PRZEGLĄD STATYSTYCZNY
R. LX – ZESZYT 4 – 2013

SABINA DENKOWSKA

NIEKLASYCZNE PROCEDURY TESTOWAŃ WIELOKROTNYCH

1. WPROWADZENIE

Testowanie wielokrotne jest powszechne w analizach statystycznych. Zdobycie i przetworzenie danych często jest czasochłonne i kosztowne, naturalną tendencją wśród badaczy jest więc testowanie znacznej liczby hipotez na raz zgromadzonych danych. Badacz stara się wykryć maksymalnie wiele zależności i formułuje dziesiątki hipotez w poszukiwaniu istotnych zależności statystycznych.

Niestety często zdarza się, że liczne testowania są prowadzone każde na poziomie istotności α, a wnioski są podsumowywane łącznie, a przecież wraz ze wzrostem liczby rozpatrywanych hipotez rośne prawdopodobieństwo wykrycia pozornie istotnych statystycznie związków. Jeśli rozważymy teoretycznie testowanie m prawdziwych, niezależnych hipotez zerowych, każdą na poziomie istotności α, to prawdopodobieństwo odrzucenia przynajmniej jednej prawdziwej hipotezy zerowej wynosi $1 – (1 – \alpha^m)$. Już w przypadku 20 niezależnych, prawdziwych hipotez zerowych, testowanych każda na poziomie istotności 0,05, prawdopodobieństwo odrzucenia co najmniej jednej prawdziwej hipotezy wynosi 0,64, a wartość oczekiwana liczby błędnych odrzuceń wynosi 1. W praktyce niezmiernie rzadko mamy do czynienia z niezależnymi testowaniami, co znacznie utrudnia kontrolę efektu testowania wielokrotnego.

Typowe sytuacje badawcze w których mamy do czynienia z testowaniem wielokrotnym to porównywanie parami wartości przeciętnych, czy też testowanie istotności kontrastów, czyli kombinacji liniowych wartości przeciętnych, dla których suma współczynników wynosi zero. Gdy są spełnione założenia modelu analizy wariancji, wówczas rozwiązaniem mogą okazać się klasyczne procedury *post-hoc* powszechnie dostępne w pakietach statystycznych. Ale nawet wówczas wybór właściwej procedury nie jest zadaniem prostym. Wielość zaawansowanych i złożonych procedur, różnorodność uzyskiwanych wyników w zależności od wybranej procedury oraz trudności interpretacyjne mogą zniechęcać praktyków do stosowania procedur wnioskowań wielokrotnych (patrz np. Denkowska, 2005). A przecież testowanie wielokrotne to nie tylko porównywanie wartości przeciętnych, czy testowanie istotności kontrastów. Z testowaniem wielokrotnym mamy również do czynienia m.in. przy badaniu istotności współczynników korelacji w macierzy korelacji, porównywaniu terapii z zabiegiem kontrolnym lub przy testowaniu istotności ocen parametrów strukturalnych w modelu

regresji. W wielu sytuacjach badawczych jedynym sposobem kontroli efektu testowania wielokrotnego jest sięgnięcie po nieklasyczne procedury testowań wielokrotnych.

Celem artykułu jest przegląd nieklasycznych procedur testowań wielokrotnych, które umożliwiają kontrolę efektu testowania wielokrotnego w sytuacjach, gdy niespełnione są restrykcyjne założenia modelowe klasycznych procedur testowań wielokrotnych lub gdy rozwiązań klasycznych po prostu brak.

Należy podkreślić, że w sytuacji, gdy testowane hipotezy nie są ze sobą powiązane ani zawartością, ani późniejszym wykorzystaniem, należy traktować je oddzielnie, a nie łącznie. W przeciwnym jednak przypadku konieczny jest łączny pomiar błędów. Gdy wniosek końcowy wysnuwany jest na podstawie przeprowadzonych testów analizowanych łącznie i jego trafność zależy od łącznego pomiaru błędów dla danego zbioru wnioskowań, wtedy taki zbiór wnioskowań powinien być rozpatrywany łącznie jako *rodzina* (Hochberg, Tamhane, 1987). Termin ten został wprowadzony ponad pół wieku temu przez Tukeya (1953) i przez lata towarzyszyła mu dyskusja dotycząca tego, jakie kwestie powinny decydować o składzie rodziny (patrz np. Hochberg, Tamhane, 1987; Miller, 1981). Temat ten do tej pory budzi sporo kontrowersji.

1.1. WYBRANE MIARY BŁĘDU I RODZAJU DLA RODZINY WNIOSKOWAŃ

W celu zaprezentowania najważniejszych miar błędu I rodzaju dla rodziny wnioskowań przyjmijmy pomocniczo następujące oznaczenia: niech V oznacza liczbę prawdziwych hipotez zerowych odrzuconych w procesie testowania m hipotez zerowych, a R – liczbę odrzuconych hipotez zerowych. V i R są to zmienne losowe. Po przeprowadzeniu testowania znana jest tylko liczba hipotez R, które odrzucamy, a tym samym liczba hipotez zerowych, dla których nie mamy podstaw do odrzucenia: $m - R$. Wartość zmiennej losowej V nie jest obserwowana.

W literaturze tematu można spotkać wiele propozycji miar błędu I rodzaju dla rodziny wnioskowań. Do najważniejszych należą FWER i FDR.

Miara FWER nawiązująca do tradycyjnego rozumienia weryfikacji hipotez zdefiniowana jest następująco:

$$\text{FWER (}\textit{Family-Wise Error Rate}\text{)}: \quad \text{FWER} = P(V > 0). \tag{1}$$

Procedury kontrolujące FWER na ustalonym poziomie α zapewniają spełnienie warunku, że prawdopodobieństwo odrzucenia co najmniej jednej prawdziwej hipotezy zerowej nie przekroczy α. W monografii „The Problem of Multiple Comparisons" Tukey (1953) porównywał różne miary kontroli błędu I rodzaju dla rodziny wnioskowań i twierdził, iż to właśnie „kontrola FWER powinna być standardem" (Hochberg, Tamhane, 1987). Niestety, wraz ze wzrostem liczby weryfikowanych hipotez, maleje moc procedur kontrolujących FWER rozumiana jako zdolność procedur do wykrywania fałszywych hipotez zerowych.

