f(X_{co}|X_e, \theta_2)} | X_e, \theta \right], \tag{6}$$

gdzie $lod(\theta_1, \theta_2|D)$ jest logarytmem szansy, że parametr θ_1 jest prawdziwy zamiast parametru θ_2 pod warunkiem posiadania danych D.

Tożsamość (6) jest rozwinięciem podstawowej tożsamości wyprowadzonej przez Dempstera, Lairda i Rubina [1977] w algorytmie EM. W szczególności, posługując się notacją Dempstera, Lairda i Rubina [1977],

$$Q\left(\theta\middle|\theta'\right)=E\left[\ell\left(\theta\middle|X_{co}\right)\middle|X_e,\theta'\right] \quad \text{i} \quad H\left(\theta\middle|\theta'\right)=E\left[\log f\left(X_{co}\middle|X_e,\theta\right)\middle|X_e,\theta'\right] \tag{7}$$

tożsamość (6) jest równoznaczna z

$$Q\left(\theta_1|\theta\right)-Q\left(\theta_2|\theta\right)=\ell_e\left(\theta_1\right)-\ell_e\left(\theta_2\right)+H\left(\theta_1|\theta\right)-H\left(\theta_2|\theta\right), \tag{8}$$

gdzie $\ell_e\left(\theta\right)\equiv\ell\left(\theta\middle|X_e\right)$.

W pracy Dempstera, Lairda i Rubina [1977] wzór (8) był podany dla przypadku $\theta=\theta_2$ oraz stanowił podstawę wyprowadzenia własności zbieżności algorytmu EM. Intuicyjnie, jeśli parametr θ_1 jest prawdziwy, wówczas jeśli dysponowalibyśmy większą ilością danych, które pochodziłyby z rozkładu opisanego funkcją $f\left(X_{co}\middle|X_e,\theta_1\right)$, otrzymalibyśmy przeciętnie wyższą wartość *lod* niż $lod\left(\theta_1,\theta_2\middle|X_e\right)$. Zakładając $\theta=\theta_1$, równanie (6) przyjmuje postać:

$$E\left[lod\left(\theta_1,\theta_2\middle|X_{co}\right)\middle|X_e,\theta_1\right]=lod\left(\theta_1,\theta_2\middle|X_e\right)+KL\left(\theta_1:\theta_2\right)\geq lod\left(\theta_1,\theta_2\middle|X_e\right), \tag{9}$$

gdzie $KL\left(\theta_1:\theta_2\right)\geq 0$ jest miarą informacji Kullbacka-Leiblera.

Miara ta świadczy o prawdziwości θ_1 zamiast θ_2, w przypadku gdy parametr θ_1, występujący w rozkładzie warunkowym względem obserwacji, jest prawdziwy. Nierówność (9) zamienia się w równość wówczas, gdy $KL\left(\theta_1:\theta_2\right)=0$, co ma miejsce wtedy i tylko wtedy, gdy $f\left(X_{co}\middle|X_e,\theta_1\right)=f\left(X_{co}\middle|X_e,\theta_2\right)$, co oznacza, że pod warunkiem posiadania obserwacji X_e, dodatkowe dane nie wniosłyby żadnej informacji pozwalającej rozróżnić θ_1 i θ_2. Odległość Kullbacka-Leiblera była szeroko wykorzystywana w teorii informacji (np.[Cover, Thomas 1991]) i statystyce matematycznej (np. [Aitchison 1975]). Ostatnie badania dotyczące wykorzystania funkcji straty K-L obejmują między innymi pozycję George'a, Fenga i Xu [2006] oraz zawarte tam odniesienia do literatury.

Podobnie jeśli parametr θ_2 jest prawdziwy, wówczas oczekujemy przeciętnie niższej wartości $lod\left(\theta_1,\theta_2\middle|X_{co}\right)$, niż gdybyśmy obserwowali X_{co}. Można to wykazać, podstawiając $\theta=\theta_2$ w równaniu (6), co prowadzi do

$$E\left[lod\left(\theta_1,\theta_2\middle|X_{co}\right)\middle|X_e,\theta_2\right]=lod\left(\theta_1,\theta_2\middle|X_e\right)-KL\left(\theta_2:\theta_1\right)\leq lod\left(\theta_1,\theta_2\middle|X_e\right). \tag{10}$$

Wówczas nierówność zamienia się w równość wtedy i tylko wtedy, gdy

$$f\left(X_{co}\middle|X_e,\theta_1\right)=f\left(X_{co}\middle|X_e,\theta_2\right).$$

Zauważmy, że wszystkie wartości oczekiwane (6), (9) i (10) są warunkowe względem X_e, możliwe jest więc dopuszczenie zależności któregokolwiek parame-

tru θ od X_e. W szczególności wartość parametru θ_0 przyjęta w hipotezie zerowej H_0 może być zarówno stałą lub, bardziej ogólnie, estymatorem metody największej wiarygodności z restrykcjami θ, przy funkcji wiarygodności $\ell\left(\theta \middle| X_e\right)$.

