Curs 2.

Biostatistică:

trecere în revistă a metodelor statistice clasice (I)

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

- Motivație
- Reminder: noțiuni de bază din probabilități și statistică
- Repartiții discrete și continue
- Estimarea parametrilor
- Verificarea ipotezelor statistice

Motivație

- Pentru două secvențe ADN (cuvinte peste alfabetul {A,C,G,T})

GGAGACTGTAGACAGCTAATGCTATA
GAACGCCCTAGCCACGAGCCCTTATC

Se pune problema dacă similaritatea dintre ele este doar întâmplătoare sau exprimă faptul ca secvențele provin de la organisme înrudite (au un strămoș comun în arborele de evoluție)

- Răspunsul necesită estimarea probabilității ca în cazul unor secvențe aleatoare de lungime 26 să existe 11 poziții identice
- E necesară definirea unui model probabilist asociat rezultatelor observării perechilor corespondente de nucleotide

Reminder: noțiuni de bază din probabilități și statistică

- Experiment aleator: experiment al cărui rezultat poate să difere de la o repetare la alta
  - Exemple:
    - Aruncarea unui zar
    - Observarea perechilor de nucleotide corespondente din două secvențe aliniate (la o repetare a experimentului se observă o singură pereche, selectată la întâmplare)

- Spațiu de selecție: mulțimea rezultatelor elementare (mutual exclusive) ce pot fi obținute prin efectuarea experimentului
  - Exemple:
    - Aruncarea unui zar: {1,2,3,4,5,6}
    - Observarea perechilor de nucleotide: {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}

Reminder: noțiuni de bază din probabilități și statistică

- Eveniment aleator: rezultat al unui experiment aleator
  - Exemple:
    - E1: “la aruncarea zarului s-a obținut o valoare pară”
    - E2: “s-a observat o pereche de nucleotide identice”

- Probabilitate: măsură a șansei de realizare a unui eveniment aleator = numărul cazurilor favorabile/numărul cazurilor posibile
  - Exemple:
    - P(E1) = card{2,4,6}/card {1,2,3,4,5,6}=3/6=0.5
    - P(E2) = card{AA,CC,GG,TT}/card {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}= 4/16=0.25

Reminder: noțiuni de bază din probabilități și statistică

- Evenimente independente: realizarea unuia dintre evenimente nu este influențată de realizarea celuilalt eveniment
  - Exemplu:
    - Repetarea aruncării zarului sau aruncarea simultană a două zaruri.
  - Proprietate: dacă E și E' sunt independente atunci probabilitatea ca ambele să se realizeze este: P(E,E')=P(E)P(E')
    - Probabilitatea să se obțină (6,6): 1/36
- Evenimente condiționate: realizarea unuia dintre evenimente influențează șansa de realizare a celuilalt
  - Exemplu: se consideră că se știe că s-a observat o pereche de purine (din {A,G}). Se pune intrebarea care este probabilitatea de a se fi observat o pereche de nucleotide identice
  - Proprietate (probabilități condiționate): P(E'|E)=P(E',E)/P(E) (probabilitatea evenimentului E' conditionata de evenimentul E)
  - P(E',E)=card{AA,GG}/16=1/8 P(E)=card{AA,AG,GA,GG}/16=1/4
    P(E'|E)=(1/8)/(1/4)=1/2

Reminder: noțiuni de bază din probabilități și statistică

- Variabilă aleatoare: caracteristică asociată unui experiment aleator; o variabilă aleatoare poate lua mai multe valori, fiecare cu o anumită probabilitate
  - Exemplu:
    - Valoarea obținută la aruncarea unui zar (la aruncări diferite se obțin rezultate diferite)
    - Tensiunea arterială a unei persoane (pentru persoane selectate la întâmplare valorile pot fi diferite)
  - In funcție de natura caracteristicii variabilele aleatoare pot fi:
    - Discrete: mulțimea de valori posibile este finită sau infinită dar numărabilă
    - Continue: caracteristica ia valori dintr-un interval
  - Dpdv formal, o variabilă aleatoare este o funcție definită pe spațiul de selecție cu valori într-o mulțime (de regulă o mulțime numerică)

Reminder: noțiuni de bază din probabilități și statistică

Distribuție (repartiție) de probabilitate: asociază probabilități valorilor posibile ale variabilei aleatoare

- Exemplu:
  - Aruncarea unui zar (corect): (1/6,1/6,1/6,1/6,1/6,1/6)
  - Observarea unei nucleotide (o poziție arbitrară din secvență): (p_1,p_2,p_3,p_4) astfel încât $p_1+p_2+p_3+p_4=1$
- Variabile aleatoare discrete:
  - Pp ca variabila aleatoare X ia valori în mulțimea $\{x_1,x_2,\ldots, x_n\}$
  - Distribuția de probabilitate este $(p_1, p_2,..., p_n)$ unde $p_i=P(x=x_i)$
  - Distribuția de probabilitate acoperă toate cazurile posibile: p1+p2+...+pn=1
- Variabile aleatoare continue:
  - Funcție de densitate de probabilitate pt variabila aleatoare X (f:R->R_+):

$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)dx = 1, \qquad P(a \leq X \leq b) = \int_a^b f(x)dx$$

  - Funcție de repartiție de probabilitate pt variabila aleatoare X (F:R->[0,1])

$0 \leq F(x) \leq 1,\ F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$, F este crescătoare

Reminder: noțiuni de bază din probabilități și statistică

Valori sintetice asociate unei variabile aleatoare:

- Valoare medie:

$$E(X) = \sum_{i=1}^{n} x_i \, \mathrm{P}(x_i), \qquad \text{(variabile discrete)}$$

$$\mathrm{E(X)} = \int_{-\infty}^{\infty} xf(x)dx \qquad \text{(variabile continue)}$$