W 2005 roku Lehmann i Romano (2005) zaproponowali miarę gFWER będącą uogólnieniem FWER:

$$\text{gFWER (}\textit{generalized FWER}\text{):}\quad \text{gFWER} = P(V > k),\quad k = 0,1,\ldots,m. \tag{2}$$

Propozycja ta nie stanowi jednak rozwiązania problemu spadku mocy procedur w przypadku bardzo licznych zbiorów, złożonych z tysięcy, czy nawet setek tysięcy wnioskowań spotykanych np. w genetyce. Przy tak bogatych zbiorach wnioskowań również procedury kontrolujące gFWER nie sprawdzają się, gdyż charakteryzują się bardzo małą mocą – indywidualne poziomy istotności są tak małe, iż rzadko dochodzi do odrzuceń hipotez zerowych.

W przypadkach bardzo licznych zbiorów wnioskowań warto rozważyć kontrolę miary FDR (Hochberg, Benjamini, 1995). Hochberg i Benjamini (1995) zaproponowali, by zamiast kontroli liczby błędnych odrzuceń, kontrolować wartość oczekiwaną frakcji błędnych odrzuceń pośród wszystkich odrzuceń hipotez zerowych:

$$\text{FDR (}\textit{False Discovery Rate}\text{):}\quad \text{FDR} = \begin{cases} E\left(\frac{V}{R}\right) & dla\ R > 0, \\ 0 & dla\ R = 0. \end{cases} \tag{3}$$

Wraz z miarą FDR zaproponowali oni procedurę[1] kontrolującą wartość oczekiwaną frakcji błędnych odrzuceń na, z góry zadanym, poziomie q ($q = \alpha$). Oznacza to, że stosując tę procedurę akceptujemy q100% błędnych odrzuceń hipotez zerowych wśród wszystkich odrzuceń.

Dla unaocznienia różnicy pomiędzy FWER i FDR, rozważmy rodzinę złożoną z tysiąca hipotez zerowych oraz odpowiadających im hipotez alternatywnych. Porównajmy sytuację polegającą na odrzuceniu 100 hipotez zerowych z których 1 jest prawdziwa, z sytuacją, gdy odrzucone zostały 2 hipotezy z których 1 jest prawdziwa. Z punktu widzenia FWER obie te sytuacje są tak samo niekorzystne, bo odrzucona została jedna prawdziwa hipoteza zerowa. Natomiast z punktu widzenia FDR w drugiej sytuacji błędnych odrzuceń było aż 50%, podczas gdy w pierwszej sytuacji tylko 1%.

Kontrola FWER jest bliższa tradycyjnemu statystycznemu wnioskowaniu. Nie zawsze jest jednak satysfakcjonującym rozwiązaniem. W przypadku licznych rodzin wnioskowań kontrola FDR stanowi istotną alternatywę.

Wspomniane miary błędu I rodzaju dla zbioru wnioskowań to zaledwie niewielka część propozycji miar wymienianych w literaturze tematu (patrz np. Hochberg, Tamhane, 1987; Dudoit, van der Laan, 2008), niemniej jednak są to miary kluczowe, najczęściej spotykane w testowaniu wielokrotnym.

---

[1] Opis procedury w podrozdziale 2.2.

1.2. SKORYGOWANE PRAWDOPODOBIEŃSTWA TESTOWE

Z testowaniem wielokrotnym ściśle związane jest pojęcie skorygowanych prawdopodobieństw testowych (ang. *adjusted p-value*). Pierwsze idee dotyczące skorygowanych prawdopodobieństw testowych pojawiły się u Rosenthala i Rubina (1983). Wright (1992) propagował stosowanie skorygowanych prawdopodobieństw testowych we wnioskowaniu jednoczesnym, podkreślając zalety takiego prezentowania wyników procedur.

Analogicznie do zwykłych prawdopodobieństw testowych (ang. *p-value*), skorygowanym prawdopodobieństwem testowym $\tilde{p}_i$ dla dowolnej hipotezy $H_{0,i}$ vs. $H_{A,i}$ nazywamy najmniejszą wartość FWER, przy której dana hipoteza zerowa $H_{0,i}$ zostałaby odrzucona, gdy rozpatrywana jest cała rodzina hipotez. Analogicznie są definiowane skorygowane prawdopodobieństwa testowe w przypadku innych miar błędu I rodzaju dla rodziny wnioskowań.

Skorygowane prawdopodobieństwa mają liczne zalety. Są łatwe do interpretacji, gdyż mając podane ich wartości, decyzję o ewentualnym odrzuceniu hipotezy podejmujemy porównując odpowiadające jej skorygowane prawdopodobieństwo testowe z przyjętym łącznym poziomem istotności dla całej rodziny wnioskowań. Wskazują, jak mocne są podstawy do odrzucenia hipotezy zerowej w kontekście kontroli wybranej miary błędu I rodzaju dla całego zbioru wnioskowań. Można również łatwo porównywać różne procedury, porównując ich prawdopodobieństwa skorygowane (mniejsze wartości skorygowanych prawdopodobieństw testowych wskazują na mniej konserwatywną procedurę).

2. PROCEDURY BRZEGOWE TESTOWAŃ WIELOKROTNYCH

W ostatnich latach znaczną popularność zyskały proste obliczeniowo brzegowe procedury testowań wielokrotnych (ang. *marginal MTP*). Metody te mogą być stosowane w przypadku skończonych rodzin hipotez, składających się tylko z hipotez minimalnych. Proces testowania przy wykorzystaniu tych procedur opiera się przede wszystkim na analizie zbioru prawdopodobieństw testowych otrzymanych z indywidualnych wnioskowań. Procedury te charakteryzują się szerokim zakresem zastosowań i mogą być stosowane zarówno w przypadku porównań wartości przeciętnych (patrz np. Denkowska, 2005), jak i testowania istotności współczynników korelacji w macierzach korelacji (Denkowska, 2006) czy badania istotności współczynników regresji w modelu regresji (Denkowska, 2011a,b).

