Zagadnienie testowania zbioru hipotez (wielokrotne testowanie hipotez)

Zofia Zielińska-Kolasińska, Aleksandra Zyskowska

12 listopada 2008

Testowanie hipotez

$\Lambda \rightarrow P(x)$ na X
Zgodnie z $P(x)$ generowane są wyniki eksperymentu Λ.
$P \in \{P_\theta(x) : \theta \in \Theta\}$
Badacza interesuje weryfikacja przypuszczenia, czy nieznany parametr θ należy do zbioru $\Theta_0 \in \Theta$

- **Hipotezą zerową H_0** nazywamy przypuszczenie, że $\theta \in \Theta_0$.
- **Hipotezą alternatywną H_A** nazywamy przypuszczenie, że $\theta \notin \Theta$.

Testujemy H_0 przeciwko H_A statystyką testową $T(X)$.
B - zbiór odrzuceń
Jeśli $T(X) \in B$, to odrzucamy H_0.
Jeśli $T(X) \notin B$, to przyjmujemy H_0.
Błąd I rodzaju - odrzucenie prawdziwej H_0
Błąd II rodzaju - przyjęcie fałszywej H_0

Testowanie hipotez

	przyjmujemy	odrzucamy
H_0 prawdziwa	OK	błąd I rodzaju
H_0 fałszywa	błąd II rodzaju	OK

Testowanie gwarantuje, że $P(\text{błąd I rodzaju}) \leq \alpha$, gdzie $\alpha \in (0, 1) \Rightarrow$ kontrolowany jest błąd I rodzaju na poziomie istotności α.

P-wartość

- Prawdopodobieństwo, że uzyskalibyśmy takie jak faktycznie obserwujemy, lub bardziej oddalone od zera wartości pewnej statystyki, przy założeniu że hipoteza zerowa jest prawdziwa.
- P–wartość jest równa najmniejszemu poziomowi istotności, na którym dla obserwacji $x \in X$ przyjmiemy hipotezę H_0.

Wielokrotne testowanie hipotez

- Zbiór m eksperymentów losowych
  $\Lambda = \{\Lambda(i) : i \in I = \{1, 2, ..., m\}\}$
  $X^{(i)}$-zbiór wszystkich możliwych wyników i-tego eksperymentu
- Na $X^{(i)}$ określamy rodzinę rozkładów prawdopodobieństwa $\overline{P}^{(i)} = \{P^{(i)}_{\theta^{(i)}} : \theta^{(i)} \in \Theta^{(i)}\}$
- $\Lambda^{(i)}$ odpowiada nieznany rozkład $P^{(i)}$ na $X^{(i)}$
  Zakładamy:$P^{(i)} \in \overline{P}^{(i)}$, tzn $\exists \theta^{*(i)} \in \Theta^{(i)}$ t.że $P^{(i)} = P^{(i)}_{\theta^{*(i)}}$

- Przypuszczenie, że $\theta^{*(i)} \in \Theta^{(i)}_0 \subset \Theta^{(i)}$nazywamy i-tą hipotezą zerową $(H^{(i)}_0)$
- Przypuszczenie, że $\theta^{*(i)} \notin \Theta^{(i)}_0 \subset \Theta^{(i)}$ nazywamy i-tą hipotezą alternatywną $(H^{(i)}_A)$

Współczynniki błędów

	#przyjętych hipotez zerowych	#odrzuconych hipotez zerowych	suma
#prawdziwych hipotez zerowych	$m_0 - V$	V	m_0
#fałszywych hipotez zerowych	$m_1 - S$	S	m_1
suma	$m - R$	R	m

- **Family-wise error rate** (p-stwo odrzucenia co najmniej jednej prawdziwej hipotezy): $\mathbf{FWER} = \mathbf{P}(\mathbf{V} \geq \mathbf{1})$
- **False discovery rate** (wartość oczekiwana frakcji fałszywie odrzuconych hipotez zerowych w zbiorze wszystkich odrzuconych hipotez zerowych, przemnożona przez p-stwo odrzucenia co najmniej jednej hipotezy):
  $\mathbf{FDR} = \mathbf{E}(\mathbf{Q}|\mathbf{R} > \mathbf{0})\mathbf{P}(\mathbf{R} > \mathbf{0})$
  UWAGA: $Q = \frac{V}{R}$, gdy $R > 0$ oraz $Q = 0$, gdy $R = 0$
- **Positive false discovery rate** (wartość oczekiwana frakcji fałszywie odrzuconych hipotez zerowych w zbiorze wszystkich odrzuconych hipotez zerowych, pod warunkiem, że $R > 0$): $\mathbf{pFDR} = \mathbf{E}(\frac{\mathbf{V}}{\mathbf{R}}|\mathbf{R} > \mathbf{0})$

$pFDR = \frac{FDR}{P(R>0)}$. $P(R > 0) \to 1$ gdy $m \to \infty \Rightarrow$ FDR i pFDR są as. równoważne.

