Statystyczna Analiza Danych – laboratorium

Wielokrotne testowanie hipotez

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 5

31 marca 2022/1 kwietnia 2022

Idea zajęć – co i po co będziemy robić?

- Hasłem przewodnim dzisiejszych zajęć są “pułapki w stosowaniu testów statystycznych”
- W szczególności zajmiemy się problemem wielokrotnego testowania hipotez za pomocą indywidualnych testów
- Prezentowane dziś techniki mają praktyczne zastosowanie np. w bioinformatyce

Zadanie 4

Dla wskazanych w scenariuszu przykładów pytań badawczych wybrać zalecane metody korekcji.

Ogólne zasady stosowania metod korekcji w Z4

- **Bonferroni lub Holm**
  - wybór bardzo małego zbioru genów, które z dużym prawdopodobieństwem mają inną ekspresję w różnych rodzajach nowotworu
  - odpowiedź na pytanie, czy istnieje jakikolwiek gen który rozróżnia rodzaje nowotworu. Zwłaszcza jeśli fałszywe przyjęcie że gen odpowiada za nowotwór jest bardzo kosztowne
- **Benjamini-Hochberg**
  - zbadanie, które geny najprawdopodobniej rozróżniają rodzaje nowotworu (możemy dopuścić pewną liczbę fałszywych odkryć)
- **Brak korekcji**
  - w ogólności nie powinno się stosować
  - znalezienie jak największej liczby genów, które potencjalnie mogą się różnić ekspresją pomiędzy nowotworami (dopuszczamy że wśród kilku tysięcy znalezionych genów jedynie kilkanaście będzie rzeczywiście rozróżniać nowotwory)

Zadanie 5

Stwórz macierz o wymiarach 10x1000, taką, że obserwacje z pierwszych stu kolumn są wylosowane z rozkładu $\mathcal{N}(1,1)$, a z pozostałych 900 z rozkładu $\mathcal{N}(0,1)$. Następnie:

- Dla każdej kolumn przeprowadź test t Studenta, czy średnia jest większa od 0 (zapisz otrzymane p-wartości w wektorze)
- Przeprowadź korekcje p-wartości za pomocą metod Bonferroniego, Holma oraz Bejnamini-Hochberga, korzystając z funkcji `p.adjust`
- Obejrzyj rozkłady p-wartości oraz q-wartości (skorygowanych p-wartości) na histogramach
- Zapisz oryginalne p-wartości i trzy wektory q-wartości w ramce danych lub macierzy o 4 kolumnach.
- Dla każdej kolumny oblicz moc testu, False Discovery Rate, False Positive Rate, przyjmując poziom istotności 0.05. Zinterpretuj wyniki.