W celu zaprezentowania procedur przyjmijmy następujące założenia oraz oznaczenia. Rozpatrzmy rodzinę m minimalnych hipotez zerowych $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$ z odpowiadającymi im prawdopodobieństwami testowymi $p_1, p_2, \ldots, p_m$. Uporządkujmy prawdopodobieństwa testowe $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ i niech $H_{(0,1)}, H_{(0,2)}, \ldots, H_{(0,m)}$ oznaczają odpowiadające im hipotezy zerowe.

2.1. BRZEGOWE PROCEDURY TESTOWAŃ WIELOKROTNYCH KONTROLUJĄCE FWER

Najstarszą, a zarazem najprostszą procedurą brzegową testowań wielokrotnych jest procedura Bonferroniego. Procedura Bonferroniego jest procedurą uniwersalną, czyli można ją stosować w przypadku dowolnej rodziny wnioskowań, bez względu na typ zależności pomiędzy statystykami testowymi. Proces testowania można przedstawić następująco. Rozważmy testowanie m hipotez zerowych $H_{0,i}$ vs. $H_{A,i}$ ($i = 1, \ldots, m$). Hipotezę zerową $H_{0,i}$ odrzucamy wtedy, gdy odpowiadające jej prawdopodobieństwo testowe p_i jest nie większe od $\frac{\alpha}{m}$. Niezwykle wygodnie jest przeprowadzać proces testowania w oparciu o skorygowane prawdopodobieństwa testowe, które dla tej procedury wyznaczane są ze wzoru:

$$\tilde{p}_j = \min(mp_j; 1) \quad \text{dla} \quad j = 1, \ldots, m. \tag{4}$$

Decyzję o odrzuceniu hipotezy zerowej $H_{0,i}$, podejmujemy, gdy odpowiadające jej skorygowane prawdopodobieństwo testowe $\tilde{p}_i$ jest nie większe od α.

Procedura Bonferroniego[2] zapewnia kontrolę FWER na poziomie α, ale jest metodą bardzo konserwatywną, czyli ma małą moc. Konserwatyzm ten jest tym poważniejszy, im silniejsze są zależności pomiędzy statystykami testowymi lub im liczniejsza jest rodzina wnioskowań.

Mniej konserwatywna jest uniwersalna procedura Holma, która jest wieloetapową modyfikacją procedury Bonferroniego. Każda hipoteza odrzucona przez metodę Bonferroniego jest odrzucona również przez metodę Holma, natomiast hipotezy odrzucone przez metodę Holma mogą nie zostać odrzucone przez metodę Bonferroniego. Prawdopodobieństwa skorygowane w metodzie Holma wyznaczamy ze wzorów:

$$\tilde{p}_{(1)} = \min\left(mp_{(1)}; 1\right),$$

$$\tilde{p}_{(j)} = \min\left(\max(\tilde{p}_{(j-1)};\ (m - j + 1)\, p_{(j)}); 1\right) \text{ dla } j = 2, \ldots, m. \tag{5}$$

W przypadku, gdy rozpatrywane statystyki testowe tworzą wielowymiarowy rozkład normalny lub rozkład t-Studenta o niezależnych składowych, a rozważane hipotezy alternatywne mają dwustronne zbiory krytyczne, do kontroli efektu testowania wielokrotnego można zastosować modyfikacje procedury Bonferroniego oraz procedury Holma oparte na nierówności Šidáka (Shaffer, 1995; Hochberg, Tamhane, 1987, s. 366). I tak modyfikacja metody Bonferroniego polega na zastąpieniu $\frac{\alpha}{m}$ przez $1-\left(1-\alpha\right)^{\frac{1}{m}}$. Skorygowane prawdopodobieństwa testowe dla procedury Bon-

[2] Procedura Bonferroniego występuje również pod nazwą „poprawki Bonferroniego".

ferroniego-Šidáka, spotykanej również pod nazwą ŠidákSS (ang. *single-step*) wyznaczane są ze wzoru[3]:

$$\tilde{p}_j = \min\left(1; 1 - \left(1 - p_j\right)^m\right) \text{ dla } \quad j = 1, \ldots, m. \tag{6}$$

Skorygowane prawdopodobieństwa testowe dla procedury Holma-Šidáka, występującej w literaturze pod nazwą ŠidákSD (ang. *step-down*), obliczamy ze wzorów[4]:

$$\tilde{p}_{(1)} = \min\left(1; 1 - \left(1 - p_{(1)}\right)^m\right),$$

$$\tilde{p}_{(j)} = \min\left(1; \max\left(\tilde{p}_{(j-1)};\ 1 - \left(1 - p_{(j)}\right)^{m-j+1}\right)\right) \text{ dla } \quad j = 2, \ldots, m. \tag{7}$$

Jak wykazali Holland i Copenhaver (1987) procedury ŠidákSS oraz ŠidákSD zapewniają kontrolę FWER również w przypadku, gdy statystyki testowe mają dodatnią zależność orthantową (Denuit, Scaillet, 2004). Do grupy tej należą na przykład statystyki *t*-Studenta, wykorzystywane przy testowaniu równości wartości przeciętnych, pochodzących z populacji o rozkładach normalnych w jednoczynnikowym modelu analizy wariancji (Denuit, Scaillet, 2004, Shaffer, 1995).

Spośród procedur brzegowych największą moc mają procedury wieloetapowe typu *step-up*, zapewniające kontrolę FWER w przypadku statystyk testowych niezależnych lub silnie dodatnio skorelowanych. Skomplikowana obliczeniowo procedura Hommela daje tylko nieznacznie lepsze wyniki od procedury Hochberga (np. Westfall i in., 1999), dla której skorygowane prawdopodobieństwa testowe wyznaczane są ze wzorów:

$$\tilde{p}_{(m)} = p_{(m)} \text{ oraz } \tilde{p}_{(m-j)} = \min\left(\tilde{p}_{(m-j+1)}; (j+1)p_{(m-j)}\right) \text{ dla } j = 1, \ldots, m - 1. \tag{8}$$

Wymieniając najważniejsze procedury brzegowe zapewniające kontrolę FWER należy wspomnieć o procedurze Shaffer dla hipotez logicznie powiązanych (Shaffer, 1986). Przykładem hipotez logicznie powiązanych mogą być hipotezy o równości wartości przeciętnych parami dla co najmniej trzech populacji. Zauważmy, że w rzeczywistości niemożliwe jest, aby $\mu_1 = \mu_2$ oraz $\mu_2 = \mu_3$, ale $\mu_1 \neq \mu_3$. Shaffer uznała więc, że w przypadku porównywania wartości przeciętnych trzech populacji nie ma potrzeby rozpatrywać sytuacji, gdy odrzucamy jedną hipotezę zerową, a przy dwóch stwierdzamy, że nie mamy podstaw do ich odrzucenia i zaproponowała modyfikację uniwersalnej procedury Holma, która dzięki uwzględnieniu logicznych relacji pomiędzy hipotezami ma większą moc, a kontrola FWER na poziomie α jest nadal zagwarantowana.