Procedura testowania zbioru hipotez

Procedura testowania zbioru hipotez - reguła określająca, w jakiej kolejności oraz na jakim poziomie istotności należy testowac poszczególne hipotezy.

Równoważnie:

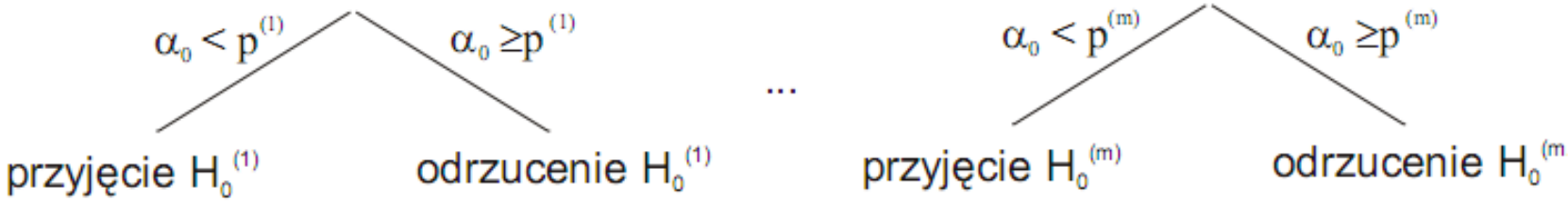
$X = B^{(i)} \dot{\cup} (B^{(i)})^C$ ze względu na każdą hipotezę $H_0^{(i)}$.

Odrzucamy $H_0^{(i)}$, jeżeli $x \in B^{(i)} \subset X$, a przyjmujemy w przeciwnym przypadku.

Procedura kontroluje w **sensie mocnym współczynnik błędu na poziomie** $\alpha > 0$, jeżeli bez względu na to ile oraz które hipotezy są prawdziwe, przyjęte obszary odrzucenia $B^{(i)}$ gwarantują, że powtarzając zbiór eksperymentów losowych dany współczynnik nie będzie średnio większy niż α.

Procedury testowania zbioru hipotez - procedury jednokrokowe

- każda z m hipotez zerowych tesowana niezależnie od wyniku testowania pozostałych
- wszystkie $H_o^{(i)}$ są testowane na tym samym poziomie istotności α_0
- odrzucane są hipotezy, dla których p-wartości $p^{(i)} \leq \alpha$

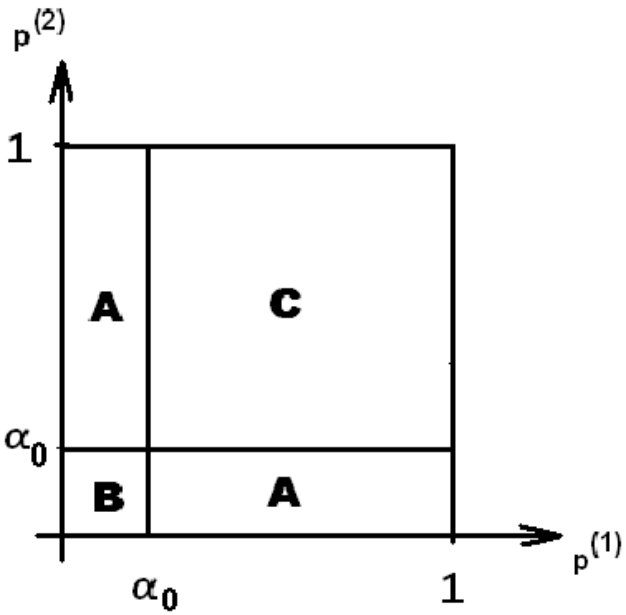


UWAGA

Procedury jednokrokowe są konserwatywne, tzn. dla dużych zbiorów hipotez prowadzą do odrzucenia bardzo małej liczby hipotez.

Procedury testowania zbioru hipotez – procedury jednokrokowe (cd.)

Obszary przyjęcia i odrzucenia dla procedury jednokrokowej dla m = 2



- *A* - obszar przyjęcia dokładnie jednej z hipotez,
- *B* - obszar odrzucenia obu hipotez,
- *C* - obszar przyjęcia obu hipotez.

