Wydajne obliczanie skorygowanych wartości p dla wielokrotnego testowania opartego na stopniowym odrzucaniu z wykorzystaniem resamplingu.

Joseph P. Romano, Michael Wolf

Dodatek do wstępu

Ostatnimi czasy krąży dużo zainteresowania wokół raportowania wartości p skorygowanych za pomocą procedur wielokrotnego testowania opartych na stopniowym odrzucaniu i **resamplingu**[1] zaproponowanych przez Romano i Wolfa (2005a,b). Oryginalne dokumenty opisują jedynie jak przeprowadzać wielokrotne testy na ustalonym **poziomie istotności**[2]. Obliczanie skorygowanych **wartości p**[3] w optymalny sposób nie jest jednak trywialne. Z tego powodu ten artykuł ma za zadanie wypełnić tą lukę opisując algorytm do tego celu.

1. Wstęp

Romano oraz Wolf(2005a,b) zaproponowali procedury wielokrotnego testowania oparte na stopniowym odrzucaniu i resamplingu aby zapanować nad **błędami I rodzaju**[4] w rodzinie testów (FWE). Opisane procedury są zaprojektowane do przeprowadzania przy przy ustalonym poziomie istotności α. Dlatego, rezultatem zastosowania takiej procedury na zbiorze danych będzie "lista" z decyzjami w formie binarnej odnoszących się do indywidualnych badanych **hipotez zerowych[5]**. Wartość ta będzie oznaczała odrzucenie lub nieodrzucenie hipotezy zerowej przy ustalonym poziomie istotności α.

Jednak, w nurcie ostatnich artykułów naukowych widać zainteresowanie w obliczaniu skorygowanych wartości p. To znaczy, dla każdej badanej hipotezy zerowej,

---

[1] (próbkowanie) tworzenie nowych próbek na podstawie już zaobserwowanej próbki repróbkowanie, tj. wielokrotne losowanie ze zwracaniem. (podręcznik str. 106)

[2] Prawdopodobieństwo błędu I rodzaju jest prawdopodobieństwem odrzucenia prawdziwej hipotezy zerowej H 0 . Nazywane jest ono również poziomem istotności testu statystycznego. (podręcznik str. 136)

[3] P-wartość to graniczny poziom istotności, czyli najmniejszy, przy którym zaobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej. Jest to więc taki poziom istotności, przy którym zmienia się decyzja testu: zaczynając od małego poziomu istotności, kiedy to nie mamy podstaw do odrzucenia hipotezy zerowej, po przekroczeniu p-wartości zaczynamy odrzucać tę hipotezę. (podręcznik strona 169)

[4] odrzucenie sprawdzanej hipotezy, wtedy gdy jest ona prawdziwa, co nazywamy błędem I rodzaju (podręcznik str. 135)

[5] weryfikowana hipoteza, którą nazywa się też hipotezą zerową i oznacza zwykle przez H0 (podręcznik str. 135)

należy obliczyć odpowiadającą jej wartość p dostosowaną do procedur wielokrotnego testowania opartych na stopniowym odrzucaniu zaproponowanych przez Romano i Wolfa (2005a,b). Przykładych takich artykułów to Heckman et al. (2010), Hein et al. (2010), Campbell et al. (2014), Gertler et al. (2014) oraz Dobbie i Fryer (2015). Niestety, opisy w tych artykułach dotyczące obliczania skorygowanych wartości p są często niejasne lub całkowicie pominięte.