Proprietăți: pt două variabile aleatoare X și Y au loc

- E(X+Y)=E(X)+E(Y)
- Dacă X și Y sunt independente atunci E(XY)=E(X)E(Y)

- Varianța (dispersia):
  - $\mathrm{Var(X)=E((X-E(X))^2)=E(X^2)-(E(X))^2}$

Repartiții discrete – trecere în revistă

- **Bernoulli**
  - se asociază experimentelor ce produc două rezultate posibile (de exemplu un eveniment de tip succes și un eveniment de tip eșec)
  - $P(Y=a)=p$, $P(Y=b)=1-p$
    (probabilitatea unui rezultat este p iar a celuilalt este 1-p)
  - Media: $E(Y)=a*p+b*(1-p)$,
  - Varianța: $Var(Y)=(a-b)^2p(1-p)$
  - **Exemplu:** evenimentul poate fi potrivirea a două nucleotide
    - eveniment de tip succes : nucleotidele coincid (a=1)
    - eveniment de tip eșec: nucleotidele nu coincid (b=0)
    - în acest caz particular: E(Y)=p, Var(Y)=p(1-p)

Repartitii discrete – trecere in revista

- **Binomială**
  - Se asociază unei succesiuni de n experimente independente de tip Bernoulli (ieșiri posibile: succes / eșec)
  - Y = nr de experimente (nu neapărat consecutive) în care se obține succes
  - **Exemplu**: nr de poziții în care două secvențe ADN coincid

(obs: aplicabil doar dacă valorile asociate nucleotidelor consecutive sunt independente)

$$P(Y = y) = C_n^y \, p^y (1-p)^{n-y}$$

$$E(Y) = np, \quad Var(Y) = np(1-p)$$

Exemplu: p=0.25, n=26, y=11

$$P(Y = 11) = C_{26}^{11} 0.25^{11} 0.75^{15} = 0.025$$

Repartitii discrete – trecere in revista

- Geometrică
  - Se asociază unei succesiuni de n experimente independente de tip Bernoulli (ieșiri posibile: succes / eșec)
  - Y = nr de experimente până la primul eșec (succese consecutive)
  - Exemplu: lungimea unei secvențe de potriviri consecutive între doua lanțuri ADN

(obs: aplicabil doar dacă valorile asociate nucleotidelor consecutive sunt independente)

$$P(Y = y) = (1 - p)\,p^y$$

$$E(Y) = \frac{p}{1-p}, \quad Var(Y) = \frac{p}{(1-p)^2}$$

Exemplu:

p=0.25, 11 potriviri consecutive

$P(Y=11)=1.7*10^{-7}$

Obs: e foarte improbabil ca în două secvențe aleatoare de câte 26 nucleotide să fie 11 potriviri consecutive

Repartitii discrete – trecere in revista

- Negativ binomială
  - Se asociază unei succesiuni de experimente independente de tip Bernoulli (ieșiri posibile: succes / esec)
  - Y = nr de repetări până la obținerea a m succese (nu neapărat consecutive)
  - Exemplu: lungimea minimă a două secvențe în care sunt m potriviri (obs: aplicabil doar dacă valorile asociate nucleotidelor consecutive sunt independente)

$$P(Y=y)=C_{y-1}^{m-1}p^m(1-p)^{y-m},\ y\in\{m,m+1,\ldots\}$$

$$E(Y)=\frac{m}{p},\ Var(Y)=\frac{m(1-p)}{p^2}$$

Exemplu:

Valoarea medie a lungimii unei secvențe în care sunt 11 potriviri nu neapărat consecutive (p=0.25): 44

Repartitii discrete – trecere in revista

- 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: rosu / alb)
  - Y = nr de bile roșii extrase
  - Exemplu: N șoareci de laborator (n masculi și N-n femele) sunt iradiați. Dupa iradiere m șoareci devin mutanți. Dacă nu ar exista corelație între apariția unei mutații și sexul șoarecelui atunci nr. de masculi mutanți are avea distribuția hipergeometrică

$$P(Y=y)=\frac{C_n^y C_{N-n}^{m-y}}{C_N^m},\ y\in\{\max\{0,n+m-N\},\ldots,\min\{m,n\}\}$$

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

Repartitii discrete – trecere in revista

- Multinomială
  - Extinderea repartiției binomiale la cazul repetării de n ori a unui experiment în care sunt m evenimente posibile (m>2)

    (Obs. în cazul binomial m=2)
  - $Y_1, Y_2, \ldots, Y_m$:variabile aleatoare dependente,
    - Y_i – numărul de apariții ele evenimentului i în n repetari ale experimentului
    - Y_i are repartiția binomială de parametru p_i
    - $Y_1+Y_2+\ldots+Y_m=n$
  - Exemplu: Numărul de prezențe ale unei nucleotide într-un șir de lungime n (m=4) în ipoteza că distribuția celor 4 tipuri de nucleotide este (p_1,p_2,p_3,p_4).

$$P(Y=(y_1,y_2,\ldots,y_m))=\frac{n!\,p_1^{y_1}\,p_2^{y_2}\ldots.p_m^{y_m}}{y_1!y_2!\ldots y_m!}$$

