Wpływ korelacji statystyk testowych na współczynniki FDR i FWER

Statystyka w medycynie 2008/09

Plan wykładu

Przypomnienie

Możliwe wyniki, gdy testujemy m hipotez, wśród których m_0 jest prawdziwych:

	Nieodrzucone	**Odrzucone**	**Razem**
H_0 prawdziwa	U	V	m_0
H_A prawdziwa	T	S	m_1
	W	R	m

$$FWER = \mathbf{Pr}(V \geqslant 1)$$

$$FDR = \mathbf{E}\Big(\frac{V}{R} \mid R > 0\Big)\mathbf{Pr}(R > 0)$$

$$pFDR = \mathbf{E}\Big(\frac{V}{R} \mid R > 0\Big)$$

Fakt

Zawsze $FDR \leqslant FWER$ *oraz* $FDR \leqslant pFDR$.

Przykład

Dla m lotnisk niech:
H_0^i: na i - tym lotniku złapany człowiek nie jest terrorystą
H_A^i: na i - tym lotniku złapany człowiek jest terrorystą.
Wówczas:

- *FWER* to prawdopodobieństwo, że co najmniej jeden niewinny zostanie uznany za terrorystę
- *FDR* to średni procent niewinnych wśród uznanych za terrorystów (gdzie brak uznanych za terrorystów oznacza procent $= 0$)
- *pFDR* to średni procent niewinnych wśród uznanych za terrorystów, gdzie średnia jest liczona tylko po tych przypadkach, gdy kogoś uznano za terrorystę.

Przykład

Dla m lotnisk niech:
H_0^i: na i - tym lotniku złapany człowiek nie jest terrorystą
H_A^i: na i - tym lotniku złapany człowiek jest terrorystą.
Wówczas:

- *FWER* to prawdopodobieństwo, że co najmniej jeden niewinny zostanie uznany za terrorystę
- *FDR* to średni procent niewinnych wśród uznanych za terrorystów (gdzie brak uznanych za terrorystów oznacza procent $= 0$)
- *pFDR* to średni procent niewinnych wśród uznanych za terrorystów, gdzie średnia jest liczona tylko po tych przypadkach, gdy kogoś uznano za terrorystę.

Przykład

Dla m lotnisk niech:
H_0^i: na i - tym lotniku złapany człowiek nie jest terrorystą
H_A^i: na i - tym lotniku złapany człowiek jest terrorystą.
Wówczas:

- *FWER* to prawdopodobieństwo, że co najmniej jeden niewinny zostanie uznany za terrorystę
- *FDR* to średni procent niewinnych wśród uznanych za terrorystów (gdzie brak uznanych za terrorystów oznacza procent $= 0$)
- *pFDR* to średni procent niewinnych wśród uznanych za terrorystów, gdzie średnia jest liczona tylko po tych przypadkach, gdy kogoś uznano za terrorystę.

Przykład

Dla m lotnisk niech:
H_0^i: na i - tym lotniku złapany człowiek nie jest terrorystą
H_A^i: na i - tym lotniku złapany człowiek jest terrorystą.
Wówczas:

- *FWER* to prawdopodobieństwo, że co najmniej jeden niewinny zostanie uznany za terrorystę
- *FDR* to średni procent niewinnych wśród uznanych za terrorystów (gdzie brak uznanych za terrorystów oznacza procent $= 0$)
- *pFDR* to średni procent niewinnych wśród uznanych za terrorystów, gdzie średnia jest liczona tylko po tych przypadkach, gdy kogoś uznano za terrorystę.

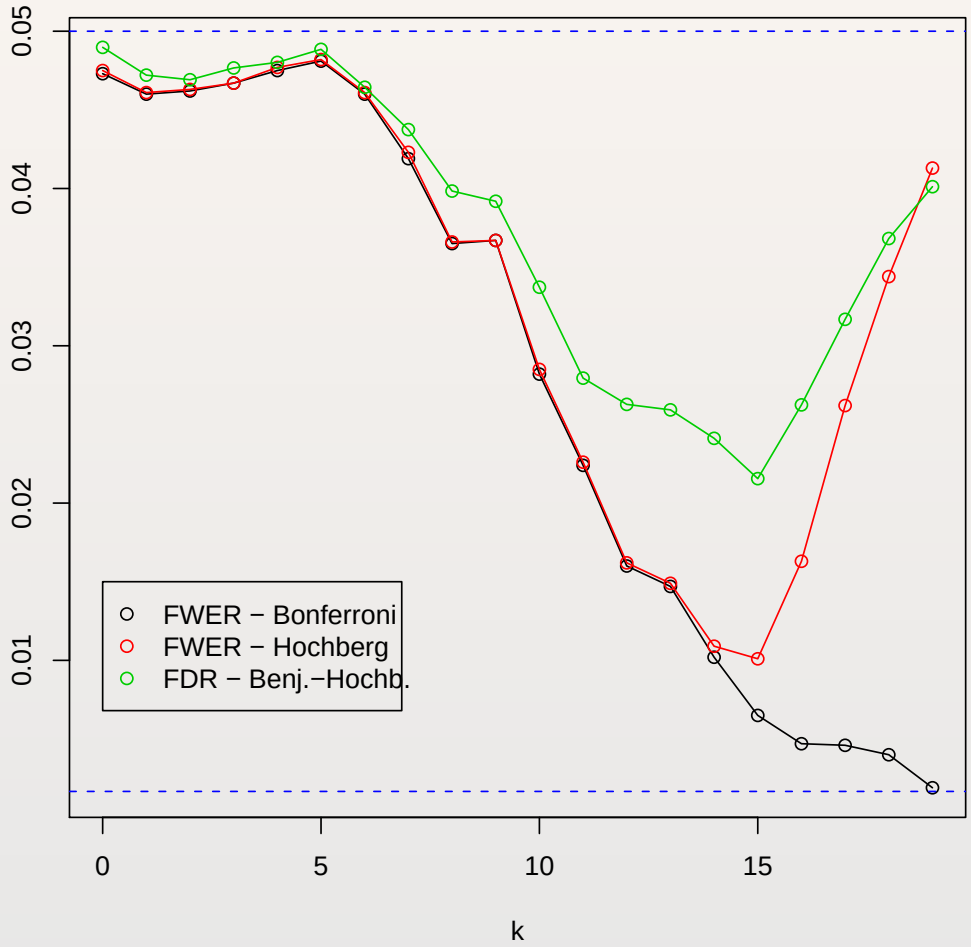
Motywacje

Rozważmy następujący przykład: Testujemy m hipotez, gdzie $H_0 : X \sim N(0,1)$, $H_A : X \sim N(1,1)$ przy użuciu testu $\mathrm{T(x)} = \mathrm{x}$, gdzie obszar krytyczny jest postaci $\{x : x \geqslant t\}$. Wówczas p - wartość dla przykładu x wynosi $p(x) = 1 - \Phi(x)$, gdzie Φ to dystrybuanta rozkładu $N(0,1)$. Mamy zatem wektor statystyk testowych $(T_1, ..., T_m)$ o rozkładzie $N(\mu, \Sigma)$, gdzie wektor średnich μ spełnia $\mu(i) \in \{0,1\}$ dla $i \in 1, ..., m$, zaś Σ jest macierzą kowariancji, która na diagonali ma same jedynki.
Wiemy już, że w przypadku pozytywnie skorelowanych $(T_1, ..., T_m)$ procedura Benjaminiego - Hochberga doboru poziomu istotności gwarantuje nam kontrolę *FDR* na ustalonym poziomie α. Pytanie co się stanie, gdy wprowadzimy skorelowanie ujemne.

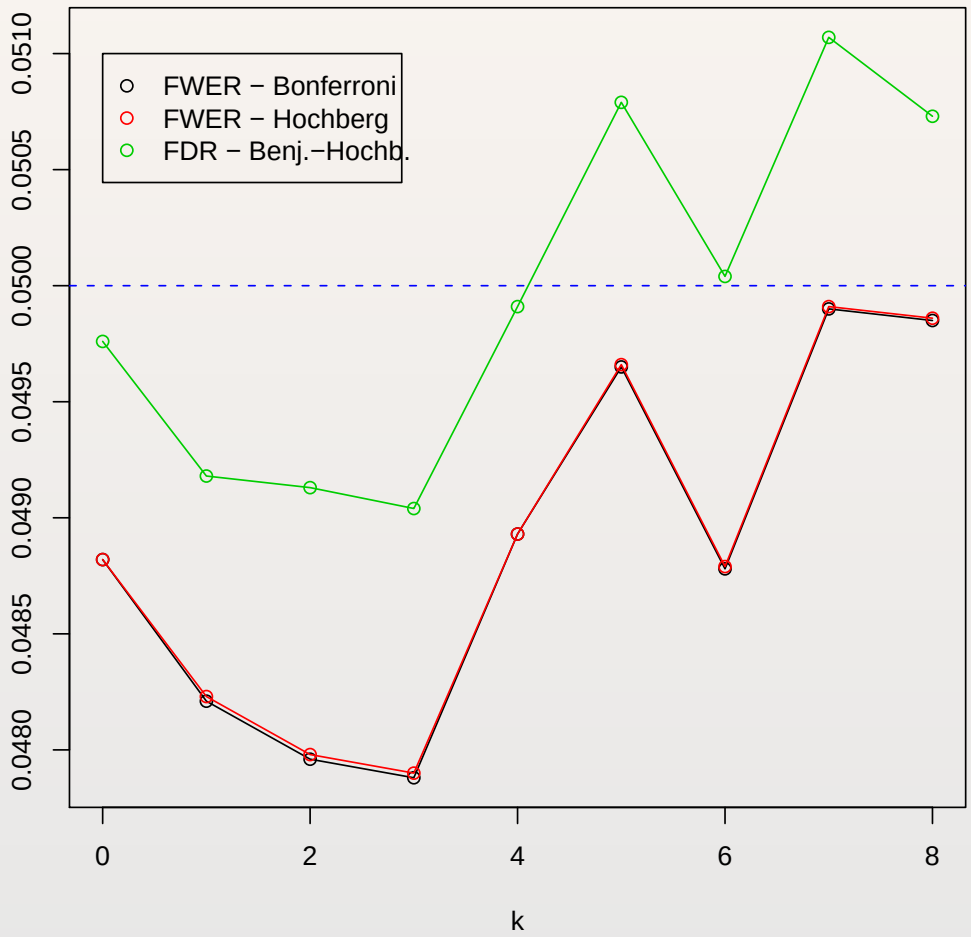
Motywacje

Przeprowadziliśmy eksperyment, w którym dla różnych parametrów m, m_0, Σ, α losowałem it razy statystyki $(T_1, ..., T_m)$ z rozkładu $N(\mu, \Sigma)$, po uśrednieniu wyników otrzymując przybliżony *FWER* w sytuacji zastosowania procedury Bonferroniego, a także Hochberga, oraz przybliżony *FDR* gdy zastosuje się metodę Benjaminiego - Hochberga. Na najbliższym slajdzie $m = 30$, $m_0 = 1$, $\alpha = 0.05$, $it = 10000$, zaś parametr k przebiega od 0 do 19, wyznaczając macierz Σ w postaci: $\Sigma = AA^T$, gdzie A jest macierzą $(0.02 + 0.05k)N + (1 - 0.02 - 0.05k)Id$ po unormowaniu wierszami, przy czym $N[i,] = e_1 \forall i$ (a zatem wraz ze wzrostem k przechodzimy od statystyk prawie niezależnych do statystyk bardzo mocno skorelowanych pozytywnie). Na drugim slajdzie $m = 10$, $m_0 = 0$, $\alpha = 0.05$, $it = 100000$, zaś parametr k biegnący od 0 do 8 wyznacza macierz Σ w postaci: $\Sigma = AA^T$, gdzie A jest macierzą $(0.02 + 0.12k)S + (1 - 0.02 - 0.12k)Id$ po unormowaniu wierszami, przy czym S to macierz, której wiersze tworzą zestaw wektorów "najmocniej rozstrzelonych"(a zatem wraz ze wzrostem k przechodzimy od statystyk prawie niezależnych do sytuacji, gdy każde dwie statystyki T_i, T_j, $i \neq j$ są negatywnie skorelowane).

Eksperyment 1



Eksperyment 2



Konkluzje

- otrzymany *FWER* jest zawsze $\geqslant$ w przypadku metody Hochberga, w porównaniu z metodą Bonferroniego (co jest ogólną własnością)
- widzimy, że w przypadku eksperymentu I *FWER* otrzymywany stosując procedurę Bonferroniego spada wraz ze wzrostem zależności statystyk do poziomu $\frac{\alpha}{m}$ (wynika to z tego, że dla największych k zmienne losowe $(T_1, ..., T_m)$ są prawie sobie równe, zatem $\mathbf{P}(V \geqslant 1) \approx P(X \in \Gamma_{\frac{\alpha}{m}}) = \frac{\alpha}{m}$, gdzie $X \sim N(0,1)$, zaś $\Gamma_{\frac{\alpha}{m}}$ to obszar krytyczny odpowiadający poziomowi istotności $\frac{\alpha}{m}$)
- w przypadku eksperymentu II zobaczyliśmy, że w przypadku ogólnej zależności procedura Benjaminiego - Hochberga przestaje gwarantować utrzymawanie *FDR* na poziomie α.

Konkluzje

- otrzymany *FWER* jest zawsze $\geqslant$ w przypadku metody Hochberga, w porównaniu z metodą Bonferroniego (co jest ogólną własnością)
- widzimy, że w przypadku eksperymentu I *FWER* otrzymywany stosując procedurę Bonferroniego spada wraz ze wzrostem zależności statystyk do poziomu $\frac{\alpha}{m}$ (wynika to z tego, że dla największych k zmienne losowe $(T_1, ..., T_m)$ są prawie sobie równe, zatem $\mathbf{P}(V \geqslant 1) \approx P(X \in \Gamma_{\frac{\alpha}{m}}) = \frac{\alpha}{m}$, gdzie $X \sim N(0,1)$, zaś $\Gamma_{\frac{\alpha}{m}}$ to obszar krytyczny odpowiadający poziomowi istotności $\frac{\alpha}{m}$)
- w przypadku eksperymentu II zobaczyliśmy, że w przypadku ogólnej zależności procedura Benjaminiego - Hochberga przestaje gwarantować utrzymawanie *FDR* na poziomie α.

Konkluzje

- otrzymany *FWER* jest zawsze $\geqslant$ w przypadku metody Hochberga, w porównaniu z metodą Bonferroniego (co jest ogólną własnością)
- widzimy, że w przypadku eksperymentu I *FWER* otrzymywany stosując procedurę Bonferroniego spada wraz ze wzrostem zależności statystyk do poziomu $\frac{\alpha}{m}$ (wynika to z tego, że dla największych k zmienne losowe $(T_1, ..., T_m)$ są prawie sobie równe, zatem $\mathbf{P}(V \geqslant 1) \approx P(X \in \Gamma_{\frac{\alpha}{m}}) = \frac{\alpha}{m}$, gdzie $X \sim N(0,1)$, zaś $\Gamma_{\frac{\alpha}{m}}$ to obszar krytyczny odpowiadający poziomowi istotności $\frac{\alpha}{m}$)
- w przypadku eksperymentu II zobaczyliśmy, że w przypadku ogólnej zależności procedura Benjaminiego - Hochberga przestaje gwarantować utrzymawanie *FDR* na poziomie α.

Cele estymacji *FDR* i *pFDR*

- estymator *FDR* jest fundamentalnym składnikiem algorytmu wykorzystującego *q* - wartości, pozwalającego na kontrolę *FDR* w przypadku tzw. "słabej zależności"
- nie da się stworzyć metody pozwalającej na kontrolę *pFDR* na ustalonym poziomie $\alpha < 1$ (bo w przypadku, gdy $m_0 = m$ cokolwiek byśmy zrobili $pFDR = 1$). Mając jednak konserwatywny estymator *pFDR* jesteśmy w stanie kontrolować ten współczynnik na tyle, na ile wyestymowana wartość nam pozwoli.

Cele estymacji *FDR* i *pFDR*

- estymator *FDR* jest fundamentalnym składnikiem algorytmu wykorzystującego *q* - wartości, pozwalającego na kontrolę *FDR* w przypadku tzw. "słabej zależności"
- nie da się stworzyć metody pozwalającej na kontrolę *pFDR* na ustalonym poziomie $\alpha < 1$ (bo w przypadku, gdy $m_0 = m$ cokolwiek byśmy zrobili $pFDR = 1$). Mając jednak konserwatywny estymator *pFDR* jesteśmy w stanie kontrolować ten współczynnik na tyle, na ile wyestymowana wartość nam pozwoli.

Wyprowadzenie estymatorów *FDR* i *pFDR*

Wprowadźmy oznaczenia:

- Γ_α - obszar krytyczny dla testu na poziomie istotności α
- $V(\Gamma)$, $W(\Gamma)$, $R(\Gamma)$ - to nasze zmienne losowe V, W i R w sytuacji gdy testujemy wszystkie m hipotez dla obszaru krytycznego Γ (zajmujemy się sytuacją, gdy mamy m identycznych testów, a zatem wspólny obszar krytyczny oznacza wspólny poziom istotności)
- $V^0(\Gamma)$, $W^0(\Gamma)$, $R^0(\Gamma)$ - jak wyżej, tylko przy założeniu, że $m_0 = m$
- $FDR(\Gamma_\alpha)$, $pFDR(\Gamma_\alpha)$ - odpowiednio *FDR* i *pFDR* gdy testujemy wszystkie m hipotez na poziomie istotności α.

Na kolejnych slajdach znajdują się następujące 3 podejścia prowadzące do tych samych estymatorów $\widehat{FDR}(\Gamma_\alpha)$ i $\widehat{pFDR}(\Gamma_\alpha)$:

- intuicujne
- histogramowe
- bayesowskiege.

Wyprowadzenie estymatorów *FDR* i *pFDR*

Wprowadźmy oznaczenia:

- Γ_α - obszar krytyczny dla testu na poziomie istotności α
- $V(\Gamma)$, $W(\Gamma)$, $R(\Gamma)$ - to nasze zmienne losowe V, W i R w sytuacji gdy testujemy wszystkie m hipotez dla obszaru krytycznego Γ (zajmujemy się sytuacją, gdy mamy m identycznych testów, a zatem wspólny obszar krytyczny oznacza wspólny poziom istotności)
- $V^0(\Gamma)$, $W^0(\Gamma)$, $R^0(\Gamma)$ - jak wyżej, tylko przy założeniu, że $m_0 = m$
- $FDR(\Gamma_\alpha)$, $pFDR(\Gamma_\alpha)$ - odpowiednio *FDR* i *pFDR* gdy testujemy wszystkie m hipotez na poziomie istotności α.