W zasadzie, dla danej pojedynczej hipotezy, skorygowaną wartość p można uzyskać metodą "prób i błędów" jako najmniejszy poziom istotności α przy którym hipoteza może zostać odrzucona przez procedurę wielokrotnego testowania opartą na stopniowym odrzucaniu. Oczywiście ten sposób obliczania skorygowanych wartości p byłby raczej kłopotliwy. W zamian za to, pożądane jest posiadanie wydajnego (lub usprawnionego) algorytmu do obliczania skorygowanych wartości p. Ten artykuł określa taki właśnie algorytm

Oczywistym jest stwierdzenie że taki algorytm do obliczania skorygowanych wartości p był już wcześniej opisywany. Za przykłady można podać Westfall and Young (1993) i rozmaite odniesienia do wcześniejszych prac określonych w sekcji 1.3 tej publikacji. Wkładem tego artykułu jest opisanie algorytmu, który jest specyficznie dostosowany do procedury wielokrotnego testowania opartej na stopniowym odrzucaniu opisanej w Romano and Wolf (2005a,b), co uczyni zrozumienie i implementację tego algorytmu łatwiejszą dla przyszłych adeptów.

2. Notacja i nieskorygowane wartości p

Teraz zostanie podana wystylizowana, wysokopoziomowa definicja badanego problemu wielokrotnego testowania. Szczegóły takie jak konstrukcja **statystyk testowych**[6] oraz wystarczające warunki dla (asymptotycznej) poprawności proponowanych procedur stopniowego odrzucania zależą od kontekstu (zobacz Romano and Wolf (2005a,b) oraz Romano et al. (2008, Sekcja3)).

Jest S problemów testowania pojedynczych hipotez:

Hs przeciwko Hs' dla s= 1, 2, ..., S

Hs- hipoteza zerowa

Hs'- **hipoteza alternatywna**[7]

---

[6] Test statystyczny wykorzystuje odpowiednio skonstruowaną statystykę
Tn - h (X x, .. .,X n), zwaną statystykę testową. Powinna ona być tak dobrana,
żeby jej wartości tn e R wyraźnie wskazywały na prawdziwość lub fałszywość
weryfikowanej hipotezy H0. (podręcznik str. 135)

[7] wyróżnia się pewną hipotezę alternatywną, oznaczaną zazwyczaj przez H1, którą przyjmuje się w przypadku odrzucenia hipotezy zerowej.

Odpowiadające im statystyki testowe są oznaczane odpowiednio t1, t2, ..., ts. Są one zaprojektowane w taki sposób, że duże wartości wskazują na alternatywę. (W szczególności, dla problemów **testowania dwustronnego**[8], statystyki testowe zazwyczaj opierają się na wartościach bezwzględnych.)

Procedury wielokrotnego testowania oparte na stopniowym odrzucaniu są w większości oparte na zbiorze resamplingowych statystyk testowych hipotezy zerowej

$t^{*,m} := (t_1^{*,m}, \ldots, t_S^{*,m})$ dla m = 1, ..., M gdzie M oznacza liczbę powtórzeń resamplingu. Zależnie od kontekstu , resampling może być wykonany **metodą bootstrapu**[9], **metodą permutacji**[10], albo metodą randomizacji. Szczegóły dla metody bootstrapu można znaleźć w Romano and Wolf (2005a, Sekcja 4.2), Romano and Wolf (2005b), oraz Romano et al. (2008, Sekcja 4.3). Szczegóły dla metod permutacji i randomizacji można znaleźć w Romano and Wolf (2005a, Sekcja 3.2). W odniesieniu do Davison and Hinkley (1997, rozdział 4), nieskorygowana (albo marginalna) wartość p dla Hs określona jako $\hat{p}_s$ definiowana jest jako:

Wzór (2.1)

$$\hat{p}_s := \frac{\#\{t_s^{*,m} \geq t_s\} + 1}{M + 1}$$

- $\#\{t_s^{*,m} \geq t_s\}$: Liczba razy, kiedy statystyka testowa resamplingu $t_s^{*,m}$ jest większa lub równa zaobserwowanej statystyce testowej t_s.
- M: Liczba powtórzeń próbkowania.
- $+1$ w liczniku i $+1$ w mianowniku: Korekta wprowadzona w celu uniknięcia problemów związanych z małymi próbkami i zapewnienia, że wartość p nigdy nie będzie dokładnie zerowa.