Repartitii discrete – trecere in revista

- Poisson
  - Caz limită al repartiției binomiale:
    - n tinde la infinit,
    - p tinde la 0
    - np=λ (parametrul repartiției)
  - Exemplu 1: Y=Numărul de proteine din celulă care se degradează spontan în unitatea de timp

$$P(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

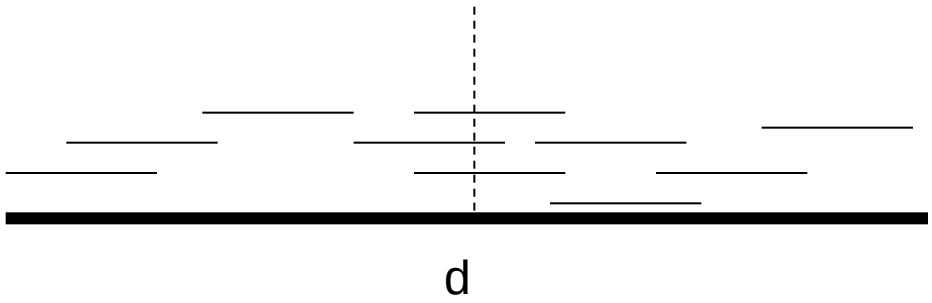
$$E(Y) = Var(Y) = \lambda$$

Repartitii discrete – trecere in revista

- Poisson. Exemplu 2 (Shotgun sequence – secvențiere ADN pe bază de fragmentare în poziții aleatoare)
  - problema reconstruirii secvenței ADN pornind de la fragmente cu suprapuneri nevide
  - Notații:
    - G – lungimea secvenței
    - N – număr fragmente
    - L – lungime fragment
  - Y = numărul de fragmente ADN avand extremitatea stanga in intervalul [d-L,d] unde d este o poziție arbitrară în secvență are repartiția Poisson de parametru a=NL/G (numit grad de acoperire al secvenței)

$$P(Y=y)=\frac{\lambda^y \exp(-\lambda)}{y!}$$

$$E(Y)=Var(Y)=\lambda$$



G=7, L=1, N=10
a=10/7

P(Y=3)=0.11

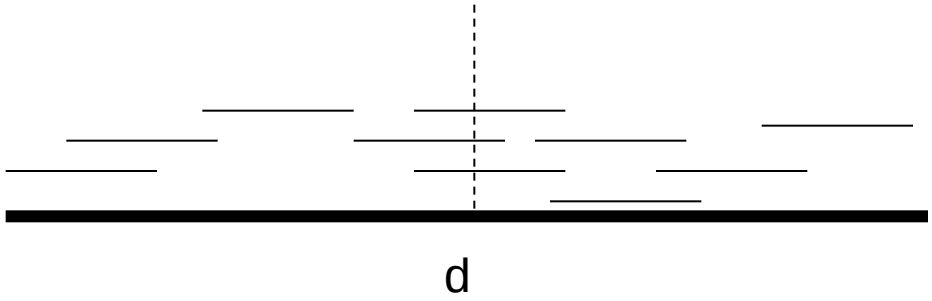
P(Y>=1)=1-P(Y=0)=1-0.24=0.76

Obs: In proiectul Human Genome gradul de acoperire utilizat a fost 8

Repartitii discrete – trecere in revista

- Poisson. Exemplu 2
  - Problema inversă: să se determine gradul minim de acoperire (a) care asigură faptul că probabilitatea ca fiecare poziție să fie "acoperită" de către cel puțin un fragment este cel puțin P_0
  - Se caută cea mai mică valoare a lui a pentru care

    $1-\exp(-a)>P_0$
  - Daca $P_0=0.9$ atunci $a>-\ln(1-P_0)=2.3$ (adică lungimea totală a fragmentelor trebuie să fie cel puțin de 2.3 ori mai mare decât lungimea secvenței)



$$P(Y=y)=\frac{\lambda^y \exp(-\lambda)}{y!}$$

$$E(Y)=Var(Y)=\lambda$$

Repartitii continue – trecere in revista

- Uniformă pe [a,b]: U(a,b)
- Normală: N(m,δ^2)
  - Bin(n,p) -> N(np,np(1-p)) (n tinde la infinit, p tinde la 0)
  - Poisson(λ) -> N(λ, λ) (lambda mare)
- Exponențială
  - Durata până la producerea unui eveniment (varianta continuă a repartiției geometrice)

$$U(a,b):\ f(x)=\frac{1}{b-a},\ E(X)=\frac{a+b}{2},\ Var(X)=\frac{(b-a)^2}{12}$$

$$N(m,\sigma^2):\ f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right),\ E(X)=m,\ Var(X)=\sigma^2$$

$$Expon(\lambda):\ f(x)=\lambda\exp(-\lambda x),\ E(X)=\frac{1}{\lambda},\ Var(X)=\frac{1}{\lambda^2}$$

Probabilități și statistică

- Probabilități:
  - Se presupune cunoscut modelul probabilist asociat unui proces/ fenomen/ experiment
  - Pe baza modelului probabilist se obțin informații despre desfășurarea procesului/ fenomenului /experimentului
  - Model probabilist -> ce ar trebui să observăm în ipoteza că procesul se desfășoară în conformitate cu modelul

- Statistică
  - Se selectează, în mod aleator, elemente dintr-o populație (eșantion) care se comportă după un model necunoscut sau cunoscut incomplet
  - Pe baza eșantionului se emit ipoteze despre modelul probabilist asociat populației și se cuantifică gradul de verosimilitate sau de încredere a acestor ipoteze
  - Eșantion -> care ar putea fi modelul (sau parametrii modelului) care explică ceea ce am observat

Estimarea parametrilor

- Date intrare:
  - Valori ale mărimii de interes determinate pe baza unui eșantion (obținut prin selecție aleatoare)
  - Ipoteze privind repartiția asociată mărimii (forma repartiției)

- Ieșire:
  - Estimarea parametrului/ parametrilor asociați repartiției

- Exemple de parametri:
  - Medie, varianță (caracteristici continue)
  - Proporție (caracteristici discrete)

- Modalități de estimare:
  - Punctuală (metoda verosimilității maxime)
  - Prin intervale de încredere

