

# Curs 1

Cristian Niculescu

## 1 Introducere

Bibliografie:

Jeremy Orloff, Jonathan Bloom, *Introduction to Probability and Statistics*,  
MIT OpenCourseWare, 2014

Cristian Niculescu, *Probabilități și statistică*, Editura Universității din București, 2015

### 1.1 Probabilități vs. statistică

Probabilitățile și statistica sunt strâns legate deoarece toate afirmațiile statistice sunt la bază afirmații despre probabilități. În ciuda acestui fapt, cele 2 domenii se simt uneori ca subiecte foarte diferite. Probabilitățile sunt independente logic; sunt câteva reguli și toate răspunsurile rezultă logic din aceste reguli, cu toate că pot fi calcule complicate. În statistică aplicăm probabilitățile pentru a trage concluzii din date.

#### Exemplu probabilistic

Avem o monedă corectă (probabilități egale pentru avers sau revers). O aruncăm de 100 de ori. Care este probabilitatea a cel puțin 60 de aversuri? Există un singur răspuns (aproximativ 0.02844397) și vom învăța cum să-l calculăm.

#### Exemplu statistic

Avem o monedă de proveniență necunoscută. Pentru a investiga dacă este corectă o aruncăm de 100 de ori și numărăm aversurile. Să spunem că numărăm 60 de aversuri. Sarcina noastră ca statisticieni este să tragem o concluzie (inferență) din aceste date. Sunt multe moduri de a proceda, atât în termeni de forma pe care o ia concluzia, cât și în ce privește calculele probabilităților folosite pentru a justifica concluzia. De fapt, statisticieni diferiți pot trage concluzii diferite.

Observăm că în primul exemplu procesul aleator este deplin cunoscut (probabilitatea aversului = 0.5). Obiectivul este să găsim probabilitatea unui anume rezultat (cel puțin 60 de aversuri) provenind din procesul aleator. În

al 2-lea exemplu, rezultatul este cunoscut (60 de aversuri) și obiectivul este să clarificăm procesul aleator necunoscut (probabilitatea aversului).

## 1.2 Interpretări frecvenționiste vs. Bayesiene

Există 2 școli proeminente și uneori conflictuale de statistică: [Bayesiană](#) și [frecvenționistă](#). Abordările lor sunt bazate pe interpretări diferite ale sensului probabilității.

**Frecvenționiștii** spun că probabilitatea măsoară [frecvența diferitelor rezultate ale unui experiment](#). De exemplu, spunând că o monedă corectă are o probabilitate de 50% a aversurilor înseamnă că, dacă o vom arunca de multe ori, atunci ne așteptăm ca aproximativ jumătate din aruncări să fie aversuri. **Bayesienii** spun că probabilitatea este un concept abstract care măsoară [o stare de cunoaștere sau un grad de încredere](#) într-o propoziție dată. În practică, Bayesienii nu atribuie o singură valoare pentru probabilitatea aversului unei monede. Mai degrabă ei consideră un domeniu de valori, fiecare cu propria probabilitate de a fi adevărată.

Abordarea frecvenționistă a fost mult timp dominantă în domeniile ca biologia, medicina, sănătatea publică și științele sociale. Abordarea Bayesiană s-a bucurat de o renaștere în era computerelor performante și datelor mari. Este utilă în special când se încorporează date noi într-un model statistic existent, de exemplu, când se antrenează un sistem de recunoaștere a vorbirii sau a feței. Astăzi, statisticienii creează instrumente puternice folosind ambele abordări în moduri complementare.

## 1.3 Aplicații, modele didactice și simulare

Probabilitățile și statistica sunt larg folosite în științele naturii, inginerie, medicină, științele sociale, economie și informatică. Lista aplicațiilor este mare: teste ale unui tratament medical contra altuia (sau a unui placebo), măsuri ale legăturii genetice, căutarea particulelor elementare, învățarea automată pentru vedere sau vorbire, strategii și probabilități de jocuri de noroc, meteorologie, prognoză economică, epidemiologie, marketing, căutare pe net...

Date fiind atât de multe aplicații captivante, poate vă mirați de ce vom consuma atât de mult timp gândindu-ne la [modele didactice](#) ca monede și zaruri. Înțelegându-le pe acestea cu desăvârșire, vom dezvolta un bun simț pentru fondul simplu interior multor probleme complexe din lumea reală. De fapt, modesta monedă este un model realist pentru orice situații cu 2 rezultate posibile: succes sau eșec al unui tratament, motor de avion, pariu, sau chiar al unui curs.

Uneori o problemă este atât de complicată încât cel mai bun mod de înțelegere a ei este prin simulare pe computer. Aici folosim software pentru a face experimente *virtuale* de multe ori pentru a estima probabilități. La acest curs vom folosi R pentru simulare, calcul și vizualizare.

## 2 Numărare și mulțimi

### 2.1 Scopurile învățării

1. Să știe definițiile și notațiile pentru mulțimi, intersecție, reuniune, complementară.
2. Să poată să vizualizeze operațiile cu mulțimi folosind diagrame Venn.
3. Să înțeleagă cum este folosită numărarea la calculul probabilităților.
4. Să poată folosi regula produsului, principiul incluzerii și al excluderii, permutările, aranjamentele și combinările pentru a calcula numărul elementelor unei mulțimi.

### 2.2 Numărare

#### 2.2.1 Întrebări motivante

**Exemplul 1.** O monedă este *corectă* dacă, aruncând-o, obținem avers sau revers cu aceeași probabilitate. Aruncăm o monedă de 3 ori. Care este probabilitatea să obținem exact un avers?

**Răspuns:** Cu 3 aruncări, putem lista ușor cele 8 **cazuri** posibile (avers =  $H$ , revers =  $T$ ):

$$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

3 din aceste cazuri au exact 1 avers:

$$\{HTT, THT, TTH\}.$$

Deoarece toate cazurile sunt egal probabile, avem

$$P(1 \text{ avers în } 3 \text{ aruncări}) = \frac{\text{numărul de cazuri cu 1 avers}}{\text{numărul total de cazuri}} = \frac{3}{8}.$$

**Gândiți:** Ar fi practică listarea cazurilor la 10 aruncări?

Un pachet de 52 de cărți de joc are 13 **ranguri** (2,3,...,9,10,J,Q,K,A) și 4 **culori** (, , , ). O mână de poker constă în 5 cărți. *O pereche* este o mână care constă în 2 cărți având un rang și 3 cărți având alte 3 ranguri, de

exemplu,  $\{3\clubsuit, 3\diamondsuit, 5\heartsuit, 8\spadesuit, 10\spadesuit\}$ .

**Testați-vă intuiția:** probabilitatea unei perechi este:

- (a) sub 5%
- (b) între 5% și 10%
- (c) între 10% și 20%
- (d) între 20% și 40%
- (e) peste 40%.

În acest moment putem doar să ghicim această probabilitate. Unul din scopurile noastre este să învățăm cum s-o calculăm exact. Pentru a începe, observăm că, deoarece fiecare mulțime de 5 cărți este **egal probabilă**, putem calcula probabilitatea unei perechi astfel

$$P(\text{o pereche}) = \frac{\text{numărul de "o pereche"}}{\text{numărul total de mâini}}.$$

Deci, pentru a afla probabilitatea exactă, avem nevoie să calculăm numărul de elemente din fiecare din aceste mulțimi. și trebuie s-o facem îscusit, deoarece sunt prea multe elemente pentru a le lista pe toate.

Am observat deja de câteva ori că toate cazurile posibile erau egal probabile și am folosit asta pentru a afla o probabilitate numărând. Afirmăm aceasta riguros în următorul principiu.

**Principiu:** Presupunem că sunt  $n$  cazuri posibile pentru un experiment și fiecare caz este egal probabil. Dacă sunt  $k$  cazuri favorabile, atunci probabilitatea unui caz favorabil este  $k/n$ .

Vă puteți gândi la un scenariu în care cazurile posibile nu sunt egal probabile?

Iată un scenariu: la un examen, puteți lua orice notă de la 1 la 10. Sunt 10 cazuri. Este probabilitatea de a lua sub 5 egală cu  $4/10$ ?

### 2.2.2 Mulțimi și notații

#### Definiții

O **mulțime**  $S$  este o colecție de elemente.

**Element:** Scriem  $x \in S$  pentru faptul că elementul  $x$  este în mulțimea  $S$ .

**Submulțime:** Spunem că mulțimea  $A$  este submulțime a lui  $S$  dacă toate elementele lui  $A$  sunt în  $S$ . Scriem  $A \subset S$ .

**Complementară:** Complementara lui  $A$  față de  $S$  este mulțimea elementelor lui  $S$  care **nu** sunt în  $A$ . Scriem  $A^c$ ,  $C_A$ ,  $\bar{A}$  sau  $S \setminus A$ .

**Reuniune:** Reuniunea lui  $A$  și  $B$  este mulțimea tuturor elementelor din  $A$  sau  $B$  (sau din ambele). Scriem  $A \cup B$ .

**Intersecție:** Intersecția lui  $A$  și  $B$  este mulțimea tuturor elementelor din  $A$  și

B. Scriem  $A \cap B$ .

**Mulțimea vidă:** Mulțimea vidă este mulțimea care nu are niciun element. O notăm cu  $\emptyset$ .

**Disjuncte:**  $A$  și  $B$  sunt **disjuncte** dacă n-au niciun element comun. Adică  $A \cap B = \emptyset$ .

**Diferență:** Diferența lui  $A$  și  $B$  este mulțimea elementelor din  $A$  care nu sunt în  $B$ . Scriem  $A \setminus B$  sau  $A - B$ .

**Exemplul 2.** Plecăm cu o mulțime de 10 animale

$$S = \{\text{antilopă, albină, pisică, câine, elefant, broască, Tânăr, hienă, iguană, jaguar}\}.$$

Considerăm 2 submulțimi:

$$M = \{\text{antilopă, pisică, câine, elefant, hienă, jaguar}\}$$

$$W = \{\text{antilopă, albină, elefant, broască, Tânăr, hienă, iguană, jaguar}\}.$$

Scopul nostru aici este să cercetăm diferite operații cu mulțimi.

**Intersecție:**  $M \cap W = \{\text{antilopă, elefant, hienă, jaguar}\}$ .

**Reuniune:**

$$M \cup W = \{\text{antilopă, albină, pisică, câine, elefant, broască, Tânăr, hienă, iguană, jaguar}\}.$$

**Complementară:**  $M^c = \{\text{albină, broască, Tânăr, iguană}\}$ .

**Diferență:**  $M \setminus W = \{\text{pisică, câine}\}$ .

Adesea sunt multe moduri de a obține aceeași mulțime, de exemplu,  $M^c = S \setminus M$ ,  $M \setminus W = M \cap C_W$ .

Relațiile dintre reuniune, intersecție și complementară sunt date de **legile lui DeMorgan**:

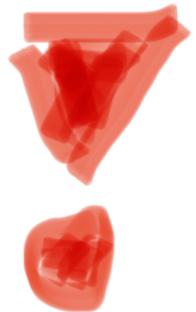
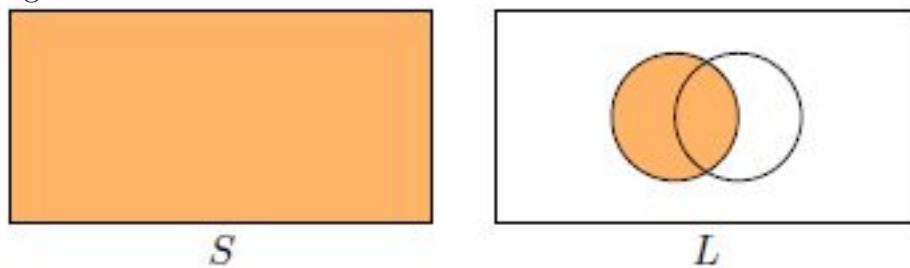
$$(A \cup B)^c = A^c \cap B^c$$

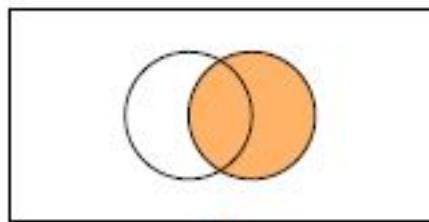
$$(A \cap B)^c = A^c \cup B^c.$$

### Diagrame Venn-Euler

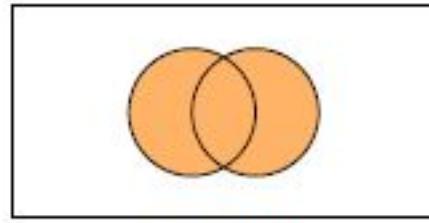
**Diagramele Venn-Euler** dau un mod ușor de a vizualiza operațiile cu mulțimi.

În toate figurile,  $S$  este regiunea din interiorul dreptunghiului mare,  $L$  este regiunea din interiorul cercului din stânga și  $R$  este regiunea din interiorul cercului din dreapta. Regiunea colorată arată mulțimea scrisă sub fiecare figură.

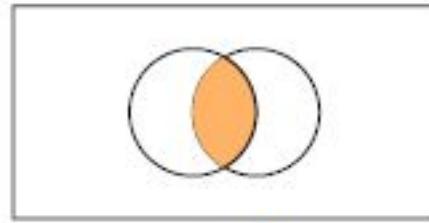




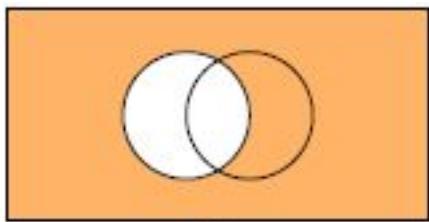
$R$



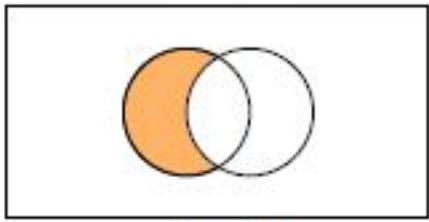
$L \cup R$



$L \cap R$

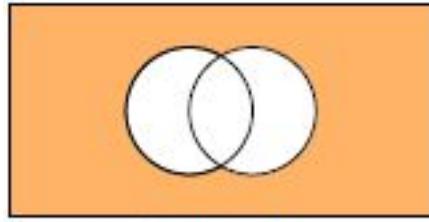


$L^c$

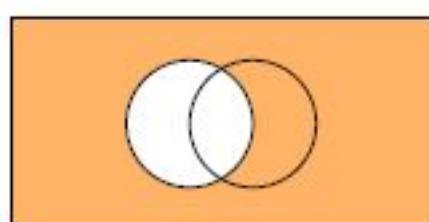


$L - R$

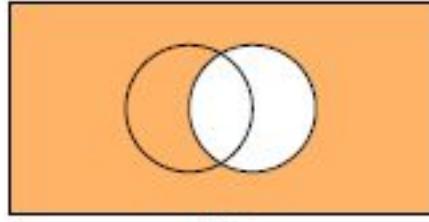
Demonstrația legilor lui DeMorgan



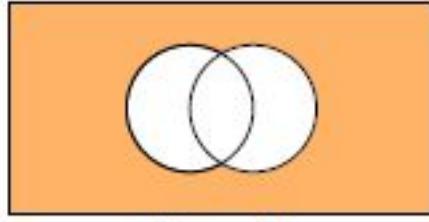
$(L \cup R)^c$



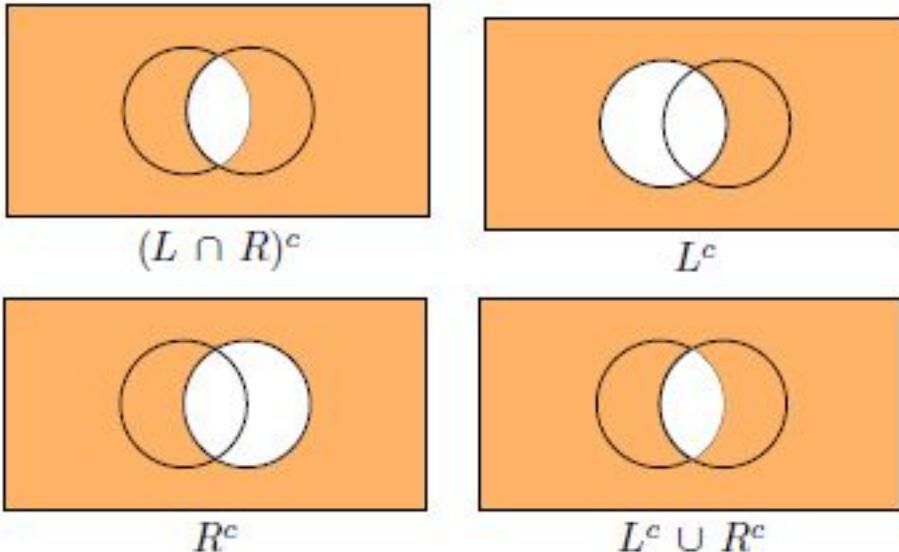
$L^c$



$R^c$



$L^c \cap R^c$



**Exemplul 3.** Verificați legile lui DeMorgan pentru submulțimile  $A = \{1, 2, 3\}$  și  $B = \{3, 4\}$  ale mulțimii  $S = \{1, 2, 3, 4, 5\}$ .

**Răspuns:** Pentru fiecare lege, calculăm ambii membri ai relației și arătăm că sunt egali.

$$1. (A \cup B)^c = A^c \cap B^c :$$

Membrul stâng:  $A \cup B = \{1, 2, 3, 4\} \Rightarrow (A \cup B)^c = \{5\}$ .

Membrul drept:  $A^c = \{4, 5\}, B^c = \{1, 2, 5\} \Rightarrow A^c \cap B^c = \{5\}$ .

Cei 2 membri sunt egali, q.e.d.

$$2. (A \cap B)^c = A^c \cup B^c :$$

Membrul stâng:  $A \cap B = \{3\} \Rightarrow (A \cap B)^c = \{1, 2, 4, 5\}$ .

Membrul drept:  $A^c = \{4, 5\}, B^c = \{1, 2, 5\} \Rightarrow A^c \cup B^c = \{1, 2, 4, 5\}$ .

Cei 2 membri sunt egali, q.e.d.

### Produsul cartezian

Produsul cartezian al mulțimilor  $S$  și  $T$  este mulțimea perechilor ordonate:

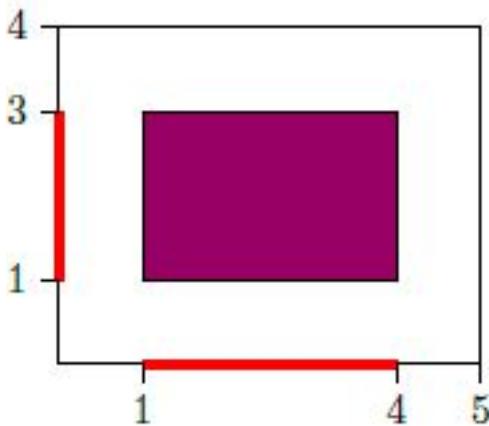
$$S \times T = \{(s, t) | s \in S, t \in T\}.$$

În cuvinte, membrul drept se citește "mulțimea perechilor ordonate  $(s, t)$  cu proprietatea că  $s$  este în  $S$  și  $t$  este în  $T$ ".

Următoarele diagrame arată 2 exemple de produs cartezian.

$\times$	1	2	3	4
1	(1,1)	(1,2)	(1,3)	(1,4)
2	(2,1)	(2,2)	(2,3)	(2,4)
3	(3,1)	(3,2)	(3,3)	(3,4)

$$\{1, 2, 3\} \times \{1, 2, 3, 4\}$$



$$[1, 4] \times [1, 3] \subset [0, 5] \times [0, 4]$$

Ultima figură ilustrează de asemenea faptul că dacă  $A \subset S$  și  $B \subset T$ , atunci  $A \times B \subset S \times T$ .

### 2.2.3 Numărare

Dacă  $S$  este finită, folosim  $|S|$  sau  $\#S$  pentru a nota **numărul elementelor lui  $S$**  (cardinalul lui  $S$ ).

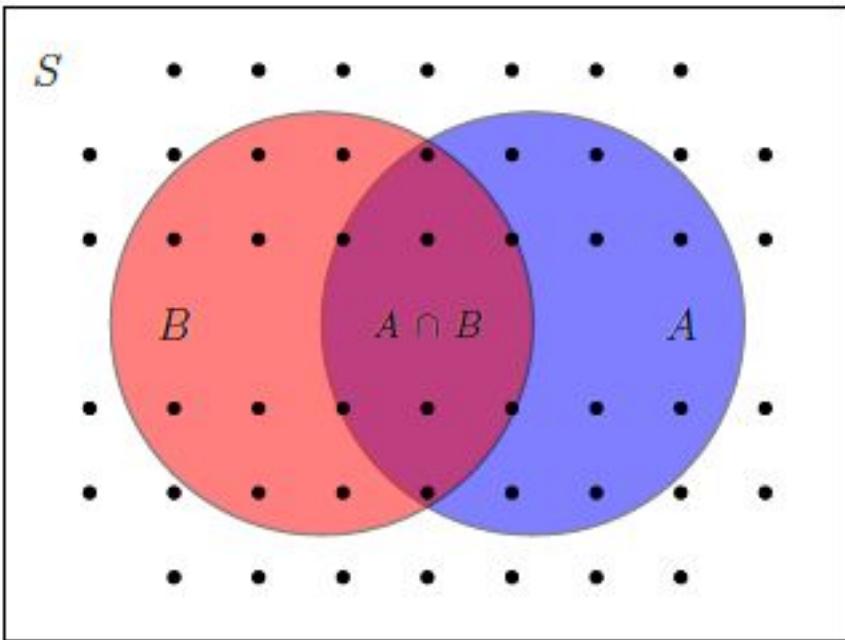
2 principii de numărare utile sunt *principiul includerii și excluderii* și *regula produsului*.

#### Principiul includerii și excluderii

**Principiul includerii și excluderii** spune

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Putem ilustra asta printr-o diagramă Venn-Euler.  $S$  este mulțimea tuturor punctelor,  $A$  este mulțimea punctelor din cercul albastru, iar  $B$  este mulțimea punctelor din cercul roșu.



$|A|$  este numărul de puncte din  $A$  și analog pentru celelalte mulțimi. Figura arată că  $|A| + |B|$  numără *de 2 ori*  $|A \cap B|$ , de aceea  $|A \cap B|$  este scăzut în formula includerii și excluderii.

**Exemplul 4.** Într-o trupă de cântăreți și chitariști sunt 7 cântăreți, 4 chitariști și 2 ambele. Cât de mare este trupa?

**Răspuns:** Fie  $S$  mulțimea cântăreților și  $G$  mulțimea chitariștilor. Prinzipiul includerii și excluderii spune că

$$\text{mărimea trupei} = |S \cup G| = |S| + |G| - |S \cap G| = 7 + 4 - 2 = 9.$$

### Regula produsului

**Regula produsului** spune:

Dacă există  $n$  moduri de a face acțiunea 1 și apoi  $m$  moduri de a face acțiunea 2, atunci există  $n \cdot m$  moduri de a face acțiunea 1 urmată de acțiunea 2.

O numim de asemenea regula **multiplicării**.

**Exemplul 5.** Dacă avem 3 cămași și 4 pantaloni, atunci putem face  $3 \cdot 4 = 12$  costumații.

**Gândiți:** Un punct extrem de important este că regula produsului are loc chiar dacă modurile de a face acțiunea 2 depind de acțiunea 1, atât timp cât *numărul* de moduri de a face acțiunea 2 este independent de acțiunea 1. Pentru a ilustra asta:

**Exemplul 6.** Există 5 competitori în finala de 100m la olimpiadă. În câte moduri pot fi date medaliile de aur, argint și bronz? (Presupunem că

egalitățile sunt excluse.)

**Răspuns:** Există 5 moduri de a da medalia de aur. Odată ce aceasta este dată, există 4 moduri de a da medalia de argint și apoi 3 moduri de a da medalia de bronz. Răspuns  $5 \cdot 4 \cdot 3 = 60$  de moduri.

Observăm că alegerea medaliatului cu aur afectează câștigătorul argintului, dar numărul posibililor medaliați cu argint este totdeauna 4.

#### 2.2.4 Permutări, aranjamente și combinări

##### Permutări și aranjamente

O **permutare** a unei mulțimi este o ordonare particulară a elementelor sale. De exemplu, mulțimea  $\{a, b, c\}$  are 6 permutări:  $abc, acb, bac, bca, cab, cba$ . Am aflat numărul permutărilor listându-le pe toate. Puteam de asemenea să aflăm numărul permutărilor folosind regula produsului. Adică, există 3 moduri de a alege primul element, apoi 2 moduri pentru al 2-lea și 1 pentru al 3-lea. Asta dă un total de  $3 \cdot 2 \cdot 1 = 6$  permutări.

În general, regula produsului ne spune că numărul de permutări ale unei mulțimi de  $k$  elemente este

$$k! = k \cdot (k - 1) \cdot \dots \cdot 3 \cdot 2 \cdot 1.$$

Vorbim de asemenea de aranjamentele a  $k$  elemente dintr-o mulțime cu  $n$  elemente. Arătăm ce înseamnă asta cu un exemplu.

**Exemplul 7.** Listați toate aranjamentele de 3 elemente ale mulțimii  $\{a, b, c, d\}$ .

**Răspuns:** Este o listă mai lungă,

$abc$	$acb$	$bac$	$bca$	$cab$	$cba$
$abd$	$adb$	$bad$	$bda$	$dab$	$dba$
$acd$	$adc$	$cad$	$cda$	$dac$	$dca$
$bcd$	$bdc$	$cbd$	$cdb$	$dbc$	$dcb$

Observăm că  $abc$  și  $acb$  se numără ca aranjamente distințe. Adică, **pentru aranjamente ordinea contează**.

Există 24 de aranjamente. Observăm că regula produsului ne-ar fi spus că sunt  $4 \cdot 3 \cdot 2 = 24$  permutări fără a ne mai deranja să le listăm.

##### Combinări

În contrast cu aranjamentele, **la combinări ordinea nu contează**: **aranjamentele sunt liste, iar combinările sunt mulțimi**. Arătăm ce vrem să spunem printr-un exemplu.

**Exemplul 8.** Listați toate combinările de 3 elemente din mulțimea  $\{a, b, c, d\}$ .

**Răspuns:** O astfel de combinare este o colecție de 3 elemente fără a consi-

dera ordinea. Astfel,  $abc$  și  $cab$  reprezintă ambele aceeași combinare. Putem lista toate combinările listând toate submulțimile de exact 3 elemente.

$$\{a, b, c\} \quad \{a, b, d\} \quad \{a, c, d\} \quad \{b, c, d\}.$$

Există doar 4 combinări, în contrast cu cele 24 de aranjamente din exemplul anterior. Factorul 6 apare deoarece fiecare combinare de 3 elemente poate fi scrisă în 6 ordini diferite.

### Formule

Folosim următoarele notații.

$A_n^k =$  numărul aranjamentelor (listelor) de  $k$  elemente distințe dintr-o mulțime cu  $n$  elemente.

$C_n^k = \binom{n}{k} =$  numărul combinărilor (submulțimilor) de  $k$  elemente dintr-o mulțime cu  $n$  elemente.

Subliniem că prin numărul de combinări de  $k$  elemente înțelegem numărul de submulțimi de  $k$  elemente.

ACEstea au următoarele notații și formule:

$$\text{Aranjamente: } A_n^k = \frac{n!}{(n-k)!} = n(n-1)\dots(n-k+1).$$

$$\text{Combinări: } C_n^k = \frac{n!}{k!(n-k)!} = \frac{A_n^k}{k!}.$$

Notația  $A_n^k$  se citește "aranjamente de  $n$  luate câte  $k$ ". Notația  $C_n^k$  se citește "combinări de  $n$  luate câte  $k$ ". Formula pentru  $A_n^k$  rezultă din regula produsului. Ea implică de asemenea formula pentru  $C_n^k$ , deoarece o submulțime de  $k$  elemente poate fi ordonată în  $k!$  moduri.

Putem ilustra relația dintre aranjamente și combinări aliniind rezultatele din cele 2 exemple anterioare.

$abc$	$acb$	$bac$	$bca$	$cab$	$cba$	$\{a, b, c\}$
$abd$	$adb$	$bad$	$bda$	$dab$	$dba$	$\{a, b, d\}$
$acd$	$adc$	$cad$	$cda$	$dac$	$dca$	$\{a, c, d\}$
$bcd$	$bdc$	$cbd$	$cdb$	$dbc$	$dcb$	$\{b, c, d\}$

Aranjamente:  $A_4^3$

Combinări:  $C_4^3$

Observăm că fiecare linie din lista de aranjamente constă din toate cele  $3!$  permutări ale mulțimii corespunzătoare din lista de combinări.

### Exemple

**Exemplul 9.** Calculați:

(i) Numărul de moduri de a alege 2 din 4 obiecte (ordinea nu contează).

(ii) Numărul de moduri de a lista 2 din 4 obiecte.

(iii) Numărul de moduri de a alege 3 din 10 obiecte.

**Răspuns:** (i)  $C_4^2 = \frac{4!}{2! \cdot 2!} = 6$ .

(ii)  $A_4^2 = \frac{4!}{2!} = 12$ .

(iii)  $C_{10}^3 = \frac{10!}{3! \cdot 7!} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$ .

**Exemplul 10.** (i) Calculați numărul de moduri de a obține 3 aversuri într-o secvență de 10 aruncări ale unei monede.

(ii) Dacă moneda este corectă, care este probabilitatea a exact 3 aversuri din 10 aruncări?

**Răspuns:** (i) Se cere numărul de secvențe de 10 aruncări (aversuri sau reversuri) cu exact 3 aversuri. Adică, trebuie să alegem exact 3 din 10 aruncări să fie aversuri. Asta este aceeași întrebare ca ultima din exemplul precedent.

$$C_{10}^3 = \frac{10!}{3! \cdot 7!} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120.$$

(ii) Fiecare aruncare are 2 cazuri posibile (avers sau revers). Astfel, regula produsului spune că există  $2^{10} = 1024$  secvențe de 10 aruncări. Deoarece moneda este corectă, fiecare secvență este egal probabilă. Deci probabilitatea a 3 aversuri este

$$\frac{120}{1024} = 0.1171875.$$

# Curs 2

Cristian Niculescu

## 1 Probabilitate: terminologie și exemple

### 1.1 Scopurile învățării

1. Să știe definițiile spațiului probelor, evenimentului și probabilității.
2. Să poată organiza un scenariu cu caracter aleator într-un experiment și spațiu al probelor.
3. Să poată face calcule de bază folosind o probabilitate.

### 1.2 Terminologie

#### 1.2.1 Lista de distribuție a probabilității

**Experiment:** o procedură repetabilă cu cazuri (rezultate) posibile bine definite.

**Spațiul probelor:** mulțimea tuturor cazurilor posibile. De obicei notăm spațiul probelor cu  $\Omega$ , uneori cu  $S$ .

**Eveniment:** o submulțime a spațiului probelor.

**Funcție de probabilitate:** o funcție dând probabilitatea pentru fiecare caz.

#### 1.2.2 Exemple simple

**Exemplul 1.** Aruncarea unei monede corecte.

**Experiment:** aruncăm moneda, raportăm dacă aterizează avers sau revers.

**Spațiul probelor:**  $\Omega = \{H, T\}$ .

**Funcție de probabilitate:**  $P(H) = 0.5$ ,  $P(T) = 0.5$ .

**Exemplul 2.** Aruncarea de 3 ori a unei monede corecte.

**Experiment:** aruncăm moneda de 3 ori, listăm rezultatele.

**Spațiul probelor:**  $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ .

**Funcție de probabilitate:** fiecare caz este egal probabil, cu probabilitatea  $1/8$ .

Pentru spații mici de probe putem pune mulțimea cazurilor și probabilitățile într-un **tabel de probabilitate**.

Caz	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Probabilitate	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

**Exemplul 3.** Măsurarea masei unui proton

**Experiment:** se urmează o procedură definită pentru a măsura masa și a raporta rezultatul.

**Spațiul probelor:**  $\Omega = [0, \infty)$ , i.e. în principiu putem obține orice valoare pozitivă.

**Funcție de probabilitate:** deoarece există un continuum de cazuri posibile, nu există o funcție de probabilitate. În locul ei avem nevoie să folosim o *densitate de probabilitate*.

**Exemplul 4.** Taxiuri (Un spațiu de probe discret infinit)

**Experiment:** se numără taxiurile care trec pe lângă facultate pe durata cursului.

**Spațiul probelor:**  $\Omega = \mathbb{N}$ .

Acesta este modelat adesea cu următoarea funcție de probabilitate cunoscută ca repartiția (distribuția) Poisson:

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!},$$

unde  $\lambda$  este numărul mediu de taxiuri. Putem pune asta într-un tabel:

Caz	0	1	2	3	...	k	...
Probabilitate	$e^{-\lambda}$	$e^{-\lambda}\lambda$	$e^{-\lambda}\lambda^2/2$	$e^{-\lambda}\lambda^3/3!$	...	$e^{-\lambda}\lambda^k/k!$	...

**Întrebare:** Acceptând că aceasta este o funcție de probabilitate validă, cât este  $\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!}$ ?

**Răspuns:** Aceasta este probabilitatea totală a tuturor cazurilor posibile, deci suma este 1. (Observăm că aceasta rezultă de asemenea din seria Taylor  $e^\lambda = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}$ .)

Într-o schemă dată poate fi mai mult de o alegere rezonabilă a spațiului probelor. Iată un exemplu simplu.

**Exemplul 5.** 2 zaruri corecte (Alegerea spațiului probelor)

Presupunem că aruncăm un zar corect. Atunci spațiul probelor și funcția de probabilitate sunt

Caz	1	2	3	4	5	6
Probabilitate	1/6	1/6	1/6	1/6	1/6	1/6

Acum presupunem că aruncăm 2 zaruri corecte. Care ar trebui să fie spațiul probelor? Iată 2 opțiuni.

- Înregistrăm perechea numerelor de pe zaruri (primul zar, al 2-lea zar).
  - Înregistrăm suma numerelor de pe zaruri. În acest caz sunt 11 cazuri  $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ . Aceste cazuri nu sunt toate egal probabile.
- Ca mai sus, putem pune această informație în tabele. Pentru prima opțiune, spațiul probelor este produsul cartezian al spațiilor probelor pentru fiecare zar

$$\{(1, 1), (1, 2), (1, 3), \dots, (6, 6)\}.$$

Fiecare din cele 36 de cazuri este egal probabil. (De ce 36 de cazuri?) Pentru funcția de probabilitate vom face un tabel 2 dimensional cu liniile corespunzând numerelor de pe primul zar, coloanele numerelor de pe al 2-lea zar, iar în tabel sunt probabilitățile.

		Zar 2					
		1	2	3	4	5	6
Zar 1	1	1/36	1/36	1/36	1/36	1/36	1/36
	2	1/36	1/36	1/36	1/36	1/36	1/36
	3	1/36	1/36	1/36	1/36	1/36	1/36
	4	1/36	1/36	1/36	1/36	1/36	1/36
	5	1/36	1/36	1/36	1/36	1/36	1/36
	6	1/36	1/36	1/36	1/36	1/36	1/36
		2 zaruri corecte într-un tabel 2-dimensional					

În a 2-a opțiune prezentăm cazurile și probabilitățile în tabelul nostru usual.

Caz	2	3	4	5	6	7	8	9	10	11	12
Probabilitate	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Suma a 2 zaruri corecte

**Gândiți:** Care este relația dintre cele 2 tabele de probabilități de mai sus? Vom vedea că cea mai bună alegere a spațiului probelor depinde de context. Deocamdată observăm că, dacă dăm rezultatul ca o pereche de numere, este ușor să aflăm suma.

**Observație.** Listarea experimentului, spațiului probelor și a funcției de probabilitate este un bun mod de a începe lucrul sistematic cu probabilitatea. Vă poate ajuta să evitați unele din capcanele obișnuite din subiect.

## Evenimente

Un **eveniment** este o colecție de cazuri, i.e. un eveniment este o submulțime a spațiului probelor  $\Omega$ . Asta sună ciudat, dar chiar corespunde înțelesului obișnuit al cuvântului.

**Exemplul 6.** Folosind schema din exemplul 2 vom descrie evenimentul să obținem exact 2 aversuri în cuvinte prin  $E = \text{"exact 2 aversuri"}$ . Scris ca o

submulțime, acesta devine

$$E = \{HHT, HTH, THH\}.$$

Ar trebui să mișcați confortabil între descrierea evenimentelor în cuvinte și ca submulțimi ale spațiului probelor.

Probabilitatea unui eveniment  $E$  este calculată adunând probabilitățile tuturor cazurilor din  $E$ . În acest exemplu fiecare caz are probabilitatea  $1/8$ , deci avem  $P(E) = 3/8$ .

### 1.2.3 Definiția unui spațiu al probelor discret

**Definiție.** Un **spațiu al probelor discret** este unul care este cel mult numărabil (listabil), putând fi ori finit ori infinit.

**Exemplu.**  $\{H, T\}$ ,  $\{1, 2, 3\}$ ,  $\mathbb{N}^*$ ,  $\{2, 3, 5, 7, 11, 13, 17, \dots\}$  sunt toate mulțimi discrete. Primele 2 sunt finite și ultimele 2 sunt infinite.

**Exemplu.** Intervalul  $[0,1]$  nu este discret, ci *continuu*.

### 1.2.4 Funcția de probabilitate

Până acum am folosit o definiție ocasională a funcției de probabilitate. Să dăm una mai precisă.

#### Definiția funcției de probabilitate

Pentru un spațiu al probelor discret  $S$ , o **funcție de probabilitate**  $P$  atribuie fiecărui caz  $\omega$  un număr  $P(\omega)$  numit probabilitatea lui  $\omega$ .  $P$  trebuie să satisfacă 2 reguli:

Regula 1.  $0 \leq P(\omega) \leq 1$  (probabilitățile sunt între 0 și 1).

Regula 2. Suma probabilităților tuturor cazurilor posibile este 1 (ceva trebuie să aibă loc).

În simboluri, regula 2 spune: dacă  $S = \{\omega_1, \omega_2, \dots, \omega_n\}$ , atunci  $P(\omega_1) + P(\omega_2) + \dots + P(\omega_n) = 1$ . Sau  $\sum_{j=1}^n P(\omega_j) = 1$ .

Probabilitatea unui eveniment  $E$  este suma probabilităților tuturor cazurilor din  $E$ . Adică,

$$P(E) = \sum_{\omega \in E} P(\omega).$$

**Gândiți:** Verificați regulile 1 și 2 pe exemplele 1 și 2 de mai sus.

**Exemplul 7.** Aruncarea până la primul avers (un exemplu clasic)

Presupunem că avem o monedă cu probabilitatea  $p$  a aversurilor și avem următorul scenariu.

**Experiment:** Aruncăm moneda până apare primul avers. Raportăm numărul aruncărilor.

# De făcut!

Spațiul probelor:  $\Omega = \mathbb{N}^*$ .

Funcția de probabilitate:  $P(n) = (1 - p)^{n-1}p$ .

**Provocarea 1:** arătați că suma tuturor probabilităților este 1 (indicație: serii geometrice).

**Provocarea 2:** justificați formula pentru  $P(n)$ .

**Probleme de oprire.** Exemplul didactic anterior este o versiune a unei clase generale de probleme numite **probleme de regula opririi**. O regulă de oprire este o regulă care ne spune când să terminăm un anumit proces. În exemplul didactic de mai sus procesul a fost aruncarea unei monezi și l-am oprit după primul avers. Un exemplu mai practic este o regulă pentru terminarea unei serii de tratamente medicale. O astfel de regulă ar putea depinde de cât de bine funcționează tratamentele, cum le tolerează pacientul și probabilitatea să fie eficace continuarea tratamentelor. Ne putem informa despre probabilitatea opririi într-un anumit număr de tratamente sau numărul mediu de tratamente la care să ne așteptăm înaintea opririi.

## 1.3 Câteva reguli ale probabilității

Pentru evenimentele  $A$ ,  $L$  și  $R$  incluse într-un spațiu al probelor  $\Omega$ :

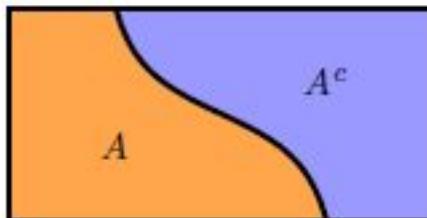
**Regula 1.**  $P(A^c) = 1 - P(A)$ .

**Regula 2.** Dacă  $L$  și  $R$  sunt disjuncte, atunci  $P(L \cup R) = P(L) + P(R)$ .

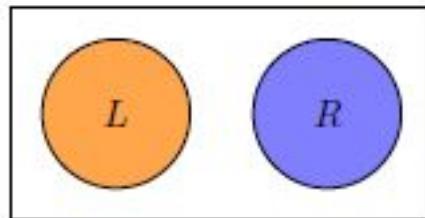
**Regula 3.** Dacă  $L$  și  $R$  nu sunt disjuncte, avem **principiul includerii și excluderii**:

$$P(L \cup R) = P(L) + P(R) - P(L \cap R).$$

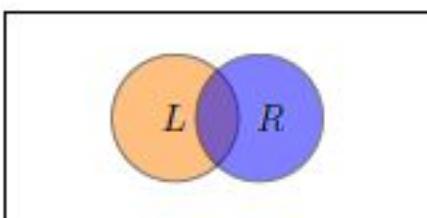
Vizualizăm aceste reguli folosind diagrame Venn-Euler.



$\Omega = A \cup A^c$ , nicio suprapunere



$L \cup R$ , nicio suprapunere



$L \cup R$ , suprapunerea =  $L \cap R$

Le putem de asemenea justifica logic.

Regula 1:  $A$  și  $A^c$  împart  $\Omega$  în 2 regiuni disjuncte. Deoarece probabilitatea totală  $P(\Omega) = 1$ , această regulă spune că probabilitatea lui  $A$  și probabilitatea lui "non  $A$ " sunt complementare, adică au suma 1.

Regula 2:  $L$  și  $R$  împart  $L \cup R$  în 2 regiuni disjuncte. Deci probabilitatea lui  $L \cup R$  este împărțită între  $P(L)$  și  $P(R)$ .

Regula 3: În suma  $P(L) + P(R)$  suprapunerea  $P(L \cap R)$  este numărată de 2 ori. Deci  $P(L) + P(R) - P(L \cap R)$  număra totul din reuniune exact o dată.

**Gândiți:** Regula 2 este un caz special al regulii 3.

Pentru următoarele exemple presupunem că avem un experiment care produce un întreg aleator între 1 și 20. Probabilitățile nu sunt necesar uniforme, i.e., nu sunt necesar aceleași pentru fiecare caz.

**Exemplul 8.** Dacă probabilitatea unui număr par este 0.6, care este probabilitatea unui număr impar?

**Răspuns:** Deoarece a fi impar este complementar cu a fi par, probabilitatea unui număr impar este  $1 - 0.6 = 0.4$ .

Să refacem acest exemplu un pic mai formal. Întâi, pentru a ne putea referi la el, să numim întregul aleator  $X$ . De asemenea, numim  $A$  evenimentul "X este par". Atunci evenimentul "X este impar" este  $A^c$ . Ne este dată  $P(A) = 0.6$ . De aceea  $P(A^c) = 1 - 0.6 = 0.4$ .

**Exemplul 9.** Considerăm 2 evenimente,  $A$ : "X este multiplu de 2";  $B$ : "X este impar și mai mic ca 10". Presupunem că  $P(A) = 0.6$  și  $P(B) = 0.25$ .

(i) Cine este  $A \cap B$ ?

(ii) Care este probabilitatea lui  $A \cup B$ ?

**Răspuns:** (i) Deoarece toate numerele din  $A$  sunt pare și toate numerele din  $B$  sunt impare, aceste evenimente sunt disjuncte. Adică  $A \cap B = \emptyset$ .

(ii) Deoarece  $A$  și  $B$  sunt disjuncte,  $P(A \cup B) = P(A) + P(B) = 0.85$ .

**Exemplul 10.** Fie  $A$ ,  $B$  și  $C$  evenimentele  $X$  este multiplu de 2, 3, respectiv 6. Dacă  $P(A) = 0.6$ ,  $P(B) = 0.3$  și  $P(C) = 0.2$ , cât este  $P(A$  sau  $B$ )?

**Răspuns:** Observăm 2 lucruri. Întâi, am folosit cuvântul "sau", care înseamnă reuniune: " $A$  sau  $B$ " =  $A \cup B$ . Al 2-lea, un întreg este divizibil cu 6 dacă și numai dacă este divizibil cu 2 și cu 3. De aici rezultă  $C = A \cap B$ . Deci principiul incluzerii și excluderii spune

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.3 - 0.2 = 0.7.$$

## 2 Probabilitatea condiționată, independență și teorema lui Bayes

### 2.1 Scopurile învățării

1. Să știe definițiile probabilității condiționate și independenței evenimentelor.
2. Să poată calcula probabilitatea condiționată direct din definiție.
3. Să poată folosi regula multiplicării pentru a calcula probabilitatea totală a unui eveniment.
4. Să poată verifica dacă 2 evenimente sunt independente.
5. Să poată folosi formula lui Bayes pentru a "inversa" probabilitățile condiționate.
6. Să poată organiza calculul probabilităților condiționate folosind arbori și tabele.
7. Să înțeleagă complet eroarea ratei de bază.

### 2.2 Probabilitatea condiționată

Probabilitatea condiționată răspunde la întrebarea "cum se schimbă probabilitatea unui eveniment dacă avem informație suplimentară".

**Exemplul 1.** Aruncăm o monedă corectă de 3 ori.

(a) Care este probabilitatea a 3 aversuri?

**Răspuns:** Spațiul probelor  $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ .

Toate cazurile sunt egal probabile, deci  $P(3 \text{ aversuri}) = 1/8$ .

(b) Presupunem că ni se spune că prima aruncare a fost avers. Fiind dată această informație, cum ar trebui să calculăm probabilitatea a 3 aversuri?

**Răspuns:** Avem un nou spațiu al probelor (redus):  $\Omega' = \{HHH, HHT, HTH, HTT\}$ .

Toate cazurile sunt egal probabile, deci

$$P(3 \text{ aversuri} \mid \text{prima aruncare este avers}) = 1/4.$$

Aceasta este numită **probabilitate condiționată**, deoarece ia în considerare condiții suplimentare. Pentru a dezvolta notația, reformulăm (b) în termeni de *evenimente*.

**(b) reformulat.** Fie  $A$  evenimentul "toate 3 aruncările sunt aversuri" =  $\{HHH\}$ .

Fie  $B$  evenimentul "prima aruncare este avers" =  $\{HHH, HHT, HTH, HTT\}$ .

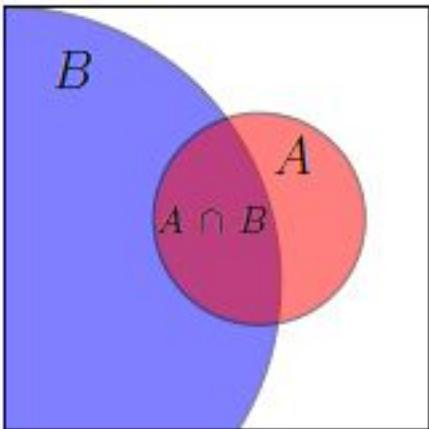
**Probabilitatea condiționată** a lui  $A$  cunoscând că  $B$  a avut loc este scrisă

$$P(A|B).$$

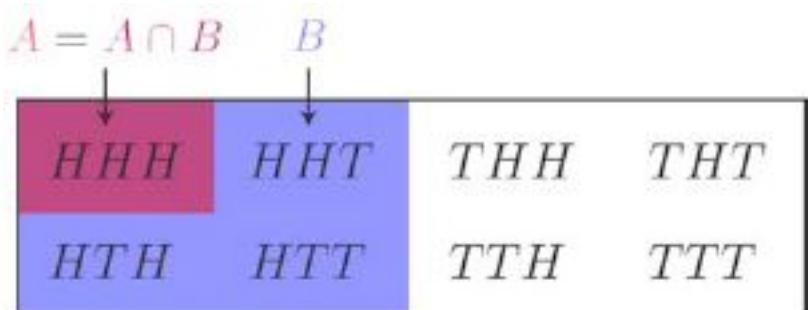
Aceasta se citește "probabilitatea condiționată a lui  $A$  **dat fiind**  $B$ " sau "probabilitatea lui  $A$  **condiționată** de  $B$ " sau, mai simplu, "probabilitatea lui  $A$

dat fiind  $B$ ".

Putem vizualiza probabilitatea condiționată după cum urmează. Gândiți  $P(A)$  ca proporția ariei ocupată de  $A$  din *tot* spațiul probelor. Pentru  $P(A|B)$  ne restrângem atenția la  $B$ . Adică,  $P(A|B)$  este proporția ocupată de  $A$  din aria lui  $B$ , i.e.  $P(A \cap B)/P(B)$ .



Probabilitatea condiționată: vizualizare abstractă



Probabilitatea condiționată: exemplul cu moneda

Observăm că  $A \subset B$  în ultima figură, deci apar doar 2 culori.

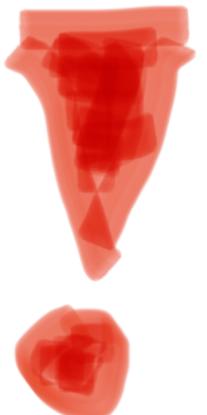
Definiția formală a probabilității condiționate prinde esența exemplului și vizualizărilor de mai sus.

### Definiția formală a probabilității condiționate

Fie  $A$  și  $B$  evenimente. Definim **probabilitatea condiționată** a lui  $A$  dat fiind  $B$  ca

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ dacă } P(B) \neq 0. \quad (1)$$

În continuare, când scriem o probabilitate condiționată, presupunem implicit că probabilitatea evenimentului care condiționează este nenulă, fără a mai



scrie aceasta.

Să refacem exemplul cu aruncarea monezii folosind definiția din relația (1). Reamintim că  $A = \text{"3 aversuri"}$  și  $B = \text{"prima aruncare este avers"}$ . Avem  $P(A) = 1/8$  și  $P(B) = 1/2$ . Deoarece  $A \cap B = A$ , avem de asemenea  $P(A \cap B) = 1/8$ . Acum, din (1),  $P(A|B) = \frac{1/8}{1/2} = 1/4$ , la fel ca răspunsul din exemplul 1b.

## 2.3 Regula multiplicării

Următoarea formulă este numită [regula multiplicării](#).

$$P(A \cap B) = P(A|B) \cdot P(B). \quad (2)$$

Aceasta este doar o rescriere a definiției din relația (1) a probabilității condiționate. Regula multiplicării este doar o versiune a regulii produsului.

Începem cu un exemplu simplu unde putem verifica direct toate probabilitățile numărând.

**Exemplul 2.** Tragem 2 cărți fără revenire dintr-un pachet de cărți de joc fără jokeri. Definim evenimentele:  $S_1 = \text{"prima carte este o pică"}$  și  $S_2 = \text{"a 2-a carte este o pică"}$ . Cât este  $P(S_2|S_1)$ ?

**Răspuns:** Putem face asta direct prin numărare: dacă prima carte este o pică, atunci din cele 51 de cărți rămase, 12 sunt pici.

$$P(S_2|S_1) = 12/51 = 4/17.$$

Acum, să recalcu-lăm asta folosind formula (1). Trebuie să calculăm  $P(S_1)$ ,  $P(S_2)$  și  $P(S_1 \cap S_2)$ . Știm că  $P(S_1) = 1/4$  deoarece sunt 52 de moduri egal posibile de a trage prima carte și 13 din ele sunt pici. Aceeași logică spune că sunt 52 de moduri egal posibile de a trage a 2-a carte, deci  $P(S_2) = 1/4$ .

**Comentariu:** Probabilitatea  $P(S_2) = 1/4$  poate părea surprinzătoare deoarece valoarea primei cărți afectează cu siguranță probabilitățile pentru a 2-a carte. Dar, dacă ne uităm la *toate* secvențele posibile de 2 cărți, vom vedea că fiecare carte din pachet are probabilitate egală de a fi a 2-a carte. Deoarece 13 din cele 52 de cărți sunt pici obținem  $P(S_2) = 13/52 = 1/4$ . Un alt fel de a spune asta este: dacă nu avem dată valoarea pentru prima carte, atunci trebuie să considerăm toate posibilitățile pentru a 2-a carte.

Continuând, vedem că

$$P(S_1 \cap S_2) = \frac{13 \cdot 12}{52 \cdot 51} = 1/17.$$

(Aceasta a fost aflată numărând modurile de a trage o pică urmată de a 2-a pică și împărțind prin numărul de moduri de a trage orice carte urmată de

orice altă carte.) Acum, folosind (1) obținem

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(S_1)} = \frac{1/17}{1/4} = 4/17.$$

În sfârșit, verificăm regula multiplicării calculând ambii membri din (2).

$$P(S_1 \cap S_2) = \frac{13 \cdot 12}{52 \cdot 51} = \frac{1}{17} \text{ și } P(S_2|S_1) \cdot P(S_1) = \frac{4/17}{1/4} = \frac{1}{17}, \text{ q.e.d.}$$

**Gândiți:** Pentru  $S_1$  și  $S_2$  din exemplul precedent, cât este  $P(S_2|S_1^c)$ ?

## 2.4 Legea probabilității totale

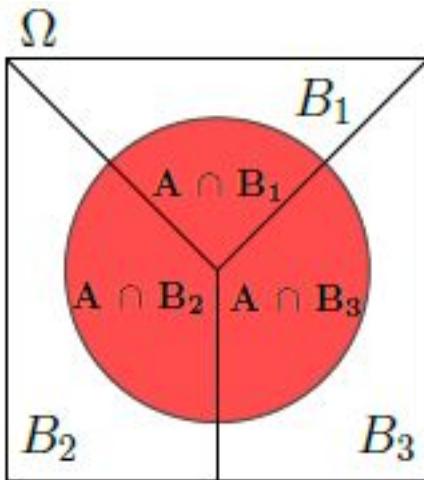
Legea probabilității totale ne permite să folosim regula multiplicării pentru a afla probabilități în exemple mai interesante. Necesară multă notație, dar ideea este destul de simplă. Enunțăm legea când spațiul probelor este împărțit în 3 părți. Este o chestiune simplă a extinde regula când sunt mai mult de 3 părți.

### Legea probabilității totale

Presupunem că spațiul probelor  $\Omega$  este împărțit în 3 evenimente disjuncte 2 câte 2  $B_1, B_2, B_3$  (vezi figura de mai jos). Atunci pentru orice eveniment  $A$ :

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3); \\ P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3). \end{aligned} \quad (3)$$

Prima relație spune "dacă  $A$  este împărțit în 3 părți disjuncte 2 câte 2, atunci  $P(A)$  este suma probabilităților părților". Relația (3) se numește **legea probabilității totale**. Ea este doar o rescriere a relației de deasupra folosind regula multiplicării.



Spațiul probelor  $\Omega$  și evenimentul  $A$  sunt fiecare împărțite în 3 părți disjuncte 2 câte 2.

Legea este valabilă dacă împărțim  $\Omega$  în orice număr de evenimente, atât timp cât ele sunt *disjuncte 2 câte 2* și acoperă tot  $\Omega$  (adică reuniunea lor este  $\Omega$ ). O astfel de împărțire este numită adesea o *partiție* a lui  $\Omega$ .

Primul nostru exemplu va fi unul unde știm deja răspunsul și putem verifica legea.

**Exemplul 3.** O urnă conține 5 bile roșii și 2 bile verzi. 2 bile sunt scoase una după alta, fără revenire. Care este probabilitatea ca a 2-a bilă să fie roșie?

**Răspuns:** Spațiul probelor este  $\Omega = \{rr, rg, gr, gg\}$ . (Am pus "r" pentru roșu și "g" pentru verde.)

Fie  $R_1$  evenimentul "prima bilă este roșie",  $G_1$  = "prima bilă este verde",  $R_2$  = "a 2-a bilă este roșie",  $G_2$  = "a 2-a bilă este verde". Ni se cere să aflăm  $P(R_2)$ .

Modul rapid de a calcula aceasta este la fel ca  $P(S_2)$  din exemplul cu cărți de joc de mai sus. Fiecare bilă este egal probabilă să fie a 2-a bilă. Deoarece 5 din 7 bile sunt roșii,  $P(R_2) = 5/7$ .

Să calculăm această valoare folosind legea probabilității totale (3). Înțâi, aflăm probabilitățile condiționate. Acesta este un exercițiu simplu de numărare.

$$P(R_2|R_1) = 4/6 = 2/3, \quad P(R_2|G_1) = 5/6.$$

Deoarece  $R_1$  și  $G_1$  partiționează  $\Omega$ , legea probabilității totale spune

$$\begin{aligned} P(R_2) &= P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) \\ &= \frac{2}{3} \cdot \frac{5}{7} + \frac{5}{6} \cdot \frac{2}{7} = \frac{10}{21} + \frac{5}{21} = \frac{15}{21} = \frac{5}{7}. \end{aligned} \tag{4}$$

### Urne de probabilitate

Exemplul precedent a folosit urne de probabilitate. Utilizarea lor se întoarce la începutul subiectului și am fi neglijenți dacă nu le-am introduce. Acest model didactic este foarte util.

În probabilități și statistică, o problemă a urnei este un exercițiu mental idealizat în care unele obiecte de real interes (ca atomi, oameni, mașini, etc.) sunt reprezentate ca bile colorate într-o urnă sau alt container. Se pretinde a se extrage una sau mai multe bile din urnă; scopul este a se determina probabilitatea extragerii unei culori sau a alteia, sau alte proprietăți. Un parametru cheie este dacă fiecare bilă este returnată în urnă sau nu după fiecare extragere.

Nu ne ia mult să facem un exemplu unde (3) este într-adevăr cel mai bun mod de a calcula probabilitatea. Iată un joc cu reguli puțin mai complicate.

**Exemplul 4.** O urnă conține 5 bile roșii și 2 bile verzi. O bilă este extrasă. Dacă este verde, o bilă roșie este adăugată în urnă, iar dacă este roșie, o bilă verde este adăugată în urnă. (Bila originală nu este returnată în urnă.) Apoi o a 2-a bilă este extrasă. Care este probabilitatea ca a 2-a bilă să fie roșie?

**Răspuns.** Cu notațiile de la exemplul 3, legea probabilității totale spune că  $P(R_2)$  poate fi calculată folosind relația (4). Doar valorile probabilităților condiționate se schimbă. Avem

$$P(R_2|R_1) = 4/7, \quad P(R_2|G_1) = 6/7.$$

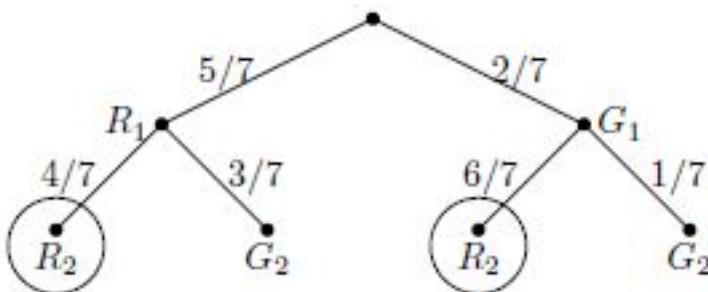
De aceea,

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) = \frac{4}{7} \cdot \frac{5}{7} + \frac{6}{7} \cdot \frac{2}{7} = \frac{32}{49}.$$

## 2.5 Utilizarea arborilor pentru a organiza calculul

Arborii sunt o metodă grozavă de a organiza calcule cu probabilități condiționate și legea probabilității totale. Figurile și exemplele vor clarifica ce înțelegem printr-un arbore. Ca la regula produsului, cheia este să organizăm procesul de bază într-o secvență de acțiuni.

Începem refăcând exemplul 4. Secvența de acțiuni este: întâi extragem bila 1 (și adăugăm bila corespunzătoare în urnă) și apoi extragem bila 2.



Interpretăm arboarele după cum urmează. Fiecare punct este numit **nod**. Arboarele este organizat pe nivele. Nodul de la vîrf (**nodul rădăcină**) este la nivelul 0. Următorul strat de mai jos este nivelul 1 și aşa mai departe. Fiecare nivel arată cazurile la un stadiu al jocului. Nivelul 1 arată cazurile posibile la prima extragere. Nivelul 2 arată cazurile posibile la a 2-a extragere, plecând din fiecare nod din nivelul 1.

Probabilitățile sunt scrise de-a lungul ramurilor. Probabilitatea lui  $R_1$  (bilă roșie la prima extragere) este  $5/7$ . Ea este scrisă de-a lungul ramurii de la nodul rădăcină la nodul etichetat  $R_1$ . La următorul nivel punem probabilitățile **condiționate**. Probabilitatea de-a lungul ramurii de la  $R_1$  la  $R_2$  este

$P(R_2|R_1) = 4/7$ . Ea reprezintă probabilitatea de a merge în nodul  $R_2$  dat fiind că suntem deja în  $R_1$ .

Regula multiplicării spune că probabilitatea de a ajunge în orice nod este tocmai produsul probabilităților aflate de-a lungul drumului de a ajunge acolo din nodul rădăcină. De exemplu, nodul etichetat  $R_2$  din extrema stângă reprezintă în realitate evenimentul  $R_1 \cap R_2$  deoarece vine din nodul  $R_1$ . Regula multiplicării spune acum

$$P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1) = \frac{5}{7} \cdot \frac{4}{7},$$

ceea ce este exact înmulțirea probabilităților aflate de-a lungul drumului de la nodul rădăcină la nodul  $R_2$  din extrema stângă.

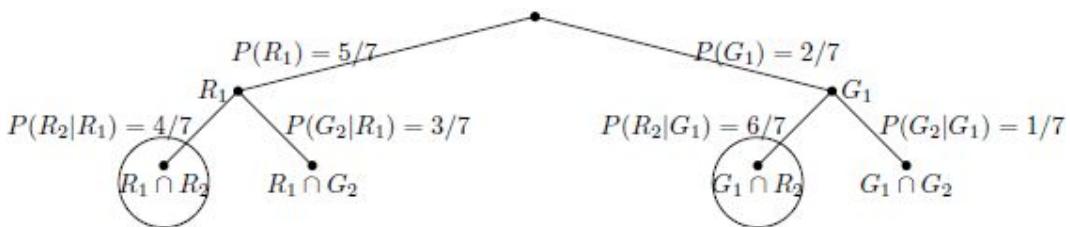
Legea probabilității totale este chiar afirmația că  $P(R_2)$  este suma probabilităților tuturor drumurilor conducând de la nodul rădăcină la  $R_2$  (cele 2 noduri încercuite din figură). În acest caz,

$$P(R_2) = \frac{5}{7} \cdot \frac{4}{7} + \frac{2}{7} \cdot \frac{6}{7} = \frac{32}{49},$$

exact ca în exemplul precedent.