4. Miara względnej informacji w dużej próbie świadcząca przeciwko hipotezie zerowej

Przypuśćmy, że wartość parametru zakładana w hipotezie zerowej wynosi θ_0 oraz że estymator MNW parametru θ przy prawdziwości hipotezy alternatywnej dla danych X_e wynosi θ_e. Ponadto zakładamy, że (I) parametr θ_e jest jednoznaczny – jest to warunek, który automatycznie zachodzi dla dużych prób oraz (II) $\theta_e \neq \theta_0$, co prawie zawsze jest spełnione w praktyce.

Następnie, ponieważ dążymy do zmierzenia informacji zawartej w brakujących danych, by odrzucić hipotezę zerową, przyjmując założenie dużej próby, naturalnym sposobem postępowania jest potraktowanie θ_e jako prawdziwy parametr oraz pomiar wartości oczekiwanej funkcji straty względem wartości oczekiwanej logarytmu ilorazu szans (*lod score*) przy kompletnych danych. Zatem definiujemy

$$RI_1 = \frac{lod\left(\theta_e, \theta_0 \middle| X_e\right)}{E\left[lod\left(\theta_e, \theta_0 \middle| X_{co}\right) \middle| X_e, \theta_e\right]} = \frac{\ell_e\left(\theta_e\right) - \ell_e\left(\theta_0\right)}{Q\left(\theta_e \middle| \theta_e\right) - Q\left(\theta_0 \middle| \theta_e\right)}. \tag{11}$$

Wzór (11) pokazuje, że obliczenie RI_1 wymaga jedynie ocen logarytmów funkcji wiarygodności $\ell_e\left(\theta\right)$ opartych na posiadanych obserwacjach oraz funkcji Q, dla $\theta = \theta_0$ oraz $\theta = \theta_e$, przy czym ta ostatnia wartość może zostać zaczerpnięta z algorytmu EM.

Przy założeniach (I) i (II) wskaźnik RI_1 jest dobrze zdefiniowany oraz, biorąc pod uwagę równanie (9), $0 < RI_1 \leq 1$. Współczynnik ten osiąga wartość 1 wtedy i tylko wtedy, gdy $KL\left(\theta_e : \theta_0\right) = 0$, co oznacza, że brakujące dane nie wnoszą informacji pozwalającej na rozróżnienie θ_e i θ_0, zatem nie ma brakującej informacji w zbiorze obserwacji X_e. Współczynnik ten osiąga 0 wtedy i tylko wtedy, gdy $lod\left(\theta_e, \theta_0 \middle| X_e\right) / KL\left(\theta_e : \theta_0\right) \to 0$, co można wytłumaczyć w ten sposób, że jeśli funkcja wiarygodności oparta na obserwacjach ma coraz mniejszą zdolność do rozróżnienia parametrów θ_e i θ_0 w stosunku do modelu z brakującymi danymi [mierzonymi $KL\left(\theta_e : \theta_0\right)$], to, wskazując na odrzucenie hipotezy zerowej, brakująca informacja zbliża się do 100%.

5. Względna miara informacji przy założeniu prawdziwości hipotezy zerowej dla dużych prób

Z nierówności (10) wynika zastosowanie względnej miary informacji przy założeniu prawdziwości hipotezy zerowej H_0 dla dużych prób. Podstawiając $\theta_1 = \theta$ oraz $\theta_2 = \theta_0$ w (10), otrzymujemy

$$E\left[lod\left(\theta,\theta_0 \middle| X_{co}\right) \middle| X_e,\theta_0 \right] = lod\left(\theta,\theta_0 \middle| X_e\right) - KL\left(\theta_0 : \theta\right) \le lod\left(\theta,\theta_0 \middle| X_e\right). \quad (12)$$

Zatem, gdy dodatkowe dane pochodzą z rozkładu opisanego funkcją $f\left(X_{co} \middle| X_e,\theta_0\right)$, oczekiwana całkowita wartość „lod" nie może przewyższyć wartości uzyskanej dla danych zaobserwowanych dla żadnego θ. Jako najlepszy estymator wartości „lod" dla pełnych danych możemy wykorzystać $\max_\theta E\left[lod\left(\theta,\theta_0 \middle| X_{co}\right) \middle| X_e,\theta_0 \right]$, co nie może przewyższyć $lod\left(\theta_e,\theta_0 \middle| X_e\right)$, tak jak to pokazuje nierówność (12); wykorzystanie estymatora punktowego wartości „lod" dla pełnych danych, bez uwzględnienia stopnia niepewności związanego z jego oceną, można uzasadnić założeniem o dużym rozmiarze próby. W konsekwencji możemy zdefiniować

$$RI_0 = \frac{\max_\theta E[lod\left(\theta,\theta_0 \middle| X_{co}\right) \middle| X_e,\theta_0]}{lod\left(\theta_e,\theta_0 \middle| X_e\right)} = \frac{\max_\theta \left[Q\left(\theta \middle| \theta_0\right) - Q\left(\theta_0 \middle| \theta_0\right) \right]}{\ell_e\left(\theta_e\right) - \ell_e\left(\theta_0\right)}. \quad (13)$$