Estimarea parametrilor

- Exemplul 1: estimarea înălțimii medii a studenților din universitate
  - Date de intrare: valori ale înălțimii măsurate pentru un eșantion
  - Ipoteză privind repartiția valorii înălțimii: repartiție normală
  - Parametrul de estimat: media repartiției
- Exemplul 2: estimarea proporției de Adenină din genomul uman
  - Date intrare: eșantion = secvență de nucleotide
  - Ipoteza privind repartiția: repartiția binomială (succes: prezența adenină; eșec: alt tip de nucleotidă)
  - Parametrul de estimat: proporția = parametrul p de la repartiția binomială

Estimarea parametrilor

- Estimare punctuală a parametrilor
  - **Scop:** estimarea valorii parametrului/parametrilor care explică cel mai bine datele observate în eșantion ($X_1, X_2, \ldots, X_n$); elementele din eșantion sunt considerate variabile aleatoare independente
  - **Metoda verosimilității maxime** – estimația parametrului trebuie să maximizeze probabilitatea observării valorilor efective ale eșantionului

Funcția de verosimilitate

functia densitate de repartiție

$$L(X_1,\ldots,X_n;\theta)=\prod_{i=1}^{n} f_X(X_i;\theta)$$

$$\log L(X_1,\ldots,X_n;\theta)=\sum_{i=1}^{n}\log f_X(X_i;\theta)$$

Determinarea parametrului care maximizează log-verosimilitatea ⇔ rezolvarea unei probleme de optimizare

Log-verosimilitate

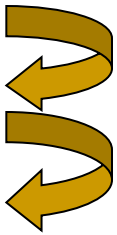
Valori observate

Parametru de estimat

Estimarea parametrilor

- Estimare prin intervale de încredere:
  - Intervale de încredere: interval aleator ce conține valoarea parametrului cu o probabilitate prestabilită (1-α); α corespunde probabilității de eroare (când valoarea reală nu se află în interval)

$$P(\theta \in (I(X_1,..., X_n), S(X_1,...., X_n))) = 1-\alpha$$
$$P(I(X_1,..., X_n) < \theta < S(X_1,...., X_n)) = 1-\alpha$$
$$P(a_1 < T(X_1,..., X_n; \theta) < a_2) = 1-\alpha$$



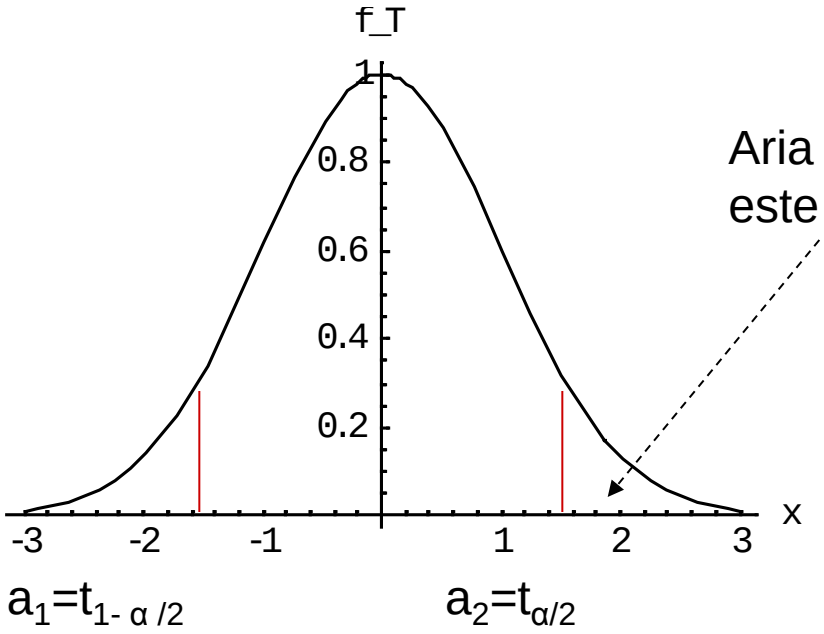
- T – variabilă pivotală (care depinde de variabilele aleatoare corespunzătoare eșantionului și de parametrul de estimat) a cărei funcție de repartiție este cunoscută (funcția densitate de repartiție corespunzătoare este în continuare notată cu t)
- a_1 și a_2 sunt denumite valori critice ale repartiției lui T corespunzătoare valorilor α_1 și α_2 (cu $\alpha_1 + \alpha_2 = \alpha$)

Estimarea parametrilor

Intervale de incredere:

- De obicei: $\alpha_1 = \alpha_2$
- Limitele se determină rezolvând setul de inecuații în raport cu parametrul necunoscut (theta)

$$T(X_1,...,X_n;\theta) > t_{1-\alpha/2}$$
$$T(X_1,...,X_n;\theta) < t_{\alpha/2}$$



Obs: în cazul repartițiilor simetrice:
$t_{1-\alpha/2} = t_{\alpha/2}$

t_α are proprietatea : $\int_{t_\alpha}^{\infty} T(u)du = \alpha$

Estimarea parametrilor

- Intervale de încredere pentru medie:
  - Eșantioane mici (n<30) – necesită ipoteza de normalitate (se consideră că valorile înregistrate sunt generate de o variabilă cu repartiția normală)
  - **Caz 1: varianța repartiției (sigma) e cunoscută**

$$T(X_1,\ldots,X_n;\mu)=\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\in N(0,1)$$

μ este media populatiei (parametrul necunoscut),

σ este abaterea standard a populatiei (cunoscuta)

$$\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}$ = valoare critica corespunzatoare repartitiei normale

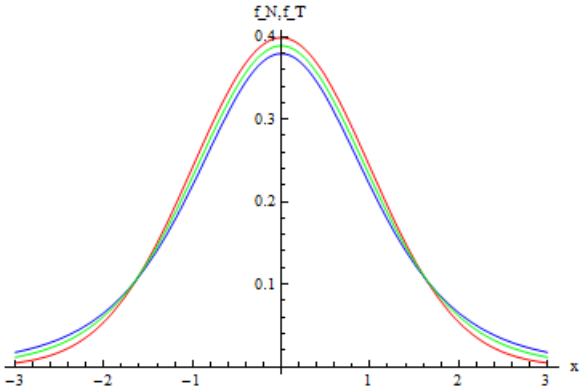
$\alpha=0.05\Rightarrow z_{\alpha/2}=1.96$

