Curs 3.

Biostatistica:

trecere in revista a metodelor statistice clasice (II)

Biblio:

W.Ewens, G.R. Grant – Statistical methods in bioinformatics, Springer, 2005

Cap. 1-3, cap.5

S.A. Glantz – Primer of biostatistics, McGraw Hill, 2005

Structura

- Teste de asociere (independență)
- Teste de concordanță
- Teste neparametrice
- Corelație și regresie

Teste de asociere

- Problema:
  - Se consideră că datele sunt grupate în categorii după două criterii
  - Aceasta presupune existența unui tabel de frecvențe (numit tabel de contingență) în care liniile sunt asociate cu un criteriu iar coloanele sunt asociate cu alt criteriu
  - Se pune întrebarea dacă există vreo legatură între cele două criterii

Teste de asociere

- Exemplu
  - N șoareci de laborator (dintre care n masculi și N-n femele) au fost iradiați
  - In urma iradierii m (m_1 masculi și m_2 femele) dintre șoareci au suferit mutații
  - Se pune întrebarea dacă există vreo asociere între sexul șoarecelui și riscul de apariție a unei mutații

	Mutant	Non-mutant	Total
mascul	m_1	$n-m_1$	n
femela	m_2	$N-n-m_2$	N-n
Total	m	N-m	N

Teste de asociere

- Dacă nu ar exista asociere între sex și apariția unei mutații atunci variabila aleatoare corespunzătoare numărului de masculi cu mutații ar avea distribuția hipergeometrică
- Remember: Repartiția hipergeometrică
  - Se asociază unei succesiuni de m experimente dependente (ex: extragere fără revenire dintr-o urnă cu n bile roșii și N-n bile albe) de tip Bernoulli (ieșiri posibile: roșu / alb)
  - Y = nr de bile roșii extrase

$$P(Y = y) = \frac{C_n^y C_{N-n}^{m-y}}{C_N^m}$$

$$E(Y) = \frac{mn}{N}$$

$$Var(Y) = \frac{mn(N-m)(N-n)}{N^2(N-1)}$$

Testul Fisher

- **Ipotezele:**
  - H_0: “nu există asociere între sex și apariția unei mutații”
  - H_A: ”șoarecii masculi sunt mai predispuși (sau mai puțin predispuși) mutațiilor decat femelele”

- **Testul Fisher**:
  - Statistica: numărul de șoareci masculi care au suferit mutație
  - Daca H_0 este adevarată atunci statistica are repartiția hipergeometrică
  - Se calculează probabilitatea ca nr. de masculi mutanți să fie mai mare sau cel puțin egal cu valoarea înregistrată
  - Daca probabilitatea obținută este mai mică decât nivelul de semnificație a testului atunci ipoteza nulă se respinge

Testul Fisher

- Exemplu (N, n și m sunt fixate)

	Mutant	Non-mutant	Total
mascul	y=6	2	n=8
femela	3	9	12
Total	m=9	11	N=20

$$P(Y = y) = \frac{C_n^y C_{N-n}^{m-y}}{C_N^m}$$

$$P(Y \geq 6) = \frac{C_8^6 C_{12}^3}{C_{20}^9} + \frac{C_8^7 C_{12}^2}{C_{20}^9} + \frac{C_8^8 C_{12}^1}{C_{20}^9} = 0.039$$

- Pentru nivelul de semnificație 0.05 ipoteza nulă se respinge (adică nu se poate afirma că nu există asociere între sexul șoarecelui și predispoziția către mutații)

Testul Fisher

- Se poate aplica doar în cazul tabelelor 2x2
- Este important ca valorile din tabel să corespundă unor evenimente independente între ele (ex: evenimentul ca un șoarece să fie mutant este independent de evenimentul ca alt șoarece să fie mutant)
- In cazul unui număr mai mare de categorii (valori posibile) pentru fiecare criteriu se aplică testul chi-patrat

Testarea asocierii cu testul chi-patrat

- Se consideră cazul a
  - r linii (r=număr de valori posibile asociate primului criteriu)
  - c coloane (c=număr de valori posibile asociate celui de al doilea criteriu)
- Statistica:

$$\sum_{j,k} \frac{(Y_{jk} - E_{jk})^2}{E_{jk}}$$

Y_{jk} = nr elementelor care au valoarea j pt primul criteriu si valoarea k pt al doilea

$$E_{jk} = \frac{y_{j*} y_{*k}}{y}$$

y_{j*} = suma valorilor de pe linia j

y_{*k} = suma valorilor de pe coloana k

y = suma tuturor valorilor din tabel

Daca ipoteza nulă este adevarată (nu există asociere între grupările corespunzătoare celor două criterii) atunci statistica are repartiția chi-patrat cu (r-1)(c-1) grade de libertate