Na podstawie wzoru (13) można wnioskować o efektywności algorytmu, ponieważ $\max_\theta Q\left(\theta \middle| \theta_0\right)$ jest wartością, która występuje w kroku E oraz kroku M algorytmu EM (por. [Stuart, Ord 1991]), jeśli przyjmiemy, że wartością z poprzedniej iteracji był parametr $\theta = \theta_0$.

Podobnie jak współczynnik RI_1, $0 \le RI_0 \le 1$. Natomiast, w odróżnieniu od RI_1, badanie sytuacji, kiedy RI_0 zbliża się do zera lub jedności, jest dość złożone, zwłaszcza przy dużych różnicach pomiędzy θ_e i θ_0. Częściowo wynika to z faktu, że RI_0 jest zdefiniowany przy założeniu prawdziwości (w przybliżeniu) hipotezy zerowej, które zostałoby odrzucone przy dużych wartościach $\delta = \theta_e - \theta_0$.

Zauważmy, że przy dodatkowym założeniu, że θ_e jest jednoznacznym punktem stacjonarnym $\ell_e\left(\theta\right)$, licznik wskaźnika RI_0 jest równy zeru wtedy i tylko wtedy, gdy mianownik jest równy zeru, zatem gdy $\ell_e\left(\theta_e\right) = \ell_e\left(\theta_0\right)$.

Istnieje związek pomiędzy niską wartością wskaźnika RI_0 a prawdopodobieństwem związanym z obserwacjami i jego niewielką możliwością do rozróżnienia między θ_0 a θ_e, tak samo jak w przypadku RI_1.

6. Uwagi końcowe

Przedstawione problemy dotyczące wnioskowania statystycznego przy niepełnej informacji wskazują na potrzebę dalszych badań. Autorzy Nicolae, Meng i Kong [2008] podali zastosowania przedstawionych miar (por. p. 4 i 5) w genetyce w szcze-

gólności dotyczące mapowania genów, badania związków genomów z cechami osobniczymi w celu określenia lokalizacji genów odpowiedzialnych za wrażliwość na pewne zachowania. Wydaje się, że zastosowanie można rozszerzyć na procesy demograficzne i ekonomiczno-społeczne.

Literatura

Aitchison J., *Goodness of prediction fit*, „Biometrika" 1975, 62, ss. 547-554.
Mathematical Reviews (MathSciNet): MR391353.
Chernoff H., *Sequential Analysis and Optimal Design,* SIAM, Philadelphia 1979, PA.
Dempster A.P., *The direct use of likelihood for significance testing*, „Statist. Comput." 1997, 7, ss. 247--252.
Dempster A.P., Laird N.M. and Rubin D.B., *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, „J. Roy. Statist. Soc. Ser. B" 1977, 39, ss. 1-37.
George E., Feng L. and Xu, X., *Improved minimax predictive densities under Kullback–Leibler loss*, „Ann. Statist." 2006, 34, ss. 78-91. MR2275235.
Meng X.-L., *A congenial overview and investigation of multiple imputation inference under uncongeniality*, [w:] (R. Groves, D. Dillman, J. Eltinge and R. Little eds.), *Survey Nonresponse*, Wiley, New York 2001, ss. 343-356.
Meng X.-L. and Rubin D. B., *Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm*, „J. Amer. Statist. Assoc" 1991, 86, ss. 899–909.
Nicolae D.L., Meng X-L and Kong A., *Quantifying the Fraction of Missing Information for Hypothesis Testing in Statistical and Genetic Studies,* „Statistical Science" 2008, 23, ss. 287-312.
Rubin D.B., *Inference and missing data*, „Biometrika" 1976, 63, ss. 581-592.
Mathematical Reviews (MathSciNet): MR455196.
Stuart A., Ord K., *Kendall's Advanced Theory of Statistics*, Vol. 2, Edward Arnold, London 1991.

SOME REMARKS ON THE PROCEDURES OF THE VERIFICATION OF HYPOTHESES UNDER INCOMPLETE INFORMATION

Summary: Due to the hard access data a necessity appears to verify hypotheses based on samples with missing information. That implies the analysis of the method of measuring the information quantity contained in the sample used in a given test. The result is the problem of choosing the best test procedure in the case of possessing complete or incomplete information. An example of such incomplete research is genetic science. In genetics the measurement of relative information results from the necessity of experiment planning, technology comparisons, data interpretation and the understanding of the properties of inference methods. In the paper some problems concerning statistical inference with incomplete information are considered.

Keywords: relative information measures, statistical tests, hypotheses verification with incomplete information.