Kontrolle der FDX

Jens Stange

16. Januar 2011

Gliederung

Einstieg

Gegeben sei ein multiples Testproblem $(H_i, \varphi_i)_{i=1,\ldots,m}$ mit folgendem Schema:

Hypothese \ Test	Nicht-Ablehnungen	Ablehnungen
wahr	$m_0 - V$	V
falsch	$m - m_0 - (R - V)$	$R - V$

wobei:

m_0	(unbekannte) Anzahl wahrer Hypothesen
V	(zufällige, unbekannte) Anzahl fälschlich verworfener Hypothesen
R	(zufällige) Anzahl verworfener Hypothesen

Weiterhin bezeichne $\mathbb{P}$ das (unbekannte) zugrundeliegende Wahrscheinlichkeitsmaß.

Einstieg

Mithilfe von Verteilungsannahmen (parametrische, nichtparametrische Ansätze $\Rightarrow$ Statistiken, p-Werte) erhalten wir die Möglichkeit zur Kontrolle verschiedenster Fehlentscheidungswahrscheinlichkeiten.

Beispielsweise:

- Family-Wise Error Rate, $FWER_{\mathbb{P}} = \mathbb{P}[V > 0]$
  Der Fehler mindestens eine Hypothese fälschlich abzulehnen (allgemeiner auch für $k > 0$ k-$FWER_{\mathbb{P}} = \mathbb{P}[V > k]$)
- False Discovery Rate, $FDR := FDR_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[FDP]$
  Der Erwartungswert der False Discovery Proportion FDP:

$$FDP = \begin{cases} \frac{V}{R}, & \text{falls } R > 0 \\ 0, & \text{sonst} \end{cases}$$

- False Discovery Exceedance Rate $FDX := FDX_{\mathbb{P}}(c) = \mathbb{P}[FDP > c]$
  Die Wahrscheinlichkeit, dass die FDP einen gewissen Schwellwert $c \in (0, 1)$ überschreitet.

Ein technisches Hilfsmittel:

Die Zufallsgröße FDP kann auch als Funktion von Indexmengen $C = C(p_1, \ldots, p_m) \subseteq \{1, \ldots, m\}$ aufgefasst werden. Eine Möglichkeit die FDX zu kontrollieren ist die Konstruktion einer sogenannten $(1-\alpha)$-confindence-envelope $\overline{FDP} = \overline{FDP}(C) \in (0, 1]$, so dass zu vorgegebenem Signifikanzniveau α

$$\mathbb{P}\left[\overline{FDP}(C) \geq FDP(C) \forall C\right] \geq 1 - \alpha$$

$\overline{FDP}(C)$ ist in Abhängigkeit der p-Werte als eine zufällige Konfidenzfunktion zu verstehen.

mit folgendem Hintergrund:

Bezeichne mit $\mathcal{R} \subseteq \{1, \dots, m\}$ die Indizes von abgelehnten Hypothesen. Wenn man „Ablehnbereich“ $\mathcal{R}$ bestimmt, so dass $\overline{FDP}(\mathcal{R}) \leq c$ so folgt:

$$\mathbb{P}\left[FDP(\mathcal{R}) > c\right] \leq \mathbb{P}\left[\overline{FDP}(\mathcal{R}) < FDP(\mathcal{R})\right]$$

$$= 1 - \mathbb{P}\left[\overline{FDP}(C) \geq FDP(C) \,\forall\, C\right] \leq \alpha$$

Und damit Kontrolle der *FDX* mit Schwellwert c zum Niveau α

Bei einseitiger Betrachtung:

Falls der Ablehnbereich eines Tests die Form

$$\mathcal{R} = \{j \in \{1, \dots, m\} | p_j \leq T\}$$

hat, so lässt sich auch eine Schwellwertfunktion

$$\overline{fdp}(t) = \overline{FDP}\left(\{j | p_j \leq t\}\right)$$

definieren, um so einen Schwellwert $T^* = \sup_t\{\overline{fdp}(t) \leq c\}$ zur FDX-Kontrolle zu berechnen.

Bemerkung:
T ist als Funktion in den p-Werten, auch als Zufallsvariable zu interpretieren.