Testarea asocierii cu testul chi-patrat

Exemplu:

- Se pune problema dacă într-o secvență ADN există asociere între nucleotidele consecutive sau nu
- Tabelul de contingență va fi constituit din 4 linii și 4 coloane, cele 4 categorii corespunzând celor 4 tipuri de nucleotide
- Liniile corespund nucleotidei prezente pe poziția i iar coloanele corespund nucleotidei prezente pe poziția următoare (i+1)
- Daca pozițiile succesive sunt independente atunci statistica va avea repartiția chi-pătrat cu

  (4-1)*(4-1)=9 grade de libertate

Nucleotida de pe pozitia i+1

Nucleotida de pe pozitia i

	A	G	C	T	
A	Y_{11}	Y_{12}	Y_{13}	Y_{14}	y_{1*}
G	Y_{21}	Y_{22}	Y_{23}	Y_{24}	y_{2*}
C	Y_{31}	Y_{32}	Y_{33}	Y_{34}	y_{3*}
T	Y_{41}	Y_{42}	Y_{43}	Y_{44}	y_{4*}
	y_{*1}	y_{*2}	y_{*3}	y_{*4}	y

Obs:

- Y_{32} reprezintă numărul de perechi (C,G) din secvență
- Y_{43} reprezintă numărul de perechi (T,C) din secvență

....

Testarea asocierii cu testul chi-patrat

- Utilitate: identificarea secvențelor codante (exoni) și a celor necodante (introni)
- Se consideră că cele două categorii de secvențe ADN au proprietăți statistice diferite
- Identificarea “amprentei” statistice se face diferit în funcție de prezența/absența unor asocieri între nucleotide succesive
- In cazul absenței unor asocieri între nucleotidele succesive acestea sunt considerate independente și amprenta este determinată de distribuția individuală a fiecărui tip de nucleotidă (estimarea probabilităților specifice distribuției multinomiale)
- In cazul prezenței unei asocieri trebuie extras un model de dependență (de exemplu dependența markoviană)

Testarea asocierii cu testul chi-patrat

- Exemplu:

  - verificarea ipotezei ca nucleotidele dintr-o secvență sunt independente
  - se pornește de la contorizarea dinucleotidelor (dimerilor)
  - se calculează valoarea statisticii
  - se compară cu valoarea critică corespunzătoare repartiției chi-pătrat cu 9 grade de libertate și nivel de semnificație 0.05 (valoarea este: 16.92)

  Obs: Exemplu în laborator 2

Teste de concordanta

- Au ca scop să verifice dacă populația din care sunt extrase datele are o anumită repartiție

- Tip problemă: testarea ipotezei ca repartiția nucleotidelor este uniformă (pentru fiecare poziție, fiecare nucleotidă apare cu aceeași probabilitate, 0.25)

- Exemple:
  - Testul chi-pătrat
  - Testul Kolmogorov-Smirnov

Teste de concordanta

- Testul chi-pătrat
- H_0: $F_X(x)=F_0(x)$ H_A:$F_X(x)<>F_0(x)$

Condiții preliminare:

- Domeniul de definiție al lui F este [a,b]
- Eșantionul este de volum n

Etape:

- Discretizare [a,b] în k subintervale (dacă este cazul):

  $a=t_0<t_1<\ldots<t_k=b$; clasa i: $[t_{i-1},t_i)$
- Calcul probabilitate teoretică pt. fiecare clasă ($p_i=F_0(t_i)-F_0(t_{i-1})$)
- Calcul frecvență pt. fiecare clasă

  n_i = nr. de date din eșantion ce aparțin lui $[t_{i-1},t_i)$

  (frecvența absolută corespunzătoare intervalului)

Teste de concordanta

- Testul chi-pătrat
- Statistica

$$T = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \in \chi^2(k-1)$$

$$T > \chi^2(k-1, \alpha) \Rightarrow \text{se respinge ipoteza nula}$$

- Obs. In cazul variabilelor discrete nu mai este necesară discretizarea:

k = reprezintă numărul de valori posibile (ex: 6 - zar, 4 - ADN)

n_i = nr. de date din eșantion care au valoarea i

p_i= 1/k (în cazul repartiției uniforme)