---

[3] Por. np. Westfall i in. (1999).

[4] Ibid.

2.2. BRZEGOWE PROCEDURY TESTOWAŃ WIELOKROTNYCH KONTROLUJĄCE FDR

Procedury kontrolujące FDR to procedury zapewniające kontrolę wartości oczekiwanej frakcji błędnych odrzuceń wśród wszystkich odrzuceń, na z góry zadanym poziomie. Wybór przez badacza tej miary błędu I rodzaju dla rodziny wnioskowań oznacza, iż dopuszcza on i akceptuje pewien niewielki odsetek $q\,(q = \alpha)$ błędnych odrzuceń wśród wszystkich odrzuceń.

Hochberg i Benjamini (1995) zaproponowali procedurę gwarantującą kontrolę FDR w przypadku niezależnych statystyk testowych, dla której skorygowane prawdopodobieństwa testowe wyznaczane są z następujących wzorów:

$$\tilde{p}_{(m)} = p_{(m)} \quad \text{oraz} \quad \tilde{p}_{(m-j)} = \min\left(\tilde{p}_{(m-j+1)}; \frac{m}{m-j} p_{(m-j)}\right) \quad \text{dla} \quad j = 1, \ldots, m-1. \tag{9}$$

Bejamini i Yekutieli (2001) pokazali, że procedura ta zapewnia kontrolę FDR również w przypadku statystyk testowych o zależności dodatnio regresyjnej.

Uniwersalną modyfikację procedury Hochberga i Benjaminiego, która zapewnia kontrolę FDR bez względu na typ zależności pomiędzy statystykami testowymi zaproponowali Benjamini i Yekutieli (2001). Prawdopodobieństwa skorygowane wyznaczane są ze wzorów:

$$\tilde{p}_{(m)} = min\left(1;\ p_{(m)} \sum_{i=1}^{m} \frac{1}{i}\right), \tag{10}$$

$$\tilde{p}_{(m-j)} = min\left(p_{(m-j+1)};\ p_{(m-j)} \frac{m}{m-j} \sum_{i=1}^{m} \frac{1}{i}\right) \quad dla \quad j = 1, \ldots, m-1.$$

Niestety, procedura Benjaminiego-Yekutieliego jest bardzo konserwatywna.

2.3. ZALETY I WADY PROCEDUR BRZEGOWYCH

Niepodważalną zaletą brzegowych procedur testowań wielokrotnych jest ich prostota obliczeniowa. W przypadku niektórych procedur testowań wielokrotnych np. procedur Hochberga–Benjaminiego, Hommela, Hochberga, czy też procedur opartych na nierówności Šidáka, istotną kwestią jest typ zależności pomiędzy statystykami testowymi dla którego procedury te zapewniają kontrolę wybranej miary błędu I rodzaju dla rodziny wnioskowań. Procedury te charakteryzują się zazwyczaj większą mocą w stosunku do procedur uniwersalnych Bonferroniego, czy Holma, ale wymogi dotyczące zależności pomiędzy statystykami znacznie ograniczają zakres ich zastosowań i komplikują zastosowanie. Procedury uniwersalne można stosować w przypadku

dowolnego zbioru wnioskowań, bez względu na typ zależności pomiędzy statystykami testowymi, ale ich wadą jest mniejsza moc w porównaniu do procedur, które uwzględniają łączny rozkład statystyk testowych. Uniwersalne procedury brzegowe testowań wielokrotnych umożliwiają więc kontrolę efektu testowania wielokrotnego w wielu sytuacjach badawczych, w których rozwiązań klasycznych brak. W sytuacjach, gdy klasyczne procedury testowań wielokrotnych są dostępne np. w modelu jednoczynnikowej analizy wariancji, zarówno badania empiryczne, jak i badania symulacyjne efektywności procedur klasycznych oraz procedur brzegowych zastosowanych do wyodrębniania jednorodnych podgrup wartości przeciętnych pokazują, że procedury brzegowe stanowią poważną konkurencję dla rozwiązań klasycznych, a ich efektywność jest nieznacznie gorsza, zaś w niektórych przypadkach wręcz porównywalna z efektywnością procedur klasycznych (Denkowska, 2007b, Denkowska 2005).

W programie R w pakiecie *multtest* dostępna jest funkcja *mt.rawp2adjp*, która zwraca skorygowane prawdopodobieństwa testowe dla wybranych brzegowych procedur testowań wielokrotnych.

3. PROCEDURY ŁĄCZNE TESTOWAŃ WIELOKROTNYCH

Procedury łączne testowań wielokrotnych (ang. *joint MTP*) uwzględniają łączny rozkład statystyk testowych i dzięki temu są mniej konserwatywne niż procedury brzegowe.

3.1. PROCEDURY ŁĄCZNE YOUNGA, WESTFALLA KONTROLUJĄCE FWER

Young i Westfall (1993) zaproponowali procedury łączne testowań wielokrotnych wykorzystujące repróbkowanie. Procedury te oparte są na regule domknięcia (Domański, Pruska, 2000, s. 201; Hochberg, Tamhane, 1987). Zastosowanie resamplingu umożliwia przeprowadzanie testowania wielokrotnego mimo braku normalności, czy też braku znajomości struktury kowariancyjnej danych. Procedury Westfalla i Younga oparte są na maksimach statystyk testowych *maxT* lub minimach prawdopodobieństw testowych *minP*.