### 2.5.1 Arbori prescurtați vs. precisi

Arborele dat mai sus implică o prescurtare. De exemplu, nodul marcat  $R_2$  din extrema stângă reprezintă în realitate evenimentul  $R_1 \cap R_2$ , deoarece termină drumul de la rădăcină prin  $R_1$  la  $R_2$ . Iată același arbore cu totul etichetat precis.



După cum se vede, acest arbore este mai greu de făcut. De obicei folosim versiunea prescurtată a arborilor.

## 2.6 Independență

2 evenimente sunt independente dacă informația că unul a avut loc nu schimbă probabilitatea ca celălalt să fi avut loc. Informal, evenimentele sunt independente dacă ele nu se influențează unul pe celălalt.

**Exemplul 5.** Aruncăm o monedă de 2 ori. Ne așteptăm ca rezultatele celor 2 aruncări să fie independente unul față de celălalt. În experimentele reale acest lucru trebuie totdeauna verificat. Dacă moneda mea aterizează în miere și nu mă deranjez s-o curăț, atunci a 2-a aruncare poate fi afectată de rezultatul primei aruncări.

Mai serios, independența experimentelor poate fi subminată de eșecul curățării sau recalibrării echipamentului între experimente, sau al izolării observatorilor presupuși independenți unul față de altul sau de o influență comună. Toți am experimentat aflarea același ”fapt” de la oameni diferiți. Aflându-l din surse diferite tindem să-i dăm crezare până aflăm că toți l-au aflat de la o sursă comună. Adică, sursele noastre nu erau independente.

Traducerea descrierii verbale în simboluri dă

$$A \text{ este independent de } B \iff P(A|B) = P(A). \quad (5)$$

Adică, cunoașterea faptului că  $B$  a avut loc nu schimbă probabilitatea ca  $A$  să fi avut loc. În termeni de evenimente ca submulțimi, cunoașterea faptului că rezultatul realizat este în  $B$  nu schimbă probabilitatea că el este în  $A$ . Dacă  $A$  și  $B$  sunt independente în sensul de mai sus, atunci regula multiplicării dă  $P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$ . Aceasta justifică următoarea definiție tehnică a independenței.

**Definiția formală a independenței:** 2 evenimente  $A$  și  $B$  sunt **independente**  $\iff$

$$P(A \cap B) = P(A) \cdot P(B). \quad (6)$$

Această definiție simetrică clarifică faptul că  $A$  este independent de  $B \iff B$  este independent de  $A$ . Spre deosebire de relația (5), această definiție are sens și când  $P(B) = 0$ . În termeni de probabilități condiționate, avem:

1. Dacă  $P(B) \neq 0$ , atunci  $A$  și  $B$  sunt independente  $\iff P(A|B) = P(A)$ .
2. Dacă  $P(A) \neq 0$ , atunci  $A$  și  $B$  sunt independente  $\iff P(B|A) = P(B)$ .

Evenimentele independente apar de obicei ca probe diferite într-un experiment, ca în exemplul următor.

**Exemplul 6.** Aruncăm o monedă corectă de 2 ori. Fie  $H_1$  = ”avers la prima aruncare” și  $H_2$  = ”avers la a 2-a aruncare”. Sunt  $H_1$  și  $H_2$  independente?

**Răspuns.** Deoarece  $H_1 \cap H_2$  este evenimentul ”ambele aruncări sunt averșuri”, avem

$$P(H_1 \cap H_2) = 1/4 = (1/2) \cdot (1/2) = P(H_1)P(H_2).$$

De aceea evenimentele  $H_1$  și  $H_2$  sunt independente.

Putem întreba despre independentă oricărora 2 evenimente, ca în următoarele 2 exemple.

**Exemplul 7.** Aruncăm o monedă corectă de 3 ori. Fie  $H_1$  = "avers la prima aruncare" și  $A$  = "2 aversuri în total". Sunt  $H_1$  și  $A$  independente?

**Răspuns.** Știm că  $P(A) = 3/8$ . Deoarece  $P(H_1) = 1/2 \neq 0$ , putem verifica dacă formula din relația (5) este îndeplinită. Acum,  $H_1 = \{HHH, HHT, HTH, HTT\}$  conține exact 2 cazuri ( $HHT, HTH$ ) din  $A$ , deci avem  $P(A|H_1) = 2/4 = 1/2$ . Deoarece  $P(A|H_1) \neq P(A)$ , evenimentele  $H_1$  și  $A$  nu sunt independente.

**Exemplul 8.** Tragem o carte dintr-un pachet de cărți de joc standard (adică fără jokeri). Să examinăm independentă 2 câte 2 a 3 evenimente: "cartea este as", "cartea este cupă" și "cartea este roșie".

Notăm evenimentele  $A$  = "as",  $H$  = "cupă",  $R$  = "roșie".

a) Știm că  $P(A) = 4/52 = 1/13$ ,  $P(A|H) = 1/13$ . Deoarece  $P(A) = P(A|H)$ ,  $A$  este independent de  $H$ .

b)  $P(A|R) = 2/26 = 1/13 = P(A)$ . Deci  $A$  este independent de  $R$ .

c) În sfârșit, ce putem spune despre  $H$  și  $R$ ? Deoarece  $P(H) = 13/52 = 1/4$  și  $P(H|R) = 13/26 = 1/2$ , avem  $P(H) \neq P(H|R)$ , deci  $H$  și  $R$  nu sunt independente. Puteam de asemenea vedea asta invers:  $P(R) = 26/52 = 1/2$  și  $P(R|H) = 13/13 = 1$ , deci  $P(R) \neq P(R|H)$ , de aceea  $H$  și  $R$  nu sunt independente.

### 2.6.1 Paradoxurile independenței

Un eveniment  $A$  cu probabilitatea 0 este independent de el însuși, deoarece în acest caz ambii membri ai relației (6) sunt 0. Aceasta apare paradoxal, deoarece cunoașterea faptului că  $A$  a avut loc sigur dă informație despre faptul dacă  $A$  a avut loc. Rezolvăm paradoxul observând că deoarece  $P(A) = 0$ , afirmația " $A$  a avut loc" este fără sens.

**Gândiți:** Pentru ce altă valoare a lui  $P(A)$  este  $A$  independent de el însuși?

## 2.7 Teorema lui Bayes

Teorema lui Bayes este foarte importantă pentru probabilități și statistică. Pentru 2 evenimente  $A$  și  $B$ , **teorema lui Bayes** (de asemenea numită **regula lui Bayes** și **formula lui Bayes**) spune

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}. \quad (7)$$

**Comentarii:** 1. Regula lui Bayes ne spune cum să "inversăm" probabilitățile condiționate, i.e. să aflăm  $P(B|A)$  din  $P(A|B)$ .

2. În practică,  $P(A)$  este adesea calculată folosind legea probabilității totale.

## Demonstrația regulii lui Bayes

Punctul cheie este că  $A \cap B$  este simetrică în  $A$  și  $B$ . Deci regula multiplicării spune

$$P(B|A) \cdot P(A) = P(A \cap B) = P(A|B) \cdot P(B).$$

Acum împărțim cu  $P(A)$  pentru a obține regula lui Bayes.

O greșală obișnuită este a confunda  $P(A|B)$  cu  $P(B|A)$ . Ele pot fi foarte diferite. Această fapt este ilustrat în următorul exemplu.

**Exemplul 9.** Aruncăm o monedă corectă de 5 ori. Fie  $H_1$  = "prima aruncare este avers" și  $H_A$  = "toate cele 5 aruncări sunt aversuri". Atunci  $P(H_1|H_A) = 1$ , dar  $P(H_A|H_1) = 1/16$ .

Pentru practică, să folosim teorema lui Bayes pentru a calcula  $P(H_1|H_A)$  din  $P(H_A|H_1)$ . Avem  $P(H_A|H_1) = 1/16$ ,  $P(H_1) = 1/2$ ,  $P(H_A) = 1/32$ . Deci,

$$P(H_1|H_A) = \frac{P(H_A|H_1)P(H_1)}{P(H_A)} = \frac{(1/16) \cdot (1/2)}{1/32} = 1,$$

de acord cu calculul nostru anterior.

### 2.7.1 Eroarea ratei de bază

Eroarea ratei de bază este unul din multele exemple care arată că este ușor de confundat semnificația lui  $P(B|A)$  cu  $P(A|B)$  când o situație este descrisă în cuvinte. Aceasta este unul din exemplele cheie în probabilități. Ar trebui să vă străduiți să-l înțelegeți complet.

#### Exemplul 10. Eroarea ratei de bază

Considerăm un test de rutină pentru o boală. Presupunem că frecvența bolii în populație (**rata de bază**) este 0.5%. Testul este extrem de exact cu o rată "fals pozitiv" de 5% și o rată "fals negativ" de 10%.

Faceți testul și ieșe pozitiv. Care este probabilitatea să aveți boala?

**Răspuns.** Vom face calculul de 3 ori: folosind arbori, tablele și simboluri.

Vom utiliza următoarele notații pentru evenimentele relevante:

$D+$  = "aveți boala",

$D-$  = "nu aveți boala",

$T+$  = "ați fost testat pozitiv",

$T-$  = "ați fost testat negativ".

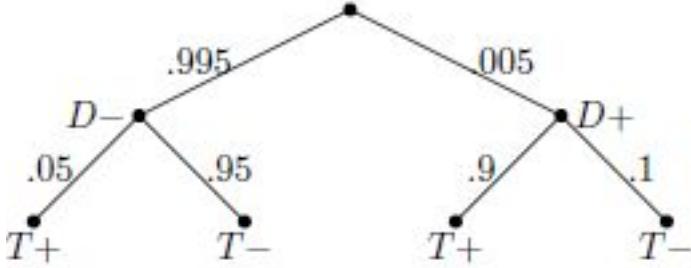
Avem dat  $P(D+) = 0.005$  și de aceea  $P(D-) = 0.995$ . Ratele "fals pozitiv" și "fals negativ" sunt (prin definiție) probabilități condiționate.

$$P(\text{fals pozitiv}) = P(T+|D-) = 0.05 \text{ și } P(\text{fals negativ}) = P(T-|D+) = 0.1.$$

Probabilitățile complementare sunt cunoscute ca ratele "adevărat negativ" și "adevărat pozitiv":

$$P(T-|D-) = 1 - P(T+|D-) = 0.95 \text{ și } P(T+|D+) = 1 - P(T-|D+) = 0.9.$$

**Arbore:** Toate aceste probabilități pot fi prezentate într-un arbore.



Se cere probabilitatea să aveți boala dat fiind că ați fost testat pozitiv, i.e.  $P(D+|T+)$ . Nu avem dată această valoare, dar știm  $P(T+|D+)$ , deci putem utiliza teorema lui Bayes.

$$P(D+|T+) = \frac{P(T+|D+) \cdot P(D+)}{P(T+)}$$

Cele 2 probabilități de la numărător sunt date. Calculăm numitorul folosind legea probabilității totale. Folosind arborele, trebuie doar să adunăm probabilitățile pentru fiecare din nodurile etichetate  $T+$

$$P(T+) = 0.995 \cdot 0.05 + 0.005 \cdot 0.9 = 0.05425.$$

Astfel,

$$P(D+|T+) = \frac{0.9 \cdot 0.005}{0.05425} \approx 0.08294931 \approx 8.3\%.$$

**Observații.** Aceasta este numită eroarea ratei de bază deoarece rata de bază a bolii în populație este atât de mică încât marea majoritate a oamenilor care fac testul sunt sănătoși și chiar cu un test precis majoritatea celor testați pozitiv vor fi oameni sănătoși.

Pentru a rezuma eroarea ratei de bază cu numere particulare  
*faptul că 95% din toate testele sunt precise nu implică faptul că 95% din testele pozitive sunt precise.*

### Alte moduri de a lucra exemplul 10

**Tabele:** Alt truc util pentru a calcula probabilități este a face un tabel. Să refacem exemplul precedent folosind un tabel construit pentru un total de 10000 de oameni împărțiți conform probabilităților din acest exemplu.

Construim tabelul după cum urmează. Alegem un număr, să zicem 10000 de

oameni, și îl plasăm ca marele total în dreapta jos. Folosind  $P(D+) = 0.005$  calculăm că 50 din cei 10000 de oameni sunt bolnavi ( $D+$ ). Analog 9950 de oameni sunt sănătoși. În acest moment tabelul arată aşa:

	$D+$	$D-$	total
$T+$			
$T-$			
total	50	9950	10000

Folosind  $P(T+|D+) = 0.9$  putem calcula că numărul oamenilor bolnavi care au fost testați pozitiv este 90% din 50, adică 45. Celealte date sunt similare. În acest moment tabelul arată aşa:

	$D+$	$D-$	total
$T+$	45	498	
$T-$	5	9452	
total	50	9950	10000

În sfârșit, adunăm elementele de pe liniile  $T+$  și  $T-$  pentru a completa coloana din dreapta.

	$D+$	$D-$	total
$T+$	45	498	543
$T-$	5	9452	9457
total	50	9950	10000

Folosind tabelul complet putem calcula

$$P(D+|T+) = \frac{|D+ \cap T+|}{|T+|} = \frac{45}{543} \approx 8.3\%.$$

**Simboluri:** Pentru completitudine, arătăm soluția scrisă direct cu simboluri.

$$\begin{aligned} P(D+|T+) &= \frac{P(T+|D+) \cdot P(D+)}{P(T+)} \\ &= \frac{P(T+|D+) \cdot P(D+)}{P(T+|D+) \cdot P(D+) + P(T+|D-) \cdot P(D-)} \\ &= \frac{0.9 \cdot 0.005}{0.9 \cdot 0.005 + 0.05 \cdot 0.995} \\ &\approx 8.3\%. \end{aligned}$$

**Vizualizare:** Figura de mai jos ilustrează eroarea ratei de bază. Aria albastră mare reprezintă toți oamenii sănătoși. Aria roșie mult mai mică reprezintă oamenii bolnavi. Dreptunghiul mai închis reprezintă oamenii care au fost testați pozitiv. Aria mai închisă acoperă cea mai mare parte a ariei roșii și doar o mică parte din aria albastră. Chiar și aşa, cea mai mare parte a ariei mai închise este peste albastru. Adică, cele mai multe teste pozitive sunt ale oamenilor sănătoși.



# Curs 3

Cristian Niculescu

## 1 Variabile aleatoare discrete

### 1.1 Scopurile învățării

1. Să știe definiția unei variabile aleatoare discrete.
2. Să știe distribuțiile (repartițiile) Bernoulli, binomială și geometrică și exemple de ce modelează ele.
3. Să poată să descrie funcția masă de probabilitate și funcția de distribuție cumulativă folosind tabele și formule.
4. Să poată să construiască noi variabile aleatoare din cele vechi.
5. Să știe cum să calculeze media.

### 1.2 Variabile aleatoare

Acest subiect este foarte mult despre introducerea unei terminologii utile, construită pe noțiunile de spațiu al probelor și funcție de probabilitate. Cuvintele cheie sunt

1. Variabilă aleatoare
2. Funcție masă de probabilitate (pmf)
3. Funcție de distribuție cumulativă (cdf), numită și funcție de repartiție.

#### 1.2.1 Recapitulare

Un **spațiu al probelor discret**  $\Omega$  este o mulțime finită sau numărabilă de cazuri (rezultate)  $\{\omega_1, \omega_2, \dots\}$ . Probabilitatea cazului  $\omega$  este notată  $P(\omega)$ .

Un **eveniment**  $E$  este o submulțime a lui  $\Omega$ . **Probabilitatea evenimentului**  $E$  este  $P(E) = \sum_{\omega \in E} P(\omega)$ .

#### 1.2.2 Variabilele aleatoare ca funcții de plată

**Exemplul 1.** Un joc cu 2 zaruri corecte.

Aruncăm 2 zaruri corecte și înregistram cazurile ca  $(i, j)$ , unde  $i$  este rezul-

tatul primului zar și  $j$  rezultatul celui de-al 2-lea. Putem lua spațiul probelor

$$\Omega = \{(1, 1), (1, 2), (1, 3), \dots, (6, 6)\} = \{(i, j) | i, j = 1, \dots, 6\}.$$

Funcția de probabilitate este  $P(i, j) = 1/36$ .

În acest joc, câștigăm 500 de dolari dacă suma este 7 și pierdem 100 altfel. Dăm acestei funcții de plată numele  $X$  și o descriem formal prin

$$X(i, j) = \begin{cases} 500 & \text{dacă } i + j = 7 \\ -100 & \text{dacă } i + j \neq 7. \end{cases}$$

**Exemplul 2.** Putem schimba jocul folosind o funcție de plată diferită. De exemplu

$$Y(i, j) = ij - 10.$$

În acest exemplu, dacă dăm (6,2), atunci câștigăm 2 dolari. Dacă dăm (2,3), atunci câștigăm  $-4$  dolari (i.e., pierdem 4 dolari).

Aceste funcții de plată sunt exemple de variabile aleatoare. O variabilă aleatoare atribuie un număr fiecărui caz dintr-un spațiu al probelor. Mai formal:

**Definiție.** Fie  $\Omega$  un spațiu al probelor. O variabilă aleatoare discretă este o funcție

$$X : \Omega \rightarrow \mathbb{R},$$

care ia o mulțime discretă de valori. (Reamintim că  $\mathbb{R}$  este mulțimea numerelor reale.)

De ce este  $X$  numită variabilă aleatoare? Este ”aleatoare” deoarece valoarea ei depinde de un caz aleator al unui experiment. Si tratăm pe  $X$  ca pe o variabilă ușuală: putem să-o adunăm la alte variabile aleatoare, să-o ridicăm la pătrat, și.a.m.d.

### 1.2.3 Evenimente și variabile aleatoare

Pentru orice valoare  $a$  scriem  $X = a$  pentru evenimentul format din toate cazurile  $\omega$  cu  $X(\omega) = a$ .

**Exemplul 3.** În exemplul 1 am aruncat 2 zaruri corecte și  $X$  a fost variabila aleatoare

$$X(i, j) = \begin{cases} 500 & \text{dacă } i + j = 7 \\ -100 & \text{dacă } i + j \neq 7. \end{cases}$$

Evenimentul  $X = 500$  este mulțimea  $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ , i.e. mulțimea tuturor cazurilor cu suma componentelor 7.  $P(X = 500) =$

$$6/36 = 1/6.$$

Permitem lui  $a$  să ia orice valoare, chiar și valori pe care  $X$  nu le ia niciodată. În exemplul 1, am putea considera evenimentul  $X = 1000$ . Deoarece  $X$  nu este egal niciodată cu 1000, acesta este tocmai **evenimentul vid** (sau multimea vidă)

$$\text{"}X = 1000\text{"} = \emptyset \quad P(X = 1000) = 0.$$

#### 1.2.4 Funcția masă de probabilitate

Devine obositor și greu să citim și să scriem  $P(X = a)$  pentru probabilitatea lui  $X = a$ . Când știm că vorbim despre  $X$  vom scrie simplu  $p(a)$ . Dacă vrem să-l facem explicit pe  $X$  vom scrie  $p_X(a)$ .

**Definiție.** **Funcția masă de probabilitate (pmf)** a unei variabile aleatoare discrete este funcția  $p(a) = P(X = a)$ .

Observăm:

1. Avem totdeauna  $0 \leq p(a) \leq 1$ .
2. Permitem lui  $a$  să fie orice număr real. Dacă  $a$  este o valoare pe care  $X$  n-o ia niciodată, atunci  $p(a) = 0$ .

**Exemplul 4.** Fie  $\Omega$  spațiul probelor de mai devreme pentru aruncarea a 2 zaruri corecte. Definim variabila aleatoare  $M$  ca **valoarea maximă a celor 2 zaruri**:

$$M(i, j) = \max(i, j).$$

De exemplu, aruncarea (3,5) are maximul 5, i.e.  $M(3, 5) = 5$ .

Putem descrie o variabilă aleatoare listând valorile ei posibile și probabilitățile asociate acestor valori. Pentru exemplul de mai sus avem:

valoarea	$a :$	1	2	3	4	5	6
pmf	$p(a) :$	1/36	1/12	5/36	7/36	1/4	11/36

Pe scurt, scriem

$$M \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/36 & 1/12 & 5/36 & 7/36 & 1/4 & 11/36 \end{pmatrix}$$

De exemplu,  $p(2) = 1/12$ .

**Întrebare:** Cât este  $p(8)$ ?

**Răspuns:**  $p(8) = 0$ .

**Gândiți:** Care este pmf pentru  $Z(i, j) = i + j$ ? Arată familiar?

#### 1.2.5 Evenimente și inegalități

Inegalitățile cu variabile aleatoare descriu evenimente. De exemplu,  $X \leq a$  este multimea tuturor cazurilor  $\omega$  a.i.  $X(\omega) \leq a$ .

**Exemplul 5.** Dacă spațiul probelor este mulțimea tuturor perechilor  $(i, j)$  provenind din aruncarea a 2 zaruri corecte și  $Z(i, j) = i + j$  este suma zarurilor, atunci

$$Z \leq 4 = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)\}.$$

### 1.2.6 Funcția de distribuție cumulativă (cdf)

**Definiție.** Funcția de distribuție cumulativă (cdf) a unei variabile aleatoare  $X$  este funcția  $F$  dată de  $F(a) = P(X \leq a)$ . Adesea vom prescurta aceasta **funcția de repartiție**.

Observăm că definiția lui  $F(a)$  utilizează simbolul mai mic **sau egal**. Acest fapt va fi important pentru a face calculele exact.

**Exemplu.** Continuând exemplul cu  $M$ , avem

valoarea	$a :$	1	2	3	4	5	6
pmf	$p(a) :$	1/36	1/12	5/36	7/36	1/4	11/36
cdf	$F(a) :$	1/36	1/9	1/4	4/9	25/36	1

$F$  e numită funcție de distribuție **cumulativă** deoarece  $F(a)$  dă totalul probabilității care se acumulează adunând probabilitățile  $p(b)$  când  $b$  merge de la  $-\infty$  la  $a$ . De exemplu, în tabelul de mai sus, elementul 4/9 din coloana 4 pentru cdf este suma valorilor pmf de la coloana 1 la coloana 4. Scriem:

$$\text{"}M \leq 4\text{"} = \{(i, j) \in \Omega | M(i, j) \leq 4\}; F(4) = P(M \leq 4) = 1/36 + 1/12 + 5/36 + 7/36 = 4/9.$$

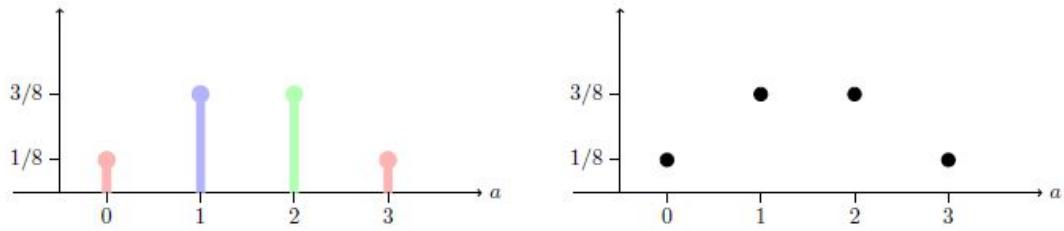
La fel ca funcția masă de probabilitate,  $F(a)$  este definită pentru toate valorile reale  $a$ . În exemplul de mai sus,  $F(8) = 1$ ,  $F(-2) = 0$ ,  $F(2.5) = 1/9$  și  $F(\pi) = 1/4$ .

### 1.2.7 Graficele lui $p$ și $F$

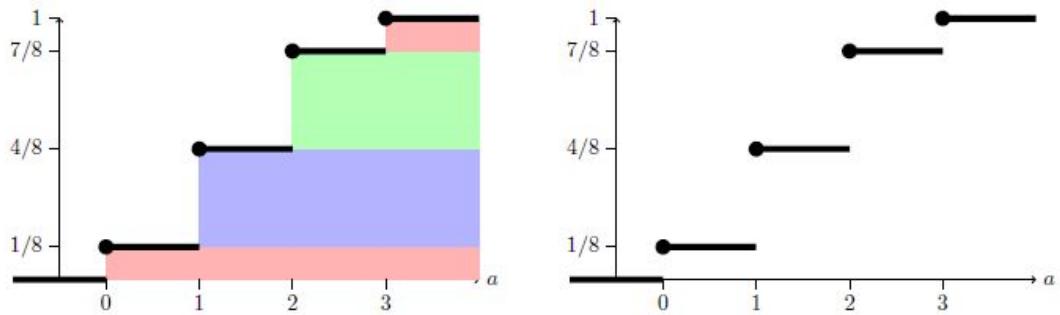
Putem vizualiza pmf și cdf cu grafice. De exemplu, fie  $X$  numărul de aversuri din 3 aruncări ale unei monede corecte:

valoarea	$a :$	0	1	2	3
pmf	$p(a) :$	1/8	3/8	3/8	1/8
cdf	$F(a) :$	1/8	1/2	7/8	1

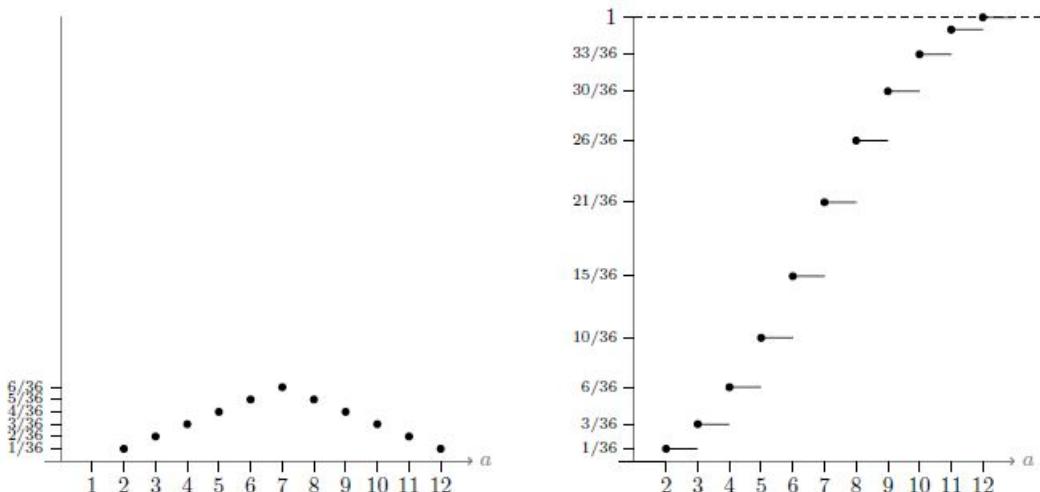
Graficele colorate arată cum funcția de distribuție cumulativă este construită prin **acumularea** probabilității când  $a$  crește. Graficele alb-negru sunt prezentări standard.



Funcția masă de probabilitate pentru  $X$



Funcția de distribuție cumulativă a lui  $X$



Pmf și cdf pentru suma a 2 zaruri corecte (Exemplul 5)

### 1.2.8 Proprietăți ale cdf $F$

Cdf  $F$  a unei variabile aleatoare satisfac câteva proprietăți:

1.  $F$  este **crescătoare**. Adică, graficul ei niciodată nu se duce în jos, sau, simbolic, dacă  $a \leq b$ , atunci  $F(a) \leq F(b)$ .
2.  $0 \leq F(a) \leq 1, \forall a \in \mathbb{R}$ .
3.  $\lim_{a \rightarrow -\infty} F(a) = 0, \lim_{a \rightarrow \infty} F(a) = 1$ .

În cuvinte, (1) spune că  $F(a)$  crește sau rămâne constantă când  $a$  crește, dar niciodată nu descrește; (2) spune că probabilitatea acumulată este totdeauna între 0 și 1; (3) spune că atunci când  $a$  ajunge negativ foarte mic, devine din ce în ce mai sigur faptul că  $X > a$ , iar când  $a$  ajunge foarte mare, devine din ce în ce mai sigur faptul că  $X \leq a$ .

**Gânditi:** De ce o cdf satisface fiecare dintre aceste proprietăți?

### 1.3 Repartiții (distribuții) clasice

#### 1.3.1 Repartiții Bernoulli

**Modelul:** Repartiția Bernoulli modelează o probă într-un experiment care poate avea ca rezultat ori **succes** ori **eșec**. O variabilă aleatoare  $X$  are o **repartiție Bernoulli** cu parametrul  $p \iff$

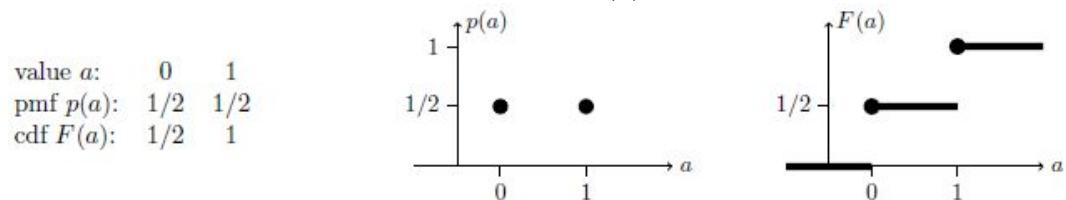
1.  $X$  ia valorile 0 și 1.
2.  $P(X = 1) = p$  și  $P(X = 0) = 1 - p$ .

Vom scrie  $X \sim \text{Bernoulli}(p)$  sau  $\text{Ber}(p)$ , ceea ce se citește " **$X$  are o repartiție Bernoulli cu parametrul  $p$** ".

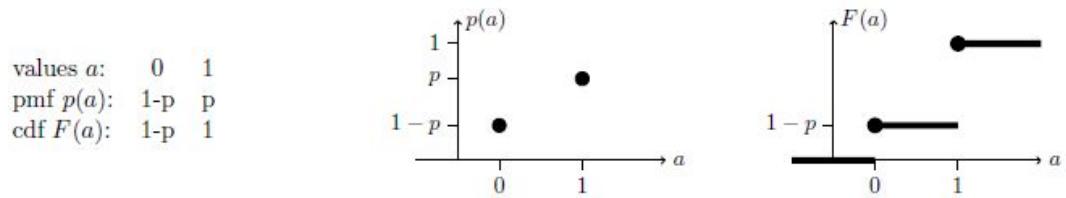
Un model simplu pentru repartiția Bernoulli este aruncarea unei monede cu probabilitatea  $p$  a aversului, cu  $X = 1$  pe avers și  $X = 0$  pe revers. Terminologia generală este a spune că  $X$  este 1 la **succes** și 0 la **eșec**, cu succesul și eșecul definite de context.

Multe decizii pot fi modelate ca o alegere binară, cum ar fi voturile pentru sau împotriva unei propuneri. Dacă  $p$  este proporția din populația care votează care favorizează propunerea, atunci votul unui individ aleator este modelat prin  $\text{Bernoulli}(p)$ .

Iată tabelul și graficele pentru pmf și cdf ale repartiției Bernoulli(1/2) și mai jos cele pentru repartiția generală Bernoulli( $p$ ).



Tabelul, pmf și cdf pentru repartiția Bernoulli(1/2)



Tabelul, pmf și cdf pentru repartiția Bernoulli( $p$ )

### 1.3.2 Repartiții binomiale

**Repartiția binomială** Binomial( $n, p$ ), sau Bin( $n, p$ ) modelează numărul de succese din  $n$  probe independente Bernoulli( $p$ ).

Aici este o ierarhie. O singură probă Bernoulli este, să zicem, o aruncare a unei monede. O singură probă binomială constă din  $n$  probe Bernoulli. Pentru aruncări de monede, spațiul probelor Bernoulli este  $\{H, T\}$ . Spațiul probelor binomiale este format din toate **sevențele** de aversuri și reversuri de lungime  $n$ . O variabilă aleatoare Bernoulli ia valorile 0 și 1. O variabilă aleatoare binomială ia valorile  $0, 1, 2, \dots, n$ .

**Exemplul 6.** Binomial( $1, p$ ) este la fel ca Bernoulli( $p$ ).

**Exemplul 7.** Numărul de aversuri din  $n$  aruncări ale unei monede cu probabilitatea  $p$  a aversului are o repartiție Binomial( $n, p$ ).

Descriem  $X \sim \text{Binomial}(n, p)$  dând valorile și probabilitățile ei. Considerăm  $k \in \{0, 1, \dots, n\}$ .

Reamintim că numărul de moduri de a alege  $k$  obiecte dintr-o colecție de  $n$  obiecte este "combinări de  $n$  luate câte  $k$ " =  $C_n^k$  și are formula

$$C_n^k = \frac{n!}{k!(n-k)!}. \quad (1)$$

(Este numit și **coeficient binomial**.) Iată un tabel pentru pmf a unei variabile aleatoare Binomial( $n, k$ ).

valorile $a$ :	0	1	2	...	$k$	...	$n$
pmf $p(a)$ :	$(1-p)^n$	$C_n^1 p^1 (1-p)^{n-1}$	$C_n^2 p^2 (1-p)^{n-2}$	...	$C_n^k p^k (1-p)^{n-k}$	...	$p^n$

**Exemplul 8.** Care este probabilitatea a 3 sau mai multe aversuri în 5 aruncări ale unei monede corecte?

**Răspuns.** Coeficienții binomiali asociați cu  $n = 5$  sunt

$$C_5^0 = 1, \quad C_5^1 = \frac{5!}{1!4!} = 5, \quad C_5^2 = \frac{5!}{2!3!} = \frac{5 \cdot 4}{2} = 10,$$

și analog

$$C_5^3 = 10, \quad C_5^4 = 5, \quad C_5^5 = 1.$$

Folosind aceste valori obținem următorul tabel pentru  $X \sim \text{Binomial}(5, p)$ .

valorile $a$ :	0	1	2	3	4	5
pmf $p(a)$ :	$(1-p)^5$	$5p(1-p)^4$	$10p^2(1-p)^3$	$10p^3(1-p)^2$	$5p^4(1-p)$	$p^5$

Ni s-a spus că  $p = 1/2$ , deci

$$P(X \geq 3) = 10 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 + 5 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^5 = \frac{16}{32} = \frac{1}{2}.$$

### 1.3.3 Explicarea probabilităților binomiale

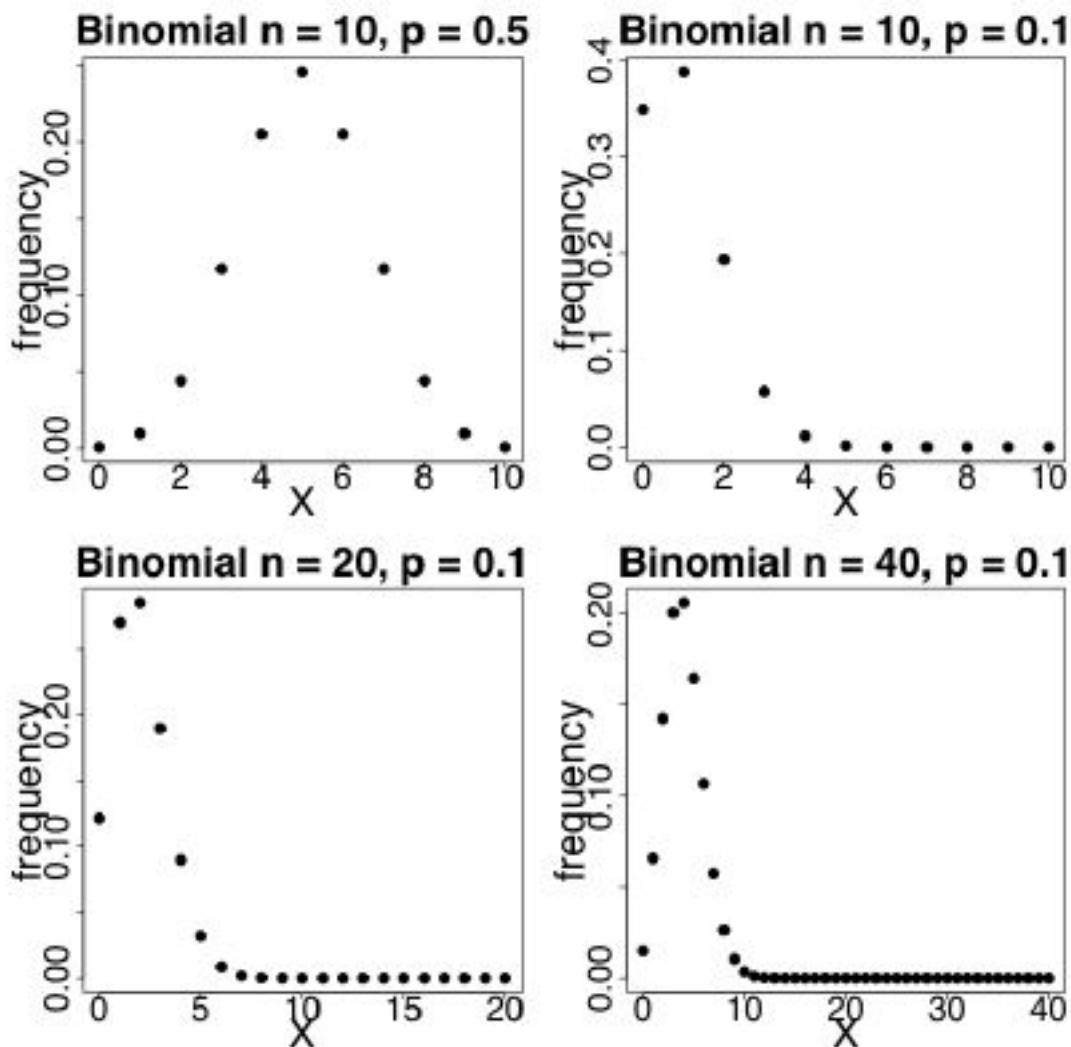
Pentru concretețe, fie  $n = 5$  și  $k = 2$  (argumentul pentru  $n$  și  $k$  arbitrari este identic). Deci  $X \sim \text{Binomial}(5, p)$  și vrem să calculăm  $p(2)$ . Calea lungă de a calcula  $p(2)$  este a lista toate modurile de a obține exact 2 aversuri în 5 aruncări de monedă și a aduna probabilitățile lor. Lista are 10 elemente: HHTTT, HTHTT, HTTHT, HTTTH, THHTT, THTHT, THTTH, TTHHT, TTHTH, TTTHH.

Fiecare element are aceeași probabilitate de a avea loc, și anume

$$p^2(1 - p)^3.$$

Aceasta este deoarece fiecare dintre cele 2 aversuri are probabilitatea  $p$  și fiecare dintre cele 3 reversuri are probabilitatea  $1 - p$ . Deoarece aruncările individuale sunt independente, putem înmulți probabilitățile. De aceea, probabilitatea totală a exact 2 aversuri este suma celor 10 probabilități identice, i.e.  $p(2) = 10p^2(1 - p)^3$ , aşa cum se arată în tabel.

Aceasta ne ghidează la calea scurtă de a face calculul. Trebuie să calculăm numărul de secvențe cu exact 2 aversuri. Pentru a face aceasta avem nevoie să alegem 2 dintre aruncări să fie aversuri și cele 3 rămase vor fi reversuri. Numărul de astfel de secvențe este numărul de moduri de a alege 2 din 5 obiecte, i.e.  $C_5^2$ . Deoarece fiecare astfel de secvență are aceeași probabilitate,  $p^2(1 - p)^3$ , obținem probabilitatea a exact 2 aversuri  $p(2) = C_5^2 p^2(1 - p)^3$ . Iată câteva pmf binomiale (aici, "frequency" înseamnă "probabilitate").



#### 1.3.4 Repartiții geometrice

O **repartiție geometrică** modelează numărul de reversuri dinaintea primului avers într-o secvență de aruncări de monedă (probe Bernoulli).

**Exemplul 9.** (a) Aruncăm repetat o monedă. Fie  $X$  numărul de reversuri dinaintea primului avers. Deci,  $X$  poate fi 0, i.e. prima aruncare este avers, 1, 2, .... În principiu poate lua orice valoare naturală.

(b) Dăm unui revers valoarea 0 și unui avers valoarea 1. În acest caz,  $X$  este numărul de 0-uri dinainte de primul 1.

(c) Dăm unui revers valoarea 1 și unui avers valoarea 0. În acest caz,  $X$  este numărul de 1-uri dinainte de primul 0.

(d) Numim reversul succes și aversul eșec. Astfel,  $X$  este numărul de succese

dinaintea primului eșec.

(e) Numim reversul eșec și aversul succes. Astfel,  $X$  este numărul de eșecuri dinaintea primului succes.

Putem vedea că aceasta modelează multe scenarii diferite de acest tip.

**Definiția formală.** Variabila aleatoare  $X$  are o **repartiție geometrică cu parametrul  $p \iff$**

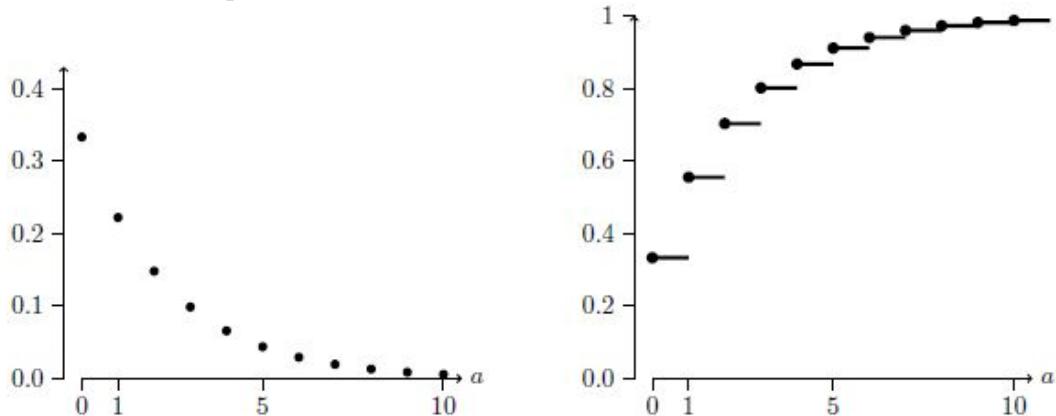
- $X$  ia valorile  $0, 1, 2, 3, \dots$
- are pmf  $p(k) = P(X = k) = p(1 - p)^k$ .

Notăm  $X \sim \text{geometric}(p)$  sau  $\text{geo}(p)$ . În formă de tabel avem:

valoarea $a$	0	1	2	3	...	$k$	...
pmf $p(a)$	$p$	$p(1 - p)$	$p(1 - p)^2$	$p(1 - p)^3$	...	$p(1 - p)^k$	...

Tabel:  $X \sim \text{geometric}(p)$ :  $X$  = numărul de 0-uri dinaintea primului 1.

Repartiția geometrică este un exemplu de repartiție discretă care ia un număr infinit de valori posibile.



Pmf și cdf pentru repartiția geometric( $1/3$ )

**Exemplul 10. Calculul probabilităților geometrice.** Presupunem că locuitorii unei insule își planifică familiile având copii până este născută prima fată. Presupunem că probabilitatea de a avea o fată este de 0.5 la fiecare sarcină, independent de alte sarcini, că toți copiii supraviețuiesc, că orice familie poate face oricât de mulți copii și că nu sunt gemeni. Care este probabilitatea ca o familie să aibă  $k$  băieți?

**Răspuns.** Numărul de băieți dintr-o familie este numărul de băieți dinaintea primei fete.

Fie  $X$  numărul de băieți dintr-o familie (aleasă aleator).  $X$  este o variabilă aleatoare geometrică. Ni se cere să aflăm  $p(k) = P(X = k)$ . O familie are  $k$  băieți dacă secvența de copii din familie de la cel mai mare la cel mai mic

este

$$BBB\dots BF,$$

primii  $k$  copii fiind băieți. Probabilitatea acestei secvențe este chiar produsul probabilităților pentru fiecare copil, i.e.  $(1/2)^k \cdot (1/2) = (1/2)^{k+1}$ . (Notă: Cele 4 presupuneri sunt simplificări ale realității.)

**Gândiți:** Care este raportul dintre băieți și fete pe insulă?

**Mai multă confuzie geometrică.** Altă definiție pentru repartiția geometrică este numărul de aruncări până la primul avers. În acest caz  $X$  poate lua valorile 1, i.e. prima aruncare este avers, 2, 3, .... Aceasta este chiar variabila aleatoare geometrică a noastră plus 1. Metodele de calcul cu ea sunt analoage celor folosite mai sus.

### 1.3.5 Repartiția uniformă

Repartiția uniformă modeleză orice situație unde toate cazurile sunt egal probabile.

$$X \sim \text{uniform}(N).$$

$X$  ia valorile  $1, 2, 3, \dots, N$ , fiecare cu probabilitatea  $1/N$ . Această repartiție apare, de exemplu, când modelăm aruncarea unei monede corecte ( $N = 2$ ), aruncarea unui zar corect ( $N = 6$ ), zile de naștere ( $N = 365$ ) și mâini de poker ( $N = C_{52}^5$ ).

### 1.3.6 Un site cu graficele unor repartiții

Adresa <http://mathlets.org/mathlets/probability-distributions/> dă o vedere dinamică a 6 repartiții. Graficele se schimbă lin pe măsură ce mișcăm diferiți cursori.

Graficele au un cod de culori ales cu grijă. 2 lucruri cu aceeași culoare reprezintă noțiuni identice sau foarte strâns legate.

### 1.3.7 Alte repartiții

Sunt multe alte repartiții cu nume apărute în diferite contexte. De exemplu, [http://en.wikipedia.org/wiki/Hypergeometric\\_distribution](http://en.wikipedia.org/wiki/Hypergeometric_distribution).

## 1.4 Aritmetică variabilelor aleatoare

Pot face operații aritmetice cu variabile aleatoare. De exemplu, le pot aduna, scădea, înmulți sau ridica la patrat.

Există o idee simplă, dar **extrem de importantă** pentru numărare. Spune că dacă avem o secvență de numere care sunt ori 0 ori 1, atunci suma secvenței

este numărul de 1-uri.

**Exemplul 11.** Considerăm secvența cu 5 1-uri

$$1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0.$$

Este ușor de văzut că suma acestei secvențe este 5, numărul de 1-uri.

Ilustrăm această idee calculând numărul de aversuri în  $n$  aruncări ale unei monede.

**Exemplul 12.** Aruncăm o monedă corectă de  $n$  ori. Fie  $X_j$  1, dacă a  $j$ -a aruncare este avers și 0, dacă este revers. Deci  $X_j$  este o variabilă aleatoare Bernoulli( $1/2$ ). Fie  $X$  numărul total de aversuri din  $n$  aruncări. Presupunând că aruncările sunt independente, știm că  $X \sim \text{binomial}(n, 1/2)$ . Putem scrie de asemenea

$$X = X_1 + X_2 + X_3 + \dots + X_n.$$

Din nou, aceasta este deoarece termenii din suma din dreapta sunt toți ori 0 ori 1. Deci, suma este exact numărul de  $X_j$ -uri care sunt 1, i.e. numărul de aversuri.

Lucrul important de văzut în exemplul de mai sus este că am scris variabila aleatoare mai complicată  $X$  ca suma variabilelor aleatoare extrem de simple  $X_j$ . Aceasta ne permite să manipulăm algebric pe  $X$ .

**Gândiți:** Presupunem că  $X$  și  $Y$  sunt independente,  $X \sim \text{binomial}(n, 1/2)$  și  $Y \sim \text{binomial}(m, 1/2)$ . Ce fel de repartiție are  $X + Y$ ?

**(Răspuns:**  $\text{binomial}(n + m, 1/2)$ . De ce?)

**Exemplul 13.** Presupunem că  $X$  și  $Y$  sunt variabile aleatoare independente.

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1/10 & 2/10 & 3/10 & 4/10 \end{pmatrix}$$

$$Y \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1/15 & 2/15 & 3/15 & 4/15 & 5/15 \end{pmatrix}$$

Verificați că probabilitatea totală pentru fiecare variabilă aleatoare este 1. Ce repartiție are variabila aleatoare  $X + Y$ ?

**Răspuns.**

$$(1/10) + (2/10) + (3/10) + (4/10) = 1.$$

Analog pentru  $Y$ .

Facem un tabel 2-dimensional pentru spațiul probelor produs constând din perechile  $(x, y)$ , unde  $x$  este o valoare posibilă a lui  $X$  și  $y$  una a lui  $Y$ . Pentru a ajuta calculul, probabilitățile pentru valorile lui  $X$  sunt puse în coloana din extrema dreaptă și cele pentru  $Y$  sunt pe linia de jos. Deoarece  $X$  și  $Y$

sunt independente, probabilitatea pentru perechea  $(x, y)$  este chiar produsul probabilităților individuale. ("X values" = "valorile lui X", "Y values" = "valorile lui Y").

		Y values					
		1	2	3	4	5	
X values	1	1/150	2/150	3/150	4/150	5/150	1/10
	2	2/150	4/150	6/150	8/150	10/150	2/10
	3	3/150	6/150	9/150	12/150	15/150	3/10
	4	4/150	8/150	12/150	16/150	20/150	4/10
		1/15	2/15	3/15	4/15	5/15	

Benzile diagonale arată mulțimile pătratelor unde  $X + Y$  este aceeași. Pentru a calcula repartiția pentru  $X + Y$  trebuie să adunăm probabilitățile pentru fiecare bandă.

$$X+Y \sim \left( \begin{array}{ccccccc} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1/150 & 4/150 & 10/150 & 20/150 & 30/150 & 34/150 & 31/150 & 20/150 \end{array} \right)$$

Simplificând fracțiile,

$$X + Y \sim \left( \begin{array}{ccccccc} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1/150 & 2/75 & 1/15 & 2/15 & 1/5 & 17/75 & 31/150 & 2/15 \end{array} \right)$$

Când tabelele sunt prea mari pentru a putea fi scrise avem nevoie să folosim tehnici pur algebrice pentru a calcula probabilitățile sumei.

## 2 Variabile aleatoare discrete: Media

### 2.1 Media

**Exemplul 1.** Presupunem că avem un zar corect cu 6 fețe marcat cu 5 3-uri și un 6. Cât ne-am aștepta să fie media a 6000 de aruncări?

**Răspuns.** Dacă am fi știut valoarea fiecărei aruncări, am fi calculat media adunând cele 6000 de valori și împărțind la 6000. Fără să știm valorile, putem calcula **media** după cum urmează.

Deoarece sunt 5 3-uri și un 6 ne așteptăm ca aproximativ  $5/6$  dintre aruncări să fie 3 și  $1/6$  dintre ele să fie 6. Presupunând acest fapt exact adevărat, avem următorul tabel de valori și numărări:

valoarea:	3	6
numărări așteptate:	5000	1000

Atunci media celor 6000 de valori este

$$\frac{5000 \cdot 3 + 1000 \cdot 6}{6000} = \frac{5}{6} \cdot 3 + \frac{1}{6} \cdot 6 = 3.5.$$

Considerăm aceasta media așteptată în sensul că ”ne așteptăm” ca fiecare dintre valorile posibile să apară cu frecvențele date.

**Exemplul 2.** Aruncăm 2 zaruri corecte. Câștigăm 1000 de dolari dacă suma este 2 și pierdem 100 de dolari altfel. Cât ne așteptăm să câștigăm în medie pe aruncare?

**Răspuns.** Probabilitatea unui 2 este  $1/36$ . Dacă jucăm de  $N$  ori, ne putem ”aștepta” ca  $\frac{1}{36} \cdot N$  dintre aruncări să dea 2 și  $\frac{35}{36} \cdot N$  dintre aruncări să dea altceva. Astfel, câștigurile așteptate totale sunt

$$1000 \cdot \frac{N}{36} - 100 \cdot \frac{35N}{36}.$$

Pentru a obține media așteptată pe aruncare, împărțim totalul la  $N$ :

$$\text{media așteptată} = 1000 \cdot \frac{1}{36} - 100 \cdot \frac{35}{36} = -69.(4).$$

**Gândită:** Ați vrea să jucați acest joc o dată? De mai multe ori?

Observăm că în ambele exemple suma pentru media așteptată constă din termeni care sunt o valoare a unei variabile aleatoare ori probabilitatea ei. Aceasta conduce la următoarea definiție.

**Definiție.** Presupunem că  $X$  este o variabilă aleatoare discretă care ia valourile  $x_1, x_2, \dots, x_n$  cu probabilitățile  $p(x_1), p(x_2), \dots, p(x_n)$ . **Media** lui  $X$  este

notată  $E(X)$  și definită prin

$$E(X) = \sum_{j=1}^n x_j p(x_j) = x_1 p(x_1) + x_2 p(x_2) + \dots + x_n p(x_n).$$

### Observații:

1. Media se mai notează adesea cu  $\mu$  ("miu") sau cu  $m$ .
2. După cum am văzut în exemplele de mai sus, media nu trebuie să fie o valoare posibilă a variabilei aleatoare. Mai degrabă este o medie ponderată a valorilor posibile.
3. Media dă o măsură a **tendinței centrale** a unei variabile aleatoare.
4. Dacă toate valorile sunt egale probabile, media este chiar media aritmetică a valorilor.
5. Fie  $X$  cu o mulțime numărabilă de valori  $x_1, x_2, \dots$  luate cu probabilitățile  $p(x_1), p(x_2), \dots$ . Atunci, media lui  $X$  este

$$E(X) = \sum_{j=1}^{\infty} x_j p(x_j),$$

dacă seria este convergentă.

**Exemplul 3.** Aflați  $E(X)$  pentru

$$X \sim \begin{pmatrix} 1 & 3 & 5 \\ 1/6 & 1/6 & 2/3 \end{pmatrix}.$$

**Răspuns.**  $E(X) = 1 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 5 \cdot \frac{2}{3} = \frac{24}{6} = 4$ .

**Exemplul 4.** Fie  $X \sim \text{Bernoulli}(p)$ . Aflați  $E(X)$ .

**Răspuns.**  $X$  ia valorile 0 și 1 cu probabilitățile  $1 - p$  și  $p$ , deci

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

**Important:** Media unei variabile aleatoare Bernoulli( $p$ ) este  $p$ .

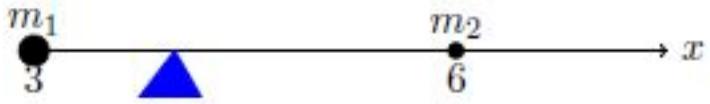
**Gândiți:** Care este media sumei a 2 zaruri corecte?

#### 2.1.1 Media și centrul de masă

Poate v-ați întrebat de ce folosim numele "funcție masă de probabilitate". Iată motivul: dacă plasăm câte un obiect de masă  $p(x_j)$  la poziția  $x_j, \forall j$ , atunci  $E(X)$  este poziția centrului de masă.

**Exemplul 5.** Fie 2 mase de-a lungul axei  $Ox$ , masa  $m_1 = 500$  la poziția  $x_1 = 3$  și masa  $m_2 = 100$  la poziția  $x_2 = 6$ . Unde este centrul de masă?

**Răspuns:** Intuitiv, centrul de masă este mai aproape de masa mai mare.



Centrul de masă este

$$\bar{x} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} = \frac{500 \cdot 3 + 100 \cdot 6}{600} = 3.5.$$

Numim această formulă o medie "ponderată" a lui  $x_1$  și  $x_2$ . Aici  $x_1$  este ponderat mai greu deoarece are mai multă masă.

Media  $E(X)$  este o medie ponderată a valorilor lui  $X$  cu ponderile fiind probabilitățile  $p(x_j)$  în locul maselor. Putem spune că "media este punctul în care repartitia s-ar echilibra". Observați similaritatea dintre exemplul din fizică și exemplul 1.

### 2.1.2 Proprietăți algebrice ale lui $E(X)$

Când adunăm, scalăm sau translatăm variabilele aleatoare, mediile fac același lucru. Modul matematic prescurtat de a spune aceasta este că  $E(X)$  este liniară.

**1.** Dacă  $X$  și  $Y$  sunt variabile aleatoare pe un spațiu al probelor  $\Omega$ , atunci

$$E(X + Y) = E(X) + E(Y).$$

**2.** Dacă  $a$  și  $b$  sunt constante, atunci

$$E(aX + b) = aE(X) + b.$$

$aX + b$  se obține prin **scalarea** lui  $X$  cu  $a$  și apoi **translatarea** cu  $b$ .

**Exemplul 6.** Aruncăm 2 zaruri corecte și fie  $X$  suma lor. Aflați  $E(X)$ .

**Răspuns:** Fie  $X_1$  numărul de pe primul zar și  $X_2$  numărul de pe cel de-al 2-lea zar. Deoarece  $X = X_1 + X_2$ , avem  $E(X) = E(X_1) + E(X_2)$ . Avem  $E(X_1) = E(X_2) = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$ , de aceea  $E(X) = 7$ .

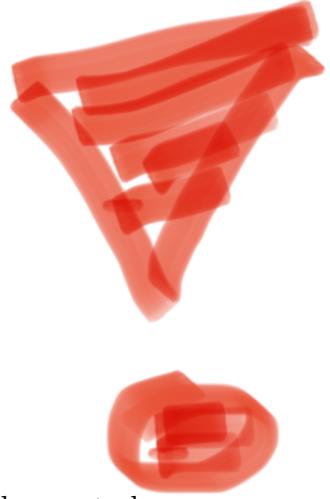
**Exemplul 7.** Fie  $X \sim \text{binomial}(n, p)$ . Aflați  $E(X)$ .

**Răspuns:** Reamintim că  $X$  modeleză numărul de succese din  $n$  variabile aleatoare Bernoulli( $p$ ), pe care le notăm  $X_1, \dots, X_n$ . Faptul cheie este că

$$X = \sum_{j=1}^n X_j.$$

Acum putem utiliza proprietatea algebrică (1) pentru a face calculele simple.

$$X = \sum_{j=1}^n X_j \implies E(X) = \sum_{j=1}^n E(X_j) = \sum_{j=1}^n p = np.$$



Puteam calcula  $E(X)$  direct

$$E(X) = \sum_{k=0}^n kp(k) = \sum_{k=0}^n kC_n^k p^k (1-p)^{n-k}.$$

Este posibil să arătăm că ultima sumă este într-adevăr  $np$ , dar metoda folosind proprietatea (1) este mult mai ușoară.

**Exemplul 8.** (Pentru variabile aleatoare discrete cu un număr infinit de valori media nu există totdeauna.) Fie  $X$  cu o mulțime numărabilă de valori,

$$X \sim \begin{pmatrix} 2 & 2^2 & 2^3 & \dots & 2^k & \dots \\ 1/2 & 1/2^2 & 1/2^3 & \dots & 1/2^k & \dots \end{pmatrix}.$$

Încercați să calculați media.

**Răspuns:** Media ar fi

$$E(X) = \sum_{k=1}^{\infty} 2^k \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty.$$

Media nu există, deoarece seria nu este convergentă.

### 2.1.3 Demonstrațiile proprietăților algebrice ale lui $E(X)$

Demonstrația proprietății (1) este simplă, dar are o subtilitate în chiar înțelegerea a ceea ce înseamnă adunarea a 2 variabile aleatoare. Reamintim că o valoare a unei variabile aleatoare este un număr determinat de rezultatul unui experiment. A aduna  $X$  și  $Y$  înseamnă a aduna valorile lui  $X$  și  $Y$  pentru același rezultat. În formă de tabel, în cazul finit:

rezultatul $\omega$ :	$\omega_1$	$\omega_2$	$\omega_3$	$\dots$	$\omega_n$
valoarea lui $X$ :	$x_1$	$x_2$	$x_3$	$\dots$	$x_n$
valoarea lui $Y$ :	$y_1$	$y_2$	$y_3$	$\dots$	$y_n$
valoarea lui $X + Y$ :	$x_1 + y_1$	$x_2 + y_2$	$x_3 + y_3$	$\dots$	$x_n + y_n$
probabilitatea $P(\omega)$ :	$P(\omega_1)$	$P(\omega_2)$	$P(\omega_3)$	$\dots$	$P(\omega_n)$

Demonstrația lui (1):

$$E(X + Y) = \sum_i (x_i + y_i) P(\omega_i) = \sum_i x_i P(\omega_i) + \sum_i y_i P(\omega_i) = E(X) + E(Y).$$

Demonstrația proprietății (2):

$$E(aX + b) = \sum_i (ax_i + b)p(x_i) = a \sum_i x_i p(x_i) + b \sum_i p(x_i) = aE(X) + b.$$

Termenul  $b$  din ultima expresie rezultă deoarece  $\sum_i p(x_i) = 1$ .

**Exemplul 9.** Media unei repartiții geometrice

Fie  $X \sim \text{geo}(p)$ . Reamintim că aceasta înseamnă că  $X$  ia valorile  $k \in \mathbb{N}$  cu probabilitățile  $p(k) = p(1-p)^k$ . ( $X$  modelează numărul de reversuri dinaintea primului avers într-o secvență de probe Bernoulli.) Media este

$$E(X) = \frac{1-p}{p}.$$

Demonstrația necesită un truc.

$$E(X) = \sum_{k=0}^{\infty} kp(1-p)^k.$$

Suma seriei geometrice:  $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ .

Derivăm ambii membri:  $\sum_{k=0}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$ .

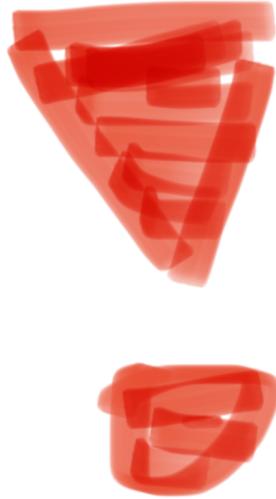
Înmulțim cu  $x$ :  $\sum_{k=0}^{\infty} kx^k = \frac{x}{(1-x)^2}$ .

Înlocuim  $x$  cu  $1-p$ :  $\sum_{k=0}^{\infty} k(1-p)^k = \frac{1-p}{p^2}$ .

Înmulțim cu  $p$ :  $\sum_{k=0}^{\infty} kp(1-p)^k = \frac{1-p}{p}$ .

Ultima expresie este media.

$$E(X) = \frac{1-p}{p}.$$



**Exemplul 10.** Aruncăm o monedă corectă până obținem un avers pentru prima dată. Care este numărul așteptat de reversuri?

**Răspuns:** Numărul de reversuri dinaintea primului avers este modelat de  $X \sim \text{geo}(1/2)$ . Din exemplul precedent,  $E(X) = \frac{1/2}{1/2} = 1$ .

**Exemplul 11.** Michael Jordan, baschetbalist, a marcat 80% dintre aruncările lui libere. Într-un joc, care este numărul așteptat de aruncări libere pe care le-ar marca înaintea primei ratări?