Na kolejnych slajdach znajdują się następujące 3 podejścia prowadzące do tych samych estymatorów $\widehat{FDR}(\Gamma_\alpha)$ i $\widehat{pFDR}(\Gamma_\alpha)$:

- intuicujne
- histogramowe
- bayesowskiege.

Wyprowadzenie estymatorów *FDR* i *pFDR*

Wprowadźmy oznaczenia:

- Γ_α - obszar krytyczny dla testu na poziomie istotności α
- $V(\Gamma)$, $W(\Gamma)$, $R(\Gamma)$ - to nasze zmienne losowe V, W i R w sytuacji gdy testujemy wszystkie m hipotez dla obszaru krytycznego Γ (zajmujemy się sytuacją, gdy mamy m identycznych testów, a zatem wspólny obszar krytyczny oznacza wspólny poziom istotności)
- $V^0(\Gamma)$, $W^0(\Gamma)$, $R^0(\Gamma)$ - jak wyżej, tylko przy założeniu, że $m_0 = m$
- $FDR(\Gamma_\alpha)$, $pFDR(\Gamma_\alpha)$ - odpowiednio *FDR* i *pFDR* gdy testujemy wszystkie m hipotez na poziomie istotności α.

Na kolejnych slajdach znajdują się następujące 3 podejścia prowadzące do tych samych estymatorów $\widehat{FDR}(\Gamma_\alpha)$ i $\widehat{pFDR}(\Gamma_\alpha)$:

- intuicujne
- histogramowe
- bayesowskiege.

Wyprowadzenie estymatorów *FDR* i *pFDR*

Wprowadźmy oznaczenia:

- Γ_α - obszar krytyczny dla testu na poziomie istotności α
- $V(\Gamma)$, $W(\Gamma)$, $R(\Gamma)$ - to nasze zmienne losowe V, W i R w sytuacji gdy testujemy wszystkie m hipotez dla obszaru krytycznego Γ (zajmujemy się sytuacją, gdy mamy m identycznych testów, a zatem wspólny obszar krytyczny oznacza wspólny poziom istotności)
- $V^0(\Gamma)$, $W^0(\Gamma)$, $R^0(\Gamma)$ - jak wyżej, tylko przy założeniu, że $m_0 = m$
- $FDR(\Gamma_\alpha)$, $pFDR(\Gamma_\alpha)$ - odpowiednio *FDR* i *pFDR* gdy testujemy wszystkie m hipotez na poziomie istotności α.

Na kolejnych slajdach znajdują się następujące 3 podejścia prowadzące do tych samych estymatorów $\widehat{FDR}(\Gamma_\alpha)$ i $\widehat{pFDR}(\Gamma_\alpha)$:

- intuicujne
- histogramowe
- bayesowskiege.

Wyprowadzenie estymatorów *FDR* i *pFDR*

Wprowadźmy oznaczenia:

- Γ_α - obszar krytyczny dla testu na poziomie istotności α
- $V(\Gamma)$, $W(\Gamma)$, $R(\Gamma)$ - to nasze zmienne losowe V, W i R w sytuacji gdy testujemy wszystkie m hipotez dla obszaru krytycznego Γ (zajmujemy się sytuacją, gdy mamy m identycznych testów, a zatem wspólny obszar krytyczny oznacza wspólny poziom istotności)
- $V^0(\Gamma)$, $W^0(\Gamma)$, $R^0(\Gamma)$ - jak wyżej, tylko przy założeniu, że $m_0 = m$
- $FDR(\Gamma_\alpha)$, $pFDR(\Gamma_\alpha)$ - odpowiednio *FDR* i *pFDR* gdy testujemy wszystkie m hipotez na poziomie istotności α.

Na kolejnych slajdach znajdują się następujące 3 podejścia prowadzące do tych samych estymatorów $\widehat{FDR}(\Gamma_\alpha)$ i $\widehat{pFDR}(\Gamma_\alpha)$:

- intuicujne
- histogramowe
- bayesowskiege.

Podejście intuicyjne

- przyjmując $\frac{V}{R} = 0$ dla $R = 0$ mamy, że $FDR(\Gamma_\alpha) = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]$
- dla dużych m mamy $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$
- kładąc jako estymator $\mathbf{E}[R(\Gamma_\alpha)]$ obserwowalną zmienną $R(\Gamma_\alpha)$, robiąc konieczną korektę, mamy na razie formułę przybliżającą *FDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{R(\Gamma_\alpha)\vee 1}$ oraz formułę przybliżającą *pFDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{(R(\Gamma_\alpha)\vee 1)\mathbf{Pr}(R(\Gamma_\alpha)>0)}$
- oznaczając $\pi_0 = \frac{m_0}{m}$ zauważmy, że intuicyjnie:
  $\mathbf{E}[V(\Gamma_\alpha)] \approx \pi_0 \mathbf{E}[R^0(\Gamma_\alpha)] = \pi_0 m\alpha$,
  $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{Pr}(R(\Gamma_\alpha)>0)} \approx \pi_0 \frac{\mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)} = \frac{\pi_0 m\alpha}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- dla odpowiednio dużego $\lambda \in [0,1]$ możnaby się spodziewać, że: $\pi_0 = \frac{W(\Gamma_\lambda)}{\mathbf{E}(W^0(\Gamma_\lambda))}$, zatem kładziemy $\widehat{\pi_0} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$ (obserwowalne)
- składając wszystko razem dostajemy sparametryzowane estymatory: $\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha)\vee 1]}$, $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$ otrzymujemy metodą Monte Carlo, o ile umiemy **sensownie** symulować m hipotez zerowych

Podejście intuicyjne

- przyjmując $\frac{V}{R} = 0$ dla $R = 0$ mamy, że $FDR(\Gamma_\alpha) = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]$
- dla dużych m mamy $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$
- kładąc jako estymator $\mathbf{E}[R(\Gamma_\alpha)]$ obserwowalną zmienną $R(\Gamma_\alpha)$, robiąc konieczną korektę, mamy na razie formułę przybliżającą *FDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{R(\Gamma_\alpha)\vee 1}$ oraz formułę przybliżającą *pFDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{(R(\Gamma_\alpha)\vee 1)\mathbf{Pr}(R(\Gamma_\alpha)>0)}$
- oznaczając $\pi_0 = \frac{m_0}{m}$ zauważmy, że intuicyjnie:
  $\mathbf{E}[V(\Gamma_\alpha)] \approx \pi_0 \mathbf{E}[R^0(\Gamma_\alpha)] = \pi_0 m\alpha$,
  $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{Pr}(R(\Gamma_\alpha)>0)} \approx \pi_0 \frac{\mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)} = \frac{\pi_0 m\alpha}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- dla odpowiednio dużego $\lambda \in [0,1]$ możnaby się spodziewać, że:
  $\pi_0 = \frac{W(\Gamma_\lambda)}{\mathbf{E}(W^0(\Gamma_\lambda))}$, zatem kładziemy $\widehat{\pi_0} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$ (obserwowalne)
- składając wszystko razem dostajemy sparametryzowane estymatory:
  $\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha)\vee 1]}$, $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$ otrzymujemy metodą Monte Carlo, o ile umiemy **sensownie** symulować m hipotez zerowych

Podejście intuicyjne

- przyjmując $\frac{V}{R} = 0$ dla $R = 0$ mamy, że $FDR(\Gamma_\alpha) = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]$
- dla dużych m mamy $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$
- kładąc jako estymator $\mathbf{E}[R(\Gamma_\alpha)]$ obserwowalną zmienną $R(\Gamma_\alpha)$, robiąc konieczną korektę, mamy na razie formułę przybliżającą *FDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{R(\Gamma_\alpha)\vee 1}$ oraz formułę przybliżającą *pFDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{(R(\Gamma_\alpha)\vee 1)\mathbf{Pr}(R(\Gamma_\alpha)>0)}$
- oznaczając $\pi_0 = \frac{m_0}{m}$ zauważmy, że intuicyjnie:
  $\mathbf{E}[V(\Gamma_\alpha)] \approx \pi_0 \mathbf{E}[R^0(\Gamma_\alpha)] = \pi_0 m\alpha$,
  $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{Pr}(R(\Gamma_\alpha)>0)} \approx \pi_0 \frac{\mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)} = \frac{\pi_0 m\alpha}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- dla odpowiednio dużego $\lambda \in [0,1]$ możnaby się spodziewać, że:
  $\pi_0 = \frac{W(\Gamma_\lambda)}{\mathbf{E}(W^0(\Gamma_\lambda))}$, zatem kładziemy $\widehat{\pi_0} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$ (obserwowalne)
- składając wszystko razem dostajemy sparametryzowane estymatory:
  $\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha)\vee 1]}$, $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$ otrzymujemy metodą Monte Carlo, o ile umiemy **sensownie** symulować m hipotez zerowych

Podejście intuicyjne

- przyjmując $\frac{V}{R} = 0$ dla $R = 0$ mamy, że $FDR(\Gamma_\alpha) = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]$
- dla dużych m mamy $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$
- kładąc jako estymator $\mathbf{E}[R(\Gamma_\alpha)]$ obserwowalną zmienną $R(\Gamma_\alpha)$, robiąc konieczną korektę, mamy na razie formułę przybliżającą *FDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{R(\Gamma_\alpha)\vee 1}$ oraz formułę przybliżającą *pFDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{(R(\Gamma_\alpha)\vee 1)\mathbf{Pr}(R(\Gamma_\alpha)>0)}$
- oznaczając $\pi_0 = \frac{m_0}{m}$ zauważmy, że intuicyjnie:
  $\mathbf{E}[V(\Gamma_\alpha)] \approx \pi_0 \mathbf{E}[R^0(\Gamma_\alpha)] = \pi_0 m\alpha$,
  $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{Pr}(R(\Gamma_\alpha)>0)} \approx \pi_0 \frac{\mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)} = \frac{\pi_0 m\alpha}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- dla odpowiednio dużego $\lambda \in [0,1]$ możnaby się spodziewać, że: $\pi_0 = \frac{W(\Gamma_\lambda)}{\mathbf{E}(W^0(\Gamma_\lambda))}$, zatem kładziemy $\widehat{\pi_0} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$ (obserwowalne)
- składając wszystko razem dostajemy sparametryzowane estymatory: $\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha)\vee 1]}$, $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$ otrzymujemy metodą Monte Carlo, o ile umiemy **sensownie** symulować m hipotez zerowych

Podejście intuicyjne

- przyjmując $\frac{V}{R} = 0$ dla $R = 0$ mamy, że $FDR(\Gamma_\alpha) = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]$
- dla dużych m mamy $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$
- kładąc jako estymator $\mathbf{E}[R(\Gamma_\alpha)]$ obserwowalną zmienną $R(\Gamma_\alpha)$, robiąc konieczną korektę, mamy na razie formułę przybliżającą *FDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{R(\Gamma_\alpha)\vee 1}$ oraz formułę przybliżającą *pFDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{(R(\Gamma_\alpha)\vee 1)\mathbf{Pr}(R(\Gamma_\alpha)>0)}$
- oznaczając $\pi_0 = \frac{m_0}{m}$ zauważmy, że intuicyjnie:
  $\mathbf{E}[V(\Gamma_\alpha)] \approx \pi_0 \mathbf{E}[R^0(\Gamma_\alpha)] = \pi_0 m\alpha$,
  $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{Pr}(R(\Gamma_\alpha)>0)} \approx \pi_0 \frac{\mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)} = \frac{\pi_0 m\alpha}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- dla odpowiednio dużego $\lambda \in [0,1]$ możnaby się spodziewać, że:
  $\pi_0 = \frac{W(\Gamma_\lambda)}{\mathbf{E}(W^0(\Gamma_\lambda))}$, zatem kładziemy $\widehat{\pi_0} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$ (obserwowalne)
- składając wszystko razem dostajemy sparametryzowane estymatory:
  $\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha)\vee 1]}$, $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$ otrzymujemy metodą Monte Carlo, o ile umiemy **sensownie** symulować m hipotez zerowych

Podejście intuicyjne

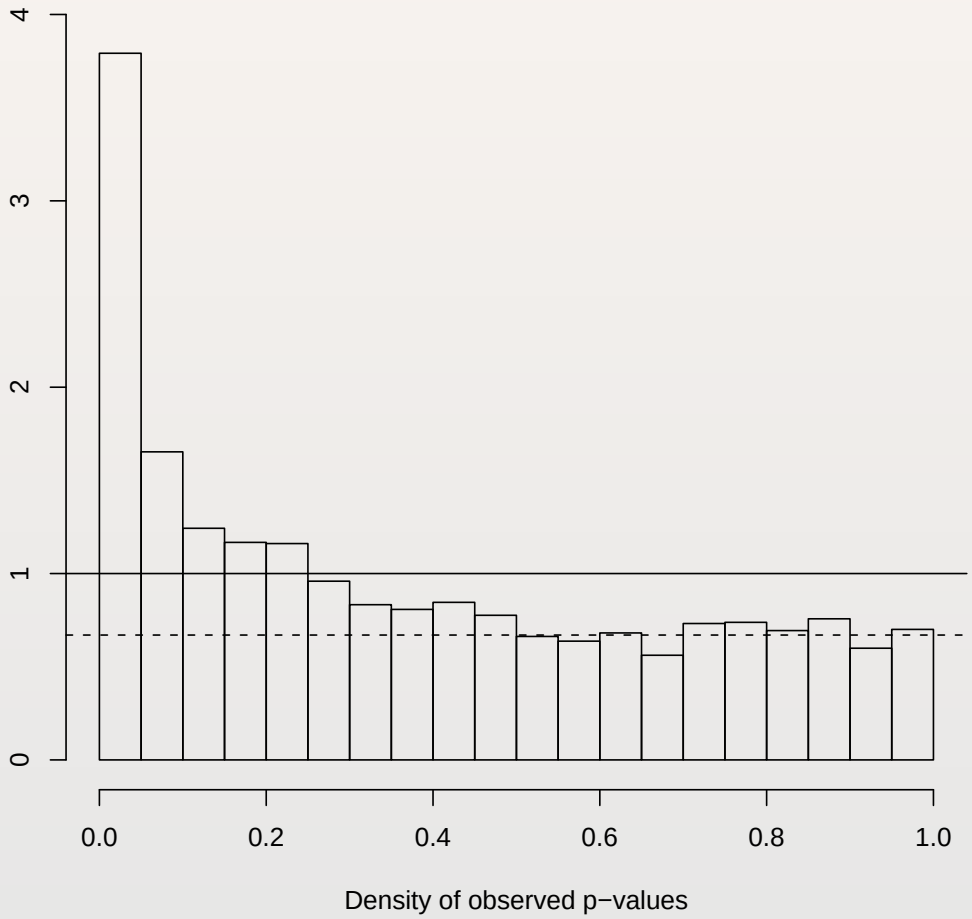
- przyjmując $\frac{V}{R} = 0$ dla $R = 0$ mamy, że $FDR(\Gamma_\alpha) = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]$
- dla dużych m mamy $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$
- kładąc jako estymator $\mathbf{E}[R(\Gamma_\alpha)]$ obserwowalną zmienną $R(\Gamma_\alpha)$, robiąc konieczną korektę, mamy na razie formułę przybliżającą *FDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{R(\Gamma_\alpha)\vee 1}$ oraz formułę przybliżającą *pFDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{(R(\Gamma_\alpha)\vee 1)\mathbf{Pr}(R(\Gamma_\alpha)>0)}$
- oznaczając $\pi_0 = \frac{m_0}{m}$ zauważmy, że intuicyjnie:
  $\mathbf{E}[V(\Gamma_\alpha)] \approx \pi_0 \mathbf{E}[R^0(\Gamma_\alpha)] = \pi_0 m\alpha$,
  $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{Pr}(R(\Gamma_\alpha)>0)} \approx \pi_0 \frac{\mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)} = \frac{\pi_0 m\alpha}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- dla odpowiednio dużego $\lambda \in [0,1]$ możnaby się spodziewać, że:
  $\pi_0 = \frac{W(\Gamma_\lambda)}{\mathbf{E}(W^0(\Gamma_\lambda))}$, zatem kładziemy $\widehat{\pi_0} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$ (obserwowalne)
- składając wszystko razem dostajemy sparametryzowane estymatory:
  $\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha)\vee 1]}$, $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$ otrzymujemy metodą Monte Carlo, o ile umiemy **sensownie** symulować m hipotez zerowych

Podejście intuicyjne

- przyjmując $\frac{V}{R} = 0$ dla $R = 0$ mamy, że $FDR(\Gamma_\alpha) = \mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right]$
- dla dużych m mamy $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$
- kładąc jako estymator $\mathbf{E}[R(\Gamma_\alpha)]$ obserwowalną zmienną $R(\Gamma_\alpha)$, robiąc konieczną korektę, mamy na razie formułę przybliżającą *FDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{R(\Gamma_\alpha)\vee 1}$ oraz formułę przybliżającą *pFDR*: $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{(R(\Gamma_\alpha)\vee 1)\mathbf{Pr}(R(\Gamma_\alpha)>0)}$
- oznaczając $\pi_0 = \frac{m_0}{m}$ zauważmy, że intuicyjnie:
  $\mathbf{E}[V(\Gamma_\alpha)] \approx \pi_0 \mathbf{E}[R^0(\Gamma_\alpha)] = \pi_0 m\alpha$,
  $\frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{Pr}(R(\Gamma_\alpha)>0)} \approx \pi_0 \frac{\mathbf{E}[R^0(\Gamma_\alpha)]}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)} = \frac{\pi_0 m\alpha}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- dla odpowiednio dużego $\lambda \in [0,1]$ możnaby się spodziewać, że:
  $\pi_0 = \frac{W(\Gamma_\lambda)}{\mathbf{E}(W^0(\Gamma_\lambda))}$, zatem kładziemy $\widehat{\pi_0} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$ (obserwowalne)
- składając wszystko razem dostajemy sparametryzowane estymatory:
  $\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha)\vee 1]}$, $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha)>0)}$
- $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$ otrzymujemy metodą Monte Carlo, o ile umiemy **sensownie** symulować m hipotez zerowych

Podejście histogramowe

Przykładowy histogram gęstości p - wartości dla ok. 3000 testów:



Podejście histogramowe

Kluczowy w tym podejściu jest:

Fakt

p - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$.