Procedury testowania zbioru hipotez - procedury wielokrokowe

- Procedury wielokrokowe wykorzystują informację o łącznym rozkładzie p-wartości dla wszystkich hipotez, co prowadzi do wyższej liczby odrzuconych hipotez.
- Niech $p^{(i:m)}$ oznacza **i-tą statystykę pozycyjną** dla p-wartości, czyli $p^{(1:m)}$ oznacza najmniejszą p-wartość.

$$H_0^{(i:m)} \text{ - hipoteza zerowa odpowiadająca } p^{(i:m)}$$

$$\alpha^{(i:m)} \text{ - poziom istotności dla } H_0^{(i:m)}$$

Zakładamy, że poziomy istotności tworzą niemalejący ciąg

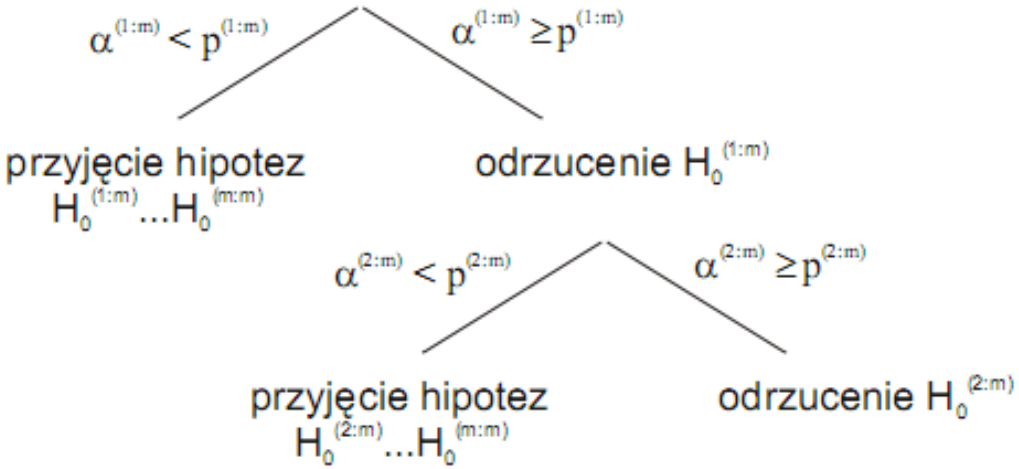
$$0 < \alpha^{(1:m)} \leq \alpha^{(2:m)} \leq \ldots \leq \alpha^{(m:m)}.$$

- Przedstwimy metody wyznaczania $\alpha^{(i:m)}$.
- W procedurze wielokrokowej kolejność testowania hipotez zależy od kolejności odpowiadających im p-wartości.

Procedury testowania zbioru hipotez – procedury wielokrokowe - procedura step-down

Procedura step-down

- testowanie rozpoczynamy od $H_0^{(1:m)}$
- $p^{(i:m)} > \alpha^{(i:m)} \Rightarrow$ przyjmujemy wszystkie hipotezy $H^{(i:m)}, \ldots, H^{(m:m)} \rightarrow$ KONIEC
- $p^{(i:m)} \leq \alpha^{(i:m)} \Rightarrow$ odrzucamy $H_0^{(i:m)}$ i postępowanie powtarzamy dla zbioru hipotez $\{H_0^{(i+1:m)}, \ldots, H_0^{(m:m)}\}$