So gehts

1. Führe multiplen Test $\tilde{\varphi} = (\tilde{\varphi}_i)_{i=1,\dots,m}$ durch, der die *FWER* zum Niveau α kontrolliert.
   Seien $(\tilde{p}_{[i]})_{i=1,\dots,m}$ die geordneten (adjustierten) p-Werte, so dass ein Ablehnbereich $\mathcal{R}_0 = \{\tilde{p}_{[1]}, \dots, \tilde{p}_{[R_0]}\}$ vorliegt.
2. Mit $R_0 = |\mathcal{R}_0|$ der Zahl abgelehnten Hypothesen, bestimme zu gegebenem $c \in (0, 1)$

$$k^* = \max \left\{ k \in \{1, \dots, m - R_0\} \middle| \text{ so dass } \frac{k}{R_0 + k} \leq c \right\}$$

sowie eine Indexmenge

$$K = \{j \in \{1, \dots, m\} | \tilde{p}_j = \tilde{p}_{[i]} \text{ für } R_0 < i \leq, R_0 + k^*\}$$

dann augmentiere: $\mathcal{R}^+ = \mathcal{R}_0 \cup K$

..., denn

Bezeichne V die Zahl der von $\tilde{\varphi}$ fälschlich abgelehnten Hypothesen und $V^+ \leq V + k^*$ die Zahl der insgesamt mehr fälschlich abgelehnten Hypothesen, sowie $R^+ = R_0 + k^*$ die Zahl der insgesamt verworfenen Hypothesen.
Sei $c^* = \frac{k^*}{R+k^*}$

Dann gilt:

$$\mathbb{P}\left[V > 0\right] \leq \alpha \Longrightarrow \mathbb{P}\left[\frac{V^+}{R^+} > c^*\right] \leq \alpha$$

Also der Test $\varphi = (\varphi_j = \mathbb{I}_{\{j \in \mathcal{R}^+\}})_{j=1,\dots,m}$ kontrolliert die *FDX*.

Bemerkung:

Zu dieser Prozedur lässt sich folgende $(1-\alpha)$-confidence-envelope angeben:

Sei $\mathcal{R}_0 = \{j | p_j \leq q\}$ für einen kritischen Wert q, so dass die *FWER* durch α kontrolliert ist. So ist

$$\overline{FDP}(C) = \begin{cases} \frac{|C \setminus \mathcal{R}_0|}{|C|}, & \text{falls } C \neq \emptyset \\ 0, & \text{sonst} \end{cases}$$

Eine Step-Down-Prozedur

Gegeben multiples Testproblem $(H_i, \varphi_i)_{i=1,\dots,m}$
Seien $(p_{[j]})_{j=1,\dots,m}$ die geordneten marginalen p-Werte.
Folgende Step-Down-Prozedur φ_{SD} kontrolliert die *FDX* bei c zum Niveau α:

Beginne mit j=1:

1. Überprüfe:

$$p_{[j]} \leq \alpha_j := \frac{(\lceil cj \rceil + 1)\alpha}{m + \lceil cj \rceil + 1 - j}$$

2. JA Lehne die Hypothese zu $p_{[j]}$ ab, gehe zurück zu Schritt 1 mit $j = j + 1$

   NEIN Lehne die entsprechenden Hypothesen zu den p-Werten $\{p_{[j]}, \dots, p_{[m]}\}$ nicht ab und beende die Prozedur.

Die Inversionsmethode:

1. Zu jeder Teilmenge $W \subseteq \{1, \dots, m\}$ führe Test φ_W zum Niveau α auf die Hypothese $\{\mathcal{P}_W \sim UNI(0,1)\}$ durch.
   (Abkürzung: $\mathcal{P}_U \sim UNI(0,1)$ für $(p_i)_{i \in U} \overset{\text{i.i.d.}}{\sim} UNI(0,1)$)
2. Bestimme die Indexmengen
$$\mathcal{U} = \{U \subseteq \{1, \dots, m\} | \text{ Test: } \varphi_U = 0\}$$
3. Definiere:
$$\overline{FDP}(C) = \begin{cases} \max_{\{U \in \mathcal{U}\}} \frac{|U \cap C|}{|C|}, & \text{falls } C \neq \emptyset \\ 0, & \text{sonst} \end{cases}$$
4. Finde Ablehnbereich $\mathcal{R}$, so dass $\overline{FDP}(\mathcal{R}) \leq c$

$\Longrightarrow$ Prozedur zur Kontrolle der *FDX* mit Schwellwert c zum Niveau α

etwas genauer:

Es sei vorausgesetzt, dass die p-Werte von wahren Hypothesen uniform auf $(0,1)$ verteilt sind.
Bezeichne $V \subseteq C$ die Indizes wahrer Hypothesen in einer Teilmenge $C \subseteq \{1,\ldots,m\}$. Es gilt:

$$\mathbb{P}\left[\varphi_V = 1\right] \leq \alpha, \text{ also } \mathbb{P}[V \in \mathcal{U}] \geq 1-\alpha$$

$$\mathbb{P}\left[FDP(C) = \frac{|V|}{|C|} \leq \max_{\{U \in \mathcal{U}\}} \frac{|U \cap C|}{|C|} = \overline{FDP}(C) \,\forall\, C\right] \geq 1-\alpha$$

Also $\overline{FDP}$ ist ein $1-\alpha$-confidence-envelope.
Aber wie genau soll man nun alle Teilmengen $W \subseteq \{1,\ldots,m\}$ testen und einen entsprechenden Ablehnbereich finden?

Ein Vorschlag:

Zum Testen kann man den $p_{[k]}$-Test verwenden:
Zu $W = \{w(1), \dots, w(r)\} \subseteq \{1, \dots, m\}$ teste $\mathcal{P}_W \sim UNI(0,1)$ mit:

$$\varphi_W(p_{w(1)}, \dots, p_{w(r)}) = \begin{cases} 0, & \text{falls } p_{w([k])} \geq q_{B(k,r-k+1)}(\alpha) \text{ oder } r < k \\ 1, & \text{sonst} \end{cases}$$

$q_{B(a,b)}(\alpha)$ ist α-Quantil der Beta-Verteilung.

$$\{U_1, \dots, U_r\} \sim UNI(0,1) \Rightarrow U_{[k]} \sim B(k, r-k+1)$$

$$\text{bzw. } \mathbb{P}\left[U_{[k]} \leq q_{B(k,r-k+1)}(\alpha)\right] = \alpha$$

Falls also der k-kleinste p-Wert dieses Quantil unterschreitet, kann eben $\mathbb{P}[\mathcal{P}_W \sim UNI(0,1)] < \alpha$ geschlossen werden.
Mit geordneten p-Werten $(p_{[i]})_{i=1,\dots,m}$ ist folgende Vorgehensweise auf Grundlage dieser $p_{[k]}$-Tests vorgeschlagen:

1. Bestimme $J(k) = \min_{j=1,\dots,m}\left\{p_{[j]} \geq q_{B(k,m-j+1)}(\alpha)\right\}$
2. Definiere: $\overline{FDP}(C) = \frac{\left|\left\{i_k,\dots,i_{J(k)-1}\right\}^C \cap C\right|}{|C|}$
   wobei $\left\{i_k,\dots,i_{J(k)-1}\right\}$ die Indizes der p-Werte $\left\{p_{[k]},\dots,p_{[J(k)-1]}\right\}$
3. Die Schwellwertfunktion $\overline{fdp}(t)$ ist hier:
   $$\overline{fdp}(t) = \begin{cases} 1, & \text{falls } t \leq p_{[k-1]} \\ \frac{k-1}{m\hat{F}(t)}, & \text{falls } p_{[k-1]} < t \leq p_{[J(k)]} \\ \frac{m\hat{F}(t)-(J(k)-k)}{m\hat{F}(t)}, & \text{sonst} \end{cases}$$
   ($\hat{F}$ ist empirische Verteilungsfunktion der p-Werte)
4. zu $c \in (0,1)$ bestimme $T = \sup_t\{\overline{fdp}(t) \leq c\}$, und den Bereich
   $$\mathcal{R} = \{j \in \{1,\dots,m\} | p_j \leq T\}$$

$\Longrightarrow$ Der Test $\varphi = (\varphi_j = \mathbb{I}_{\{j\in\mathcal{R}\}})_{j=1,\dots,m}$ kontrolliert die *FDX*

Beweisansatz:

Die p-Werte $(p_j)_{j=1,\dots,m}$ liegen o.B.d.A. geordnet vor.
Bezeichne $C_t := \{j | p_j \leq t\} = \{p_1, \dots, p_{j(t)}\}$.

Behauptung

$\forall t \in [0,1], \forall U \in \mathcal{U}$:

$$\frac{|U \cap C_t|}{|C_t|} \leq \frac{|U^* \cap C_t|}{|C_t|}$$

für $U^* = \{k, \dots, J(k) - 1\}^C$

Das heißt, für Ablehnbereiche der Form C_t ist $\overline{FDP}$ eine $1 - \alpha$-Konfidenzschranke für die *FDP*.