Wnioski z niego są takie:

- $\mathbf{E}[V(\Gamma_\alpha)] = m_0\alpha = \pi_0 m\alpha$
- skoro *p* - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$, zaś *p* - wartości odpowiadające hipotezom alternatywnym są przede wszystkim skupione w okolicach 0, to należy się spodziewać, że pole pod kropkowaną kreską na obrazku odpowiada mniej więcej proporcji hipotez zerowych wśród wszystkich, a więc wysokość tej kreski jest etymatorem π_0
- to na jakiej wysokości umieścimy kreskę zależy od naszego postrzegania, od którego miejsca histogram się "wypłaszczył". Jeśli uznamy, że tym miejscem jest pewne λ, to wówczas krótka analiza obrazka prowadzi nas do estymatora:
$\widehat{\pi}_0(\lambda) = \frac{\#\{p_i>\lambda; i=1,\ldots,m\}}{m(1-\lambda)} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$

Podejście histogramowe

Kluczowy w tym podejściu jest:

Fakt

p - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$.

Wnioski z niego są takie:

- $\mathbf{E}[V(\Gamma_\alpha)] = m_0\alpha = \pi_0 m\alpha$
- skoro *p* - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$, zaś *p* - wartości odpowiadające hipotezom alternatywnym są przede wszystkim skupione w okolicach 0, to należy się spodziewać, że pole pod kropkowaną kreską na obrazku odpowiada mniej więcej proporcji hipotez zerowych wśród wszystkich, a więc wysokość tej kreski jest etymatorem π_0
- to na jakiej wysokości umieścimy kreskę zależy od naszego postrzegania, od którego miejsca histogram się "wypłaszczył". Jeśli uznamy, że tym miejscem jest pewne λ, to wówczas krótka analiza obrazka prowadzi nas do estymatora:
$\widehat{\pi}_0(\lambda) = \frac{\#\{p_i>\lambda; i=1,\ldots,m\}}{m(1-\lambda)} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$

Podejście histogramowe

Kluczowy w tym podejściu jest:

Fakt

p - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$.

Wnioski z niego są takie:

- $\mathbf{E}[V(\Gamma_\alpha)] = m_0\alpha = \pi_0 m\alpha$
- skoro *p* - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$, zaś *p* - wartości odpowiadające hipotezom alternatywnym są przede wszystkim skupione w okolicach 0, to należy się spodziewać, że pole pod kropkowaną kreską na obrazku odpowiada mniej więcej proporcji hipotez zerowych wśród wszystkich, a więc wysokość tej kreski jest etymatorem π_0
- to na jakiej wysokości umieścimy kreskę zależy od naszego postrzegania, od którego miejsca histogram się "wypłaszczył". Jeśli uznamy, że tym miejscem jest pewne λ, to wówczas krótka analiza obrazka prowadzi nas do estymatora:
$\widehat{\pi}_0(\lambda) = \frac{\#\{p_i>\lambda; i=1,\ldots,m\}}{m(1-\lambda)} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$

Podejście histogramowe

Kluczowy w tym podejściu jest:

Fakt

p - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$.

Wnioski z niego są takie:

- $\mathbf{E}[V(\Gamma_\alpha)] = m_0\alpha = \pi_0 m\alpha$
- skoro *p* - wartości odpowiadające hipotezom zerowym mają rozkład $U(0,1)$, zaś *p* - wartości odpowiadające hipotezom alternatywnym są przede wszystkim skupione w okolicach 0, to należy się spodziewać, że pole pod kropkowaną kreską na obrazku odpowiada mniej więcej proporcji hipotez zerowych wśród wszystkich, a więc wysokość tej kreski jest etymatorem π_0
- to na jakiej wysokości umieścimy kreskę zależy od naszego postrzegania, od którego miejsca histogram się "wypłaszczył". Jeśli uznamy, że tym miejscem jest pewne λ, to wówczas krótka analiza obrazka prowadzi nas do estymatora:
$\widehat{\pi_0}(\lambda) = \frac{\#\{p_i > \lambda; i=1,\ldots,m\}}{m(1-\lambda)} = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$

Podejście histogramowe

Dalej tak jak poprzednio korzystamy z tego, że dla dużych m: $\mathbf{E}\left[\frac{V(\Gamma_\alpha)}{R(\Gamma_\alpha)}\right] \approx \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]}$, $\mathbf{E}[R(\Gamma_\alpha)]$ estymujemy poprzez $R(\Gamma_\alpha) \vee 1$, zaś na mocy poprzedniego slajdu $\mathbf{E}[V(\Gamma_\alpha)]$ estymujemy jako $\widehat{\pi_0}(\lambda)m\alpha$ dostając łącznie te same estymatory:

$$\widehat{FDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha) \vee 1]}, \widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)}$$

Parametr λ wymaga optymalizacji, o czym będzie mowa jeszcze później.

Podejście bayesowskie

Na poprzednich zajęciach pojawiło się:

Twierdzenie

Załóżmy, że wykonujemy m identycznych testów przy użyciu statystyk $T_1, ..., T_m$ *dla obszaru krytycznego* Γ. *Niech* $H_1, ..., H_n$ *będą zmiennymi losowymi, t. że* $H_i = 0$ *oznacza prawdziwość i - tej hipotezy, zaś* $H_i = 1$ *- nieprawdziwość. Załóżmy, że* (T_i, H_i) *są i.i.d., oraz że dla każdego i:* $(T_i \mid H_i = 0) \sim F_0$, $(T_i \mid H_i = 1) \sim F_1$. *Dalej załóżmy, że dla wszystkich i* $H_i \sim Bernoulli(1 - \pi_0)$. *Wówczas:*

$$pFDR(\Gamma) = \frac{\pi_0 \mathbf{Pr}(T \in \Gamma \mid H = 0)}{\mathbf{Pr}(T \in \Gamma)} = \mathbf{Pr}(H = 0 \mid T \in \Gamma)$$

.

Podejście bayesowskie

Okazuje się, że można dokonać w pewnym stopniu uogólnienia tego faktu na sytuację gdy nie zakładamy niezależności (T_i, H_i), za to wprowadzamy luźniejsze założenie, że zachodzi tzw. **słaba zależność**, tzn:
$\frac{\sum_{i=1}^{m} \mathbf{1}(T_i \in \Gamma)(1-H_i)}{\sum_{i=1}^{m}(1-H_i)} \xrightarrow{P} \mathbf{Pr}(T \in \Gamma \mid H = 0)$ i
$\frac{\sum_{i=1}^{m} \mathbf{1}(T_i \in \Gamma)H_i}{\sum_{i=1}^{m} H_i} \xrightarrow{P} \mathbf{Pr}(T \in \Gamma \mid H = 1)$ gdy $m \to \infty$.
Wówczas zachodzi:

Twierdzenie

Przy powyższym poluźnieniu założeń:

$$pFDR_m(\Gamma) \xrightarrow{p.n.} \frac{\pi_0 \mathbf{Pr}(T \in \Gamma \mid H = 0)}{\mathbf{Pr}(T \in \Gamma)}$$

gdzie $pFDR_m(\Gamma)$ *oznacza* $pFDR(\Gamma)$ *dla pierwszych m statystyk.*

Podejście bayesowskie

- w przypadku niezależności (T_i, H_i) luźna zależność wynika z MPWL
- z luźną zależnością mamy do czynienia w sytuacji, gdy zależności występują lokalnie w grupach wzajemnie niezależnych
- w praktycznych zastosowaniach mamy często do czynienia właśnie z taką luźną zależnością, sensowne wydaje się więc budowanie estymatora $pFDR(\Gamma_\alpha)$ na bazie formuły $\frac{\pi_0 \mathbf{Pr}(T \in \Gamma_\alpha | H=0)}{\mathbf{Pr}(T \in \Gamma_\alpha)}$
- kładąc w miejsce π_0 nasz estymator $\widehat{\pi_0}(\lambda) = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$, w miejsce $\mathbf{Pr}(T \in \Gamma_\alpha)$ po prostu $\frac{R(\Gamma_\alpha) \vee 1}{m}$ i mając $\mathbf{Pr}(T \in \Gamma_\alpha \mid H = 0) = \alpha$, dostajemy estymator: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha) \vee 1]}$
- ponieważ jednak przy $m \to \infty$ *FDR* i *pFDR* są tym samym, zaś w praktyce nie mamy $m \to \infty$, tylko jakieś ustalone m, skłonni jesteśmy przydzielić powyższy wzór jako estymator $FDR(\Gamma_\alpha)$, zaś jako estymator $pFDR(\Gamma_\alpha)$ położyć na tej samej zasadzie co w poprzednich podejściach: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)}$.

Podejście bayesowskie

- w przypadku niezależności (T_i, H_i) luźna zależność wynika z MPWL
- z luźną zależnością mamy do czynienia w sytuacji, gdy zależności występują lokalnie w grupach wzajemnie niezależnych
- w praktycznych zastosowaniach mamy często do czynienia właśnie z taką luźną zależnością, sensowne wydaje się więc budowanie estymatora $pFDR(\Gamma_\alpha)$ na bazie formuły $\frac{\pi_0 \mathbf{Pr}(T \in \Gamma_\alpha | H=0)}{\mathbf{Pr}(T \in \Gamma_\alpha)}$
- kładąc w miejsce π_0 nasz estymator $\widehat{\pi_0}(\lambda) = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$, w miejsce $\mathbf{Pr}(T \in \Gamma_\alpha)$ po prostu $\frac{R(\Gamma_\alpha) \vee 1}{m}$ i mając $\mathbf{Pr}(T \in \Gamma_\alpha \mid H = 0) = \alpha$, dostajemy estymator: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha) \vee 1]}$
- ponieważ jednak przy $m \to \infty$ *FDR* i *pFDR* są tym samym, zaś w praktyce nie mamy $m \to \infty$, tylko jakieś ustalone m, skłonni jesteśmy przydzielić powyższy wzór jako estymator $FDR(\Gamma_\alpha)$, zaś jako estymator $pFDR(\Gamma_\alpha)$ położyć na tej samej zasadzie co w poprzednich podejściach: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)}$.

Podejście bayesowskie

- w przypadku niezależności (T_i, H_i) luźna zależność wynika z MPWL
- z luźną zależnością mamy do czynienia w sytuacji, gdy zależności występują lokalnie w grupach wzajemnie niezależnych
- w praktycznych zastosowaniach mamy często do czynienia właśnie z taką luźną zależnością, sensowne wydaje się więc budowanie estymatora $pFDR(\Gamma_\alpha)$ na bazie formuły $\frac{\pi_0 \mathbf{Pr}(T \in \Gamma_\alpha | H=0)}{\mathbf{Pr}(T \in \Gamma_\alpha)}$
- kładąc w miejsce π_0 nasz estymator $\widehat{\pi_0}(\lambda) = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$, w miejsce $\mathbf{Pr}(T \in \Gamma_\alpha)$ po prostu $\frac{R(\Gamma_\alpha) \vee 1}{m}$ i mając $\mathbf{Pr}(T \in \Gamma_\alpha \mid H = 0) = \alpha$, dostajemy estymator: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha) \vee 1]}$
- ponieważ jednak przy $m \to \infty$ *FDR* i *pFDR* są tym samym, zaś w praktyce nie mamy $m \to \infty$, tylko jakieś ustalone m, skłonni jesteśmy przydzielić powyższy wzór jako estymator $FDR(\Gamma_\alpha)$, zaś jako estymator $pFDR(\Gamma_\alpha)$ położyć na tej samej zasadzie co w poprzednich podejściach: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)}$.

Podejście bayesowskie

- w przypadku niezależności (T_i, H_i) luźna zależność wynika z MPWL
- z luźną zależnością mamy do czynienia w sytuacji, gdy zależności występują lokalnie w grupach wzajemnie niezależnych
- w praktycznych zastosowaniach mamy często do czynienia właśnie z taką luźną zależnością, sensowne wydaje się więc budowanie estymatora $pFDR(\Gamma_\alpha)$ na bazie formuły $\frac{\pi_0 \mathbf{Pr}(T \in \Gamma_\alpha | H=0)}{\mathbf{Pr}(T \in \Gamma_\alpha)}$
- kładąc w miejsce π_0 nasz estymator $\widehat{\pi_0}(\lambda) = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$, w miejsce $\mathbf{Pr}(T \in \Gamma_\alpha)$ po prostu $\frac{R(\Gamma_\alpha) \vee 1}{m}$ i mając $\mathbf{Pr}(T \in \Gamma_\alpha \mid H = 0) = \alpha$, dostajemy estymator: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha) \vee 1]}$
- ponieważ jednak przy $m \to \infty$ *FDR* i *pFDR* są tym samym, zaś w praktyce nie mamy $m \to \infty$, tylko jakieś ustalone m, skłonni jesteśmy przydzielić powyższy wzór jako estymator $FDR(\Gamma_\alpha)$, zaś jako estymator $pFDR(\Gamma_\alpha)$ położyć na tej samej zasadzie co w poprzednich podejściach: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)}$.

Podejście bayesowskie

- w przypadku niezależności (T_i, H_i) luźna zależność wynika z MPWL
- z luźną zależnością mamy do czynienia w sytuacji, gdy zależności występują lokalnie w grupach wzajemnie niezależnych
- w praktycznych zastosowaniach mamy często do czynienia właśnie z taką luźną zależnością, sensowne wydaje się więc budowanie estymatora $pFDR(\Gamma_\alpha)$ na bazie formuły $\frac{\pi_0 \mathbf{Pr}(T \in \Gamma_\alpha | H=0)}{\mathbf{Pr}(T \in \Gamma_\alpha)}$
- kładąc w miejsce π_0 nasz estymator $\widehat{\pi_0}(\lambda) = \frac{W(\Gamma_\lambda)}{m(1-\lambda)}$, w miejsce $\mathbf{Pr}(T \in \Gamma_\alpha)$ po prostu $\frac{R(\Gamma_\alpha) \vee 1}{m}$ i mając $\mathbf{Pr}(T \in \Gamma_\alpha \mid H = 0) = \alpha$, dostajemy estymator: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{W(\Gamma_\lambda)\alpha}{(1-\lambda)[R(\Gamma_\alpha) \vee 1]}$
- ponieważ jednak przy $m \to \infty$ *FDR* i *pFDR* są tym samym, zaś w praktyce nie mamy $m \to \infty$, tylko jakieś ustalone m, skłonni jesteśmy przydzielić powyższy wzór jako estymator $FDR(\Gamma_\alpha)$, zaś jako estymator $pFDR(\Gamma_\alpha)$ położyć na tej samej zasadzie co w poprzednich podejściach: $\widehat{pFDR}_\lambda(\Gamma_\alpha) = \frac{\widehat{FDR}_\lambda(\Gamma_\alpha)}{\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)}$.

Najważniejsze własności $\widehat{FDR}_\lambda$ i $\widehat{pFDR}_\lambda$

1. Jeśli statystyki testowe są niezależne, to niezależnie od λ:

$$\mathbf{E}[\widehat{FDR}_\lambda(\Gamma_\alpha)] \geqslant FDR(\Gamma_\alpha), \mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geqslant pFDR(\Gamma_\alpha).$$

Własność prawdziwa również w sytuacji ogólnej zależności, jeśli:
$\mathbf{E}[R(\Gamma_\lambda) \mid R(\Gamma_\alpha)] \leqslant \frac{\mathbf{E}[R(\Gamma_\lambda)]}{\mathbf{E}[R(\Gamma_\alpha)]} R(\Gamma_\alpha)$ oraz
$\mathbf{E}[V(\Gamma_\alpha) \mid R(\Gamma_\alpha)] \leqslant \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]} R(\Gamma_\alpha)$.

2. Jeśli (T_i, H_i) są niezależne, to:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{p.n.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha) \geqslant pFDR(\Gamma_\alpha),$$

gdzie $\pi_1 = 1 - \pi_0$, a $g(\lambda)$ to moc testu T_i dla poziomu istotności λ.
W sytuacji luźnej zależności prawdziwy jest analogiczny wzór:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{P}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} \lim_{m\to\infty} pFDR_m(\Gamma_\alpha).$$

Najważniejsze własności $\widehat{FDR}_\lambda$ i $\widehat{pFDR}_\lambda$

1. Jeśli statystyki testowe są niezależne, to niezależnie od λ:

$$\mathbf{E}[\widehat{FDR}_\lambda(\Gamma_\alpha)] \geqslant FDR(\Gamma_\alpha), \mathbf{E}[\widehat{pFDR}_\lambda(\Gamma_\alpha)] \geqslant pFDR(\Gamma_\alpha).$$

Własność prawdziwa również w sytuacji ogólnej zależności, jeśli:
$\mathbf{E}[R(\Gamma_\lambda) \mid R(\Gamma_\alpha)] \leqslant \frac{\mathbf{E}[R(\Gamma_\lambda)]}{\mathbf{E}[R(\Gamma_\alpha)]} R(\Gamma_\alpha)$ oraz
$\mathbf{E}[V(\Gamma_\alpha) \mid R(\Gamma_\alpha)] \leqslant \frac{\mathbf{E}[V(\Gamma_\alpha)]}{\mathbf{E}[R(\Gamma_\alpha)]} R(\Gamma_\alpha)$.

2. Jeśli (T_i, H_i) są niezależne, to:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{p.n.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha) \geqslant pFDR(\Gamma_\alpha),$$

gdzie $\pi_1 = 1 - \pi_0$, a $g(\lambda)$ to moc testu T_i dla poziomu istotności λ.
W sytuacji luźnej zależności prawdziwy jest analogiczny wzór:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{P}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} \lim_{m\to\infty} pFDR_m(\Gamma_\alpha).$$

Uwagi odnośnie własności

- własność 1 daje nam gwarancję konserwatywności estymatorów nawet w przypadku zależności, o ile zachodzą pewne specyficzne własności nasuwające na myśl martyngały. Niemniej w praktyce bardzo rzadko się zdarza, aby konserwatywność nie zadziałała
- własność 2 daje niestety tylko złudzenie tego, że w przypadku niezależności bądź słabej zależności i dużych m estymator *pFDR* jest dokładnym, blisko trzymającym się prawdy przybliżeniem, czego przykłady zobaczymy niebawem. Niemniej własność ta może pozwolić sensownie dostroić parametr λ, o ile znamy wykres $g(\lambda)$ (wybieramy wówczas takie λ, aby wartość $\frac{1-g(\lambda)}{1-\lambda}$ była jak najmniejsza)
- poza tym druga własność budzi w nas domysły, że im mniejsze π_0, tym mniej dokładny będzie nasz estymator - co również potwierdzi się w dalszych przykładach.