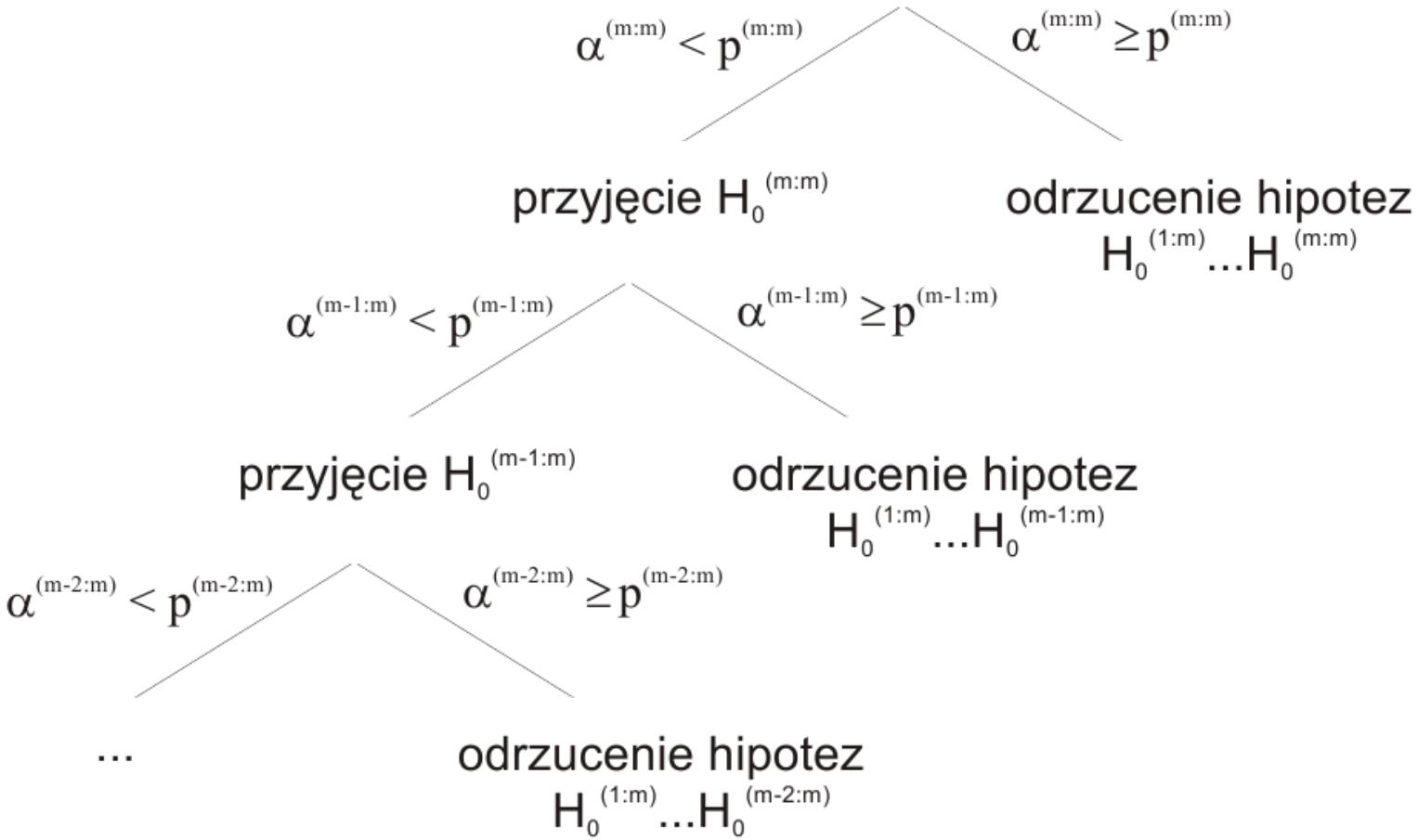


Procedury testowania zbioru hipotez – procedury wielokrokowe - procedura step-up

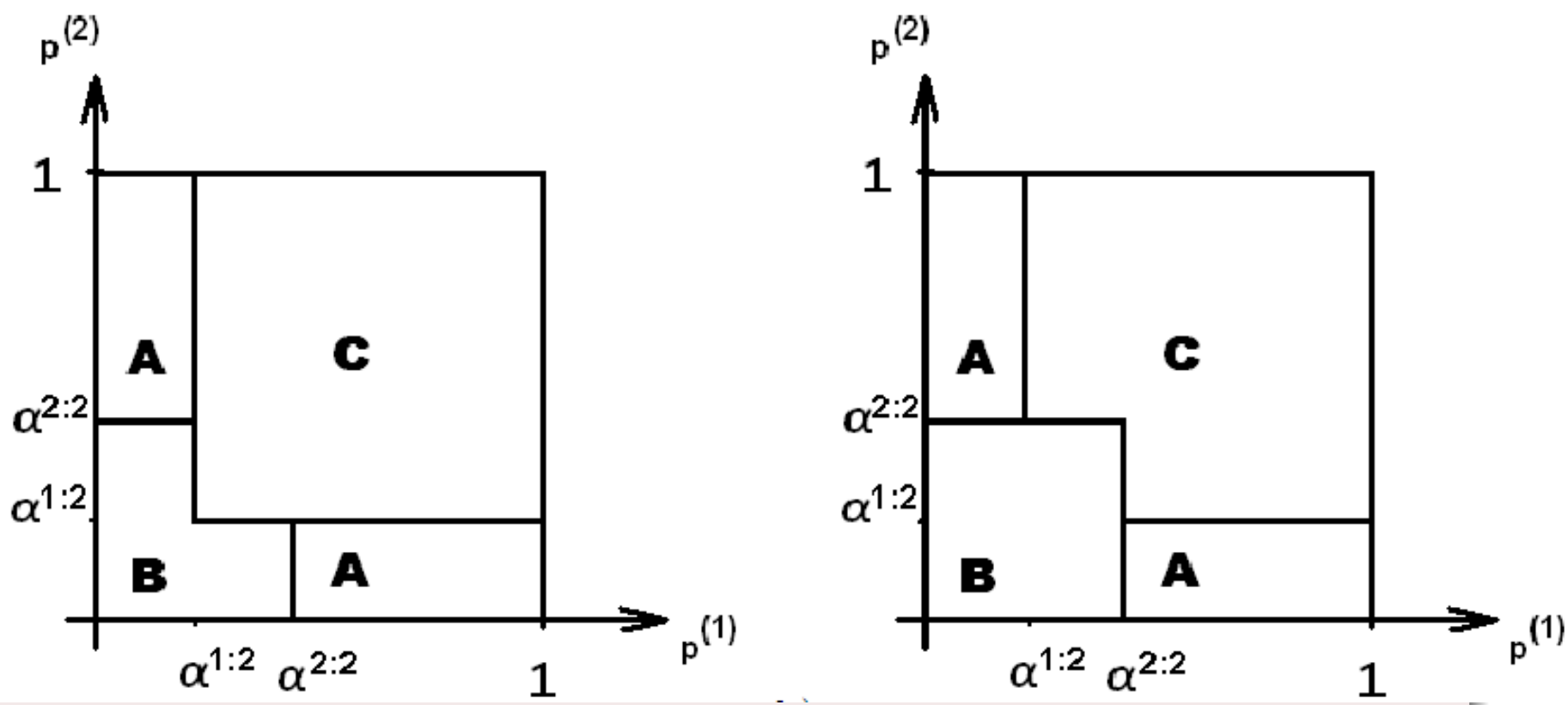
Procedura step-up

- testowanie rozpoczynamy od $H_0^{(m:m)}$
- $p^{(m:m)} \leq \alpha^{(m:m)} \Rightarrow$ odrzucamy wszystkie hipotezy $\rightarrow$ KONIEC
- $p^{(m:m)} > \alpha^{(m:m)} \Rightarrow$ przyjmujemy $H_0^{(m:m)}$ i postępowanie powtarzamy dla zbioru hipotez $\{H_0^{(1:m)}, \ldots, H_0^{(m-1:m)}\}$
- i-ty krok:
  - $p^{(m-i+1:m)} \leq \alpha^{(m-i+1:m)} \Rightarrow$ odrzucamy $\{H_0^{(1:m)}, \ldots, H_0^{(m-i+1:m)}\} \rightarrow$ KONIEC
  - $p^{(m-i+1:m)} > \alpha^{(m-i+1:m)} \Rightarrow$ przyjmujemy $H_0^{(m-i+1:m)}$ i przechodzimy do kroku $i+1$

Procedury testowania zbioru hipotez – procedury wielokrokowe - procedura step-up (cd.)



Procedury testowania zbioru hipotez - procedury wielokrokowe (cd.)



Obszary przyjęcia i odrzucenia (oznaczenia jak wcześniej) dla procedury step–down dla m = 2 (po lewej) oraz dla procedury step–up (po prawej).

Procedury kontroli współczynników błędu - FWER

Procedura Bonferroniego