**Răspuns:** Aceasta este un exemplu unde vrem numărul de succese dinaintea primului eșec. Folosind limbajul de aversuri și reversuri: succesul este revers (probabilitate =  $1-p$ ) și eșecul este avers (probabilitate =  $p$ ). De aceea  $p = 0.2$  și numărul de reversuri (aruncări libere marcate) dinaintea primului avers (aruncare liberă ratată) este modelat de  $X \sim \text{geo}(0.2)$ . Din exemplul 9,

$$E(X) = \frac{1-p}{p} = \frac{0.8}{0.2} = 4.$$

#### 2.1.4 Medii ale funcțiilor de o variabilă aleatoare

(Formula schimbării de variabile)

Dacă  $X$  este o variabilă aleatoare luând valorile  $x_1, x_2, \dots$  și  $h$  este o funcție, atunci  $h(X)$  este o nouă variabilă aleatoare. Media ei este

$$E(h(X)) = \sum_j h(x_j)p(x_j).$$

**Exemplul 12.** Fie  $X$  valoarea aruncării unui zar corect și  $Y = X^2$ . Aflați  $E(Y)$ .

**Răspuns:** Deoarece există un număr mic de valori, putem face un tabel.

$X$	1	2	3	4	5	6
$Y$	1	4	9	16	25	36
probabilitatea	1/6	1/6	1/6	1/6	1/6	1/6

Observăm că probabilitatea pentru fiecare valoare a lui  $Y$  este aceeași ca cea pentru valoarea corespunzătoare a lui  $X$ . Deci,

$$E(Y) = E(X^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \dots + 6^2 \cdot \frac{1}{6} = 15.1(6).$$

**Exemplul 13.** Aruncăm 2 zaruri corecte și fie  $X$  suma. Presupunem că funcția de plată este dată de  $Y = X^2 - 6X + 1$ . Este acesta un pariu bun din punct de vedere financiar?

**Răspuns:** Avem  $E(Y) = \sum_{j=2}^{12} (j^2 - 6j + 1)p(j)$ , unde  $p(j) = P(X = j)$ .

Tabelul:

$X$	2	3	4	5	6	7	8	9	10	11	12
$Y$	-7	-8	-7	-4	1	8	17	28	41	56	73
probabilitatea	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Este mai ușor să calculăm în R. Iată codul:

```
x=2:12
y=x^2-6*x+1
p=c(1,2,3,4,5,6,5,4,3,2,1)/36
sum(p*y)
```

Obținem 13.83333, deci  $E(Y) = 13.8(3)$ .

Deoarece plata așteptată este pozitivă, pariu este bun din punct de vedere financiar.

**Întrebare:** Dacă  $Y = h(X)$ , rezultă  $E(Y) = h(E(X))$ ?

**Răspuns:** NU!!! Nu este adevărat în general!

**Gândiți:** Este adevărat în exemplul precedent?

**Întrebare:** Dacă  $Y = 3X + 77$ , rezultă  $E(Y) = 3E(X) + 77$ ?

**Răspuns:** Da. Din proprietatea (2), scalarea și translatarea se comportă astfel.

# Curs 4

Cristian Niculescu

## 1 Dispersia (varianța) variabilelor aleatoare discrete

### 1.1 Scopurile învățării

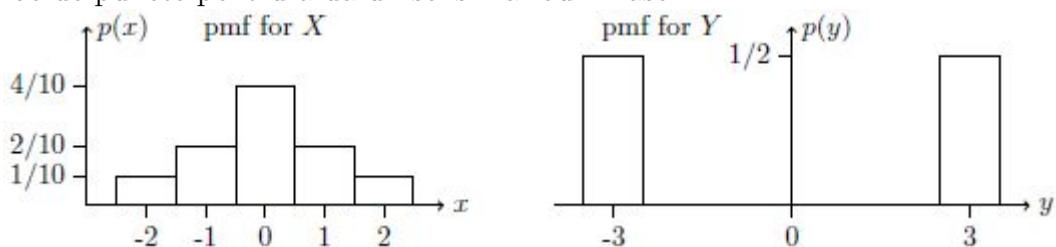
1. Să poată să calculeze dispersia și deviația standard a unei variabile aleatoare discrete.
2. Să înțeleagă că deviația standard este o măsură a scalării sau împrăștierii.
3. Să poată să calculeze dispersia folosind proprietățile de scalare și liniaritate.

### 1.2 Împrăștierea

Media unei variabile aleatoare este o măsură a [tendinței centrale](#). Dacă ar trebui să rezumăm cu un singur număr o variabilă aleatoare, media ar fi o alegere bună. Totuși, media nu cuprinde toată informația. De exemplu, variabilele aleatoare  $X$  și  $Y$  de mai jos au ambele media 0, dar masa lor de probabilitate este împărțită în jurul mediei complet diferit.

$$X \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 1/10 & 2/10 & 4/10 & 2/10 & 1/10 \end{pmatrix} \quad Y \sim \begin{pmatrix} -3 & 3 \\ 1/2 & 1/2 \end{pmatrix}.$$

Pentru a vedea împrăștierile diferite, reprezentăm pmf-urile. Folosim bare în loc de puncte pentru a da un sens mai bun masei.



Pmf-urile pentru 2 repartiții diferențiale, ambele cu media 0

### 1.3 Dispersia și deviația standard

Luând media centrul repartiției unei variabile aleatoare, dispersia este o măsură a cât de mult masa de probabilitate este **împrăștiată** în jurul acestui centru.

**Definiție.** Dacă  $X$  este o variabilă aleatoare cu media  $E(X) = \mu$ , atunci **dispersia** lui  $X$  este

$$Var(X) = E((X - \mu)^2).$$

**Deviația standard**  $\sigma$  a lui  $X$  este

$$\sigma = \sqrt{Var(X)}.$$

Dacă variabila aleatoare relevantă este clară din context, atunci dispersia și deviația standard sunt adesea notate cu  $\sigma^2$  și  $\sigma$  ("sigma"), la fel cum media este  $\mu$  ("miu").

Rescriem definiția explicit ca o sumă. Dacă  $X$  ia valorile  $x_1, x_2, \dots, x_n$  cu pmf  $p$ , atunci

$$Var(X) = E((X - \mu)^2) = \sum_{i=1}^n p(x_i)(x_i - \mu)^2.$$

Formula pentru  $Var(X)$  este o medie ponderată a pătratelor distanțelor la medie. Prin ridicare la pătrat, ne asigurăm că mediem numai valori nenegative, astfel încât împărțirea la dreapta mediei să nu se anuleze cu cea de la stânga. Prin folosirea mediei, ponderăm valorile cu probabilitate mai mare mai mult decât valorile cu probabilitate mai mică.

Observații despre unități de măsură:

1.  $\sigma$  are aceleași unități de măsură ca  $X$ .
2.  $Var(X)$  are aceleași unități de măsură ca pătratul lui  $X$ . Astfel, dacă  $X$  este în metri, atunci  $Var(X)$  este în metri pătrați.

Deoarece  $\sigma$  și  $X$  au aceleași unități de măsură, deviația standard este o măsură firească a împărăstierii.

**Exemplul 1.** Calculați media, dispersia și deviația standard ale variabilei aleatoare

$$X \sim \begin{pmatrix} 1 & 3 & 5 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

**Răspuns:** Întâi calculăm  $E(X) = 1 \cdot (1/4) + 3 \cdot (1/4) + 5 \cdot (1/2) = 7/2$ . Apoi completăm tabelul cu  $(X - 7/2)^2$ .

valoarea $x$	1	3	5
$p(x)$	1/4	1/4	1/2
$(x - 7/2)^2$	25/4	1/4	9/4

Acum calculul dispersiei este similar celui al mediei:

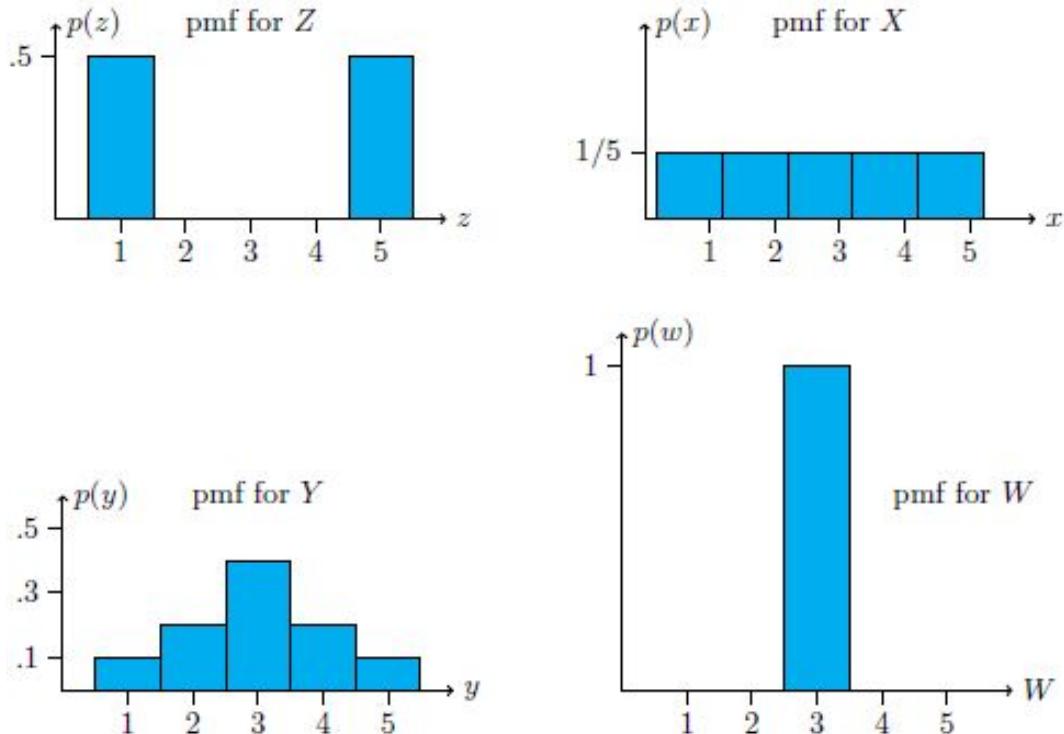
$$Var(X) = \frac{25}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} + \frac{9}{4} \cdot \frac{1}{2} = \frac{11}{4}.$$

Extrăgând radicalul, avem deviația standard  $\sigma = \sqrt{11}/2$ .

**Exemplul 2.** Pentru fiecare variabilă aleatoare  $X, Y, Z$  și  $W$  reprezentați pmf și calculați media și dispersia.

- (i)  $Z \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$ ;
- (ii)  $Y \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1/10 & 1/5 & 2/5 & 1/5 & 1/10 \end{pmatrix}$ ;
- (iii)  $Z \sim \begin{pmatrix} 1 & 5 \\ 1/2 & 1/2 \end{pmatrix}$ ;
- (iv)  $W \sim \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ .

**Răspuns:** Fiecare variabilă aleatoare are media 3, dar probabilitatea este împărațială diferit. În reprezentările de mai jos, le ordonăm de la cea mai mare la cea mai mică dispersie:  $Z, X, Y, W$ .



Calculăm dispersia fiecărei variabile. Toate au media  $\mu = 3$ . Deoarece dispersia este definită ca o medie, o putem calcula folosind tabele.

	valoarea $x$	1	2	3	4	5
(i)	pmf $p(x)$	1/5	1/5	1/5	1/5	1/5
	$(X - \mu)^2$	4	1	0	1	4

$$Var(X) = E((X - \mu)^2) = \frac{4}{5} + \frac{1}{5} + 0 + \frac{1}{5} + \frac{4}{5} = 2.$$

	valoarea $y$	1	2	3	4	5
(ii)	pmf $p(y)$	1/10	1/5	2/5	1/5	1/10
	$(Y - \mu)^2$	4	1	0	1	4

$$Var(Y) = E((Y - \mu)^2) = \frac{2}{5} + \frac{1}{5} + 0 + \frac{1}{5} + \frac{2}{5} = 1.2.$$

	valoarea $z$	1	5
(iii)	pmf $p(z)$	1/2	1/2
	$(Z - \mu)^2$	4	4

$$Var(Z) = E((Z - \mu)^2) = 2 + 2 = 4.$$

	valoarea $w$	3
(iv)	pmf $p(w)$	1
	$(W - \mu)^2$	0

$Var(W) = E((W - \mu)^2) = 0$ . Observăm că  $W$  nu variază, de aceea are dispersia 0.

### 1.3.1 Dispersia unei variabile aleatoare Bernoulli( $p$ )

Dacă  $X \sim \text{Bernoulli}(p)$ , atunci

$$Var(X) = p(1 - p).$$

**Demonstratie:** Știm că  $E(X) = p$ . Calculăm  $Var(X)$  folosind un tabel.

valorile lui $X$	0	1
pmf $p(x)$	1 - $p$	$p$
$(X - \mu)^2$	$(0 - p)^2$	$(1 - p)^2$

$$Var(X) = (1 - p)p^2 + p(1 - p)^2 = p(1 - p)(p + 1 - p) = p(1 - p).$$

**Gândiți:** Pentru ce valoare a lui  $p$  Bernoulli( $p$ ) are cea mai mare dispersie? Încercați să răspundeți aplicând inegalitatea mediilor sau aflând maximul funcției de gradul 2  $f(p) = p - p^2$  pe  $(0, 1)$ .

### 1.3.2 Variabile aleatoare discrete independente

Până acum am folosit noțiunea de variabile aleatoare independente fără să o definim riguros. De exemplu, o repartiție binomială este sumă de probe Bernoulli **independente**. Sigur, avem un sens intuitiv despre ce înseamnă independentă pentru probe experimentale. Avem de asemenea sensul probabilistic că variabilele aleatoare  $X$  și  $Y$  sunt independente când cunoașterea

valorii lui  $X$  nu ne dă informații despre valoarea lui  $Y$ .

**Definiție:** Variabilele aleatoare discrete  $X$  și  $Y$  sunt **independente**  $\iff$

$$P(X = a, Y = b) = P(X = a)P(Y = b), \forall a, b.$$

### 1.3.3 Proprietățile dispersiei

Cele mai utile 3 proprietăți pentru calculul dispersiei sunt:

1. Dacă  $X$  și  $Y$  sunt **independente**, atunci  $Var(X + Y) = Var(X) + Var(Y)$ .
2. Pentru constantele  $a$  și  $b$ ,  $Var(aX + b) = a^2Var(X)$ .
3.  $Var(X) = E(X^2) - E(X)^2$ .

Pentru proprietatea 1, observăm cu atenție cerința că  $X$  și  $Y$  sunt independente.

Proprietatea 3 dă o formulă pentru  $Var(X)$  care este adesea mai ușor de utilizat în calcule.

**Exemplul 3.** Presupunem că  $X$  și  $Y$  sunt independente și  $Var(X) = 3$  și  $Var(Y) = 5$ . Aflați:

- (i)  $Var(X + Y)$ ,
- (ii)  $Var(3X + 4)$ ,
- (iii)  $Var(X + X)$ ,
- (iv)  $Var(X + 3Y)$ .

**Răspuns:** Pentru a calcula aceste dispersii, folosim proprietățile 1 și 2.

- (i) Deoarece  $X$  și  $Y$  sunt **independente**,  $Var(X + Y) = Var(X) + Var(Y) = 8$ .
- (ii) Folosind proprietatea 2,  $Var(3X + 4) = 9Var(X) = 27$ .
- (iii) Nu vă lasați păcăliți! Proprietatea 1 nu poate fi aplicată deoarece sigur  $X$  nu este independentă de ea însăși. Putem folosi proprietatea 2:  $Var(X + X) = Var(2X) = 4Var(X) = 12$ . (Observație: dacă din greșală am fi folosit proprietatea 1, am fi răspuns greșit 6.)
- (iv) Folosim ambele proprietăți 1 și 2.

$$Var(X + 3Y) = Var(X) + Var(3Y) = 3 + 9 \cdot 5 = 48.$$

**Exemplul 4.** Folosiți proprietatea 3 pentru a calcula dispersia lui  $X \sim \text{Bernoulli}(p)$ .

**Răspuns:** Din tabelul

$X$	0	1
$p(x)$	1 - $p$	$p$
$X^2$	0	1

avem  $E(X^2) = p$ . Deci proprietatea 3 dă

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p).$$

Rezultatul coincide cu cel din calculul anterior.

**Exemplul 5.** Refață exemplul 1 folosind proprietatea 3.

**Răspuns:** Din tabelul

$X$	1	3	5
$p(x)$	1/4	1/4	1/2
$X^2$	1	9	25

avem  $E(X)=7/2$  și

$$E(X^2) = 1 \cdot \frac{1}{4} + 9 \cdot \frac{1}{4} + 25 \cdot \frac{1}{2} = \frac{60}{4} = 15.$$

Deci  $Var(X) = 15 - (7/2)^2 = 11/4$ , ca în exemplul 1.

#### 1.3.4 Dispersia pentru binomial( $n, p$ )

Presupunem  $X \sim \text{binomial}(n, p)$ . Deoarece  $X$  este sumă de  $n$  variabile Bernoulli( $p$ ) *independente* și fiecare variabilă Bernoulli are dispersia  $p(1-p)$ , avem

$$X \sim \text{binomial}(n, p) \Rightarrow Var(X) = np(1-p).$$

#### 1.3.5 Demonstrațiile proprietăților 2 și 3

**Demonstrația proprietății 2:** Rezultă din proprietățile lui  $E(X)$ .

Fie  $\mu = E(X)$ . Atunci  $E(aX + b) = a\mu + b$  și

$$\begin{aligned} Var(aX + b) &= E((aX + b - (a\mu + b))^2) = E((aX - a\mu)^2) = E(a^2(X - \mu)^2) \\ &= a^2 E((X - \mu)^2) = a^2 Var(X). \end{aligned}$$

**Demonstrația proprietății 3:** Folosim proprietățile lui  $E(X)$ . Reamintim că  $\mu$  este o constantă și că  $E(X) = \mu$ .

$$\begin{aligned} E((X - \mu)^2) &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - E(X)^2, \text{ q.e.d.} \end{aligned}$$

## 1.4 Tabele de repartiții și proprietăți

Repartiția	valorile lui $X$	pmf $p(x)$	media $E(X)$	dispersia $Var(X)$
Bernoulli( $p$ )	0,1	$p(0) = 1 - p, p(1) = p$	$p$	$p(1 - p)$
Binomial( $n, p$ )	$0, 1, \dots, n$	$p(k) = C_n^k p^k (1 - p)^{n-k}$	$np$	$np(1 - p)$
Uniform( $n$ )	$1, 2, \dots, n$	$p(k) = \frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Geometric( $p$ )	$0, 1, 2, \dots$	$p(k) = p(1 - p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

Fie  $X$  o variabilă aleatoare discretă cu valorile  $x_1, x_2, \dots$  și pmf  $p(x_j)$ .

Media:	Dispersia:
Sinonime:	varianță
Notății: $E(X), \mu, m$	$Var(X), \sigma^2$
Definiție: $E(X) = \sum_j x_j p(x_j)$	$E((X - \mu)^2) = \sum_j p(x_j)(x_j - \mu)^2$
Scalare, translatare: $E(aX + b) = aE(X) + b$	$Var(aX + b) = a^2 Var(X)$
Aditivitate: $E(X + Y) = E(X) + E(Y)$	$X, Y$ ind. $\Rightarrow Var(X + Y) = Var(X) + Var(Y)$
Funcții de $X$ : $E(h(X)) = \sum_j h(x_j) p(x_j)$	
Formulă alternativă:	$Var(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$

## 2 Variabile aleatoare continue

### 2.1 Scopurile învățării

1. Să știe definiția unei variabile aleatoare continue.
2. Să știe definiția funcției densitate de probabilitate (pdf) și funcției de distribuție cumulativă (cdf).
3. Să poată să explice de ce folosim densitatea de probabilitate pentru variabilele aleatoare continue.

### 2.2 Introducere

Variabilele aleatoare atribuie un număr fiecărui rezultat posibil dintr-un spațiu al probelor. În timp ce variabilele aleatoare discrete iau o mulțime discretă de valori posibile, variabilele aleatoare continue au o mulțime continuă de valori.

Computațional, pentru a merge de la discret la continuu, pur și simplu înlocuim sumele cu integrale.

**Exemplul 1.** Deoarece timpul este continuu, cantitatea de timp cu care Ion vine mai devreme (sau întârzie) la curs este o variabilă aleatoare continuă. Presupunem că măsurăm cât de devreme sosește Ion la curs în fiecare zi (în minute). Adică, rezultatul unei probe din experimentul nostru este un timp

în minute. Presupunem că sunt fluctuații aleatoare în timpul exact în care el apare. Deoarece în principiu Ion ar putea sosi, să zicem, cu 3.43 minute mai devreme, sau cu 2.7 minute mai târziu (corespunzând rezultatului  $-2.7$ ), sau la orice alt timp, spațiul probelor constă din toate numerele reale. Deci variabila aleatoare care dă rezultatul are ea însăși un **domeniu continuu** de valori posibile.

## 2.3 Încălzire de analiză matematică

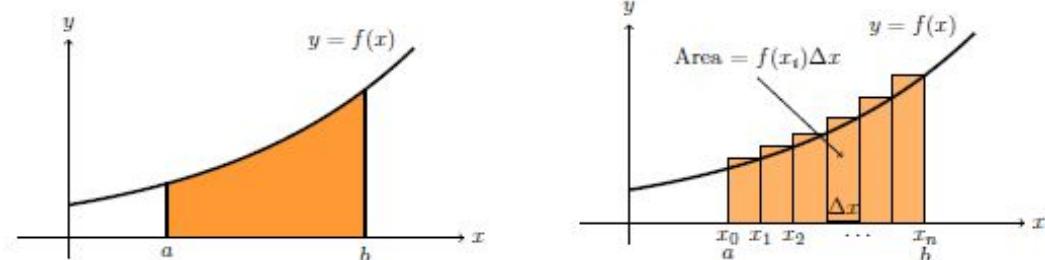
Dacă  $f$  este o funcție integrabilă și  $f(x) \geq 0, \forall x \in [a, b]$ , atunci

1.  $\int_a^b f(x)dx$  = aria de sub curba  $y = f(x)$ .
2.  $\int_a^b f(x)dx$  = "suma din  $f(x)dx$ ".

Coneziunea dintre cele 2 este:

$$\text{aria} \approx \text{suma ariilor dreptunghiurilor} = f(x_1)\Delta x + f(x_2)\Delta x + \dots + f(x_n)\Delta x = \sum_{i=1}^n f(x_i)\Delta x.$$

Pe măsură ce lungimea  $\Delta x$  a intervalor devine mai mică, aproximarea devine mai bună.



Aria este aproximativ suma ariilor dreptunghiurilor

**Observație:** Interesul nostru în integrale vine în primul rând din interpretarea lor ca "sumă" și într-o mai mică măsură din interpretarea lor ca arie.

## 2.4 Variabile aleatoare continue și funcții densitate de probabilitate

O variabilă aleatoare continuă ia un **domeniu de valori** a cărui lungime poate fi finită sau infinită. Iată câteva exemple de domenii de valori:  $[0, 1]$ ,  $[0, \infty)$ ,  $\mathbb{R}$ ,  $[a, b]$ .

**Definiție:** O variabilă aleatoare  $X$  este **continuă**  $\iff \exists$  o funcție  $f(x)$  astfel încât  $\forall c \leq d$  avem

$$P(c \leq X \leq d) = \int_c^d f(x)dx. \quad (1)$$

Funcția  $f$  este numită **funcția densitate de probabilitate (pdf)**.

Pdf satisface totdeauna următoarele proprietăți:

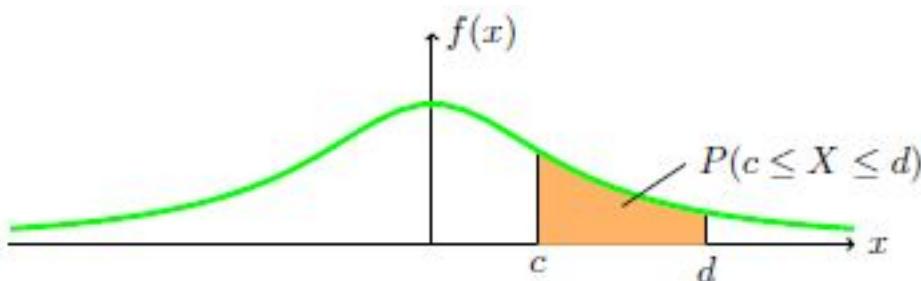
1.  $f(x) \geq 0$  ( $f$  este nenegativă).
  2.  $\int_{-\infty}^{\infty} f(x)dx = 1$  (Aceasta este echivalentă cu:  $P(-\infty < X < \infty) = 1$ ).
- Funcția densitate de probabilitate  $f$  a unei variabile aleatoare continue este analoaga funcției masă de probabilitate  $p$  a unei variabile aleatoare discrete. Iată 2 diferențe importante:
1. Spre deosebire de  $p$ , pdf  $f$  nu este o probabilitate. Trebuie să-o integrăm pentru a obține probabilitate.
  2. Deoarece  $f$  nu este o probabilitate, nu există restricția  $f(x) \leq 1$ .

**Observație:** În proprietatea 2, am integrat de la  $-\infty$  la  $\infty$  deoarece n-am știut domeniul de valori luate de  $X$ . Formal, aceasta are sens deoarece definim  $f(x) = 0$  în afara domeniului de valori al lui  $X$ . În practică, vom integra între marginile date de domeniul de valori al lui  $X$ .

#### 2.4.1 Vederea grafică a probabilității

Dacă facem graficul pdf a unei variabile aleatoare continue  $X$ , atunci

$$P(c \leq X \leq d) = \text{aria de sub graficul pdf dintre } c \text{ și } d.$$



**Gândiți:** Cât este aria totală de sub pdf  $f$ ?

#### 2.4.2 Termenii ”masă de probabilitate” și ”densitate de probabilitate”

De ce folosim termenii masă și densitate pentru a descrie pmf și pdf? Care este diferența dintre cele 2? Răspunsul simplu este că acești termeni sunt complet analogi masei și densității din fizică.

**Masa ca o sumă:**

Dacă masele  $m_1, m_2, m_3$  și  $m_4$  sunt puse pe o linie în pozițiile  $x_1, x_2, x_3$  și  $x_4$ , atunci masa totală este  $m_1 + m_2 + m_3 + m_4$ .

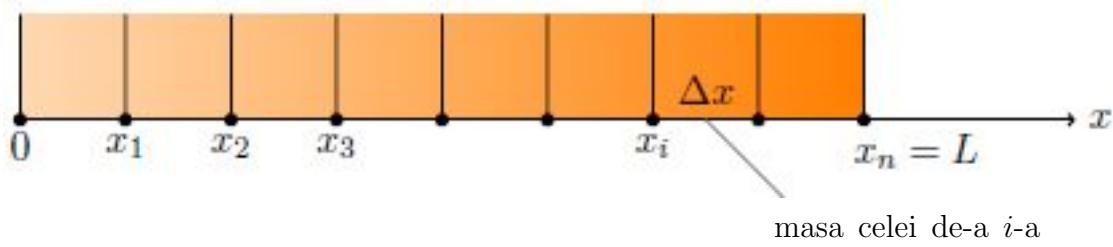


Puteam defini o "funcție de masă"  $p$  cu  $p(x_j) = m_j$  pentru  $j = 1, 2, 3, 4$  și  $p(x) = 0$  altfel. Cu această notație, masa totală este  $p(x_1) + p(x_2) + p(x_3) + p(x_4)$ .

**Funcția masă de probabilitate** se comportă în exact același mod, cu excepția faptului că are dimensiunea probabilității în loc de cea a masei.

**Masa ca o integrală a densității:**

Presupunem că avem o tijă de lungime  $L$  metri cu densitate variabilă  $f(x)$  kg/m. (Observăm că unitățile de măsură sunt masă/lungime.)



bucăți  $\approx f(x_i)\Delta x$

Dacă densitatea variază continuu, trebuie să aflăm masa totală a tijei prin integrare:

$$\text{masa totală} = \int_0^L f(x)dx.$$

Această formulă provine din împărțirea tijei în bucăți mici și "adunarea" masei fiecărei bucăți. Adică:

$$\text{masa totală} \approx \sum_{i=1}^n f(x_i)\Delta x.$$

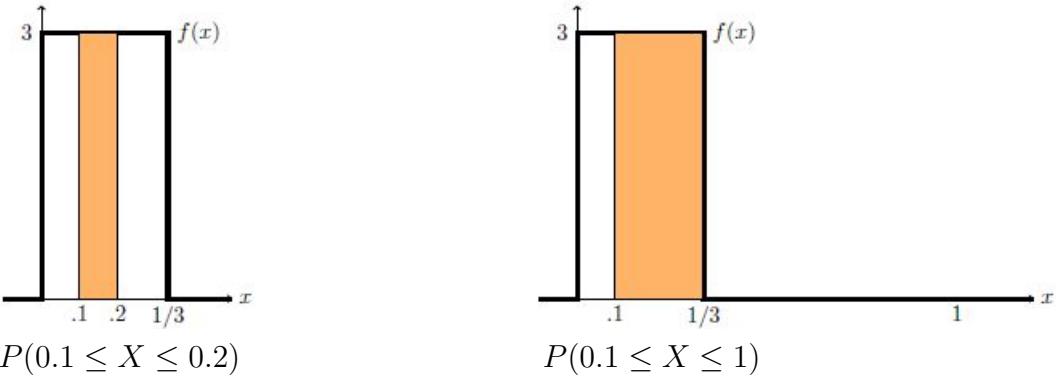
La limită când  $\Delta x$  tinde la 0, suma tinde la integrală.

**Funcția densitate de probabilitate** se comportă exact la fel, exceptând faptul că are ca unități de măsură probabilitatea/(unitatea lui  $x$ ) în loc de kg/m. Într-adevăr, relația (1) este exact analoaga integralei de mai sus pentru masa totală.

Pentru ambele variabile aleatoare, discretă și continuă, media este **centralul masei** sau punctul de echilibru.

**Exemplul 2.** Presupunem că  $X$  are pdf  $f(x) = 3$  pe  $[0, 1/3]$  (aceasta înseamnă că  $f(x) = 0$  în afara lui  $[0, 1/3]$ ). Reprezentați pdf și calculați  $P(0.1 \leq X \leq 0.2)$  și  $P(0.1 \leq X \leq 1)$ .

**Răspuns:**



$$P(0.1 \leq X \leq 0.2) = \int_{0.1}^{0.2} f(x)dx = \int_{0.1}^{0.2} 3dx = 3x|_{0.1}^{0.2} = 0.6 - 0.3 = 0.3.$$

Sau putem afla probabilitatea geometric:

$$P(0.1 \leq X \leq 0.2) = \text{aria dreptunghiului} = 3 \cdot 0.1 = 0.3.$$

Deoarece  $P(0.1 \leq X \leq 1)$  este aria de sub  $f(x)$  de la 0.1 doar până la  $1/3$ , avem  $P(0.1 \leq X \leq 1) = 3(1/3 - 0.1) = 3 \cdot (7/30) = 0.7$ .

**Gânditi:** În exemplul anterior  $f(x)$  ia valori mai mari ca 1. De ce aceasta nu încalcă regula că probabilitățile sunt totdeauna între 0 și 1?

**Observație asupra notației.** Putem defini o variabilă aleatoare dându-i domeniul de valori și funcția densitate de probabilitate. De exemplu, putem spune: fie  $X$  o variabilă aleatoare cu domeniul de valori  $[0, 1]$  și pdf  $f(x) = x/2$ . Implicit, aceasta înseamnă că  $X$  nu are densitate de probabilitate în afara domeniului de valori dat. Dacă am fi vrut să fim absolut riguroși, am fi scris explicit și că  $f(x) = 0$  în afara lui  $[0, 1]$ , dar în practică aceasta nu este necesar.

**Exemplul 3.** Fie  $X$  o variabilă aleatoare cu domeniul de valori  $[0, 1]$  și pdf  $f(x) = Cx^2$ . Cât este  $C$ ?

**Răspuns:** Deoarece probabilitatea totală trebuie să fie 1, avem

$$\int_0^1 f(x)dx = 1 \Leftrightarrow \int_0^1 Cx^2 dx = 1.$$

Calculând integrala, ultima relație devine

$$((Cx^3)/3)|_0^1 = 1 \Rightarrow C/3 = 1 \Rightarrow C = 3.$$

Observație: Avem nevoie de constanta  $C$  de mai sus pentru a **norma** densitatea astfel încât probabilitatea totală să fie 1.

**Exemplul 4.** Fie  $X$  variabila aleatoare din exemplul 3. Aflați  $P(X \leq 1/2)$ .

**Răspuns:**  $P(X \leq 1/2) = \int_0^{1/2} 3x^2 dx = x^3|_0^{1/2} = \frac{1}{8}$ .

**Gândiți:** Pentru acest  $X$  (sau orice variabilă aleatoare continuă):

- Cât este  $P(a \leq X \leq a)$ ?
- Cât este  $P(X = 0)$ ?
- $P(X = a) = 0$  înseamnă că  $X$  nu poate fi niciodată egal cu  $a$ ?

În cuvinte, întrebările de mai sus duc la faptul că probabilitatea ca înălțimea unei persoane aleatoare să fie exact 1.7526 m (la precizie infinită, i.e. fără rotunjiri!) este 0. Totuși, este posibil ca înălțimea cuiva să fie exact 1.7526 m. Deci răspunsurile la întrebările de gândire sunt 0, 0 și Nu.

#### 2.4.3 Funcția de distribuție cumulativă (funcția de repartiție)

Funcția de distribuție cumulativă (cdf) a unei variabile aleatoare continue  $X$  este definită în exact același mod ca cdf a unei variabile aleatoare discrete.

$$F(b) = P(X \leq b).$$

Observăm că definiția este despre probabilitate. Când folosim cdf ar trebui să ne gândim mai întâi la ea ca la o probabilitate. Apoi, când o calculăm putem folosi

$$F(b) = P(X \leq b) = \int_{-\infty}^b f(x)dx, \text{ unde } f(x) \text{ este pdf a lui } X.$$

#### Observații:

1. Pentru variabile aleatoare discrete, am definit funcția de distribuție cumulativă, dar nu am avut prea mult ocazia să o folosim. Cdf joacă un rol mult mai proeminent pentru variabile aleatoare continue.
2. Ca mai înainte, am integrat de la  $-\infty$  deoarece n-am știut precis domeniul de valori al lui  $X$ . Formal, aceasta are sens deoarece  $f(x) = 0$  în afara domeniului de valori al lui  $X$ . În practică, știm domeniul de valori și integrăm de la începutul lui.
3. În practică spunem adesea "X are repartiția  $F(x)$ " mai degrabă decât "X are funcția de distribuție cumulativă  $F(x)$ ".

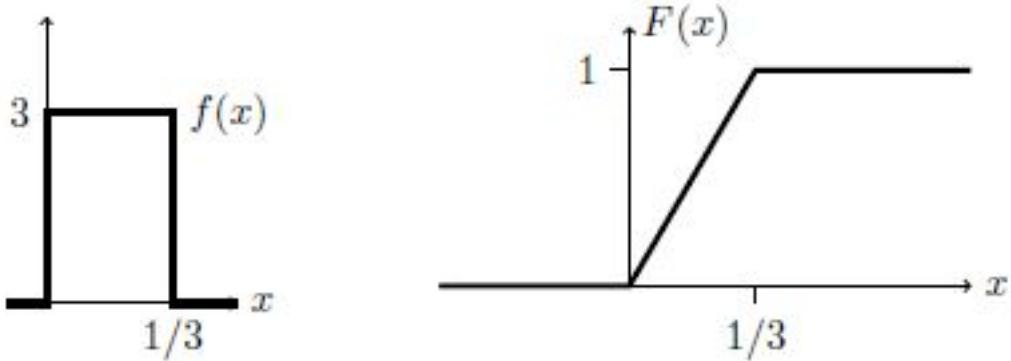
**Exemplul 5.** Aflați cdf pentru densitatea din exemplul 2.

**Răspuns:** Pentru  $a \in [0, 1/3]$  avem  $F(a) = \int_0^a f(x)dx = \int_0^a 3dx = 3x|_0^a = 3a$ .

Deoarece  $f(x) = 0$  în afara lui  $[0,1/3]$  avem  $F(a) = P(X \leq a) = 0$  pentru  $a < 0$  și  $F(a) = 1$  pentru  $a > 1/3$ . Punând toate acestea împreună, avem

$$F(a) = \begin{cases} 0, & \text{dacă } a < 0, \\ 3a, & \text{dacă } 0 \leq a \leq 1/3, \\ 1, & \text{dacă } a > 1/3. \end{cases}$$

Iată graficele lui  $f$  și  $F$ .



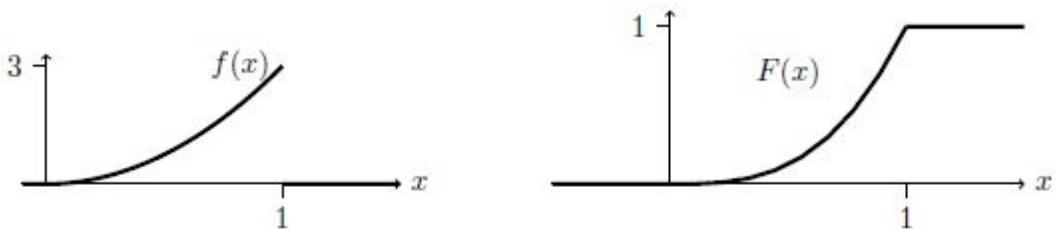
Observați scalele diferite pe axele verticale. Reamintim că axa verticală pentru pdf reprezintă densitatea de probabilitate, iar cea pentru cdf reprezintă probabilitatea.

**Exemplul 6.** Aflați cdf pentru pdf din exemplul 3,  $f(x) = 3x^2$  pe  $[0,1]$ . Presupunem că  $X$  este o variabilă aleatoare cu această repartiție. Aflați  $P(X < 1/2)$ .

**Răspuns:**  $f(x) = 3x^2$  pe  $[0, 1] \Rightarrow F(a) = \int_0^a 3x^2 dx = x^3|_0^a = a^3$  pe  $[0,1]$ . De aceea,

$$F(a) = \begin{cases} 0, & \text{dacă } a < 0, \\ a^3, & \text{dacă } 0 \leq a \leq 1, \\ 1, & \text{dacă } a > 1. \end{cases}$$

Astfel,  $P(X < 1/2) = F(1/2) = (1/2)^3 = 1/8$ . Iată graficele lui  $f$  și  $F$ :



#### 2.4.4 Proprietăți ale funcției de distribuție cumulative

Iată un rezumat al celor mai importante proprietăți ale funcției de distribuție cumulative (cdf) pentru o variabilă aleatoare continuă:

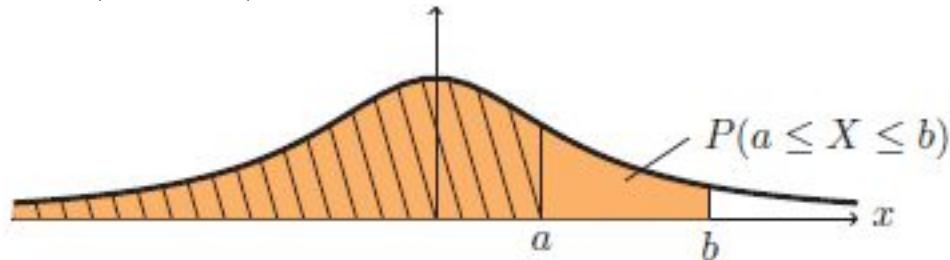
1. (Definiția)  $F(x) = P(X \leq x)$ .
2.  $0 \leq F(x) \leq 1, \forall x \in \mathbb{R}$ .
3.  $F$  este crescătoare, i.e. dacă  $a \leq b$ , atunci  $F(a) \leq F(b)$ .
4.  $\lim_{x \rightarrow -\infty} F(x) = 0$  și  $\lim_{x \rightarrow \infty} F(x) = 1$ .
5.  $P(a \leq X \leq b) = F(b) - F(a)$ .
6.  $F'(x) = f(x), \forall x$  în care  $F$  este derivabilă.

Proprietățile 2, 3 și 4 sunt identice cu cele pentru repartiții discrete. Graficele din exemplele anterioare le ilustrează.

Proprietatea 5 poate fi demonstrată:

$$\begin{aligned} \int_{-\infty}^b f(x)dx &= \int_{-\infty}^a f(x)dx + \int_a^b f(x)dx \\ \Leftrightarrow \int_a^b f(x)dx &= \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx \\ \Leftrightarrow P(a \leq X \leq b) &= F(b) - F(a). \end{aligned}$$

Proprietatea 5 poate fi de asemenea văzută geometric. Regiunea colorată de mai jos reprezintă  $F(b)$  și regiunea hașurată reprezintă  $F(a)$ . Diferența lor este  $P(a \leq X \leq b)$ .



Proprietatea 6 este o teoremă fundamentală a analizei matematice.

#### 2.4.5 Densitatea de probabilitate ca o tablă de darts

Ne gândim la prelevarea de valori ale unei variabile aleatoare continue ca la aruncarea de darts-uri într-o tablă de darts. Considerăm regiunea de sub graficul pdf ca o tablă de darts. Împărțim regiunea în pătrate mici de aceeași mărime și presupunem că, atunci când aruncăm un dart, este egal probabil ca el să aterizeze în orice pătrat. Probabilitatea ca dartul să aterizeze într-o regiune dată este fracția din totalul ariei de sub curbă preluată de regiune.

Deoarece totalul ariei este 1, această fracție este chiar aria regiunii. Dacă  $X$  reprezintă coordonata  $x$  a dartului, atunci probabilitatea ca dartul să aterizeze cu coordonata  $x$  între  $a$  și  $b$  este chiar

$$P(a \leq X \leq b) = \text{aria de sub } f(x) \text{ dintre } a \text{ și } b = \int_a^b f(x)dx.$$

# Curs 5

Cristian Niculescu

## 1 Galerie a variabilelor aleatoare continue

### 1.1 Scopurile învățării

1. Să poată da exemple de ceea ce modeleză repartițiile uniformă, exponențială și normală.
2. Să poată da domeniile de valori și pdf-urile repartițiilor uniformă, exponențială și normală.

### 1.2 Introducere

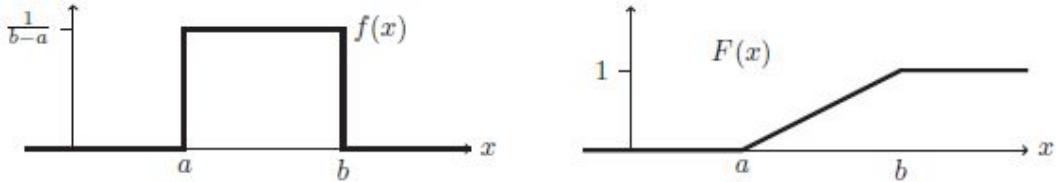
Introducem câteva repartiții continue fundamentele. Pentru fiecare repartitie dăm domeniul de valori, pdf, cdf și o scurtă descriere a situațiilor pe care le modeleză. Toate aceste repartiții depind de parametri, pe care îi specificăm.

Cu toate că o abordăm spre final, repartitia normală este cea mai importantă din cele definite aici. Când dăm funcția de repartitie (cdf, prescurtat "repartitia") se subînțelege, dacă este cazul, că la stânga intervalului pe care este dată este 0, iar la dreapta este 1.

### 1.3 Repartitia uniformă

1. Parametri:  $a < b$ .
2. Domeniu de valori:  $[a, b]$ .
3. Notație:  $\text{uniform}(a, b)$  sau  $U(a, b)$ .
4. Densitate:  $f(x) = \frac{1}{b-a}$  pentru  $a \leq x \leq b$ .
5. Repartitia:  $F(x) = (x - a)/(b - a)$  pentru  $a \leq x \leq b$ .
6. Modele: Toate rezultatele din domeniul de valori au probabilitate egală (mai precis, toate rezultatele au aceeași densitate de probabilitate).

Grafice pentru pdf și cdf:



**Exemplu.** 1. Presupunem că avem o ruletă marcată în milimetri. Dacă măsurăm (până la cel mai apropiat marcaj) lungimea unor articole care au aproximativ 1 metru, eroarea de rotunjire va fi repartizată uniform între -0.5 și 0.5 milimetri.

2. Multe jocuri de tablă folosesc săgeți care se învârt pentru a introduce hazardul. Când este învârtită, săgeata se oprește la un unghi care este repartizat uniform între 0 și  $2\pi$  radiani.

3. În cele mai multe generatoare de numere pseudo-aleatoare, generatorul de bază simulează o repartiție uniformă și toate celelalte repartiții sunt construite transformând generatorul de bază.

## 1.4 Repartiția exponențială

1. Parametru:  $\lambda > 0$ .
2. Domeniu de valori:  $[0, \infty)$ .
3. Notație:  $\text{exponential}(\lambda)$  sau  $\exp(\lambda)$ .
4. Densitate:  $f(x) = \lambda e^{-\lambda x}$  pentru  $x \geq 0$ .
5. Repartiție:  $F(x) = 1 - e^{-\lambda x}$  pentru  $x \geq 0$ .
6. Repartiția cozii drepte:  $P(X > x) = 1 - F(x) = e^{-\lambda x}$ .
7. Modele: Timpul de așteptare pentru un proces continuu de schimbare a stării.

**Exemplu.** 1. Dacă ies afară după curs și aștept un taxi, timpul meu de așteptare în minute este repartizat exponențial. În acest caz  $\lambda$  este dat de 1 supra numărul mediu de taxiuri care trec pe minut (în această perioadă de timp din zilele lucrătoare).

2. Repartiția exponențială modelează timpul de așteptare până când un izotop instabil suferă o descompunere nucleară. În acest caz  $\lambda$  este legat de timpul de înjumătărire al izotopului.

**Proprietatea de lipsă a memoriei:** Sunt și alte repartiții care modelează timpii de așteptare, dar repartiția exponențială are proprietatea adițională că este lipsită de memorie. Iată ce înseamnă aceasta în contextul exemplului 1. Presupunem că probabilitatea ca un taxi să sosească în primele 5 minute este  $p$ . Dacă aștept 5 minute și de fapt nu sosește niciun taxi, atunci probabilitatea ca un taxi să sosească în următoarele 5 minute este tot  $p$ .

Din contra, să presupunem că merg la o stație de metrou și aștept următorul tren. Deoarece trenurile sunt coordonate să urmeze un program (de exemplu, trenul 1 pleacă la ora 8:00 și următorul să fie la ora 8:15), atunci probabilitatea ca următorul să plece la ora 8:15 este de 1.

plu, cel mult 12 minute între trenuri), dacă aştept 5 minute fără să văd un tren, atunci este o probabilitate mult mai mare ca trenul să sosescă în următoarele 5 minute. În particular, timpul de aşteptare pentru metrou nu este fără memorie și un model mai bun pentru el ar fi repartiția uniformă pe intervalul  $[0,12]$ .

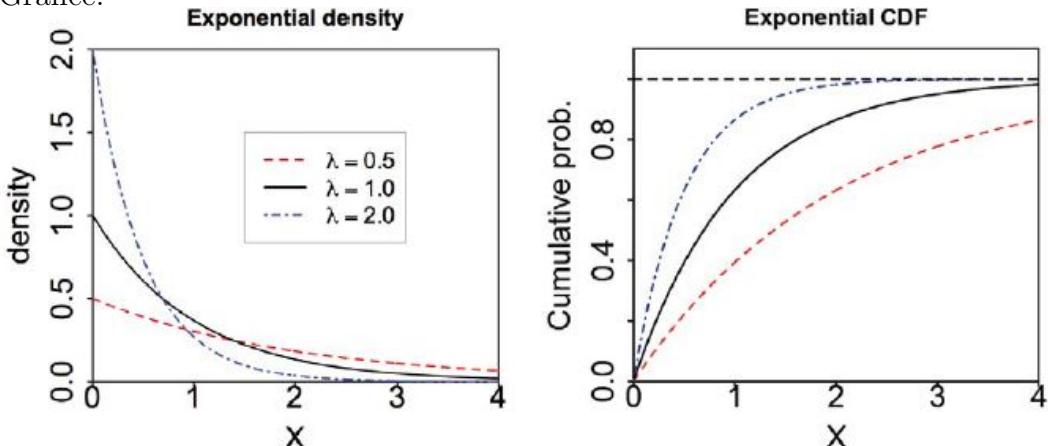
Lipsa de memorie a repartiției exponențiale este analoagă lipsei de memorie a repartiției geometrice (discrete), unde faptul că am aruncat 5 reversuri la rând nu ne dă nicio informație despre următoarele 5 aruncări.

Într-adevăr, repartiția exponențială este precis perechea continuă a repartiției geometrice, care modeleză timpul de aşteptare pentru un proces discret de schimbare de stări. Mai formal, lipsa de memorie înseamnă că probabilitatea de a aştepta mai mult de  $t$  minute este neafectată de faptul că am aşteptat deja  $s$  minute fără ca evenimentul să se producă. În simboluri,  $P(X > s + t | X > s) = P(X > t)$ .

**Demonstrația lipsei de memorie:** Deoarece  $(X > s + t) \cap (X > s) = (X > s + t)$ , avem

$$P(X > s + t | X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t), \text{ q.e.d.}$$

Graifice:



## 1.5 Repartiția normală

În 1809, Carl Friedrich Gauss a publicat o monografie introducând câteva noțiuni care au devenit fundamentale în statistică: repartiția normală, estimarea de verosimilitate maximă și metoda celor mai mici pătrate. Din acest motiv, repartiția normală mai este numită și repartiția *Gaussiană*. Ea este cea mai importantă repartiție continuă.

1. Parametri:  $\mu \in \mathbb{R}, \sigma > 0$ .

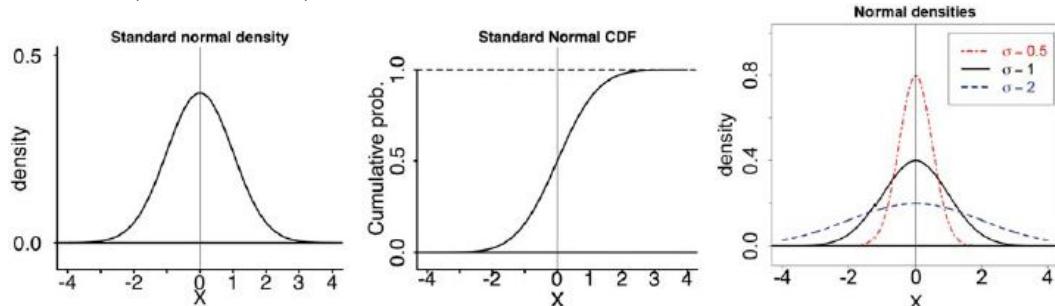
2. Domeniul de valori:  $\mathbb{R}$ .
3. Notație:  $\text{normal}(\mu, \sigma^2)$  sau  $N(\mu, \sigma^2)$ .
4. Densitate:  $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$ .
5. Repartiție:  $F(x)$  nu are formulă, aşa că folosim tabele sau comenzi ca `pnorm` în R pentru a calcula  $F(x)$ .
6. Modele: măsurarea erorii, inteligenței/abilității, înălțimii, mediilor laturilor de date.

**Repartiția normală standard**  $N(0, 1)$  are media 0 și dispersia 1. Rezervăm  $Z$  pentru o variabilă aleatoare normală standard,  $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$  pentru densitatea normală standard și  $\Phi(z)$  pentru repartiția normală standard.

Observație: Media și dispersia variabilelor aleatoare continue au aceeași interpretare ca în cazul discret. Repartiția normală  $N(\mu, \sigma^2)$  are media  $\mu$ , dispersia  $\sigma^2$  și deviația standard  $\sigma$ .

Iată câteva grafice ale repartiției normale. Observăm că densitățile au graficele curbe în formă de *clopot*. Mai observăm că pe măsură ce  $\sigma$  crește, ele devin din ce în ce mai împrăștiate.

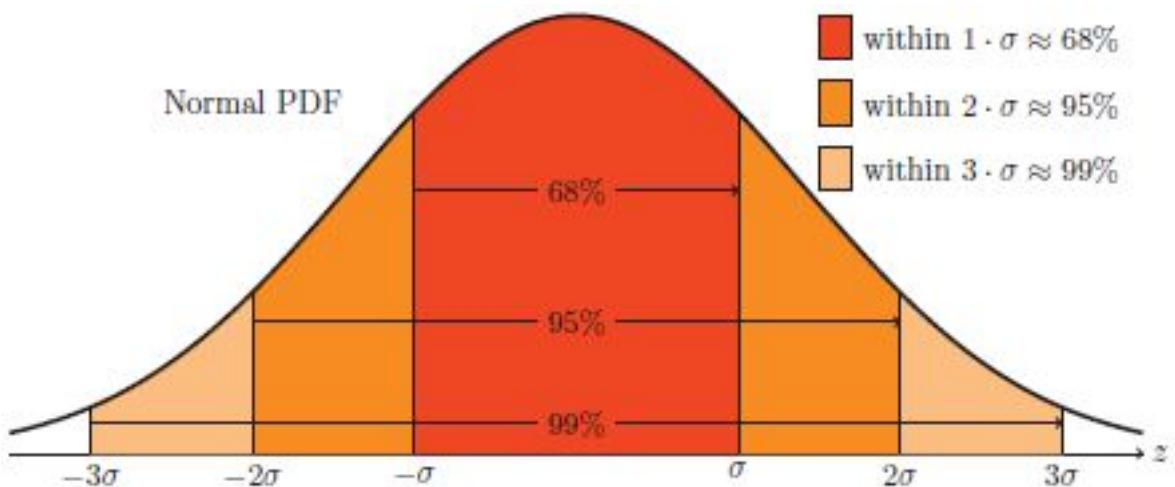
Grafice: (*curba clopot*):



### 1.5.1 Probabilități normale

Pentru a face aproximări este util să ne reamintim următoarea regulă a degetului mare pentru 3 probabilități aproximative

$$P(-1 \leq Z \leq 1) \approx 0.68, \quad P(-2 \leq Z \leq 2) \approx 0.95, \quad P(-3 \leq Z \leq 3) \approx 0.99.$$

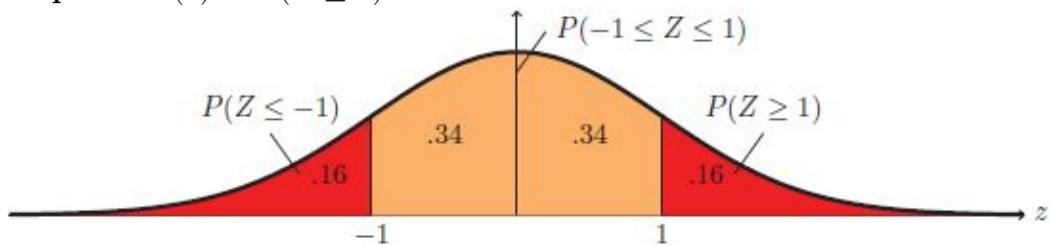


### Calcule cu simetrie

Putem folosi simetria graficului densității normale standard față de dreapta  $x = 0$  pentru a face unele calcule.

**Exemplul 1.** Regula degetului mare spune  $P(-1 \leq Z \leq 1) \approx 0.68$ . Folosiți aceasta pentru a estima  $\Phi(1)$ .

**Răspuns:**  $\Phi(1) = P(Z \leq 1)$ .



În figură, cele 2 cozi (în roșu) au împreună aria  $1 - 0.68 = 0.32$ . Din simetrie, coada stângă are aria 0.16 (jumătate din 0.32), deci  $P(Z \leq 1) \approx 0.16 + 0.68 = 0.84$ .

### 1.5.2 Folosirea lui R pentru a calcula $\Phi(z)$

```
# Folosim funcția R pnorm(x, μ, σ) pentru a calcula F(x) pentru N(μ, σ²).
pnorm(1,0,1)
[1] 0.8413447
pnorm(0,0,1)
[1] 0.5
pnorm(1,0,2)
[1] 0.6914625
pnorm(1,0,1)-pnorm(-1,0,1)
[1] 0.6826895
```

```

pnorm(5,0,5)-pnorm(-5,0,5)
[1] 0.6826895
# Desigur, z poate fi un vector de valori:
pnorm(c(-3,-2,-1,0,1,2,3),0,1)
[1] 0.001349898 0.022750132 0.158655254 0.500000000 0.841344746 0.977249868
[7] 0.998650102

```

**Observație:** Funcția R  $\text{pnorm}(x, \mu, \sigma)$  utilizează  $\sigma$  în timp ce notația noastră pentru repartiția normală  $N(\mu, \sigma^2)$  folosește  $\sigma^2$ .

Iată un tabel de valori cu o acuratețe cu mai puține zecimale:

$z:$	-2	-1	0	.3	.5	1	2	3
$\Phi(z):$	0.0228	0.1587	0.5000	0.6179	0.6915	0.8413	0.9772	0.9987

**Exemplul 2.** Utilizați R pentru a calcula  $P(-1.5 \leq Z \leq 2)$ .

**Răspuns:**  $P(-1.5 \leq Z \leq 2) = \Phi(2) - \Phi(-1.5) = \text{pnorm}(2, 0, 1) - \text{pnorm}(-1.5, 0, 1) = 0.9104427$ .

## 1.6 Repartiția Pareto

1. Parametri:  $m > 0$  și  $\alpha > 0$ .
2. Domeniul de valori:  $[m, \infty)$ .
3. Notație:  $\text{Pareto}(m, \alpha)$ .
4. Densitate:  $f(x) = \frac{\alpha m^\alpha}{x^{\alpha+1}}$ .
5. Repartiția:  $F(x) = 1 - \frac{m^\alpha}{x^\alpha}$ , pentru  $x \geq m$ .
6. Repartiția coadă:  $P(X > x) = m^\alpha/x^\alpha$ , pentru  $x \geq m$ .
7. Modele: Repartiția Pareto modeleză o **lege de putere**, unde probabilitatea ca un eveniment să aibă loc variază ca o putere a unui atribut al evenimentului. Multe fenomene urmează o lege de putere, ca mărimea meteoziilor, nivelurile veniturilor într-o populație și nivelurile populației în orașe. Vezi [http://en.wikipedia.org/wiki/Pareto\\_distribution#Applications](http://en.wikipedia.org/wiki/Pareto_distribution#Applications).

## 2 Manipularea variabilelor aleatoare continue

### 2.1 Scopul învățării

Să poată să afle pdf și cdf ale unei variabile aleatoare definite în termeni de o variabilă aleatoare cu pdf și cdf cunoscute.

### 2.2 Transformări de variabile aleatoare

Dacă  $Y = aX + b$ , atunci proprietățile mediei și dispersiei ne spun că  $E(Y) = aE(X) + b$  și  $Var(Y) = a^2Var(X)$ . Dar care este funcția de

repartiție a lui  $Y$ ? Dacă  $Y$  este continuă, care este pdf a ei?

Adesea, pentru transformarea variabilelor aleatoare discrete, lucrăm cu tabele. Pentru variabile aleatoare continue transformarea pdf este chiar schimbarea de variabilă de la analiza matematică. Transformarea cdf folosește direct definiția cdf.

Reamintim:

1. Cdf a lui  $X$  este  $F_X(x) = P(X \leq x)$ .
2. Pdf a lui  $X$  este legată de  $F_X$  prin  $f_X(x) = F'_X(x)$ .

**Exemplul 1.** Fie  $X \sim U(0, 2)$ , deci  $f_X(x) = 1/2$  și  $F_X(x) = x/2$  pe  $[0, 2]$ . Care sunt domeniul de valori, pdf și cdf ale lui  $Y = X^2$ ?

**Răspuns:** Domeniul de valori al lui  $Y$  este  $[0, 4]$ .

Pentru a afla cdf folosim definiția:

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}) = \sqrt{y}/2.$$

Pentru a afla pdf derivăm cdf:

$$f_Y(y) = F'_Y(y) = \left(\frac{\sqrt{y}}{2}\right)' = \frac{1}{4\sqrt{y}}.$$

Un mod alternativ de a afla pdf este prin schimbare de variabilă. Trucul aici este să ne reamintim că  $f_X(x)dx$  dă probabilitatea ( $f_X(x)$  este densitatea de probabilitate). Iată calculul:

$$\begin{aligned} y &= x^2 \Rightarrow dy = 2xdx \Rightarrow dx = \frac{dy}{2\sqrt{y}} \\ f_X(x)dx &= \frac{dx}{2} = \frac{dy}{4\sqrt{y}} = f_Y(y)dy. \end{aligned}$$

De aceea,  $f_Y(y) = \frac{1}{4\sqrt{y}}$ .

**Exemplul 2.** Fie  $X \sim \exp(\lambda)$ , deci  $f_X(x) = \lambda e^{-\lambda x}$  pe  $[0, \infty)$ . Care este densitatea lui  $Y = X^2$ ?

**Răspuns:** Folosim schimbarea de variabilă:

$$\begin{aligned} y &= x^2 \Rightarrow dy = 2xdx \Rightarrow dx = \frac{dy}{2\sqrt{y}} \\ f_X(x)dx &= \lambda e^{-\lambda x}dx = \lambda e^{-\lambda\sqrt{y}} \frac{dy}{2\sqrt{y}} = f_Y(y)dy. \end{aligned}$$

De aceea,  $f_Y(y) = \frac{\lambda}{2\sqrt{y}} e^{-\lambda\sqrt{y}}$ .

**Exemplul 3.** Presupunem  $X \sim N(5, 3^2)$ . Arătați că  $Z = \frac{X-5}{3}$  este normală standard, i.e.,  $Z \sim N(0, 1)$ .

**Răspuns:** Din nou folosind schimbarea de variabilă și formula lui  $f_X(x)$  avem:

$$z = \frac{x - 5}{3} \Rightarrow dz = \frac{dx}{3} \Rightarrow dx = 3dz$$

$$f_X(x)dx = \frac{1}{3\sqrt{2\pi}}e^{-(x-5)^2/(2\cdot 3^2)}dx = \frac{1}{3\sqrt{2\pi}}e^{-z^2/2}3dz = f_Z(z)dz.$$

De aceea,  $f_Z(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ . Deoarece aceasta este exact densitatea pentru  $N(0, 1)$ , am arătat că  $Z$  este normală standard.

Acest exemplu a arătat o proprietate generală importantă a variabilelor aleatoare normale pe care o dăm în exemplul următor.

**Exemplul 4.** Presupunem  $X \sim N(\mu, \sigma^2)$ . Arătați că  $Z = \frac{X-\mu}{\sigma}$  este normală standard, i.e.,  $Z \sim N(0, 1)$ .

**Răspuns.** Este exact același calcul ca la exemplul trecut cu  $\mu$  înlocuind pe 5 și  $\sigma$  pe 3:

$$z = \frac{x - \mu}{\sigma} \Rightarrow dz = \frac{dx}{\sigma} \Rightarrow dx = \sigma dz$$

$$f_X(x)dx = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\cdot\sigma^2)}dx = \frac{1}{\sigma\sqrt{2\pi}}e^{-z^2/2}\sigma dz = f_Z(z)dz.$$

De aceea,  $f_Z(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ . Aceasta arată că  $Z$  este normală standard.

# Curs 6

Cristian Niculescu

## 1 Media, dispersia și deviația standard pentru variabile aleatoare continue

### 1.1 Scopurile învățării

1. Să poată să calculeze și să interpreteze media, dispersia și deviația standard pentru variabile aleatoare continue.
2. Să poată să calculeze și să interpreteze quantilele pentru variabile aleatoare continue sau discrete.

