Statystyczna analiza Danych

Dla bioinformatyków

Wykład pierwszy:

O testowaniu hipotez

Plan na dziś

- Quiz!
- Cele wykładu
- Plan na semestr
- Kryteria zaliczenia
- Sprawy organizacyjne

Quiz (15 minut)

- Jakie znasz rozkłady prawdopodobieństwa?
- Co to jest wartość oczekiwana zmiennej losowej?
- Jakie jest prawdopodobieństwo (w przybliżeniu) wylosowania liczby >=2 z rozkładu normalnego N(0,1)?
- Co to jest hipoteza statystyczna i jakie błędy można popełnić przy jej testowaniu?

Cele wykładu

- Podstawowe pojęcia i intuicje statystyczne
- Praktyczne metody statystyczne w zadaniach bioinformatycznych
- Praktyczna nauka środowiska Gnu R do analiz statystycznych
- Umiejętność przetwarzania i wizualizacji danych
- Umiejętność wyboru metod statystycznych do pytania badawczego

Plan semestru

- Testy statystyczne
- Testowanie wielu hipotez
- Analiza Wariancji
- Korelacja i regresja
- Analiza dyskryminacyjna (LDA, QDA)
- Analiza skupień
- Modele mieszane

Plan semestru (c.d.)

- Zagadnienie klasyfikacji, metoda kNN
- Składowe główne i kanoniczne korelacje
- Klasyfikatory drzewowe
- Mieszanki klasyfikatorów
- Klasyfikacja Bayesowska
- Metody jądrowe (np. SVM, KCCA)

Kryteria zaliczenia

- Ćwiczenia/Laboratorium – 2 projekty:
  - Mały projekt (testowanie hipotez): 10 pkt
  - Duży projekt (klasyfikacja): 20 pkt, prezentacja na zajęciach
- Wykład – egz. pisemny (komputerowy) – 30 pkt
- Punkty sumują się – razem do zdobycia 60 pkt
- Do zaliczenia potrzeba 60% (36 punktów)

Sprawy organizacyjne

- Kontakt:

  bartek@mimuw.edu.pl

  - Pokój 5770, najlepiej po umówieniu się,
  - Konsultacje Pon, 8.30-10.
- Strona przedmiotu:

  http://bioputer.mimuw.edu.pl/~bartek/SAD

Literatura

- Wykłady – brak oficjalnego podręcznika, slajdy **nie są** notatkami z wykładu
- Metody - http://statweb.stanford.edu/~tibs/ElemStatLearn
- R
  - www.biecek.pl/R/
  - http://www.biecek.pl/NaPrzelajPrzezDataMining/

Statystyka klasyczna i Bayesowska

Thomas Bayes

(1701-1762)

Ronald A. Fisher

(1890-1962)

Przykład “z życia”

Ronald Fisher użył w swojej książce takiegoprzykładu, opartego na ponoć faktycznym eksperymenie przeprowadzonym na Muriel Bristol, która twierdziła, że potrafi rozpoznać, czy do filiżanki najpierw wlano herbatę czy najpierw mleko.

	Twierdzi, że najpierw herbata	Twierdzi, że najpierw mleko
Najpierw herbata	3	1
Najpierw mleko	1	3