$\overline{X}$ = media calculata pe baza esantionului

Obs: în cazul esantioanelor mari (n>30) nu e necesară ipoteza de normalitate

Estimarea parametrilor

- Intervale de încredere pentru medie:



N(0,1) – rosu
t(5) – albastru
t(10) - verde

Caz 2: varianța repartiției e necunoscută

- Necesită ipoteza de normalitate
- Obs: în cazul eșantioanelor mari (n>30 sau mai degraba n>60) repartiția lui T poate fi aproximată cu cea normală

$$T(X_1,...,X_n;\mu)=\frac{\sqrt{n}(\overline{X}-\mu)}{s}\in t(n-1)$$ - repartitia Student

$$\overline{X}-t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}<\mu<\overline{X}+t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$t_{n-1,\alpha/2}$ = valoare critica corespunzatoare repartitiei Student cu n -1 grade de libertate

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$ = dispersia de selectie calculata pe baza esantionului

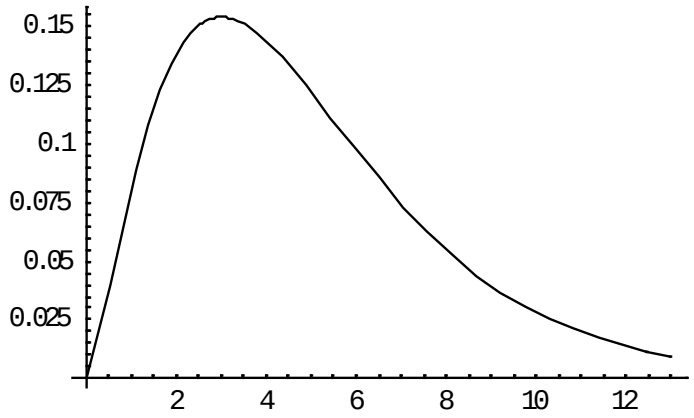
Estimarea parametrilor

- Interval de incredere pentru dispersie:
  - Ipoteza: caracteristica analizată (populația) are repartiție normală

$$T(X_1,\ldots,X_n;\sigma)=\frac{(n-1)s^2}{\sigma^2}\in\chi^2(n-1)$$ - repartitia chi patrat

$$\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}<\sigma<\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}$$

cu n -1 grade de libertate



Estimarea parametrilor

- Interval de încredere pentru proporție (π):
  - Ipoteze: caracteristică cu repartiție binomială, eșantion mare (np>10, n(1-p)>10) - permite aproximarea repartiției binomiale cu cea normală

$$T(X_1,\ldots,X_n;\pi)=\frac{\sqrt{n}(p-\pi)}{\sqrt{\pi(1-\pi)}}\in N(0,1)\ \text{(aproximativ)}$$

Valoare calculată pe baza eșantionului

Parametru de estimat

$$\text{Aproximare}:\ \pi(1-\pi)/n\cong p(1-p)/n$$

$$p-\frac{\sqrt{p(1-p)}}{\sqrt{n}}z_{\alpha/2}<\pi<p+\frac{\sqrt{p(1-p)}}{\sqrt{n}}z_{\alpha/2}$$

- p= valoarea proporției estimată pe baza selecției

Exemplu: estimarea pe baza unui eșantion a proporției de nucleotide de un anumit tip în genomul unei specii.

Verificarea ipotezelor statistice

Exemplu:

Fie GGAGACTGTAGACAGCTAATGCTATA

GAACGCCCTAGCCACGAGCCCTTATC

două secvențe ADN

In ce măsură se poate afirma că potrivirile dintre cele două secvențe sunt pur întâmplătoare ?

Statistica ofera un instrument (testele statistice) care permite determinarea unor răspunsuri la întrebări ca cea de mai sus

Procesul de verificare a unei ipoteze statistice constă în parcurgerea câtorva etape

Verificarea ipotezelor statistice

- Etape (varianta bazată pe utilizarea unui nivel de semnificație):
  - Stabilirea ipotezei nule (ipoteza care va fi respinsă sau nu) și a celei alternative. Notații uzuale: H_0 (ipoteza nulă); H_A (ipoteza alternativă)
  - Stabilirea nivelului de semnificație (măsură a probabilității ca ipoteza nulă să fie respinsă când ea este de fapt adevărată – probabilitatea de a face o eroare de tip I)
  - Identificarea statisticii testului = variabilă aleatoare a cărei repartiție este cunoscută dacă ipoteza nulă este adevărată
  - Identificarea valorilor critice (pe baza repartiției statisticii testului și a nivelului de semnificație)
  - Construirea regiunii critice (regiunea de respingere)
  - Luarea deciziei: ipoteza nulă se respinge dacă valoarea statisticii calculată pentru elementele eșantionului aparține regiunii de respingere

Verificarea ipotezelor statistice

- Exemplu (potrivirea a doua secvențe ADN):

Ipoteza nulă: “potrivirile dintre cele două secvențe sunt pur întâmplătoare”
⇔ “probabilitatea ca două poziții să fie identice este ¼”

Adică:

H_0: p=0.25

H_A: p<>0.25

Statistica testului: “numărul de potriviri între cele două secvențe”

(are repartiția binomială B(n,p) - dacă se presupune că nucleotidele succesive sunt interpretate ca realizări ale unor variabile aleatoare independente)

Verificarea ipotezelor statistice

Tipuri de teste statistice (clasificare în funcție de informațiile cunoscute despre populație)