### 1.2 Introducere

Până acum am studiat media, deviația standard și dispersia pentru variabile aleatoare discrete. Aceste statistici de rezumat au același înțeles pentru variabile aleatoare continue:

- Media  $\mu = E(X)$  este o măsură a tendinței centrale.
- Deviația standard  $\sigma$  este o măsură a împrăștierii.
- Dispersia  $\sigma^2 = Var(X)$  este pătratul deviației standard.

Pentru a trece de la discret la continuu, pur și simplu înlocuim sumele din formule cu integrale. Un alt tip de [statistică de rezumat](#) sunt [quantilele](#). 0.5 quantila unei repartiții se mai numește [mediană](#) sau a 50-a percentilă.

### 1.3 Media unei variabile aleatoare continue

**Definiție:** Fie  $X$  o variabilă aleatoare continuă cu domeniul de valori  $[a, b]$  (cu convenția că, dacă  $a = -\infty$  sau  $b = \infty$ , intervalul este deschis în acel capăt) și funcția densitate de probabilitate  $f(x)$ . [Media](#) lui  $X$  este

$$E(X) = \int_a^b x f(x) dx.$$

Să vedem cum se compară aceasta cu formula pentru o variabilă aleatoare discretă:

$$E(X) = \sum_{i=1}^n x_i p(x_i).$$

Formula discretă spune să luăm o sumă ponderată a valorilor  $x_i$  ale lui  $X$ , unde ponderile sunt probabilitățile  $p(x_i)$ . Reamintim că  $f(x)$  este o **densitate** de probabilitate. Unitățile ei de măsură sunt prob/(unitatea de măsură a lui  $X$ ). Deci  $f(x)dx$  reprezintă probabilitatea că  $X$  este într-un domeniu de valori infinitezimal de lățime  $dx$  în jurul lui  $x$ . Astfel, putem interpreta formula pentru  $E(X)$  ca o integrală ponderată de valorile  $x$  ale lui  $X$ , unde ponderile sunt probabilitățile  $f(x)dx$ .

### 1.3.1 Exemple

**Exemplul 1.** Fie  $X \sim U(0, 1)$ . Aflați  $E(X)$ .

**Răspuns:**  $X$  are domeniul de valori  $[0,1]$  și densitatea  $f(x) = 1$ . De aceea,

$$E(X) = \int_0^1 xf(x)dx = \int_0^1 xdx = \frac{x^2}{2} \Big|_0^1 = \frac{1}{2}.$$

Media este la mijlocul domeniului de valori.

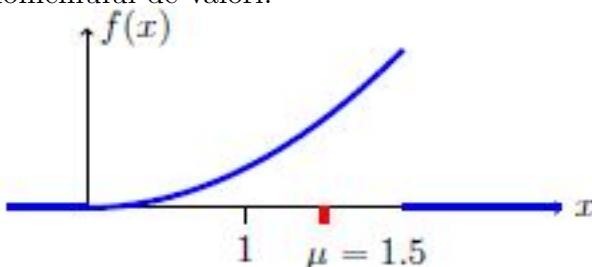
**Exemplul 2.** Fie  $X$  cu domeniul de valori  $[0,2]$  și densitatea  $f(x) = \frac{3}{8}x^2$ . Aflați  $E(X)$ .

**Răspuns:**

$$E(X) = \int_0^2 xf(x)dx = \int_0^2 \frac{3}{8}x^3 dx = \frac{3x^4}{32} \Big|_0^2 = \frac{3}{2}.$$

Are sens că  $X$  are media în jumătatea dreaptă a domeniului lui de valori?

**Răspuns:** Da. Deoarece densitatea de probabilitate crește când  $x$  crește în domeniul de valori, media lui  $X$  ar trebui să fie în jumătatea dreaptă a domeniului de valori.



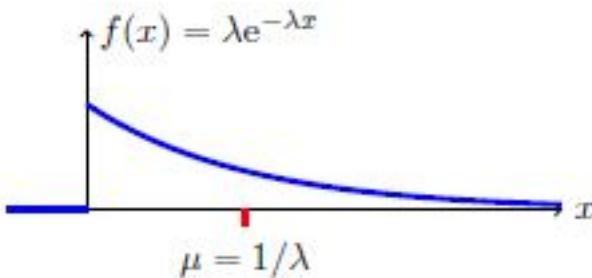
$\mu$  este "tras" la dreapta mijlocului 1 deoarece este mai multă masă la dreapta.

**Exemplul 3.** Fie  $X \sim exp(\lambda)$ . Aflați  $E(X)$ .

**Răspuns:** Domeniul de valori al lui  $X$  este  $[0, \infty)$  și pdf este  $f(x) = \lambda e^{-\lambda x}$ . Deci

$$\begin{aligned} E(X) &= \int_0^\infty x f(x) dx = \int_0^\infty \lambda x e^{-\lambda x} dx = \int_0^\infty x (-e^{-\lambda x})' dx \\ &= -x e^{-\lambda x} \Big|_0^\infty + \int_0^\infty x' e^{-\lambda x} dx = 0 + \int_0^\infty e^{-\lambda x} dx = -\frac{e^{-\lambda x}}{\lambda} \Big|_0^\infty = \frac{1}{\lambda}. \end{aligned}$$

Am folosit integrarea prin părți și faptul că  $\lim_{x \rightarrow \infty} x e^{-\lambda x} = \lim_{x \rightarrow \infty} e^{-\lambda x} = 0$ .



Media unei variabile aleatoare exponențiale

**Exemplul 4.** Fie  $Z \sim N(0, 1)$ . Aflați  $E(Z)$ .

**Răspuns:** Domeniul de valori al lui  $Z$  este  $\mathbb{R}$  și pdf este  $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ . Avem  $E(Z) = \int_{-\infty}^{\infty} z \phi(z) dz$ . Arătăm că integrala converge, i.e. media chiar există (unele variabile aleatoare nu au medie).

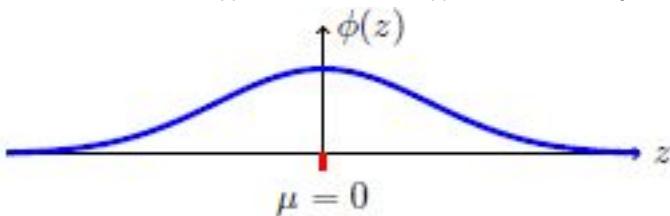
$$\int_0^\infty z \phi(z) dz = \frac{1}{\sqrt{2\pi}} \int_0^\infty z e^{-z^2/2} dz.$$

Substituția  $u = z^2/2$  dă  $du = zdz$ . Deci integrala devine

$$\frac{1}{\sqrt{2\pi}} \int_0^\infty z e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-u} du = -\frac{1}{\sqrt{2\pi}} e^{-u} \Big|_0^\infty = \frac{1}{\sqrt{2\pi}}.$$

Analog,  $\int_{-\infty}^0 z \phi(z) dz = -\frac{1}{\sqrt{2\pi}}$ .

Deci  $E(Z) = \int_{-\infty}^{\infty} z \phi(z) dz = \int_{-\infty}^0 z \phi(z) dz + \int_0^{\infty} z \phi(z) dz = -\frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} = 0$ .



Repartiția normală standard este simetrică și are media 0.

### 1.3.2 Proprietățile lui $E(X)$

Proprietățile lui  $E(X)$  pentru variabile aleatoare continue sunt aceleași ca pentru cele discrete:

1. Dacă  $X$  și  $Y$  sunt variabile aleatoare continue pe un spațiu al probelor  $\Omega$ , atunci

$$E(X + Y) = E(X) + E(Y). \quad (\text{liniaritate I})$$

2. Dacă  $a$  și  $b$  sunt constante, atunci

$$E(aX + b) = aE(X) + b. \quad (\text{liniaritate II})$$

**Exemplul 5.** Verificați că pentru  $X \sim N(\mu, \sigma^2)$  avem  $E(X) = \mu$ .

**Răspuns:** În exemplul 4 am arătat că pentru  $Z$  normală standard,  $E(Z) = 0$ . Am putea mima calculul de acolo pentru a arăta că  $E(X) = \mu$ . În loc de aceasta folosim proprietățile de liniaritate ale lui  $E(X)$ . La manipularea variabilelor aleatoare am arătat că, dacă  $X \sim N(\mu, \sigma^2)$ , o putem **standardiza**:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

De aici avem  $X = \sigma Z + \mu$ . Liniaritatea mediei dă

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \sigma \cdot 0 + \mu = \mu.$$

### 1.3.3 Media funcțiilor de $X$

Aceasta merge exact ca în cazul discret. Dacă  $h$  este o funcție continuă, atunci  $Y = h(X)$  este o variabilă aleatoare și

$$E(Y) = E(h(X)) = \int_{-\infty}^{\infty} h(x)f_X(x)dx.$$

**Exemplul 6.** Fie  $X \sim \exp(\lambda)$ . Aflați  $E(X^2)$ .

**Răspuns.** Integrăm prin părți de 2 ori:

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2 f(x)dx = \int_0^{\infty} \lambda x^2 e^{-\lambda x} dx = \int_0^{\infty} x^2 (-e^{-\lambda x})' dx \\ &= -x^2 e^{-\lambda x} \Big|_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x} dx = 0 + \int_0^{\infty} 2x \left(-\frac{e^{-\lambda x}}{\lambda}\right)' dx \\ &= -2x \frac{e^{-\lambda x}}{\lambda} \Big|_0^{\infty} + 2 \int_0^{\infty} \frac{e^{-\lambda x}}{\lambda} dx = 0 - 2 \frac{e^{-\lambda x}}{\lambda^2} \Big|_0^{\infty} = \frac{2}{\lambda^2}. \end{aligned}$$

## 1.4 Dispersia

Definiția dispersiei este identică cu cea pentru variabile aleatoare discrete.

**Definiție:** Fie  $X$  o variabilă aleatoare continuă cu media  $\mu$ . **Dispersia** lui  $X$  este

$$Var(X) = E((X - \mu)^2).$$

### 1.4.1 Proprietăți ale dispersiei

Acestea sunt exact aceleași ca în cazul discret.

1. Dacă  $X$  și  $Y$  sunt **independente**, atunci  $Var(X + Y) = Var(X) + Var(Y)$ .
2. Pentru constantele  $a$  și  $b$ ,  $Var(aX + b) = a^2Var(X)$ .
3. **Teorema:**  $Var(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$ .

Proprietatea 3 dă o formulă pentru  $Var(X)$  care este adesea mai ușor de folosit la calcule. Demonstrațiile proprietăților 2 și 3 sunt analoage celor din cazul discret.

**Exemplul 7.** Fie  $X \sim U(0, 1)$ . Aflați  $Var(X)$  și  $\sigma_X$ .

**Răspuns:** În exemplul 1 am aflat  $\mu = 1/2$ .

$$\begin{aligned} Var(X) &= E((X - \mu)^2) = \int_0^1 (x - 1/2)^2 dx = \frac{(x - 1/2)^3}{3} \Big|_0^1 = \frac{1}{12}. \\ \sigma_X &= \sqrt{Var(X)} = \frac{1}{\sqrt{12}} = \frac{1}{2\sqrt{3}} = \frac{\sqrt{3}}{6}. \end{aligned}$$

**Exemplul 8.** Fie  $X \sim exp(\lambda)$ . Aflați  $Var(X)$  și  $\sigma_X$ .

**Răspuns:** În exemplele 3 și 6 am calculat:

$$E(X) = \frac{1}{\lambda} \text{ și } E(X^2) = \frac{2}{\lambda^2}.$$

Deci, din proprietatea 3,

$$Var(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \text{ și } \sigma_X = \sqrt{Var(X)} = \frac{1}{\lambda}.$$

Puteam să calculăm și direct din  $Var(X) = \int_0^\infty (x - 1/\lambda)^2 \lambda e^{-\lambda x} dx$ .

**Exemplul 9.** Fie  $Z \sim N(0, 1)$ . Arătați că  $Var(Z) = 1$ .

**Răspuns:** Deoarece  $E(Z) = 0$ , avem

$$\begin{aligned} Var(Z) &= E(Z^2) - E(Z)^2 = E(Z^2) - 0^2 = E(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z (-e^{-z^2/2})' dz = \frac{1}{\sqrt{2\pi}} \left( -ze^{-z^2/2} \Big|_{-\infty}^{\infty} \right) + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz. \end{aligned}$$

Primul termen este 0 deoarece exponențiala tinde la 0 mai repede decât  $z$  tinde la  $\pm\infty$ . Al 2-lea termen este 1 deoarece este exact integrala probabilității totale a pdf  $\phi(z)$  pentru  $N(0, 1)$ . Deci  $Var(Z) = 1$ .

**Exemplul 10.** Fie  $X \sim N(\mu, \sigma^2)$ . Arătați că  $Var(X) = \sigma^2$ .

**Răspuns:** Făcând schimbarea de variabilă  $z = (x - \mu)/\sigma$ , avem

$$\begin{aligned} Var(X) &= E((X - \mu)^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \sigma^2. \end{aligned}$$

Ultima integrală a fost calculată la exemplul 9.

## 1.5 Quantile

**Definiție:** **Mediana** lui  $X$  este valoarea  $x$  pentru care  $P(X \leq x) = 0.5$ , i.e. valoarea lui  $x$  astfel încât  $P(X \leq x) = P(X \geq x)$ . Cu alte cuvinte,  $X$  are probabilitate egală de a fi deasupra sau sub mediană și de aceea fiecare probabilitate este  $1/2$ . În termeni de cdf  $F(x) = P(X \leq x)$ , putem defini echivalent mediana ca valoarea lui  $x$  satisfăcând  $F(x) = 0.5$ .

**Gândiți:** Care este mediana lui  $Z$ ?

**Răspuns:** Din simetrie, mediana este 0.

**Exemplul 11.** Aflați mediana lui  $X \sim exp(\lambda)$ .

**Răspuns:** Cdf a lui  $X$  este  $F(x) = 1 - e^{-\lambda x}$  pe  $[0, \infty)$ . Mediana este valoarea lui  $x$  pentru care  $F(x) = 1 - e^{-\lambda x} = 0.5$ . Rezolvând ecuația obținem  $x = (\ln 2)/\lambda$ .

În acest caz mediana este mai mică decât media  $\mu = 1/\lambda$ .

**Definiție:** A  $p$ -a quantilă a lui  $X$  este valoarea  $q_p$  astfel încât  $P(X \leq q_p) = p$ .

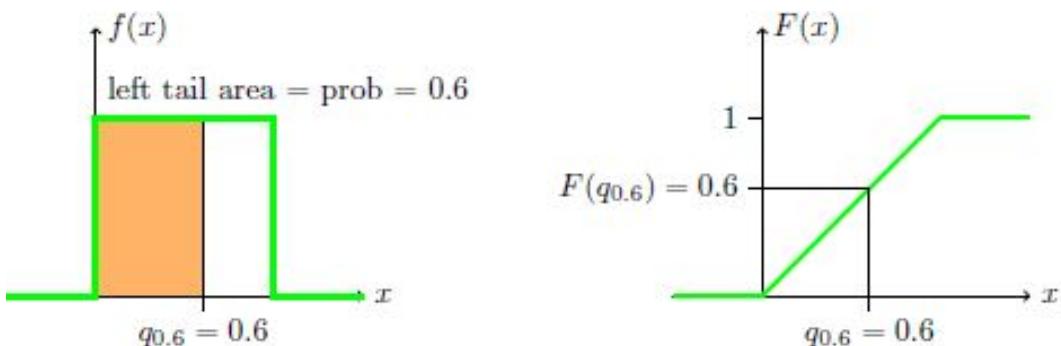
**Observații.** 1. Cu această notație mediana este  $q_{0.5}$ .

2. În termeni de cdf:  $F(q_p) = p$ .

În raport cu pdf  $f$ , quantila  $q_p$  este valoarea astfel încât sub graficul lui  $f$  este o arie de  $p$  la stânga lui  $q_p$  și o arie de  $1 - p$  la dreapta lui  $q_p$ . În exemplele de mai jos, observați cum putem reprezenta quantila grafic folosind aria de sub graficul pdf sau înălțimea cdf.

**Exemplul 12.** Aflați 0.6 quantila pentru  $X \sim U(0, 1)$ .

**Răspuns.** Cdf pentru  $X$  este  $F(x) = x$  pe domeniul de valori  $[0, 1]$ . Din  $F(q_{0.6}) = 0.6 \Rightarrow q_{0.6} = 0.6$ .

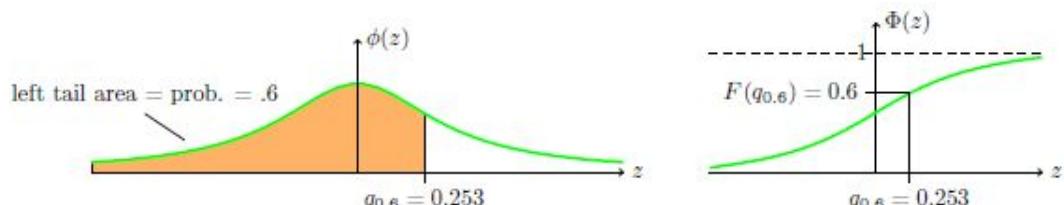


$q_{0.6}$ : aria cozii stângi  $= 0.6 \Leftrightarrow F(q_{0.6}) = 0.6$

**Exemplul 13.** Aflați 0.6 quantila repartiției normale standard.

**Răspuns.** Nu avem o formulă pentru cdf, deci vom folosi "funcția quantilă" din R, `qnorm`.

$$q_{0.6} = \text{qnorm}(0.6, 0, 1) = 0.2533471$$



$q_{0.6}$ : aria cozii stângi  $= 0.6 \Leftrightarrow F(q_{0.6}) = 0.6$

Quantilele dau o măsură utilă a **locației** pentru o variabilă aleatoare.

### 1.5.1 Percentile, decile, quartile

Pentru confort, quantilele sunt adesea descrise în termeni de percentile, decile sau quartile. A 60-a **percentilă** este 0.6 quantila. De exemplu, sunteți în a 60-a percentilă pentru înălțime dacă sunteți mai înalt(ă) decât 60% din populație, i.e. **probabilitatea** să fiți mai înalt(ă) decât o persoană aleasă aleator este 60%.

De asemenea, **decilele** reprezintă pași de 1/10. A 3-a decilă este 0.3 cuantila.

**Quartilele** sunt în pași de 1/4. A 3-a quartilă este 0.75 quantila și a 75-a percentilă.

## 2 Teorema limită centrală și legea numerelor mari

### 2.1 Scopurile învățării

1. Să înțeleagă enunțul legii numerelor mari.
2. Să înțeleagă enunțul teoremei limită centrală.
3. Să poată să utilizeze teorema limită centrală pentru a aproxima probabilități ale mediilor și sumelor de variabile aleatoare independente identic distribuite.

### 2.2 Introducere

Media multor măsurări ale aceleiași cantități necunoscute tinde să dea o estimare mai bună decât o singură măsurare. Intuitiv, aceasta este deoarece eroarea aleatoare a fiecărei măsurări se reduce în medie. 2 moduri de a face precisă această intuiție sunt legea numerelor mari (LoLN) și teorema limită centrală (CLT).

Scurt, atât legea numerelor mari cât și teorema limită centrală sunt despre multe date independente din aceeași repartiție. LoLN ne spune 2 lucruri:

1. Media multor date independente este (cu mare probabilitate) aproape de media repartiției de bază.
2. Histograma densității multor date independente este (cu mare probabilitate) aproape de graficul densității repartiției de bază.

Pentru a fi absolut corectă matematic trebuie să facem aceste afirmații mai precise, dar, aşa cum au fost făcute, sunt un bun mod de a gândi despre legea numerelor mari.

Teorema limită centrală spune că suma sau media multor copii independente ale unei variabile aleatoare este aproximativ o variabilă aleatoare normală. CLT continuă dând valori precise pentru media și deviația standard ale variabilei normale.

Adesea în practică  $n$  nu trebuie să fie foarte mare. Valorile  $n > 30$  sunt adesea suficiente.

#### 2.2.1 Este mai multă experimentare decât matematică

Matematica LoLN spune că media unui lot de date independente dintr-o variabilă aleatoare se va apropia aproape sigur de media variabilei. Matematica **nu ne spune** dacă instrumentul sau experimentul produce date care merită să li se facă media. De exemplu, dacă dispozitivul de măsurare este defect sau

prost calibrat, atunci media multor măsurări va fi o estimare foarte precisă a unui lucru greșit! Acesta este un exemplu de [eroare sistematică](#) sau [deplasare a datelor](#), în contrast cu [eroarea aleatoare](#) controlată de legea numerelor mari.

### 2.3 Legea numerelor mari

Presupunem că  $X_1, X_2, \dots, X_n$  sunt variabile aleatoare independente cu aceeași repartiție. În acest caz, spunem că  $X_i$  sunt [independente și identic distribuite](#) sau [i.i.d.](#). În particular,  $X_i$  au aceeași medie  $\mu$  și deviație standard  $\sigma$ .

Fie  $\bar{X}_n$  media lui  $X_1, X_2, \dots, X_n$ :

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i.$$

$\bar{X}_n$  este ea însăși o variabilă aleatoare. Legea numerelor mari și teorema limită centrală ne spun despre valoarea, respectiv, repartiția lui  $\bar{X}_n$ .

**LoLN:** Când  $n$  crește, probabilitatea că  $\bar{X}_n$  este aproape de  $\mu$  tinde la 1.

**CLT:** Când  $n$  crește, repartiția lui  $\bar{X}_n$  converge la repartiția normală  $N(\mu, \sigma^2/n)$ .

**Exemplul 1. Medii ale variabilelor aleatoare Bernoulli**

Presupunem că fiecare  $X_i$  este o aruncare independentă a unei monede corecte, deci  $X_i \sim \text{Bernoulli}(0.5)$  și  $\mu = 0.5$ . Atunci  $\bar{X}_n$  este proporția de aversuri în  $n$  aruncări și ne așteptăm că această proporție este aproape de 0.5 pentru  $n$  mare. Acest fapt nu este garantat; de exemplu, putem obține 1000 de aversuri în 1000 de aruncări, deși probabilitatea ca aceasta să aibă loc este foarte mică.

[Cu mare probabilitate](#) media de selecție  $\bar{X}_n$  este aproape de media 0.5 pentru  $n$  mare. Vom verifica asta făcând unele calcule în R.

La început, vom considera probabilitatea de a fi în 0.1 a mediei. Putem exprima această probabilitate ca

$$P(|\bar{X}_n - 0.5| \leq 0.1) = P(0.4 \leq \bar{X}_n \leq 0.6).$$

Legea numerelor mari spune că această probabilitate tinde la 1 când numărul de aruncări  $n$  devine mare. Codul R produce următoarele valori pentru  $P(0.4 \leq \bar{X}_n \leq 0.6)$ :

$$n = 10 : \text{pbinom}(6, 10, 0.5) - \text{pbinom}(3, 10, 0.5) = 0.65625$$

$$n = 50 : \text{pbinom}(30, 50, 0.5) - \text{pbinom}(19, 50, 0.5) = 0.8810795$$

$$n = 100 : \text{pbinom}(60, 100, 0.5) - \text{pbinom}(39, 100, 0.5) = 0.9647998$$

$$n = 500 : \text{pbinom}(300, 500, 0.5) - \text{pbinom}(199, 500, 0.5) = 0.9999941$$

$$n = 1000 : \text{pbinom}(600, 1000, 0.5) - \text{pbinom}(399, 1000, 0.5) = 1$$

După cum s-a prezis de LoLN, probabilitatea tinde la 1 când  $n$  crește. Refacem aceste calcule pentru a vedea probabilitatea de a fi în 0.01 a mediei. Codul R produce următoarele valori pentru  $P(0.49 \leq \bar{X}_n \leq 0.51)$ :

$$\begin{aligned} n = 10 : \text{pbinom}(5, 10, 0.5) - \text{pbinom}(4, 10, 0.5) &= 0.2460937 \\ n = 100 : \text{pbinom}(51, 100, 0.5) - \text{pbinom}(48, 100, 0.5) &= 0.2356466 \\ n = 1000 : \text{pbinom}(510, 1000, 0.5) - \text{pbinom}(489, 1000, 0.5) &= 0.49334 \\ n = 10000 : \text{pbinom}(5100, 10000, 0.5) - \text{pbinom}(4899, 10000, 0.5) &= 0.9555742 \end{aligned}$$

Din nou, vedem probabilitatea de a fi aproape de medie tinzând la 1 când  $n$  crește. Deoarece  $0.01 < 0.1$  este nevoie de valori mai mari ale lui  $n$  pentru a crește probabilitatea să se apropie de 1.

Convergența probabilității la 1 este LoLN în acțiune! Astfel, conform LoLN, **cu probabilitate mare**, media unui mare număr de date independente din aceeași repartiție va fi foarte aproape de media repartiției.

### 2.3.1 Enunțul formal al legii numerelor mari

**Teoremă (Legea numerelor mari):** Presupunem că  $X_1, X_2, \dots, X_n, \dots$  sunt variabile aleatoare i.i.d. cu media  $\mu$  și dispersia  $\sigma^2$ . Pentru fiecare  $n$ , fie  $\bar{X}_n$  media primelor  $n$  variabile. Atunci pentru orice  $a > 0$ , avem

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < a) = 1.$$

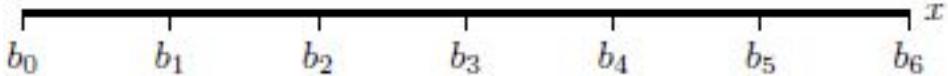
Aceasta spune precis că atunci când  $n$  crește, probabilitatea de a fi în  $a$  a mediei tinde la 1. Ne gândim la  $a$  ca la o mică toleranță a erorii de la adevărată medie  $\mu$ . În exemplul nostru, dacă vrem ca probabilitatea să fie cel puțin  $p = 0.99999$  ca proporția de aversuri  $\bar{X}_n$  este în  $a = 0.1$  a lui  $\mu = 0.5$ , atunci  $n > N = 500$  va fi suficient de mare. Dacă descreștem toleranța  $a$  și/sau creștem probabilitatea  $p$ , atunci  $N$  trebuie să fie mai mare.

## 2.4 Histograme

Putem rezuma date multiple  $x_1, \dots, x_n$  ale unei variabile aleatoare într-o **histogramă**. Aici vrem să construim histogramele cu grija astfel încât ele să semene cu aria de sub pdf.

Instrucțiunile pas cu pas pentru a construi o histogramă de densitate sunt următoarele.

1. Alegem un interval de pe dreapta reală și-l împărțim în  $m$  intervale, cu capetele  $b_0, b_1, \dots, b_m$ . De obicei aceste intervale au lungimi egale, deci presupunem asta de la început.



Fiecare din intervale este numit **coș** (în engleză bin). De exemplu, în figura de mai sus, primul coș este  $[b_0, b_1]$  și ultimul coș este  $[b_5, b_6]$ . Fiecare coș are o **lățime a coșului**, de exemplu  $b_1 - b_0$  este lățimea primului coș. De obicei toate coșurile au aceeași lățime, numită lățimea coșului histogramei.

2. Plasăm fiecare  $x_i$  în coșul care-i conține valoarea. Dacă  $x_i$  este pe frontieră a 2 coșuri, îl punem în coșul din stânga (aceasta este modul implicit în R, deși poate fi schimbat).

3. Pentru a desena o **histogramă de frecvențe**: punem un dreptunghi vertical deasupra fiecărui coș. **Înălțimea** dreptunghiului ar trebui să fie egală cu numărul de  $x_i$ -uri din coș.

4. Pentru a desena o **histogramă de densitate**: punem un dreptunghi vertical deasupra fiecărui coș. **Aria** dreptunghiului ar trebui să fie egală cu fracția din toate datele care sunt în coș.

### Observații:

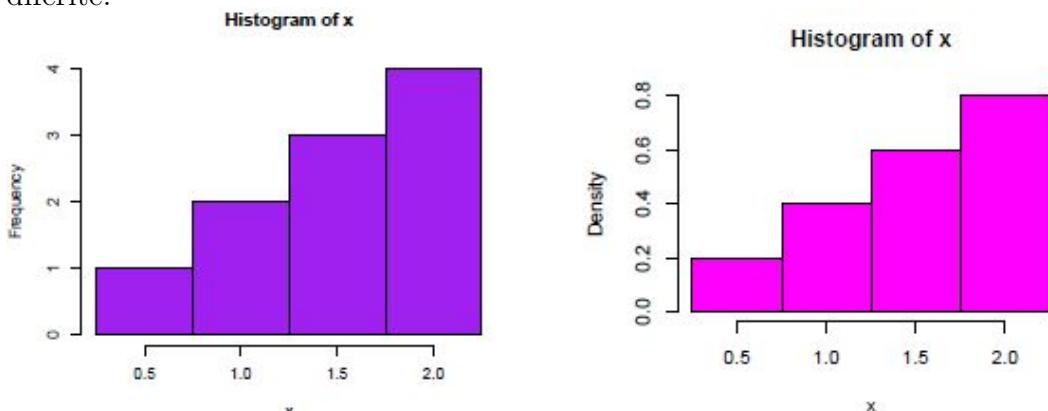
1. Când toate coșurile au aceeași lățime, dreptunghiurile histogramei de frecvențe au aria proporțională cu numărul. Deci histograma de densitate rezultă simplu din împărțirea înălțimii fiecărui dreptunghi la aria totală a histogramei de frecvențe. **Ignorând scala verticală, cele 2 histograme arată identic.**

2. **Atenție:** dacă lărimile coșurilor diferă, histogramele de frecvență și de densitate pot arăta foarte diferit.

În general, preferăm histograma de densitate deoarece scala ei verticală este aceeași cu cea a pdf.

**Exemplu.** Iată câteva exemple de histograme, toate cu datele  $[0.5, 1, 1, 1.5, 1.5, 1.5, 2, 2, 2, 2]$ .

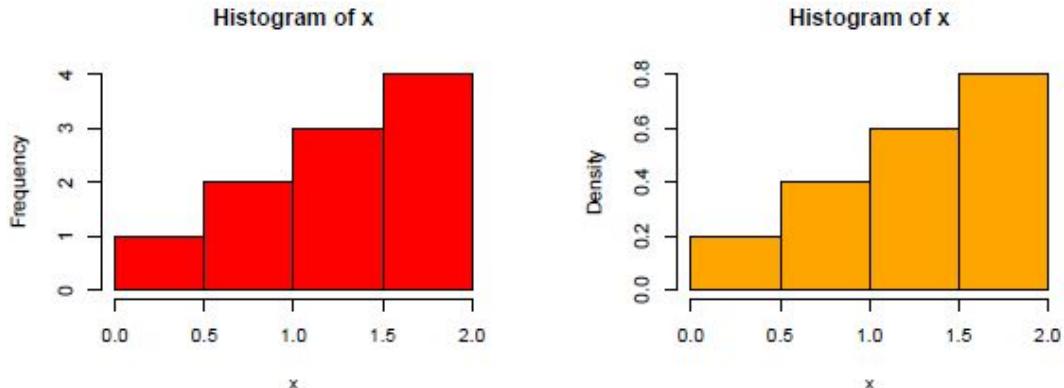
1. Aici reprezentările frecvenței și densității arată la fel, dar au scale verticale diferite.



Coșuri centrate în 0.5, 1, 1.5, 2, i.e. lățime 0.5, margini la 0.25, 0.75, 1.25,

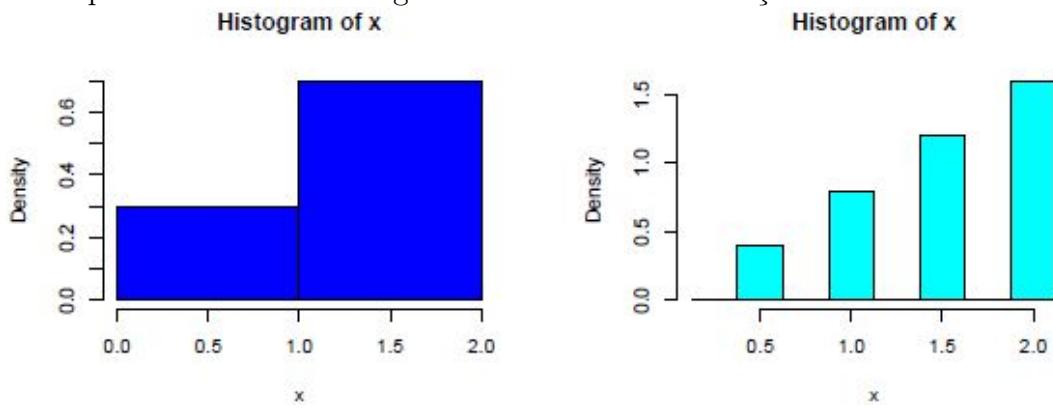
1.75, 2.25.

2. Aici valoarile sunt toate pe frontierele coșurilor și sunt puse în coșul din stânga. Adică, coșurile sunt [închise la dreapta](#), de exemplu primul coș este intervalul închis la dreapta  $(0, 0.5]$ .



Marginile coșurilor la 0, 0.5, 1, 1.5, 2.

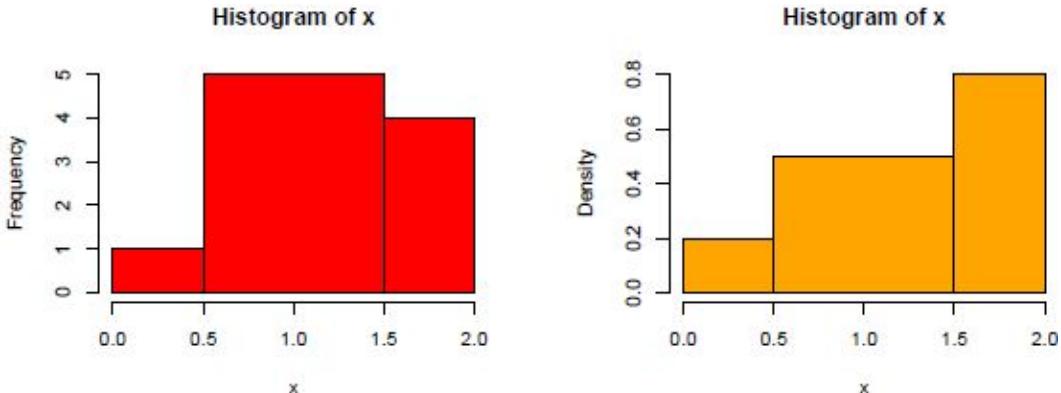
3. Aici sunt histograme de densitate bazate pe diferite lățimi de coșuri. Scala păstrează aria totală egală cu 1. Golurile sunt coșuri cu zero date.



Stânga: coșuri late.

Dreapta: coșuri înguste.

4. Aici folosim lățimi diferite de coșuri, aşa că histogramele de frecvență și densitate arată diferit.



Nu vă lăsați păcăliți! Acestea sunt bazate pe aceleasi date.

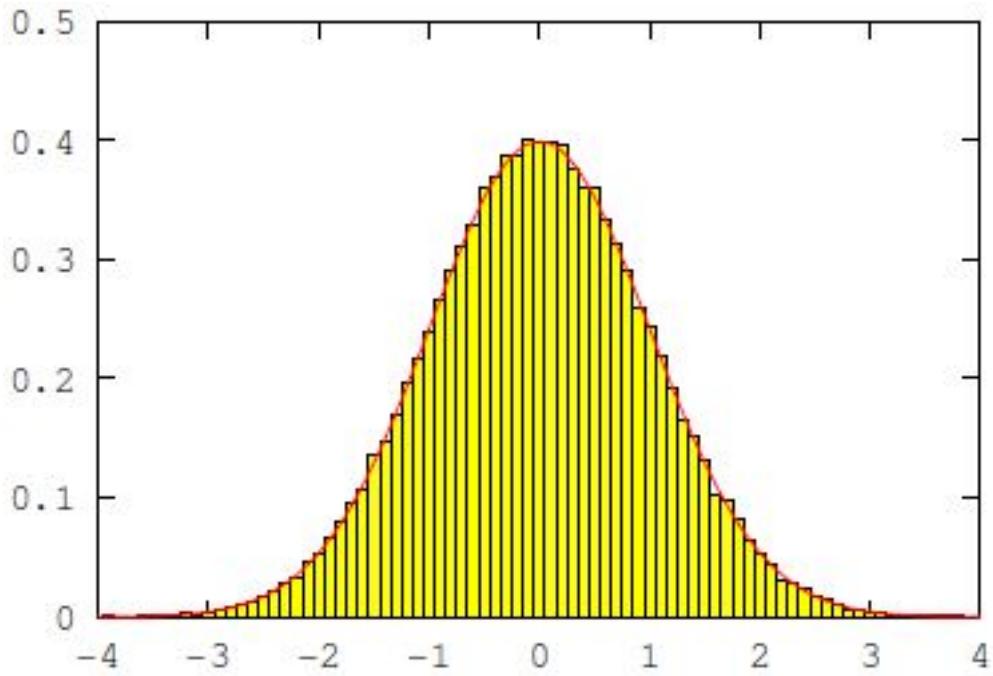
Histograma de densitate este alegerea mai bună la lățimi diferite de coșuri. De fapt, R se va plânge dacă încercăm să facem o histogramă de frecvențe cu lățimi diferite de coșuri. Comparați histograma de frecvențe cu lățimi diferite de coșuri cu toate celelalte histograme pe care le-am desenat pentru aceste date. Arată clar diferit. Ce s-a întâmplat este că prin combinarea datelor din coșurile  $(0.5, 1]$  și  $(1, 1.5]$  într-un coș  $(0.5, 1.5]$  am făcut mai mare înălțimea ambelor coșuri mai mici.

#### 2.4.1 Legea numerelor mari și histogramele

Legea numerelor mari are o consecință importantă pentru histograme de densitate.

**LoLN pentru histograme:** Cu probabilitate mare histograma de densitate a unui **mare** număr de date dintr-o repartiție este o bună aproximare a graficului pdf  $f$  a repartiției.

Ilustrăm aceasta generând o histogramă de densitate cu lățimea coșului 0.1 din 10000 de date dintr-o repartiție normală standard. Histograma de densitate urmărește foarte aproape graficul pdf normală standard  $\phi$ .



Histograma de densitate a 10000 de date dintr-o repartiție normală standard, cu graficul lui  $\phi$  în roșu.

## 2.5 Teorema limită centrală

### 2.5.1 Standardizare

Fiind dată o variabilă aleatoare  $X$  cu media  $\mu$  și deviația standard  $\sigma$ , definim **standardizarea** lui  $X$  ca noua variabilă aleatoare

$$Z = \frac{X - \mu}{\sigma}.$$

$Z$  are media 0 și deviația standard 1. Dacă  $X$  are o repartiție normală, atunci standardizarea lui  $X$  are repartiția normală standard  $Z$  cu media 0 și dispersia 1. Aceasta explică termenul "standardizare" și notația  $Z$  de mai sus.

### 2.5.2 Enunțul teoremei limită centrală

Presupunem că  $X_1, X_2, \dots, X_n, \dots$  sunt variabile aleatoare i.i.d., fiecare având media  $\mu$  și deviația standard  $\sigma$ . Pentru fiecare  $n$  notăm cu  $S_n$  suma și cu  $\bar{X}_n$

media lui  $X_1, \dots, X_n$ .

$$S_n = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i$$

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{S_n}{n}.$$

Proprietățile mediei și dispersiei arată că

$$E(S_n) = n\mu, \quad Var(S_n) = n\sigma^2, \quad \sigma_{S_n} = \sigma\sqrt{n}$$

$$E(\bar{X}_n) = \mu, \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}, \quad \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}.$$

Deoarece  $S_n$  este multiplu al lui  $\bar{X}_n$ ,  $S_n$  și  $\bar{X}_n$  au aceeași standardizare

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

**Teorema limită centrală:** Pentru  $n$  mare,

$$\bar{X}_n \approx N(\mu, \sigma^2/n), \quad S_n \approx N(n\mu, n\sigma^2), \quad Z_n \approx N(0, 1).$$

**Observații:** 1. În cuvinte:  $\bar{X}_n$  este aproximativ o repartiție normală cu aceeași medie ca  $X$ , dar o dispersie mai mică.

2.  $S_n$  este aproximativ normală.

3.  $\bar{X}_n$  și  $S_n$  standardizate sunt aproximativ normale standard.

Teorema limită centrală ne permite să aproximăm o sumă sau medie de variabile aleatoare i.i.d. printr-o variabilă aleatoare normală. Ea este extrem de folositoare deoarece de obicei este ușor să facem calcule cu repartiția normală.

Un enunț precis al CLT este: cdf-urile lui  $Z_n$  converg la  $\Phi$ :

$$\lim_{n \rightarrow \infty} F_{Z_n}(z) = \Phi(z).$$

### 2.5.3 Probabilități normale standard

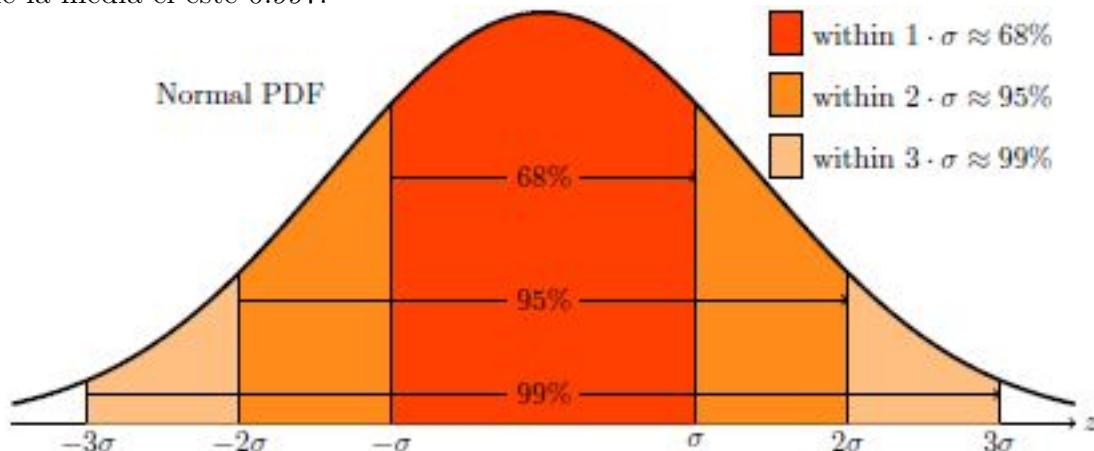
Pentru a aplica CLT vrem să avem la îndemână unele probabilități normale standard. Reamintim: dacă  $Z \sim N(0, 1)$ , atunci cu rotunjire avem:

1.  $P(|Z| < 1) = 0.68$ .
2.  $P(|Z| < 2) = 0.95$ ; mai precis,  $P(|Z| < 1.96) \approx 0.95$ .
3.  $P(|Z| < 3) = 0.997$ .

Aceste numere sunt ușor de calculat în R folosind `pnorm`. Oricum, ele sunt

demne de reamintit ca regula degetului mare.

1. Probabilitatea că o variabilă aleatoare normală este într-o deviație standard de la media ei este 0.68. (Adică,  $X \sim N(\mu, \sigma^2) \Rightarrow P(|X - \mu| < \sigma) \approx 0.68$ .)
2. Probabilitatea că o variabilă aleatoare normală este în 2 deviații standard de la media ei este 0.95.
3. Probabilitatea că o variabilă aleatoare normală este în 3 deviații standard de la media ei este 0.997.



**Consecințe:**

1.  $P(Z < 1) \approx 0.84$ .
2.  $P(Z < 2) \approx 0.977$ .
3.  $P(Z < 3) \approx 0.999$ .

**Demonstrație:** 1. Știm că  $P(|Z| < 1) = 0.68$ . Probabilitatea rămasă de 0.32 este în 2 regiuni  $Z > 1$  și  $Z < -1$ . Aceste regiuni sunt numite **coada dreaptă**, respectiv **coada stângă**. Din simetrie, fiecare coadă are 0.16. Deci,

$$P(Z < 1) = P(|Z| < 1) + P(\text{coada stângă}) \approx 0.68 + 0.16 = 0.84.$$

Celelalte 2 consecințe se fac similar.

#### 2.5.4 Aplicații ale CLT

**Exemplul 2.** Aruncăm o monedă corectă de 100 de ori. Estimați probabilitatea a mai mult de 55 de aversuri.

**Răspuns:** Fie  $X_j$  rezultatul celei de-a  $j$ -a aruncări, deci  $X_j = 1$  pentru avers și  $X_j = 0$  pentru revers. Numărul total de aversuri este

$$S = X_1 + X_2 + \dots + X_{100}.$$

Știm că  $E(X_j) = 0.5$  și  $Var(X_j) = 1/4$ . Deoarece  $n = 100$ , avem

$$E(S) = 50, \quad Var(S) = 25, \quad \text{și } \sigma_S = 5.$$

Teorema limită centrală spune că standardizarea lui  $S$  este aproximativ  $N(0, 1)$ . Se cere estimarea  $P(S > 55)$ . Standardizând și folosind CLT obținem

$$P(S > 55) = P\left(\frac{S - 50}{5} > \frac{55 - 50}{5}\right) \approx P(Z > 1) \approx 0.16.$$

Aici  $Z$  este o variabilă aleatoare normală și  $P(Z > 1) = 1 - P(Z < 1) \approx 1 - 0.84 = 0.16$ .

**Exemplul 3.** Estimați probabilitatea a mai mult de 220 de aversuri în 400 de aruncări.

**Răspuns:** Este aproape identic cu exemplul precedent. Acum  $\mu_S = 200$  și  $\sigma_S = 10$  și se cere estimarea  $P(S > 220)$ . Standardizând și folosind CLT obținem:

$$P(S > 220) = P\left(\frac{S - \mu_S}{\sigma_S} > \frac{220 - 200}{10}\right) \approx P(Z > 2) \approx 0.023.$$

Din nou,  $Z \sim N(0, 1)$  și regulile degetului mare arată că  $P(Z > 2) \approx 0.023$ .

**Observație:** Cu toate că  $55/100 = 220/400$ , probabilitatea a mai mult de 55 de aversuri în 100 de aruncări este mai mare decât probabilitatea a mai mult de 220 de aversuri în 400 de aruncări. Aceasta este datorată LoLN și valorii mai mari a lui  $n$  în ultimul caz.

**Exemplul 4.** Estimați probabilitatea unui număr între 40 și 60 de aversuri în 100 de aruncări.

**Răspuns:** Ca în primul exemplu,  $E(S) = 50$ ,  $Var(S) = 25$  și  $\sigma_S = 5$ . Deci

$$P(40 \leq S \leq 60) = P\left(\frac{40 - 50}{5} \leq \frac{S - 50}{5} \leq \frac{60 - 50}{5}\right) \approx P(-2 \leq Z \leq 2).$$

Putem calcula ultimul membru folosind regula degetului mare. Pentru un răspuns mai precis folosim R:

$$\text{pnorm}(2) - \text{pnorm}(-2) = 0.9544997.$$

Reamintim că am folosit repartiția binomială pentru a calcula un răspuns de 0.9647998. Deci răspunsul nostru aproximativ folosind CLT are o eroare de aproximativ 1%.

**Gândiți:** Vă așteptați ca metoda CLT să dea o aproximare mai bună sau mai rea pentru  $P(200 < S < 300)$  cu  $n = 500$ ?

Verificați răspunsul folosind R.

**Exemplul 5. Sondaj.** Când se face un sondaj politic, rezultatele sunt adesea raportate ca un număr cu o margine de eroare. De exemplu,  $52\% \pm 3\%$  susțin

candidatul A. Regula degetului mare este că dacă sondazi  $n$  oameni, atunci marginea de eroare este  $\pm \frac{1}{\sqrt{n}}$ . Vom vedea acum exact ce înseamnă aceasta și că este o aplicație a teoremei limită centrală.

Presupunem că sunt 2 candidați A și B. Presupunem mai departe că fracția din populație care preferă pe A este  $p_0$ . Adică, dacă întrebî o persoană aleatoare pe cine preferă, probabilitatea să răspundă A este  $p_0$ .

Pentru a face sondajul, un sondator alege aleator  $n$  persoane și le întreabă "Susțineți candidatul A sau candidatul B?" Astfel, putem vedea sondajul ca o secvență de  $n$  date Bernoulli( $p_0$ ) independente,  $X_1, X_2, \dots, X_n$ , unde  $X_i$  este 1 dacă a  $i$ -a persoană preferă pe A și 0 dacă preferă pe B. Fracția de oameni sondăți că-l preferă pe A este chiar media  $\bar{X}$ .

Stim că fiecare  $X_i \sim \text{Bernoulli}(p_0)$ , deci

$$E(X_i) = p_0 \text{ și } \sigma_{X_i} = \sqrt{p_0(1 - p_0)}.$$

De aceea, teorema limită centrală ne spune că

$$\bar{X} \approx N(p_0, \sigma/\sqrt{n}), \text{ unde } \sigma = \sqrt{p_0(1 - p_0)}.$$

Într-o repartiție normală 95% din probabilitate este în 2 deviații de la medie. Aceasta înseamnă că în 95% din sondajele a  $n$  oameni media de selecție  $\bar{X}$  va fi în  $2\sigma/\sqrt{n}$  de la adevărata medie  $p_0$ . Pasul final este să observăm că pentru orice valoare  $p_0$  avem  $\sigma \leq 1/2$ . (De exemplu, se poate aplica inegalitatea mediilor.) Aceasta înseamnă că putem spune că în 95% din sondajele a  $n$  oameni media de selecție este în  $1/\sqrt{n}$  de la adevărata medie. Statisticianul frecvenționist ia intervalul  $[\bar{X} - 1/\sqrt{n}, \bar{X} + 1/\sqrt{n}]$  și-l numește intervalul de 95% încredere pentru  $p_0$ .

**Un cuvânt de precauție:** este tentant și ușor, dar gresit, să gândești că este o probabilitate de 95% ca adevărata fracție  $p_0$  să fie în intervalul de încredere. Acest fapt este subtil, dar eroarea este aceeași cu a gândești că ai boala dacă un test 95% precis iese pozitiv. Este adevărat că 95% din oamenii care fac testul primesc rezultatul corect. Nu este necesar adevărat că 95% dintre toate testele pozitive sunt corecte.

### 2.5.5 De ce folosim CLT?

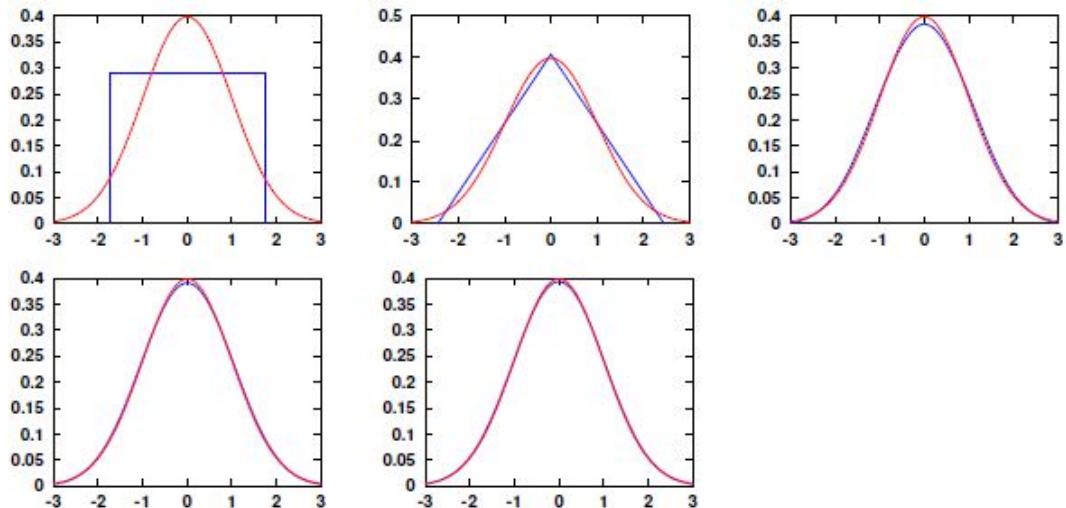
Deoarece probabilitățile din exemplele de mai sus puteau fi calculate folosind repartiția binomială, poate vă întrebați care este rostul aflării unui răspuns aproximativ folosind CLT. De fapt, puteam să calculăm exact aceste probabilități deoarece  $X_i$  erau Bernoulli și de aceea suma  $S$  era binomială. În general, repartiția lui  $S$  nu va fi familiară, de aceea nu vom putea calcula exact probabilitățile pentru  $S$ ; se poate întâmpla ca să fie posibil calculul

exact în teorie, dar să fie prea greu de calculat în practică, chiar și pentru un computer. Puterea CLT este că se poate aplica când  $X_i$  are aproape orice repartiție.

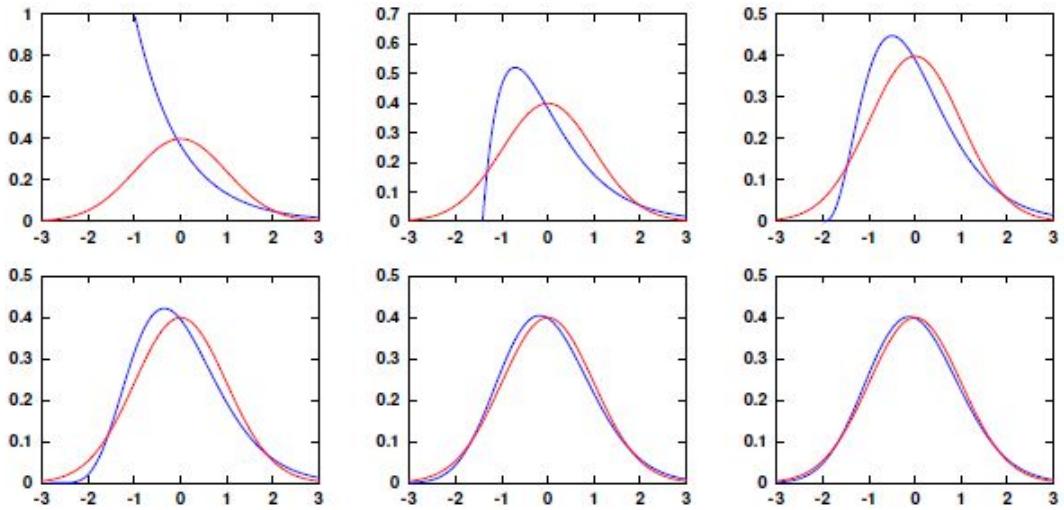
### 2.5.6 Cât de mare trebuie să fie $n$ pentru a aplica CLT?

Răspuns scurt: adesea, nu aşa mare.

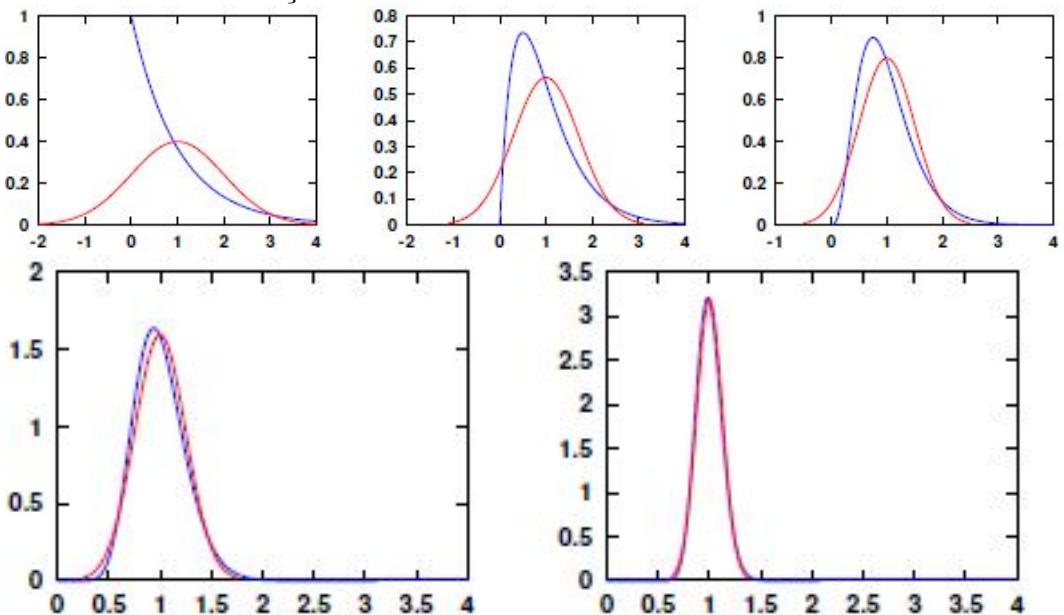
Următoarele secvențe de imagini arată convergența la o repartiție normală. Înțâi arătăm media standardizată a  $n$  variabile aleatoare **uniforme** i.i.d. cu  $n = 1, 2, 4, 8, 12$ . Pdf a mediei este cu albastru și pdf normală standard este în roșu. La  $n = 12$  potrivirea dintre media standardizată și adevărata normală arată foarte bine.



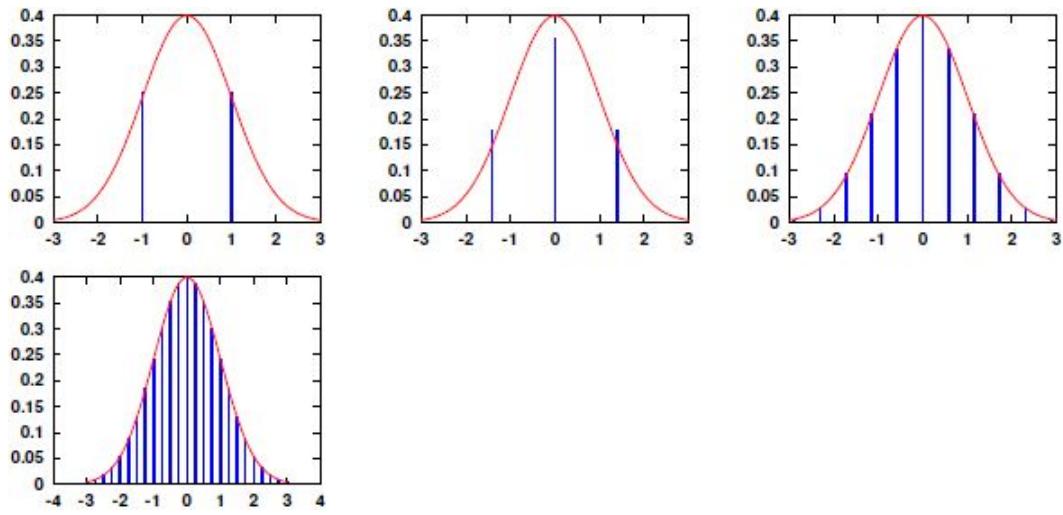
Apoi arătăm media standardizată a  $n$  variabile aleatoare **exponențiale** i.i.d. cu  $n = 1, 2, 4, 8, 16, 64$ . Această densitate asimetrică ia mai mulți termeni pentru a se apropia de densitatea normală.



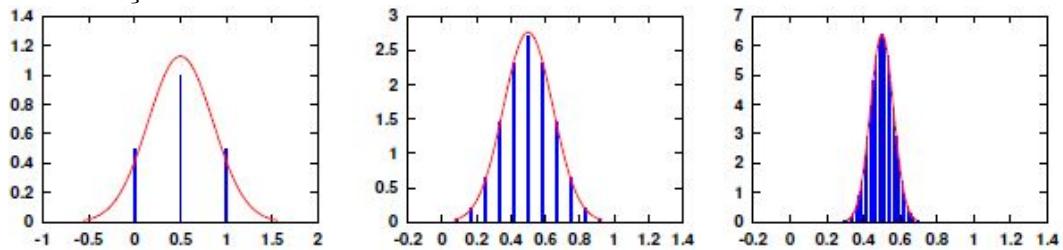
Apoi arătăm media (nestandardizată) a  $n$  variabile aleatoare exponențiale cu  $n = 1, 2, 4, 16, 64$ . Deviația standard de micșorează când  $n$  crește, rezultând o densitate mai ascuțită.



Teorema limită centrală funcționează și pentru variabile aleatoare discrete. Aici este media standardizată a  $n$  variabile aleatoare Bernoulli(0.5) i.i.d. cu  $n = 1, 2, 12, 64$ . Când  $n$  crește, media poate lua mai multe valori, ceea ce permite repartiției discrete să "umple" densitatea normală.



La sfârșit arătăm media (nestandardizată) a  $n$  variabile aleatoare Bernoulli(0.5), cu  $n = 4, 12, 64$ . Deviația standard devine mai mică rezultând o densitate mai ascuțită.



# Curs 7

Cristian Niculescu

## 1 Demonstrații

### 1.1 Introducere

Dăm mai mult material matematic formal pentru a sublinia că atunci când facem matematică ar trebui să avem grijă să specificăm complet ipotezele noastre și să dăm argumente deductive clare pentru a demonstra afirmațiile noastre.

### 1.2 Cu probabilitate mare histograma de densitate seamănă cu graficul funcției densitate de probabilitate

Am afirmat că o consecință a legii numerelor mari este că, atunci când numărul datelor crește, histograma de densitate a datelor are o probabilitate crescătoare de a se potrivi cu graficul pdf sau pmf. Aceasta este o bună regulă a degetului mare, dar este mai degrabă imprecisă. Putem face afirmații mai precise.

Presupunem că avem un experiment care produce date conform variabilei aleatoare  $X$  și presupunem că generăm  $n$  date independente din  $X$ . Le notăm

$$x_1, x_2, \dots, x_n.$$

Prinț-un coș înțelegem o mulțime de valori, de exemplu  $(b_k, b_{k+1}]$ . Pentru a face o histogramă de densitate împărțim domeniul de valori al lui  $X$  în  $m$  coșuri și calculăm fracția din date din fiecare coș.

Fie  $p_k$  probabilitatea ca o dată aleatoare să fie în al  $k$ -lea coș. Aceasta este probabilitatea pentru o variabilă aleatoare (Bernoulli) **indicator**  $B_{k,j}$ , care este 1 dacă a  $j$ -a dată este în coș și 0 altfel.

**Propoziția 1.** Fie  $\bar{p}_k = (\text{numărul datelor din coșul } k)/n$ . Când numărul  $n$  al datelor devine mare, probabilitatea că  $\bar{p}_k$  este aproape de  $p_k$  se apropie de 1. Altfel spus, fiind dat un număr pozitiv oricât de mic, să-l numim  $a$ ,

probabilitatea  $P(|\bar{p}_k - p_k| < a)$  depinde de  $n$  și, când  $n \rightarrow \infty$ , această probabilitate tinde la 1.