Uwagi odnośnie własności

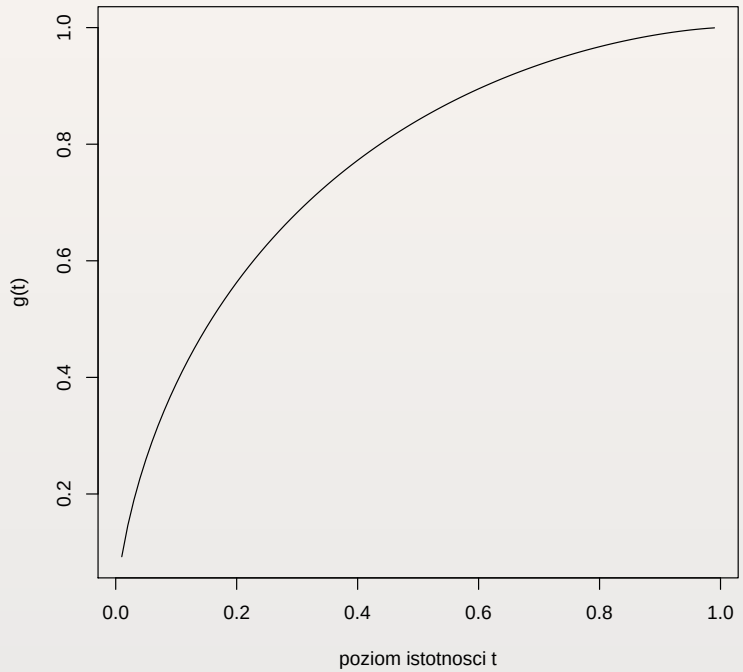
- własność 1 daje nam gwarancję konserwatywności estymatorów nawet w przypadku zależności, o ile zachodzą pewne specyficzne własności nasuwające na myśl martyngały. Niemniej w praktyce bardzo rzadko się zdarza, aby konserwatywność nie zadziałała
- własność 2 daje niestety tylko złudzenie tego, że w przypadku niezależności bądź słabej zależności i dużych m estymator *pFDR* jest dokładnym, blisko trzymającym się prawdy przybliżeniem, czego przykłady zobaczymy niebawem. Niemniej własność ta może pozwolić sensownie dostroić parametr λ, o ile znamy wykres $g(\lambda)$ (wybieramy wówczas takie λ, aby wartość $\frac{1-g(\lambda)}{1-\lambda}$ była jak najmniejsza)
- poza tym druga własność budzi w nas domysły, że im mniejsze π_0, tym mniej dokładny będzie nasz estymator - co również potwierdzi się w dalszych przykładach.

Uwagi odnośnie własności

- własność 1 daje nam gwarancję konserwatywności estymatorów nawet w przypadku zależności, o ile zachodzą pewne specyficzne własności nasuwające na myśl martyngały. Niemniej w praktyce bardzo rzadko się zdarza, aby konserwatywność nie zadziałała
- własność 2 daje niestety tylko złudzenie tego, że w przypadku niezależności bądź słabej zależności i dużych m estymator *pFDR* jest dokładnym, blisko trzymającym się prawdy przybliżeniem, czego przykłady zobaczymy niebawem. Niemniej własność ta może pozwolić sensownie dostroić parametr λ, o ile znamy wykres $g(\lambda)$ (wybieramy wówczas takie λ, aby wartość $\frac{1-g(\lambda)}{1-\lambda}$ była jak najmniejsza)
- poza tym druga własność budzi w nas domysły, że im mniejsze π_0, tym mniej dokładny będzie nasz estymator - co również potwierdzi się w dalszych przykładach.

Przykład $g(\lambda)$

Tak wygląda wykres $g(\lambda)$ w przypadku omawianego eksperymentu:



Spodziewalibyśmy się więc, że im większe λ, tym lepszy dostaniemy estymator w tym eksperymencie.

Przykłady estymacji *FDR*

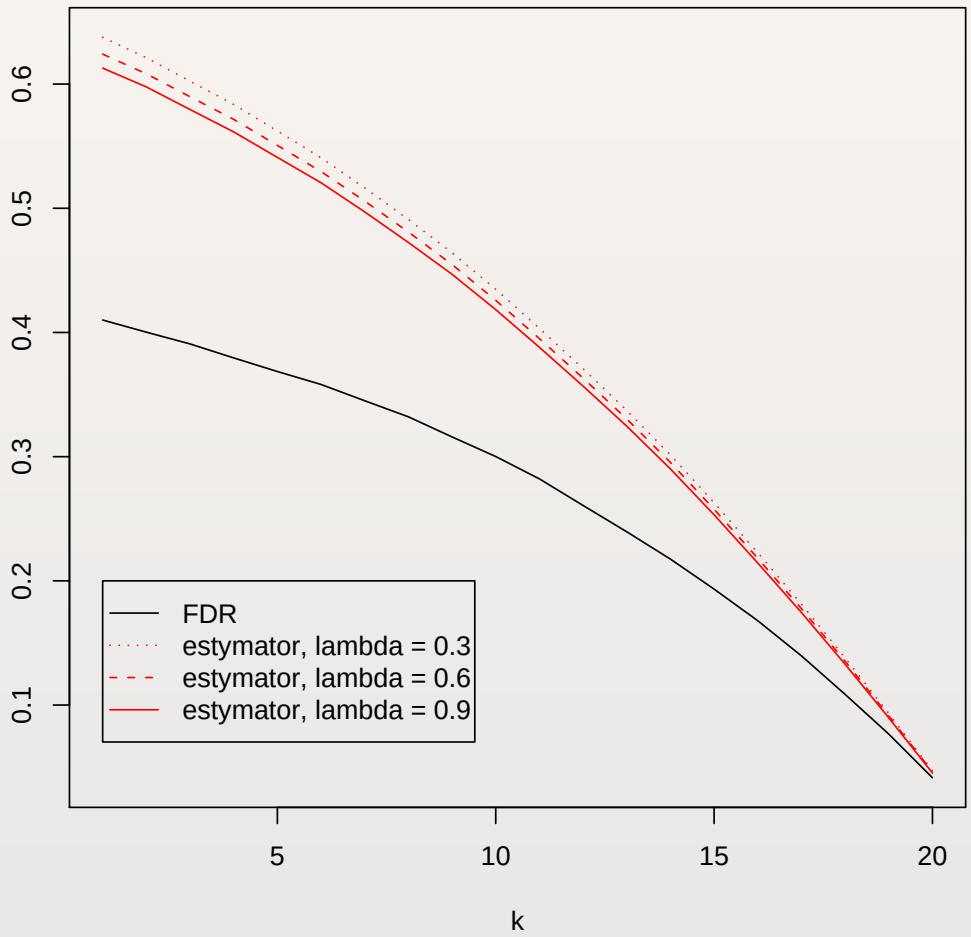
Kontynuując omawiany już eksperyment, postanowiliśmy zweryfikować zachowanie $\widehat{FDR}_\lambda$ gdy testujemy przy różnych poziomach istotności i w sytuacji różnych rodzajów zależności statystyk $(T_1, ..., T_m)$. Podobnie więc jak poprzednio dla ustalonych wartości m, m_0, α, Σ przeprowadziliśmy $it = 10000$ losowań statystyk $(T_1, ..., T_m)$ z rozkładu $N(\mu, \Sigma)$, po uśrednieniu dostając empirycznie wartość FDR w sytuacji gdy testujemy na poziomie istotności α. Skupiliśmy się na 3 rodzajach zależności:

- niezależność, tj. $\Sigma = Id$
- słaba zależność - tu zasymulowaliśmy silne zależności w kolejnych czwórkach statystyk, tzn. $\Sigma = AA^T$, gdzie A jest macierzą $0.9N + 0.1Id$ po unormowaniu wierszami, przy czym N jest macierzą, w której kolejnych wierszach znajdują się kolejno 4 razy e_1, potem 4 razy e_2, ... aż do 4 razy $e_{\frac{m}{4}}$
- ekstremalnie silna zależność - czyli $\Sigma = AA^T$, gdzie A jest macierzą $0.9N + 0.1Id$ po unormowaniu wierszami, przy czym $N[i,] = e_1 \forall i$.

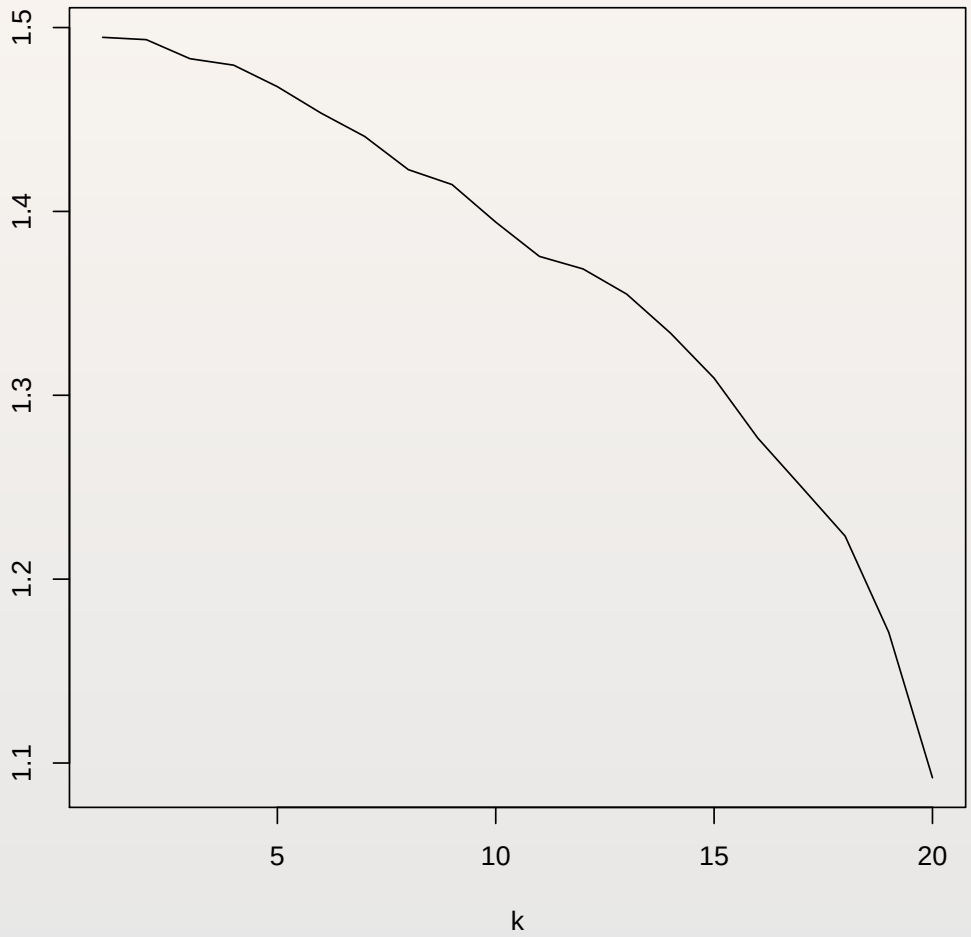
Przykłady estymacji *FDR*

Dla każdego z tych trzech rodzajów zależności przeprowadziliśmy 2 eksperymenty - jeden dla $m = 100$ i $m_0 = 10$, a drugi dla $m = 100$ i $m_0 = 90$. W wyniku każdego takiego eksperymentu otrzymaliśmy dwa wykresy. Na obu wykresach parametr k przebiega od 1 do 20 wyznaczając stosowany w testach poziom istotności $\alpha_k = 0.0005 * (21 - k)$. Dla danego k pierwszy wykres przedstawia odpowiednio: empiryczny $FDR(\Gamma_{\alpha_k})$ oraz empiryczną $\mathbf{E}[\widehat{FDR}_\lambda(\Gamma_{\alpha_k})]$ dla $\lambda = 0.3$, 0.6 oraz 0.9, zaś drugi wykres przedstawia dla poszczególnych k iloraz $\frac{\mathbf{E}[\widehat{FDR}_{0.9}(\Gamma_{\alpha_k})]}{FDR(\Gamma_{\alpha_k})}$.