Jednokrokowa procedura testowania z poziomem istotności $\alpha_0 = \frac{\alpha}{m}$, kontroluje współczynnik FWER na poziomie α.

UWAGA:

- Także dla skorelowanych statystyk testowych.
- Zapewnia mocną kontrolę FWER.
- Bardzo konserwatywna, trudno odrzucić jakiekolwiek hipotezy.

Dopasowanie p-wartości

$\widetilde{p}_i = \min(m \cdot p_i, 1)$

Procedury kontroli współczynników błędu - FWER (cd.)

Procedura Holma

Jeżeli statystyki testowe testowanych hipotez są *niezależne*, to procedura **wielokrokowa step-down** z poziomem istotności

$$\alpha^{(i:m)} = \frac{\alpha}{m - i + 1},$$

kontroluje współczynnik FWER na poziomie α.

Dopasowanie p-wartości

$$\widetilde{p}^{(i:m)} = \max_{k=1,\ldots,i} \min\left((m - k + 1) \cdot p^{(i:m)}, 1)\right)$$

Procedury kontroli współczynników błędu - FWER (cd.)

Procedura Hochberga

Jeżeli statystyki testowe testowanych hipotez są *niezależne*, to procedura **wielokrokowa step-up** z poziomem istotności

$$\alpha^{(i:m)} = \frac{\alpha}{m - i + 1},$$

kontroluje współczynnik FWER na poziomie α.
UWAGA: Procedura Hochberga ma większe obszary odrzuceń niż procedura Holma, przez co odrzuca średnio więcej fałszywych hipotez zerowych.