**Demonstrație.** Fie  $\bar{B}_k$  media lui  $B_{k,j}$ . Deoarece  $E(B_{k,j}) = p_k$ , legea numerelor mari spune exact că

$$P(|\bar{B}_k - p_k| < a) \rightarrow 1 \text{ când } n \rightarrow \infty.$$

Dar, deoarece  $B_{k,j}$  sunt variabile indicator, media lor este exact  $\bar{p}_k$ . Înlocuirea lui  $\bar{B}_k$  cu  $\bar{p}_k$  în relația de mai sus dă

$$P(|\bar{p}_k - p_k| < a) \rightarrow 1 \text{ când } n \rightarrow \infty.$$

Aceasta este exact ce s-a cerut în propoziția 1.

**Propoziția 2.** Aceeași propoziție este valabilă simultan pentru un număr finit de coșuri. Adică, pentru coșurile de la 1 la  $m$  avem

$$P(|\bar{B}_1 - p_1| < a, |\bar{B}_2 - p_2| < a, \dots, |\bar{B}_m - p_m| < a) \rightarrow 1 \text{ când } n \rightarrow \infty.$$

**Demonstrație.** Mai întâi scriem următoarea regulă de probabilitate, care este o consecință a principiului includerii și excluderii: dacă 2 evenimente  $A$  și  $B$  au  $P(A) = 1 - \alpha_1$  și  $P(B) = 1 - \alpha_2$ , atunci  $P(A \cap B) \geq 1 - (\alpha_1 + \alpha_2)$ . Acum, propoziția 1 spune că  $\forall \alpha > 0$ , de la un  $n$  suficient de mare avem  $P(|\bar{B}_k - p_k| < a) > 1 - \alpha/m$  separat pentru fiecare coș. Din regula de probabilitate, probabilitatea intersecției tuturor acestor evenimente este cel puțin  $1 - \alpha$ . Deoarece putem lua  $\alpha > 0$  oricât de mic vrem, lăsând  $n$  să tindă la  $\infty$ , la limită obținem probabilitatea 1 așa cum am afirmat.

**Propoziția 3.** Dacă  $f$  este o densitate de probabilitate continuă cu domeniul de valori  $[a, b]$ , atunci, luând suficiente date și având lățimea coșurilor suficient de mică, putem garanta că, cu mare probabilitate, histograma de densitate este cât de aproape vrem de graficul lui  $f$ .

**Demonstrație.** (Schiță.) Presupunem că lățimea coșului din jurul lui  $x$  este  $\Delta x$ . Dacă  $\Delta x$  este suficient de mic, atunci probabilitatea că o dată este în coș este aproximativ  $f(x)\Delta x$ . Propoziția 1 garantează că, dacă  $n$  este suficient de mare, atunci, cu probabilitate mare,  $(\text{numărul datelor din coș})/n$  este de asemenea aproximativ  $f(x)\Delta x$ . Deoarece aceasta este aria coșului în histogramă, vedem că înălțimea coșului va fi aproximativ  $f(x)$ . Adică, cu mare probabilitate, înălțimea histogramei peste orice punct  $x$  este aproape de  $f(x)$ . Aceasta este ceea ce a afirmat propoziția 3.

**Observație.** Dacă domeniul de valori este nemărginit sau densitatea tinde la  $\infty$  în vreun punct, trebuie să fim mai atenți. Există rezultate și pentru aceste cazuri.

### 1.3 Inegalitatea lui Cebâşev

O demonstrație a LoLN rezultă din următoarea inegalitate cheie.

**Inegalitatea lui Cebâşev.** Presupunem că  $Y$  este o variabilă aleatoare cu media  $\mu$  și dispersia  $\sigma^2$ . Atunci,  $\forall a > 0$  avem

$$P(|Y - \mu| \geq a) \leq \frac{Var(Y)}{a^2}.$$

În cuvinte, inegalitatea lui Cebâşev spune că probabilitatea ca  $Y$  să difere de medie cu cel puțin  $a$  este mărginită superior de  $Var(Y)/a^2$ . Practic, cu cât mai mică este dispersia lui  $Y$ , cu atât mai mică este probabilitatea că  $Y$  este departe de media lui.

**Demonstrația legii numerelor mari.** Deoarece  $Var(\bar{X}_n) = Var(X)/n$ ,  $\lim_{n \rightarrow \infty} Var(\bar{X}_n) = 0$ . Astfel, din inegalitatea lui Cebâşev pentru  $Y = \bar{X}_n$  și  $a$  fixat  $\implies \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \geq a) = 0 \implies \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < a) = 1$ , ceea ce este LoLN.

**Demonstrația inegalității lui Cebâşev.** Demonstrația este în esență aceeași pentru  $Y$  discretă și  $Y$  continuă. Presupunem  $Y$  continuă și  $\mu = 0$ , deoarece înlocuind  $Y$  cu  $Y - \mu$ , dispersia nu se schimbă. Astfel,

$$\begin{aligned} P(|Y| \geq a) &= \int_{-\infty}^{-a} f(y)dy + \int_a^{\infty} f(y)dy \leq \int_{-\infty}^{-a} \frac{y^2}{a^2} f(y)dy + \int_a^{\infty} \frac{y^2}{a^2} f(y)dy \\ &\leq \int_{-\infty}^{\infty} \frac{y^2}{a^2} f(y)dy = \frac{Var(Y)}{a^2}. \end{aligned}$$

Prima inegalitate utilizează faptul că  $y^2/a^2 \geq 1$  pe intervalele de integrare. A 2-a inegalitate rezultă deoarece includerea intervalului  $[-a, a]$  nu micșorează integrala, deoarece integrantul este nenegativ.

### 1.4 Nevoia de dispersie

Am trecut cu vederea un fapt tehnic. Peste tot am presupus că repartițiile aveau dispersie. De exemplu, demonstrația legii numerelor mari a folosit dispersia prin intermediul inegalității lui Cebâşev. Dar sunt repartiții care nu au dispersie deoarece suma sau integrala pentru dispersie nu converge. Pentru astfel de repartiții legea numerelor mari poate fi falsă.

## 2 Repartiții comune, independentă

### 2.1 Scopurile învățării

1. Să înțeleagă ce înseamnă o pmf, pdf și cdf comună a 2 variabile aleatoare.
2. Să poată calcula probabilități și marginale dintr-o pmf sau pdf comună.
3. Să poată testa dacă 2 variabile aleatoare sunt independente.

### 2.2 Introducere

În știință și viața reală, suntem adesea interesați de 2 (sau mai multe) variabile aleatoare în același timp. De exemplu, putem măsura înălțimea și greutatea girafelor, sau IQ-ul și greutatea la naștere a copiilor, sau frecvența exercițiilor și rata bolilor de inimă la adulți, sau nivelul poluării aerului și rata bolilor respiratorii în orașe, sau numărul de prieteni pe Facebook și vârsta membrilor Facebook.

**Gândiți:** Ce relație ne-am aștepta să fie în fiecare din cele 5 exemple de mai sus? De ce?

În astfel de situații variabilele aleatoare au o **repartiție comună** care ne permite să calculăm probabilități ale evenimentelor implicând ambele variabile și să înțelegem relația dintre variabile. Cel mai simplu este când variabilele sunt **independente**. Când nu sunt, folosim **covarianța** și **corelația** ca măsuri ale naturii dependenței dintre ele.

### 2.3 Repartiția comună

#### 2.3.1 Cazul discret

Presupunem că  $X$  și  $Y$  sunt 2 variabile aleatoare discrete și că  $X$  ia valorile din  $\{x_1, x_2, \dots, x_n\}$ , iar  $Y$  ia valorile din  $\{y_1, y_2, \dots, y_m\}$ . Perechea ordonată  $(X, Y)$  ia valorile din produsul cartezian  $\{(x_1, y_1), (x_1, y_2), \dots, (x_n, y_m)\}$ . **funcția masă de probabilitate comună** (pmf comună) a lui  $X$  și  $Y$  este funcția

$$p(x_i, y_j) = P(X = x_i, Y = y_j), i = \overline{1, n}, j = \overline{1, m}.$$

O organizăm într-un tabel de probabilitate comună:

$X \setminus Y$	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_m$
$x_1$	$p(x_1, y_1)$	$p(x_1, y_2)$	$\dots$	$p(x_1, y_j)$	$\dots$	$p(x_1, y_m)$
$x_2$	$p(x_2, y_1)$	$p(x_2, y_2)$	$\dots$	$p(x_2, y_j)$	$\dots$	$p(x_2, y_m)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_i$	$p(x_i, y_1)$	$p(x_i, y_2)$	$\dots$	$p(x_i, y_j)$	$\dots$	$p(x_i, y_m)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_n$	$p(x_n, y_1)$	$p(x_n, y_2)$	$\dots$	$p(x_n, y_j)$	$\dots$	$p(x_n, y_m)$

**Exemplul 1.** Aruncăm 2 zaruri corecte. Fie  $X$  numărul de pe primul zar și  $Y$  numărul de pe cel de-al 2-lea. Atunci atât  $X$  cât și  $Y$  iau valori de la 1 la 6, iar pmf comună este  $p(i, j) = 1/36, \forall i, j = \overline{1, 6}$ . Iată tabelul de probabilitate comună:

$X \setminus Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

**Exemplul 2.** Aruncăm 2 zaruri corecte. Fie  $X$  numărul de pe primul zar și  $T$  totalul de pe cele 2 zaruri. Iată tabelul de probabilitate comună:

$X \setminus T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

O funcție masă de probabilitate comună trebuie să satisfacă 2 proprietăți:

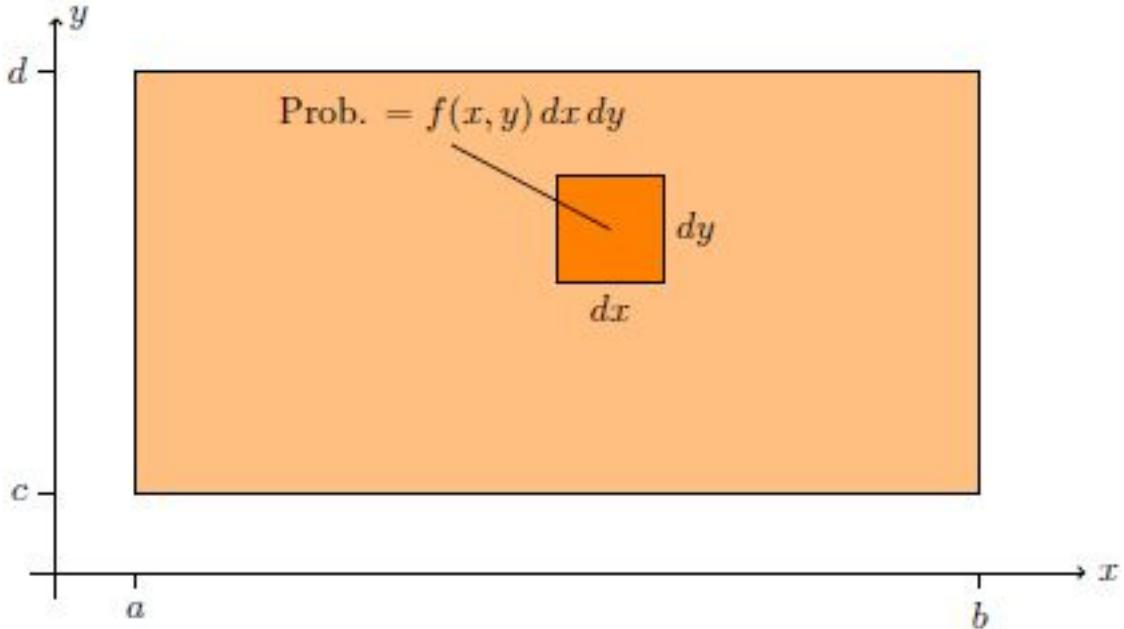
1.  $0 \leq p(x_i, y_j) \leq 1, \forall i = \overline{1, n}, j = \overline{1, m}.$
2. Probabilitatea totală este 1. Putem exprima aceasta cu o **sumă dublă**:

$$\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) = 1.$$

### 2.3.2 Cazul continuu

Cazul continuu este în esență la fel: doar înlocuim mulțimile discrete de valori cu intervale, funcția masă de probabilitate comună cu o **funcție densitate de probabilitate comună** și sumele cu integrale.

Dacă  $X$  ia valori în  $[a, b]$  și  $Y$  ia valori în  $[c, d]$ , atunci perechea  $(X, Y)$  ia valori în produsul cartezian  $[a, b] \times [c, d]$ . **Funcția densitate de probabilitate comună** (pdf comună) a lui  $X$  și  $Y$  este o funcție  $f(x, y)$  dând densitatea de probabilitate în  $(x, y)$ . Adică, probabilitatea că  $(X, Y)$  este într-un mic dreptunghi de lățime  $dx$  și înălțime  $dy$  în jurul lui  $(x, y)$  este  $f(x, y)dx dy$ .



O funcție densitate comună de probabilitate trebuie să satisfacă 2 proprietăți:

1.  $f(x, y) \geq 0, \forall x, y.$
2. Probabilitatea totală este 1. Exprimăm aceasta cu o **integrală dublă**:

$$\int_c^d \int_a^b f(x, y) dx dy = 1.$$

Observație: ca la pdf a unei singure variabile aleatoare, pdf comună  $f(x, y)$  poate lua valori  $> 1$ ; este o densitate de probabilitate, **nu** o probabilitate.

- Ar trebui să puteți calcula integrale duble pe dreptunghiuri.
- Pe o regiune nedreptunghiulară, când  $f(x, y) = c$  este constantă, integrala dublă este egală cu  $c \cdot$ (aria regiunii).

### 2.3.3 Evenimente

Variabilele aleatoare sunt utile pentru a descrie evenimente. Reamintim că un eveniment este o mulțime de rezultate și că variabilele aleatoare atribuie numere rezultatelor. De exemplu, evenimentul " $X > 1$ " este mulțimea tuturor rezultatelor pentru care  $X$  este  $> 1$ . Aceste concepte se extind la perechi de variabile aleatoare și rezultate comune.

**Exemplul 3.** În exemplul 1, descrieți evenimentul  $B = "Y - X \geq 2"$  și aflați-i probabilitatea.

**Răspuns.** Putem descrie  $B$  ca o mulțime de perechi  $(x, y)$ :

$$B = \{(1, 3), (1, 4), (1, 5), (1, 6), (2, 4), (2, 5), (2, 6), (3, 5), (3, 6), (4, 6)\}.$$

Putem să-l descriem și vizual:

$X \setminus Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

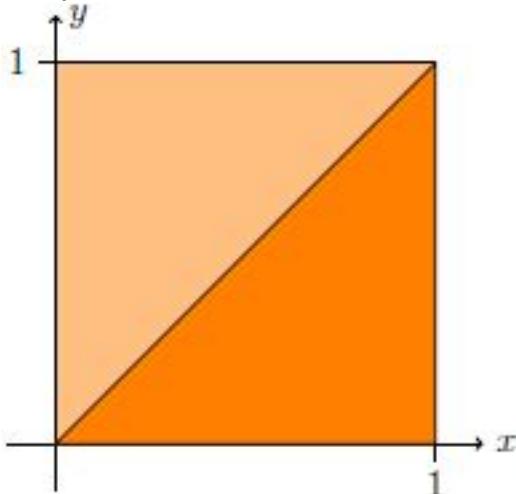
Evenimentul  $B$  constă din rezultatele corespunzătoare dreptunghiurilor portocalii.

Probabilitatea lui  $B$  este suma probabilităților din dreptunghiurile portocalii, astfel,  $P(B) = 10/36 = 5/18$ .

**Exemplul 4.** Presupunem că atât  $X$  cât și  $Y$  iau valori în  $[0,1]$  cu densitatea uniformă  $f(x, y) = 1$ . Vizualizați evenimentul " $X > Y$ " și aflați-i probabilitatea.

**Răspuns.** În comun,  $X$  și  $Y$  iau valori în pătratul unitate. Evenimentul

” $X > Y$ ” corespunde triunghiului mai închis din dreapta-jos din figura de mai jos.

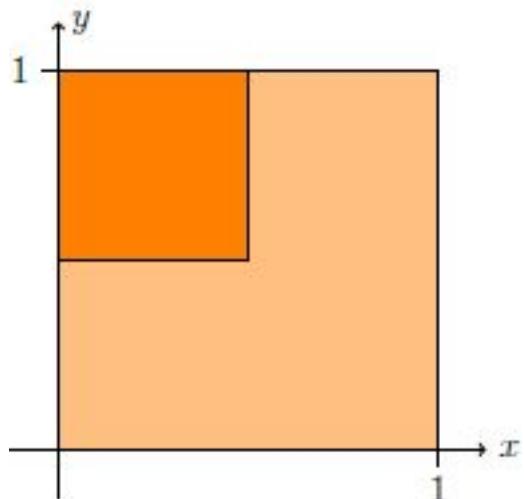


Evenimentul ” $X > Y$ ” în pătratul unitate.

Deoarece densitatea este constantă, probabilitatea este chiar fracția din aria totală ocupată de eveniment. În acest caz este 0.5.

**Exemplul 5.** Presupunem că atât  $X$  cât și  $Y$  iau valori în  $[0,1]$  cu densitatea  $f(x,y) = 4xy$ . Arătați că  $f$  este o pdf comună validă, vizualizați evenimentul  $A = ”X < 0.5 \text{ și } Y > 0.5”$  și aflați-i probabilitatea.

**Răspuns.** În comun,  $X$  și  $Y$  iau valori în pătratul unitate.



Evenimentul  $A$  în pătratul unitate.

Pentru a arăta că  $f$  este o pdf validă trebuie să verificăm că este nenegativă

(ceea ce este clar) și că probabilitatea totală este 1.

$$\begin{aligned}\text{Probabilitatea totală} &= \int_0^1 \int_0^1 4xy dx dy = \int_0^1 2x^2 y \Big|_{x=0}^{x=1} dy = \int_0^1 2y dy \\ &= y^2 \Big|_0^1 = 1, \text{ q.e.d.}\end{aligned}$$

Evenimentul  $A$  este chiar pătratul din stânga sus. Deoarece densitatea nu este constantă, trebuie să calculăm o integrală pentru a afla probabilitatea.

$$P(A) = \int_0^{0.5} \int_{0.5}^1 4xy dy dx = \int_0^{0.5} 2xy^2 \Big|_{y=0.5}^{y=1} dx = \int_0^{0.5} \frac{3x}{2} dx = \frac{3x^2}{4} \Big|_0^{0.5} = \frac{3}{16}.$$

### 2.3.4 Funcția de repartiție comună

Presupunem că  $X$  și  $Y$  sunt variabile aleatoare repartizate comun. Folosim notația " $X \leq x, Y \leq y$ " pentru evenimentul " $X \leq x$  și  $Y \leq y$ ". **Funcția de repartiție comună** (cdf comună) este definită astfel

$$F(x, y) = P(X \leq x, Y \leq y)$$

**Cazul continuu:** Dacă  $X$  și  $Y$  sunt variabile aleatoare continue cu densitatea comună  $f(x, y)$  pe domeniul de valori  $[a, b] \times [c, d]$ , atunci cdf comună este dată de integrală dublă

$$F(x, y) = \int_c^y \int_a^x f(u, v) du dv.$$

Pentru a recupera pdf comună, derivăm cdf comună. Deoarece sunt 2 variabile, avem nevoie să folosim derivate partiale:

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y).$$

**Cazul discret:** Dacă  $X$  și  $Y$  sunt variabile aleatoare discrete cu pmf comună  $p(x_i, y_j)$ , atunci cdf comună este dată de suma dublă

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j).$$

### 2.3.5 Proprietățile cdf comună

Cdf comună  $F(x, y)$  a lui  $X$  și  $Y$  trebuie să satisfacă proprietățile:

1.  $F$  este crescătoare: i.e. dacă  $x$  sau  $y$  cresc, atunci  $F$  trebuie să rămână constantă sau să crească.

2.  $F(x, y) = 0$  la stânga jos a domeniului de valori comun.

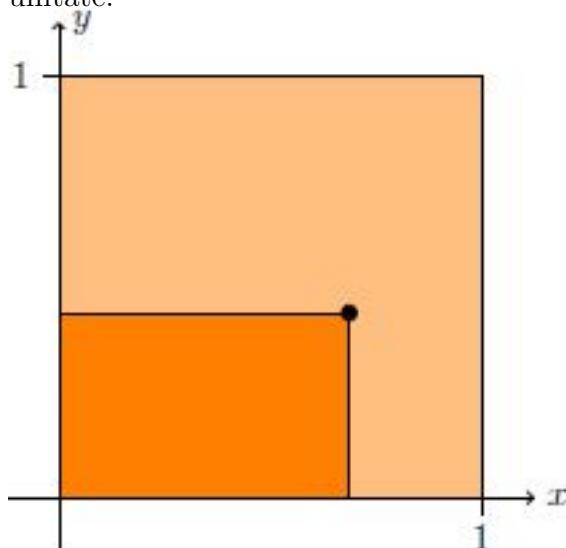
$$\lim_{(x,y) \rightarrow (-\infty, -\infty)} F(x, y) = 0.$$

3.  $F(x, y) = 1$  la dreapta sus a domeniului de valori comun.

$$\lim_{(x,y) \rightarrow (\infty, \infty)} F(x, y) = 1.$$

**Exemplul 6.** Aflați cdf comună pentru variabilele aleatoare din exemplul 5.

**Răspuns.** Evenimentul " $X \leq x$  și  $Y \leq y$ " este un dreptunghi în pătratul unitate.



Pentru a afla cdf  $F$ , calculăm o integrală dublă:

$$F(x, y) = \int_0^y \int_0^x 4uv du dv = \int_0^y 2u^2 v|_{u=0}^{u=x} dv = \int_0^y 2x^2 v dv = x^2 v^2|_{v=0}^{v=y} = x^2 y^2.$$

**Exemplul 7.** În exemplul 1, calculați  $F(3.5, 4)$ .

**Răspuns.** Redesenăm tabelul de probabilitate comună. Figura este similară cu cea din exemplul precedent.

$F(3.5, 4)$  este probabilitatea evenimentului " $X \leq 3.5$  și  $Y \leq 4$ ". Putem vizualiza acest eveniment ca dreprunghiurile colorate din tabel:

$X \setminus Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Evenimentul " $X \leq 3.5$  și  $Y \leq 4$ ".

Adunând probabilitățile din dreptunghiurile colorate obținem  $F(3.5, 4) = 12/36 = 1/3$ .

**Observație.** O diferență nefericită între vizualizările continuă și discretă este că pentru variabile continue valoarea crește când mergem în sus în direcția verticală în timp ce contrarul este adevărat pentru cazul discret.

### 2.3.6 Repartiții marginale

Când  $X$  și  $Y$  sunt variabile aleatoare repartizate comun, putem considera doar una din ele, să zicem  $X$ . În acest caz avem nevoie să aflăm pmf (sau pdf sau cdf) a lui  $X$  fără  $Y$ . Aceasta este numită o **pmf** (sau pdf sau cdf) **marginală**. Următorul exemplu ilustreză modul de a calcula aceasta și motivul pentru termenul "marginal".

### 2.3.7 Pmf marginală

**Exemplul 8.** În exemplul 2 am aruncat 2 zaruri corecte și fie  $X$  valoarea primului zar și  $T$  totalul ambelor zaruri. Calculați pmf-urile marginale ale lui  $X$  și  $T$ .

**Răspuns.** În tabel fiecare linie reprezintă o singură valoare a lui  $X$ . Astfel, evenimentul " $X = 3$ " este linia colorată a tabelului. Pentru a afla  $P(X = 3)$  trebuie să adunăm probabilitățile din această linie. Punem suma în **marginea dreaptă** a tabelului. Analog,  $P(T = 5)$  este chiar suma coloanei cu  $T = 5$ . Punem suma în **marginea de jos** a tabelului.

$X \setminus T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0	1/6
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	1/6
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	1/6
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	1/6
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	1/6
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(t_j)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

Calculul probabilităților marginale  $P(X = 3) = 1/6$  și  $P(T = 5) = 4/36 = 1/9$ .

**Observație.** În acest caz știam deja pmf-urile lui  $X$  și  $T$ . Rezultatele de aici coincid cu cele anterioare.

Pmf-urile marginale se obțin din pmf comună prin adunare:

$$p_X(x_i) = \sum_j p(x_i, y_j), \quad p_Y(y_j) = \sum_i p(x_i, y_j).$$

Termenul "marginal" se referă la faptul că valorile sunt scrise pe marginea tabelului.

### 2.3.8 Pdf marginală

Pentru densitatea comună continuă  $f(x, y)$  cu domeniul  $[a, b] \times [c, d]$ , pdf-urile marginale sunt:

$$f_X(x) = \int_c^d f(x, y) dy, \quad f_Y(y) = \int_a^b f(x, y) dx.$$

Comparând aceasta cu pmf-urile marginale de mai sus, ca de obicei sumele sunt înlocuite de integrale.

Spunem că pentru a obține marginala pentru  $X$ , integrăm după  $y$  pdf comună și viceversa.

**Exemplul 9.** Presupunem că  $(X, Y)$  ia valori în pătratul  $[0, 1] \times [1, 2]$  cu pdf comună  $f(x, y) = \frac{8}{3}x^3y$ . Aflați pdf-urile marginale  $f_X(x)$  și  $f_Y(y)$ .

**Răspuns.** Pentru a afla  $f_X(x)$  integrăm după  $y$  și pentru a afla  $f_Y(y)$  integrăm după  $x$ .

$$f_X(x) = \int_1^2 \frac{8}{3}x^3y dy = \frac{4}{3}x^3y^2 \Big|_{y=1}^{y=2} = \frac{16}{3}x^3 - \frac{4}{3}x^3 = 4x^3.$$

$$f_Y(y) = \int_0^1 \frac{8}{3} x^3 y dx = \frac{2}{3} x^4 y \Big|_{x=0}^{x=1} = \frac{2}{3} y.$$

**Exemplul 10.** Presupunem că  $(X, Y)$  ia valori în pătratul unitate  $[0, 1] \times [0, 1]$  cu pdf comună  $f(x, y) = \frac{3}{2}(x^2 + y^2)$ . Aflați pdf marginală  $f_X(x)$  și utilizați-o pentru a afla  $P(X < 0.5)$ .

**Răspuns.**

$$f_X(x) = \int_0^1 \frac{3}{2} (x^2 + y^2) dy = \left( \frac{3}{2} x^2 y + \frac{y^3}{2} \right) \Big|_{y=0}^{y=1} = \frac{3}{2} x^2 + \frac{1}{2}.$$

$$P(X < 0.5) = \int_0^{0.5} f_X(x) dx = \int_0^{0.5} \left( \frac{3}{2} x^2 + \frac{1}{2} \right) dx = \left( \frac{1}{2} x^3 + \frac{1}{2} x \right) \Big|_0^{0.5} = \frac{1}{16} + \frac{1}{4} = \frac{5}{16}.$$

### 2.3.9 Cdf marginală

Aflarea cdf marginală din cdf comună se face astfel: dacă  $X$  și  $Y$  iau comun valori pe  $[a, b] \times [c, d]$ , atunci

$$F_X(x) = F(x, d), \quad F_Y(y) = F(b, y).$$

Dacă  $d = \infty$  (caz în care  $[c, d]$  se consideră deschis în  $d$ ), atunci  $F_X(x) = \lim_{y \rightarrow \infty} F(x, y)$ . Analog pentru  $F_Y(y)$ .

**Exemplul 11.** Cdf comună în ultimul exemplu a fost  $F(x, y) = \frac{1}{2}(x^3 y + x y^3)$  pe  $[0, 1] \times [0, 1]$ . Aflați cdf-urile marginale și utilizați  $F_X(x)$  pentru a calcula  $P(X < 0.5)$ .

**Răspuns.** Avem  $F_X(x) = F(x, 1) = \frac{1}{2}(x^3 + x)$  și  $F_Y(y) = F(1, y) = \frac{1}{2}(y + y^3)$ . Deci  $P(X < 0.5) = F_X(0.5) = \frac{1}{2}(0.5^3 + 0.5) = \frac{5}{16}$ , exact la fel ca la exemplul 10.

### 2.3.10 Vizualizare 3D

Am vizualizat  $P(a < x < b)$  ca aria de sub pdf  $f(x)$  pe intervalul  $[a, b]$ . Deoarece domeniul de valori al lui  $(X, Y)$  este deja o regiune bidimensională în plan, graficul lui  $f(x, y)$  este o suprafață peste acea regiune. Putem atunci vizualiza probabilitatea ca **volumul** de sub suprafață.

## 2.4 Independență

Reamintim că evenimentele  $A$  și  $B$  sunt independente  $\iff$

$$P(A \cap B) = P(A)P(B).$$

Variabilele aleatoare  $X$  și  $Y$  definesc evenimente ca ” $X \leq 2$ ” sau ” $Y > 5$ ”. Astfel,  $X$  și  $Y$  sunt independente dacă orice eveniment definit de  $X$  este independent de orice eveniment definit de  $Y$ . Iată definiția care garantează aceasta:

**Definiție.** Variabilele aleatoare repartizate comun  $X$  și  $Y$  sunt independente  $\iff$  cdf comună a lor este produsul cdf-urilor marginale:

$$F(x, y) = F_X(x)F_Y(y).$$

Pentru variabile discrete aceasta este echivalent cu pmf comună să fie produsul pmf-urilor marginale:

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j).$$

Pentru variabile continue aceasta este echivalent cu pdf comună să fie produsul pdf-urilor marginale:

$$f(x, y) = f_X(x)f_Y(y).$$

**Exemplul 12.** Pentru variabile discrete independență înseamnă că probabilitatea din orice celulă trebuie să fie produsul dintre probabilitățile marginale de pe linia și coloana ei. În primul tabel de mai jos aceasta este adevărat: fiecare probabilitate marginală este  $1/6$  și fiecare celulă are  $1/36$ , adică produsul probabilităților marginale. De aceea  $X$  și  $Y$  sunt independente.

În al 2-lea tabel de mai jos, există probabilități ale celulelor care nu sunt produsul probabilităților marginale. De exemplu, nicio probabilitate marginală nu este 0, deci nicio probabilitate 0 dintr-o celulă nu poate fi produsul marginalelor. De aceea  $X$  și  $T$  nu sunt independente.

$X \setminus Y$	1	2	3	4	5	6	$p(x_i)$
$p(y_j)$	1/6	1/6	1/6	1/6	1/6	1/6	1
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6

$X \setminus T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
$p(y_j)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0	1/6
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	1/6
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	1/6
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	1/6
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	1/6
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	1/6

**Exemplul 13** Pentru variabile continue independentă înseamnă că putem factoriza pdf sau cdf comună ca produsul unei funcții de  $x$  cu o funcție de  $y$ .

- (i) Presupunând că  $X$  are domeniul  $[0, 1/2]$ ,  $Y$  are domeniul  $[0, 1]$  și  $f(x, y) = 96x^2y^3$ , atunci  $X$  și  $Y$  sunt independente. Densitățile marginale sunt  $f_X(x) = 24x^2$  și  $f_Y(y) = 4y^3$ .
- (ii) Dacă  $f(x, y) = 1.5(x^2 + y^2)$  pe pătratul unitate, atunci  $X$  și  $Y$  nu sunt independente deoarece nu există niciun mod de a factoriza  $f(x, y)$  într-un produs  $f_X(x)f_Y(y)$ .
- (iii) Dacă  $F(x, y) = \frac{1}{2}(x^3y + xy^3)$  pe pătratul unitate, atunci  $X$  și  $Y$  nu sunt independente deoarece cdf nu se factorizează într-un produs  $F_X(x)F_Y(y)$ .

# Curs 8

Cristian Niculescu

## 1 Covarianță și corelație

### 1.1 Scopurile învățării

1. Să înțeleagă covarianță și corelația.
2. Să poată calcula covarianță și corelația a 2 variabile aleatoare.

### 1.2 Covarianță

Covarianța este o măsură a cât de mult 2 variabile aleatoare variază împreună. De exemplu, înălțimea și greutatea girafelor au covarianță pozitivă deoarece, când una este mare, celalaltă tinde de asemenea să fie mare.

**Definiție.** Presupunem că  $X$  și  $Y$  sunt variabile aleatoare cu mediile  $\mu_X$  și  $\mu_Y$ . **Covarianța** lui  $X$  și  $Y$  este

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

#### 1.2.1 Proprietățile covarianței

1.  $Cov(aX + b, cY + d) = acCov(X, Y), \forall a, b, c, d$  constante.
2.  $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y).$
3.  $Cov(X, X) = Var(X).$
4.  $Cov(X, Y) = E(XY) - \mu_X\mu_Y.$
5.  $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y), \forall X, Y.$
6. Dacă  $X$  și  $Y$  sunt independente, atunci  $Cov(X, Y) = 0$ .

**Avertizare.** Reciproca este falsă: covarianță 0 nu implică totdeauna independentă.

**Observații.** 1. Proprietatea 4 este ca proprietatea similară pentru dispersie.

Într-adevăr, dacă  $X = Y$ , este exact acea proprietate:

$$Var(X) = E(X^2) - \mu_X^2.$$

Proprietatea 6 reduce formula din proprietatea 5 la formula anterioară  
 $Var(X + Y) = Var(X) + Var(Y)$  când  $X$  și  $Y$  sunt independente.

2.  $Cov(X, Y) = Cov(Y, X), \forall X, Y.$

### 1.2.2 Sume și integrale pentru calculul covarianței

Deoarece covarianța este definită ca o medie, o calculăm ca de obicei ca o sumă sau integrală.

**Cazul discret.** Dacă  $X$  și  $Y$  are pmf  $p(x_i, y_j)$  atunci

$$Cov(X, Y) = \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) (x_i - \mu_X) (y_j - \mu_Y) = \left( \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) x_i y_j \right) - \mu_X \mu_Y.$$

**Cazul continuu.** Dacă  $X$  și  $Y$  au pdf comună  $f(x, y)$  pe domeniul  $[a, b] \times [c, d]$ , atunci

$$Cov(X, Y) = \int_c^d \int_a^b (x - \mu_X) (y - \mu_Y) f(x, y) dx dy = \left( \int_c^d \int_a^b xy f(x, y) dx dy \right) - \mu_X \mu_Y.$$

### 1.2.3 Exemple

**Exemplul 1.** Aruncăm o monedă corectă de 3 ori. Fie  $X$  numărul de aversuri în primele 2 aruncări și  $Y$  numărul de aversuri din ultimele 2 aruncări (astfel încât este o suprapunere peste aruncarea din mijloc). Calculați  $Cov(X, Y)$ .

**Răspuns.** Metoda I (folosind tabelul probabilităților comune și definiția covarianței).

Cu 3 aruncări sunt doar 8 rezultate

$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ , deci putem face direct tabelul de probabilitate comună.

$X \setminus Y$	0	1	2	$p(x_i)$
0	1/8	1/8	0	1/4
1	1/8	2/8	1/8	1/2
2	0	1/8	1/8	1/4
$p(y_j)$	1/4	1/2	1/4	1

Din marginale calculăm  $E(X) = 1 = E(Y)$ . Acum folosim [definiția](#):

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = \sum_{i,j} p(x_i, y_j) (x_i - 1)(y_j - 1).$$

Scriem suma eliminând termenii care sunt 0, i.e. toți termenii unde  $x_i = 1$  sau  $y_j = 1$  sau probabilitatea este 0.

$$Cov(X, Y) = \frac{1}{8}(0-1)(0-1) + \frac{1}{8}(2-1)(2-1) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

Metoda II (folosind tabelul probabilităților comune și proprietatea 4).

Ca la metoda I se calculează tabelul probabilităților comune și mediile lui  $X$  și  $Y$ . Din tabel calculăm

$$E(XY) = \sum_{i,j} p(x_i, y_j) x_i y_j = 1 \cdot \frac{2}{8} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 4 \cdot \frac{1}{8} = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{2} = \frac{5}{4}.$$

Din proprietatea 4,

$$Cov(X, Y) = E(XY) - \mu_X \mu_Y = \frac{5}{4} - 1 = \frac{1}{4}.$$

Metoda III (folosind proprietățile covarianței). Ca de obicei, fie  $X_i$  rezultatul celei de-a  $i$ -a aruncări, deci  $X_i \sim \text{Bernoulli}(0.5)$ . Avem

$$X = X_1 + X_2 \text{ și } Y = X_2 + X_3.$$

Știm că  $E(X_i) = 1/2$  și  $Var(X_i) = 1/4, \forall i = 1, 2, 3$ . De aceea, folosind proprietatea 2 a covarianței și observația 2, avem

$$Cov(X, Y) = Cov(X_1 + X_2, X_2 + X_3) = Cov(X_1, X_2) + Cov(X_1, X_3) + Cov(X_2, X_2) + Cov(X_2, X_3).$$

Deoarece aruncările diferite sunt independente, din proprietatea 6,

$$Cov(X_1, X_2) = Cov(X_1, X_3) = Cov(X_2, X_3) = 0.$$

Deci, din proprietatea 3,

$$Cov(X, Y) = Cov(X_2, X_2) = Var(X_2) = \frac{1}{4}.$$

**Exemplul 2.** (Covarianță 0 nu implică independentă.) Fie  $X \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$ .

Fie  $Y = X^2$ . Arătați că avem  $Cov(X, Y) = 0$ , dar  $X$  și  $Y$  nu sunt independente.

**Răspuns.** Facem un tabel de probabilitate comună

$Y \setminus X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	$1/5$	0	0	$1/5$
1	0	$1/5$	0	$1/5$	0	$2/5$
4	$1/5$	0	0	0	$1/5$	$2/5$
$p(x_i)$	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$	1

Folosind marginalele calculăm  $E(X) = 0$  și  $E(Y) = 2$ .

Calculăm covarianța folosind proprietatea 4:

$$Cov(X, Y) = \sum_{i,j} p(x_i, y_j) x_i y_j - \mu_X \mu_Y = \frac{1}{5}(-8 - 1 + 1 + 8) - 0 \cdot 2 = 0.$$

Arătăm că  $X$  și  $Y$  nu sunt independente. Pentru a face asta, trebuie să găsim un loc unde regula produsului eșuează, i.e. unde  $p(x_i, y_j) \neq p(x_i)p(y_j)$ :

$$P(X = -2, Y = 0) = 0 \text{ dar } P(X = -2)P(Y = 0) = (1/5)(1/5) = 1/25.$$

Deoarece acestea nu sunt egale,  $X$  și  $Y$  nu sunt independente.

**Discuție.** Acest exemplu arată  $Cov(X, Y) = 0$  nu implică  $X$  și  $Y$  sunt independente. De fapt,  $X$  și  $X^2$  sunt atât de dependente cât pot fi variabilele aleatoare: dacă știm valoarea lui  $X$ , atunci știm valoarea lui  $X^2$  cu 100% certitudine.

$Cov(X, Y)$  măsoară relația liniară dintre  $X$  și  $Y$ . În exemplul de mai sus  $X$  și  $X^2$  au o relație pătratică complet pierdută de  $Cov(X, Y)$ .

#### 1.2.4 Demonstrațiile proprietăților covarianței

Proprietățile 1 și 2 rezultă din proprietățile similare pentru medie.

3.

$$Cov(X, X) = E((X - \mu_X)(X - \mu_X)) = E((X - \mu_X)^2) = Var(X).$$

4.

$$\begin{aligned} Cov(X, Y) &= E((X - \mu_X)(Y - \mu_Y)) \\ &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y. \end{aligned}$$

5. Folosind proprietățile 3 și 2 și observația 2 obținem

$$\begin{aligned} Var(X + Y) &= Cov(X + Y, X + Y) = Cov(X, X) + 2Cov(X, Y) + Cov(Y, Y) \\ &= Var(X) + Var(Y) + 2Cov(X, Y). \end{aligned}$$

6. Reamintim că  $E(X - \mu_X) = 0$ . Dacă  $X$  și  $Y$  sunt independente, atunci  $f(x, y) = f_X(x)f_Y(y)$ . De aceea

$$\begin{aligned} Cov(X, Y) &= \int \int (x - \mu_X)(y - \mu_Y)f_X(x)f_Y(y)dxdy \\ &= \int (x - \mu_X)f_X(x)dx \int (y - \mu_Y)f_Y(y)dy \\ &= E(X - \mu_X)E(Y - \mu_Y) \\ &= 0. \end{aligned}$$

### 1.3 Corelația

Unitățile de măsură ale covarianței  $Cov(X, Y)$  sunt ”unitățile lui  $X$  ori unitățile lui  $Y$ ”. Aceasta face grea compararea covarianțelor: dacă schimbăm scalele, atunci se schimbă și covarianța. Corelația este un mod de a elimina scara din covarianță.

**Definiție.** Coeficientul de corelație între  $X$  și  $Y$  este definit de

$$Cor(X, Y) = \rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

#### 1.3.1 Proprietăți ale corelației

1.  $\rho$  este covarianța standardizărilor lui  $X$  și  $Y$ .
2.  $\rho$  este fără dimensiune.
3.  $-1 \leq \rho \leq 1$ . Mai mult,

$$\rho = 1 \iff Y = aX + b \text{ cu } a > 0,$$

$$\rho = -1 \iff Y = aX + b \text{ cu } a < 0.$$

Proprietatea 3 arată că  $\rho$  măsoară relația liniară între variabile. Când corelația este pozitivă, atunci, dacă  $X$  este mare,  $Y$  va fi mare de asemenea. Când corelația este negativă, atunci, dacă  $X$  este mare,  $Y$  va fi mic.

Exemplul 2 de mai sus arată că relațiile de ordin mai mare pot fi ratate complet de corelație.

#### 1.3.2 Demonstrația proprietății 3 a corelației

$$0 \leq Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) - 2Cov\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = 2 - 2\rho \implies \rho \leq 1.$$

$$\text{Analog } 0 \leq Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \implies \rho \geq -1.$$

$$\rho = 1 \iff Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0 \iff \frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c \iff Y = aX + b \text{ cu}$$

$a = \frac{\sigma_Y}{\sigma_X} > 0$  și  $b = -c\sigma_Y$ .

Analog  $\rho = -1 \iff Y = aX + b$  cu  $a = -\frac{\sigma_Y}{\sigma_X} < 0$ .

**Exemplu.** Continuăm exemplul 1. Pentru a calcula corelația împărțim covarianța la deviațiile standard. În exemplul 1 am aflat  $Cov(X, Y) = 1/4$  și  $Var(X) = 2Var(X_j) = 1/2$ . Deci,  $\sigma_X = 1/\sqrt{2}$ . Analog  $\sigma_Y = 1/\sqrt{2}$ . Astfel,

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{1/4}{1/2} = \frac{1}{2}.$$

Vedem o corelație pozitivă, ceea ce înseamnă că  $X$  mai mare tinde să meargă cu  $Y$  mai mare și  $X$  mai mic cu  $Y$  mai mic. În exemplul 1 aceasta se întâmplă deoarece aruncarea 2 este inclusă atât în  $X$  cât și în  $Y$ , deci contribuie la mărimea ambelor.

### 1.3.3 Repartiții normale bivariate

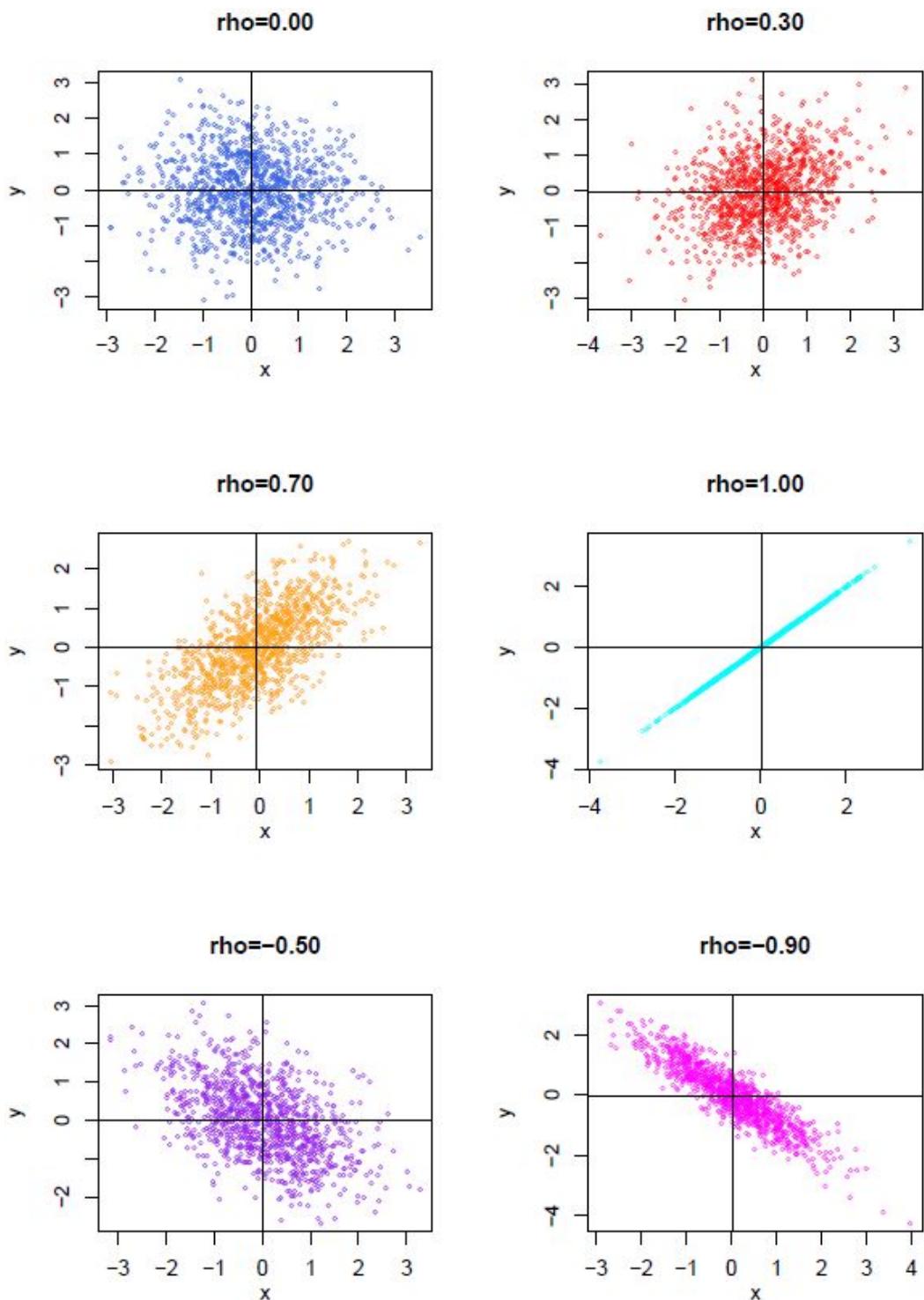
Repartiția normală bivariată are densitatea

$$f(x, y) = \frac{e^{\frac{-1}{2(1-\rho^2)} \left[ \frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} \right]}}{2\pi\sigma_X \sigma_Y \sqrt{1-\rho^2}}.$$

Pentru această repartiție, repartițiile marginale pentru  $X$  și  $Y$  sunt normale și corelația între  $X$  și  $Y$  este  $\rho$ .

În figurile de mai jos a fost utilizat R pentru a simula repartiția pentru diferite valori ale lui  $\rho$ . Individual  $X$  și  $Y$  sunt normale standard, i.e.  $\mu_X = \mu_Y = 0$  și  $\sigma_X = \sigma_Y = 1$ . Figurile arată reprezentările rezultatelor.

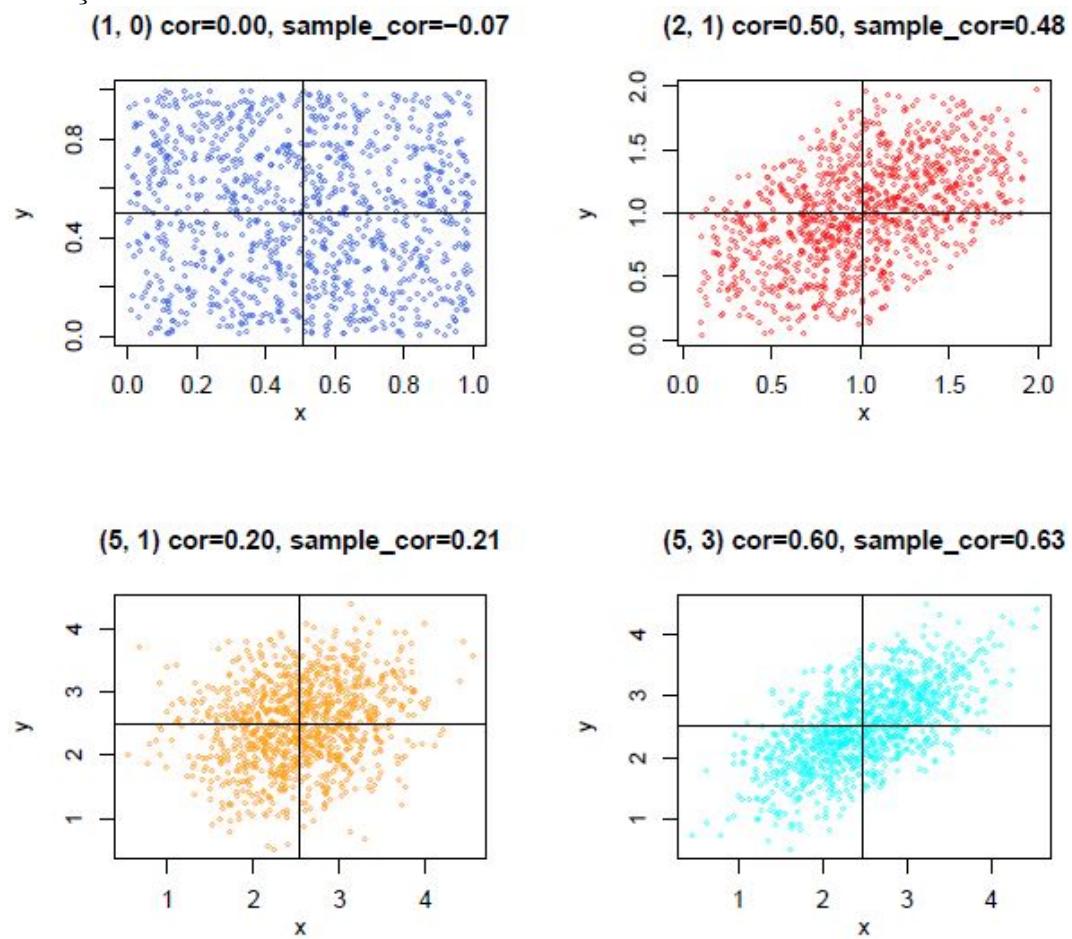
Aceste reprezentări și setul următor arată o proprietate importantă a corelației. Împărțim datele în cadrane desenând o linie orizontală la media datelor  $y$  și una verticală la media datelor  $x$ . O corelație pozitivă corespunde tendinței datelor de a se afla în cadranele 1 și 3. O corelație negativă corespunde tendinței datelor de a se afla în cadranele 2 și 4. Putem vedea datele adunându-se de-a lungul unei linii când  $\rho$  devine apropiat de  $\pm 1$ .

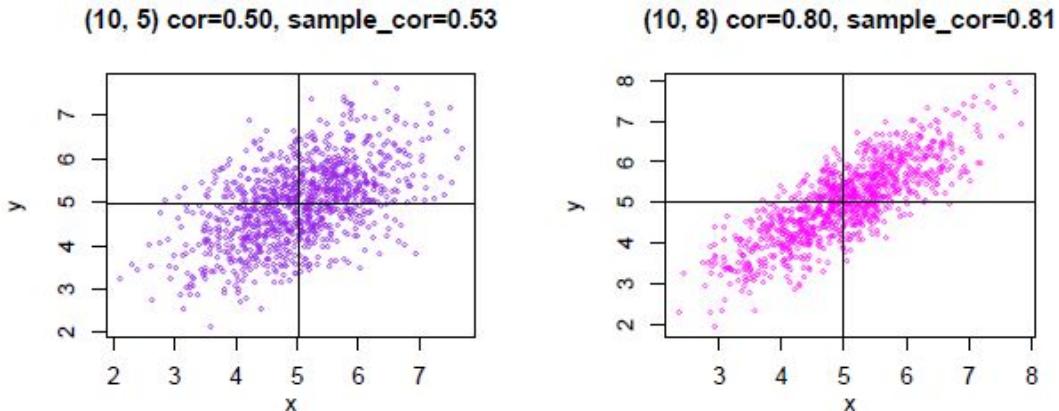


### 1.3.4 Suprapunerea repartițiilor uniforme

Fie  $X_1, X_2, \dots, X_{20}$  i.i.d repartizate  $U(0, 1)$ .  $X$  și  $Y$  sunt ambele sume de același număr de  $X_i$ -uri. Numim numărul de  $X_i$ -uri comune lui  $X$  și  $Y$  suprapunerea. Notația din figurile de mai jos indică numărul de  $X_i$ -uri adunate și numărul care se suprapun. De exemplu, 5,3 arată că  $X$  și  $Y$  erau fiecare sumă de 5  $X_i$ -uri și că 3 dintre  $X_i$ -uri erau comune ambelor sume. (Datele au fost generate în R folosind `rand(1, 1000)` ;.)

Folosind liniaritatea covarianței se calculează corelația teoretică. Pentru fiecare reprezentare sunt date atât corelația teroretică, cât și corelația datelor din eșantionul simutat.





## 2 Probleme recapitulative

### 2.1 Numărare și probabilități

- 1) a) În câte moduri putem aranja literele cuvântului STATISTICS? (De exemplu, SSSTTTIAC este un aranjament).
- b) Dacă toate aranjamentele sunt egal posibile, care este probabilitatea ca "I"-urile să fie unul după altul?

### 2.2 Probabilitate condiționată și teorema lui Bayes

- 2) Corupt de puterea lui, juriul showului tv *Următorul matematician de top al Americii* a luat mită de la unii participanți. În fiecare episod, fiecare participant rămâne în show sau este eliminat.

Dacă un participant a mituit juriul, rămâne în show cu probabilitatea 1. Dacă un participant nu a mituit juriul, rămâne în show cu probabilitatea  $1/3$ .

De-a lungul a 2 episoade, presupunem că  $1/4$  dintre participanți au mituit juriul. Aceiași participanți mituiesc juriul în ambele runde, i.e., dacă un participant dă mită în prima rundă, dă mită și în a 2-a (și viceversa).

- a) Dacă alegem aleator un participant care a rămas în show după primul episod, care este probabilitatea să fi mituit juriul?
- b) Dacă alegem aleator un participant, care este probabilitatea să rămână în show după ambele episoade?
- c) Dacă alegem aleator un participant care a rămas în show după primul episod, care este probabilitatea să fie eliminat în al 2-lea episod?

## 2.3 Independență

- 3) Aruncăm un zar corect cu 20 de fețe, numerotate de la 1 la 20. Determinați dacă următoarele perechi de evenimente sunt independente.
- "Număr par" și "Număr  $\leq 10$ ".
  - "Număr par" și "Număr prim".

## 2.4 Medie și dispersie

- 4) Fie  $X \sim \begin{pmatrix} -1 & 0 & 1 \\ 1/8 & 1/4 & 5/8 \end{pmatrix}$ .
- Calculați  $E(X)$ .
  - Dați pmf pentru  $Y = X^2$  și folosiți-o pentru a calcula  $E(Y)$ .
  - În loc de aceasta, calculați  $E(X^2)$  direct dintr-un tabel extins.
  - Calculați  $\text{Var}(X)$ .
  - Calculați media și dispersia unei variabile aleatoare Bernoulli( $p$ ).
  - Presupunem că 100 de oameni își aruncă pălăria într-o cutie și apoi aleg aleator câte o pălărie din cutie. Care este media numărului de oameni care își iau înapoi pălăria proprie?

Indicație: Exprimăți numărul de oameni care își iau înapoi propria pălărie ca o sumă de variabile aleatoare a căror medie este ușor de calculat.

## 2.5 Funcții masă de probabilitate, funcții densitate de probabilitate și funcții de repartiție

- 7) a) Presupunem că  $X$  are funcția densitate de probabilitate  $f_X(x) = \lambda e^{-\lambda x}$  pentru  $x \geq 0$ . Calculați cdf,  $F_X(x)$ .
- b) Dacă  $Y = X^2$ , calculați pdf și cdf ale lui  $Y$ .
- 8) Presupunem că aruncăm un zar corect de 100 de ori (independent) și primim 3 dolari de fiecare dată când dăm 6. Fie  $X_1$ , numărul de dolari primiți în aruncările de la 1 la 25.

Fie  $X_2$ , numărul de dolari primiți în aruncările de la 26 la 50.

Fie  $X_3$ , numărul de dolari primiți în aruncările de la 51 la 75.

Fie  $X_4$ , numărul de dolari primiți în aruncările de la 76 la 100.

Fie  $X = X_1 + X_2 + X_3 + X_4$ , numărul de dolari primiți în 100 de aruncări.

a) Care este funcția masă de probabilitate a lui  $X$ ?

b) Care este media și dispersia lui  $X$ ?

c) Fie  $Y = 4X_1$ . (Deci, în loc să aruncăm de 100 de ori, aruncăm doar de 25 de ori și înmulțim câștigurile cu 4.)

i) Care sunt media și dispersia lui  $Y$ ?

- ii) Cum sunt media și dispersia lui  $Y$  în comparație cu ale lui  $X$ ? (I.e., sunt mai mari, mai mici, sau egale?) Explicați pe scurt de ce aceasta are sens.

## 2.6 Probabilitate comună, covarianță, corelație

- 9) (**Puzzle aritmetic**) Pmf-urile comune și marginale ale lui  $X$  și  $Y$  sunt parțial date în următoarul tabel.

$X \setminus Y$	1	2	3	
1	1/6	0	...	1/3
2	...	1/4	...	1/3
3	...	...	1/4	...
	1/6	1/3	...	1

- a) Completați tabelul.  
b) Sunt  $X$  și  $Y$  independente?

10) **Covarianță și independentă**

Fie  $X \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$ .

Fie  $Y = X^2$ .

- a) Completați următorul tabel dând funcția de frecvență comună pentru  $X$  și  $Y$ . Includeți probabilitățile marginale.

$X$	-2	-1	0	1	2	total
$Y$						
0						
1						
4						
total						

- b) Aflați  $E(X)$  și  $E(Y)$ .  
c) Arătați că  $X$  și  $Y$  nu sunt independente.  
d) Arătați  $\text{Cov}(X, Y) = 0$ .

Acesta este un exemplu de variabile aleatoare necorelate dar dependente. Motivul pentru care aceasta se poate întâmpla este că doar dependența liniară este măsurată de corelație. În acest caz,  $X$  și  $Y$  nu sunt legate liniar.

11) **Repartiții comune continue.** Presupunem că  $X$  și  $Y$  sunt variabile aleatoare continue cu funcția de densitate comună  $f(x, y) = x + y$  pe pătratul unitate  $[0, 1] \times [0, 1]$ .

- a) Fie  $F(x, y)$  cdf comună. Calculați  $F(1, 1)$ . Calculați  $F(x, y)$ .

- b) Calculați densitățile marginale pentru  $X$  și  $Y$ .
- c) Sunt  $X$  și  $Y$  independente?
- d) Calculați  $E(X)$ ,  $E(Y)$ ,  $E(X^2 + Y^2)$ ,  $Cov(X, Y)$ .

## 2.7 Legea numerelor mari, teorema limită centrală

- 12) Presupunem că  $X_1, \dots, X_{100}$  sunt i.i.d. cu media  $1/5$  și dispersia  $1/9$ . Utilizați teorema limită centrală pentru a estima  $P(\sum X_i < 30)$ .
- 13) (Mai mult cu teorema limită centrală)  
IQ-ul mediu al unei populații este 100 cu deviația standard 15 (prin definiție, IQ-ul este normalizat, deci aşa este și aici). Care este probabilitatea ca un grup de 100 de oameni selectat aleator să aibă un IQ mediu peste 115?

**Tabel standard normal al probabilităților cozii stângi.**

$\Phi(z) = P(Z \leq z)$  pentru  $N(0, 1)$ . (Folosiți interpolarea pentru a estima valurile pentru a 3-a zecimală a lui  $z$ .)

$z$	$\Phi(z)$	$z$	$\Phi(z)$	$z$	$\Phi(z)$	$z$	$\Phi(z)$
-4.00	0.0000	-2.00	0.0228	0.00	0.5000	2.00	0.9772
-3.95	0.0000	-1.95	0.0256	0.05	0.5199	2.05	0.9798
-3.90	0.0000	-1.90	0.0287	0.10	0.5398	2.10	0.9821
-3.85	0.0001	-1.85	0.0322	0.15	0.5596	2.15	0.9842
-3.80	0.0001	-1.80	0.0359	0.20	0.5793	2.20	0.9861
-3.75	0.0001	-1.75	0.0401	0.25	0.5987	2.25	0.9878
-3.70	0.0001	-1.70	0.0446	0.30	0.6179	2.30	0.9893
-3.65	0.0001	-1.65	0.0495	0.35	0.6368	2.35	0.9906
-3.60	0.0002	-1.60	0.0548	0.40	0.6554	2.40	0.9918
-3.55	0.0002	-1.55	0.0606	0.45	0.6736	2.45	0.9929
-3.50	0.0002	-1.50	0.0668	0.50	0.6915	2.50	0.9938
-3.45	0.0003	-1.45	0.0735	0.55	0.7088	2.55	0.9946
-3.40	0.0003	-1.40	0.0808	0.60	0.7257	2.60	0.9953
-3.35	0.0004	-1.35	0.0885	0.65	0.7422	2.65	0.9960
-3.30	0.0005	-1.30	0.0968	0.70	0.7580	2.70	0.9965
-3.25	0.0006	-1.25	0.1056	0.75	0.7734	2.75	0.9970
-3.20	0.0007	-1.20	0.1151	0.80	0.7881	2.80	0.9974
-3.15	0.0008	-1.15	0.1251	0.85	0.8023	2.85	0.9978
-3.10	0.0010	-1.10	0.1357	0.90	0.8159	2.90	0.9981
-3.05	0.0011	-1.05	0.1469	0.95	0.8289	2.95	0.9984
-3.00	0.0013	-1.00	0.1587	1.00	0.8413	3.00	0.9987
-2.95	0.0016	-0.95	0.1711	1.05	0.8531	3.05	0.9989
-2.90	0.0019	-0.90	0.1841	1.10	0.8643	3.10	0.9990
-2.85	0.0022	-0.85	0.1977	1.15	0.8749	3.15	0.9992
-2.80	0.0026	-0.80	0.2119	1.20	0.8849	3.20	0.9993
-2.75	0.0030	-0.75	0.2266	1.25	0.8944	3.25	0.9994
-2.70	0.0035	-0.70	0.2420	1.30	0.9032	3.30	0.9995
-2.65	0.0040	-0.65	0.2578	1.35	0.9115	3.35	0.9996
-2.60	0.0047	-0.60	0.2743	1.40	0.9192	3.40	0.9997
-2.55	0.0054	-0.55	0.2912	1.45	0.9265	3.45	0.9997
-2.50	0.0062	-0.50	0.3085	1.50	0.9332	3.50	0.9998
-2.45	0.0071	-0.45	0.3264	1.55	0.9394	3.55	0.9998
-2.40	0.0082	-0.40	0.3446	1.60	0.9452	3.60	0.9998
-2.35	0.0094	-0.35	0.3632	1.65	0.9505	3.65	0.9999
-2.30	0.0107	-0.30	0.3821	1.70	0.9554	3.70	0.9999
-2.25	0.0122	-0.25	0.4013	1.75	0.9599	3.75	0.9999
-2.20	0.0139	-0.20	0.4207	1.80	0.9641	3.80	0.9999
-2.15	0.0158	-0.15	0.4404	1.85	0.9678	3.85	0.9999
-2.10	0.0179	-0.10	0.4602	1.90	0.9713	3.90	1.0000
-2.05	0.0202	-0.05	0.4801	1.95	0.9744	3.95	1.0000

# Curs 9

Cristian Niculescu

## 1 Soluțiile problemelor recapitulative

### 1.1 Numărare și probabilitate

1) a) Creăm un aranjament în etape și calculăm numărul de posibilități la fiecare etapă:

Etapa 1: alegem 3 din cele 10 locuri pentru a pune S-urile:  $C_{10}^3$ .