Wadą tych procedur jest wymóg ***obrotowości podzbioru*** (ang. *subset pivotality*), który oznacza, że łączny rozkład statystyk testowych dla dowolnego podzbioru I zbioru wszystkich rozważanych wnioskowań $\{1,\ldots,m\}$, ma być taki sam, zarówno pod warunkiem prawdziwości wszystkich hipotez zerowych $H_{0,i}$ dla $i \in I$, jak i pod warunkiem prawdziwości globalnej hipotezy zerowej H_0^C, głoszącej, że wszystkie hipotezy zerowe $H_{0,i}$ ($i \in \{1,\ldots,m\}$) są prawdziwe. W przypadku procedur opartych na maksimach statystyk testowych oznacza to, że dla dowolnego $I \subseteq \{1,\ldots,m\}$ rozkład $\max_{i \in I} T_i \,|\, H_I$ musi być taki sam, jak rozkład $\max_{i \in I} T_i \,|\, H_0^C$.

Warunek obrotowości podzbioru jest bardzo istotny, zwłaszcza gdy resampling wykorzystuje rozkład generujący dane przy założeniu prawdziwości wszystkich

hipotez zerowych, pozwala to bowiem uprościć algorytm procedury opartej na regule domknięcia i zamiast testować $2^m - 1$ przecięć hipotez zerowych, wystarczy przeprowadzić m testowań. Niestety w wielu sytuacjach badawczych warunek ten nie jest spełniony. Należy do nich np. testowanie istotności współczynników korelacji. Rozkład generujący dane przy założeniu prawdziwości hipotez zerowych może dawać łączny rozkład statystyk testowych inny od prawdziwego (rzeczywistego) rozkładu. Rozważmy badanie istotności trzech współczynników korelacji ρ_{12}, ρ_{13}, ρ_{23}. Aitken (1969, 1971) wykazał (patrz Westfall, Young, 1993), że gdy $H_{0,12}$ oraz $H_{0,13}$ są prawdziwe, a $H_{0,23}$ jest fałszywa, to łączny rozkład statystyk testowych odpowiadających prawdziwym hipotezom zerowym jest w przybliżeniu normalny, zależny od współczynnika korelacji ρ_{23}, czyli warunek obrotowości podzbioru nie jest spełniony.

Oznaczmy przez $T_1, \ldots, T_m$ statystyki testowe, a przez $P_1, \ldots, P_m$ zmienne losowe oznaczające prawdopodobieństwa testowe związane z hipotezami zerowymi $H_{0,1}, \ldots, H_{0,m}$.

W jednoetapowych procedurach łącznych Westfalla i Younga prawdopodobieństwa skorygowane wyznaczane są ze wzorów:

Procedura single-step maxT

$$\tilde{p}_i = Pr\left(\max_{1\le j\le m}|T_j| \ge |t_i| \,\middle|\, H_0^C\right) \tag{11}$$

w przypadku dwustronnych zbiorów krytycznych.

Procedura single-step minP

$$\tilde{p}_i = Pr\left(\min_{1\le j\le m} P_j \le p_i \,\middle|\, H_0^C\right). \tag{12}$$

Przyjmijmy pomocniczo, że uporządkowane prawdopodobieństwa testowe mają indeksy $r_1, \ldots, r_m$ takie, że $p_{(1)} = p_{r_1}, \ldots, p_{(m)} = p_{r_m}$. Wówczas skorygowane prawdopodobieństwa w procedurach typu step-down Westfalla i Younga można zapisać następująco:

Procedura step-down maxT

$$\tilde{p}_{(1)} = Pr\left(\max_{l\in\{r_1,\ldots,r_m\}}|T_l| \ge \left|t_{(1)}\right| \,\middle|\, H_0^C\right), \tag{13}$$

$$\tilde{p}_{(j)} = \max\left(Pr\left(\max_{l\in\{r_j,\ldots,r_m\}}|T_l| \ge \left|t_{(j)}\right| \,\middle|\, H_0^C\right); \tilde{p}_{(j-1)}\right) \text{ dla } j = 2, \ldots, m,$$

w przypadku dwustronnych zbiorów krytycznych.

Procedura step-down minP

$$\tilde{p}_{(1)} = Pr\left(\min_{l\in\{r_1,..,r_m\}} P_l \leq p_{(1)} \middle| H_0^C\right), \tag{14}$$

$$\tilde{p}_{(j)} = \max\left(Pr\left(\min_{l\in\{r_j,..,r_m\}} P_l \leq p_{(j)} \middle| H_0^C\right); \tilde{p}_{(j-1)}\right) \text{ dla} j = 2, \dots, m.$$

Procedury Westfalla i Younga uwzględniają łączny rozkład statystyk testowych i dzięki temu mają większą moc niż procedury brzegowe. Procedury te niekoniecznie muszą opierać się na resamplingu. W niektórych sytuacjach badawczych prawdopodobieństwa $Pr\left(\max_{l\in\{r_j,..,r_m\}} |T_l| \geq \left|t_{(j)}\right| \middle| H_0^C\right)$ można wyznaczyć z wielowymiarowego rozkładu normalnego lub *t*-Studenta i wówczas resampling nie jest konieczny.

W programie R w pakiecie *multtest* dostępne są zstępujące (*step-down*) procedury permutacyjne *mt.maxT* oraz *mt.minP* Westfalla i Younga (1993).