Teste de concordanta

- Exemplu: verificarea ipotezei că nucleotidele dintr-o secvență au distribuția uniformă pe setul {A,G,C,T} = secvența este aleatoare
- Etape:
  - Se determină frecvențele de apariție, n_i, ale nucleotidelor din secvență
  - Se calculează statistica T (slide anterior) pentru p_i=0.25 și k=4
  - Se calculeaza valoarea critică a repartiției chi-patrat cu 3 grade de libertate corespunzatoare nivelului de semnificație dorit (pentru 0.05 valoarea este 7.81)
  - Daca T este mai mare decat valoarea critică se respinge ipoteza că nucleotidele au o distribuție uniformă

Obs: Exemplu in Laborator 2

Teste neparametrice

- Sunt teste care permit compararea a două populații și care nu folosesc ipoteze asupra repartiției populațiilor sau parametrilor
- Se pot aplica în cazul variabilelor care nu sunt neapărat numerice (este suficient ca valorile să poată fi comparate între ele – de exemplu variabile ordinale)

- Exemple:

  - Testul semnelor
  - Testul rangurilor (Mann-Whitney-Wilcoxon)
  - Testul rangurilor cu semn (Wilcoxon)

Testul semnelor

- Test de comparare a două populații împerecheate pentru care diferența mediilor de selecție nu are repartiția normală (varianta neparametrică a testului t)

- Specific:
  - în loc să se utilizeze valorile numerice ale observațiilor se folosesc doar semnele unor diferențe sau rezultatul unor comparații între valori ale unor variabile ordinale
  - Eșantioanele din cele două populații trebuie să fie împerecheate (de exemplu valoarea unei mărimi înainte și după aplicarea unui tratament pentru același pacient)
  - Ipoteza nulă: H_0: $M_1=M_2$ (mediile celor două populații sunt egale sau nu există diferență între valorile inițiale și cele ulterioare – de exemplu: tratamentul nu are efect)

Testul semnelor

Etape:

- Se calculează diferențele dintre valorile corespunzătoare și se determină semnele acestora
  - n_1 diferențe pozitive
  - k valori egale
  - n_2 diferențe negative

- Statistica: T=numărul de diferențe pozitive (n_1)

- Daca H_0 e adevarată atunci T are repartiția binomială:
  - pe {0,..,m=n-k} (cazurile de egalitate se ignoră)
  - cu parametrul p=1/2

Testul semnelor

Etape:

- Se calculează $P(T<=n_1)$ folosind tabelul repartiției binomiale pentru B(m,1/2)

$$P(T \leq n_1) = \sum_{i=0}^{n_1} C_m^i (\frac{1}{2})^i (\frac{1}{2})^{m-i} = \frac{1}{2^m} \sum_{i=0}^{n_1} C_m^i$$

  - Daca valoarea este mai mica decat nivelul de semnificație atunci se respinge H_0

Testul semnelor

Exemplu: analiza impactului unui tratament

Valoare inițială	Valoare finală	Semnul diferenței	Efect
x_1	y_1	$Sgn(x_1-y_1)$	pozitiv
x_2	y_2	$Sgn(x_2-y_2)$	negativ
.	.	.	neutru
.	.	.	negativ
x_n	y_n	$Sgn(x_n-y_n)$	pozitiv
			pozitiv
			neutru
			...

Exemplu:

Se calculează numărul diferențelor pozitive sau a cazurilor în care efectul este pozitiv (n1) și probabilitatea P(T<=n1) (folosind formula din slide-ul anterior)

Testul rangurilor (Mann-Whitney)

- Specific: se folosește pentru compararea a două populații a căror repartiție este necunoscută (eșantioanele sunt independente și au m respectiv n elemente)
- H0: cele două populații au aceeași repartiție
- Etape:
  - Se construiește eșantionul reunit
  - Se ordonează crescător după valoare
  - Se asociază fiecărui element un rang (de la 1 la m+n)
  - Se calculează
    - R1= suma rangurilor asociate elementelor din primul eșantion
    - R2= suma rangurilor asociate elementelor din al doilea eșantion

Testul rangurilor (Mann-Whitney)

- R1 = suma rangurilor asociate elementelor din primul eșantion
- R2 = suma rangurilor asociate elementelor din al doilea eșantion

$$R_1 + R_2 = (m+n)(m+n+1)/2$$

$$U = \min\{U_1, U_2\}$$

$$U_1 = mn + \frac{m(m+1)}{2} - R_1$$

$$U_2 = mn + \frac{n(n+1)}{2} - R_2$$

Testul rangurilor (Mann-Whitney)