Etapa 2: alegem 3 din cele 7 locuri rămase pentru a pune T-urile:  $C_7^3$ .

Etapa 3: alegem 2 din cele 4 locuri rămase pentru a pune I-urile:  $C_4^2$ .

Etapa 4: alegem 1 din cele 2 locuri rămase pentru a pune A-ul:  $C_2^1$ .

Etapa 5: folosim ultimul loc pentru C:  $C_1^1$ .

Numărul de moduri de aranjare este:

$$C_{10}^3 \cdot C_7^3 \cdot C_4^2 \cdot C_2^1 \cdot C_1^1 = 120 \cdot 35 \cdot 6 \cdot 2 \cdot 1 = 50400.$$

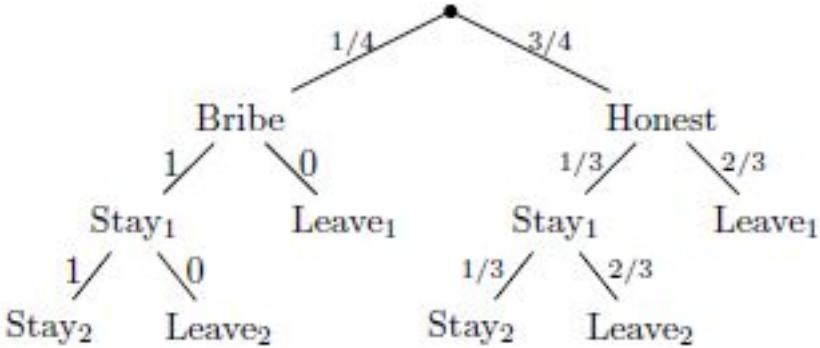
b) Sunt  $C_{10}^2 = 45$  de moduri egal posibile de a plasa cele 2 I-uri.

Sunt 9 moduri de a le plasa unul lângă altul, i.e. pe pozițiile 1 și 2, pozițiile 2 și 3,..., pozițiile 9 și 10.

Deci probabilitatea ca I-urile să fie adiacente este  $9/45 = 1/5 = 0.2$ .

### 1.2 Probabilitate condiționată și teorema lui Bayes

2) Facem un arbore. Stay<sub>1</sub> înseamnă că participantului i s-a permis să rămână în joc după primul episod și Stay<sub>2</sub> că i s-a permis să rămână în joc după al 2-lea episod.



Evenimentele relevante sunt:

$B$  = "participantul mituiește juriul"

$H$  = "participantul este cinstit (nu mituiește juriul)"

$S_1$  = "participantului i se permite să rămână în joc după primul episod"

$S_2$  = "participantului i se permite să rămână în joc după al 2-lea episod"

$L_1$  = "participantul este eliminat în primul episod"

$L_2$  = "participantul este eliminat în al 2-lea episod"

a) Din legea probabilității totale,

$$P(S_1) = P(S_1|B)P(B) + P(S_1|H)P(H) = 1 \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{2}.$$

De aceea, din regula lui Bayes,  $P(B|S_1) = P(S_1|B) \frac{P(B)}{P(S_1)} = 1 \cdot \frac{1/4}{1/2} = \frac{1}{2}$ .

b) Folosind arborele avem

$$P(S_2) = \frac{1}{4} \cdot 1 \cdot 1 + \frac{3}{4} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{3}.$$

c) Vrem să calculăm  $P(L_2|S_1) = \frac{P(L_2 \cap S_1)}{P(S_1)}$ .

a)  $\implies P(S_1) = 1/2$ . Pentru numărător avem (vezi arborele)

$$P(L_2 \cap S_1) = P(L_2 \cap S_1|B)P(B) + P(L_2 \cap S_1|H)P(H) = 0 \cdot \frac{1}{4} + \frac{2}{9} \cdot \frac{3}{4} = \frac{1}{6}.$$

De aceea  $P(L_2|S_1) = \frac{1/6}{1/2} = \frac{1}{3}$ .

### 1.3 Independență

3)  $E$  = "număr par" =  $\{2, 4, \dots, 20\}$ .

$L$  = "număr  $\leq 10$ " =  $\{1, 2, \dots, 10\}$ .

$B$  = "număr prim" =  $\{2, 3, 5, 7, 11, 13, 17, 19\}$ .

a)  $P(E) = 10/20 = 1/2$ ,  $P(E|L) = 5/10 = 1/2 \implies P(E) = P(E|L) \implies E$  și  $L$  sunt independente.

b)  $P(E) = 10/20 = 1/2$ ,  $P(E|B) = 1/8 \implies P(E) \neq P(E|B) \implies E$  și  $B$  nu sunt independente.

## 1.4 Media și dispersia

4) a)  $E(X) = (-1) \cdot \frac{1}{8} + 0 \cdot \frac{1}{4} + 1 \cdot \frac{5}{8} = -\frac{1}{8} + \frac{5}{8} = \frac{1}{2}$ .  
 b)

valorile lui $X$	-1	0	1
probabilitățile	1/8	1/4	5/8
$X^2$	1	0	1

$$Y \sim \begin{pmatrix} 0 & 1 \\ 1/4 & 3/4 \end{pmatrix} \implies E(Y) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{3}{4} = \frac{3}{4}.$$

c) Formula schimbării de variabile spune să folosim ultima linie a tabelului extins de la b):  $E(X^2) = 1 \cdot \frac{1}{8} + 0 \cdot \frac{1}{4} + 1 \cdot \frac{5}{8} = \frac{3}{4}$  (la fel ca la b)).

d)  $Var(X) = E(X^2) - E(X)^2 = \frac{3}{4} - (\frac{1}{2})^2 = \frac{1}{2}$ .

5) Facem un tabel:

$X$	0	1
probabilitățile	$1-p$	$p$
$X^2$	0	1

Din tabel,  $E(X) = 0 \cdot (1-p) + 1 \cdot p = p$ .

Deoarece  $X$  și  $X^2$  au același tabel,  $E(X^2) = E(X) = p$ .

De aceea,  $Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$ .

6) Fie  $X$  numărul de persoane care își iau înapoi propria pălărie.

Urmând indicația: fie  $X_j = 1$  dacă persoana  $j$  își ia înapoi pălăria proprie și  $X_j = 0$  dacă nu.

Avem  $X = \sum_{j=1}^{100} X_j$ , deci  $E(X) = \sum_{j=1}^{100} E(X_j)$ .

Deoarece persoana  $j$  poate lua cu aceeași probabilitate orice pălărie, avem  $P(X_j = 1) = 1/100$ .

Astfel,  $X_j \sim \text{Bernoulli}(1/100) \implies E(X_j) = 1/100 \implies E(X) = 1$ .

## 1.5 Funcții masă de probabilitate, funcții densitate de probabilitate și funcții de repartiție

7) a) Cdf a lui  $X$  este

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = (-e^{-\lambda t})|_0^x = 1 - e^{-\lambda x}.$$

b) Pentru  $y \geq 0$  avem

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = 1 - e^{-\lambda \sqrt{y}}.$$

$$f_Y(y) = F'_Y(y) = \frac{\lambda}{2\sqrt{y}} e^{-\lambda \sqrt{y}}.$$

8) Fie  $T$  numărul de apariții ale lui 6 în 100 de aruncări.

Stim că  $T \sim \text{Binomial}(100, 1/6)$ .

Deoarece primim 3 dolari de fiecare dată când dăm 6, avem  $X = 3T$ . Deci, putem scrie

$$P(X = 3k) = C_{100}^k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{100-k}, \text{ pentru } k = 0, 1, 2, \dots, 100.$$

Sau putem scrie

$$P(X = x) = C_{100}^{x/3} \left(\frac{1}{6}\right)^{x/3} \left(\frac{5}{6}\right)^{100-x/3}, \text{ pentru } x = 0, 3, 6, \dots, 300.$$

b)  $E(X) = E(3T) = 3E(T) = 3 \cdot 100 \cdot \frac{1}{6} = 50$ .

$$Var(X) = Var(3T) = 9Var(T) = 9 \cdot 100 \cdot \frac{1}{6} \cdot \frac{5}{6} = 125.$$

c) i) Fie  $T_1$  numărul total de apariții ale lui 6 în primele 25 de aruncări. Deci,  $X_1 = 3T_1$  și  $Y = 12T_1$ .

$T_1 \sim \text{Binomial}(25, 1/6)$ , deci

$$E(Y) = E(12T_1) = 12E(T_1) = 12 \cdot 25 \cdot \frac{1}{6} = 50$$

și

$$Var(Y) = Var(12T_1) = 144Var(T_1) = 144 \cdot 25 \cdot \frac{1}{6} \cdot \frac{5}{6} = 500.$$

ii) Mediile sunt aceleași din liniaritate deoarece  $X$  și  $Y$  sunt ambele  $300 \times 1$  variabilă aleatoare Bernoulli(1/6).

Pentru dispersie,  $Var(X) = 4Var(X_1)$  deoarece  $X$  este suma a 4 variabile *independente* toate identice cu  $X_1$ . Dar  $Var(Y) = Var(4X_1) = 16Var(X_1)$ . Deci  $Var(Y) = 4Var(X)$ . Aceasta are sens intuitiv, deoarece  $X$  este construită din mai multe încercări independente ca  $X_1$ .

## 1.6 Probabilitate comună, covarianță, corelație

9) (**Puzzle aritmetic**) a) Suma probabilităților marginale este 1, deci cele 2 probabilități marginale lipsă sunt  $P(X = 3) = 1/3$ ,  $P(Y = 3) = 1/2$ . Suma probabilităților de pe fiecare linie sau coloană este probabilitatea marginală corespunzătoare. De exemplu,  $1/6 + 0 + P(X = 1, Y = 3) = 1/3$ , deci  $P(X = 1, Y = 3) = 1/6$ . Iată tabelul completat:

$X \setminus Y$	1	2	3	
1	1/6	0	1/6	1/3
2	0	1/4	1/12	1/3
3	0	1/12	1/4	1/3
	1/6	1/3	1/2	1

b) Nu, nu sunt independente, deoarece, de exemplu,  
 $P(X = 2, Y = 1) = 0 \neq P(X = 2) \cdot P(Y = 1)$ .

#### 10) Covariantă și independentă

$X \setminus Y$	-2	-1	0	1	2	
0	0	0	1/5	0	0	1/5
1	0	1/5	0	1/5	0	2/5
4	1/5	0	0	0	1/5	2/5
	1/5	1/5	1/5	1/5	1/5	1

Fiecare coloană, exceptând marginea, are doar o probabilitate nenulă. De exemplu, când  $X = -2$ , atunci  $Y = 4$ , deci în coloana  $X = -2$  doar  $P(X = -2, Y = 4) \neq 0$ .

b) Folosind repartițiile marginale:

$$E(X) = \frac{1}{5}(-2 - 1 + 0 + 1 + 2) = 0.$$

$$E(Y) = 0 \cdot \frac{1}{5} + 1 \cdot \frac{2}{5} + 4 \cdot \frac{2}{5} = 2.$$

c) Arătăm că probabilitatea intersecției nu este produsul probabilităților:

$$P(X = -2, Y = 0) = 0 \neq 1/25 = P(X = -2)P(Y = 0).$$

Deoarece acestea nu sunt egale,  $X$  și  $Y$  nu sunt independente. (Este evident că  $X^2$  nu este independent de  $X$ .)

d) Folosind tabelul din a) și mediile din b), avem:

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) \\ &= \frac{1}{5} \cdot (-2) \cdot 4 + \frac{1}{5} \cdot (-1) \cdot 1 + \frac{1}{5} \cdot 0 \cdot 0 + \frac{1}{5} \cdot 1 \cdot 1 + \frac{1}{5} \cdot 2 \cdot 4 - 0 \cdot 2 \\ &= 0. \end{aligned}$$

#### 11) Repartiții comune continue

$$\begin{aligned} a) F(a, b) &= P(X \leq a, Y \leq b) = \int_0^a \int_0^b (x + y) dy dx = \int_0^a \left( xy + \frac{y^2}{2} \right) \Big|_{y=0}^{y=b} dx \\ &= \int_0^a \left( xb + \frac{b^2}{2} \right) dx = \left( \frac{x^2}{2} b + \frac{b^2}{2} x \right) \Big|_0^a = \frac{a^2 b + ab^2}{2}. \end{aligned}$$

Deci  $F(x, y) = \frac{x^2y+xy^2}{2}$  pe  $[0, 1] \times [0, 1]$  și  $F(1, 1) = 1$ .

b)  $f_X(x) = \int_0^1 f(x, y) dy = \int_0^1 (x + y) dy = \left(xy + \frac{y^2}{2}\right) \Big|_{y=0}^{y=1} = x + \frac{1}{2}$ .

Din simetrie,  $f_Y(y) = y + \frac{1}{2}$ .

c) Pentru a vedea dacă  $X$  și  $Y$  sunt independente, verificăm dacă densitatea comună este produsul densităților marginale.

$$f(x, y) = x + y, \quad f_X(x) \cdot f_Y(y) = \left(x + \frac{1}{2}\right) \left(y + \frac{1}{2}\right).$$

Deoarece acestea nu sunt egale,  $X$  și  $Y$  nu sunt independente.

d)  $E(X) = \int_0^1 \int_0^1 x(x + y) dy dx = \int_0^1 \left[ \left(x^2y + x\frac{y^2}{2}\right) \Big|_{y=0}^{y=1} \right] dx = \int_0^1 \left(x^2 + \frac{x}{2}\right) dx$   
 $= \left(\frac{x^3}{3} + \frac{x^2}{4}\right) \Big|_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}.$

(Sau, folosind b),  $E(X) = \int_0^1 x f_X(x) dx = \int_0^1 x \left(x + \frac{1}{2}\right) dx = \frac{7}{12}.$ )

Din simetrie,  $E(Y) = \frac{7}{12}$ .

$$\begin{aligned} E(X^2 + Y^2) &= \int_0^1 \int_0^1 (x^2 + y^2)(x+y) dy dx = \int_0^1 \int_0^1 (x^3 + x^2y + xy^2 + y^3) dy dx \\ &= \int_0^1 \left[ \left(x^3y + \frac{x^2y^2}{2} + \frac{xy^3}{3} + \frac{y^4}{4}\right) \Big|_{y=0}^{y=1} \right] dx = \int_0^1 \left(x^3 + \frac{x^2}{2} + \frac{x}{3} + \frac{1}{4}\right) dx \\ &= \left(\frac{x^4}{4} + \frac{x^3}{6} + \frac{x^2}{6} + \frac{x}{4}\right) \Big|_0^1 = \frac{1}{4} + \frac{1}{6} + \frac{1}{6} + \frac{1}{4} = \frac{5}{6}. \end{aligned}$$

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^1 xy(x+y) dy dx = \int_0^1 \int_0^1 (x^2y + xy^2) dy dx = \int_0^1 \left[ \left(\frac{x^2y^2}{2} + \frac{xy^3}{3}\right) \Big|_{y=0}^{y=1} \right] dx \\ &= \int_0^1 \left(\frac{x^2}{2} + \frac{x}{3}\right) dx = \left(\frac{x^3}{6} + \frac{x^2}{6}\right) \Big|_0^1 = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}. \end{aligned}$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \frac{7}{12} \cdot \frac{7}{12} = -\frac{1}{144}.$$

## 1.7 Legea numerelor mari, teorema limită centrală

12) Standardizăm:

$$\begin{aligned} P\left(\sum_i X_i < 30\right) &= P\left(\frac{\frac{1}{n} \sum X_i - \mu}{\sigma/\sqrt{n}} < \frac{30/n - \mu}{\sigma/\sqrt{n}}\right) \\ &\approx P\left(Z < \frac{30/100 - 1/5}{1/30}\right) \text{ (din teorema limită centrală)} \\ &= P(Z < 3) \\ &\approx 0.9987 \text{ (din tabelul probabilităților normale sau din R cu pnorm(3))} \end{aligned}$$

13) (Mai mult cu teorema limită centrală)

Fie  $X_j$  IQ-ul persoanei  $j$ . Avem  $E(X_j) = 100$  și  $\sigma_{X_j} = 15$ .

Fie  $\bar{X}$  media IQ-urilor a 100 de persoane selectate aleator. Atunci avem

$$E(\bar{X}) = 100 \text{ și } \sigma_{\bar{X}} = 15/\sqrt{100} = 1.5.$$

Problema cere  $P(\bar{X} > 115)$ . Standardizând, obținem  
 $P(\bar{X} > 115) \approx P(Z > 10) = 0$  (din R, cu `1-pnorm(10)`).

## 2 Introducere în statistică

### 2.1 Scopurile învățării

1. Să știe cele 3 ”faze” care se suprapun ale practicii statistice.
2. Să știe ce se înțelege prin termenul *statistică*.

### 2.2 Introducere în statistică

Statistica se ocupă cu date. Vorbind în general, scopul statisticii este de a face inferențe (deducții) bazate pe date. Putem împărți acest proces în 3 faze: colectarea datelor, descrierea datelor și analiza datelor. Aceasta se potrivește cu paradigma metodei științifice. Facem ipoteze despre ce este adevărat, colectăm date în experimente, descriem rezultatele și apoi inferăm din rezultate **puterea dovezilor** care privesc ipotezele noastre.

#### 2.2.1 Proiectarea experimentelor

Proiectarea unui experiment este crucială pentru a asigura utilitatea datelor colectate. Un experiment proiectat prost va produce date de calitate slabă, din care poate fi imposibil să tragem concluzii utile, valide. ”A consulta un statistician după ce un experiment este terminat este adesea doar a-i cere să conducă o examinare post-mortem. El poate spune probabil de ce a murit experimentul.” (R. A. Fisher, unul dintre fondatorii statisticii moderne).

#### 2.2.2 Statistică descriptivă

Datele neprelucrate iau adesea forma unei masive liste, matrice sau baze de date de etichete și numere. Pentru a da sens datelor, putem calcula **statistici de rezumat** ca media, mediana și domeniile intercuartile. Putem de asemenea să vizualizăm datele folosind dispozitive grafice ca histogramele, reprezentări ale împrăștierii și cdf empirică. Aceste metode sunt utile atât pentru comunicarea cât și pentru explorarea datelor pentru a obține o perspectivă în structura lor, de exemplu pentru a recunoaște dacă urmează o repartiție de probabilitate cunoscută.

### 2.2.3 Statistică deductivă (inferențială)

În cele din urmă vrem să tragem concluzii despre lume. Adesea aceasta ia forma specificării unui model statistic pentru procesul aleator din care provin datele. De exemplu, presupunem că datele iau forma unei serii de măsurători a căror eroare credem că urmează o repartiție normală. (Observăm că aceasta este totdeauna o aproximare deoarece știm că eroarea trebuie să aibă o margine în timp ce repartiția normală are domeniul  $\mathbb{R}$ .) Putem apoi folosi datele pentru a produce dovezi pentru sau împotriva acestei ipoteze. De exemplu, presupunând că lungimea gestației are o repartiție  $N(\mu, \sigma^2)$ , folosim datele lungimilor gestației să zicem a 500 de sarcini pentru a trage concluzii despre valorile parametrilor  $\mu$  și  $\sigma$ . Similar, putem modela rezultatul alegerii a 2 candidați printr-o repartiție Bernoulli( $p$ ) și utiliza datele sondajului pentru a trage concluzii despre valoarea lui  $p$ .

Rareori putem face afirmații definitive despre astfel de parametri deoarece chiar datele vin dintr-un proces aleator (cum ar fi pe cine să sondezi). Mai degrabă, dovezile noastre statistice vor implica totdeauna afirmații probabiliste. Din nefericire, media și publicul larg înceleg greșit sensul probabilist al afirmațiilor statistice. De fapt, cercetătorii însăși fac adesea aceeași eroare.

**Exemplul 1.** Pentru a studia eficacitatea unui nou tratament pentru cancer, pacienții sunt recrutați și apoi împărțiți într-un grup experimental și un grup de control. Grupului experimental i se dă noul tratament și grupul de control primește standardul curent de îngrijire. Datele colectate de la pacienți pot include informație demografică, istorie medicală, stadiul inițial al cancerului, progresul cancerului în timp, costul tratamentului și efectul tratamentului asupra mărimii tumorii, ratele de remisie, longevitatea și calitatea vieții. Datele vor fi folosite pentru a face deducții despre eficacitatea nouui tratament comparativ cu standardul curent de îngrijire.

Observează că acest studiu va trece prin toate cele 3 faze descrise mai sus. Proiectarea experimentului trebuie să fie specifică mărimii studiului, cine va fi eligibil să se alăture, cum vor fi alese grupurile experimental și de control, cum vor fi administrate tratamentele, dacă subiecții sau doctorii știu sau nu cine ce tratament primește și precum ce date se colectează, printre alte lucruri. Odată ce datele sunt colectate, trebuie descrise și analizate pentru a determina dacă ele susțin ipoteza că noul tratament este mai (sau mai puțin) eficient decât cel(e) curent(e) și cu cât. Aceste concluzii statistice vor fi formulate ca afirmații precise implicând probabilități.

După cum s-a observat mai sus, interpretarea greșită a înțelesului afirmațiilor statistice este o sursă obișnuită de eroare care a dus la tragedie de mai multe ori.

**Exemplul 2.** În 1999 în Marea Britanie, Sally Clark a fost condamnată

pentru uciderea celor 2 copii ai ei după ce fiecare copil a murit la câteva săptămâni după naștere (primul în 1996, al 2-lea în 1998). Condamnarea ei s-a bazat foarte mult pe o utilizare defectuoasă a statisticii pentru a exclude sindromul morții subite infantile. Cu toate că a fost achitată în 2003, a suferit în timpul și după închisoare de probleme psihiatrice foarte serioase și a murit în 2007 de intoxicație cu alcool. Vezi [http://en.wikipedia.org/wiki/Sally\\_Clark](http://en.wikipedia.org/wiki/Sally_Clark).

O discuție despre cazul Sally Clark și alte exemple de intuiție statistică proastă este la adresa <http://www.youtube.com/watch?v=kLmzxmRcUTo>.

#### 2.2.4 Ce este o statistică?

Dăm o definiție simplă al cărei înțeles este cel mai bine elucidat de exemple.

**Definiție.** O **statistică** este orice se poate calcula din datele colectate.

**Exemplul 3.** Considerăm datele a 1000 de aruncări ale unui zar. Sunt statistici: media celor 1000 de aruncări; numărul de 6-uri; suma pătratelor aruncărilor minus numărul aruncărilor pare. Este greu să ne imaginăm cum am putea utiliza ultimul exemplu, dar el este o statistică. Pe de altă parte, probabilitatea de a da 6 nu este o statistică, indiferent dacă zarul este cu adevărat corect sau nu. Mai degrabă, această probabilitate este o proprietate a zarului (și felului în care îl aruncăm) care poate fi **estimată** folosind datele. O astfel de estimare este dată de statistică ”numărul de 6-uri/1000”.

**Exemplul 4.** Presupunem că tratăm un grup de pacienți cu cancer cu o nouă procedură și colectăm datele despre cât de mult supraviețuiesc după tratament. Din date putem calcula media timpului de supraviețuire al pacienților din grup. Putem folosi această statistică ca o estimare a timpului de supraviețuire pentru viitorii pacienți cu cancer care urmeză noua procedură.

**Exemplul 5.** Presupunem că întrebăm 1000 de rezidenți dacă sprijină sau nu legalizarea marijuanei în Massachusetts. Proportia din cei 1000 care sprijină propunerea este o statistică. Proportia din toți rezidenții din Massachusetts care sprijină propunerea *nu* este o statistică deoarece nu i-am întrebat pe fiecare (observați cuvântul ”colectate” din definiție). Mai degrabă, sperăm să tragem o concluzie statistică despre proporția la nivel de stat, bazată pe datele din eșantionul nostru aleator.

Folosim 2 tipuri generale de statistici:

1. **Statistică punctuală:** o singură valoare calculată din date, ca media de selecție  $\bar{x}_n$  sau deviația standard de selecție  $s_n$ .
2. **Statistică de interval:** un interval  $[a, b]$  calculat din date. Aceasta este efectiv doar o pereche de statistici punctuale și e adesea prezentată în forma  $\bar{x} \pm s$ .

## 2.3 Importanța teoremei lui Bayes

Teorema lui Bayes este foarte importantă pentru statistica inferențială. Remintim că teorema lui Bayes ne permite să ”inversăm” probabilitățile condiționate. Adică, dacă  $H$  și  $D$  sunt evenimente, atunci teorema lui Bayes spune

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

În experimentele științifice pornim cu o ipoteză și colectăm date pentru a testa ipoteza. Adesea  $H :=$  ”ipoteza noastră e adevărată” și  $D$  sunt datele colectate. Cu aceste cuvinte, teorema lui Bayes spune

$$P(\text{ipoteza este adevărată} | \text{date}) = \frac{P(\text{date} | \text{ipoteza este adevărată})P(\text{ipoteza este adevărată})}{P(\text{date})}$$

Membrul stâng este probabilitatea că ipoteza noastră este adevărată date fiind datele colectate. Aceasta este precis ce am vrea să știm. Când toate probabilitățile din dreapta sunt cunoscute exact, putem calcula probabilitatea din stânga exact. Din păcate, în practică rareori știm valorile exacte ale tuturor probabilităților din dreapta. Statisticienii au dezvoltat un număr de moduri de a face față acestei lipse de cunoștințe și de a face totuși inferențe utile.

### Exemplul 6. Test pentru boală

Presupunem că un test pentru o boală are 1% rată fals pozitivă și 1% rată fals negativă. Presupunem de asemenea că rata bolii în populație este 0.002. În sfârșit, presupunem că o persoană selectată aleator face testul și ieșe pozitiv. În limbaj de ipoteză și date avem:

Ipoteza:  $H =$  ”persoana are boala”.

Date:  $D =$  ”testul a fost pozitiv”.

Ce vrem să știm:  $P(H|D) = P(\text{persoana are boala} | \text{un test pozitiv})$ .

În acest exemplu toate probabilitățile din dreapta sunt cunoscute, deci putem folosi teorema lui Bayes pentru a calcula ce vrem să știm.

$$\begin{aligned} P(\text{ipoteză} | \text{date}) &= P(\text{persoana are boala} | \text{un test pozitiv}) \\ &= P(H|D) \\ &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{0.99 \cdot 0.002}{0.99 \cdot 0.002 + 0.01 \cdot 0.998} \\ &= 0.1655518. \end{aligned}$$

Înaintea testului am fi spus că probabilitatea ca persoana să fi avut boala era 0.002. După test vedem că probabilitatea este 0.1655518. Adică, testul pozitiv dă unele dovezi că persoana are boala.

# Curs 10

Cristian Niculescu

## 1 Estimări de verosimilitate maximă

### 1.1 Scopurile învățării

1. Să poată să definească funcția de verosimilitate pentru date dintr-un model parametric.
2. Să poată să calculeze estimarea de verosimilitate maximă a parametrului necunoscut (parametrilor necunoscuți).

### 1.2 Introducere

Presupunem că știm că avem date constând din valori  $x_1, \dots, x_n$  dintr-o repartiție exponențială. Întrebarea rămâne: care repartiție exponențială?!

Ne-am referit uneori la repartiția exponențială sau la repartiția binomială sau la repartiția normală. De fapt, repartiția exponențială  $\exp(\lambda)$  nu este o singură repartiție, ci mai degrabă o familie de repartiții cu un parametru. Fiecare valoare  $\lambda$  definește o repartiție diferită din familie, cu pdf  $f_\lambda(x) = \lambda e^{-\lambda x}$  pe  $[0, \infty)$ . Similar, o repartiție binomială  $\text{bin}(n, p)$  este determinată de cei 2 parametri  $n$  și  $p$ , iar o repartiție normală  $N(\mu, \sigma^2)$  este determinată de cei 2 parametri  $\mu$  și  $\sigma^2$  (sau echivalent,  $\mu$  și  $\sigma$ ). Familiile parametrizate de repartiții sunt adesea numite [repartiții parametrice](#) sau [modele parametrice](#).

Adesea suntem puși în fața situației de a avea date aleatoare despre care știm (sau credem) că provin dintr-un model parametric, ai cărui parametri nu-i știm. De exemplu, într-o alegere între 2 candidați, datele sondajului sunt dintr-o repartiție Bernoulli( $p$ ) cu parametru necunoscut  $p$ . În acest caz ne-ar plăcea să folosim datele pentru a estima valoarea parametrului  $p$ , deoarece aceasta prezice rezultatul alegerii. Similar, presupunând că lungimea gestației are o repartiție normală, ne-ar plăcea să folosim datele lungimilor gestaționale dintr-o selecție aleatoare de sarcini pentru a trage concluzii despre valorile parametrilor  $\mu$  și  $\sigma^2$ .

Scopul nostru până acum a fost să calculăm [probabilitatea datelor](#) provenind

dintron model parametric cu parametri cunoscuți. Deduția statistică întoarce asta pe dos: estimăm probabilitatea parametrilor dat fiind un model parametric și date observate din el. Valorile parametrului sunt văzute în mod natural ca ipoteze, deci de fapt estimăm probabilitatea diverselor ipoteze știind datele.

### 1.3 Estimări de verosimilitate maximă

Sunt mai multe metode de a estima parametru necunoscuți din date. Întâi considerăm estimarea de verosimilitate maximă (MLE, prescurtare de la maximum likelihood estimate), care răspunde la întrebarea:

Pentru ce valoare a parametrului au datele observate cea mai mare probabilitate?

MLE este un exemplu de estimare punctuală deoarece dă o singură valoare pentru parametrul necunoscut. 2 avantaje ale MLE sunt că este adesea ușor de calculat și este de acord cu intuiția noastră în exemple simple.

**Exemplul 1.** Aruncăm o monedă de 100 de ori. Dat fiind că au fost 55 de aversuri, aflați estimarea de verosimilitate maximă pentru probabilitatea  $p$  a aversului la o singură aruncare.

Înainte de a rezolva problema, stabilim niște notații și termeni.

Putem gândi calcularea numărului de aversuri în 100 de aruncări ca un experiment. Pentru o valoare dată a lui  $p$ , probabilitatea de a obține 55 de aversuri în acest experiment este probabilitatea binomială

$$P(55 \text{ aversuri}) = C_{100}^{55} p^{55} (1-p)^{45}.$$

Probabilitatea de a obține 55 de aversuri depinde de valoarea lui  $p$ , deci hai să-l includem pe  $p$  folosind notația probabilității condiționate:

$$P(55 \text{ aversuri} | p) = C_{100}^{55} p^{55} (1-p)^{45}.$$

Citim  $P(55 \text{ aversuri} | p)$

”probabilitatea a 55 de aversuri dat fiind  $p$ ,”

sau mai precis

”probabilitatea a 55 de aversuri dat fiind că probabilitatea unui avers la o singură aruncare este  $p$ .“

Iată câțiva termeni standard în statistică:

**Experiment:** Aruncăm moneda de 100 de ori și calculăm numărul de aversuri.

**Date:** Datele sunt rezultatul experimentului. În acest caz sunt ”55 de aversuri“.

**Parametru (Parametri) de interes:** Suntem interesați de valoarea parametrului necunoscut  $p$ .

**Verosimilitatea sau funcția de verosimilitate:** aceasta este  $P(\text{date}|p)$ . Observăm că este o funcție de date și parametrul  $p$ . În acest caz verosimilitatea este

$$P(55 \text{ aversuri } | p) = C_{100}^{55} p^{55} (1-p)^{45}.$$

Observații: 1. Verosimilitatea  $P(\text{date}|p)$  se schimbă dacă dacă parametrul de interes  $p$  se schimbă.

2. Atenție la definiție! O sursă tipică de confuzie este a confunda verosimilitatea  $P(\text{date}|p)$  cu  $P(p| \text{date})$ . Știm că  $P(\text{date}|p)$  și  $P(p| \text{date})$  sunt de obicei diferite.

**Definiție:** Știind datele, **estimarea de verosimilitate maximă (MLE)** pentru parametrul  $p$  este valoarea care maximizează verosimilitatea  $P(\text{date}|p)$ . Adică, MLE este valoarea lui  $p$  pentru care datele sunt cele mai probabile.

**Răspuns:** Am văzut că verosimilitatea este

$$P(55 \text{ aversuri } | p) = C_{100}^{55} p^{55} (1-p)^{45}.$$

Folosim notația  $\hat{p}$  pentru MLE. Folosim analiza matematică pentru a o afla, luând derivata funcției de verosimilitate și egalând-o cu 0.

$$\frac{d}{dp} P(\text{date } | p) = C_{100}^{55} (55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0.$$

Rezolvând ecuația obținem

$$\begin{aligned} 55p^{54}(1-p)^{45} &= 45p^{55}(1-p)^{44} \\ 55(1-p) &= 45p \quad (\text{deoarece } p \in (0, 1)) \\ 55 &= 100p \\ \text{MLE este } \hat{p} &= 0.55. \end{aligned}$$

Observații: 1. MLE pentru  $p$  s-a dovedit a fi exact fracția de aversuri pe care am văzut-o în datele noastre.

2. MLE este calculată din date. Adică, este o statistică.

3. Oficial ar trebui să verificăm că punctul critic este într-adevăr un maxim. Putem face aceasta cu testul derivatei a 2-a.

### 1.3.1 Log verosimilitate

Adesea este mai ușor să lucrăm cu logaritmul natural al funcției de verosimilitate. Pentru scurtime acesta este numit simplu **log verosimilitate**. Deoarece  $\ln(x)$  este o funcție strict crescătoare, maximele verosimilității și log verosimilității

coincid.

**Exemplul 2.** Refacem exemplul precedent folosind log verosimilitatea.

**Răspuns:** Aveam verosimilitatea  $P(55 \text{ aversuri } | p) = C_{100}^{55} p^{55} (1-p)^{45}$ . De aceea log verosimilitatea este

$$\ln(P(55 \text{ aversuri } | p)) = \ln(C_{100}^{55}) + 55 \ln(p) + 45 \ln(1-p).$$

Maximizarea verosimilității este același lucru cu maximizarea log verosimilității.

Verificăm că analiza matematică ne dă același răspuns ca mai sus:

$$\begin{aligned} \frac{d}{dp}(\log \text{verosimilitate}) &= \frac{d}{dp}[\ln(C_{100}^{55}) + 55 \ln(p) + 45 \ln(1-p)] \\ &= \frac{55}{p} - \frac{45}{1-p} = 0 \\ \implies 55(1-p) &= 45p \\ \implies \hat{p} &= 0.55. \end{aligned}$$

### 1.3.2 Verosimilitate maximă pentru repartiții continue

Pentru repartiții continue, folosim funcția densitate de probabilitate pentru a defini verosimilitatea.

**Exemplul 3. Becuri**

Presupunem că durata de funcționare a unui anumit tip de becuri este modelată de o repartition exponențială cu parametru necunoscut  $\lambda$ . Testăm 5 becuri și aflăm că au duratele de funcționare de 2, 3, 1, 3 respectiv 4 ani. Care este MLE pentru  $\lambda$ ?

**Răspuns:** Avem nevoie să fim atenți cu notația noastră. Cu 5 valori diferite cel mai bine este să folosim indici. Fie  $X_i$  durata de funcționare al celui de-al  $i$ -lea bec și fie  $x_i$  valoarea pe care o ia  $X_i$ . Atunci fiecare  $X_i$  are pdf  $f_{X_i}(x_i) = \lambda e^{-\lambda x_i}$ . Presupunem că duratele de funcționare ale becurilor sunt independente, deci pdf comună este produsul densităților individuale:

$$f(x_1, x_2, x_3, x_4, x_5 | \lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5}) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}.$$

Observăm că scriem aceasta ca o densitate condiționată, deoarece depinde de  $\lambda$ . Văzând datele ca fixate și  $\lambda$  ca variabilă, această densitate este funcția de verosimilitate. Datele noastre au avut valorile

$$x_1 = 2, x_2 = 3, x_3 = 1, x_4 = 3, x_5 = 4.$$

Deci funcțiile de verosimilitate și log verosimilitate cu datele noastre sunt

$$f(2, 3, 1, 3, 4 | \lambda) = \lambda^5 e^{-13\lambda}, \quad \ln(f(2, 3, 1, 3, 4 | \lambda)) = 5 \ln(\lambda) - 13\lambda.$$

În sfârșit, folosim analiza matematică pentru a afla MLE:

$$\frac{d}{d\lambda}(\log \text{verosimilitate}) = \frac{5}{\lambda} - 13 = 0 \implies \hat{\lambda} = \frac{5}{13}.$$

- Observații: 1. În acest exemplu am folosit o literă mare pentru o variabilă aleatoare și litera mică corespunzătoare pentru valoarea pe care o ia.  
 2. MLE pentru  $\lambda$  s-a dovedit a fi inversa mediei de selecție  $\bar{x}$ , deci  $X \sim \exp(\hat{\lambda})$  satisfacă  $E(X) = \bar{x}$ .

#### **Exemplul 4. Repartiții normale**

Presupunem că datele  $x_1, x_2, \dots, x_n$  sunt dintr-o repartiție  $N(\mu, \sigma^2)$ , unde  $\mu$  și  $\sigma$  sunt necunoscute. Aflați estimarea de verosimilitate maximă pentru perechea  $(\mu, \sigma^2)$ .

**Răspuns:** Formulăm aceasta în termeni de variabile aleatoare și densități. Fie  $X_1, \dots, X_n$  variabile aleatoare i.i.d.  $N(\mu, \sigma^2)$  și fie  $x_i$  valoarea pe care o ia  $X_i$ . Densitatea pentru fiecare  $X_i$  este

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Deoarece  $X_i$  sunt independente pdf comună a lor este produsul pdf-urilor individuale:

$$f(x_1, \dots, x_n | \mu, \sigma) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Pentru datele fixate  $x_1, \dots, x_n$ , verosimilitatea și log verosimilitatea sunt

$$f(x_1, \dots, x_n | \mu, \sigma) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}},$$

$$\ln(f(x_1, \dots, x_n | \mu, \sigma)) = -n \ln(\sqrt{2\pi}) - n \ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Deoarece  $\ln(f(x_1, \dots, x_n | \mu, \sigma))$  este o funcție de 2 variabile  $\mu, \sigma$ , folosim derivatele parțiale pentru a afla MLE. Valoarea ușor de aflat este  $\hat{\mu}$ :

$$\frac{\partial \ln f(x_1, \dots, x_n | \mu, \sigma)}{\partial \mu} = \sum_{i=1}^n \frac{(x_i - \mu)}{\sigma^2} = 0 \implies \sum_{i=1}^n x_i = n\mu \implies \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.$$

Pentru a afla  $\hat{\sigma}$  derivăm și rezolvăm în raport cu  $\sigma$ :

$$\frac{\partial \ln f(x_1, \dots, x_n | \mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} = 0 \implies \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}.$$

Stim deja  $\hat{\mu} = \bar{x}$ , deci folosim aceasta ca valoare pentru  $\mu$  în formula pentru  $\hat{\sigma}$ . Obținem estimările de verosimilitate maximă

$$\begin{aligned}\hat{\mu} &= \bar{x} &= \text{media datelor} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 &= \text{dispersia datelor.}\end{aligned}$$

### Exemplul 5. Repartiții uniforme

Presupunem că datele noastre  $x_1, \dots, x_n$  sunt independente dintr-o repartiție uniformă  $U(a, b)$ . Aflați MLE pentru  $a$  și  $b$ .

**Răspuns:** Acest exemplu este diferit de cele precedente prin aceea că nu folosim analiza matematică pentru a afla MLE. Densitatea pentru  $U(a, b)$  este  $\frac{1}{b-a}$  pe  $[a, b]$ . De aceea funcția noastră de verosimilitate este

$$f(x_1, \dots, x_n | a, b) = \begin{cases} \left(\frac{1}{b-a}\right)^n, & \text{dacă toți } x_i \text{ sunt în intervalul } [a, b] \\ 0, & \text{altfel.} \end{cases}$$

Aceasta este maximizată făcând  $b - a$  cât mai mic posibil. Singura restricție este aceea că intervalul  $[a, b]$  trebuie să conțină toate datele. Astfel, MLE pentru perechea  $(a, b)$  este

$$\hat{a} = \min(x_1, \dots, x_n), \quad \hat{b} = \max(x_1, \dots, x_n).$$

### Exemplul 6. Metoda captură/recaptură

Metoda captură/recaptură este un mod de a estima mărimea unei populații în sălbăticie. Metoda presupune că fiecare animal din populație este egal posibil de a fi capturat de o cursă.

Presupunem că 10 animale sunt capturate, etichetate și eliberate. Câteva luni mai târziu, 20 de animale sunt capturate, examineate și eliberate. 4 dintre aceste 20 sunt găsite etichetate. Estimați mărimea populației sălbaticice folosind MLE pentru probabilitatea că un animal sălbatic este etichetat.

**Răspuns:** Parametrul nostru necunoscut  $n$  este numărul de animale din sălbăticie. Datele noastre sunt că 4 dintre cele 20 de animale capturate ultima dată au fost etichetate (și că sunt 10 animale etichetate). Funcția de verosimilitate este

$$P(\text{date} | n \text{ animale}) = \frac{C_{n-10}^{16} \cdot C_{10}^4}{C_n^{20}}.$$

(Numărătorul este numărul de moduri de a alege 16 animale din cele  $n - 10$  neetichetate ori numărul de moduri de a alege 4 din cele 10 animale etichetate. Numitorul este numărul de moduri de a alege 20 de animale din întreaga populație de  $n$ .) Putem folosi R (cu comanda `dhyper(16, n-10, 10, 20)`

pentru diversi  $n$ ) pentru a calcula că funcția de verosimilitate este maximizată când  $n = 50$ . Aceasta ar trebui să aibă un sens. Acesta spune că cea mai bună estimare a noastră este când fractia din toate animalele care sunt etichetate este  $10/50$ , care este egală cu fractia din animalele capturate ultima oară care sunt etichetate ( $4/20$ ).

**Exemplul 7. Hardy-Weinberg.** Presupunem că o genă particulară apare ca una din 2 alele ( $A$  și  $a$ ), unde alela  $A$  are frecvența  $\theta$  în populație. Adică, o copie aleatoare a genei este  $A$  cu probabilitatea  $\theta$  și  $a$  cu probabilitatea  $1 - \theta$ . Deoarece un genotip diploid este format din 2 gene, probabilitatea fiecărui genotip este dată de:

genotip	AA	Aa	aa
probabilitate	$\theta^2$	$2\theta(1 - \theta)$	$(1 - \theta)^2$

Presupunem că testăm un eșantion aleator de oameni și aflăm că:  $k_1$  sunt  $AA$ ,  $k_2$  sunt  $Aa$  și  $k_3$  sunt  $aa$ . Aflați MLE pentru  $\theta$ .

**Răspuns:** Funcția de verosimilitate este dată de

$$P(k_1, k_2, k_3 | \theta) = C_{k_1+k_2+k_3}^{k_1} \cdot C_{k_2+k_3}^{k_2} \cdot C_{k_3}^{k_3} \cdot \theta^{2k_1} (2\theta(1 - \theta))^{k_2} (1 - \theta)^{2k_3}.$$

Log verosimilitatea este dată de

$$\text{constantă} + 2k_1 \ln(\theta) + k_2 \ln(2\theta(1 - \theta)) + k_3 \ln(1 - \theta)$$

Egalăm derivata cu 0:

$$\frac{2k_1 + k_2}{\theta} - \frac{k_2 + 2k_3}{1 - \theta} = 0$$

Rezolvând ecuația în  $\theta$ , aflăm că MLE este

$$\hat{\theta} = \frac{2k_1 + k_2}{2k_1 + 2k_2 + 2k_3},$$

care este pur și simplu fracția de alele  $A$  dintre toate genele din populația selectată.

## 1.4 De ce folosim densitatea pentru a afla MLE pentru repartiții continue

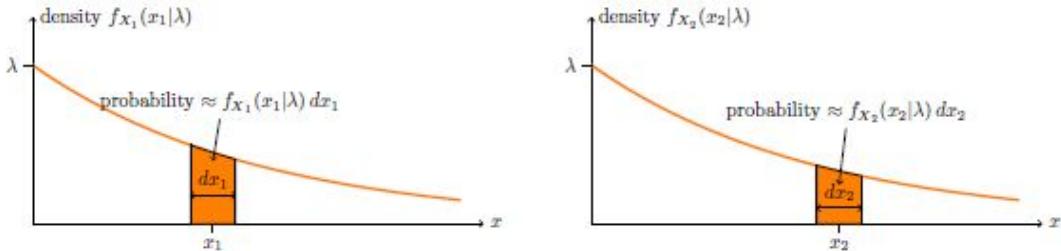
Ideea pentru estimarea de verosimilitate maximă este de a afla valoarea parametrului (parametrilor) pentru care datele au cea mai mare probabilitate.

**Exemplul 8.** Presupunem că avem 2 becuri a căror durată de funcționare

are o repartiție  $\exp(\lambda)$ . Presupunem de asemenea că măsurăm independent duratele lor de funcționare și obținem datele  $x_1 = 2$  ani și  $x_2 = 3$  ani. Aflați valoarea lui  $\lambda$  care maximizează probabilitatea acestor date.

**Răspuns:** Principalul paradox cu care avem de-a face este că pentru o repartiție continuă probabilitatea unei singure valori, să spunem  $x_1 = 2$ , este 0. Rezolvăm acest paradox reamintind că o singură măsurătoare înseamnă de fapt un domeniu de valori, de exemplu aici putem verifica băcul o dată pe zi. Deci data  $x_1 = 2$  ani înseamnă de fapt că  $x_1$  este undeva într-un domeniu de o zi în jurul a 2 ani.

Dacă domeniul este mic, îl numim  $dx_1$ . Probabilitatea că  $X_1$  este în domeniu este aproximativ  $f_{X_1}(x_1|\lambda)dx_1$ . Acest fapt este ilustrat de figura de mai jos. Valoarea  $x_2$  a datelor este tratată exact la fel.



Relația uzuală între densitate și probabilitate pentru domenii mici.

Deoarece datele sunt colectate independent, probabilitatea comună este produsul probabilităților individuale, adică

$$P(X_1 \text{ în domeniu}, X_2 \text{ în domeniu}|\lambda) \approx f_{X_1}(x_1|\lambda)dx_1 \cdot f_{X_2}(x_2|\lambda)dx_2.$$

Folosind valorile  $x_1 = 2$  și  $x_2 = 3$  și formula pentru o pdf exponențială avem

$$P(X_1 \text{ în domeniu}, X_2 \text{ în domeniu}|\lambda) \approx \lambda e^{-2\lambda}dx_1 \cdot \lambda e^{-3\lambda}dx_2 = \lambda^2 e^{-5\lambda}dx_1dx_2.$$

Acum, deoarece avem o probabilitate autentică, putem căuta valoarea pentru  $\lambda$  care o maximizează. Privind formula de mai sus vedem că factorul  $dx_1dx_2$  nu joacă niciun rol în aflarea maximului. Deci pentru MLE îl eliminăm și numim pur și simplu densitatea verosimilitate:

$$\text{verosimilitatea} = f(x_1, x_2|\lambda) = \lambda^2 e^{-5\lambda}.$$

Valoarea lui  $\lambda$  care maximizează aceasta este aflată la fel ca în exemplul de mai sus. Ea este  $\hat{\lambda} = 2/5$ .

## 1.5 Proprietăți ale MLE

MLE se comportă bine la transformări. Adică, dacă  $\hat{p}$  este MLE pentru  $p$  și  $g$  este o funcție injectivă, atunci  $g(\hat{p})$  este MLE pentru  $g(p)$ . De exemplu,

dacă  $\hat{\sigma}$  este MLE pentru deviația standard  $\sigma$ , atunci  $(\hat{\sigma})^2$  este MLE pentru dispersia  $\sigma^2$ .

În plus, MLE este **asimptotic nedeplasat** și are **asimptotic dispersie minimă**. Pentru a explica aceste noțiuni, observăm că MLE este ea însăși o variabilă aleatoare deoarece datele sunt aleatoare și MLE este calculată din date. Fie  $x_1, x_2, \dots$  un sir de date dintr-o repartiție cu parametrul  $p$ . Fie  $\hat{p}_n$  MLE pentru  $p$  bazată pe datele  $x_1, \dots, x_n$ .

Asimptotic nedeplasat înseamnă că atunci când numărul datelor crește, media MLE converge la  $p$ . În simboluri:  $E(\hat{p}_n) \rightarrow p$  când  $n \rightarrow \infty$ . Desigur, am vrea ca MLE să fie aproape de  $p$  cu probabilitate mare, nu doar în medie, deci cu cât mai mică este dispersia MLE, cu atât mai bine. Asimptotic dispersie minimă înseamnă că, atunci când numărul de date crește, MLE are dispersia minimă dintre toți estimatorii nedeplasați ai lui  $p$ . În simboluri: pentru orice estimator nedeplasat  $\tilde{p}_n$  și  $\varepsilon > 0$  avem  $Var(\tilde{p}_n) + \varepsilon > Var(\hat{p}_n)$  când  $n \rightarrow \infty$ .

## 2 Actualizare Bayesiană cu a priori discrete

### 2.1 Scopurile învățării

1. Să poată aplica teorema lui Bayes pentru a calcula probabilități.
2. Să poată defini și identifica rolurile probabilității a priori, verosimilității (termenul Bayes), probabilității a posteriori, datelor și ipotezei în aplicarea teoremei lui Bayes.
3. Să poată folosi un tabel de actualizare Bayesiană pentru a calcula probabilitățile a posteriori.

### 2.2 Recapitularea teoremei lui Bayes

Reamintim că teorema lui Bayes ne permite să ”inversăm” probabilitățile condiționate. Dacă  $\mathcal{H}$  și  $\mathcal{D}$  sunt evenimente, atunci:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}.$$

### 2.3 Terminologie și teorema lui Bayes în formă tabelară

**Exemplul 1.** Sunt 3 tipuri de monede care au probabilități diferite de avers când sunt aruncate.

Tipul A: monedele sunt corecte cu probabilitatea aversului 0.5.

Tipul B: monedele sunt cu probabilitatea aversului 0.6.

Tipul C: monedele sunt cu probabilitatea aversului 0.9.

Presupunem că am un sertar cu 5 monede: 2 de tipul  $A$ , 2 de tipul  $B$  și una de tipul  $C$ . Iau aleator o monedă din sertar. Fără a v-o arăta, o arunc o dată și obțin avers. Care este probabilitatea ca ea să fie de tipul  $A$ ? Tipul  $B$ ? Tipul  $C$ ?

**Răspuns.** Fie  $A, B$  și  $C$  evenimentul că moneda aleasă a fost de tipul  $A$ , tipul  $B$ , respectiv tipul  $C$ . Fie  $\mathcal{D}$  evenimentul că aruncarea este avers. Problema cere

$$P(A|\mathcal{D}), P(B|\mathcal{D}), P(C|\mathcal{D}).$$

Terminologie:

**Experiment:** Luăm aleator o monedă din sertar, o aruncăm și înregistram rezultatul.

**Date:** rezultatul experimentului nostru. În acest caz evenimentul  $\mathcal{D}$  = "avers". Gândim  $\mathcal{D}$  ca datele care dau probe pentru sau împotriva fiecarei ipoteze.

**Ipoteze:** testăm 3 ipoteze: moneda este de tipul  $A, B$  sau  $C$ .

**Probabilitate a priori:** probabilitatea fiecărei ipoteze înaintea aruncării monedei (colectării datelor). Deoarece sertarul are 2 monede de tipul  $A$ , 2 de tipul  $B$  și una de tipul  $C$  avem

$$P(A) = 0.4, P(B) = 0.4, P(C) = 0.2.$$

**Verosimilitate:** (Aceasta este aceeași verosimilitate pe care am folosit-o pentru MLE.) Funcția de verosimilitate este  $P(\mathcal{D}|\mathcal{H})$ , i.e., probabilitatea datelor presupunând că ipoteza este adevărată. Cel mai adesea considerăm datele ca fixate și lăsăm ipotezele să varieze. De exemplu,  $P(\mathcal{D}|A)$  = probabilitatea aversului dacă moneda este de tipul  $A$ . În cazul nostru verosimilitățile sunt

$$P(\mathcal{D}|A) = 0.5, P(\mathcal{D}|B) = 0.6, P(\mathcal{D}|C) = 0.9.$$

Numele verosimilitate este atât de bine stabilit în literatură că trebuie să vî-l predăm. Dar în limbajul colocvial verosimilitate și probabilitate sunt sinonime. Aceasta duce adesea la confundarea funcției de verosimilitate cu probabilitatea unei ipoteze. De aceea preferăm să folosim numele termenul Bayes.

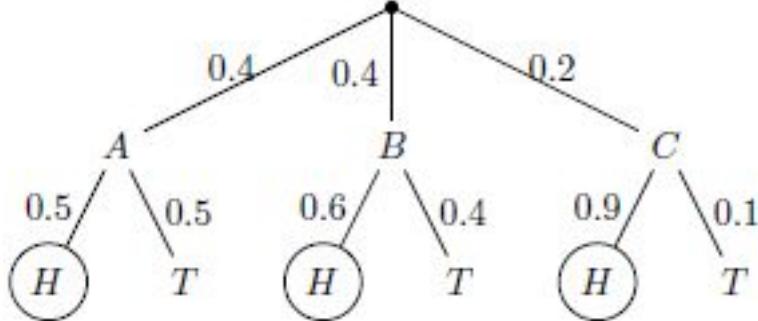
**Probabilitatea a posteriori:** probabilitatea (a posteriori) a fiecărei ipoteze cunoscând datele din aruncarea monedei.

$$P(A|\mathcal{D}), P(B|\mathcal{D}), P(C|\mathcal{D}).$$

Aceste probabilități a posteriori sunt cele cerute de problemă.

Folosim teorema lui Bayes pentru a calcula fiecare din probabilitățile a posteriori. Reamintim că datele  $\mathcal{D}$  sunt că aruncarea a fost avers.

Organizăm probabilitățile într-un arbore:



Arboarele de probabilități pentru alegera și aruncarea unei monede.

Teorema lui Bayes spune, de exemplu,  $P(A|\mathcal{D}) = \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})}$ . Numitorul  $P(\mathcal{D})$  este calculat cu legea probabilității totale:

$$P(\mathcal{D}) = P(\mathcal{D}|A)P(A) + P(\mathcal{D}|B)P(B) + P(\mathcal{D}|C)P(C) = 0.5 \cdot 0.4 + 0.6 \cdot 0.4 + 0.9 \cdot 0.2 = 0.62.$$

Acum fiecare dintre cele 3 probabilități a posteriori poate fi calculată:

$$\begin{aligned} P(A|\mathcal{D}) &= \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})} = \frac{0.5 \cdot 0.4}{0.62} = \frac{0.2}{0.62} \approx 0.3226, \\ P(B|\mathcal{D}) &= \frac{P(\mathcal{D}|B)P(B)}{P(\mathcal{D})} = \frac{0.6 \cdot 0.4}{0.62} = \frac{0.24}{0.62} \approx 0.3871, \\ P(C|\mathcal{D}) &= \frac{P(\mathcal{D}|C)P(C)}{P(\mathcal{D})} = \frac{0.9 \cdot 0.2}{0.62} = \frac{0.18}{0.62} \approx 0.2903. \end{aligned}$$

Observăm că probabilitatea totală  $P(\mathcal{D})$  este aceeași la fiecare numitor și este suma celor 3 numărători. Putem organiza toate acestea într-un **tabel de actualizare Bayesiană** ("hypothesis" = "ipoteză", "prior" = "a priori", "likelihood" = "verosimilitate", "numerator" = "numărător", "posterior" = "a posteriori"):

	Bayes			
hypothesis	prior	likelihood	numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D} \mathcal{H})$	$P(\mathcal{D} \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H} \mathcal{D})$
$A$	0.4	0.5	0.2	0.3226
$B$	0.4	0.6	0.24	0.3871
$C$	0.2	0.9	0.18	0.2903
total	1		0.62	1

**Numărătorul Bayes** este produsul dintre (probabilitatea) a priori și verosimilitate. Vedem în fiecare din calculele de mai sus cu formula lui Bayes că probabilitatea a posteriori este obținută împărțind numărătorul Bayes la  $P(\mathcal{D}) = 0.62$ . Vedem de asemenea că legea probabilității totale spune că  $P(\mathcal{D})$  este suma elementelor de pe coloana numărătorului Bayes.

**Actualizare Bayesiană:** Procesul trecerii de la probabilitatea a priori  $P(\mathcal{H})$  la a posteriori  $P(\mathcal{H}|\mathcal{D})$  se numește **actualizare Bayesiană**. Actualizarea Bayesiană utilizează datele pentru a modifica înțelegerea noastră pentru probabilitatea fiecărei ipoteze posibile.

### 2.3.1 Lucruri importante de observat

1. Sunt 2 tipuri de probabilități: Tipul 1 este probabilitatea standard a datelor, de exemplu, probabilitatea aversurilor este  $p = 0.9$ . Tipul 2 este probabilitatea ipotezelor, de exemplu, probabilitatea că moneda aleasă este de tipul  $A, B$  sau  $C$ . Al 2-lea tip are valori a priori (înainte de date) și a posteriori (după date).
2. Probabilitățile a posteriori (după date) pentru fiecare ipoteză sunt în ultima coloană. Vedem că moneda  $B$  este acum cea mai probabilă, cu toate că probabilitatea ei a scăzut de la 0.4 la 0.3871. Între timp, probabilitatea tipului  $C$  a crescut de la 0.2 la 0.2903.
3. Coloana numărătorului Bayes determină coloana probabilității a posterioiri. Pentru a o calcula pe ultima, pur și simplu am rescalat numărătorii Bayes astfel încât sumele lor să fie 1.
4. Dacă tot ce vrem este să aflăm cea mai probabilă ipoteză, numărătorul Bayes funcționează la fel de bine ca a posteriori.
5. Suma de pe coloana verosimilității nu este 1. Funcția de verosimilitate *nu* este o funcție de probabilitate.
6. Probabilitatea a posteriori reprezintă un "compromis" între verosimilitate și a priori. Când calculăm a posteriori, o a priori mare poate fi scăzută de o verosimilitate mică și o a priori mică poate fi mărită de o verosimilitate mare.
7. Estimarea de verosimilitate maximă (MLE) pentru exemplul 1 este ipoteza  $C$ , cu o verosimilitate  $P(\mathcal{D}|C) = 0.9$ . MLE este utilă, dar puteți vedea din acest exemplu că nu este totă povestea, deoarece tipul  $B$  are cea mai mare probabilitate a posteriori.

Cu terminologia la îndemână, putem exprima teorema lui Bayes în diverse moduri:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

$$P(\text{ipoteză}|date) = \frac{P(\text{date}|ipoteză)P(ipoteză)}{P(\text{date})}.$$

Cu datele fixate, numitorul  $P(\mathcal{D})$  servește doar la normalizarea probabilității a posteriori la 1. Astfel, putem exprima teorema lui Bayes ca o afirmație despre proporționalitatea a 2 funcții de  $\mathcal{H}$  (i.e., a ultimelor 2 coloane ale tabelului).

$$P(\text{ipoteză}|date) \propto P(\text{date}|ipoteză)P(ipoteză).$$

Aceasta conduce la cea mai elegantă formă a teoremei lui Bayes în contextul actualizării Bayesiene:

$$\text{a posteriori} \propto \text{verosimilitate} \times \text{a priori}.$$

### 2.3.2 Funcții masă de probabilitate a priori și a posteriori

Notății standard:

$\theta$  este valoarea ipotezei.

$p(\theta)$  este funcția masă de probabilitate a priori a ipotezei.

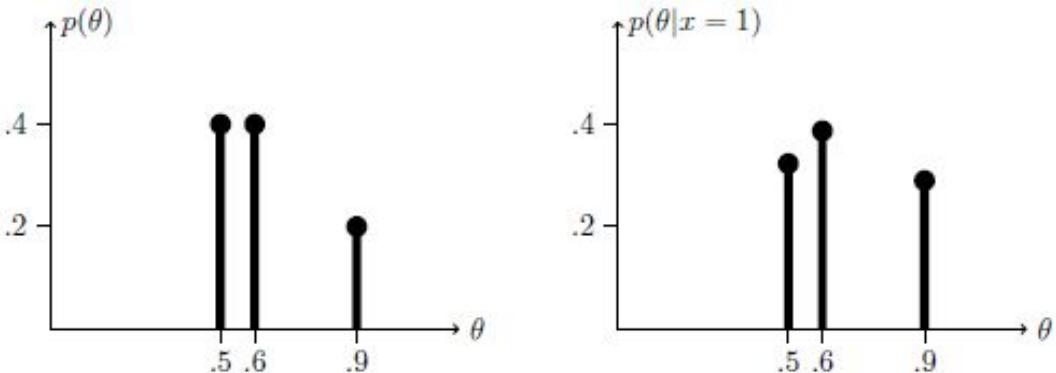
$p(\theta|\mathcal{D})$  este funcția masă de probabilitate a posteriori a ipotezei cunoscând datele.

$p(\mathcal{D}|\theta)$  este funcția de verosimilitate. (Aceasta nu este o pmf!)

În exemplul 1 putem reprezenta cele 3 ipoteze  $A$ ,  $B$  și  $C$  prin  $\theta = 0.5, 0.6, 0.9$ . Pentru date  $x = 1$  înseamnă avers și  $x = 0$  revers. Atunci probabilitățile a priori și a posteriori din tabel definesc funcțiile masă de probabilitate a priori și a posteriori. ("poster" = "a posteriori")

Hypothesis	$\theta$	prior pmf $p(\theta)$	poster pmf $p(\theta x=1)$
$A$	0.5	$P(A) = p(0.5) = 0.4$	$P(A \mathcal{D}) = p(0.5 x=1) = 0.3226$
$B$	0.6	$P(B) = p(0.6) = 0.4$	$P(B \mathcal{D}) = p(0.6 x=1) = 0.3871$
$C$	0.9	$P(C) = p(0.9) = 0.2$	$P(C \mathcal{D}) = p(0.9 x=1) = 0.2903$

Iată reprezentările pmf-urilor a priori și a posteriori din exemplu:



Pmf a priori  $p(\theta)$  și pmf a posteriori  $p(\theta|x = 1)$  pentru exemplul 1.