Zahlenbeispiel:

Folgende Situation: $m = 7, k = 3$ und $J(k) = 6$

p_1	p_2	p_3	p_4	p_5	p_6	p_7
0	0	1	1	1	0	0

bzw. $\mathcal{P}_{U^*} = \{p_1, p_2, p_6, p_7\}$

und daraus abzulesen:

- Der 3-kleinste in p-Wert $\mathcal{P}_{U^*}$ ist:

$$p_6 \geq q_{3,7-6+1} = q_{3,4-3+1} \Rightarrow U^* \in \mathcal{U}$$

- $p_3 \leq p_4 \leq p_5 < q_{3,7-5+1} = q_{3,5-3+1}$, damit werden alle 5-elementigen Teilmengen $p_{[3]}$-Test abgelehnt.
- $p_3 \leq p_4 < q_{3,7-4+1} = q_{3,6-3+1}$, damit werden alle 6-elementigen Teilmengen abgelehnt.
- $p_3 < q_{3,7-3+1}$, also $\{p_1, \ldots, p_7\}$ wird vom $p_{[3]}$-Test ebenfalls abgelehnt.

FDX-Kontrolle mit Resampling

Gegeben: Daten $X = (X_1, ..., X_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$
Betrachte Teststatistiken $(T_j(X))_{j=1,\ldots,m}\ \mathbb{Q}$
für mulitplen Test $\varphi = \left(\varphi_j = \mathbb{I}_{\{T_j(X)>K_j\}}\right)_{j=1,\ldots,m}$.

$$\text{Sei } K^* = inf\left\{K \geq 0 \middle| \mathbb{P}\left[\frac{\sum_{i\in I_0} \mathbb{I}_{\{T_i(X)>K\}}}{\sum_{i=1}^{m} \mathbb{I}_{\{T_i(X)>K\}}} > c\right] \leq \alpha\right\}$$

Mit Bootstrap-samples $(X_b^{\#})_{b=1,\ldots,B}$ lassen sich

- die Nullverteilung $\mathbb{Q}_0$,
- die Verteilung $\mathbb{Q}$,
- eine Indexmenge $\hat{\mathcal{I}}_0 \subseteq \{1, \ldots, m\}$ von wahren Hypothesen

schätzen.

Schätzung von $\hat{\mathcal{I}}_0$

Betrachte $(H_i) \stackrel{\text{i.i.d.}}{\sim}$ Bernoulli(p_0) mit $p_0 = \frac{m_0}{m}$
Sowie $T_j \sim p_0 f_0 + (1 - p_0) f_1 = f$,
wobei f_0 Dichte von T_j unter Q_0,
entsprechend sei f_1 Dichte gegeben $H_j = 1$
Dann ergibt sich die a posteriori-Wahrscheinlichkeit (Bayes-Theorem):

$$\Rightarrow \mathbb{P}\left[H_j = 0 | T_j = t\right] = p_0 \frac{f_0(t)}{f(t)}$$

Dann mit Schätzungen $\hat{p}_0, \hat{f}_0, \hat{f}$ für p_0, f_0, f

$$\text{Sei } \hat{\mathcal{I}}_0 = \{i | Y_i = 0\}$$

für Zufallsvariablen

$$(Y_j)_{j=1,\ldots,m} \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\min(1, \hat{p}_0 \frac{\hat{f}_0(T_j)}{\hat{f}(T_j)}))$$

Für $b = 1, \ldots B$ definiere:

$$r_b(K) = \frac{\sum_{i \in \hat{\mathcal{I}}_0} \mathbb{I}_{\{T^{\#}_{i,b} > K\}}}{\sum_{j=1}^{m} \mathbb{I}_{\{T^{\#}_{j,b} > K\}}} \quad K \geq 0$$

$\Longrightarrow$ Schätzung für K^*:

$$\hat{K} = \inf\left\{ K \geq 0 \,\middle|\, \frac{1}{B} \sum_{b=1}^{B} r_b(K) \leq \alpha \right\}$$

Die daraus resultierende (common-cut-off)-Testprozedur:
$\left(\varphi_j = \mathbb{I}_{\{T_j(X) > \hat{K}\}}\right)_{j=1,\ldots,m}$ ist *FDX*-kontrollierend.