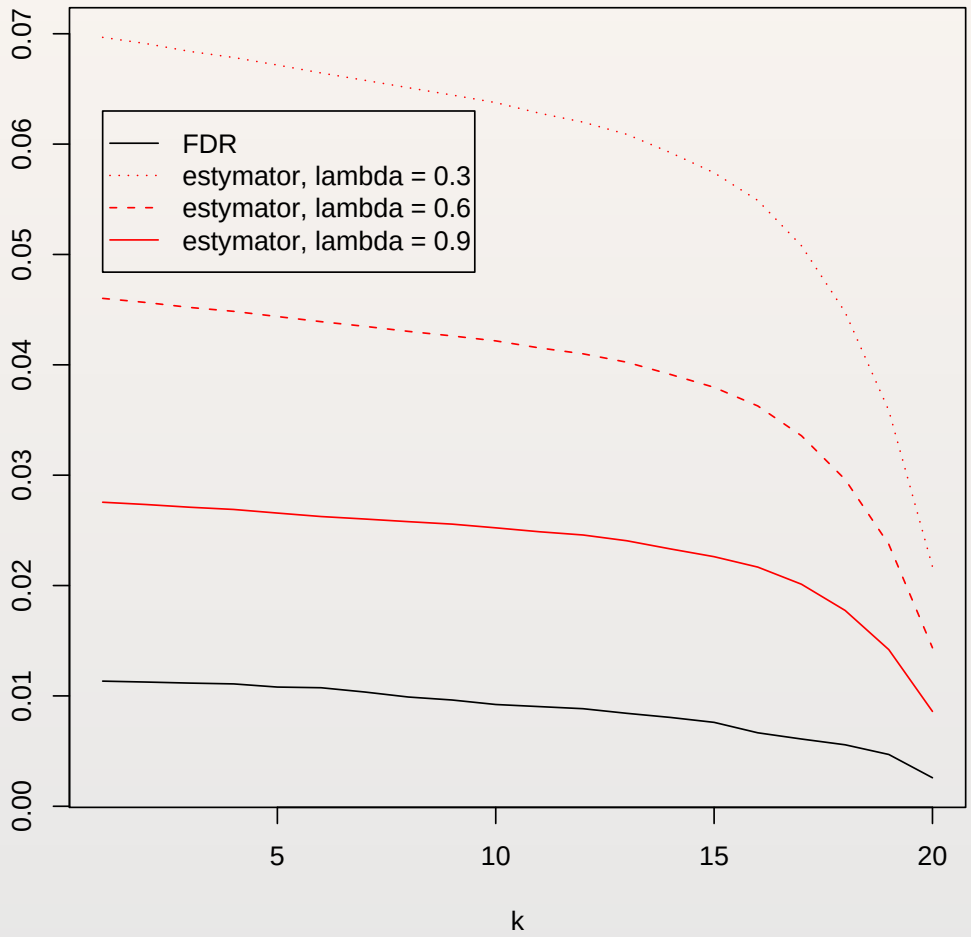
Niezależność, $m = 100$, $m_0 = 90$



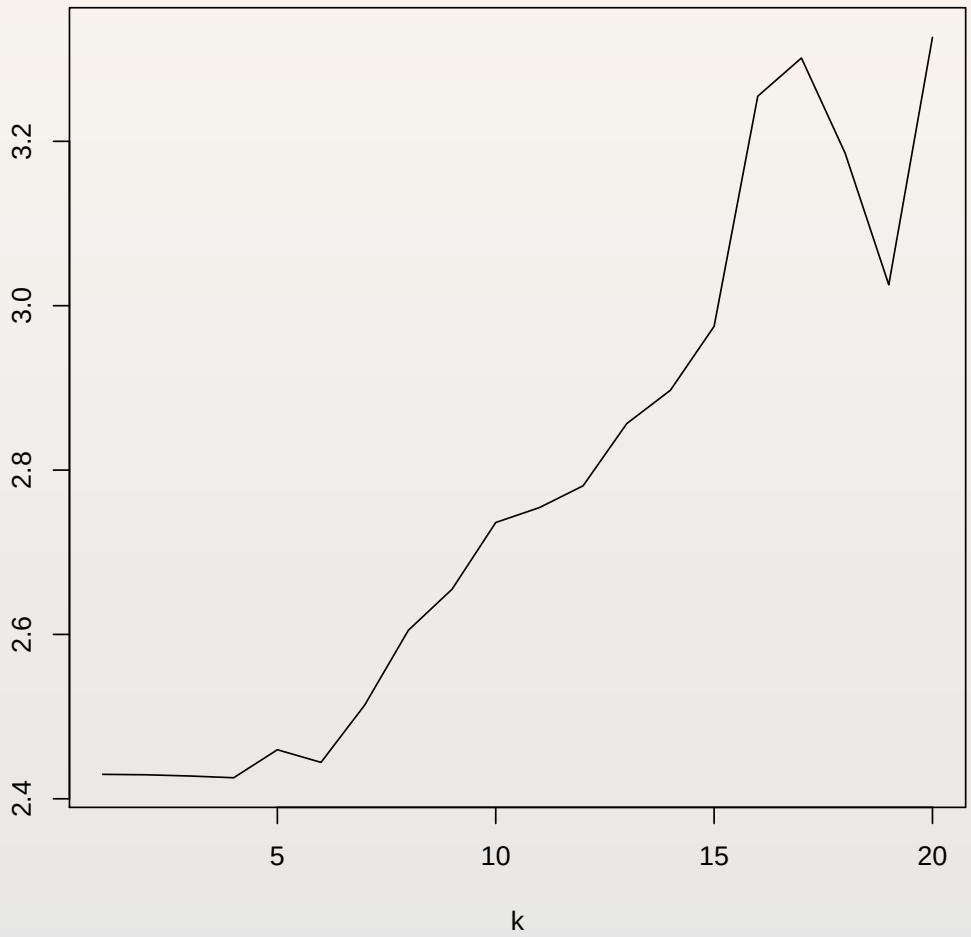
Niezależność, $m = 100$, $m_0 = 90$



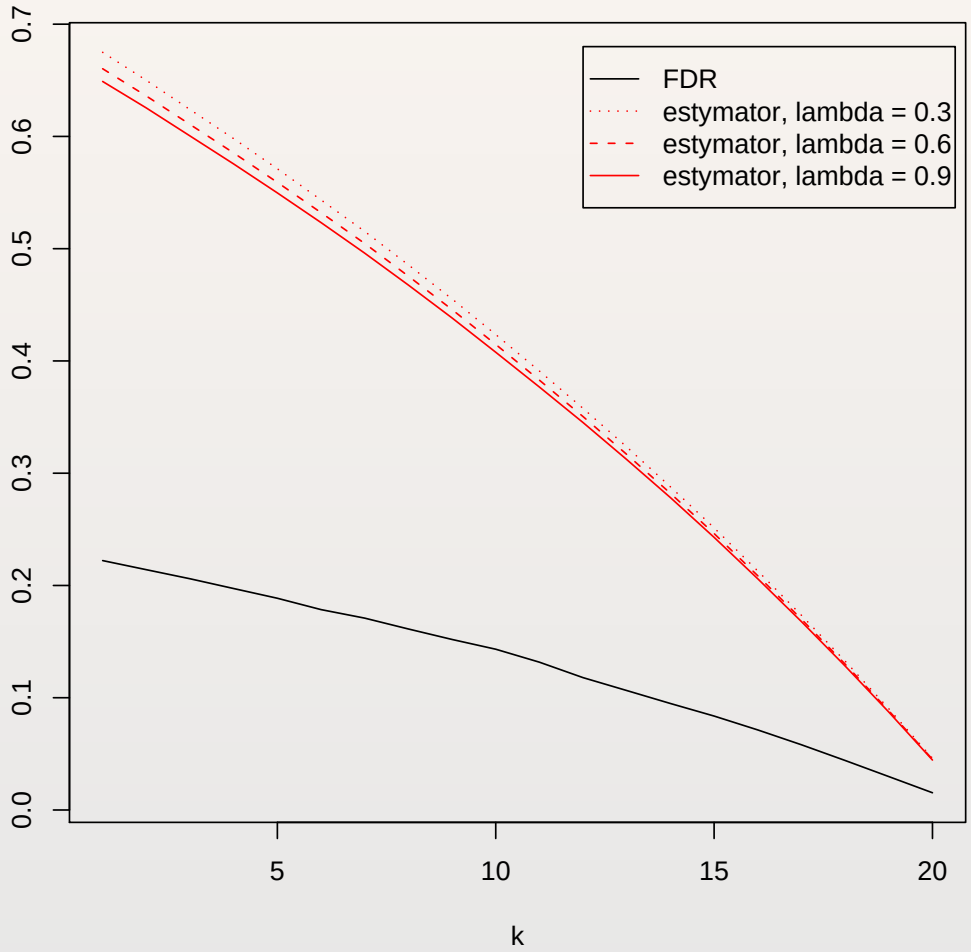
Niezależność, $m = 100$, $m_0 = 10$



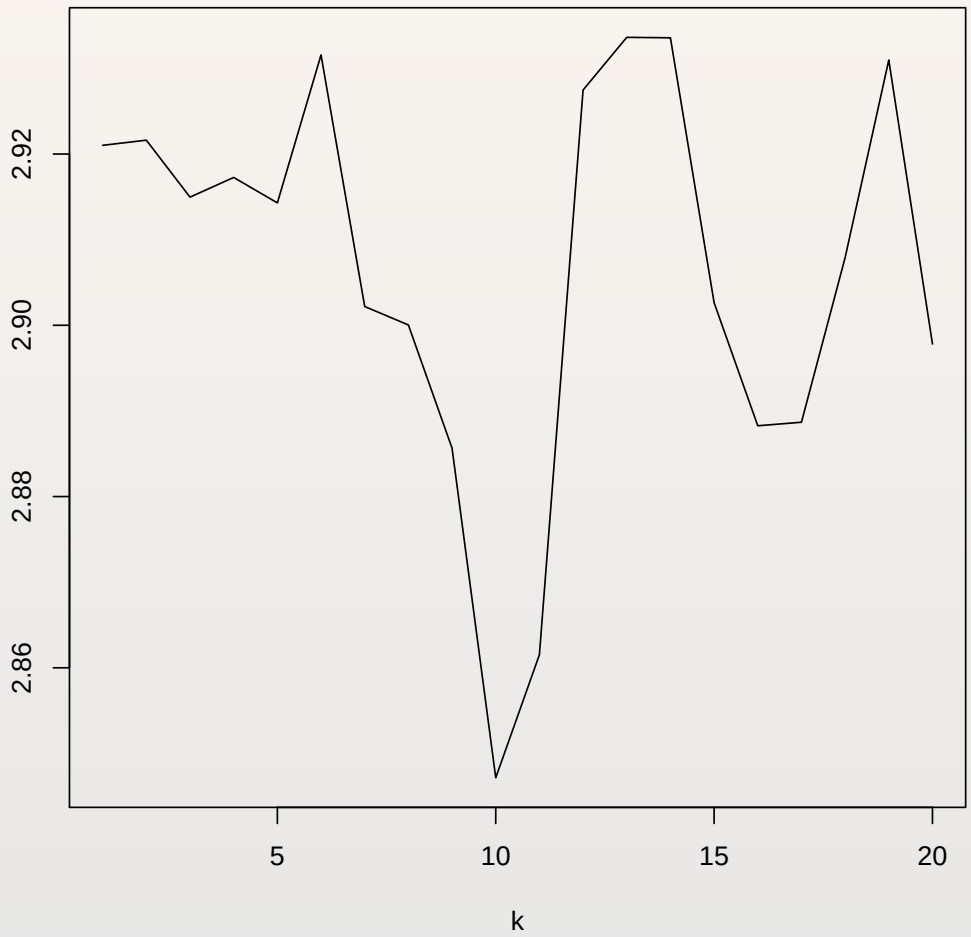
Niezależność, $m = 100$, $m_0 = 10$



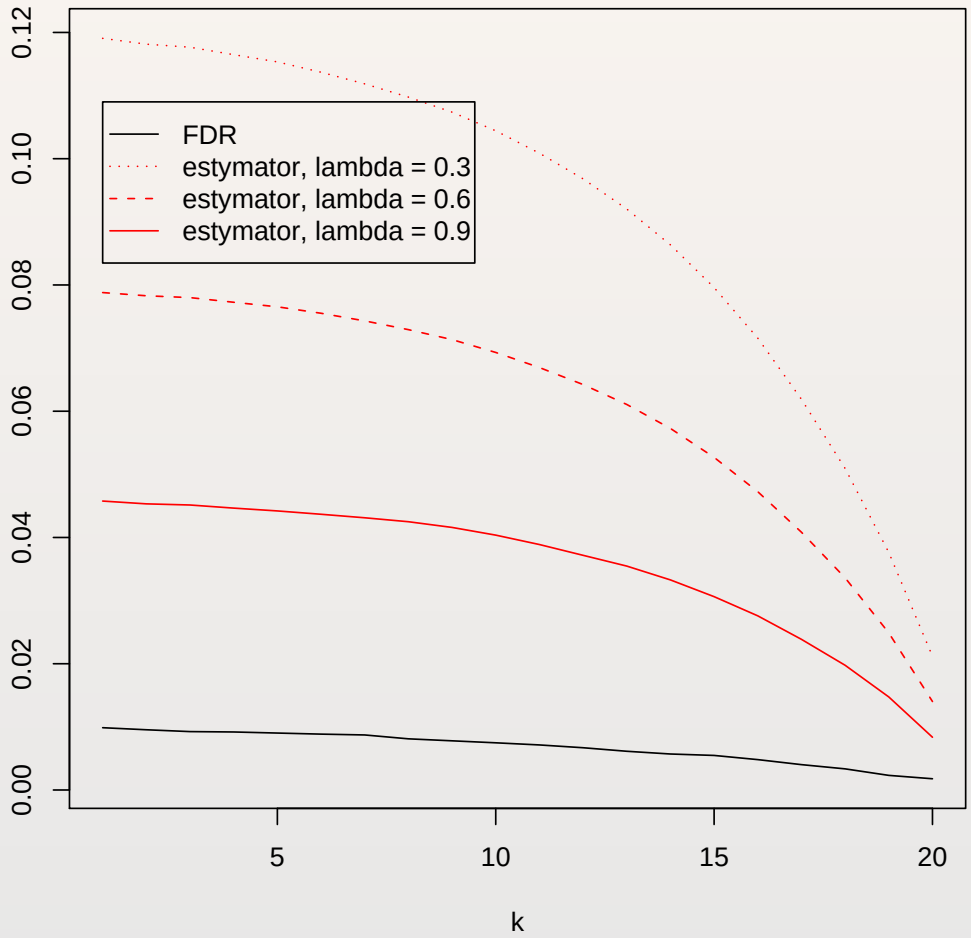
Słaba zależność, $m = 100$, $m_0 = 90$



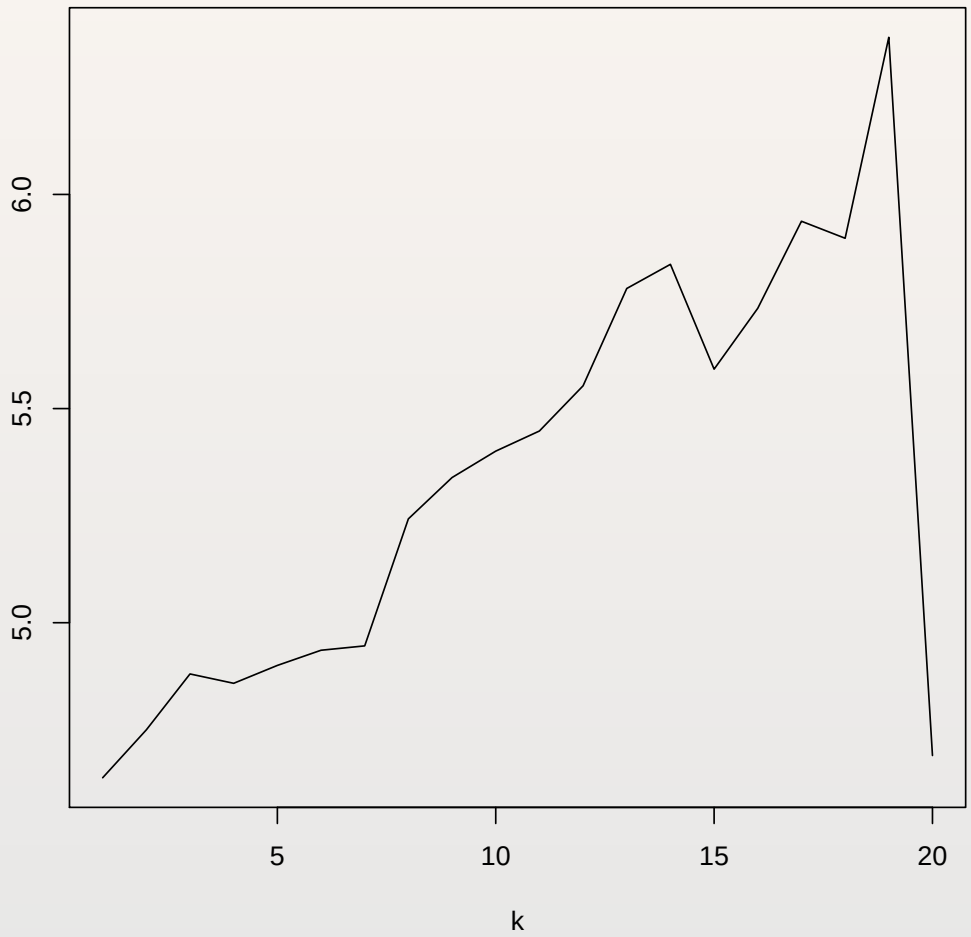
Słaba zależność, $m = 100$, $m_0 = 90$



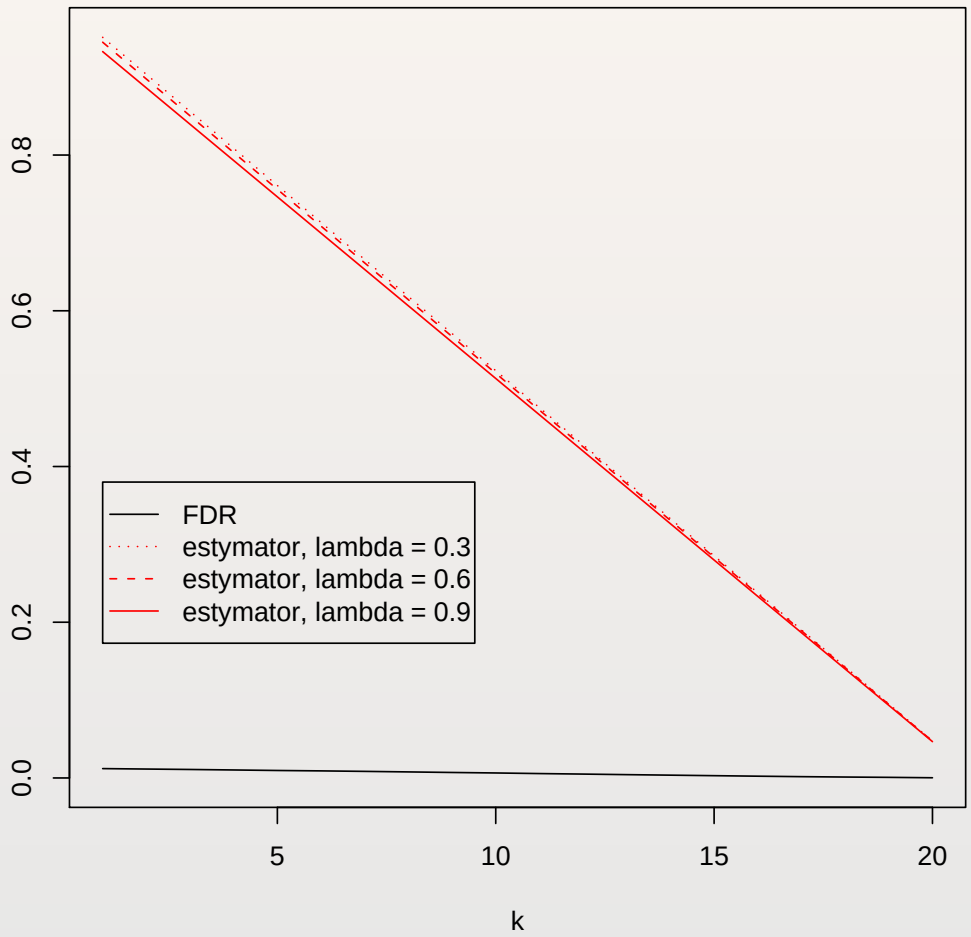
Słaba zależność, $m = 100$, $m_0 = 10$



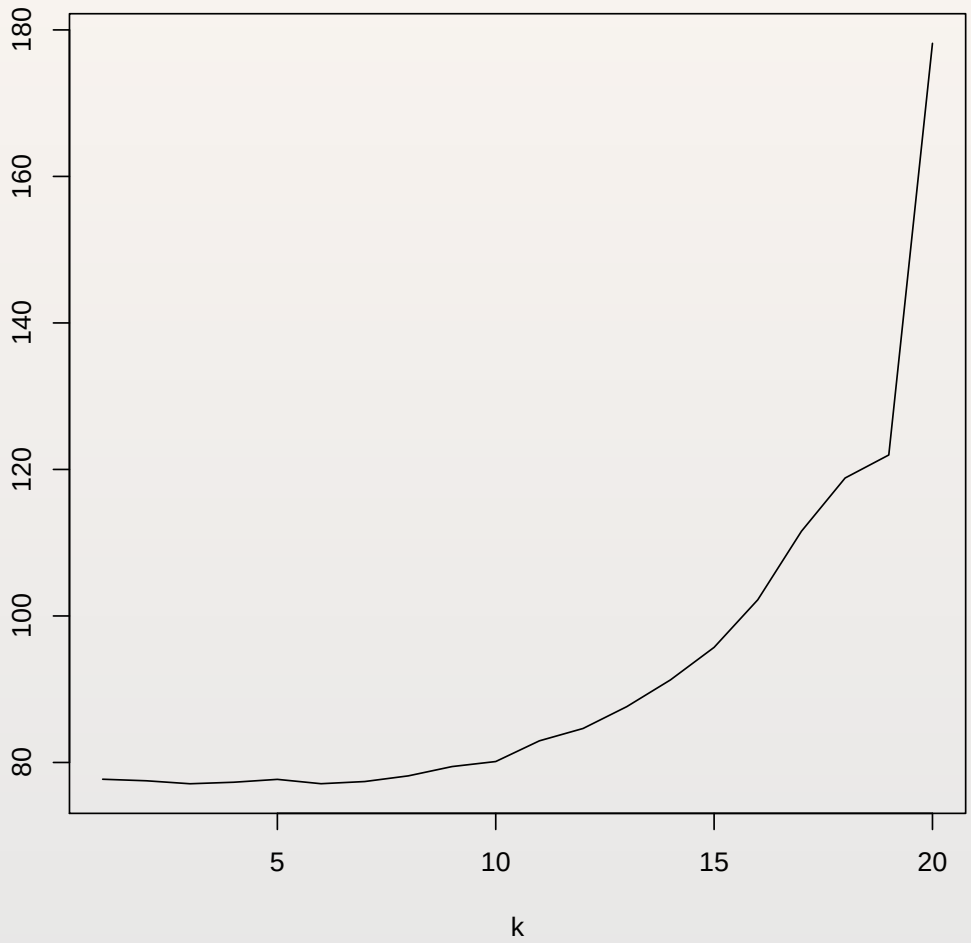
Słaba zależność, $m = 100$, $m_0 = 10$



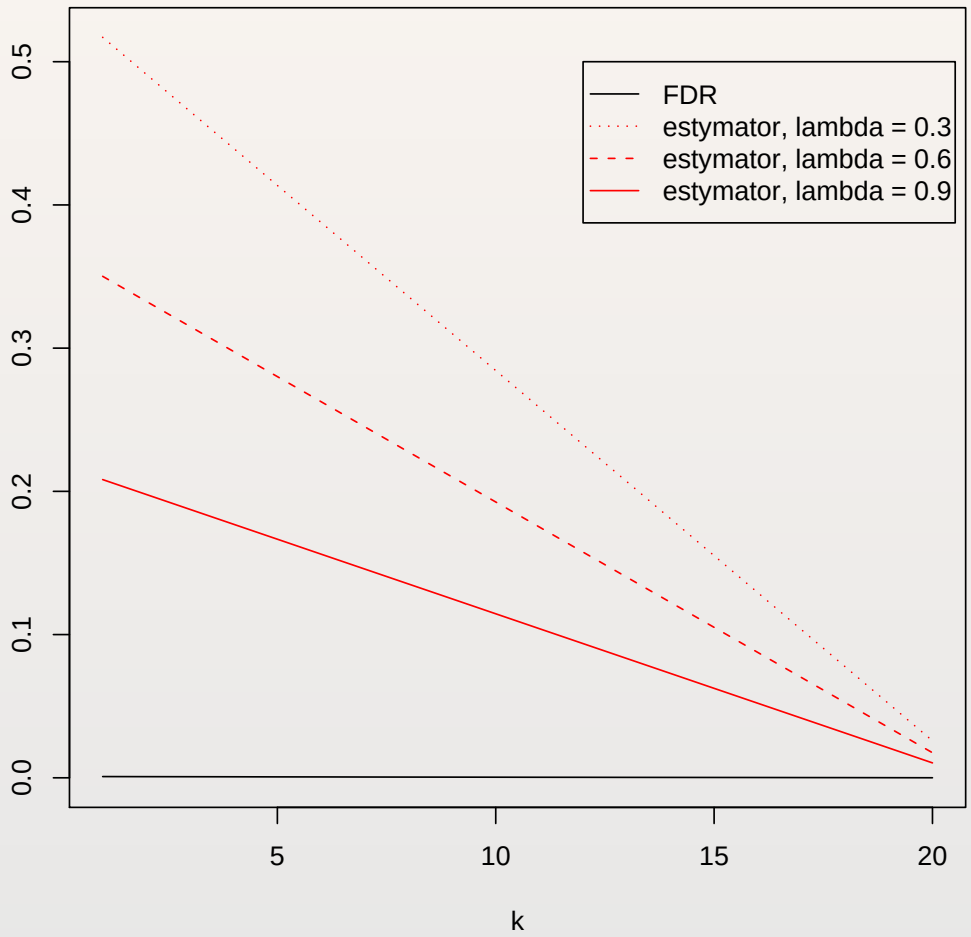
Silna zależność, $m = 100$, $m_0 = 90$



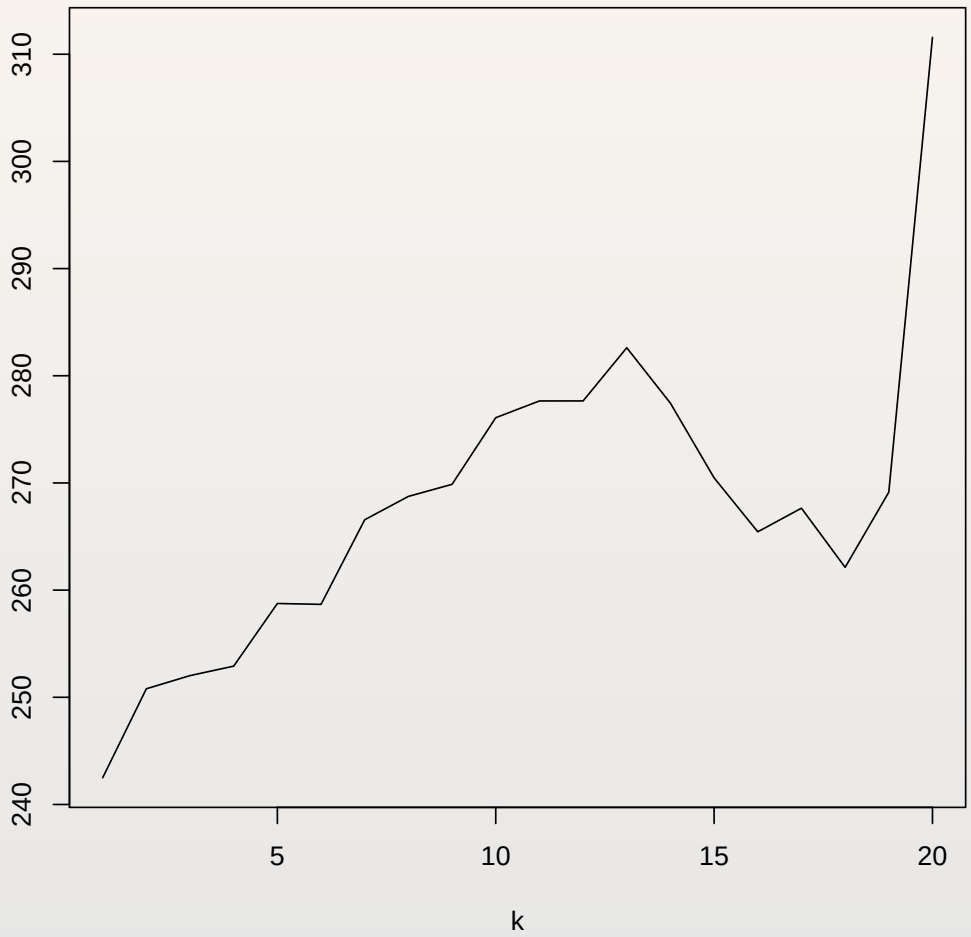
Silna zależność, $m = 100$, $m_0 = 90$



Silna zależność, $m = 100$, $m_0 = 10$



Silna zależność, $m = 100$, $m_0 = 10$



Wnioski

- im większa zależność tym estymacja jest mniej dokładna
- w naszym eksperymencie im większe λ tym dokładność większa (bo funkcja mocy $g(\lambda)$ jest wklęsła)
- potwierdza się konserwatywność estymatora, tzn. że $\mathbf{E}[\widehat{FDR}_\lambda(\Gamma_\alpha)] \geqslant FDR_\lambda(\Gamma_\alpha)$
- zauważmy, że niezależnie od rodzaju zależności, w sytuacji gdy $m = 100$ a $m_0 = 90$, czyli gdy $\pi_0 = 0.9$ estymacja jest dużo dokładniejsza niż wtedy, gdy $m = 100$ a $m_0 = 10$, czyli gdy $\pi_0 = 0.1$. Takie obserwacje są logiczne w kontekście wzoru:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{p.n.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha)$$

działającego dla niezależności i w analogicznej wersji dla słabej zależności. Generalnie im większe π_0, tym większej można się spodziewać dokładności.

Wnioski

- im większa zależność tym estymacja jest mniej dokładna
- w naszym eksperymencie im większe λ tym dokładność większa (bo funkcja mocy $g(\lambda)$ jest wklęsła)
- potwierdza się konserwatywność estymatora, tzn. że $\mathbf{E}[\widehat{FDR}_\lambda(\Gamma_\alpha)] \geqslant FDR_\lambda(\Gamma_\alpha)$
- zauważmy, że niezależnie od rodzaju zależności, w sytuacji gdy $m = 100$ a $m_0 = 90$, czyli gdy $\pi_0 = 0.9$ estymacja jest dużo dokładniejsza niż wtedy, gdy $m = 100$ a $m_0 = 10$, czyli gdy $\pi_0 = 0.1$. Takie obserwacje są logiczne w kontekście wzoru:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{p.n.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha)$$

działającego dla niezależności i w analogicznej wersji dla słabej zależności. Generalnie im większe π_0, tym większej można się spodziewać dokładności.