- Dacă ipoteza nulă este adevarată atunci variabila U are repartiția U și are proprietățile:

$$E(U) = \frac{mn}{2}$$

$$Var(U) = \frac{mn(m+n+1)}{12}$$

- Dacă m și n sunt suficient de mari (m>20, n>20) atunci :

$$T = \frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \in \mathrm{N}(0,1)$$

Testul rangurilor (Mann-Whitney)

- Pentru a se lua decizia se parcurg următoarele etape:
  - Se calculează valoarea statisticii T
  - Se determină valoarea critică corespunzătoare repartiției normale standard pt. nivelul de semnificație dorit (pentru 0.05 valoarea este 1.65)
  - Daca T este în regiunea critică (pentru nivelul de semnificație 0.05, aceasta înseamnă să fie în afara intervalului [-1.65, 1.65]) atunci ipoteza nula se respinge

Testul rangurilor cu semn (Wilcoxon)

- Specific: testarea medianei unei populații (cu repartiție asimetrică)
- H0: Mediana=M_0

- Etape:
  - Se calculează modulele diferențelor x_i-M_0
  - Se ordonează crescător și li se asignează ranguri (valorilor identice li se asociază același rang)
  - Statistica: T=suma rangurilor diferențelor inițial pozitive

- Daca H0 e adevarată atunci T are proprietățile
  - E(T) = n(n+1)/4
  - Var(T) = n(n+1)(2n+1)/24

- Daca n este mare atunci T are repartiția normală cu parametrii de mai sus

Testul rangurilor cu semn (Wilcoxon)

- Alte aplicații:
  - Se poate utiliza pentru compararea a două selecții împerecheate (similar testului semnelor)
    - De exemplu pentru a compara comportarea a doi algoritmi aleatori de optimizare în ipoteza că se rulează ambii algoritmi de mai multe ori pornind de la aceeași aproximație inițială

Corelatie si regresie

- Scop: analiza dependenței dintre una sau mai multe mărimi predictor și o mărime prezisă
  - Dependența dintre greutate și înălțime
  - Dependența dintre indicele de masă corporală și vârstă
- Coeficienți de corelație: permit analiza cantitativă a gradului de dependență între mărimi
- Regresie: permite determinarea tipului de dependență și a parametrilor acesteia:
  - Regresie liniară simplă / multiplă
  - Regresie neliniară simplă / multiplă
  - Regresie logistică

Coeficienti de corelatie

n = volum eșantion pt fiecare variabilă

Coeficient de corelație (Pearson)

- Util pentru variabile numerice
- măsură a gradului de dependență liniară
- Valori între -1 și 1
- Valoare apropiată de +1/ -1: corelație liniară pozitivă/ negativă semnificativă
- Valoare apropiată de 0: nu există corelație liniară între variabile

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Coeficient de pe baza rangurilor (Spearman)

- Se ordonează crescător valorile corespunzătoare fiecărei mărimi
- Se calculează diferența dintre ranguri (d_i)
- E adecvat pt variabile ordinale (nu neapărat numerice) în cazul în care valorile asociate celor două mărimi sunt distincte
- Valoare de +1/ -1: corelație pozitivă/ negativă semnificativă (nu neapărat liniară)
- Valoare apropiată de 0: nu există corelație între variabile

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Regresie liniară

- Regresie liniară simplă
  - Date de intrare: $(x_1,x_2,...,x_n)$ și $(y_1,y_2,...,y_n)$
  - x_i și y_i sunt valori scalare
  - Ieșire: estimarea parametrilor a si b ai modelului de dependență liniară Y=aX+b
  - Scopul estimării: determinarea valorilor lui a și b care minimizează suma pătratelor erorilor

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

- Regresie logistică simplă
  - Scop: analiza dependența dintre o variabilă nominală (în cazul cel mai simplu, având 2 valori) și o variabilă cantitativă
  - Idee: se consideră că raportul șanselor celor două valori posibile (odds ratio: p/(1-p)) depinde de variabila predictor cf unui model log-liniar; notând y=log(p/(1-p)) se ajunge la o problemă de regresie liniară simplă

$$\frac{p}{1-p} = \exp(a + bx)$$

$$\log\left(\frac{p}{1-p}\right) = a + bx$$