Ta definicja nieskorygowanych wartości p nie jest unikalna. Niektórzy wolą używać tej definicji:

---

[8] Test dwustronny to rodzaj testu statystycznego, dla którego skrajne wartości obserwowanej zmiennej znajdują się po obydwu stronach jej rozkładu. (dobrebadania.pl) (wspomniany podręcznik str. 167)

[9] Realizację
(x*',...,Jt*') próby bootstrapowej generuje się przez losowanie ze zwracaniem
spośród posiadanych elementów próby (x p ...,xn), które traktujemy jako
populację. Przez N-krotne generowanie takich realizacji uzyskujemy ciąg N
wartości statystyki Tn (podręcznik str. 125)

[10] Testy permutacyjne stanowią podejście do testowania hipotez statystycznych oparte na wykorzystaniu wielu permutacji jednej wejściowej próby losowej dla oceny nieznanego rozkładu statystyki testowej (podręcznik str. 155)

Wzór (2.2)

$$\hat{p}_s := \frac{\#\{t_s^{*,m} \geq t_s\}}{M}$$

- $\#\{t_s^{*,m} \geq t_s\}$: Liczba razy, kiedy statystyka testowa resamplingu $t_s^{*,m}$ jest większa lub równa zaobserwowanej statystyce testowej t_s.
- M: Liczba powtórzeń próbkowania.

Oczywiście gdy M jest odpowiednio duże (na przykład M=1000), różnica pomiędzy (2.1) a (2.2) nie jest istotna.

3. Procedury wielokrotnego testowania oparte na stopniowym odrzucaniu

Wygodnie będzie pierwsze opisać ogólną procedurę wielokrotnego testowania opartą na stopniowym odrzucaniu która pozwala kontrolować błąd I rodzaju dla rodziny testów (FWE) na ustalonym poziomie istotności α w stylizowanej notacji używanej w tym artykule. W ten sposób, algorytm obliczający skorygowane wartości p w następnej sekcji będzie prostszy do zrozumienia.

Hipotezy są ponownie oznaczane w porządku malejącym na podstawie zaobserwowanych statystyk testowych. Dokładniej mówiąc, oznaczmy permutację

$\{r_1, r_2, ..., r_S\}$ jako permutację zbioru {1, 2, ..., S}, która spełnia warunek: $t_{r1}>=t_{r2}>=...>=t_{rs}$. W ten sposób H_{r1} jest "najbardziej znaczącą" hipotezą oraz H_{rs} jest "najmniej znaczącą" hipotezą.

Definicja maksymalnej wartości w wektorze statystyk testowych resamplingu:

- Niech $\max_{t,j}^*$ oznacza największą wartość wektora $(t_{r_j}^*, \ldots, t_{r_S}^*)$.
- Mamy:

$$\max_{t,j}^* := \max\{t_{r_j}^*, \ldots, t_{r_S}^*\} \quad \text{dla} \quad j = 1, \ldots, S \quad \text{i} \quad m = 1, \ldots, M$$

Co więcej,

- $\hat{c}(1-\alpha, j)$ oznacza empiryczny kwantyl $1-\alpha$ dla zbioru $\{\max_{t,j}^{*,m}\}_{m=1}^{M}$.

Nie ma unikalnej definicji **kwantyla empirycznego**[11] ale jeżeli M jest odpowiednio duże , różnice są praktycznie nieistotne. Algorytm procedury wielokrotnego testowania opartej na stopniowym odrzucaniu podano poniżej:

Algorytm 3.1 (Stepdown Multiple Testing at Significance Level α)

1. Dla $s = 1, \ldots, S$, odrzuć H_{r_s} wtedy i tylko wtedy, gdy $t_{r_s} > \hat{c}(1-\alpha, 1)$.

2. Oznacz przez R_1 liczbę odrzuconych hipotez. Jeśli $R_1 = 0$, zakończ; w przeciwnym razie ustaw $j = 2$.

3. Dla $s = R_{j-1} + 1, \ldots, S$, odrzuć H_{r_s} wtedy i tylko wtedy, gdy $t_{r_s} > \hat{c}(1-\alpha, R_{j-1}+1)$.

4. (a) Jeśli żadna dalsza hipoteza nie zostanie odrzucona, zakończ.