Dacă datele ar fi diferite, atunci coloana verosimilității din tabloul de actualizare Bayesiană ar fi diferită. Putem planifica pentru date diferite construind întregul [tabel de verosimilitate](#) în avans. În exemplul cu moneda sunt 2 posibilități pentru date: aruncarea este avers sau aruncarea este revers. Astfel, tabelul de verosimilitate complet are 2 coloane de verosimilitate:

hypothesis	likelihood $p(x \theta)$	
$\theta$	$p(x = 0 \theta)$	$p(x = 1 \theta)$
0.5	0.5	0.5
0.6	0.4	0.6
0.9	0.1	0.9

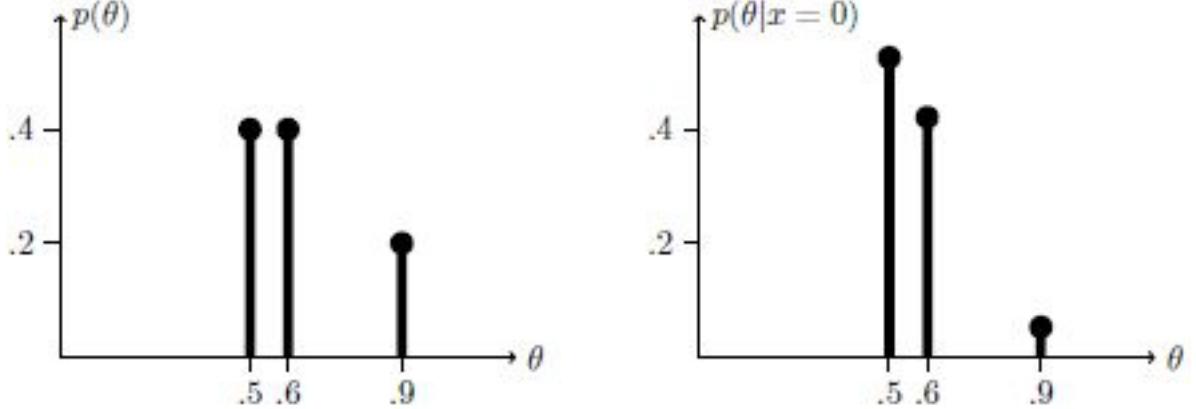
**Exemplul 2.** Folosind notația  $p(\theta)$ , etc., refață exemplul 1 presupunând că aruncarea a fost revers.

**Răspuns:** Deoarece datele s-au schimbat, coloana verosimilității din tabelul de actualizare Bayesiană este acum pentru  $x = 0$ . Adică, trebuie să luăm coloana  $p(x = 0|\theta)$  din tabelul de verosimilitate.

hypothesis	Bayes			
	prior	likelihood	numerator	posterior
$\theta$	$p(\theta)$	$p(x = 0 \theta)$	$p(x = 0 \theta)p(\theta)$	$p(\theta x = 0)$
0.5	0.4	0.5	0.2	0.5263
0.6	0.4	0.4	0.16	0.4211
0.9	0.2	0.1	0.02	0.0526
<b>total</b>	<b>1</b>		<b>0.38</b>	<b>1</b>

Acum probabilitatea tipului A a crescut de la 0.4 la 0.5263, în timp ce probabilitatea tipului C a scăzut de la 0.2 la 0.0526. Iată reprezentările core-

spunzătoare:



### 2.3.3 Hrană pentru minte

Presupunem că în exemplul 1 nu știm câte monede de fiecare tip sunt în sertar. Aruncăm una aleasă aleator și obținem avers. Cum veți decide care ipoteză (tip de monedă), dacă este vreuna, a fost cea mai susținută de date?

## 2.4 Reactualizare

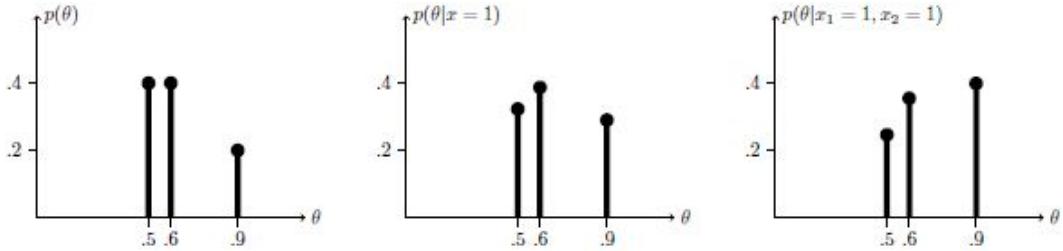
În deducția Bayesiană, după actualizarea de la a priori la a posteriori, putem să luăm mai multe date și să actualizăm iar. Pentru a 2-a actualizare, a posteriori din primele date devine a priori pentru ultimele date.

**Exemplul 3.** Presupunem că am ales o monedă ca în exemplul 1. O aruncăm o dată și obținem avers. Apoi aruncăm aceeași monedă și obținem iar avers. Care este probabilitatea ca moneda să fi fost de tipul A? Tipul B? Tipul C?

**Răspuns:** Tabelul devine mare:

hypothesis	prior	Bayes		Bayes		posterior 2
		likelihood 1	numerator 1	likelihood 2	numerator 2	
$\theta$	$p(\theta)$	$p(x_1 = 1 \theta)$	$p(x_1 = 1 \theta)p(\theta)$	$p(x_2 = 1 \theta)$	$p(x_2 = 1 \theta)p(x_1 = 1 \theta)p(\theta)$	$p(\theta x_1 = 1, x_2 = 1)$
0.5	0.4	0.5	0.2	0.5	0.1	0.2463
0.6	0.4	0.6	0.24	0.6	0.144	0.3547
0.9	0.2	0.9	0.18	0.9	0.162	0.3990
total	1				0.406	1

Al 2-lea numărător Bayes este calculat înmulțind primul numărător Bayes cu a 2-a verosimilitate; deoarece suntem interesați doar de a posteriori finală, nu trebuie să normalizăm până la ultimul pas. După cum este arătat în ultima coloană și reprezentare, după 2 aversuri ipoteza tipul C a luat în sfârșit conducerea.



A priori  $p(\theta)$ , prima a posteriori  $p(\theta|x_1 = 1)$ , a 2-a a posteriori  $p(\theta|x_1 = 1, x_2 = 1)$

## 2.5 Eroarea ratei de bază

**Exemplul 4.** Un test pentru o boală este atât sensibil cât și specific. Prin aceasta înțelegem că el este de obicei pozitiv când este testată o persoană care are boala și de obicei negativ când este testat cineva care nu are boala. Presupunem că rata adevărat pozitivă este 99% și rata fals pozitivă este 2%. Presupunem că răspândirea bolii în populația totală este 0.5%. Dacă o persoană aleatoare este testată pozitiv, care este probabilitatea ca ea să aibă boala?

**Răspuns:** Ca o recapitulare, întâi facem calculul folosind un arbore. Apoi refacem calculul folosind un tabel.

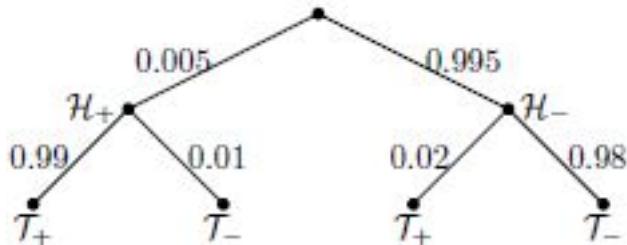
Folosim notațiile stabilite anterior pentru ipoteze și date: fie  $\mathcal{H}_+$  ipoteza (evenimentul) că persoana are boala și  $\mathcal{H}_-$  ipoteza că nu o are. Analog, fie  $\mathcal{T}_+$  și  $\mathcal{T}_-$  datele unui test pozitiv, respectiv negativ. Ni se cere să calculăm  $P(\mathcal{H}_+|\mathcal{T}_+)$ . Ni s-a dat

$$P(\mathcal{T}_+|\mathcal{H}_+) = 0.99, \quad P(\mathcal{T}_+|\mathcal{H}_-) = 0.02, \quad P(\mathcal{H}_+) = 0.005.$$

Din acestea putem calcula ratele fals negativă și adevărat negativă:

$$P(\mathcal{T}_-|\mathcal{H}_+) = 0.01, \quad P(\mathcal{T}_-|\mathcal{H}_-) = 0.98.$$

Toate aceste probabilități pot fi prezentate într-un arbore.



Teorema lui Bayes dă

$$P(\mathcal{H}_+|\mathcal{T}_+) = \frac{P(\mathcal{T}_+|\mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{T}_+)} = \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.02 \cdot 0.995} \approx 0.1992 \approx 20\%.$$

Acum refacem calculele folosind un tabel de actualizare Bayesiană:

hypothesis	Bayes			
	prior	likelihood	numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{T}_+ \mathcal{H})$	$P(\mathcal{T}_+ \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H} \mathcal{T}_+)$
$\mathcal{H}_+$	0.005	0.99	0.00495	0.19920
$\mathcal{H}_-$	0.995	0.02	0.01990	0.80080
total	1	NO SUM	0.02485	1

Tabelul arată că probabilitatea a posteriori  $P(\mathcal{H}_+|\mathcal{T}_+)$  ca o persoană cu test pozitiv să aibă boala este de aproximativ 20%. Aceasta este mult sub cea a sensibilității testului (99%), dar mult mai mare decât răspândirea bolii în populație (0.5%).

# Curs 11

Cristian Niculescu

## 1 Actualizare Bayesiană: predicție probabilistică

### 1.1 Scopul învățării

Să poată să utilizeze legea probabilității totale pentru a calcula probabilitățile predictive a priori și a posteriori.

### 1.2 Introducere

Am văzut actualizarea probabilității ipotezelor bazată pe date. Putem de asemenea folosi datele pentru a actualiza probabilitatea fiecărui rezultat posibil al unui experiment viitor.

#### 1.2.1 Predicție probabilistică; cuvinte de probabilitate estimativă (WEP)

Moduri de predicție:

Predicție: "Mâine va ploua."

Predicție folosind cuvinte de probabilitate estimativă (WEP): "Este probabil să plouă mâine."

Predicție probabilistică: "Mâine va ploua cu probabilitatea de 60% (și nu plouă cu probabilitatea de 40%)."

Fiecare tip de formulare este adecvat în diverse momente. Puteți vedea [http://en.wikipedia.org/wiki/Words\\_of\\_Estimative\\_Probability](http://en.wikipedia.org/wiki/Words_of_Estimative_Probability) pentru o discuție despre utilizarea adecvată a cuvintelor de probabilitate estimativă. Articolul conține de asemenea o listă de *cuvinte nevăstuică (weasel words)* ca "ar putea (might)", "nu poate exclude (cannot rule out)", "este de conceput (it's conceivable)" care ar trebui evitate deoarece aproape sigur produc confuzie.

Sunt multe locuri unde vrem să facem o predicție probabilistică. Exemple: Rezultate ale tratamentului medical

Prognoza meteo  
 Schimbarea climei  
 Pariuri sportive  
 Alegeri  
 ...

Acstea sunt toate situații unde există incertitudine despre rezultat și am vrea o descriere cât mai precisă posibil a ce se poate întâmpla.

### 1.3 Probabilități predictive

Predicția probabilistică înseamnă atribuirea unei probabilități fiecărui rezultat posibil al unui experiment.

Reconsiderăm exemplul cu monedele: sunt 3 tipuri de monede care sunt de nedistins în afară de probabilitățile lor de avers la aruncare.

Monedele de tip *A* sunt corecte, cu probabilitatea aversului 0.5.

Monedele de tip *B* au probabilitatea aversului 0.6.

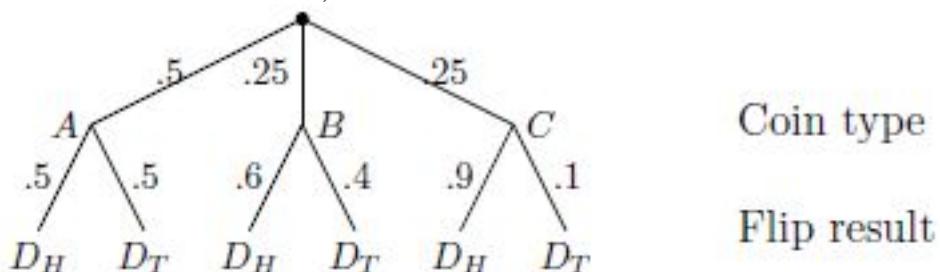
Monedele de tip *C* au probabilitatea aversului 0.9.

Avem un sertar cu 4 monede: 2 de tip *A*, una de tip *B* și una de tip *C*.

Alegem o monedă din sertar. Fie A evenimentul "moneda aleasă este de tipul *A*". Analog pentru *B* și *C*.

#### 1.3.1 Probabilități predictive a priori

Înainte de a colecta datele, calculăm probabilitatea ca moneda aleasă de noi să dea avers (sau revers) dacă o aruncăm. Fie  $D_H$  evenimentul "avers" și  $D_T$  evenimentul "revers". Putem folosi **legea probabilității totale** pentru a determina probabilitățile acestor evenimente. Desenând un arbore sau calculând cu formula, obținem ("Coin type" = "tipul monedei", "Flip result" = "rezultatul aruncării"):



$$\begin{aligned}
P(D_H) &= P(D_H|A)P(A) + P(D_H|B)P(B) + P(D_H|C)P(C) \\
&= 0.5 \cdot 0.5 + 0.6 \cdot 0.25 + 0.9 \cdot 0.25 = 0.625 \\
P(D_T) &= P(D_T|A)P(A) + P(D_T|B)P(B) + P(D_T|C)P(C) \\
&= 0.5 \cdot 0.5 + 0.4 \cdot 0.25 + 0.1 \cdot 0.25 = 0.375.
\end{aligned}$$

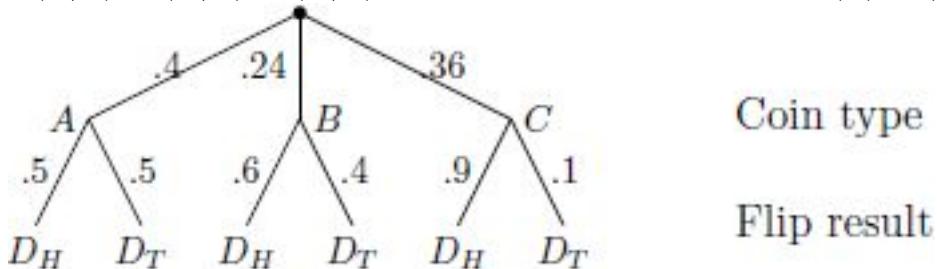
**Definiție:** Aceste probabilități dau o predicție (probabilistică) a ce se va întâmpla dacă moneda este aruncată. Deoarece ele sunt calculate înainte de a colecta orice dată, ele se numesc **probabilități predictive a priori**.

### 1.3.2 Probabilități predictive a posteriori

Presupunem că aruncăm moneda o dată și obținem avers. Acum avem datele  $D$ , pe care le putem folosi pentru a actualiza probabilitățile a priori ale ipotezelor noastre la probabilități a posteriori. Folosim un tabel Bayes pentru a facilita acest calcul:

		Bayes		
hypothesis	prior	likelihood	numerator	posterior
$H$	$P(H)$	$P(D H)$	$P(D H)P(H)$	$P(H D)$
$A$	0.5	0.5	0.25	0.4
$B$	0.25	0.6	0.15	0.24
$C$	0.25	0.9	0.225	0.36
total	1		0.625	1

Dacă am aruncat moneda o dată și am obținut avers, putem calcula probabilitatea că moneda aleasă de noi va da avers (sau revers) dacă o aruncăm a 2-a oară. Procedăm ca mai sus, dar folosind probabilitățile a posteriori  $P(A|D)$ ,  $P(B|D)$ ,  $P(C|D)$  în locul probabilităților a priori  $P(A)$ ,  $P(B)$ ,  $P(C)$ .



$$\begin{aligned}
P(D_H|D) &= P(D_H|A)P(A|D) + P(D_H|B)P(B|D) + P(D_H|C)P(C|D) \\
&= 0.5 \cdot 0.4 + 0.6 \cdot 0.24 + 0.9 \cdot 0.36 = 0.668 \\
P(D_T|D) &= P(D_T|A)P(A|D) + P(D_T|B)P(B|D) + P(D_T|C)P(C|D) \\
&= 0.5 \cdot 0.4 + 0.4 \cdot 0.24 + 0.1 \cdot 0.36 = 0.332.
\end{aligned}$$

**Definiție:** Aceste probabilități dau o predicție (probabilistică) a ceea ce se va întâmpla dacă moneda este aruncată iar. Deoarece sunt calculate după colectarea datelor și actualizarea a priori la a posteriori, ele sunt numite **probabilități predictive a posteriori**.

Observăm că aversul la prima aruncare crește probabilitatea aversului la a 2-a aruncare.

### 1.3.3 Recapitulare

Fiecare ipoteză dă o probabilitate diferită a aversului, deci probabilitatea totală a aversului este o medie ponderată. Pentru probabilitatea predictivă a priori a aversului, ponderile sunt date de probabilitățile a priori ale ipotezelor. Pentru probabilitatea predictivă a posteriori a aversului, ponderile sunt date de probabilitățile a posteriori ale ipotezelor.

**Rețineți:** Probabilitățile a priori și a posteriori sunt pentru ipoteze. Probabilitățile predictive a priori și a posteriori sunt pentru date. Ultimele prezic datele viitoare.

## 2 Actualizare Bayesiană: șanse

### 2.1 Scopurile învățării

1. Să poată converti șansele în probabilitate și invers.
2. Să poată actualiza șansele a priori în șanse a posteriori folosind factorii Bayes.
3. Să înțeleagă cum factorii Bayes măsoară gradul în care datele dau dovezi pentru sau împotriva ipotezelor.

### 2.2 Șanse

Când comparăm 2 evenimente, putem formula afirmațiile probabilistice în termeni de șanse.

**Definiție.** **Şansele** unui eveniment  $E$  versus evenimentul  $E'$  sunt raportul probabilităților lor  $P(E)/P(E')$ . Dacă nu se specifică evenimentul  $E'$ , al

2-lea eveniment este presupus a fi evenimentul complementar  $E^c$ . Astfel, **șansele** lui  $E$  sunt:

$$O(E) = \frac{P(E)}{P(E^c)}.$$

De exemplu,  $O(\text{ploaie}) = 2$  înseamnă că probabilitatea de a ploua este de 2 ori probabilitatea de a nu ploua ( $2/3$  versus  $1/3$ ). Putem spune că ”șansele de ploaie sunt 2 la 1”.

**Exemplu.** Pentru o monedă corectă,  $O(\text{avers}) = \frac{1/2}{1/2} = 1$ . Putem spune că șansele aversului sunt **1 la 1** sau **50-50**.

**Exemplu.** Pentru un zar standard, șansele de a da 4 sunt  $\frac{1/6}{5/6} = \frac{1}{5}$ . Putem spune că șansele sunt ”1 la 5 pentru” sau ”5 la 1 împotriva” a da un 4.

**Exemplu.** Probabilitatea unei perechi într-o mână de poker de 5 cărți este 0.42257. Astfel, șansele unei perechi sunt  $0.42257/(1 - 0.42257) = 0.73181$ .

**Formule de conversie:** Dacă  $P(E) = p$ , atunci  $O(E) = \frac{p}{1-p}$ . Dacă  $O(E) = q$ , atunci  $P(E) = \frac{q}{1+q}$ .

Observații:

1. A 2-a formulă provine din rezolvarea ecuației în  $p$   
 $q = p/(1 - p)$ .
2. Probabilitățile sunt între 0 și 1, în timp ce șansele sunt între 0 și  $\infty$ .
3. Proprietatea  $P(E^c) = 1 - P(E)$  devine  $O(E^c) = 1/O(E)$ .

**Exemplu.** Fie  $F$  evenimentul că o mână de poker să fie ful (3 cărți de un rang și alte 2 cărți de un alt rang). Atunci  $P(F) = 0.0014521$  deci  $O(F) = 0.0014521/(1 - 0.0014521) = 0.0014542$ .

Şansele de a nu avea ful sunt

$$O(F^c) = (1 - 0.0014521)/0.0014521 = 687.6578 = 1/O(F).$$

4. Dacă  $P(E)$  sau  $O(E)$  este mic, atunci  $O(E) \approx P(E)$ . Aceasta rezultă din formulele de conversie.

**Exemplu.** La exemplul din poker, unde  $F$  = ”ful”, am văzut că  $P(F)$  și  $O(F)$  diferă abia la a 6-a zecimală.

## 2.3 Actualizarea şanselor

### 2.3.1 Introducere

În actualizarea Bayesiană, am utilizat verosimilitatea datelor pentru a actualiza probabilitățile a priori ale ipotezelor la probabilități a posteriori. În limbaj de șanse, vom actualiza **șansele a priori** la **șanse a posteriori**. Datele dau dovezi pentru sau împotriva unei ipoteze dacă șansele ei a posteriori sunt mai mari sau mai mici decât șansele ei a priori.

### 2.3.2 Exemplu: sindromul Marfan

Sindromul Marfan este o boală genetică a ţesutului conjunctiv care apare la 1 din 15000 de oameni. Principalele caracteristici oculare ale sindromului Marfan includ ectopia de cristalin bilaterală, miopie și detașarea retinei. Aproximativ 70% din oamenii cu sindrom Marfan au cel puțin una din aceste caracteristici oculare, în timp ce doar 7% din oamenii fără sindrom Marfan au. (Nu garantăm acuratețea acestor numere.)

Dacă o persoană are cel puțin una din aceste caracteristici oculare, care sunt şansele să aibă sindromul Marfan?

**Răspuns:** Aceasta este o problemă standard de actualizare Bayesiană. Ipotezele noastre sunt:

$M$  = "persoana are sindromul Marfan",

$M^c$  = "persoana nu are sindromul Marfan".

Datele sunt:

$F$  = "persoana are cel puțin o caracteristică oculară".

Stim probabilitatea a priori a lui  $M$  și verosimilitățile lui  $F$  dat fiind  $M$  sau  $M^c$ :

$$P(M) = 1/15000, \quad P(F|M) = 0.7, \quad P(F|M^c) = 0.07.$$

Puteți calcula probabilitățile a posteriori folosind un tabel:

Bayes				
hypothesis	prior	likelihood	numerator	posterior
$H$	$P(H)$	$P(F H)$	$P(F H)P(H)$	$P(H F)$
$M$	0.000067	0.7	0.0000467	0.00066
$M^c$	0.999933	0.07	0.069995	0.99933
total	1		0.07004	1

Şansele a priori sunt:

$$O(M) = \frac{P(M)}{P(M^c)} = \frac{1/15000}{14999/15000} = \frac{1}{14999} \approx 0.000067.$$

Şansele a posteriori sunt date de raportul probabilităților a posteriori sau al numărătorilor Bayes, deoarece factorul normalizator este același la numărător și la numitor.

$$O(M|F) = \frac{P(M|F)}{P(M^c|F)} = \frac{P(F|M)P(M)}{P(F|M^c)P(M^c)} = 0.000667.$$

Şansele a posteriori sunt de 10 ori mai mari decât şansele a priori. În acest sens, a avea o caracteristică oculară este o dovadă puternică în favoarea

ipotezei  $M$ . Totuși, deoarece şansele a priori sunt atât de mici, este încă foarte puțin probabil ca persoana să aibă sindromul Marfan.

## 2.4 Factorii Bayes și puterea dovezii

Factorul 10 din exemplul precedent se numește factor Bayes.

**Definiție:** Pentru ipoteza  $H$  și datele  $D$ , **factorul Bayes** este raportul verosimilităților:

$$\text{factorul Bayes} = \frac{P(D|H)}{P(D|H^c)}.$$

Să vedem exact unde factorul Bayes apare în actualizarea şanselor. Avem

$$\begin{aligned} O(H|D) &= \frac{P(H|D)}{P(H^c|D)} \\ &= \frac{P(D|H)P(H)}{P(D|H^c)P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot \frac{P(H)}{P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot O(H). \end{aligned}$$

şansele a posteriori = **factorul Bayes** × şansele a priori.

Din această formulă vedem că factorul Bayes ( $BF$ ) ne spune dacă datele dau dovezi pentru sau împotriva ipotezei.

Dacă  $BF > 1$ , atunci şansele a posteriori sunt mai mari decât şansele a priori. Astfel, datele dau dovezi pentru ipoteză.

Dacă  $BF < 1$ , atunci şansele a posteriori sunt mai mici decât şansele a priori. Astfel, datele dau dovezi împotriva ipotezei.

Dacă  $BF = 1$ , atunci şansele a priori și a posteriori sunt egale. Astfel, datele nu dau nicio dovedă pentru sau împotriva ipotezei.

Următorul exemplu este din cartea *Information Theory, Inference, and Learning Algorithms* de David J. C. Mackay, care spune cu privire la probele dintr-un proces:

”În opinia mea, sarcina unui juriu ar trebui în general să fie a înmulți împreună rapoarte de verosimilitate evaluate cu grijă de la fiecare parte independentă de dovedă admisibilă cu o probabilitate a priori motivată la fel de grijuliu. Acest punct de vedere este împărtășit de mulți statisticieni, dar învățații judecători de apel britanici nu au fost de acord recent și în realitate au răsturnat verdictul unui proces deoarece jurații fuseseră învățați să utilizeze teorema lui Bayes pentru a trata dovezi ADN complicate.”

**Exemplul 1.** 2 oameni au lăsat urme de sânge la locul unei crime. Un suspect, Oliver, are grupa de sânge "0". Grupele de sânge ale celor 2 urme sunt "0" (o grupă obișnuită în populația locală, având frecvența 60%) și "AB" (o grupă rară, cu frecvența 1%). Dau aceste date (grupele "0" și "AB" găsite la locul crimei) dovezi în favoarea propoziției că "Oliver a fost una dintre cele 2 persoane prezente la locul crimei"?

**Răspuns:** Sunt 2 ipoteze:

$S =$  "Oliver și o altă persoană necunoscută au fost la locul crimei";

$S^c =$  "2 persoane necunoscute au fost la locul crimei".

Datele sunt:

$D =$  "grupele de sânge "0" și "AB" au fost găsite".

Factorul Bayes pentru prezența lui Oliver este  $BF_{\text{Oliver}} = \frac{P(D|S)}{P(D|S^c)}$ . Calculăm numărătorul și numitorul separat.

Datele spun că ambele grupe de sânge "0" și "AB" au fost aflate la locul crimei. Dacă Oliver a fost la locul crimei, grupa "0" ar fi acolo. Deci,  $P(D|S)$  este probabilitatea că cealaltă persoană avea grupa "AB". Ni s-a spus că aceasta este 0.01, deci  $P(D|S) = 0.01$ .

Dacă Oliver nu a fost la locul crimei, atunci au fost 2 persoane aleatoare, una cu grupa "0" și una cu grupa "AB". Probabilitatea acestui fapt este  $2 \cdot 0.6 \cdot 0.01$ . Factorul 2 este deoarece sunt 2 moduri în care se poate întâmpla asta - prima persoană are grupa 0 și a 2-a grupa AB sau viceversa. (Am presupus că grupele de sânge ale celor 2 persoane sunt independente. Aceasta nu este exact adevărat, dar pentru o populație suficient de mare este destul de aproape de adevăr. Fie  $N_0$ , numărul de oameni cu grupa 0,  $N_{AB}$ , numărul de oameni cu grupa AB și  $N$ , mărimea populației. Probabilitatea exactă este  $\frac{(N_0-1) \cdot N_{AB}}{C_{N-1}^2} = \frac{2 \cdot (N_0-1) \cdot N_{AB}}{(N-1)(N-2)} = 2 \cdot \frac{0.6N-1}{N-1} \cdot \frac{0.01N}{N-2} \xrightarrow{N \rightarrow \infty} 2 \cdot 0.6 \cdot 0.01$ . Aceasta arată că pentru  $N$  suficient de mare, probabilitatea este aproximativ  $2 \cdot 0.6 \cdot 0.01$ .)

Factorul Bayes pentru prezența lui Oliver la locul crimei este

$$BF_{\text{Oliver}} = \frac{P(D|S)}{P(D|S^c)} \approx \frac{0.01}{2 \cdot 0.6 \cdot 0.01} = 0.8(3).$$

Deoarece  $BF_{\text{Oliver}} < 1$ , datele dau dovezi (slabe) împotriva prezenței lui Oliver la locul crimei.

**Exemplul 2.** Un alt suspect, Alberto, are grupa de sânge AB. Dau aceleași date dovezi în favoarea propoziției că Alberto a fost una din cele 2 persoane prezente la locul crimei?

**Răspuns:** Refolosind notațiile de mai sus cu Alberto în locul lui Oliver, avem:

$$BF_{\text{Alberto}} = \frac{P(D|S)}{P(D|S^c)} \approx \frac{0.6}{2 \cdot 0.6 \cdot 0.01} = 50.$$

Deoarece  $BF_{\text{Alberto}} \gg 1$ , datele dă o dovedă puternică în favoarea prezenței lui Alberto la locul crimei.

Observații:

1. În ambele exemple, am calculat doar factorul Bayes, nu şansele a posterio-ri. Pentru a le calcula pe ultimele, am avea nevoie să ştim şansele a priori pentru prezența lui Oliver (sau Alberto) la locul crimei, pe baza altor dovezi.
2. Dacă 50% din populație ar fi avut grupa 0 de sânge în loc de 60%, atunci factorul Bayes al lui Oliver ar fi fost 1 (nici pentru, nici împotrivă). Mai general, punctul de echilibru pentru dovezile cu grupe de sânge este când proporția grupei de sânge a suspectului din populația generală egalează proporția grupei de sânge a suspectului dintre cele aflate la locul crimei.

#### 2.4.1 Reactualizare

Presupunem că adunăm datele în 2 etape, întâi  $D_1$ , apoi  $D_2$ . A posteriori finale pot fi calculate toate o dată sau în 2 etape, unde întâi actualizăm a priori folosind verosimilitățile pentru  $D_1$  și apoi actualizăm a posteriori rezultate după prima actualizare folosind verosimilitățile pentru  $D_2$ . Ultima abordare merge ori de câte ori verosimilitățile se înmulțesc:

$$P(D_1, D_2 | H) = P(D_1 | H)P(D_2 | H).$$

Deoarece verosimilitățile sunt condiționate de ipoteze, spunem că  $D_1$  și  $D_2$  sunt **condiționat independente** dacă relația de mai sus are loc pentru orice ipoteză  $H$ .

**Exemplu.** Sunt 5 zaruri într-un sertar, cu 4, 6, 8, 12 și 20 de fețe (acestea sunt ipotezele). Luăm aleator un zar și-l aruncăm de 2 ori, obținând întâi 7, apoi 11. Sunt aceste rezultate condiționat independente? Dar independente?

**Răspuns.** Rezultatele sunt condiționat independente. De exemplu, pentru ipoteza zarului cu 8 fețe avem:

$$\begin{aligned} P(7 \text{ la aruncarea } 1 | \text{zar cu } 8 \text{ fețe}) &= 1/8, \\ P(11 \text{ la aruncarea } 2 | \text{zar cu } 8 \text{ fețe}) &= 0, \\ P(7 \text{ la aruncarea } 1, 11 \text{ la aruncarea } 2 | \text{zar cu } 8 \text{ fețe}) &= 0. \end{aligned}$$

Pentru ipoteza zarului cu 20 de fețe avem:

$$\begin{aligned} P(7 \text{ la aruncarea } 1 | \text{zar cu } 20 \text{ fețe}) &= 1/20, \\ P(11 \text{ la aruncarea } 2 | \text{zar cu } 20 \text{ fețe}) &= 1/20, \\ P(7 \text{ la aruncarea } 1, 11 \text{ la aruncarea } 2 | \text{zar cu } 20 \text{ fețe}) &= (1/20)^2. \end{aligned}$$

Totuși, rezultatele nu sunt independente. Adică:

$$P(7 \text{ la aruncarea } 1, 11 \text{ la aruncarea } 2) \neq P(7 \text{ la aruncarea } 1)P(11 \text{ la aruncarea } 2).$$

Intuitiv, aceasta este deoarece 7 la aruncarea 1 ne permite să eliminăm zarurile cu 4 și 6 fețe, făcând un 11 la aruncarea 2 mai probabil. Ne verificăm intuiția calcuând ambii membri. La membrul drept avem:

$$\begin{aligned} P(7 \text{ la aruncarea } 1) &= \frac{1}{5} \cdot \frac{1}{8} + \frac{1}{5} \cdot \frac{1}{12} + \frac{1}{5} \cdot \frac{1}{20} = \frac{31}{600}, \\ P(11 \text{ la aruncarea } 2) &= \frac{1}{5} \cdot \frac{1}{12} + \frac{1}{5} \cdot \frac{1}{20} = \frac{2}{75}. \end{aligned}$$

Deci membrul drept este  $\frac{31}{600} \cdot \frac{2}{75} = 0.0013(7)$ .

Membrul stâng este:

$$\begin{aligned} P(7 \text{ la aruncarea } 1, 11 \text{ la aruncarea } 2) &= P(11 \text{ la aruncarea } 2|7 \text{ la aruncarea } 1) \cdot P(7 \text{ la aruncarea } 1) \\ &= \left( \frac{10}{31} \cdot \frac{1}{12} + \frac{6}{31} \cdot \frac{1}{20} \right) \cdot \frac{31}{600} \\ &= \frac{17}{9000} = 0.001(8). \end{aligned}$$

Aici  $\frac{10}{31}$  și  $\frac{6}{31}$  sunt probabilitățile a posteriori ale zarurilor cu 12 și cu 20 de fețe dat fiind un 7 la aruncarea 1 (se pot calcula cu un tabel de actualizare Bayesiană).

Concluzionăm că, fără condiționarea de ipoteze, aruncările nu sunt independente.

Întorcându-ne la schema generală, dacă  $D_1$  și  $D_2$  sunt condiționat independente pentru  $H$  și  $H^c$ , atunci are sens să considerăm fiecare factor Bayes independent:

$$BF_i = \frac{P(D_i|H)}{P(D_i|H^c)}, i = 1, 2.$$

Şansele a priori ale lui  $H$  sunt  $O(H)$ . Șansele a posteriori după  $D_1$  sunt

$$O(H|D_1) = BF_1 \cdot O(H).$$

Şansele a posteriori după  $D_1$  și  $D_2$  sunt

$$\begin{aligned} O(H|D_1, D_2) &= BF_2 \cdot O(H|D_1) \\ &= BF_2 \cdot BF_1 \cdot O(H). \end{aligned}$$

Actualizarea şanselor cu date noi, condiționat independente de cele vechi, se face prin înmulțirea şanselor a posteriori curente cu factorul Bayes al noilor

date.

### **Exemplul 3. Alte simptome ale sindromului Marfan**

Reamintim din exemplul de mai devreme că factorul Bayes pentru cel puțin o caracteristică oculară ( $F$ ) este

$$BF_F = \frac{P(F|M)}{P(F|M^c)} = \frac{0.7}{0.07} = 10.$$

Semnul încieturii mâinii ( $W$ ) este capacitatea de a înfășura o mâna în jurul celeilalte încieturi a mâinii astfel încât unghia degetului mic să poată fi acoperită de degetul mare. Presupunem că 10% din populație are semnul încieturii mâinii, în timp ce 90% din oamenii cu sindromul Marfan îl au. De aceea, factorul Bayes pentru semnul încieturii mâinii este

$$BF_W = \frac{P(W|M)}{P(W|M^c)} = \frac{0.9}{0.1} = 9.$$

Presupunem că  $F$  și  $W$  sunt simptome condiționat independente. Adică, printre oamenii cu sindromul Marfan, caracteristicile oculare și semnul încieturii mâinii sunt independente și printre oamenii fără sindromul Marfan, caracteristicile oculare și semnul încieturii mâinii sunt independente. Cu această presupunere, şansele a posteriori ale sindromului Marfan pentru cineva care are atât caracteristicile oculare cât și semnul încieturii mâinii sunt

$$O(M|F, W) = BF_W \cdot BF_F \cdot O(M) = 9 \cdot 10 \cdot \frac{1}{14999} \approx \frac{6}{1000}.$$

Putem converti şansele posterioare în probabilitate, dar, deoarece şansele sunt atât de mici, rezultatul este aproape același:

$$P(M|F, W) \approx \frac{6}{1000 + 6} \approx 0.596\%.$$

Așadar, caracteristicile oculare și semnul încieturii sunt ambele dovezi puternice în favoarea ipotezei  $M$  și, luate împreună, sunt dovezi foarte puternice. Din nou, deoarece şansele a priori sunt atât de mici, este încă improbabil că persoana are sindromul Marfan, dar în această situație merită să fie făcute alte teste date fiind consecințele potențial fatale ale bolii (ca anevrismul aortic sau disecția aortică).

Observăm de asemenea că, dacă o persoană are exact unul din cele 2 simptome, atunci produsul factorilor Bayes este aproape de 1 (9/10 sau 10/9). Astfel, cele 2 părți ale datelor se anulează una pe alta în ce privește dovezile pe care le dău pentru sindromul Marfan.

## 2.5 Log şanse

În practică, este adesea convenabil a lucra cu logaritmii naturali ai şanselor în locul şanselor. Destul de firesc, aceştia sunt numiți **log şanse**. Formula de actualizare Bayesiană

$$O(H|D_1, D_2) = BF_2 \cdot BF_1 \cdot O(H)$$

devine

$$\ln(O(H|D_1, D_2)) = \ln(BF_2) + \ln(BF_1) + \ln(O(H)).$$

Putem interpreta formula de mai sus pentru log şansele a posteriori ca suma log şanselor a priori cu toate dovezile  $\ln(BF_i)$  furnizate de date. Luând logaritmii, dovezile pentru ( $BF_i > 1$ ) sunt pozitive și dovezile împotriva ( $BF_i < 1$ ) sunt negative. Log şansele joacă de asemenea un rol central în regresia logistică, un model statistic legat de regresia liniară.

# Curs 12

Cristian Niculescu

## 1 Actualizare Bayesiană cu a priori continue

### 1.1 Scopurile învățării

1. Să înțeleagă o familie parametrizată de repartiții ca reprezentând un domeniu continuu de ipoteze pentru datele observate.
2. Să poată să enunțe teorema lui Bayes și legea probabilității totale pentru densități continue.
3. Să poată să aplice teorema lui Bayes pentru a actualiza o funcție densitate de probabilitate a priori la o pdf a posteriori cunoscând datele și funcția de verosimilitate.
4. Să poată interpreta și calcula probabilități predictive a posteriori.

### 1.2 Introducere

Până acum am făcut actualizare Bayesiană când aveam un număr finit de ipoteze, de pildă exemplul nostru cu zaruri avea 5 ipoteze (4, 6, 8, 12 sau 20 de fețe). Acum vom studia actualizare Bayesiană când există un **domeniu continuu de ipoteze**. Procesul de actualizare Bayesiană va fi în esență același ca în cazul discret. Ca de obicei când ne mișcăm de la discret la continuu vom avea nevoie să înlocuim funcția masă de probabilitate cu o funcție densitate de probabilitate și sumele cu integrale.

Primele câteva secțiuni ale acestui curs sunt dedicate lucrului cu pdf-uri. În particular, vom cuprinde legea probabilității totale și teorema lui Bayes. Acestea sunt în esență identice cu versiunile discrete. Apoi, vom aplica teorema lui Bayes și legea probabilității totale pentru actualizare Bayesiană.

### 1.3 Exemple cu domenii continue de ipoteze

Iată 3 exemple standard cu domenii continue de ipoteze.

**Exemplul 1.** Presupunem că avem un sistem care poate reuși (sau eşua) cu

probabilitatea  $p$ . Atunci putem face ipotezele că  $p$  este oriunde în domeniul  $[0,1]$ . Adică, avem un domeniu continuu de ipoteze. Adesea vom modela acest exemplu cu o monedă cu probabilitatea  $p$  a aversului necunoscută.

**Exemplul 2.** Durata de viață a unui anumit izotop este modelată de o repartiție exponențială  $\exp(\lambda)$ . În principiu, media duratei de viață  $1/\lambda$  poate fi orice număr în  $(0, \infty)$ .

**Exemplul 3.** Nu suntem restricționați la un singur parametru. În principiu, parametrii  $\mu$  și  $\sigma$  ai unei repartiții normale pot fi orice numere reale în  $(-\infty, \infty)$ , respectiv în  $(0, \infty)$ . Dacă modelăm durata sarcinii fără gemeni printr-o repartiție normală, atunci din milioane de date știm că  $\mu$  este aproximativ 40 de săptămâni și  $\sigma$  este aproximativ o săptămână.

În toate aceste exemple am modelat procesele aleatoare dând naștere la date printr-o repartiție cu parametri - numită **repartiție parametrizată**. Fiecare posibilă **alegere a parametrului (parametrilor)** este o ipoteză, de pildă putem face ipoteza că probabilitatea succesului în exemplul 1 este  $p = 0.7313$ . Avem o multime continuă de ipoteze deoarece am putea lua orice valoare între 0 și 1.

## 1.4 Convenții de notație

### 1.4.1 Modele parametrizate

Ca în exemplele de mai sus ipotezele noastre adesea iau forma **un anumit parametru are valoarea  $\theta$** . Vom utiliza adesea litera  $\theta$  pentru o ipoteză arbitrară. Aceasta va lăsa simboluri ca  $p$ ,  $f$  și  $x$  să ia sensurile lor uzuale de pmf, pdf și date. De asemenea, decât să spunem "ipoteza că parametrul de interes are valoarea  $\theta$ " vom spune simplu **ipoteza  $\theta$** .

### 1.4.2 Litere mari și mici

Avem 2 notații paralele pentru rezultate și probabilități:

1. (**Litere mari**) Evenimentul  $A$ , funcția de probabilitate  $P(A)$ .
2. (**Litere mici**) Valoarea  $x$ , pmf  $p(x)$ , pdf  $f(x)$ .

Aceste notații sunt legate prin  $P(X = x) = p(x)$ , unde  $x$  este o valoare a variabilei aleatoare discrete  $X$  și " $X = x$ " este evenimentul corespunzător.

Ducem aceste notății la probabilitățile folosite în actualizarea Bayesiană.

1. (**Litere mari**) Din ipotezele  $\mathcal{H}$  și datele  $\mathcal{D}$  calculăm câteva probabilități asociate:

$$P(\mathcal{H}), P(\mathcal{D}), P(\mathcal{H}|\mathcal{D}), P(\mathcal{D}|\mathcal{H}).$$

În exemplul cu moneda putem avea  $\mathcal{H}$  = "moneda aleasă are probabilitatea aversului 0.6",  $\mathcal{D}$  = "aruncarea a fost avers" și  $P(\mathcal{D}|\mathcal{H}) = 0.6$ .

2. (Litere mici) Valorile ipotezei  $\theta$  și valorile datelor  $x$  au ambele probabilități sau densități de probabilitate:

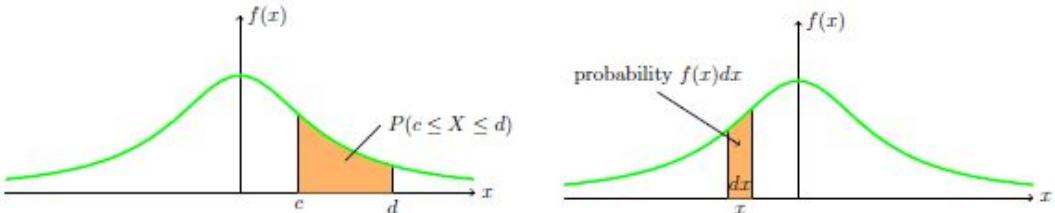
$$\begin{aligned} p(\theta), \quad p(x), \quad p(\theta|x), \quad p(x|\theta) \\ f(\theta), \quad f(x), \quad f(\theta|x), \quad f(x|\theta) \end{aligned}$$

În exemplul cu moneda putem avea  $\theta = 0.6$  și  $x = 1$ , astfel  $p(x|\theta) = 0.6$ . Putem de asemenea scrie  $p(x = 1|\theta = 0.6)$  pentru a evidenția valorile lui  $x$  și  $\theta$ , dar niciodată nu vom scrie doar  $p(1|0.6)$  deoarece este neclar care valoare este  $x$  și care este  $\theta$ .

Cu toate că încă vom folosi ambele tipuri de notății, de acum înainte vom folosi mai ales notăția cu litere mici implicând pmf-uri și pdf-uri. Ipotezele vor fi de obicei parametri reprezentați de litere grecești ( $\theta, \lambda, \mu, \sigma, \dots$ ) în timp ce valorile datelor vor fi de obicei reprezentate de litere românești ( $x, x_i, y, \dots$ ).

## 1.5 Recapitulare rapidă a pdf și probabilității

Presupunem că  $X$  este o variabilă aleatoare cu pdf  $f(x)$ . Reamintim că  $f(x)$  este o densitate; unitățile ei sunt probabilitate/(unitățile lui  $x$ ).



Probabilitatea că valoarea lui  $X$  este în  $[c, d]$  este dată de

$$\int_c^d f(x)dx.$$

Probabilitatea că  $X$  este într-un interval infinitezimal  $dx$  în jurul lui  $x$  este  $f(x)dx$ . De fapt, formula integrală este doar "suma" acestor probabilități infinitezimale. Putem vizualiza aceste probabilități văzând integrala ca aria de sub graficul lui  $f(x)$ . Pentru a manipula probabilități în loc de densități în cele ce urmează, vom utiliza frecvent noțiunea că  $f(x)dx$  este probabilitatea că  $X$  este într-un interval infinitezimal în jurul lui  $x$  de lungime  $dx$ .

## 1.6 A priori continue, verosimilități discrete

În cadrul Bayesian avem probabilități ale ipotezelor - numite probabilități a priori și a posteriori - și probabilități ale datelor cunoscând o ipoteză - numite verosimilități. În cursurile precedente atât ipotezele cât și datele

aveau domenii discrete de valori. Am văzut în introducere că putem avea un domeniu continuu de ipoteze. Același lucru este valabil și pentru date, dar pentru acum vom presupune că datele noastre pot lua doar o mulțime discretă de valori. În acest caz, verosimilitatea datelor  $x$  cunoscând ipoteza  $\theta$  este scrisă folosind o pmf:  $p(x|\theta)$ .

Vom utiliza următorul exemplu cu monedă pentru a explica aceste noțiuni.

**Exemplul 4.** Presupunem că avem o monedă cu probabilitatea aversurilor necunoscută  $\theta$ . Valoarea lui  $\theta$  este aleatoare și poate fi oriunde între 0 și 1. Pentru acest exemplu și pentru următoarele vom presupune că valoarea lui  $\theta$  are o repartiție cu **densitatea de probabilitate a priori continuă**  $f(\theta) = 2\theta$ . Avem o **verosimilitate discretă** deoarece aruncarea monedei are doar 2 rezultate,  $x = 1$  pentru avers și  $x = 0$  pentru revers.

$$p(x = 1|\theta) = \theta, \quad p(x = 0|\theta) = 1 - \theta.$$

**Gândiți:** Aceasta poate fi dificil de înțeles. Avem o monedă cu o probabilitate necunoscută  $\theta$  a aversului. Valoarea parametrului  $\theta$  este ea însăși aleatoare și are o pdf a priori  $f(\theta)$ . Poate ajuta să vedem că exemplele discrete din cursurile anterioare sunt similare. De exemplu, am avut o monedă care putea avea probabilitatea aversului 0.5, 0.6 sau 0.9. Deci, puteam numi ipotezele noastre  $H_{0.5}$ ,  $H_{0.6}$ ,  $H_{0.9}$  și acestea aveau probabilitățile a priori  $P(H_{0.5})$ , etc. Cu alte cuvinte, aveam o monedă cu o probabilitate necunoscută a aversurilor, aveam ipoteze despre acea probabilitate și fiecare dintre acele ipoteze aveau o probabilitate a priori.

## 1.7 Legea probabilității totale

Legea probabilității totale pentru repartiții de probabilitate continue este în esență aceeași ca pentru repartiții discrete. Înlocuim pmf a priori cu o pdf a priori și suma cu o integrală. Începem prin a recapitula legea pentru cazul discret.

Reamintim că pentru o mulțime discretă de ipoteze  $\{\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_n\}$  legea probabilității totale spune

$$P(\mathcal{D}) = \sum_{i=1}^n P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i). \quad (1)$$

Aceasta este **probabilitatea a priori** totală a lui  $\mathcal{D}$  deoarece am folosit probabilitățile a priori  $P(\mathcal{H}_i)$ .

În notațiile cu litere mici cu  $\theta_1, \theta_2, \dots, \theta_n$  pentru ipoteze și  $x$  pentru date, legea

probabilității totale este scrisă

$$p(x) = \sum_{i=1}^n p(x|\theta_i)P(\theta_i). \quad (2)$$

De asemenea am numit aceasta **probabilitatea predictivă a priori** a rezultatului  $x$  pentru a o distinge de probabilitatea a priori a ipotezei  $\theta$ .

Analog, este o lege a probabilității totale pentru pdf continue. O formulăm ca o teoremă folosind notația cu litere mici.

**Teoremă.** **Legea probabilității totale.** Presupunem că avem un parametru continuu  $\theta$  în domeniul  $[a, b]$  și datele discrete aleatoare  $x$ . Presupunem că  $\theta$  este el însuși aleator cu densitatea  $f(\theta)$  și că  $x$  și  $\theta$  au verosimilitatea  $p(x|\theta)$ . În acest caz, probabilitatea totală a lui  $x$  este dată de formula

$$p(x) = \int_a^b p(x|\theta)f(\theta)d\theta. \quad (3)$$

**Demonstrație.** Demonstrația noastră va fi analoagă versiunii discrete: termenul de probabilitate  $p(x|\theta)f(\theta)d\theta$  este perfect analog termenului  $p(x|\theta_i)p(\theta_i)$  din relația 2 (sau termenului  $P(\mathcal{D}|\mathcal{H}_i)P(\mathcal{H}_i)$  din relația 1). Continuând analogia: suma din relația 2 devine integrala din relația 3.

Ca în cazul discret, când gândim  $\theta$  ca o ipoteză care explică probabilitatea datelor, numim  $p(x)$  **probabilitatea predictivă a priori pentru  $x$** .

**Exemplul 5.** Continuarea exemplului 4. Avem o monedă cu probabilitatea  $\theta$  a aversului. Valoarea lui  $\theta$  este aleatoare cu pdf a priori  $f(\theta) = 2\theta$  pe  $[0,1]$ . Presupunem că aruncăm moneda o dată. Care este probabilitatea totală a aversului?

**Răspuns:** În exemplul 4 am observat că verosimilitățile sunt  $p(x = 1|\theta) = \theta$  și  $p(x = 0|\theta) = 1 - \theta$ . Probabilitatea totală a lui  $x = 1$  este

$$p(x = 1) = \int_0^1 p(x = 1|\theta)f(\theta)d\theta = \int_0^1 \theta \cdot 2\theta d\theta = \int_0^1 2\theta^2 d\theta = \frac{2\theta^3}{3} \Big|_0^1 = \frac{2}{3}.$$

## 1.8 Teorema lui Bayes pentru densități de probabilitate continue

Enunțul teoremei lui Bayes pentru pdf-uri continue este în esență identic cu cel pentru pmf-uri. O enunțăm incluzând  $d\theta$  pentru a avea probabilități autentice:

**Teoremă.** **Teorema lui Bayes.** Folosim aceleași presupuneri ca în legea probabilității totale, i.e.,  $\theta$  este un parametru continuu cu pdf  $f(\theta)$  și domeniul

$[a, b]$ ;  $x$  sunt datele discrete aleatoare; împreună au verosimilitatea  $p(x|\theta)$ . Cu aceste presupuneri:

$$f(\theta|x)d\theta = \frac{p(x|\theta)f(\theta)d\theta}{p(x)} = \frac{p(x|\theta)f(\theta)d\theta}{\int_a^b p(x|\theta)f(\theta)d\theta}. \quad (4)$$

**Demonstrație.** Fie  $\Theta$  variabila aleatoare care produce valoarea  $\theta$ . Considerăm evenimentele

$H = " \Theta \text{ este într-un interval de lungime } d\theta \text{ în jurul valorii } \theta "$

și

$D = " \text{valoarea datelor este } x "$ .

Atunci  $P(H) = f(\theta)d\theta$ ,  $P(D) = p(x)$  și  $P(D|H) = p(x|\theta)$ . Din forma uzuală a teoremei lui Bayes avem

$$f(\theta|x)d\theta = P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{p(x|\theta)f(\theta)d\theta}{p(x)}, \text{ q.e.d.}$$

Păstrarea factorului  $d\theta$  în enunțul teoremei lui Bayes este mai propice pentru a gândi în termeni de probabilități. Dar deoarece apare în toți membrii relației 4 se poate simplifica și putem scrie teorema lui Bayes în termeni de densități astfel:

$$f(\theta|x) = \frac{p(x|\theta)f(\theta)}{p(x)} = \frac{p(x|\theta)f(\theta)}{\int_a^b p(x|\theta)f(\theta)d\theta}.$$

## 1.9 Actualizare Bayesiană cu a priori continue

Caracteristici ale tabelului de actualizare Bayesiană pentru a priori continue:

1. Tabelul pentru a priori continue este foarte simplu: deoarece nu putem avea o linie pentru fiecare dintr-un număr infinit de ipoteze, vom avea doar **o linie care folosește o variabilă pentru toate ipotezele  $\theta$** .
2. Incluzând  $d\theta$ , toate aparițiile din tabel sunt probabilități și se aplică toate regulile de probabilități obișnuite.

**Exemplul 6.** ([Actualizare Bayesiană](#).) Continuarea exemplelor 4 și 5. Avem o monedă cu probabilitate necunoscută  $\theta$  a aversului. Valoarea lui  $\theta$  este aleatoare cu pdf a priori  $f(\theta) = 2\theta$ . Presupunem că aruncăm moneda o dată și obținem avers. Calculați pdf a posteriori pentru  $\theta$ .

**Răspuns:** Facem un tabel de actualizare cu coloanele obișnuite. Deoarece acesta este primul nostru exemplu, prima linie este versiunea abstractă a actualizării Bayesiene în general și a 2-a linie este actualizarea Bayesiană

pentru acest exemplu particular. "hypothesis" = "ipoteză", "prior" = "a priori", "likelihood" = "verosimilitate", "Bayes numerator" = "numărător Bayes", "posterior" = "a posteriori".

hypothesis	prior	likelihood	Bayes numerator	posterior
$\theta$	$f(\theta) d\theta$	$p(x = 1 \theta)$	$p(x = 1 \theta)f(\theta) d\theta$	$f(\theta x = 1) d\theta$
$\theta$	$2\theta d\theta$	$\theta$	$2\theta^2 d\theta$	$3\theta^2 d\theta$
total	$\int_a^b f(\theta) d\theta = 1$		$p(x = 1) = \int_0^1 2\theta^2 d\theta = 2/3$	1

De aceea pdf a posteriori (după vederea unui avers) este  $f(\theta|x) = 3\theta^2$ .

Comentarii:

1. Deoarece am folosit probabilitatea a priori  $f(\theta)d\theta$ , ipoteza ar fi trebuit să fie: "parametrul necunoscut este într-un interval de lungime  $d\theta$  în jurul lui  $\theta$ ". Chiar dacă este prea mult de scris pentru noi, asta trebuie să gândim de fiecare dată când scriem că ipoteza este  $\theta$ .
2. Pdf a posteriori pentru  $\theta$  se află înlăturând  $d\theta$  din probabilitatea a posteriori din tabel.

$$f(\theta|x) = 3\theta^2.$$

3. i)  $p(x)$  este **probabilitatea totală**. Deoarece avem o repartiție continuă, în loc de o sumă calculăm o integrală.
- ii) Incluzând  $d\theta$  în tabel, este clar ce integrală avem nevoie să calculăm pentru a afla probabilitatea totală  $p(x)$ .
4. Tabelul organizează versiunea continuă a teoremei lui Bayes. Anume, pdf a posteriori este legată de pdf a priori și verosimilitate prin

$$f(\theta|x)d\theta = \frac{p(x|\theta)f(\theta)d\theta}{\int_a^b p(x|\theta)f(\theta)d\theta} = \frac{p(x|\theta)f(\theta)d\theta}{p(x)}.$$

Împărțind la  $d\theta$  avem enunțul în termeni de densități.

5. Putem exprima teorema lui Bayes în forma:

$$f(\theta|x) \propto p(x|\theta) \cdot f(\theta)$$

a posteriori  $\propto$  verosimilitatea  $\times$  a priori.

### 1.9.1 A priori plate

O **a priori plată sau uniformă** presupune că fiecare ipoteză este egal probabilă. De exemplu, dacă  $\theta$  are domeniul  $[0,1]$ , atunci  $f(\theta) = 1$  este o a priori plată.

**Exemplul 7.** (A priori plate.) Avem o monedă cu probabilitatea necunoscută  $\theta$  a aversului. Presupunem că o aruncăm o dată și obținem revers. Presupunând o a priori plată, aflați probabilitatea a posteriori pentru  $\theta$ .

**Răspuns:** Procedăm ca la exemplul 6, cu a priori și verosimilitatea modificate.

hypothesis $\theta$	prior $f(\theta) d\theta$	likelihood $p(x = 0 \theta)$	Bayes numerator	posterior $f(\theta x = 0) d\theta$
$\theta$	$1 \cdot d\theta$	$1 - \theta$	$(1 - \theta) d\theta$	$2(1 - \theta) d\theta$
total	$\int_a^b f(\theta) d\theta = 1$		$p(x = 0) = \int_0^1 (1 - \theta) d\theta = 1/2$	$1$

### 1.9.2 Folosirea pdf a posteriori

**Exemplul 8.** În exemplul anterior probabilitatea a priori a fost plată. Întâi arătați că aceasta înseamnă că a priori moneda este egal părtinitoare spre avers sau revers. Apoi, după observarea unui avers, care este probabilitatea (a posteriori) ca moneda să fie părtinitoare spre avers?

**Răspuns.** Deoarece parametrul  $\theta$  este probabilitatea aversului, întâi ni se cere să arătăm că  $P(\theta > 0.5) = 0.5$ , apoi ni se cere să calculăm  $P(\theta > 0.5|x = 1)$ . Acestea se calculează din pdf-urile a priori, respectiv a posteriori.

Probabilitatea a priori ca moneda să fie părtinitoare spre avers este

$$P(\theta > 0.5) = \int_{0.5}^1 f(\theta) d\theta = \int_{0.5}^1 1 d\theta = \theta|_{0.5}^1 = \frac{1}{2}.$$

Probabilitatea de  $1/2$  înseamnă că moneda este egal părtinitoare spre avers sau revers.

Ca în exemplul 7 se calculează că  $f(\theta|x = 1) = 2\theta$ .

Probabilitatea a posteriori că moneda este părtinitoare spre avers este

$$P(\theta > 0.5|x = 1) = \int_{0.5}^1 f(\theta|x = 1) d\theta = \int_{0.5}^1 2\theta d\theta = \theta^2|_{0.5}^1 = \frac{3}{4}.$$

Vedem că observarea unui avers a crescut probabilitatea ca moneda să fie părtinitoare spre avers de la  $1/2$  la  $3/4$ .

## 1.10 Probabilități predictive

La fel ca în cazul discret, suntem de asemenea interesați să folosim probabilitățile a posteriori ale ipotezelor pentru a face predicții pentru ce se va

întâmplă mai departe.

**Exemplul 9.** (*Predicție a priori și a posteriori.*) Continuarea exemplelor 4, 5, 6: avem o monedă cu probabilitate necunoscută  $\theta$  a aversului și valoarea lui  $\theta$  are pdf a priori  $f(\theta) = 2\theta$ . Aflați probabilitatea predictivă a priori a aversului. Apoi presupunând că prima aruncare a fost avers, aflați probabilitățile predictive a posteriori atât pentru avers cât și pentru revers la a 2-a aruncare.

**Răspuns:** Fie  $x_1$  rezultatul primei aruncări și  $x_2$  al celei de-a 2-a. Probabilitatea predictivă a priori este exact probabilitatea totală calculată în exemplele 5 și 6.

$$p(x = 1) = \int_0^1 p(x = 1|\theta)f(\theta)d\theta = \int_0^1 \theta \cdot 2\theta d\theta = \int_0^1 2\theta^2 d\theta = \frac{2\theta^3}{3} \Big|_0^1 = \frac{2}{3}.$$

Probabilitățile predictive a posteriori sunt probabilitățile totale calculate folosind pdf a posteriori. Din exemplul 6 știm că pdf a posteriori este  $f(\theta|x_1 = 1) = 3\theta^2$ . Probabilitățile predictive a posteriori sunt

$$\begin{aligned} p(x_2 = 1|x_1 = 1) &= \int_0^1 p(x_2 = 1|\theta, x_1 = 1)f(\theta|x_1 = 1)d\theta = \int_0^1 \theta \cdot 3\theta^2 d\theta \\ &= \frac{3\theta^4}{4} \Big|_0^1 = \frac{3}{4}, \\ p(x_2 = 0|x_1 = 1) &= \int_0^1 p(x_2 = 0|\theta, x_1 = 1)f(\theta|x_1 = 1)d\theta = \int_0^1 (1 - \theta) \cdot 3\theta^2 d\theta \\ &= \left( \theta^3 - \frac{3\theta^4}{4} \right) \Big|_0^1 = \frac{1}{4}. \end{aligned}$$

(Mai simplu, puteam calcula

$$p(x_2 = 0|x_1 = 1) = 1 - p(x_2 = 1|x_1 = 1) = 1 - 3/4 = 1/4.)$$

## 1.11 De la actualizarea Bayesiană discretă la actualizarea Bayesiană continuă

Pentru a dezvolta intuiția pentru tranziția de la actualizarea Bayesiană discretă la actualizarea Bayesiană continuă, vom:

- i) aproxima domeniul continuu de ipoteze printr-un număr finit de ipoteze;
  - ii) crea tabelul de actualizare discretă pentru numărul finit de ipoteze;
  - iii) consideră cum se schimbă tabelul când numărul de ipoteze tinde la infinit.
- În acest fel, vom vedea că pmf-urile a priori și a posteriori converg la pdf-urile a priori și a posteriori.

**Exemplul 10.** Pentru a păstra lucrurile concrete, vom lucra cu moneda cu o a priori plată  $f(\theta) = 1$  din exemplul 7. Scopul nostru este să mergem de la discret la continuu crescând numărul de ipoteze.