Wnioski

- im większa zależność tym estymacja jest mniej dokładna
- w naszym eksperymencie im większe λ tym dokładność większa (bo funkcja mocy $g(\lambda)$ jest wklęsła)
- potwierdza się konserwatywność estymatora, tzn. że $\mathbf{E}[\widehat{FDR}_\lambda(\Gamma_\alpha)] \geqslant FDR_\lambda(\Gamma_\alpha)$
- zauważmy, że niezależnie od rodzaju zależności, w sytuacji gdy $m = 100$ a $m_0 = 90$, czyli gdy $\pi_0 = 0.9$ estymacja jest dużo dokładniejsza niż wtedy, gdy $m = 100$ a $m_0 = 10$, czyli gdy $\pi_0 = 0.1$. Takie obserwacje są logiczne w kontekście wzoru:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{p.n.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha)$$

działającego dla niezależności i w analogicznej wersji dla słabej zależności. Generalnie im większe π_0, tym większej można się spodziewać dokładności.

Wnioski

- im większa zależność tym estymacja jest mniej dokładna
- w naszym eksperymencie im większe λ tym dokładność większa (bo funkcja mocy $g(\lambda)$ jest wklęsła)
- potwierdza się konserwatywność estymatora, tzn. że $\mathbf{E}[\widehat{FDR}_\lambda(\Gamma_\alpha)] \geqslant FDR_\lambda(\Gamma_\alpha)$
- zauważmy, że niezależnie od rodzaju zależności, w sytuacji gdy $m = 100$ a $m_0 = 90$, czyli gdy $\pi_0 = 0.9$ estymacja jest dużo dokładniejsza niż wtedy, gdy $m = 100$ a $m_0 = 10$, czyli gdy $\pi_0 = 0.1$. Takie obserwacje są logiczne w kontekście wzoru:

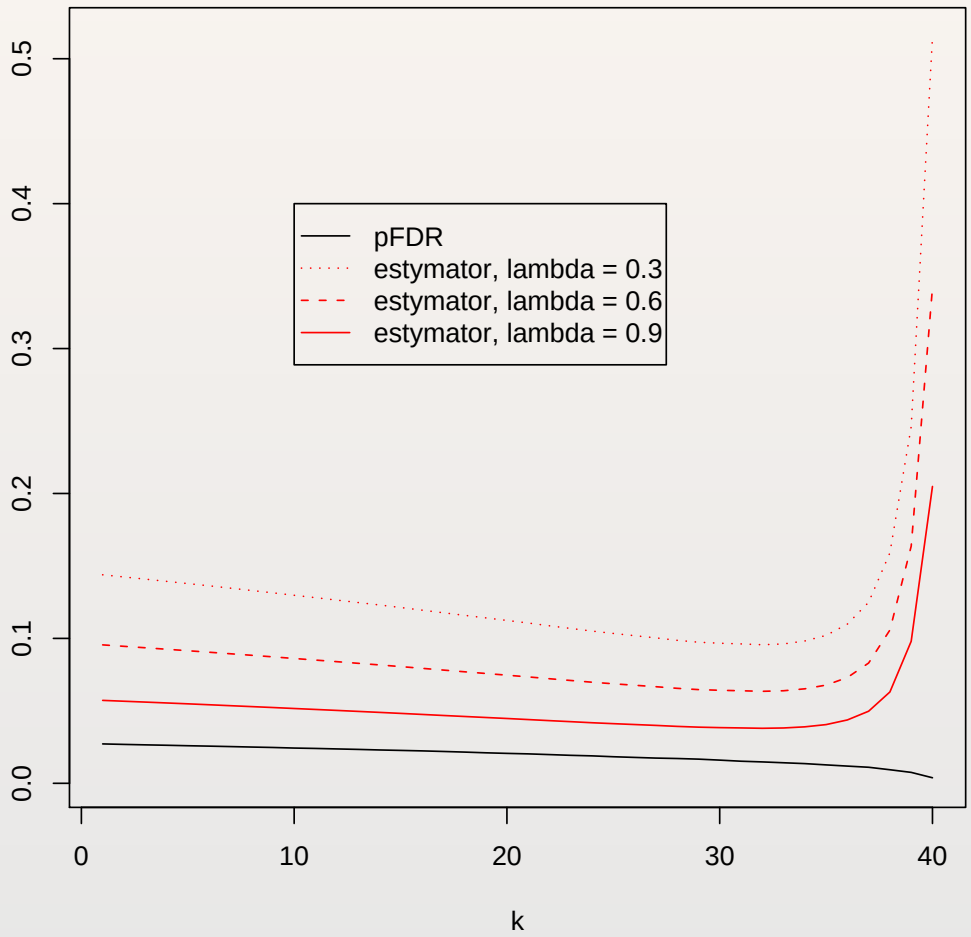
$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{p.n.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha)$$

działającego dla niezależności i w analogicznej wersji dla słabej zależności. Generalnie im większe π_0, tym większej można się spodziewać dokładności.

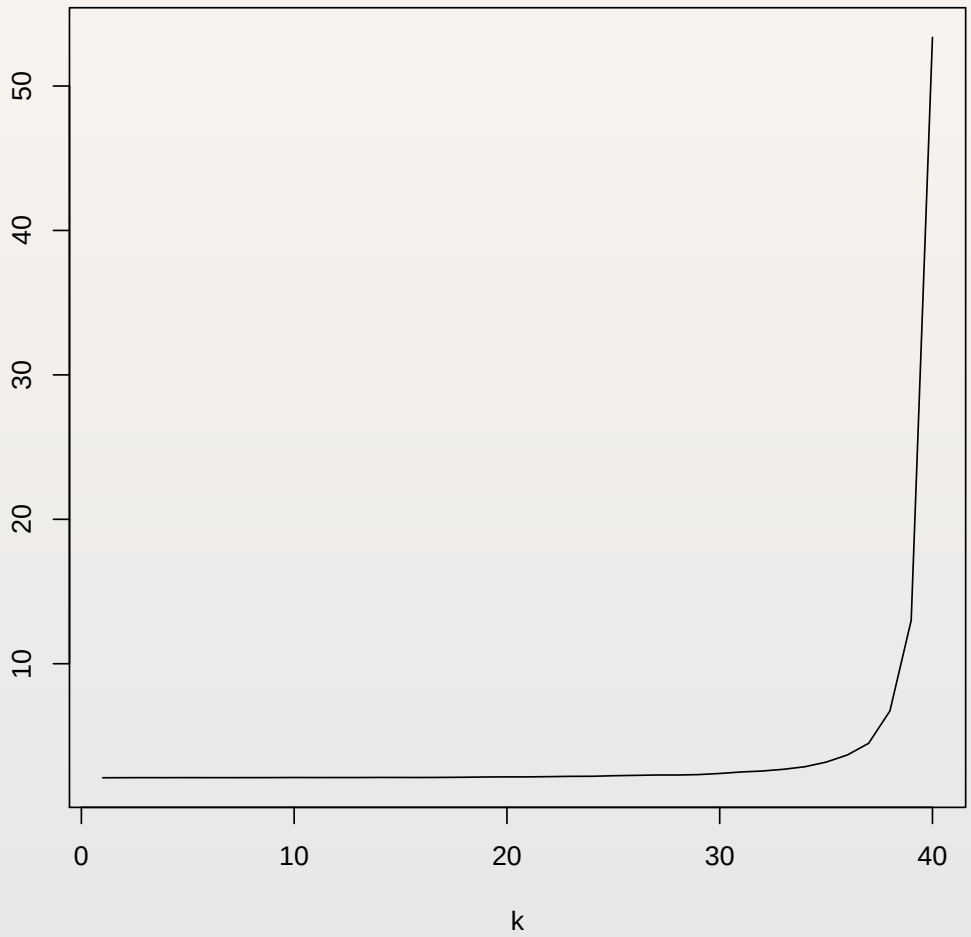
Przykłady estymacji *pFDR*

Znów odwołaliśmy się do naszego bazowego eksperymentu. Przeprowadziliśmy w przypadku niezależności dla $m = 100, m_0 = 10$ oraz $m = 100, m_0 = 90$ analogiczną procedurę co dla estymacji *FDR*, tym razem porównując empiryczny *pFDR* z estymatorem, dokonując $it = 10000$ iteracji. Parametr k przebiega od 1 do 40 zadając poziom istotności $\alpha = 0.0001 + 0.0025 * (40 - k)$.
Potrzebne $\mathbf{Pr}(R^0(\Gamma_{\alpha_k}) > 0)$ dla estymatora *pFDR* otrzymaliśmy symulując m hipotez zerowych w następujący sposób: jeśli X oznacza macierz $it \times m$ rezultatów otrzymanych z powyższych iteracji, to it - krotną symulację m hipotez zerowych dostajemy automatycznie po wycentrowaniu po kolumnach macierzy X.