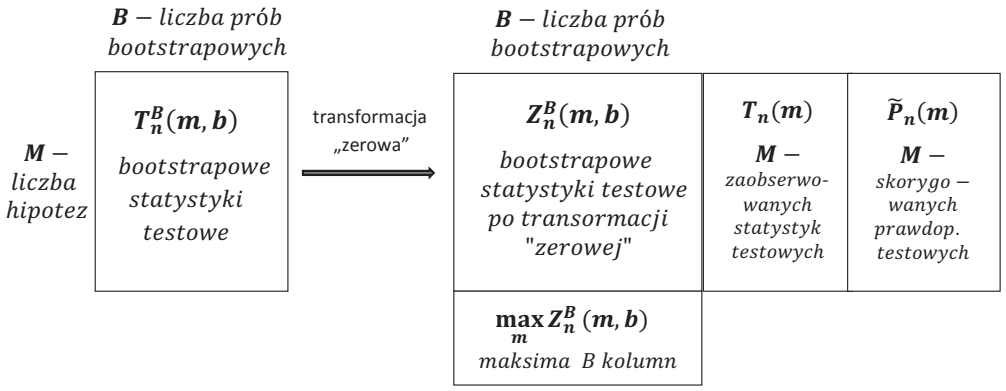
3.2. PROCEDURY ŁĄCZNE TESTOWAŃ WIELOKROTNYCH DUDOIT I VAN DER LAANA

Ciekawą, nową propozycję w literaturze tematu stanowią procedury łączne oparte na bootstrapie (Dudoit, van der Laan, 2008), które w odróżnieniu od propozycji Westfalla i Younga, nie opierają się na rozkładzie[5] generującym dane który spełnia hipotezę, że wszystkie hipotezy zerowe są prawdziwe, ale na rozkładzie „zerowym" statystyk testowych, czyli rozkładzie statystyk testowych przy założeniu prawdziwości hipotezy zupełnej. Autorzy zaproponowali dwa rodzaje „zerowych" transformacji bootstrapowych statystyk testowych: przesunięcie i skalowanie (ang. *null shift and scale*) oraz kwantylowe przekształcenie (ang. *null quantile*). Bootstrapowym estymatorem „zerowego" rozkładu statystyk testowych jest rozkład otrzymany w wyniku transformacji „zerowej" bootstrapowych statystyk testowych.

Kontrolę miary FWER zapewnia zastosowanie łącznych procedur jednoetapowych (ang. *single-step* SS) lub wieloetapowych zstępujących (ang. *step–down* SD), opartych na maksimach statystyk testowych *maxT* lub na minimach prawdopodobieństw testowych *minP*.

Na rys. 1 przedstawiony jest schemat wyznaczania skorygowanych prawdopodobieństw testowych w przypadku zastosowania jednoetapowej procedury *maxT* (dla prawostronnych zbiorów krytycznych). Skorygowane prawdopodobieństwo testowe $\tilde{P}_n(m)$ $(m = 1, \dots, M)$ dla hipotezy $H_0(m)$ jest frakcją maksimów bootstrapowych $\max_m Z_n^B(m, b)$ nie mniejszych od zaobserwowanej wartości statystyki testowej $T_n(m)$.

[5] Rozkład generujący dane może dać w efekcie rozkład łączny statystyk testowych o innej strukturze zależnościowej niż ich prawdziwy rozkład (gdy niespełniony jest warunek obrotowości podzbioru).



Rysunek 1. Schemat wyznaczania skorygowanych prawdopodobieństw testowych w przypadku zastosowania procedury jednoetapowej *SS.maxT* (dla prawostronnych zbiorów krytycznych)

Źródło: opracowane na postawie Dudoit, van der Laan (2008).

Procedury zaproponowane przez Dudoit, van der Laana (2008) są zaimplementowane w pakiecie *multtest* w R pod nazwą MTP. Procedurę MTP można zastosować do porównań parami wartości przeciętnych, do testowania istotności współczynników regresji w modelu regresji, testowania istotności współczynników korelacji oraz w wielu innych sytuacjach badawczych. Użytkownik określa *statystyki testowe* (statystyki są determinowane poprzez wybór testu np. *t.twosamp.equalvar*, *t.cor*, *f*), *miarę błędu I rodzaju* (FWER, gFWER, TPPFP[6], FDR), *metodę estymacji rozkładu „zerowego” statystyk testowych* (bootstrap z centrowaniem i skalowaniem *boot.sc*, bootstrap z transformacją kwantylową *boot.qt*, permutacje) oraz *procedurę łączną testowań wielokrotnych* (SSmaxT, SSminP, SDmaxT, SDminP).

Dudoit i van der Laan (2008) dedykowali procedurę MTP badaniom genetycznym, specyficznym ze względu na bardzo liczne rodziny wnioskowań, składające się z tysięcy hipotez zerowych. Wstępne eksperymenty symulacyjne (opisane poniżej) pokazują jednak na konieczność dodatkowych badań nad procedurą MTP.

3.3. EKSPERYMENT SYMULACYJNY

W celu sprawdzenia kontroli miary błędu I rodzaju FWER przez procedurę MTP przeprowadzono eksperyment symulacyjny. Eksperyment polegał na generowaniu M prób n-elementowych z rozkładu normalnego $N(0,1)$ i testowaniu hipotez postaci: $H_{0,i}: \mu_i = 0$ vs. $H_{A,i}: \mu_i \neq 0$ ($i = 1, \ldots, M$).

Eksperyment powtarzano co najmniej 3000 razy i szacowano prawdopodobieństwo właściwych decyzji, czyli rozpoznania, że wszystkie hipotezy zerowe są prawdziwe.

[6] TPPFP (Tail Probability for Proportion of False Positives): $TPPFP = P\left(\frac{V}{R} > q\right)$ gdzie $q \in (0,1)$.

W badaniu wykorzystano funkcję MTP zaimplementowaną w pakiecie *multtest* w R. Parametrami funkcji MTP były m.in. test t-Studenta dla wartości oczekiwanej (*t.one.samp)*, wartość weryfikowana ustawiona domyślnie na 0, liczba prób bootstrapowych B równa 1000 oraz FWER = 0,05.
Prawdopodobieństwo właściwego rozpoznania, że wszystkie hipotezy zerowe są prawdziwe, szacowano w zależności od:
– metody estymacji rozkładu „zerowego" statystyk testowych (*boot.sc*, *boot.qt*),
– procedury łącznej testowań wielokrotnych (*SSmaxT*, *SSminP*, *SDmaxT*, *SDminP*).
W badaniu rozpatrywano:
– próby o liczebności *n*: 10, 30, 100,
– liczby testowanych hipotez zerowych *M* wynosiły: 100, 200, 400.

Rezultaty badań symulacyjnych przedstawiono w tabeli 1.

Tabela 1.

Wyniki badań symulacyjnych

n	*M*	*SSmaxT boot.cs*	*SSminP boot.cs*	*SDmaxT boot.cs*	*SDminP boot.cs*	*SSmaxT boot.qt*	*SSminP boot.qt*	*SDmaxT boot.qt*	*SDminP boot.qt*
10	100	0,977	0,972	0,975	0,974	0,937	0,903	0,936	0,907
30	100	0,948	0,918	0,948	0,917	0,948	0,901	0,947	0,904
100	100	0,935	0,899	0,935	0,895	0,945	0,902	0,944	0,903
10	200	0,985	0,953	0,984	0,953	0,936	0,819	0,935	0,818
30	200	0,954	0,848	0,953	0,846	0,949	0,820	0,949	0,826
100	200	0,937	0,814	0,938	0,817	0,946	0,820	0,946	0,812
10	400	0,987	0,909	0,988	0,884	0,940	0,646	0,937	0,661
30	400	0,956	0,717	0,954	0,711	0,947	0,662	0,947	0,681
100	400	0,948	0,671	0,949	0,683	0,967	0,682	0,960	0,699

Źródło: opracowanie własne.