- Teste parametrice
  - Se aplică atunci când forma repartiției populației este cunoscută
  - Se referă la parametrii repartiției
  - Variante:
    - pentru o populație (ex: ipoteză asupra mediei/ dispersiei unei populații)
    - pentru două populații (ex: ipoteza asupra relației dintre mediile a două populații)
    - pentru mai mult de două populații (ex: analiza impactului mai multor tratamente) -> analiza variației
  - Exemple de teste clasice:
    - testul z
    - testul Student (t)

Verificarea ipotezelor statistice

Tipuri de teste statistice (clasificare in functie de informațiile cunoscute despre populație)

Teste neparametrice

- Se aplică atunci când repartiția populației nu este cunoscută
- Variante:
  - Teste de concordanță (ex: testul chi-patrat) – verificarea unei ipoteze privind distribuția populației (ex: se poate afirma că o secvență ADN este generată aleator?; se poate considera că populația analizată are distribuție normală?)
  - Teste de independență /asociere (ex: testul Fisher, testul chi-patrat) – verificarea unei ipoteze privind independența dintre anumite caracteristici (ex: sunt valorile nucleotidelor consecutive dintr-o secvență independente?)
  - Teste de comparare în cazul caracteristicilor nominale/ordinale (ex: testul semnelor, testul rangurilor, testul rangurilor cu semn etc) – este un tratament mai efectiv decât altul?

Testul z (ipoteze asupra mediei unei populatii)

- Exemplu: verificarea unei ipoteze privind media unei populații cu repartiție normală de dispersie cunoscută

  - Variante de ipoteze

$$H_0{:}\mu = \mu_0 \qquad H_0 : \mu = \mu_0 \qquad H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0 \qquad H_A : \mu > \mu_0 \qquad H_A : \mu < \mu_0$$

Ipoteză alternativă bilaterală

Ipoteze alternative unilaterale

  - Nivel de semnificație: alpha = 0.05
  - Statistica (corespunzătoare testului z):

$$T(X_1,\ldots,X_n) = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \in N(0,1)$$

Testul z (ipoteze asupra mediei unei populatii)

- Valori critice și regiuni critice

$a) H_0: \mu = \mu_0 \quad H_A: \mu \neq \mu_0$

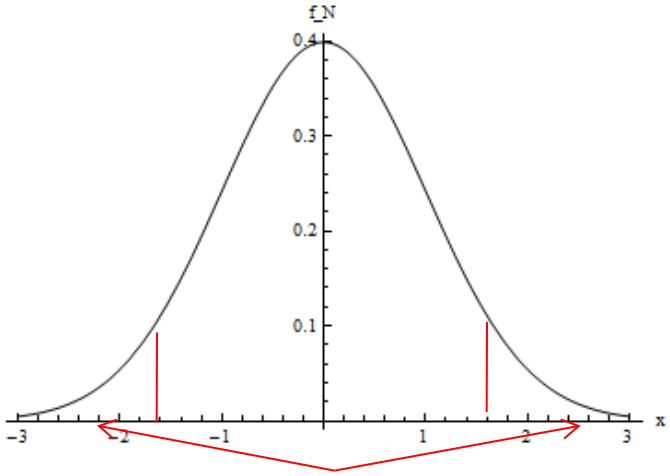
$b) H_0: \mu = \mu_0 \quad H_A: \mu > \mu_0$

$c) H_0: \mu = \mu_0 \quad H_A: \mu < \mu_0$

$a) \quad R = (-\infty, z_{1-\alpha/2}) \cup (z_{\alpha/2}, \infty)$

$b) \quad R = (z_{\alpha}, \infty)$

$c) \quad R = (-\infty, z_{1-\alpha})$



Regiune respingere (bilateral)

- Daca $T(x_1, x_2, \ldots, x_n)$ apartine lui R atunci se respinge ipoteza H_0

Testul t (ipoteze asupra mediei unei populatii)

- Exemplu: verificarea unei ipoteze privind media unei populații cu repartiție normală de dispersie necunoscută

  - Variante de ipoteze

$$H_0: \mu = \mu_0 \qquad H_0 : \mu = \mu_0 \qquad H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0 \qquad H_A : \mu > \mu_0 \qquad H_A : \mu < \mu_0$$

Ipoteză alternativă bilaterală

Ipoteze alternative unilaterale

  - Nivel de semnificație: alpha = 0.05
  - Statistica (corespunzatoare testului Student (notatie: t)):

$$T(X_1,\ldots,X_n) = \frac{\sqrt{n}(\overline{X} - \mu_0)}{s} \in t(n-1)$$

Testul t (ipoteze asupra mediei unei populatii)

- Valori critice și regiuni critice

$a) H_0: \mu = \mu_0 \quad H_A: \mu \neq \mu_0$

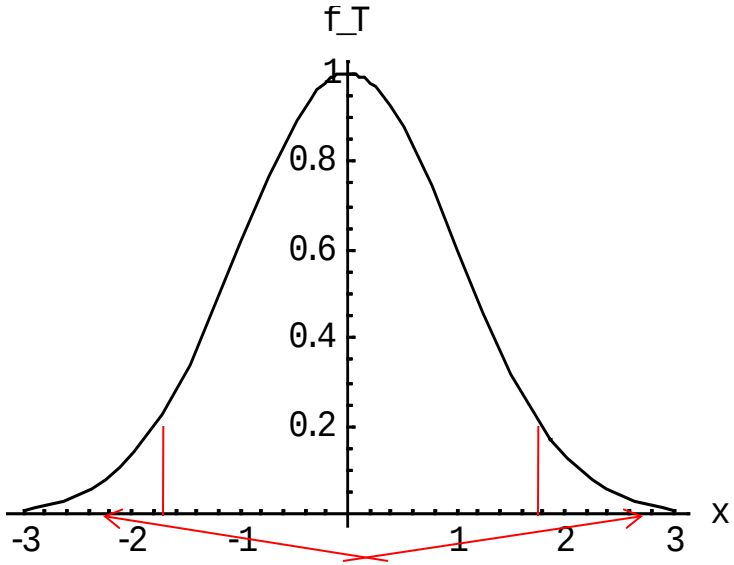
$b) H_0: \mu = \mu_0 \quad H_A: \mu > \mu_0$