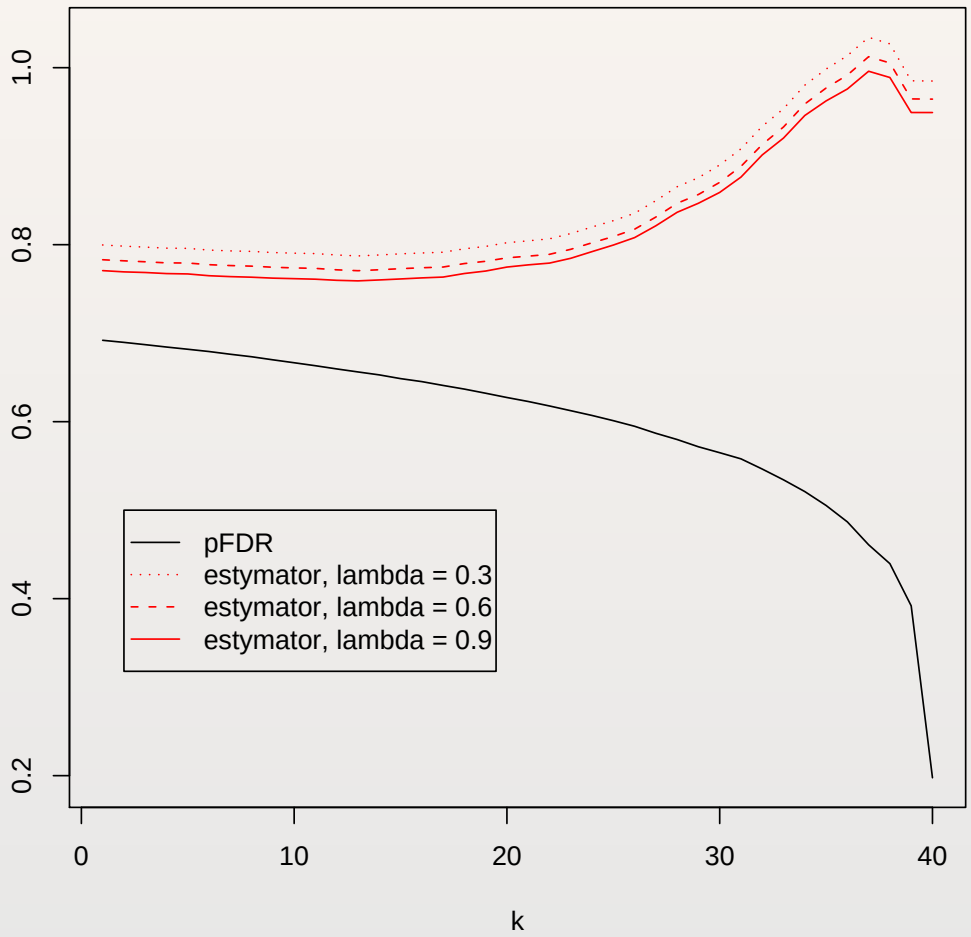
Niezależność, $m = 100$, $m_0 = 10$



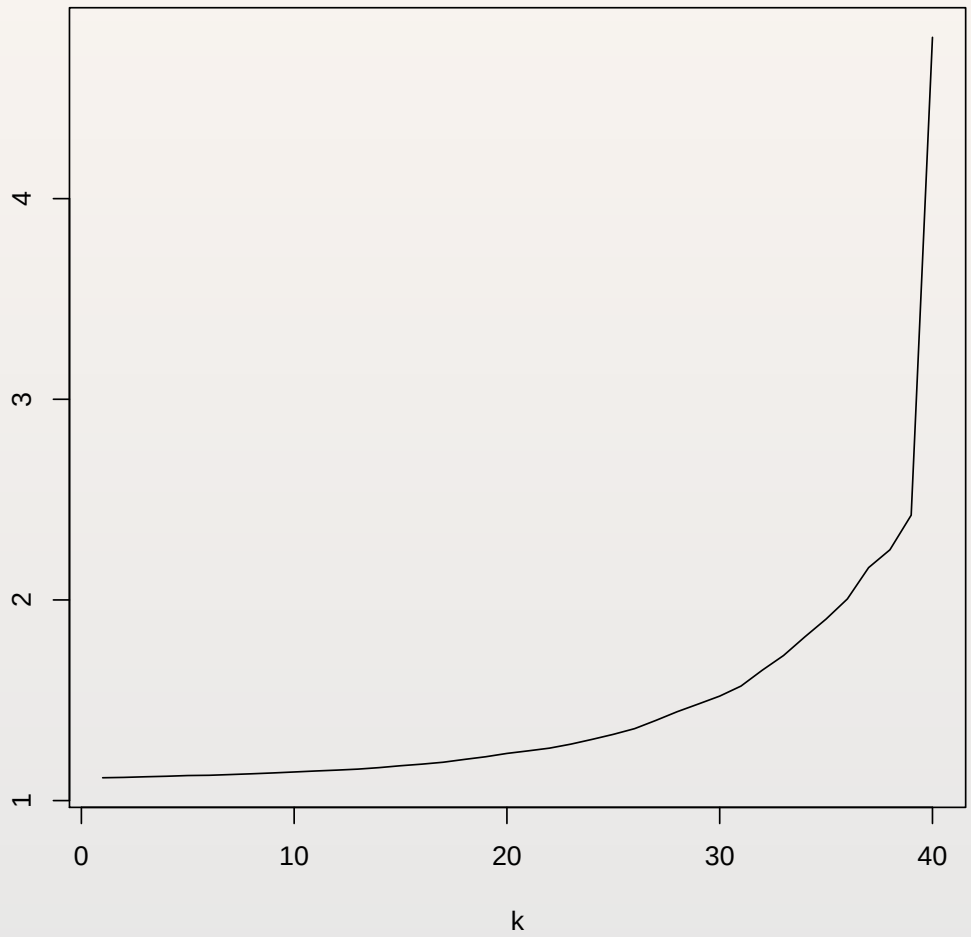
Niezależność, $m = 100$, $m_0 = 10$



Niezależność, $m = 100$, $m_0 = 90$



Niezależność, $m = 100$, $m_0 = 90$



Wnioski

- konserwatywność, wzrost dokładności wraz ze wzrostem π_0
- ale dla α dostatecznie małych estymator zaczyna dziwnie uciekać. Nie jest to jednak przypadek. Przypomnijmy zachodzący w przypadku niezależności wzór:

$$\lim_{m\to\infty} \widehat{pFDR}_\lambda(\Gamma_\alpha) \overset{p.n.}{=} \frac{\pi_0 + \frac{1-g(\lambda)}{1-\lambda}\pi_1}{\pi_0} pFDR(\Gamma_\alpha)$$

który mówi o sytuacji, gdy α noże być nawet bardzo małe ale ustalone, zaś $m \to \infty$. Tu mamy do czynienia z sytuacją odwrotną: m choćby bardzo duże - jest ustalone, zaś $\alpha \to 0$. Okazuje się, że w takiej sytuacji zachodzi:

Fakt

$\lim_{\alpha\to 0} \widehat{FDR}_\lambda(\Gamma_\alpha) = 0$, *a jeśli założyć (co jest sensowne), że* $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0) \geqslant \frac{1}{B}$, *gdzie* B *to ilość iteracji służących do wyestymowania* $\mathbf{Pr}(R^0(\Gamma_\alpha) > 0)$, *to wtedy:* $\lim_{\alpha\to 0} \widehat{pFDR}_\lambda(\Gamma_\alpha) = \widehat{\pi}_0$.

Mikromacierz DNA

Rozpatrzmy macierz X rozmiaru $m \times n$, gdzie pierwsze n_0 kolumn odpowiada przypadkom typu A, zaś pozostałe n_1 kolumn - przypadkom typu B. Wiersz i - ty odpowiada ekspresji i - tego genu dla poszczególnych n przypadków. Naszym zadaniem jest wychwycić te geny, których ekspresja jest istotnie różna w zależności od typu przypadku. Dokonujemy więc dla każdego z m genów testu t - studenta przy ustalonym obszarze krytycznym Γ dla przypadku nieznanych wariancji i hipotezy zerowej: równe średnie, przeciwko alternatywie: nierówne.

Możemy też zastosować nasz estymator $\widehat{pFDR}_\lambda$. Stają przed nami wówczas następujące pytania:

- skąd wziąć sensowne symulacje m hipotez zerowych w celu wyliczenia $\mathbf{Pr}(R^0(\Gamma) > 0)$?
- w jaki sposób wybrać λ ?

Skupimy się na razie na drugim problemie.

Wybór optymalnego λ

- Interesuje nas
  $\lambda_{best} = \arg\min_{\lambda\in[0,1]} \mathbf{E}[(\widehat{pFDR}_\lambda(\Gamma_\alpha) - pFDR(\Gamma_\alpha))^2] =$
  $\arg\min_{\lambda\in[0,1]}[wariancja(\widehat{pFDR}_\lambda(\Gamma_\alpha)) + (obciazenie(\widehat{pFDR}_\lambda(\Gamma_\alpha)))^2].$
- W przypadku naszej mikromacierzy DNA: X spodziewamy się, że kolumny - przypadki są od siebie niezależne, podczas gdy być może silne zależności występują pomiędzy wierszami, czyli ekspresjami poszczególnych genów. W takiej sytuacji sensownym estymatorem wariancji jest tzw. "estymator scyzorykowy":

$$\widehat{var}_\lambda = \frac{\sum_{i=1}^{n}\left(\widehat{pFDR}_\lambda^{(-i)}(\Gamma_\alpha) - \widehat{pFDR}_\lambda(\Gamma_\alpha)\right)^2}{n}$$

gdzie $\widehat{pFDR}_\lambda^{(-i)}(\Gamma_\alpha)$ to wyestymowany $pFDR$ na bazie naszej tabeli X z wyrzuconą i - tą kolumną.

Wybór optymalnego λ

- Interesuje nas
  $\lambda_{best} = \arg\min_{\lambda\in[0,1]} \mathbf{E}[(\widehat{pFDR}_\lambda(\Gamma_\alpha) - pFDR(\Gamma_\alpha))^2] =$
  $\arg\min_{\lambda\in[0,1]}[wariancja(\widehat{pFDR}_\lambda(\Gamma_\alpha)) + (obciazenie(\widehat{pFDR}_\lambda(\Gamma_\alpha)))^2]$.
- W przypadku naszej mikromacierzy DNA: X spodziewamy się, że kolumny - przypadki są od siebie niezależne, podczas gdy być może silne zależności występują pomiędzy wierszami, czyli ekspresjami poszczególnych genów. W takiej sytuacji sensownym estymatorem wariancji jest tzw. "estymator scyzorykowy":

$$\widehat{var}_\lambda = \frac{\sum_{i=1}^{n}\left(\widehat{pFDR}_\lambda^{(-i)}(\Gamma_\alpha) - \widehat{pFDR}_\lambda(\Gamma_\alpha)\right)^2}{n}$$

gdzie $\widehat{pFDR}_\lambda^{(-i)}(\Gamma_\alpha)$ to wyestymowany *pFDR* na bazie naszej tabeli X z wyrzuconą i - tą kolumną.

Wybór optymalnego λ

- Dla wyestymowania obciążenia, wobec braku wiedzy ile wynosi rzeczywiste $pFDR(\Gamma_\alpha)$, musimy je czymś zastąpić. Ponieważ nasz estymator spełnia $\min_{\lambda'} \mathbf{E}[\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)] \geqslant pFDR(\Gamma_\alpha)$, to najsensowniej zastąpić $pFDR(\Gamma_\alpha)$ przez $\min_{\lambda' \in [0,1]}(\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)$. Dostajemy więc estymator obciążenia w kwadracie:

$$\widehat{bias}_\lambda^2 = \left(\widehat{pFDR}_\lambda(\Gamma_\alpha) - \min_{\lambda' \in [0,1]}(\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)\right)^2$$

- następnie dla uwspólnienia skali tworzymy:
$\widehat{bias}_\lambda^{2*} = \frac{\widehat{bias}_\lambda^2}{median_{\lambda' \in [0,1]}(\widehat{bias}_{\lambda'}^2)}$, $\widehat{var}_\lambda^* = \frac{\widehat{var}_\lambda}{median_{\lambda' \in [0,1]}(\widehat{var}_{\lambda'})}$
- oznaczając $\widehat{MSE}(\lambda) = \widehat{bias}_\lambda^{2*} + \widehat{var}_\lambda^*$, zadajemy estymator optymalnego λ jako: $\widehat{\lambda} = \arg\min_{\lambda \in [0,1]} \widehat{MSE}(\lambda)$, który w praktyce możemy wyznaczyć zastępując przedział $[0, 1]$ siatką, np. $\{0, 0.05, 0.10, ..., 0.95\}$.

Wybór optymalnego λ

- Dla wyestymowania obciążenia, wobec braku wiedzy ile wynosi rzeczywiste $pFDR(\Gamma_\alpha)$, musimy je czymś zastąpić. Ponieważ nasz estymator spełnia $\min_{\lambda'} \mathbf{E}[\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)] \geqslant pFDR(\Gamma_\alpha)$, to najsensowniej zastąpić $pFDR(\Gamma_\alpha)$ przez $\min_{\lambda' \in [0,1]}(\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)$. Dostajemy więc estymator obciążenia w kwadracie:

$$\widehat{bias}_\lambda^2 = \left(\widehat{pFDR}_\lambda(\Gamma_\alpha) - \min_{\lambda' \in [0,1]}(\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)\right)^2$$

- następnie dla uwspólnienia skali tworzymy:
$\widehat{bias}_\lambda^{2*} = \frac{\widehat{bias}_\lambda^2}{median_{\lambda' \in [0,1]}(\widehat{bias}_{\lambda'}^2)}$, $\widehat{var}_\lambda^* = \frac{\widehat{var}_\lambda}{median_{\lambda' \in [0,1]}(\widehat{var}_{\lambda'})}$
- oznaczając $\widehat{MSE}(\lambda) = \widehat{bias}_\lambda^{2*} + \widehat{var}_\lambda^*$, zadajemy estymator optymalnego λ jako: $\widehat{\lambda} = \arg\min_{\lambda \in [0,1]} \widehat{MSE}(\lambda)$, który w praktyce możemy wyznaczyć zastępując przedział $[0, 1]$ siatką, np. $\{0, 0.05, 0.10, ..., 0.95\}$.

Wybór optymalnego λ

- Dla wyestymowania obciążenia, wobec braku wiedzy ile wynosi rzeczywiste $pFDR(\Gamma_\alpha)$, musimy je czymś zastąpić. Ponieważ nasz estymator spełnia $\min_{\lambda'} \mathbf{E}[\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)] \geqslant pFDR(\Gamma_\alpha)$, to najsensowniej zastąpić $pFDR(\Gamma_\alpha)$ przez $\min_{\lambda' \in [0,1]}(\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)$. Dostajemy więc estymator obciążenia w kwadracie:

$$\widehat{bias}_\lambda^2 = \left(\widehat{pFDR}_\lambda(\Gamma_\alpha) - \min_{\lambda' \in [0,1]}(\widehat{pFDR}_{\lambda'}(\Gamma_\alpha)\right)^2$$

- następnie dla uwspólnienia skali tworzymy:
$\widehat{bias}_\lambda^{2*} = \frac{\widehat{bias}_\lambda^2}{median_{\lambda' \in [0,1]}(\widehat{bias}_{\lambda'}^2)}$, $\widehat{var}_\lambda^* = \frac{\widehat{var}_\lambda}{median_{\lambda' \in [0,1]}(\widehat{var}_{\lambda'})}$
- oznaczając $\widehat{MSE}(\lambda) = \widehat{bias}_\lambda^{2*} + \widehat{var}_\lambda^*$, zadajemy estymator optymalnego λ jako: $\widehat{\lambda} = \arg\min_{\lambda \in [0,1]} \widehat{MSE}(\lambda)$, który w praktyce możemy wyznaczyć zastępując przedział $[0, 1]$ siatką, np. $\{0, 0.05, 0.10, ..., 0.95\}$.

Definicja

p-wartość

minimalny poziom istotności, przy którym zaobserwowana wartość t statystyki testowej T spowoduje podjęcie decyzji o odrzuceniu H_0

Wprowadzamy indeksowaną parametrem α rodzinę obszarów krytycznych (obszarów odrzuceń) Γ_α taką, że:

$$\mathbb{P}(T \in \Gamma_\alpha | H_0 \textit{ jest prawdziwa}) = \alpha$$

Wtedy *p-wartość* definiujemy następująco:

$$p(t) := \inf_{\{\alpha : t \in \Gamma_\alpha\}} \mathbb{P}(T \in \Gamma_\alpha | H_0 \textit{ jest prawdziwa})$$

Im mniejsza jest *p-wartość*, tym mocniejsze staje się przekonanie testującego o fałszywości hipotezy zerowej, a prawdziwości hipotezy alternatywnej.

Definicja

p-wartość

minimalny poziom istotności, przy którym zaobserwowana wartość t statystyki testowej T spowoduje podjęcie decyzji o odrzuceniu H_0

Wprowadzamy indeksowaną parametrem α rodzinę obszarów krytycznych (obszarów odrzuceń) Γ_α taką, że:

$$\mathbb{P}(T \in \Gamma_\alpha | H_0 \textit{ jest prawdziwa}) = \alpha$$

Wtedy *p-wartość* definiujemy następująco:

$$p(t) := \inf_{\{\alpha : t \in \Gamma_\alpha\}} \mathbb{P}(T \in \Gamma_\alpha | H_0 \textit{ jest prawdziwa})$$

Im mniejsza jest *p-wartość*, tym mocniejsze staje się przekonanie testującego o fałszywości hipotezy zerowej, a prawdziwości hipotezy alternatywnej.

Definicja

p-wartość

minimalny poziom istotności, przy którym zaobserwowana wartość t statystyki testowej T spowoduje podjęcie decyzji o odrzuceniu H_0

Wprowadzamy indeksowaną parametrem α rodzinę obszarów krytycznych (obszarów odrzuceń) Γ_α taką, że:

$$\mathbb{P}(T \in \Gamma_\alpha | H_0 \textit{ jest prawdziwa}) = \alpha$$

Wtedy *p-wartość* definiujemy następująco:

$$p(t) := \inf_{\{\alpha : t \in \Gamma_\alpha\}} \mathbb{P}(T \in \Gamma_\alpha | H_0 \textit{ jest prawdziwa})$$

Im mniejsza jest *p-wartość*, tym mocniejsze staje się przekonanie testującego o fałszywości hipotezy zerowej, a prawdziwości hipotezy alternatywnej.

Definicja

q-wartość

Dla ustalonej wartości t statystyki testowej T *q-wartość* definiujemy następująco:

$$q\text{-}wartość(t) = \inf_{\{\alpha: t \in \Gamma_\alpha\}} pFDR(\Gamma_\alpha)$$

Z powyższej definicji wynika, że *q-wartość* to najmniejsza wartość miary pFDR, przy której zobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej H_0.

Czasem *q-wartość* definuje się jako:

$$q\text{-}wartość(t) = \inf_{\{\alpha: t \in \Gamma_\alpha\}} \mathbb{P}(H = 0 | T \in \Gamma_\alpha)$$

Definicja ta jest równoważna poprzedniej w sytuacji, gdy statystyki testowe są niezależne, a zbliżona dla dużych m i słabej zależności.

Definicja

q-wartość

Dla ustalonej wartości t statystyki testowej T *q-wartość* definiujemy następująco:

$$q\text{-}wartość(t) = \inf_{\{\alpha: t \in \Gamma_\alpha\}} pFDR(\Gamma_\alpha)$$

Z powyższej definicji wynika, że *q-wartość* to najmniejsza wartość miary pFDR, przy której zobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej H_0.

Czasem *q-wartość* definuje się jako:

$$q\text{-}wartość(t) = \inf_{\{\alpha: t \in \Gamma_\alpha\}} \mathbb{P}(H = 0 | T \in \Gamma_\alpha)$$

Definicja ta jest równoważna poprzedniej w sytuacji, gdy statystyki testowe są niezależne, a zbliżona dla dużych m i słabej zależności.

Definicja

q-wartość

Dla ustalonej wartości t statystyki testowej T *q-wartość* definiujemy następująco:

$$q\text{-}wartość(t) = \inf_{\{\alpha: t \in \Gamma_\alpha\}} pFDR(\Gamma_\alpha)$$

Z powyższej definicji wynika, że *q-wartość* to najmniejsza wartość miary pFDR, przy której zobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej H_0.

Czasem *q-wartość* definuje się jako:

$$q\text{-}wartość(t) = \inf_{\{\alpha: t \in \Gamma_\alpha\}} \mathbb{P}(H = 0 | T \in \Gamma_\alpha)$$

Definicja ta jest równoważna poprzedniej w sytuacji, gdy statystyki testowe są niezależne, a zbliżona dla dużych m i słabej zależności.