Zdecydowanie nie wszystkie otrzymane wyniki można uznać za satysfakcjonujące. Stosunkowo najlepsze wyniki dały procedury oparte na maksimach statystyk testowych. Bardzo słabo wypadły natomiast procedury oparte na minimach prawdopodobieństw testowych *minP*, a zwłaszcza te z nich, w których „zerowa" transformacja bootstrapowych statystyk testowych opierała się na przekształceniu kwantylowym, szczególnie w sytuacji bardzo dużej liczby testowanych hipotez. Sytuacji takiej nie zaobserwowano natomiast w równolegle prowadzonym badaniu nad procedurami brzegowymi testowań wielokrotnych, które bez względu na liczbę testowanych hipotez oraz liczebności prób potwierdziły kontrolę miar błędu I rodzaju (FWER, FDR).

W badaniach nad procedurą MTP przyjęto ustawioną domyślnie w procedurze liczbę próbkowań (B = 1000). Wstępne eksperymenty polegające na zwiększeniu liczby prób bootstrapowych pokazują, że im większa liczba testowanych hipotez tym poprawa prawdopodobieństwa właściwego rozpoznania jest wyraźniejsza. Badanie prowadzono w przypadku prób o małej liczebności (n = 10). Liczbę prób bootstrapowych zwiększono 10-krotnie (B = 10000). I tak, w przypadku procedury *sdMinP* (*boot.qt*) dla 400 hipotez prawdopodobieństwo właściwego rozpoznania wzrosło do 0,85; dla 200 hipotez zaobserwowano mniejszą poprawę – wzrost prawdopodobieństwa właściwego rozpoznania do 0,88, a dla 100 hipotez korekta prawdopodobieństwa była wręcz niezauważalna. Pomimo 10-krotnego zwiększenia liczby próbkowań, otrzymane prawdopodobieństwa właściwego rozpoznania nadal nie można uznać za satysfakcjonujące.

Rezultaty eksperymentu symulacyjnego okazały się zaskakujące, okazało się bowiem, że MTP nie zawsze gwarantuje kontrolę FWER dla zbioru wnioskowań na z góry ustalonym poziomie. Co więcej, wstępne rozpoznanie wskazuje również na pewną niestabilność wyników zależną od liczby próbkowań. W 2009 roku Werft i Benner (2009) sygnalizowali problemy z kontrolą miary FDR przez procedurę MTP w przypadku bardzo dużej liczby testowanych hipotez i małej liczebności prób. Konieczne są zatem dalsze badania nad procedurami Dudoit i van der Laana wyjaśniające przyczyny problemów z kontrolą wspomnianych miar błędu I rodzaju dla zbioru wnioskowań.

4. PODSUMOWANIE

Kontrola efektu testowań wielokrotnych jest bezsprzecznie konieczna. Niekontrolowane testowanie wielokrotne prowadzi do wykrywania wielu zupełnie przypadkowych zależności. Zależności takie są później często eksponowane w naukowych, jak również popularno-naukowych publikacjach jako ciekawe, czy wręcz zdumiewające wyniki badań. To z kolei może budzić sceptycyzm wobec metod statystycznych, podczas gdy źródłem nieporozumienia są niewłaściwie przeprowadzone badania, nieuwzględniające efektu testowania wielokrotnego.

Z powodu rygorystycznych założeń modelowych procedury klasyczne testowań wielokrotnych mają znacznie ograniczony zakres zastosowań. Co więcej, w wielu sytuacjach badawczych rozwiązań klasycznych brak. Dlatego też, coraz większe zainteresowanie zdobywają nieklasyczne brzegowe oraz łączne metody testowań wielokrotnych. Popularne, proste obliczeniowo i o szerokim zakresie zastosowań uniwersalne procedury brzegowe nie uwzględniają łącznego rozkładu statystyk testowych, przez co są bardziej konserwatywne od procedur łącznych. Z kolei zakres zastosowań procedur łącznych zaproponowanych przez Westfalla i Younga jest ograniczony ze względu na wymóg obrotowości podzbioru. Ciekawą alternatywą zatem wydaje się dedykowana badaniom genetycznym procedura łączna zaproponowana przez Dudoit

oraz van der Laana (2008). Szeroki zakres zastosowań, możliwość wyboru miary błędu I rodzaju dla zbioru wnioskowań oraz powszechnie dostępne gotowe oprogramowanie w pakiecie *multtest* w R, to jej istotne zalety. Niestety zaprezentowane w artykule badania symulacyjne pokazały, że procedura MTP nie zawsze gwarantuje kontrolę FWER dla zbioru wnioskowań na z góry ustalonym poziomie. Z kolei problemy z kontrolą miary FDR przez procedurę MTP sygnalizowali Werft i Benner (2009). Wyniki te wskazują na konieczność dalszych badań nad procedurą MTP Dudoit oraz van der Laana.