$c) H_0: \mu = \mu_0 \quad H_A: \mu < \mu_0$

$a) \quad R = (-\infty, t_{n-1,1-\alpha/2}) \cup (t_{n-1,\alpha/2}, \infty)$

$b) \quad R = (t_{n-1,\alpha}, \infty)$

$c) \quad R = (-\infty, t_{n-1,1-\alpha})$



Regiune respingere (bilateral)

- Daca $T(x_1, x_2, \ldots, x_n)$ apartine lui R atunci se respinge ipoteza H_0

Testul chi-patrat (ipoteze asupra dispersiei unei populatii)

- Testarea unei ipoteze privind dispersia unei populații normale

$$a) H_0: \sigma = \sigma_0 \qquad b) H_0: \sigma = \sigma_0 \qquad c) H_0: \sigma = \sigma_0$$
$$H_A: \sigma \neq \sigma_0 \qquad H_A: \sigma > \sigma_0 \qquad H_A: \sigma < \sigma_0$$

$$T(X_1, ..., X_n) = \frac{(n-1)s^2}{\sigma_0^2} \in \chi^2(n-1)$$

$$a) R = (0, \chi^2_{n-1,1-\alpha/2}) \cup (\chi^2_{n-1,\alpha/2}, \infty)$$

$$b) R = (\chi^2_{n-1,\alpha}, \infty)$$

$$c) R = (0, \chi^2_{n-1,1-\alpha})$$

- In cazul eșantioanelor mari se poate folosi statistica de mai jos (fără a fi necesară ipoteza de normalitate)

$$T(X_1, ..., X_n) = \frac{\sqrt{n}(s^2 - \sigma_0^2)}{\sigma_0^2\sqrt{2}} \in N(0,1) \text{ (asimptotic)}$$

Regiunea de respingere va fi similara testului z

Testul z (ipoteze asupra proportiei)

- Testarea proporției (eșantioane mari, n>30 sau $np_0>10$ și $n(1-p_0)>10$)

$$a) H_0: p = p_0 \qquad b) H_0 : p = p_0 \qquad c) H_0 : p = p_0$$
$$H_A : p \neq p_0 \qquad H_A : p > p_0 \qquad H_A : p < p_0$$

$$T(X_1,....,X_n) = \frac{\sqrt{n}(p - p_0)}{\sqrt{p_0(1-p_0)}} \in N(0,1) \text{ (asimptotic)}$$

$$a) R = (-\infty, -z_{\alpha/2}) \bigcup (z_{\alpha/2}, \infty)$$

$$b) R = (z_\alpha, \infty)$$

$$c) R = (-\infty, -z_\alpha)$$

Testul z (ipoteze asupra proportiei)

Exemplu

- Context: se consideră că 26% din copii mamelor seropozitive sunt infectați. Cercetătorii consideră că riscul ca o mama seropozitivă să infecteze copilul este corelat cu starea infecției mamei.
- Intrebare: procentul copiilor infectati diferă de 26% în cazul în care mamele sunt în stare avansată de infecție ?

- Studiu: se studiază un eșantion de 150 copii născuți de către mame cu nivel mare de infecție; 107 dintre copii sunt seropozitivi (procent 71.33%).
- Intrebare: oferă aceste rezultate suficiente argumente pentru a spune că procentul copiilor seropozitivi născuți de către mame cu stare avansată a infecției este mai mare decât 26% ?

Testul z (ipoteze asupra proportiei)

Procedura de verificare a ipotezei statistice

- Se enunta ipotezele:

  H_0: p=0.26 H_A: p>0.26

- Se calculeaza statistica testului:

$$T(X_1, X_2, ..., X_n) = \frac{\sqrt{n}(p - p_0)}{\sqrt{p_0(1-p_0)}} = \frac{\sqrt{150}(0.7133 - 0.26)}{\sqrt{0.26(1-0.26)}} = 12.6569$$

- Se compara valoarea statisticii cu valoarea critica a repartitiei normale standard corespunzatoare nivelului de semnificatie alfa=0.05 (cazul testului unilateral)

  $T(X_1,\ldots,X_n)=12.6569>1.65$

- Decizia: ipoteza nulă se respinge (în cazul mamelor cu infecție avansată nu se acceptă ipoteza că procentul copiilor infectați este 0.26)

Teste pt compararea mediei a doua populatii

- Test pentru compararea mediilor a două populații cu dispersii egale (eșantioane de volum n_1 respectiv n_2)

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

$$T(X_1,....,X_n) = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{s\sqrt{1/n_1 + 1/n_2}} = \frac{(\overline{X_1} - \overline{X_2})}{s\sqrt{1/n_1 + 1/n_2}} \in t(n_1 + n_2 - 1)$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$ (dispersia corespunzatoare esantioanelor reunite)

$$R = (-\infty, -t_{n1+n2-1,\alpha/2}) \cup (t_{n1+n2-1,\alpha/2}, \infty)$$

- In cazul eșantioanelor mici este necesară ipoteza de normalitate a populațiilor

Teste pt compararea mediei a doua populatii

Exemple:

- Se poate utiliza pentru compararea nivelului mediu de exprimare a unei gene în două tipuri de celule (una normală și una de natură canceroasă).
- Nivelul de exprimare al unei gene este corelat cu “cantitatea” de proteine sintetizată pe baza genei respective.
- Măsurarea nivelului de expresie genică se realizează prin tehnici experimentale (de exemplu, “DNA microarray”).
- In cazul particular al caracteristicilor nominale este util sa se compare proportiile a două caracteristici

Teste pt compararea mediei a doua populatii

Exemplu: se consideră două seturi de pacienți cărora li s-au administrat substanțe diferite pentru anesteziere în vederea unei intervenții chirurgicale. Primul set conține 61 de pacienți și dintre aceștia 8 au decedat, iar al doilea set conține 67 de pacienți dintre care 10 au decedat. Are tipul de anestezic influență asupra ratei de deces?