Dopasowanie p-wartości

$$\widetilde{p}^{(i:m)} = \min_{k=i,\ldots,m} \min\left((m - k + 1) \cdot p^{(i:m)}, 1\right)$$

Procedury kontroli współczynników błędu - FDR

Procedura Benjaminiego-Hochberga

Jeżeli statystyki testowe testowanych hipotez są *pozytywnie zależne* (macierz korelacji statystyk testowych jest dodatnio określona), to procedura **wielokrokowa step-up** z poziomem istotności

$$\alpha^{(i:m)} = \frac{i}{m}\alpha,$$

kontroluje współczynnik FDR na poziomie α.

Dopasowanie p-wartości

$$\widetilde{p}^{(i:m)} = \min_{k=i,\ldots,m}\left(\min(\tfrac{m}{k}\cdot p^{(i:m)}, 1)\right)$$

Kontrola FWER i FDR

Kontrolowanie FDR-kontrolowanie oczekiwanej wartości stosunku błędnie odrzuconych hipotez (do wszystkich odrzuconych hipotez).

- $FDR = FWER \Leftrightarrow$ wszystkie hipotezy są prawdziwe(poza tym $FDR \leq FWER$).
- Kontrola FDR $\Rightarrow$ słaba kontrola FWER.
- Kontrola FWER $\Rightarrow$ kontrola FDR.
- Jeśli $m = m_0$ (nawet jeśli pojedyncza hipoteza jest odrzucona), to $\frac{v}{r} = 1 \Rightarrow Q \neg$ kontrolowane $\Rightarrow (\frac{V}{R}|R > 0) \neg$ kontrolowane $\Rightarrow E(\frac{V}{R}|R > 0) \neg$ kontrolowana. Ale $FDR = P(R > 0)E(\frac{V}{R}|R > 0)$ da się kontrolować.
- Proporcja błędnych do wszystkich: $Q^{'} = \frac{E(V)}{r}$-złożenie wartości oczekiwanej i realizacji $\Rightarrow$ nawet nie $E(Q|R = r) = \frac{E(V|R=r)}{r}$,gdzie znów problem z kontrolą j.w.
- Gdy wszystkie hipotezy prawdziwe $\frac{E(V)}{E(R)} = 1 \Rightarrow$ problem z kontrolą. Środki: dodać jeden do mianownika (zniekształcony wynik), mianownik zamienić na $E(R|R > 0)$. Zmiana licznika i mianownika równocześnie-problem kontroli przy $m = m_0$.

Kontrola FDR i pFDR

- Gdy $m = m_0$, $E[\frac{V}{R}|R > 0] = 1$(niekontrolowane) oraz mnożnik $P(R > 0)$ w definicji FDR $\Rightarrow$ wprowadzenie definicji $pFDR = E[\frac{V}{R}|R > 0]$.
- Dla $m = m_0$ mamy $pFDR = 1 \Rightarrow$ do kontroli pFDR musimy go estymować dla konkretnego zbioru odrzuceń.
- Dwa sposoby kontroli pFDR:
  - Ustalamy poziom istotności (α) i estymujemy zbiór odrzuceń-ma sens dla FWER (mierzy p-ństwo popełnienia więcej niż jednego błędu I rzędu $\Rightarrow$ potrafimy powiedzieć a priori jakie powinno być α). FDR-y są bardziej "badawcze". pFDR nie jest kontrolowalny w tym sensie.
  - Ustalamy zbiór odrzuceń, estymujemy α (dobry dla pFDR).

Kontrola FDR i pFDR (cd.)

Przykład

100 hipotez. Chcemy kontrolować FDR na poziomie $\alpha = 5\%$. Czy był to dobry wybór? Zależy od liczby odrzuconych hipotez (100 odrzuconych$\Rightarrow$dobry wybór, 2 odrzucone$\Rightarrow$mniej przydatny wybór).

Wniosek

Dla FDR oraz pFDR ustalanie najpierw zbioru odrzuceń jest lepsze.

- Ustalanie zbioru odrzuceń-koncepcyjnie prostsze podejście dla skomplikowanych złożonych miar błędów jak pFDR, FDR.
- Estymacja punktowa przekłada się na kontrolowanie FDR i pFDR jednocześnie.
- pFDR -proces stochastyczny na wszystkich przedziałach odrzuceń.
- pFDR-najbardziej odpowiednia miara błędu łącząca podejście częstościowe oraz bayesowskie.