Test statystyczny

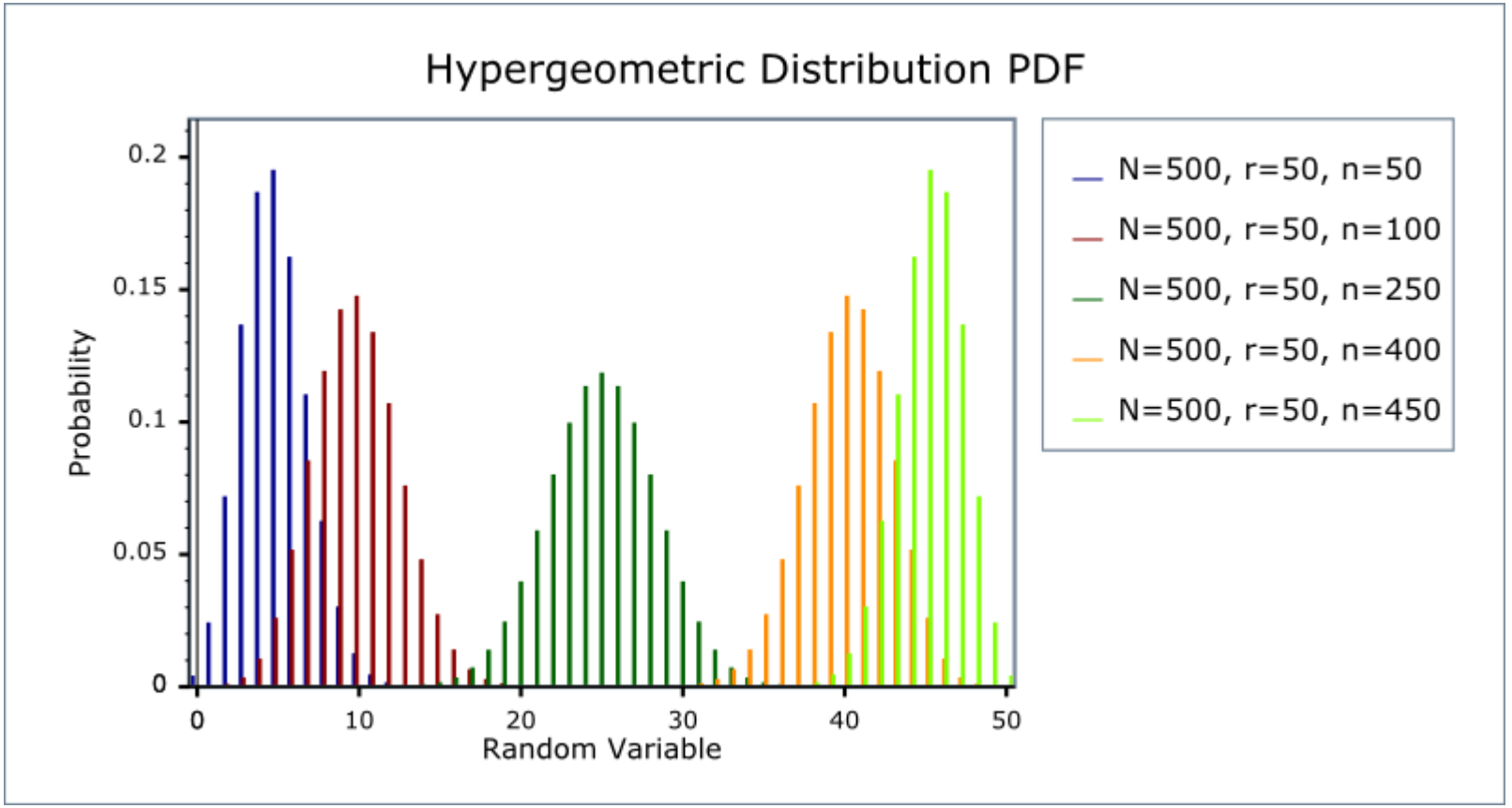
- W naszym przypadku interesuje nas, czy p. Bristol w istocie potrafi przewidzieć lepiej odpowiedź na nasze pytanie niż losowy rzut monetą.
- W tym celu definiujemy **hipotezę zerową H_0**, która mówi, że jej predykcje są losowe
- Będziemy próbowali **odrzucić** tę hipotezę, w oparciu o dane z naszej tabelki liczebności
- Kluczowe pytanie to: **jakie jest prawdopodobieństwo** uzyskania danych takich jak zaobserwowano lub "lepszych" **przy założeniu H_0**

Poziom istotności

- Dla testu statystycznego, będziemy ustalali poziom istotności, czyli prawdopodobieństwo poniżej którego będziemy odrzucać H_0
- Błędem I rodzaju (false-positive) będziemy określać sytuację, w której odrzucimy H_0, która była w rzeczywistości spełniona
- Błędem II rodzaju (false-negative) będziemy nazywali sytuację odwrotną, czyli przyjęcie H_0, które w istocie nie jest spełnione.

Rozkład hipergeometryczny

$$P(X = x) = \frac{\binom{a}{x}\binom{n-a}{r-x}}{\binom{n}{r}}$$



Test Fisher'a

- Dla zadanej tablicy liczebności (contingency table), prawdopodobieństwo osiągnięcia zadanej liczebności przy założeniu hiopotezy zerowej jest zadane rozkładem hipergeometrycznym:

$$P(X = x) = \frac{\binom{a}{x}\binom{n-a}{r-x}}{\binom{n}{r}}$$

- Nas interesuje P(X>=x)

Przykłady bardziej bioinformatyczne

- Spośród 20 000 genów, 5000 podlega ekspresji w tkance A, a 7000 w tkance B. W obu tkankach mamy ekspresję 2000 genów, czy tkanki A i B mogą mieć wspólne mechanizmy ekspresji genów?
- Mam 10 000 pozycji w genomie, gdzie może wystąpić wiązanie czynnika X, oraz 15 000 miejsc potencjalnych wystąpień wiązania czynnika Y, jeśli wiem, że wszystkich miejsc wiązania może być 100000, to czy mogę zakładać, że X I Y wiążą się niezależnie od siebie, jeśli mają wspólnych 2000 wystąpień?