Test pentru compararea proporției

$$H_0 : p_1 = p_2; \quad H_1 : p_1 \neq p_2$$

$$T(X_1, X_2, ..., X_n) = \frac{p_1 - p_2}{\sqrt{s_1^2 + s_2^2}} \in N(0,1) \quad (\text{if } n_1 \text{ and } n_2 \text{ are large enough})$$

$$s_1^2 = \frac{p_1(1-p_1)}{n_1}, \; s_2^2 = \frac{p_2(1-p_2)}{n_2}$$

$$R = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$$

Teste pt compararea dispersiei a doua populatii

- Test pentru compararea dispersiilor a două populații cu repartiție normală (testul F)

$$a) H_0 : \sigma_1 = \sigma_2$$

$$H_A : \sigma_1 \neq \sigma_2$$

$$T(X_1, ..., X_n) = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} \in F(n_1 - 1, n_2 - 1)$$

$$R = (0, F_{n1-1,n2-1,1-\alpha/2}) \bigcup (F_{n1-1,n2-1,\alpha/2}, \infty)$$

Verificarea ipotezelor statistice – P-valoare

Motivație: evită determinarea valorii critice (util în special în cazul repartițiilor discrete – vezi de exemplu repartiția binomială)

Principiu:

- Se calculează probabilitatea ca statistica asociată testului să aibă valoarea specificată în ipoteza nulă

- Daca probabilitatea calculată este mai mică decât nivelul de semnificație asociat testului (de exemplu 0.05) atunci ipoteza nulă este respinsă

Verificarea ipotezelor statistice – P-valoare

Exemplu: analiza potrivirilor dintre două secvențe ADN

Etape:

- Se calculează probabilitatea ca statistica asociată testului să aibă valoarea specificată în ipoteza nulă: potrivirile sunt întâmplătoare. In acest caz nr de potriviri ar avea o repartiție binomială cu p=0.25). Pt n=26 probabilitatea să se observe 11 potriviri ar fi: $C_{26}^{11} 0.25^{11} 0.75^{15} = 0.025$

- Daca probabilitatea calculata este mai mică decat nivelul de semnificatie asociat testului atunci ipoteza nula este respinsa. Pt. α=0.05 ipoteza nulă se respinge.

Concluzie: nu se poate afirma că potrivirile dintre cele două secvențe sunt pur întâmplătoare

Compararea a mai mult de doua populatii

Exemplu: analiza influenței dietei asupra stării de sănătate (boli cardio-vasculare)

Experiment:

- Se consideră 4 grupuri selectate aleator dintr-o populație:
  - G1: Alimentație normală (grup de control)
  - G2: Alimentație bazată pe paste făinoase
  - G3: Alimentație bazată pe carne
  - G4: Alimentație bazată pe fructe și legume

- Pentru membrii celor 4 grupuri se urmărește evoluția unei mărimi (ex: tensiunea arterială)

Intrebare: are tipul de alimentație o influență semnificativă asupra valorii tensiunii arteriale?

Compararea a mai mult de doua populatii

Abordare: analiza varianței (ANalysis Of Variance = ANOVA)

Are ca scop să decidă dacă diferențele observate între valorile corespunzătoare celor 4 grupuri sunt cauzate de tipurile de alimentație sau sunt efectul variației intrinseci dintre grupuri

H_0: nu există diferență între grupuri: $\mu_1 = \mu_2 = ... = \mu_k$

H_A: există diferență intre grupuri

Idee:

- Se calculează varianța în cadrul fiecărui grup și valoarea medie a acestor variante („within”)
- Se calculează varianța între mediile grupurilor („between”)
- In cazul în care nu ar exista diferențe între grupuri cele două variante ar fi apropiate ca valoare

Compararea a mai mult de doua populatii

Etape de calcul:

Estimari ale variantei "intra" si "inter" grupuri

$$s^2_{within} = \frac{1}{k}\sum_{i=1}^{k} s^2(G_i)$$

$$s^2_{between} = ns^2(\{\overline{G_1}, \overline{G_2}, ..., \overline{G_k}\})$$

Statistica (are distributia Fisher) :

$$\frac{s^2_{between}}{s^2_{within}} \in F(k-1, n-k)$$

Regiunea de respingere (pt ipoteza alternativa bilaterala)

$$R = (0, F_{k-1,n-k,1-\alpha/2}) \bigcup (F_{k-1,n-k,\alpha/2}, \infty)$$

Notatii : k = nr grupuri;

n = dimensiune esantion (nr elemente in fiecare grup)

Compararea a mai mult de doua populatii

Obs:

- ANOVA permite doar să se decidă dacă există diferențe semnificative între grupuri dar nu indică care dintre ele sunt diferite

- Pt a obține această informație (în cazul în care ipoteza nulă este respinsă) trebuie aplicate ulterior teste de comparare la nivel de perechi
  - Se compară toate perechile de grupuri aplicând de k(k-1)/2 testul t
  - Se compară toate grupurile față de grupul de control aplicând de k-1 ori testul t
  - Obs: deciziile în cadrul testului t se iau folosind corecții ale valorii nivelului de semnificație, alpha. De exemplu în cazul a k grupuri nivelul de semnificație utilizat pentru o comparație trebuie să fie alpha/k (corecția Bonferroni)

- Obs: exerciții cu ANOVA în lab 2

Curs urmator:

- Teste neparametrice
- Regresie