Uniwersytet Ekonomiczny w Krakowie

LITERATURA

[1] Benjamini Y., Hochberg Y., (1995), Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society*, Ser. B, 57 (1), 289–300.
[2] Benjamini Y., Yekutieli D., (2001), The Control of the False Discovery Rate in Multiple Testing Under Dependency, *Annals of Statistics*, 29, 1165–1188.
[3] Bretz F., Hothorn T, Westfall P., (2011), *Multiple Comparisons Using R*, Chapman and Hall, Boca Raton.
[4] Denkowska S., (2005), Zastosowanie procedur testowań wielokrotnych opartych na uporządkowanych prawdopodobieństwach testowych do wydzielania jednorodnych podgrup wartości przeciętnych, *Przegląd Statystyczny*, 1, 115–131.
[5] Denkowska S., (2006), Multiple Testing in a Correlation Matrix, w: Pociecha J., (red.), *A Comparative Analysis of the Socio-Economic Consequences of Transition Processes in the Central and Eastern European Countries*, Wydawnictwo AE w Krakowie, 117–134.
[6] Denkowska S., (2007a), Testowanie wielokrotne w badaniach ekonomicznych, *Folia Oeconomica Cracoviensia*, XLV, Wydawnictwo Oddziału PAN, Kraków, 119–135.
[7] Denkowska S., (2007b), Monte Carlo Analysis of the Effectiveness of Multiple Comparison Procedures, *Education of Quantitative Mathematical-Statistical Methods*, University of Economics in Bratislava, Bratislava, 117–126.
[8] Denkowska S., (2011a), Testowanie jednoczesne przy weryfikacji ocen parametrów strukturalnych liniowego modelu ekonometrycznego, *Zeszyty Naukowe „Metody Analizy Danych"*, 873, Wydawnictwo UEK, Kraków, 53–68.
[9] Denkowska S., (2011b), Testowanie wielokrotne przy budowie modelu ekonometrycznego, *Zeszyty Naukowe „Metody Analizy Danych"*, Wydawnictwo UEK, Kraków, 27–42.
[10] Denuit M., Scaillet O., (2004), Nonparametric Tests for Positive Quadrant Dependence, *Journal of Financial Econometrics*, 2, 422–450.
[11] Domański Cz., Pruska K., (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.
[12] Domański Cz., Parys D., (2007), *Statystyczne metody wnioskowania wielokrotnego*, Wydawnictwo UŁ, Łódź.
[13] Dudoit S., van der Laan M., (2008), *Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics.
[14] Hochberg Y., Tamhane A. C., (1987), *Multiple Comparison Procedures*, John Wiley & Sons, NY.

[15] Holland B., Copenhaver M. D., (1987), An Improved Sequentially Rejective Bonferroni Test Procedure, *Biometrics*, 43, 417–423.
[16] Lehmann E. L., Romanno J. P., (2005), Generalizations of the Familywise Error Rate, *Annals of Statistics*, 33 (3), 1138–1154.
[17] Rosenthal R., Rubin D. B., (1983), Ensemble-Adjusted p Values, *Psychological Bulletin*, 94 (3), 540–541.
[18] Shaffer J. P., (1986), Modified Sequentially Rejective Multiple Test Procedures, *Journal of the American Statistical Association*, 81, 826–831.
[19] Shaffer J. P., (1995), Multiple Hypothesis Testing, *Annual Review of Psychology*, 46, 561–84.
[20] Tukey J. W., (1953), The Problem of Multiple Comparisons, w: Braun H. I., (red.), (1994) *The Collected Works of John W. Tukey*, vol. VIII: *Multiple Comparisons: 1948–1983*, New York: Chapman & Hall, 1–300.
[21] Westfall, P. H., Tobias R. D., Rom D., Wolfinger R. D., Hochberg Y., (1999), *Multiple Comparisons and Multiple Tests Using the SAS System*, SAS Institute Inc., Cary, NC.
[22] Westfall P. H., Young S. S., (1993), *Resampling Based Multiple Testing*, Wiley, New York.
[23] Werft W., Benner A., (2009), www.iscb2009.info/RSystem/Soubory/Prez%20Monday/S10.4%20Werft.pdf.
[24] Wright S. P., (1992), Adjusted P-values for Simultaneous Inference, *Biometrics*, 48, 1005–1013.

NIEKLASYCZNE PROCEDURY TESTOWAŃ WIELOKROTNYCH

Streszczenie

Zakres zastosowań klasycznych procedur testowań wielokrotnych jest ograniczony z powodu założeń modelowych, a w wielu sytuacjach badawczych rozwiązań klasycznych po prostu brak. Kontrolę efektu testowania wielokrotnego umożliwiają wówczas nieklasyczne procedury testowań wielokrotnych. Proste obliczeniowo, o szerokim zakresie zastosowań, brzegowe procedury testowań wielokrotnych nie uwzględniają jednak łącznego rozkładu statystyk testowych, przez co są bardziej konserwatywne od procedur łącznych. Zakres zastosowań procedur łącznych Westfalla i Younga (1993) jest natomiast ograniczony ze względu na wymóg *obrotowości podzbioru*. Ciekawą alternatywę stanowią dedykowane badaniom genetycznym procedury łączne, zaproponowane przez Dudoit oraz van der Laana (2008). Szeroki zakres zastosowań, możliwość wyboru miary błędu I rodzaju oraz powszechnie dostępne, oprogramowanie (procedura MTP jest zaimplementowana w pakiecie *multtest* w R), to ich istotne zalety. Niestety, badania nad procedurą MTP przeprowadzone przez Werfta i Bennera (2009) pokazały problemy z kontrolą miary FDR w przypadku bardzo dużej liczby testowanych hipotez i małej liczebności prób. Z kolei zaprezentowany w artykule eksperyment symulacyjny pokazał, że procedura MTP nie zapewnia również kontroli FWER na z góry zadanym poziomie.

Słowa kluczowe: testowanie wielokrotne, FWER, FDR, repróbkowanie, MTP

NON-CLASSICAL MULTIPLE TESTING PROCEDURES

Abstract

The range of applications of classical multiple testing procedures is limited due to model assumptions, and in many cases classic solutions are non-existent. In such situations non-classical multiple testing procedures allow to control the effect of multiple testing. Although they are popular for computational simplicity and a wide range of applications, marginal multiple testing procedures do not take into account joint distribution of test statistics, which make them more conservative than joint multiple testing procedures. The range of applications of joint procedures introduced by Westfall and Young (1993) is limited due to the subset pivotality requirement. Thus, joint multiple testing procedures suggested by Dudoit and van der Laan (2008) seem very promising. A wide range of applications, the possibility of choosing the Type I error rate and easily accessible software (MTP procedure is implemented in R *multtest* package) are their obvious advantages. Unfortunately, the results of the analysis of MPT procedure obtained by Werft and Benner (2009) revealed that it does not control FDR in case of numerous sets of hypotheses and small samples. Furthermore, the simulation experiment presented in the article showed that MTP procedure does not control FWER, either.

Keywords: multiple testing, FWER, FDR, resampling, MTP