Interpretacja bayesowska pFDR

- m identycznych testów H_0 vs. H_A, statystyki testowe: $T_1, \ldots, T_m$
- Γ - ustalony "zbiór istotności" (significance region)
- $V(\Gamma) = \#\{T_i \text{ dla prawdziwej } H_0^{(i)} : T_i \in \Gamma\}$, $R(\Gamma) = \#\{T_i : T_i \in \Gamma\}$
- $H_i := 0$, gdy i-ta hipoteza zerowa prawdziwa; $H_i := 1$, gdy fałszywa.
- π_0 - rozkład a priori p-stwa, że hipoteza zerowa jest prawdziwa (zakładamy, że H_i są i.i.d. o rozkładzie Bernoulliego: $P(H_i = 0) = \pi_0$, $P(H_i = 1) = 1 - \pi_0 = \pi_1$).

$$pFDR(\Gamma) = E\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right]$$

Twierdzenie – "błąd I rodzaju a posteriori"

m identycznych testów opartych na statystykach $T_1, \ldots, T_m$ ze zbiorem istotności Γ. Załóżmy, że (T_i, H_i) są i.i.d. takie, że $T_i|H_i$ ma rozkład: $(1 - H_i) \cdot F_0 + H_i \cdot F_1$ dla pewnego rozkładu F_0 i rozkładu alternatywnego F_1, oraz dla H_i z rozkładu Bern(π_1). Wówczas rozkład a posteriori dany jest wzorem ($\pi_0 = 1 - \pi_1$ - rozkład a priori):

$$pFDR(\Gamma) = P(H = 0|T \in \Gamma) = \frac{\pi_0 P(T \in \Gamma|H = 0)}{P(T \in \Gamma)}.$$

Tło historyczne

- Termin **hipoteza** - "Teoria Matematyki"greckiego filozofa Geminusa (pierwsze dziesięciolecia naszej ery)
- Hipoteza zweryfkowana na gruncie analizy statystycznej - John Arbuthnot w pracy „An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes" w 1710.
  Roczne liczby urodzeń chłopców oraz dziewcząt (Londyn 1625 – 1710) $\rightarrow$ więcej chłopców niż dziewcząt.
  Częstość urodzin chłopców $= \frac{1}{2} \Rightarrow$ **P**(co rok, przez 86 lat, rodziło się więcej chłopców niż dziewcząt)=$\frac{1}{286} < 10^{24}$, czyli jest niezmiernie małe. Czyli częstość urodzin chłopców jest statystycznie istotnie większa niż częstość urodzin dziewcząt.
- Skrytykowane (Nicholas Bernoulli). Konstrukcja pierwszego statystycznego testu - Pierri-Simon Laplace (1749 – 1827). W 1796 fizykalno-matematyczne uzasadnienie „nebular hypothesis" (**hipotezy mgławicowej**-hipotezy Kanta-Laplacea), opisującej genezę powstania Układu Słonecznego. Laplace - zwolennik subiektywnej interpretacji prawdopodobieństwa (wnioskowanie bliskie bayesowskiemu).

Tło historyczne (cd.)

- Jerzy Spława-Neyman i Egon Pearson - latach dwudzieste XX wieku - aksjomatyczne podstawy dla zagadnienia testowania. 1928–1933 - prace o procesie testowania hipotez, testach statystycznych, testach najefektywniejszych, rozmiarach testu, poziomach istotności itp. Wprowadzili pojęcie **hipotezy alternatywnej** (dziś oczywiste i niekwestionowane). Neymanowsko-Pearsonowską teorią testowania hipotez (aksjomatyczna, częstościowa). Takie ujęcie procesu testowania hipotez dominowało od lat 30-40 do końca XX wieku.
- Ważny **wybór poziomu istotności** (dopuszczalnego p-stwa odrzucenia prawdziwej hipotezy zerowej). Najczęściej przyjmowany to 0.05, co oznacza, że średnio błędnie odrzucimy hipotezę zerową nie częściej niż raz na 20 razy.

Przykład

Rozważmy zbiór 100 hipotez zerowych, każda orzekająca, że pewien lek nie wpływa na stan pacjenta. Wykonajmy 100 testów, każdy na poziomie istotności $\alpha = 0.05$. Nawet, jeżeli żaden z rozważanych leków nie wpływa na zdrowie pacjenta, to w około 5 przypadkach test odrzuci błędnie hipotezę zerową, a przyjmie hipotezę alternatywną. Czy to oznacza, że te 5 leków istotnie wpływa na stan pacjenta? Oczywiście, że nie.