Verteilung der Zufallsvariablen *FDP*

Die 4 vorgestellten Methoden benutzen obere Schranken bzw. eine Schätzung der *FDP* um damit *FDX*-Kontrolle zu bekommen.
Ein anderer Ansatz ist es, sich die Verteilung der

$$FDP = FDP(\varphi_1(X), \ldots, \varphi_m(X)) \text{ mit } X = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$$

explizit auszurechnen, finit oder asymptotisch (d.h. für $m \to \infty$)
Damit lassen sich dann Momente (*FDR*) sowie Überschreitungswahrscheinlichkeiten (*FDX*) berechnen.

hier nur der asympotische Ansatz:

Folgende Annahmen:

Seien $(p_j)_{j=1,\ldots,m} \overset{\text{i.i.d}}{\sim} F = \pi_0 F_0 + (1-\pi_0) F_1$ mit

- $(p_i)_{i \in I_0} \overset{\text{i.i.d.}}{\sim} F_0 \sim UNI(0,1)$
- $(p_j)_{j \in \{1,\ldots,m\} \setminus I_0} \overset{\text{i.i.d.}}{\sim} F_1$,$F_1$ sei konkave C^1 Verteilungsfunktion.
- $\pi_0 = \lim_{m \to \infty} \frac{m_0}{m}$

Definiere (nicht beobachtbare) empirische Verteilungsfunktionen:
$\hat{F}_{0,m}(t) = \frac{1}{m_0} \sum_{i \in I_0} \mathbb{I}_{\{p_i \le t\}}$,$\hat{F}_{1,m}(t) = \frac{1}{m-m_0} \sum_{j \in \{1,\ldots,m\} \setminus I_0} \mathbb{I}_{\{p_j \le t\}}$
und gemischte empirische Verteilungsfunktion:
$\hat{F}_m = \pi_0 \hat{F}_{0,m} + (1-\pi_0) \hat{F}_{1,m}$

Satz von Donsker

(i)

$$\sqrt{m}\left(\begin{pmatrix}\hat{F}_{0,m}\\ \hat{F}_{1,m}\end{pmatrix}-\begin{pmatrix}F_0\\ F_1\end{pmatrix}\right)\xrightarrow{d}\begin{pmatrix}Z_0\\ Z_1\end{pmatrix}\text{ auf }[0,1]$$

wobei $Z_0 \overset{d}{=} \mathbb{B}$, $Z_1 \overset{d}{=} \mathbb{B} \circ F_1$ mit $\mathbb{B}$ der Standard Brownschen Brücke.

(ii)

$$\sqrt{m}(\hat{F}_m - F) \xrightarrow{d} Z \text{ auf } [0,1]$$

und $Z = \pi_0 Z_0 + (1-\pi_0)Z_1$ ist stetiger Gaußprozeß

Die *FDP* als stochastischer Prozess:

$$FDP_m(t) = \frac{\pi_0 \hat{F}_{0,m}(t)}{\max(\hat{F}_m(t), \frac{1}{m})} \text{bei (det.) Schwellwert } t \in [0,1]$$

Bezeichne zu einer Testprozedur $T_m = T(\hat{F}_m)$ den den zufälligen Schwellwert.
Wobei $T : D[0,1] \to [0,1]$ eine Abbildung sei,
die einer „cadlag"-Funktion F einen Schwellwert zuordnet.
Beispiel: (Benjamini-Hochberg)

$$T(F_m) = sup\{t \geq 0 : F_m(t) = 1/m \sum \mathbb{I}_{\{p_i \leq t\}} \geq t/\alpha\}$$

Satz:

Die Abbildung $T : D[0,1] \to [0,1]$ sei Hadamard-differenzierbar in der Verteilungsfunktion F, mit Hadamard-Ableitung $\dot{T}_F : C[0,1] \to \mathbb{R}$ Bezeichne $p(t) = \frac{\pi_0 t}{F(t)}$ die false dicscovery rate bei Schwellwert t, sowie $T^* = T(F)$ den Schwellwert der wahren Verteilung Dann gilt:

$$\sqrt{m}\,(FDP_m(T_m) - p(T)) \xrightarrow{d} Y$$

wobei $Y = p(T^*)(1 - p(T^*)) \left(\frac{Z_0(T^*)}{T^*} - \frac{Z_1(T^*)}{F_1(T^*)} \right) + p'(T^*)\dot{T}_F(Z)$

Bemerkung:
Y ist eine Zufallsvariable.