Definicja cd.

m hipotez
m statystyk testowych odpowiadających każdej hipotezie: $T_1, ..., T_m$
m *p-wartości* odpowiadających każdej hipotezie: $p_1, ..., p_m$
p-wartości porządkujemy w kolejności niemalejącej:

$$p_{(1)} \leqslant ... \leqslant p_{(m)}$$

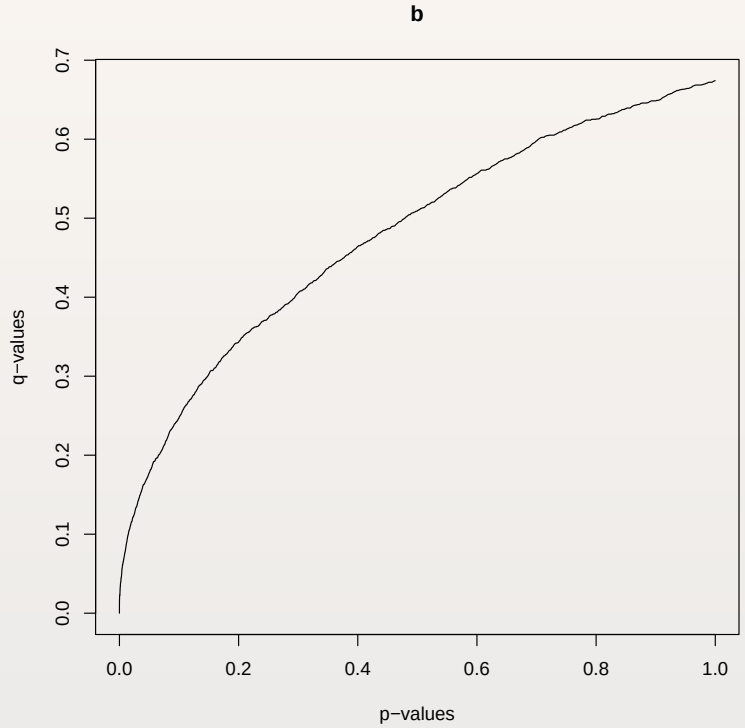
Wtedy, ponieważ dla dużych m $FDR \sim pFDR$, określamy:

$$q(p_{(i)}) = \min_{\{t \geqslant p_{(i)}\}} FDR(t)$$

co przy znajomości $\widehat{FDR}(t)$ możemy zapisać jako:

$$\hat{q}(p_{(i)}) = \min_{\{t \geqslant p_{(i)}\}} \widehat{FDR}(t)$$

Wykres zależności *q-wartości* od *p-wartości*



Uwaga!

$\hat{q}$ jest funkcją niemalejącą ze względu na niemalejące wartości $p_{(i)}$

Algorytm estymacji q-wartości

1. Porządkujemy *p-wartości* w kolejności niemalejącej:

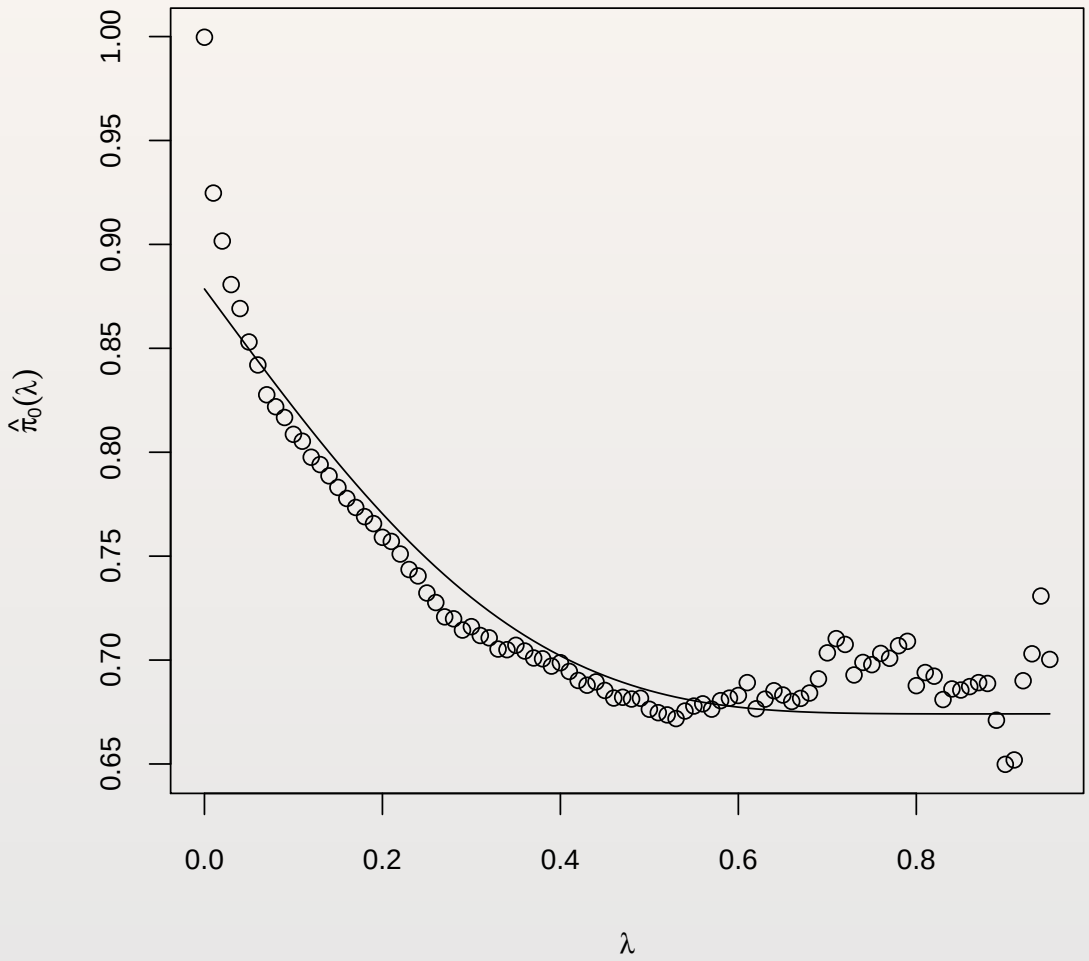
$$p_{(1)} \leqslant \ldots \leqslant p_{(m)}$$

2. Dla wartości parametru $\lambda = 0,01; 0,02; \ldots; 0,99$ obliczamy wartość:

$$\hat{\pi}_0(\lambda) = \frac{\#(p_{(i)} > \lambda)}{m(1-\lambda)}$$

i rysujemy wykres zależności $\hat{\pi}_0$ od λ:

Wykres zależności $\hat{\pi}_0$ od λ



Algorytm estymacji q-wartości cd.

3. Używając interpolacji wielomianami trzeciego stopnia (cubic splane) szukam $\hat{f}$ - najlepszego przybliżenia funkcji $\hat{\pi}_0$ zależnej od λ
4. Przyjmujemy, że: $\hat{\pi}_0 = \hat{f}(1)$
5. Liczymy:

$$\hat{q}(p_{(m)}) = \min_{t \geqslant p_{(m)}} \frac{\hat{\pi}_0 \cdot m \cdot t}{\#(j : p_j \leqslant t)} = \hat{\pi}_0 \cdot p_{(m)}$$

6. Dla $i = m-1, m-2, ..., 1$ liczymy:

$$\hat{q}(p_{(i)}) = \min_{t \geqslant p_{(i)}} \frac{\hat{\pi}_0 \cdot m \cdot t}{\#(j : p_j \leqslant t)} = \min\left(\frac{\hat{\pi}_0 \cdot m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right)$$

7. Wyestymowaną *q-wartością* dla $i-$ej hipotezy będzie $\hat{q}(p_{(i)})$.

Przykład

Mamy dwie grupy pacjentów:

Pierwsza grupa

Chorzy na nowotwór typu A
Liczebność n_1=7

Druga grupa

Chorzy na nowotwór typu B
Liczebność n_2=8

Cel

Chcemy znaleźć jak największą liczbę genów o istotnie różnych poziomach aktywności (ekspresji) *(oczywiście używając FDR i związane z nim q-wartości)*

Będziemy korzystać z testu $t - Studenta$ dla równoważnych dwóch populacji o niejednorodnych wariancjach.
Testujemy istotność różnic w poziomach ekspresji dla każdego spośród $m = 3170$ genów.

X_i-poziom ekspresji i-tego genu w pierwszej grupie pacjentów
Y_i-poziom ekspresji i-tego genu w drugiej grupie pacjentów
$i \in \{1, ..., 3170\}$

Niech:

$$X_i \sim N(m_{i1}, \sigma_{i1})$$

$$Y_i \sim N(m_{i2}, \sigma_{i2})$$

Testujemy m hipotez:

$$H_0^i : m_{i1} - m_{i2} = 0$$

wobec

$$H_1^i : m_{i1} - m_{i2} \neq 0$$

Transformujemy wartości tych poziomów ekspresji za pomocą funkcji $log_2(\cdot)$i stosujemy statystykę testową.

Statystyka testowa ma postać:

$$T_i = \frac{\overline{x}_{i2} - \overline{x}_{i1}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}},$$

gdzie:
$\bar{x}_{i1}, \bar{x}_{i2}$-próbkowa średnia odpowiednio dla pierwszej i drugiej populacji
s_{i1}^2, s_{i2}^2-próbkowa wariancja odpowiednio dla pierwszej i drugiej populacji

Szacowanie *p-wartości*

Przy założeniu prawdziwości hipotezy zerowej rozkład statystyki T_i nie jest do końca jasny, dlatego odpowiednie *p-wartości* oszacujemy za pomocą metody permutacyjnej.

W naszym przypaku $p_i = \mathbb{P}(|T_i| \geqslant |t_i| \mid H_i = 0)$, więc przydałyby się nam symulacje hipotez zerowych, w celu empirycznego wyliczenia *p-wartości*.

Jeśli X oznacza macierz, której kolejne wiersze dla $i \in \{1, ..., 3170\}$ to (X_i, Y_i), to wówczas możemy zasymulować $m = 3170$ hipotez zerowych poprzez permutację kolumn tej macierzy (zachowując strukturę zależności). Wykonując takie permutacje B razy, otrzymamy mB przykładów hipotezy zerowej.

Szacowanie *p-wartości*

Przy założeniu prawdziwości hipotezy zerowej rozkład statystyki T_i nie jest do końca jasny, dlatego odpowiednie *p-wartości* oszacujemy za pomocą metody permutacyjnej.

W naszym przypaku $p_i = \mathbb{P}(|T_i| \geqslant |t_i| \mid H_i = 0)$, więc przydałyby się nam symulacje hipotez zerowych, w celu empirycznego wyliczenia *p-wartości*.

Jeśli X oznacza macierz, której kolejne wiersze dla $i \in \{1, ..., 3170\}$ to (X_i, Y_i), to wówczas możemy zasymulować $m = 3170$ hipotez zerowych poprzez permutację kolumn tej macierzy (zachowując strukturę zależności). Wykonując takie permutacje B razy, otrzymamy mB przykładów hipotezy zerowej.

Szacowanie *p-wartości*

Przy założeniu prawdziwości hipotezy zerowej rozkład statystyki T_i nie jest do końca jasny, dlatego odpowiednie *p-wartości* oszacujemy za pomocą metody permutacyjnej.

W naszym przypaku $p_i = \mathbb{P}(|T_i| \geqslant |t_i| \mid H_i = 0)$, więc przydałyby się nam symulacje hipotez zerowych, w celu empirycznego wyliczenia *p-wartości*.

Jeśli X oznacza macierz, której kolejne wiersze dla $i \in \{1, ..., 3170\}$ to (X_i, Y_i), to wówczas możemy zasymulować $m = 3170$ hipotez zerowych poprzez permutację kolumn tej macierzy (zachowując strukturę zależności). Wykonując takie permutacje B razy, otrzymamy mB przykładów hipotezy zerowej.

Szacowanie *p-wartości*

Metoda permutacyjna

Wykonujemy $B = 100$ permutacji macierzy X otrzymując wyniki testów $(t_1^b, ..., t_m^b)$, dla $b = 1, ..., B$. Wtedy odpowiednia *p-wartość* jest dana wzorem:

$$p_i = \sum_{b=1}^{B} \frac{\#\{j : |t_j^b| \geqslant |t_i|,\ j = 1, 2, ..., m\}}{mB}$$

Następnie w celu oceny istotności różnic poziomów ekspresji badanych genów możemy zastosować algorytm *q-wartości* dla tak jak powyżej powstałych *p-wartości*.

Wykorzystanie *q-wartości* do odrzucania hipotez

Gdy już mamy wyliczone *q-wartości* $q_{(1)}, ..., q_{(m)}$, to okazuje się, że w sytuacji słabej zależnosci statystyk testowych i dużej liczby m tych statystyk, odrzucanie hipotez dla których *q-wartość* jest $\leqslant \alpha$ zapewnia nam, że $FDR \leqslant \alpha$.

Zatem gdy np. przeprowadziliśmy całą procedurę dla $\alpha = 0.05$, to jeśli 160 genów zostało zakwalifikowanych do odrzucenia, możemy się spodziewać 8 niesłusznie odrzuconych.

Wykorzystanie *q-wartości* do odrzucania hipotez

Gdy już mamy wyliczone *q-wartości* $q_{(1)}, ..., q_{(m)}$, to okazuje się, że w sytuacji słabej zależnosci statystyk testowych i dużej liczby m tych statystyk, odrzucanie hipotez dla których *q-wartość* jest $\leqslant \alpha$ zapewnia nam, że $FDR \leqslant \alpha$.

Zatem gdy np. przeprowadziliśmy całą procedurę dla $\alpha = 0.05$, to jeśli 160 genów zostało zakwalifikowanych do odrzucenia, możemy się spodziewać 8 niesłusznie odrzuconych.

Pakiet *qvalue*

1. `>qvalue(p=NULL,lambda=seq(0,0.90,0.05), fdr.level=NULL, robust = FALSE)`

   Argumenty:

   - `p` - wektor p-wartości (jedyny wymagany argument);
   - `lambda` - siatka punktów do estymacji π_0
   - `pi0.method` - metoda używana do estymacji π_0, przyjmuje
   - `fdr.level` - poziom, na którym kontrolujemy FDR
   - `robust` - dla małej ilości p-wartości używa do estymacji q-wartości pFDR (robust=TRUE)

   Co nam zwraca?

   - `$pi0` - wyestymowana wartość $\hat{\pi}_0$
   - `$qvalues` - wektor q-wartości
   - `$pvalues` - wektor p-wartości
   - `$significant` - jeśli `fdr.level` jest podany, to pokazuje,które q-wartości leżą poniżej `fdr.level`

Przykład:

```
>library(qvalue)
>data(hedenfalk)
>qwart<-qvalue(hedenfalk)
```

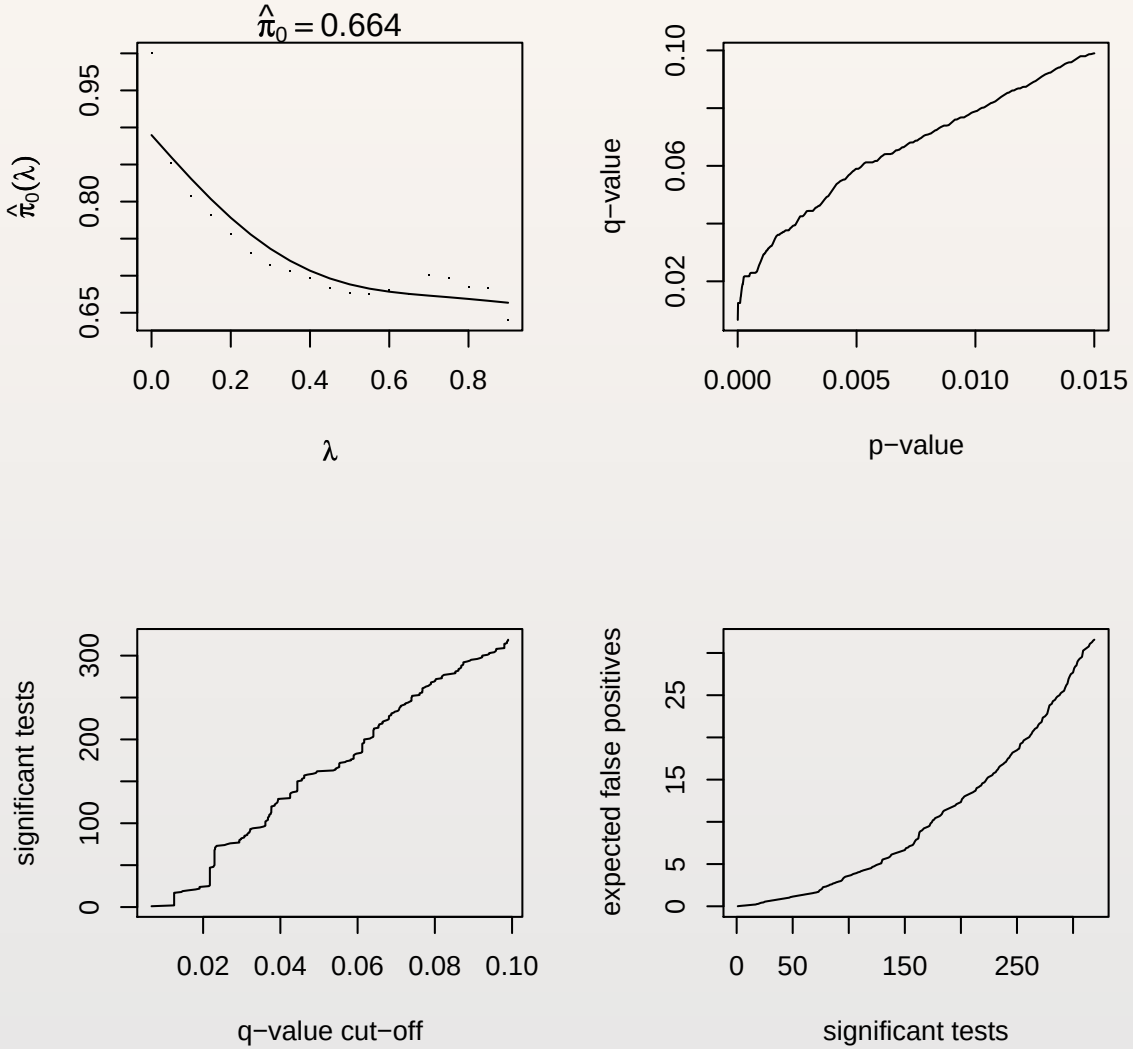
2. `>qsummary`

```
>qsummary(qwart)
```

Call: qvalue(p = hedenfalk)
pi0: 0.6635185
Cumulative number of significant calls:

	<1e-04	**<0.001**	**<0.01**	**<0.025**	**<0.05**	**<0.1**	**<1**
p-value	15	76	265	424	605	868	3170
q-value	0	0	1	73	162	319	3170

3. >qplot


$\hat{\pi}_0 = 0.664$
$\hat{\pi}_0(\lambda)$
0.65
0.80
0.95
0.0
0.2
0.4
0.6
0.8
λ
q−value
0.02
0.06
0.10
0.000
0.005
0.010
0.015
p−value
significant tests
0
100
200
300
0.02
0.04
0.06
0.08
0.10
q−value cut−off
expected false positives
0
5
15
25
0
50
150
250
significant tests

4. `>qwrite(qwart, filename="wyniki.txt")` - zwraca plik,w którym będziemy mieli zapisaną wartość $\hat{\pi}_0$, wektor p-wartości i odpowiadające im q-wartości

Bibliografia

- "Estimating FDR under dependence" John D. Storey, Robert Tibshirani
- "Statistical significance for genomewide studies" John D. Storey, Robert Tibshirani