UWAGA

Wzrost liczby przeprowadzonych testów na ustalonym poziomie istotności$\Rightarrow$ wzrost p-ństwa błędnego odrzucenia przynajmniej jednej hipotezy zerowej.

- Biostatystyk Graham Martin oskarżał badaczy stosujących metody statystyczne o powtarzanie eksperymentu wielokrotnie lub też wykorzystywanie różnych testów, tak długo, aż któryś test odrzuci hipotezę zerową i „potwierdzi” słuszność ich przypuszczeń (siatką Munchausena).

$\Rightarrow$ Przy publikacji wyników wymóg stosowania korekty poziomu istotności uwzględniającej liczbę weryfikowanych hipotez. Historycznie pierwszą i wciąż najpopularniejszą korektą jest **korekta Bonferroniego** [Bonferroni 1936]. Korekta ta była z powodzeniem wykorzystywana w przypadku, gdy testowano kilka lub kilkadziesiąt hipotez.

- Od lat 80. XX wieku w eksperymentach wykorzystywane są techniki wysokoprzepustowe, dostarczające badaczom olbrzymią ilość danych. Możliwe jest testowanie wielu tysięcy (a często setek tysięcy) hipotez. Dla tak dużej liczby hipotez korekta Bonferroniego $\Rightarrow$ znaczącego zmniejszenia poziomu istotności w pojedynczym teście$\Rightarrow$dramatyczny spadek mocy.
  Nowe procedury testowania - Multiple Hypothesis Testing („**testowanie wielokrotne**"). Prace: Terry Speed, Sandrine Dudoit, Yoseph Hochberg, John D. Storey i Brad Efron. Ujęcia testowania zbioru hipotez: klasyczne, bayesowskie (różne w interpretacji wyników i możliwości stosowania różnych metod statystycznych).

mt.rawp2adjp()

Opis działania

Na wejściu funkcja dostaje wektor p-wartości. Zwraca macierz przeskalowanych (do procedur wielokrotnego testowania) p-wartości:

- Bonferroni, Holm(1979), Hochberg(1988), Sidak - silna kontrola FWER
- Benjamini & Hochberg (1995), Benjamini & Yekutieli (2001) - silna kontrola FDR.

mt.rawp2adjp() (cd.)

mt.rawp2adjp(rawp, proc=c("Bonferroni", "Holm", "Hochberg", " SidakSS", " SidakSD", "BH", "BY"))

- rawp - wejściowy wektor p-wartości (współrzędne odpowiadające hipotezom)
- proc - wektor zawierający nazwy procedur wielokrotnego testowania ("Bonferroni", "Holm", "Hochberg", " SidakSS", " SidakSD", "BH", "BY")

Wartości funkcji

- adjp -macierz przeskalowane p-wartości, wiersze odpowiadają hipotezom, a kolumny procedurom testowania. Hipotezy są posortowane rosnąco, według początkowych p-wartości.
- index - wektor zawierający permutacje liczb od 1 do length(rawp)[z nowego porządku do starego], wskazujący w jaki sposób została zmieniona kolejność p - vartości w stosunku do kolejności początkowej. Aby otrzymać przeskalowane p-wartości w początkowym porządku stosujemy: adjp[order(index),].

mt.rawp2adjp() - przykład

```
> p = runif(10)
> p
 [1] 0.69931150 0.74574627 0.01929059 0.64262921 0.07901572
 [6] 0.99025059 0.37859570 0.85165144 0.25436628 0.95168929
> mt.rawp2adjp(p, proc=c("Bonferroni", "Holm", "Hochberg","BH"))
$adjp
             rawp Bonferroni      Holm  Hochberg        BH
 [1,] 0.01929059  0.1929059 0.1929059 0.1929059 0.1929059
 [2,] 0.07901572  0.7901572 0.7111415 0.7111415 0.3950786
 [3,] 0.25436628  1.0000000 1.0000000 0.9902506 0.8478876
 [4,] 0.37859570  1.0000000 1.0000000 0.9902506 0.9464893
 [5,] 0.64262921  1.0000000 1.0000000 0.9902506 0.9902506
 [6,] 0.69931150  1.0000000 1.0000000 0.9902506 0.9902506
 [7,] 0.74574627  1.0000000 1.0000000 0.9902506 0.9902506
 [8,] 0.85165144  1.0000000 1.0000000 0.9902506 0.9902506
 [9,] 0.95168929  1.0000000 1.0000000 0.9902506 0.9902506
[10,] 0.99025059  1.0000000 1.0000000 0.9902506 0.9902506

$index
 [1]  3  5  9  7  4  1  2  8 10  6
```

mt.rawp2adjp() - przykład c.d.