   (b) W przeciwnym razie, oznacz przez R_j liczbę wszystkich odrzuconych hipotez do tej pory, a następnie ustaw $j := j + 1$. Następnie wróć do kroku 3.

Uwaga 3.1 (Opis alternatywny)

Łatwo zauważyć, że H_{r_s} zostanie odrzucona na poziomie α przez Algorytm 3.1 wtedy i tylko wtedy, gdy:

$$t_{r_j} > \hat{c}(1-\alpha, j) \quad \text{dla wszystkich} \quad j = 1, \ldots, s.$$

Zatem, zbiór hipotez odrzuconych na poziomie α jest dany przez kolekcję $\{H_{r_1}, \ldots, H_{r_n}\}$, gdzie n jest największą liczbą całkowitą w zbiorze $\{1, \ldots, S\}$ taką, że:

$$t_{r_j} > \hat{c}(1-\alpha, j) \quad \text{dla wszystkich} \quad j = 1, \ldots, n.$$

Jeśli taka liczba n nie istnieje, żadna hipoteza nie zostaje odrzucona.

4. Skorygowanie wartości p dla procedury

Oznaczamy skorygowaną wartość p dla hipotezy Hs jako ^p_s^{adj} . Poniższy algorytm opisuje jak wartości p mogą być wyliczone w wydajny sposób.

[11] Kwantyl empiryczny to wartość, która dzieli zbiór danych w taki sposób, że określony procent danych znajduje się poniżej tej wartości.

Algorytm 4.1 (Obliczanie wartości p skorygowanych dla procedury wielokrotnego testowania opartej na stopniowym odrzucaniu)

1. Zdefiniuj

$$\hat{p}_1^{\text{adj}} := \frac{\#\{\max_{t,1}^{*,m} \geq t_{r_1}\} + 1}{M + 1}.$$

2. Dla $s = 2, \ldots, S$,

   (a) najpierw zdefiniuj

$$\hat{p}_s^{\text{initial}} := \frac{\#\{\max_{t,s}^{*,m} \geq t_{r_s}\} + 1}{M + 1},$$

   (b) następnie wymuś monotoniczność, definiując

$$\hat{p}_s^{\text{adj}} := \max\{\hat{p}_s^{\text{initial}}, \hat{p}_{s-1}^{\text{adj}}\}.$$

Krok 2(b) w algorytmie (4.1) jest konieczny. Bez niego, skorygowane wartości p dla hipotez H_{r2}, … , H_{rs} byłyby zbyt optymistyczne (w sensie dostarczania dowodów przeciwko hipotezie zerowej). Ten fakt jest najprostszy do zaobserwowania rozpatrując H_{rs}. Bez kroku 2(b), algorytm podtrzymywałby, że $\hat{p}_{rs}^{adj} = \hat{p}_{rs}$, więc skorygowana wartość p byłaby równa nieskorygowanej wartości p.

Zauważalne jest, że skorygowane wartości p są poprawne w sensie, że do momentu aż M jest dostatecznie duże, Hs zostanie odrzucone na poziomie istotności α przez algorytm 3.1 dla wszystkich zastosowań praktycznych wtedy i tylko wtedy, gdy skorygowana wartość p dla Hs obliczona przez algorytm 4.1 spełnia $\hat{p}_s^{adj} \le \alpha$. Dodatek “dla wszystkich zastosowań praktycznych” do tego zdania znajduje się tam z powodu, że jak już wspomniano, nie istnieje unikalna definicja kwantylu empirycznego używanego w algorytmie 3.1 ani dla wartości p opartych na resamplingu i używanych w algorytmie 4.1. Natomiast jeżeli M jest dostatecznie duże (na przykład M=1000), naruszenia stwierdzenia “wtedy i tylko wtedy” nie nastąpią przed trzecim miejscem po przecinku α, co sprawia że jest to praktycznie nieistotne.