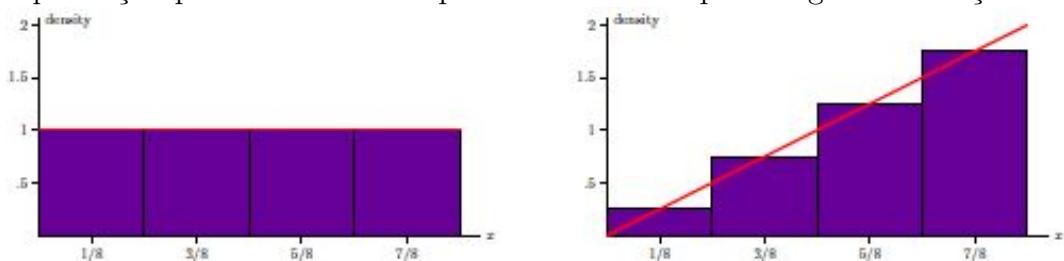
**4 ipoteze.** Împărțim  $[0,1]$  în 4 intervale egale:  $[0,1/4]$ ,  $[1/4,1/2]$ ,  $[1/2,3/4]$ ,  $[3/4,1]$ . Fiecare interval are lungimea  $\Delta\theta = 1/4$ . Punem cele 4 ipoteze ale noastre  $\theta_i$  în centrele celor 4 intervale:

$$\theta_1 : " \theta = 1/8 ", \theta_2 : " \theta = 3/8 ", \theta_3 : " \theta = 5/8 ", \theta_4 : " \theta = 7/8 ".$$

A priori plată dă fiecarei ipoteze o probabilitate de  $1/4 = 1 \cdot \Delta\theta$ . Avem tabelul:

hypothesis	prior	likelihood	Bayes num.	posterior
$\theta = 1/8$	$1/4$	$1/8$	$(1/4) \times (1/8)$	$1/16$
$\theta = 3/8$	$1/4$	$3/8$	$(1/4) \times (3/8)$	$3/16$
$\theta = 5/8$	$1/4$	$5/8$	$(1/4) \times (5/8)$	$5/16$
$\theta = 7/8$	$1/4$	$7/8$	$(1/4) \times (7/8)$	$7/16$
Total	1	-	$\sum_{i=1}^n \theta_i \Delta\theta$	1

Iată histogramele de densitate ale pmf-urilor a priori și a posteriori. Pdf-urile a priori și a posteriori din exemplul 7 sunt trasate pe histograme cu roșu.

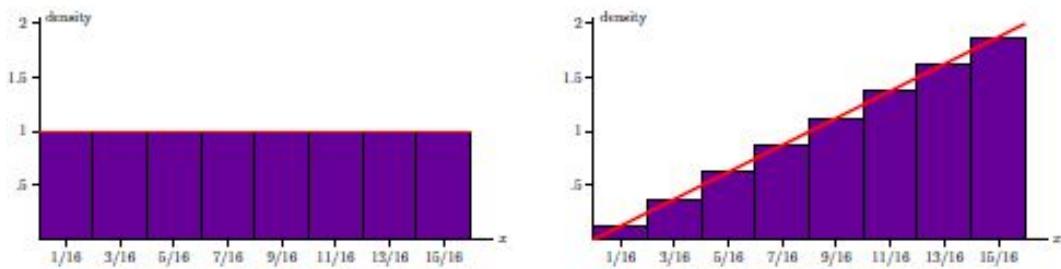


**8 ipoteze.** Acum împărțim  $[0,1]$  în 8 intervale, fiecare cu lungimea  $\Delta\theta = 1/8$  și punem cele 8 ipoteze ale noastre  $\theta_i$  în centrele celor 4 intervale:

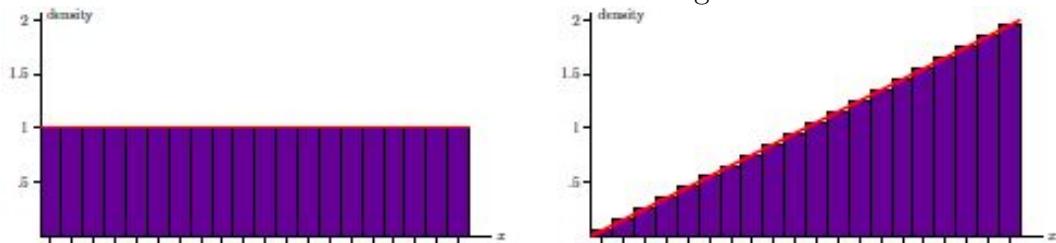
$$\begin{aligned} \theta_1 : " \theta = 1/16 ", \theta_2 : " \theta = 3/16 ", \theta_3 : " \theta = 5/16 ", \theta_4 : " \theta = 7/16 ", \\ \theta_5 : " \theta = 9/16 ", \theta_6 : " \theta = 11/16 ", \theta_7 : " \theta = 13/16 ", \theta_8 : " \theta = 15/16 ". \end{aligned}$$

A priori plată dă fiecărei ipoteze o probabilitate de  $1/8 = 1 \cdot \Delta\theta$ . Iată tabelul și histogramele de densitate:

hypothesis	prior	likelihood	Bayes num.	posterior
$\theta = 1/16$	$1/8$	$1/16$	$(1/8) \times (1/16)$	$1/64$
$\theta = 3/16$	$1/8$	$3/16$	$(1/8) \times (3/16)$	$3/64$
$\theta = 5/16$	$1/8$	$5/16$	$(1/8) \times (5/16)$	$5/64$
$\theta = 7/16$	$1/8$	$7/16$	$(1/8) \times (7/16)$	$7/64$
$\theta = 9/16$	$1/8$	$9/16$	$(1/8) \times (9/16)$	$9/64$
$\theta = 11/16$	$1/8$	$11/16$	$(1/8) \times (11/16)$	$11/64$
$\theta = 13/16$	$1/8$	$13/16$	$(1/8) \times (13/16)$	$13/64$
$\theta = 15/16$	$1/8$	$15/16$	$(1/8) \times (15/16)$	$15/64$
Total	1	—	$\sum_{i=1}^n \theta_i \Delta\theta$	1



**20 ipoteze.** Acum împărțim  $[0,1]$  în 20 de intervale. Tabelul se face analog ca în cele 2 cazuri anterioare. Sărim direct la histogramele de densitate.



Privind la reprezentări, vedem cum histogramele de densitate a priori și a posteriori converg la funcțiile densitate de probabilitate a priori și a posterio-ori.

## 2 Convenții de notație

### 2.1 Scopul învățării

Să poată lucra cu notațiile și termenii pe care îi folosim pentru a descrie probabilitățile și verosimilitatea.

### 2.2 Introducere

Am introdus un număr de notații diferite pentru probabilitate, ipoteze și date. Le adunăm aici, pentru a le avea într-un singur loc.

### 2.3 Notații și terminologie pentru date și ipoteze

Când am început cursul am vorbit despre rezultate (sau cazuri), de exemplu avers sau revers. Apoi, când am introdus variabilele aleatoare am dat rezultatelor valori numerice, de exemplu 1 pentru avers și 0 pentru revers. Aceasta ne-a permis să facem lucruri precum calculul mediilor și dispersiilor. Reamintim convențiile noastre de notație:

Evenimentele sunt notate cu majuscule, de exemplu  $A, B, C$ .

O variabilă aleatoare este majuscula  $X$  și ia valorile  $x$ .

Legătura dintre valori și evenimente: " $X = x$ " este evenimentul că  $X$  ia valoarea  $x$ .

Probabilitatea unui eveniment este majuscula  $P(A)$ .

O variabilă aleatoare discretă are funcția masă de probabilitate  $p(x)$ . Legătura dintre  $P$  și  $p$  este că  $P(X = x) = p(x)$ .

O variabilă aleatoare continuă are funcția densitate de probabilitate  $f(x)$ .

Legătura dintre  $P$  și  $f$  este că  $P(a \leq X \leq b) = \int_a^b f(x)dx$ .

Pentru o variabilă aleatoare continuă  $X$ , probabilitatea că  $x$  este într-un interval infinitesimal de lungime  $dx$  în jurul lui  $x$  este  $f(x)dx$ .

În contextul actualizării Bayesiene avem convenții similare.

Folosim litere mari caligrafice (de mâna), în special  $\mathcal{H}$ , pentru a indica o ipoteză, de exemplu  $\mathcal{H}$  = "moneda este corectă".

Folosim litere mici grecești, în special  $\theta$ , pentru a indica valoarea ipotetică a unui parametru al modelului, de exemplu probabilitatea ca moneda să ateeizeze avers este  $\theta = 0.5$ .

Folosim litere mari caligrafice (de mâna), în special  $\mathcal{D}$ , când vorbim despre date ca evenimente. De exemplu,  $\mathcal{D}$  = "secvența de aruncări a fost HTH".

Folosim litere mici, în special  $x$ , când vorbim despre date ca valori. De exemplu, secvența de date a fost  $x_1, x_2, x_3 = 1, 0, 1$ .

Când mulțimea de ipoteze este discretă, putem folosi probabilitatea ipotezelor individuale, de exemplu  $p(\theta)$ . Când mulțimea este continuă, trebuie să folosim probabilitatea pentru un domeniu infinitezimal de ipoteze, de exemplu  $f(\theta)d\theta$ .

Următorul tabel rezumă aceasta pentru  $\theta$  discretă și  $\theta$  continuă. În ambele cazuri presupunem o mulțime discretă de rezultate posibile (date)  $x$ .

		Bayes			
	hypothesis	prior	likelihood	numerator	posterior
	$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D} \mathcal{H})$	$P(\mathcal{D} \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H} \mathcal{D})$
Discrete $\theta$ :	$\theta$	$p(\theta)$	$p(x \theta)$	$p(x \theta)p(\theta)$	$p(\theta x)$
Continuous $\theta$ :	$\theta$	$f(\theta)d\theta$	$p(x \theta)$	$p(x \theta)f(\theta)d\theta$	$f(\theta x)d\theta$

Reamintim că ipoteza continuă  $\theta$  este în realitate o prescurtare pentru "parametrul  $\theta$  este într-un interval de lungime  $d\theta$  în jurul lui  $\theta$ ".

# Curs 13

Cristian Niculescu

## 1 Repartiții beta

### 1.1 Scopurile învățării

1. Să se familiarizeze cu familia cu 2 parametri a repartițiilor beta și cu normalizarea ei.
2. Să poată să actualizeze o a priori beta la o a posteriori beta în cazul unei verosimilități binomiale.

### 1.2 Repartiția beta

Repartiția beta  $\text{beta}(a, b)$  este o repartiție cu **2 parametri** cu domeniul  $[0, 1]$  și pdf

$$f(\theta) = \frac{(a+b-1)!}{(a-1)!(b-1)!} \theta^{a-1} (1-\theta)^{b-1}.$$

Următoarea aplicație ne permite să explorăm forma repartiției Beta când parametrii variază:

<http://mathlets.org/mathlets/beta-distribution/>.

Repartiția beta poate fi definită pentru orice numere reale  $a > 0$  și  $b > 0$ . Am definit-o doar pentru  $a, b \in \mathbb{N}^*$ , dar puteți vedea întreaga istorie aici: [http://en.wikipedia.org/wiki/Beta\\_distribution](http://en.wikipedia.org/wiki/Beta_distribution).

În contextul actualizării bayesiene,  $a$  și  $b$  sunt numiți adesea **hiperparametri** pentru a-i deosebi de parametrul necunoscut  $\theta$  care reprezintă ipoteza noastră. Într-un anumit sens,  $a$  și  $b$  sunt la "un nivel mai sus" decât  $\theta$ , deoarece ei parametrizează pdf.

#### 1.2.1 O observație simplă, dar importantă

Dacă o pdf  $f(\theta)$  are forma  $c\theta^{a-1}(1-\theta)^{b-1}$ , atunci  $f(\theta)$  este pdf a unei repartiții beta( $a, b$ ) și constanta de normalizare trebuie să fie

$$c = \frac{(a+b-1)!}{(a-1)!(b-1)!}.$$

Aceasta rezultă deoarece constanta  $c$  trebuie să normalizeze pdf pentru a avea probabilitatea totală 1. Există doar o asemenea constantă și ea este dată în formula pentru repartiția beta.

O observație similară este valabilă pentru repartițiile normală, exponențială, etc.

### 1.2.2 A priori și a posteriori beta pentru variabile aleatoare binomiale

**Exemplul 1.** Presupunem că avem o monedă cu probabilitate necunoscută  $\theta$  a aversului. Aruncăm moneda de 12 ori și obținem 8 aversuri și 4 reversuri. Plecând cu o a priori plată, arătați că pdf a posteriori este o repartiție beta(9, 5).

**Răspuns.** Notăm cu  $x_1$  datele din cele 12 aruncări. În tabelul următor numim  $c_2$  factorul constant din coloana a posteriori. Observația noastră simplă ne va spune că acesta trebuie să fie factorul constant din pdf beta.

Datele sunt 8 aversuri și 4 reversuri. Deoarece acestea vin dintr-o repartiție binomială( $12, \theta$ ), verosimilitatea este  $p(x_1|\theta) = C_{12}^8 \theta^8 (1-\theta)^4$ . Astfel, tabelul de actualizare Bayesiană este

hypothesis	prior	likelihood	Bayes	
			numerator	posterior
$\theta$	$1 \cdot d\theta$	$\binom{12}{8} \theta^8 (1-\theta)^4$	$\binom{12}{8} \theta^8 (1-\theta)^4 d\theta$	$c_2 \theta^8 (1-\theta)^4 d\theta$
total	1		$T = \binom{12}{8} \int_0^1 \theta^8 (1-\theta)^4 d\theta$	1

Observația noastră simplă de mai sus are loc cu  $a = 9$  și  $b = 5$ . De aceea pdf a posteriori

$$f(\theta|x_1) = c_2 \theta^8 (1-\theta)^4$$

are o repartiție beta(9, 5) și constanta de normalizare  $c_2$  trebuie să fie

$$c_2 = \frac{13!}{8! \cdot 4!}.$$

Notă. Am inclus explicit coeficientul binomial  $C_{12}^8$  în verosimilitate. Puteam să-l notăm cu  $c_1$  și să nu-i dăm valoarea explicită.

**Exemplul 2.** Acum presupunem că aruncăm aceeași monedă din nou, obținând  $n$  aversuri și  $m$  reversuri. Folosind pdf a posteriori din exemplul precedent ca nouă noastră pdf a priori, arătați că noua pdf a posteriori este cea a unei repartiții beta( $9 + n, 5 + m$ ).

**Răspuns.** Totul este în tabel. Vom numi  $x_2$  datele acestor  $n + m$  aruncări adiționale. De această dată nu vom mai face explicit coeficientul binomial. În loc de aceasta îl vom nota cu  $c_3$ . Ori de câte ori avem nevoie de o nouă

etichetă vom folosi  $c$  cu un nou indice. În loc de "Bayes posterior" se va citi "numărătorul Bayes", iar în loc de "numerator" se va citi "a posteriori".

			Bayes	
hyp.	prior	likelihood	posterior	numerator
$\theta$	$c_2 \theta^8 (1 - \theta)^4 d\theta$	$c_3 \theta^n (1 - \theta)^m$	$c_2 c_3 \theta^{n+8} (1 - \theta)^{m+4} d\theta$	$c_4 \theta^{n+8} (1 - \theta)^{m+4} d\theta$

total	1	$T = \int_0^1 c_2 c_3 \theta^{n+8} (1 - \theta)^{m+4} d\theta$	1
-------	---	----------------------------------------------------------------	---

Din nou observația noastră simplă are loc și, de aceea, pdf a posteriori

$$f(\theta|x_1, x_2) = c_4 \theta^{n+8} (1 - \theta)^{m+4}$$

este cea a unei repartiții beta( $n + 9, m + 5$ ).

**Observație.** **Beta plată.** Repartiția beta(1, 1) coincide cu repartiția uniformă pe  $[0, 1]$ , pe care am numit-o de asemenea a priori plată pentru  $\theta$ . Aceasta rezultă înlocuind  $a = 1$  și  $b = 1$  în definiția repartiției beta, dând  $f(\theta) = 1$ .

**Rezumat.** Dacă probabilitatea aversului este  $\theta$ , numărul de aversuri în  $n+m$  aruncări are o repartiție binomială( $n+m, \theta$ ). Am văzut că dacă a priori pentru  $\theta$  este o repartiție beta, atunci la fel este și a posteriori; doar parametrii  $a$  și  $b$  ai repartiției beta se schimbă! Rezumăm precis cum se schimbă într-un tabel. Presupunem că datele sunt  $n$  aversuri în  $n+m$  aruncări.

hypothesis	data	prior	likelihood	posterior
$\theta$	$x = n$	beta( $a, b$ )	binomial( $n+m, \theta$ )	beta( $a+n, b+m$ )
$\theta$	$x = n$	$c_1 \theta^{a-1} (1 - \theta)^{b-1} d\theta$	$c_2 \theta^n (1 - \theta)^m$	$c_3 \theta^{a+n-1} (1 - \theta)^{b+m-1} d\theta$

### 1.2.3 A priori conjugate

Repartiția beta este numită o **a priori conjugată** pentru repartiția binomială. Aceasta înseamnă că, dacă funcția de verosimilitate este binomială, atunci o a priori beta dă o a posteriori beta. De fapt, repartiția beta este o a priori conjugată și pentru repartițiile Bernoulli și geometrică.

Un alt exemplu important: repartiția normală își este propria a priori conjugată. În particular, dacă funcția de verosimilitate este normală cu dispersie cunoscută, atunci o a priori normală dă o a posteriori normală.

A priori conjugate sunt utile deoarece reduc actualizarea Bayesiană la modificarea parametrilor repartiției a priori (așa numiții hiperparametri) în locul calculului integralelor. Am văzut asta pentru repartiția beta în ultimul tabel. Pentru mult mai multe exemple: [http://en.wikipedia.org/wiki/Conjugate\\_prior\\_distribution](http://en.wikipedia.org/wiki/Conjugate_prior_distribution).

## 2 Date continue cu a priori continue

### 2.1 Scopurile învățării

1. Să poată construi un tabel de actualizare Bayesiană pentru ipoteze continue și date continue.
2. Să poată recunoaște pdf a unei repartiții normale și să determine media și dispersia ei.

### 2.2 Introducere

Facem actualizare Bayesiană când atât ipotezele cât și datele iau valori continue. Modelul este același cu ce-am făcut înainte; vom recapitula mai întâi celelalte 2 cazuri.

### 2.3 Cazurile anterioare

#### 2.3.1 Ipoteze discrete, date discrete

##### Notății

Ipoteze  $\mathcal{H}$

Date  $x$

A priori  $P(\mathcal{H})$

Verosimilitatea  $p(x|\mathcal{H})$

A posteriori  $P(\mathcal{H}|x)$ .

**Exemplul 1.** Presupunem că avem datele  $x$  și 3 posibile explicații (ipoteze) pentru date, pe care le vom numi  $A, B, C$ . De asemenea presupunem că datele pot lua 2 valori posibile,  $-1$  și  $1$ .

Pentru a folosi datele pentru estimarea probabilităților diverselor ipoteze, avem nevoie de o pmf a priori și un tabel de verosimilitate. Presupunem că a priori și verosimilitățile sunt date în următoarele tabele.

hypothesis $\mathcal{H}$	prior $P(\mathcal{H})$
A	0.1
B	0.3
C	0.6

hypothesis $\mathcal{H}$	likelihood $p(x \mathcal{H})$	
	$x = -1$	$x = 1$
A	0.2	0.8
B	0.5	0.5
C	0.7	0.3

Probabilități a priori

Verosimilități

Fiecare element din tabelul de verosimilitate este o verosimilitate  $p(x|\mathcal{H})$ . De exemplu,  $p(x = -1|A) = 0.2$ .

**Întrebare:** Presupunem că obținem datele  $x_1 = 1$ . Folosiți aceasta pentru

a obține probabilitățile a posteriori pentru ipoteze.

**Răspuns.** Datele aleg o coloană din tabelul de verosimilități pe care o folosim apoi în tabelul nostru de actualizare Bayesiană.

hypothesis	prior	likelihood	Bayes	
			numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$p(x = 1   \mathcal{H})$	$p(x   \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}   x) = \frac{p(x   \mathcal{H})P(\mathcal{H})}{p(x)}$
$A$	0.1	0.8	0.08	0.195
$B$	0.3	0.5	0.15	0.366
$C$	0.6	0.3	0.18	0.439
total	1		$p(x) = 0.41$	1

Pentru a rezuma: probabilitățile a priori ale ipotezelor și verosimilitățile datelor cunoscând ipotezele au fost date; numărătorul Bayes este produsul dintre a priori și verosimilitate; probabilitatea totală  $p(x)$  este suma probabilităților din coloana numărătorilor Bayes; împărțim la  $p(x)$  pentru a normaliza numărătorii Bayes.

### 2.3.2 Ipoteze continue, date discrete

Acum presupunem că avem datele  $x$  care pot lua o mulțime discretă de valori și un parametru continuu  $\theta$  care determină repartiția din care sunt extrase datele.

#### Notății

Ipoteze  $\theta$

Date  $x$

A priori  $f(\theta)d\theta$

Verosimilitate  $p(x|\theta)$

A posteriori  $f(\theta|x)d\theta$ .

Notă: Am înmulțit cu  $d\theta$  pentru a exprima a priori și a posteriori ca probabilități. Ca densități, avem pdf a priori  $f(\theta)$  și pdf a posteriori  $f(\theta|x)$ .

**Exemplul 2.** Presupunem că  $x \sim \text{Binomial}(5, \theta)$ . Deci  $\theta$  este în domeniul  $[0, 1]$  și datele  $x$  pot lua 6 valori posibile, 0, 1, ..., 5.

Deoarece există un domeniu continuu de valori, folosim o pdf pentru a descrie a priori pentru  $\theta$ . Presupunem că a priori este  $f(\theta) = 2\theta$ . Putem face totuși un tabel de verosimilitate, cu toate că are o singură linie, reprezentând o ipoteză arbitrară  $\theta$ .

hypothesis	likelihood $p(x   \theta)$						
	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	
$\theta$	$\binom{5}{0}(1 - \theta)^5$	$\binom{5}{1}\theta(1 - \theta)^4$	$\binom{5}{2}\theta^2(1 - \theta)^3$	$\binom{5}{3}\theta^3(1 - \theta)^2$	$\binom{5}{4}\theta^4(1 - \theta)$	$\binom{5}{5}\theta^5$	

Verosimilități

**Întrebare** Presupunem că obținem data  $x_1 = 2$ . Folosiți aceasta pentru a afla pdf a posteriori pentru parametrul (ipoteza)  $\theta$ .

**Răspuns.** Ca mai înainte, data alege una din coloanele din tabelul de verosimilități pe care o putem folosi în tabelul nostru de actualizare Bayesiană. Deoarece vrem să lucrăm cu probabilități scriem  $f(\theta)d\theta$  și  $f(\theta|x_1)d\theta$ . Pe ultima linie, în loc de " $\theta^2$ " se va citi " $\theta^3$ ". Pe ultima coloană, în loc de " $\frac{3!3!}{7!}$ " se va citi " $\frac{7!}{3!3!}$ ".

hypothesis	prior	likelihood	Bayes numerator	posterior
$\theta$	$f(\theta)d\theta$	$p(x=2 \theta)$	$p(x \theta)f(\theta)d\theta$	$f(\theta x)d\theta = \frac{p(x \theta)f(\theta)d\theta}{p(x)}$
$\theta$	$2\theta d\theta$	$\binom{5}{2}\theta^2(1-\theta)^3$	$2\binom{5}{2}\theta^3(1-\theta)^3d\theta$	$f(\theta x)d\theta = \frac{3!3!}{7!}\theta^3(1-\theta)^3d\theta$
total	1	$p(x) = \int_0^1 2\binom{5}{2}\theta^2(1-\theta)^3d\theta = 2\binom{5}{2}\frac{3!3!}{7!}$		1

Pentru a rezuma: probabilitățile a priori ale ipotezelor și verosimilitățile datelor cunoscând ipotezele sunt date; numărătorul Bayes este produsul dintre a priori și verosimilitate; probabilitatea totală  $p(x)$  este integrala probabilităților din coloana numărătorului Bayes; împărțim prin  $p(x)$  pentru a normaliza numărătorul Bayes.

## 2.4 Ipoteze continue și date continue

Când atât datele și ipotezele sunt continue, singura schimbare în exemplul anterior este că funcția de verosimilitate folosește o pdf  $f(x|\theta)$  în locul unei pmf  $p(x|\theta)$ . Forma generală a tabelului de actualizare Bayesiană este aceeași.

### Notății

Ipoteze  $\theta$

Date  $x$

A priori  $f(\theta)d\theta$

Verosimilitate  $f(x|\theta)dx$

A posteriori  $f(\theta|x)d\theta$ .

**Simplificarea notației.** În cazurile precedente am inclus  $d\theta$  astfel că lucram cu probabilități în loc de densități. Când atât datele cât și ipotezele sunt continue, vom avea nevoie atât de  $d\theta$  cât și de  $dx$ . Aceasta face lucrurile mai simple conceptual, dar mai greoale notațional. Pentru a simplifica notația ne vom permite să eliminăm  $dx$  din tabelele noastre. Aceasta este în regulă, deoarece datele  $x$  sunt fixate. Vom păstra  $d\theta$  deoarece ipotezei  $\theta$  i se permite să varieze.

Pentru comparație, întâi arătăm tabelul general cu notație simplificată, urmat imediat apoi de tabelul arătând infinitezimalele. La primul tabel, pe ultima coloană, în loc de ” $f(\theta|x)$ ” se va citi ” $f(\theta|x)d\theta$ ”.

			Bayes	
hypoth.	prior	likelihood	numerator	posterior
$\theta$	$f(\theta) d\theta$	$f(x \theta)$	$f(x \theta)f(\theta) d\theta$	$f(\theta x) = \frac{f(x \theta)f(\theta) d\theta}{f(x)}$
total	1		$f(x) = \int f(x \theta)f(\theta) d\theta$	1

Tabel de actualizare Bayesiană fără  $dx$

			Bayes	
hypoth.	prior	likelihood	numerator	posterior
$\theta$	$f(\theta) d\theta$	$f(x \theta) dx$	$f(x \theta)f(\theta) d\theta dx$	$f(\theta x) d\theta = \frac{f(x \theta)f(\theta) d\theta dx}{f(x) dx} = \frac{f(x \theta)f(\theta) d\theta}{f(x)}$
total	1		$f(x) dx = (\int f(x \theta)f(\theta) d\theta) dx$	1

Tabel de actualizare Bayesiană cu  $d\theta$  și  $dx$

Pentru a rezuma: probabilitățile a priori ale ipotezelor și verosimilitățile datelor cunoscând ipotezele erau date; numărătorul Bayes este produsul dintre a priori și verosimilitate; probabilitatea totală  $f(x)dx$  este integrala probabilităților din coloana numărătorului Bayes; împărțim prin  $f(x)dx$  pentru a normaliza numărătorul Bayes.

## 2.5 Ipoteză normală, date normale

Un exemplu standard de ipoteze continue și date continue presupune că atât datele cât și a priori au repartiții normale. Următorul exemplu presupune că dispersia datelor este cunoscută.

**Exemplul 3.** Presupunem că avem data  $x = 5$  care a fost extrasă dintr-o repartiție normală cu medie necunoscută  $\theta$  și deviație standard 1.

$$x \sim N(\theta, 1).$$

Presupunem mai departe că repartiția noastră a priori pentru  $\theta$  este  $\theta \sim N(2, 1)$ .

Fie  $x$  o valoare arbitrară a datelor.

- a) Faceți un tabel Bayesian cu a priori, verosimilitate și numărător Bayes.
- b) Arătați că repartiția a posteriori pentru  $\theta$  este tot normală.

c) Aflați media și dispersia repartiției a posteriori.

**Răspuns.** Cum am făcut cu tabelele de mai sus, un bun compromis asupra notației este să includem  $d\theta$ , dar nu  $dx$ . Motivul pentru aceasta este că probabilitatea totală este calculată integrând după  $\theta$  și  $d\theta$  ne reamintește asta.

Pdf a priori a noastră este

$$f(\theta) = \frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2}.$$

Funcția de verosimilitate este

$$f(x = 5 | \theta) = \frac{1}{\sqrt{2\pi}} e^{-(5-\theta)^2/2}.$$

Înmulțim a priori cu verosimilitatea.

$$\begin{aligned} \text{a priori} \cdot \text{verosimilitatea} &= \frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(5-\theta)^2/2} \\ &= \frac{1}{2\pi} e^{-(2\theta^2 - 14\theta + 29)/2} \\ &= \frac{1}{2\pi} e^{-(\theta^2 - 7\theta + 29/2)} \text{ (completăm pătratul)} \\ &= \frac{1}{2\pi} e^{-((\theta-7/2)^2 + 9/4)} \\ &= \frac{e^{-\frac{9}{4}}}{2\pi} e^{-(\theta-7/2)^2} \\ &= c_1 e^{-(\theta-7/2)^2}. \end{aligned}$$

La ultimul pas am înlocuit factorul constant complicat cu expresia mai simplă  $c_1$ . Pe linia a 3-a a tabelului, în loc de ” $e^{-(\theta-7/2)^2}$ ” se va citi ” $e^{-(\theta-7/2)^2} d\theta$ ” (de 2 ori).

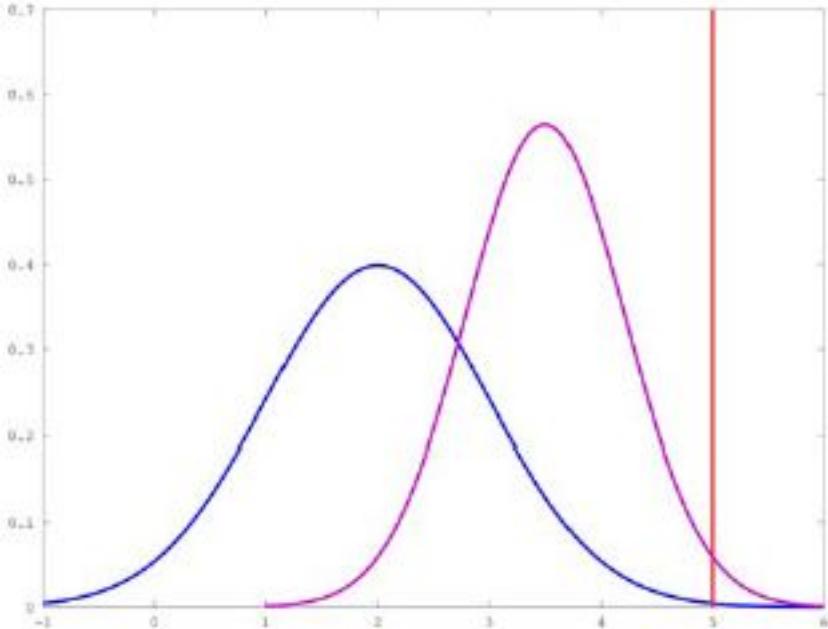
hypothesis	prior	likelihood	Bayes numerator	posterior
$\theta$	$f(\theta) d\theta$	$f(x = 5   \theta)$	$f(x = 5   \theta) f(\theta) d\theta$	$\frac{f(x = 5   \theta) f(\theta) d\theta}{f(x = 5)}$
$\theta$	$\frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2} d\theta$	$\frac{1}{\sqrt{2\pi}} e^{-(5-\theta)^2/2}$	$c_1 e^{-(\theta-7/2)^2}$	$c_2 e^{-(\theta-7/2)^2}$
total	1	$f(x = 5) = \int f(x = 5   \theta) f(\theta) d\theta$		1

Pdf a posteriori este cea a unei repartiții normale. Deoarece exponențiala unei repartiții normale este  $e^{-(\theta-\mu)^2/2\sigma^2}$ , avem media  $\mu = 7/2$  și  $2\sigma^2 = 1$ , deci

dispersia este  $\sigma^2 = 1/2$ .

Nu trebuie să calculăm probabilitatea totală; ea este folosită doar pentru normalizare și știm deja constanta de normalizare  $\frac{1}{\sigma\sqrt{2\pi}}$  pentru o repartiție normală.

Iată graficele pdf-urilor a priori și a posteriori. Observăm cum data "trage" a priori spre ea.



a priori = albastru; a posteriori = mov; data = roșu

Acum, repetăm exemplul precedent pentru  $x$  general.

**Exemplul 4.** Presupunem că data noastră  $x$  este extrasă dintr-o repartiție normală cu medie necunoscută  $\theta$  și deviație standard 1.

$$x \sim N(\theta, 1).$$

**Răspuns.** Pdf a priori și funcția de verosimilitate sunt

$$f(\theta) = \frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2}, \quad f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}.$$

Numărătorul Bayes este produsul dintre a priori și verosimilitate:

$$\begin{aligned}
 \text{a priori} \cdot \text{verosimilitatea} &= \frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} \\
 &= \frac{1}{2\pi} e^{-(2\theta^2 - (4+2x)\theta + 4+x^2)/2} \\
 &= \frac{1}{2\pi} e^{-(\theta^2 - (2+x)\theta + (4+x^2)/2)} \quad (\text{completăm pătratul}) \\
 &= \frac{1}{2\pi} e^{-((\theta-(1+x/2))^2 - (1+x/2)^2 + (4+x^2)/2)} \\
 &= c_1 e^{-(\theta-(1+x/2))^2}.
 \end{aligned}$$

Ca în ultimul exemplu, în ultimul pas am înlocuit toate constantele, inclusiv exponentialele care implică doar pe  $x$ , prin simpla constantă  $c_1$ .

Acum, tabelul Bayesian de înlocuire devine (pe linia a 3-a a tabelului, în loc de ” $e^{-(\theta-(1+x/2))^2}$ ” se va citi ” $e^{-(\theta-(1+x/2))^2} d\theta$ ” (de 2 ori)).

hypothesis	prior	likelihood	Bayes numerator	posterior $f(\theta   x) d\theta$
$\theta$	$f(\theta) d\theta$	$f(x   \theta)$	$f(x   \theta) f(\theta) d\theta$	$\frac{f(x   \theta) f(\theta) d\theta}{f(x)}$
$\theta$	$\frac{1}{\sqrt{2\pi}} e^{-(\theta-2)^2/2} d\theta$	$\frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$	$c_1 e^{-(\theta-(1+x/2))^2}$	$c_2 e^{-(\theta-(1+x/2))^2}$
total	1		$f(x) = f(x   \theta) f(\theta) d\theta$	1

Ca în exemplul precedent putem vedea din forma a posteriori că ea trebuie să fie a unei repartiții normale cu media  $1+x/2$  și dispersia  $1/2$ . (Comparați aceasta cu cazul  $x = 5$  din exemplul precedent.)

## 2.6 Probabilități predictive

Deoarece datele  $x$  sunt continue, au pdf-uri predictive a priori și a posteriori. **Pdf predictivă a priori** este densitatea de probabilitate totală calculată la marginea de jos a coloanei numărătorului Bayes:

$$f(x) = \int f(x|\theta) f(\theta) d\theta,$$

unde integrala este calculată pe întregul domeniu al lui  $\theta$ .

**Pdf predictivă a posteriori** are aceeași formă ca pdf predictivă a priori, cu excepția faptului că folosește probabilitățile a posteriori pentru  $\theta$ :

$$f(x_2|x_1) = \int f(x_2|\theta, x_1) f(\theta|x_1) d\theta.$$

Ca de obicei, presupunem că  $x_1$  și  $x_2$  sunt **condiționat independente**. Adică,

$$f(x_2|\theta, x_1) = f(x_2|\theta).$$

În acest caz formula pentru pdf predictivă a posteriori este un pic mai simplă:

$$f(x_2|x_1) = \int f(x_2|\theta)f(\theta|x_1)d\theta.$$

# Curs 14

Cristian Niculescu

## 1 A priori conjugate: Beta și normală

### 1.1 Scopurile învățării

1. Să înțeleagă beneficiile a priori conjugate.
2. Să poată să actualizeze o a priori beta dată fiind o verosimilitate Bernoulli, binomială sau geometrică.
3. Să înțeleagă și să poată folosi formula pentru actualizarea unei a priori normale fiind data o verosimilitate normală cu dispersie cunoscută.

### 1.2 Introducere și definiție

Cu o a priori conjugată, a posteriori este de același tip, de exemplu pentru verosimilitate binomială, a priori beta devine o a posteriori beta. A priori conjugate sunt utile deoarece ele reduc actualizarea Bayesiană la modificarea parametrilor repartiției a priori (așa numiții hiperparametri) în locul calculului de integrale.

Ne vom concentra pe 2 exemple importante de a priori conjugate: beta și normală. O listă mult mai cuprinzătoare este în tabelele din [http://en.wikipedia.org/wiki/Conjugate\\_prior\\_distribution](http://en.wikipedia.org/wiki/Conjugate_prior_distribution).

**Definiție.** Presupunem că avem date cu funcția de verosimilitate  $f(x|\theta)$  depinzând de un parametru  $\theta$ . Mai presupunem că repartiția a priori pentru  $\theta$  este una dintr-o familie de repartiții parametrizate. Dacă repartiția a posteriori pentru  $\theta$  este în aceeași familie ca repartiția a priori, spunem că a priori este o [a priori conjugată](#) pentru verosimilitate.

### 1.3 Repartiția beta

Arătăm că repartiția beta este o a priori conjugată pentru verosimilități binomiale, Bernoulli și geometrice.

### 1.3.1 Verosimilitate binomială

Am văzut că repartiția beta este o a priori conjugată pentru repartiția binomială. Aceasta înseamnă că dacă funcția de verosimilitate este binomială și repartiția a priori este beta, atunci repartiția a posteriori este tot beta.

Mai concret, presupunem că funcția de verosimilitate are o repartiție binomială( $N, \theta$ ), unde  $N$  este cunoscut și  $\theta$  este parametrul (necunoscut) de interes. Avem de asemenea că data  $x$  este un întreg între 0 și  $N$ . Atunci, pentru o a priori beta avem următorul tabel:

hypothesis	data	prior	likelihood	posterior
$\theta$	$x$	$\text{beta}(a, b)$	$\text{binomial}(N, \theta)$	$\text{beta}(a + x, b + N - x)$
$\theta$	$x$	$c_1 \theta^{a-1} (1-\theta)^{b-1}$	$c_2 \theta^x (1-\theta)^{N-x}$	$c_3 \theta^{a+x-1} (1-\theta)^{b+N-x-1}$

Tabelul este simplificat scriind coeficienții de normalizare ca  $c_1, c_2$  și  $c_3$ .

$$c_1 = \frac{(a+b-1)!}{(a-1)!(b-1)!}, \quad c_2 = C_N^x = \frac{N!}{x!(N-x)!}, \quad c_3 = \frac{(a+b+N-1)!}{(a+x-1)!(b+N-x-1)!}.$$

### 1.3.2 Verosimilitate Bernoulli

Repartiția beta este o a priori conjugată pentru repartiția Bernoulli. Acesta este de fapt un caz special al repartiției binomiale, deoarece Bernoulli( $\theta$ ) este aceeași ca binomiala( $1, \theta$ ). În tabelul de mai jos, arătăm actualizările corespunzând succesului ( $x = 1$ ) și eșecului ( $x = 0$ ) pe linii separate.

hypothesis	data	prior	likelihood	posterior
$\theta$	$x$	$\text{beta}(a, b)$	$\text{Bernoulli}(\theta)$	$\text{beta}(a+1, b)$ or $\text{beta}(a, b+1)$
$\theta$	$x = 1$	$c_1 \theta^{a-1} (1-\theta)^{b-1}$	$\theta$	$c_3 \theta^a (1-\theta)^{b-1}$
$\theta$	$x = 0$	$c_1 \theta^{a-1} (1-\theta)^{b-1}$	$1-\theta$	$c_3 \theta^{a-1} (1-\theta)^b$

Constantele  $c_1$  și  $c_3$  au aceleași formule ca în cazul precedent (al verosimilității binomiale) cu  $N = 1$ .

### 1.3.3 Verosimilitate geometrică

Repartiția geometrică( $\theta$ ) descrie probabilitatea a  $x$  eșecuri înaintea primului succes, unde probabilitatea succesului în fiecare încercare independentă este  $\theta$ . Pmf corespunzătoare este  $p(x) = \theta(1-\theta)^x$ .

Acum presupunem că avem o dată  $x$  și ipoteza noastră  $\theta$  este că  $x$  este extrasă dintr-o repartiție geometrică( $\theta$ ). Din tabel vedem că repartiția beta este o a priori conjugată pentru o verosimilitate geometrică:

ipoteza	data	a priori	verosimilitatea	a posteriori
$\theta$	$x$	$\text{beta}(a, b)$	$\text{geometrică}(\theta)$	$\text{beta}(a+1, b+x)$
$\theta$	$x$	$c_1 \theta^{a-1} (1-\theta)^{b-1}$	$\theta(1-\theta)^x$	$c_3 \theta^a (1-\theta)^{b+x-1}$

**Exemplul 1.** În timp ce călătoreau prin Regatul Ciupercilor, Mario și Luigi

au găsit niște monede neobișnuite. Ei au căzut de acord asupra unei a priori  $f(\theta) \sim \text{beta}(5, 5)$  pentru probabilitatea aversului, dar nu au fost de acord ce experiment să facă pentru a investiga  $\theta$ .

- a) Mario decide să arunce o monedă de 5 ori. El obține un avers în 5 aruncări.
- b) Luigi decide să arunce o monedă până la primul avers. El obține 4 reversuri înaintea primului avers.

Arătați că Mario și Luigi ajung la aceeași a posteriori pentru  $\theta$  și calculați această a posteriori.

**Răspuns.** Tabelul lui Mario:

ipoteza	data	a priori	verosimilitatea	a posteriori
$\theta$	$x = 1$	beta(5, 5)	binomială( $\theta$ )	???
$\theta$	$x = 1$	$c_1\theta^4(1 - \theta)^4$	$C_5^1\theta(1 - \theta)^4$	$c_3\theta^5(1 - \theta)^8$

Tabelul lui Luigi:

ipoteza	data	a priori	verosimilitatea	a posteriori
$\theta$	$x = 4$	beta(5, 5)	geometrică( $\theta$ )	???
$\theta$	$x = 4$	$c_1\theta^4(1 - \theta)^4$	$\theta(1 - \theta)^4$	$c_3\theta^5(1 - \theta)^8$

Atât a posteriori a lui Mario cât și a lui Luigi au forma unei repartiții beta(6, 9). Factorul de normalizare este același în ambele cazuri deoarece este determinat cerând ca probabilitatea totală să fie 1.

## 1.4 Normala generează normală

Repartiția normală este a priori conjugată cu ea însăși. În particular, dacă funcția de verosimilitate este normală cu dispersie cunoscută, atunci o a priori normală dă o a posteriori normală. Acum atât ipotezele și datele sunt continue.

Presupunem că avem o măsurare  $x \sim N(\theta, \sigma^2)$ , unde dispersia  $\sigma^2$  este cunoscută. Adică, media  $\theta$  este parametrul nostru necunoscut de interes și știm că verosimilitatea vine dintr-o repartiție normală cu dispersia  $\sigma^2$ . Dacă alegem o pdf a priori normală

$$f(\theta) \sim N(\mu_{\text{prior}}, \sigma_{\text{prior}}^2),$$

atunci pdf a posteriori este de asemenea normală:  $f(\theta|x) \sim N(\mu_{\text{post}}, \sigma_{\text{post}}^2)$ , unde

$$\frac{\mu_{\text{post}}}{\sigma_{\text{post}}^2} = \frac{\mu_{\text{prior}}}{\sigma_{\text{prior}}^2} + \frac{x}{\sigma^2}, \quad \frac{1}{\sigma_{\text{post}}^2} = \frac{1}{\sigma_{\text{prior}}^2} + \frac{1}{\sigma^2}. \quad (1)$$

Următoarea formă a acestor formule este mai ușor de citit și arată că  $\mu_{\text{post}}$  este o medie ponderată între  $\mu_{\text{prior}}$  și data  $x$ .

$$a = \frac{1}{\sigma_{\text{prior}}^2}, \quad b = \frac{1}{\sigma^2}, \quad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a + b}, \quad \sigma_{\text{post}}^2 = \frac{1}{a + b}. \quad (2)$$

Cu aceste formule în minte, putem exprima actualizarea prin tabelul:

hypothesis	data	prior	likelihood	posterior
$\theta$	$x$	$f(\theta) \sim N(\mu_{\text{prior}}, \sigma_{\text{prior}}^2)$	$f(x \theta) \sim N(\theta, \sigma^2)$	$f(\theta x) \sim N(\mu_{\text{post}}, \sigma_{\text{post}}^2)$
$\theta$	$x$	$c_1 \exp\left(\frac{-(\theta - \mu_{\text{prior}})^2}{2\sigma_{\text{prior}}^2}\right)$	$c_2 \exp\left(\frac{-(x - \theta)^2}{2\sigma^2}\right)$	$c_3 \exp\left(\frac{-(\theta - \mu_{\text{post}})^2}{2\sigma_{\text{post}}^2}\right)$

Demonstrația formulelor generale se face analog ca în următorul exemplu numeric.

**Exemplul 2.** Presupunem că avem a priori  $\theta \sim N(4, 8)$  și verosimilitatea  $x \sim N(\theta, 5)$ . Presupunem de asemenea că avem o măsurare  $x_1 = 3$ . Arătați că repartiția a posteriori este normală.

**Răspuns.**

a priori:  $f(\theta) = c_1 e^{-(\theta-4)^2/16}$ ; verosimilitatea:  $f(x_1|\theta) = c_2 e^{-(x_1-\theta)^2/10} = c_2 e^{-(3-\theta)^2/10}$ .

Înmulțim a priori cu verosimilitatea pentru a obține a posteriori:

$$f(\theta|x_1) = c_1 c_2 e^{-(\theta-4)^2/16} e^{-(3-\theta)^2/10} = c_1 c_2 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right).$$

Completăm pătratul din exponent

$$\begin{aligned} -\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} &= -\frac{5(\theta-4)^2 + 8(3-\theta)^2}{80} \\ &= -\frac{13\theta^2 - 88\theta + 152}{80} \\ &= -\frac{\theta^2 - \frac{88}{13}\theta + \frac{152}{13}}{80/13} \\ &= -\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}. \end{aligned}$$

De aceea, a posteriori este

$$f(\theta|x_1) = c_1 c_2 e^{-\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}} = c_3 e^{-\frac{(\theta - 44/13)^2}{80/13}}.$$

Aceasta are forma pdf pentru  $N(44/13, 40/13)$ , q.e.d.

Verificăm aceasta cu formulule (2).

$$\mu_{\text{prior}} = 4, \sigma_{\text{prior}}^2 = 8, \sigma^2 = 5 \implies a = \frac{1}{8}, b = \frac{1}{5}.$$

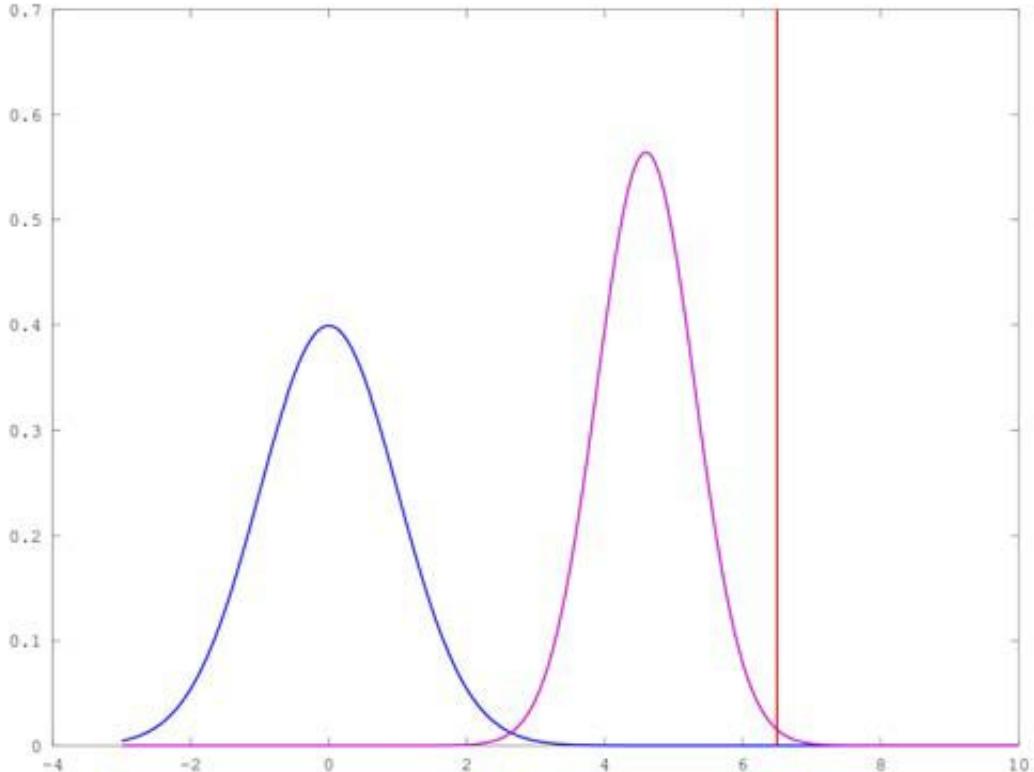
De aceea

$$\mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{\frac{1}{8} \cdot 4 + \frac{1}{5} \cdot 3}{\frac{1}{8} + \frac{1}{5}} = \frac{44}{13} \approx 3.38,$$

$$\sigma_{\text{post}}^2 = \frac{1}{a+b} = \frac{1}{\frac{1}{8} + \frac{1}{5}} = \frac{40}{13} \approx 3.08.$$

**Exemplul 3.** Presupunem că știm datele  $x \sim N(\theta, \sigma^2)$  și avem a priori  $N(0, 1)$ . Obținem o valoare a datelor  $x = 6.5$ . Descrieți schimbările pdf pentru  $\theta$  în actualizarea de la a priori la a posteriori.

**Răspuns.** Iată graficul pdf-urilor a priori, a posteriori cu data marcată cu o linie roșie.



A priori în albastru, a posteriori în mov, data în roșu.

Media a posteriori va fi o medie ponderată dintre media a priori și dată.  
Vârful pdf a posteriori va fi între vârful a priori și linia roșie. Avem

$$\sigma_{\text{post}}^2 = \frac{1}{1/\sigma_{\text{prior}}^2 + 1/\sigma^2} = \sigma_{\text{prior}}^2 \cdot \frac{\sigma^2}{\sigma_{\text{prior}}^2 + \sigma^2} < \sigma_{\text{prior}}^2.$$

Adică a posteriori are dispersie mai mică decât a priori, i.e. data ne face mai siguri despre unde este  $\theta$  în domeniul său.

#### 1.4.1 Mai mult de o dată

**Exemplul 4.** Presupunem că avem datele  $x_1, x_2, x_3$ . Folosiți formulele (1) pentru a actualiza succesiv.

**Răspuns.** Notăm media și dispersia a priori cu  $\mu_0$ , respectiv  $\sigma_0^2$ . Mediile și

dispersiile actualizate vor fi  $\mu_i$ , respectiv  $\sigma_i$ . Avem succesiv

$$\begin{aligned}\frac{1}{\sigma_1^2} &= \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}; \quad \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1}{\sigma^2} \\ \frac{1}{\sigma_2^2} &= \frac{1}{\sigma_1^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{2}{\sigma^2}; \quad \frac{\mu_2}{\sigma_2^2} = \frac{\mu_1}{\sigma_1^2} + \frac{x_2}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1 + x_2}{\sigma^2} \\ \frac{1}{\sigma_3^2} &= \frac{1}{\sigma_2^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{3}{\sigma^2}; \quad \frac{\mu_3}{\sigma_3^2} = \frac{\mu_2}{\sigma_2^2} + \frac{x_3}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1 + x_2 + x_3}{\sigma^2}.\end{aligned}$$

Exemplul se generalizează la  $n$  valori ale datelor  $x_1, \dots, x_n$ :

### Formule de actualizare normală-normală pentru $n$ date

$$\frac{\mu_{\text{post}}}{\sigma_{\text{post}}^2} = \frac{\mu_{\text{prior}}}{\sigma_{\text{prior}}^2} + \frac{n\bar{x}}{\sigma^2}, \quad \frac{1}{\sigma_{\text{post}}^2} = \frac{1}{\sigma_{\text{prior}}^2} + \frac{n}{\sigma^2}, \quad \bar{x} = \frac{x_1 + \dots + x_n}{n}. \quad (3)$$

Din nou dăm o formă mai simplu de citit, arătând că  $\mu_{\text{post}}$  este o medie ponderată între  $\mu_{\text{prior}}$  și media de selecție  $\bar{x}$ :

$$a = \frac{1}{\sigma_{\text{prior}}^2}, \quad b = \frac{n}{\sigma^2}, \quad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + b\bar{x}}{a + b}, \quad \sigma_{\text{post}}^2 = \frac{1}{a + b}. \quad (4)$$

**Interpretare:**  $\mu_{\text{post}}$  este o medie ponderată între  $\mu_{\text{prior}}$  și  $\bar{x}$ . Dacă numărul datelor este mare, atunci ponderea  $b$  este mare și  $\bar{x}$  va avea o puternică influență asupra a posteriori. Dacă  $\sigma_{\text{prior}}^2$  este mică, atunci ponderea  $a$  este mare și  $\mu_{\text{prior}}$  va avea o puternică influență asupra a posteriori. Pentru a rezuma:

1. Multe date au o mare influență asupra a posteriori.
2. Siguranța mare (dispersia mică) în a priori are o mare influență asupra a posteriori.

## 2 Alegerea a priori

### 2.1 Scopurile învățării

1. Să învețe că alegerea a priori afectează a posteriori.
2. Să vadă că o a priori prea rigidă poate face dificilă folosirea datelor.
3. Să vadă că mai multe date scad dependența a posteriori de a priori.
4. Să poată face o alegere rezonabilă a a priori, bazată pe înțelegerea a priori a sistemului considerat.

### 2.2 Introducere

Până acum niciun pdf sau pmf a priori nu a fost dat. În acest caz, deducția statistică din date este în esență o aplicație a teoremei lui Bayes. Când a

priori este cunoscută, nu sunt controverse despre cum trebuie procedat. Arta statisticii începe când a priori nu este cunoscută cu siguranță. Sunt 2 școli principale despre cum să procedăm în acest caz: **Bayesiană** și **frecvenționistă**. Acum urmăram abordarea Bayesiană. Vom învăța și abordarea frecvenționistă. Reamintim că fiind cunoscute datele  $D$  și ipoteza  $H$  am folosit teorema lui Bayes pentru a scrie

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

a posteriori  $\propto$  verosimilitate  $\cdot$  a priori

**Bayesiană:** Bayesianii fac deducții folosind a posteriori  $P(H|D)$  și de aceea au nevoie totdeauna de o a priori  $P(H)$ . Dacă a priori nu este cunoscută cu certitudine, Bayesianul trebuie să încerce să facă o alegere rezonabilă. Sunt multe feluri de a face asta și oameni rezonabili pot face alegeri diferite. În general este o practică bună justificarea alegerii și explorarea unui domeniu de a priori pentru a vedea dacă toate indică aceeași concluzie.

**Frecvenționistă:** Foarte scurt, frecvenționiștii nu încearcă să creeze o a priori. În schimb, ei fac deducții folosind verosimilitatea  $P(D|H)$ .

2 beneficii ale abordării Bayesiene:

1. Probabilitatea a posteriori  $P(H|D)$  pentru ipoteză date fiind dovezile este de obicei exact ce am vrea să știm. Bayesianul poate spune ceva ca "parametrul de interes are probabilitatea 0.95 de a fi între 0.49 și 0.51".
2. Presupunerile care se iau în considerare pentru alegerea a priori pot fi clar precizate.

**Mai multe date bune:** Totdeauna **mai multe date bune** permit concluzii mai puternice și scad influența a priori. Accentul ar trebui să fie atât pe date bune (calitate) cât și pe mai multe date (cantitate).

## 2.3 Exemplu: zaruri

Presupunem că avem un sertar plin de zaruri, fiecare dintre acestea având 4, 6, 8, 12 sau 20 de fețe. De această dată nu știm câte zaruri de fiecare tip sunt în sertar. Un zar este ales la întâmplare din sertar și aruncat de 5 ori. Rezultatele sunt în ordine 4, 2, 4, 7 și 5.

### 2.3.1 A priori uniformă

Presupunem că nu avem idee care poate fi repartitia zarurilor din sertar. În acest caz este rezonabil să folosim o a priori plată. Iată tabelul de actualizare pentru probabilitățile a posteriori care rezultă din actualizarea după fiecare

aruncare. Pentru a încăpea toate coloanele, am eliminat a posteriori nenormalizate.

hyp.	prior	lik <sub>1</sub>	post <sub>1</sub>	lik <sub>2</sub>	post <sub>2</sub>	lik <sub>3</sub>	post <sub>3</sub>	lik <sub>4</sub>	post <sub>4</sub>	lik <sub>5</sub>	post <sub>5</sub>
$H_4$	1/5	1/4	0.370	1/4	0.542	1/4	0.682	0	0.000	0	0.000
$H_6$	1/5	1/6	0.247	1/6	0.241	1/6	0.202	0	0.000	1/6	0.000
$H_8$	1/5	1/8	0.185	1/8	0.135	1/8	0.085	1/8	0.818	1/8	0.876
$H_{12}$	1/5	1/12	0.123	1/12	0.060	1/12	0.025	1/12	0.161	1/12	0.115
$H_{20}$	1/5	1/20	0.074	1/20	0.022	1/20	0.005	1/20	0.021	1/20	0.009

Cunoscând datele, a posteriori finală este puternic ponderată spre ipoteza  $H_8$  că a fost ales un zar cu 8 fețe.

### 2.3.2 Alte a priori

Pentru a vedea cât de mult a posteriori de mai sus depinde de alegerea noastră a a priori, încercăm alte a priori. Presupunem că avem un motiv de a crede că sunt de 10 ori mai multe zaruri cu 20 de fețe în sertar decât de fiecare alt tip. Tabelul devine:

hyp.	prior	lik <sub>1</sub>	post <sub>1</sub>	lik <sub>2</sub>	post <sub>2</sub>	lik <sub>3</sub>	post <sub>3</sub>	lik <sub>4</sub>	post <sub>4</sub>	lik <sub>5</sub>	post <sub>5</sub>
$H_4$	0.071	1/4	0.222	1/4	0.453	1/4	0.650	0	0.000	0	0.000
$H_6$	0.071	1/6	0.148	1/6	0.202	1/6	0.193	0	0.000	1/6	0.000
$H_8$	0.071	1/8	0.111	1/8	0.113	1/8	0.081	1/8	0.688	1/8	0.810
$H_{12}$	0.071	1/12	0.074	1/12	0.050	1/12	0.024	1/12	0.136	1/12	0.107
$H_{20}$	0.714	1/20	0.444	1/20	0.181	1/20	0.052	1/20	0.176	1/20	0.083

Chiar și aici a posteriori finală este puternic ponderată spre ipoteza  $H_8$ .

Dar dacă zarurile cu 20 de fețe sunt de 100 de ori mai probabile decât fiecare din celelalte?

hyp.	prior	lik <sub>1</sub>	post <sub>1</sub>	lik <sub>2</sub>	post <sub>2</sub>	lik <sub>3</sub>	post <sub>3</sub>	lik <sub>4</sub>	post <sub>4</sub>	lik <sub>5</sub>	post <sub>5</sub>
$H_4$	0.0096	1/4	0.044	1/4	0.172	1/4	0.443	0	0.000	0	0.000
$H_6$	0.0096	1/6	0.030	1/6	0.077	1/6	0.131	0	0.000	1/6	0.000
$H_8$	0.0096	1/8	0.022	1/8	0.043	1/8	0.055	1/8	0.266	1/8	0.464
$H_{12}$	0.0096	1/12	0.015	1/12	0.019	1/12	0.016	1/12	0.053	1/12	0.061
$H_{20}$	0.9615	1/20	0.889	1/20	0.689	1/20	0.354	1/20	0.681	1/20	0.475

Cu o astfel de convingere a priori puternică în zarurile cu 20 de fețe, a posteriori finală dă o mare pondere teoriei că datele sunt dintr-un zar cu 20 de fețe, chiar dacă este extrem de improbabil ca un zar cu 20 de fețe să producă un maxim de 7 în 5 aruncări. A posteriori dă acum șanse aproximativ egale ca să fi fost ales un zar cu 8 fețe versus un zar cu 20 de fețe.

### 2.3.3 A priori rigide

**Disonanță cognitivă ușoară.** O convingere a priori prea rigidă poate copleși orice cantitate de date. Presupunem că suntem convinși că zarul trebuie să fie cu 20 de fețe. Deci punem a priori a noastră  $P(H_{20}) = 1$  cu

celealte 4 ipoteze având probabilitatea 0. Iată ce se întâmplă în tabelul de actualizare.

hyp.	prior	lik <sub>1</sub>	post <sub>1</sub>	lik <sub>2</sub>	post <sub>2</sub>	lik <sub>3</sub>	post <sub>3</sub>	lik <sub>4</sub>	post <sub>4</sub>	lik <sub>5</sub>	post <sub>5</sub>
$H_4$	0	1/4	0	1/4	0	1/4	0	0	0	0	0
$H_6$	0	1/6	0	1/6	0	1/6	0	0	0	1/6	0
$H_8$	0	1/8	0	1/8	0	1/8	0	1/8	0	1/8	0
$H_{12}$	0	1/12	0	1/12	0	1/12	0	1/12	0	1/12	0
$H_{20}$	1	1/20	1	1/20	1	1/20	1	1/20	1	1/20	1

Indiferent care sunt datele, o ipoteză cu probabilitatea a priori 0 va avea probabilitatea a posteriori 0. În acest caz nu vom scăpa niciodată de ipoteza  $H_{20}$ , cu toate că putem experimenta o ușoară disonanță cognitivă.

**Disonanță cognitivă severă.** A priori rigide pot de asemenea duce la absurdități. Presupunem că suntem convinși că zarul trebuie să fie cu 4 fețe. Deci punem  $P(H_4) = 1$  și celealte probabilități a priori 0. Cu datele cunoscute, la a 4-a aruncare intrăm în impas. 7 nu poate veni de la un zar cu 4 fețe. Totuși, aceasta este singura ipoteză pe care o permitem. A posteriori a noastră nenormalizată este o coloană de zerouri care nu poate fi normalizată.

hyp.	prior	lik <sub>1</sub>	post <sub>1</sub>	lik <sub>2</sub>	post <sub>2</sub>	lik <sub>3</sub>	post <sub>3</sub>	lik <sub>4</sub>	unnorm.	post <sub>4</sub>	post <sub>4</sub>
$H_4$	1	1/4	1	1/4	1	1/4	1	0	0	0	???
$H_6$	0	1/6	0	1/6	0	1/6	0	0	0	0	???
$H_8$	0	1/8	0	1/8	0	1/8	0	1/8	0	0	???
$H_{12}$	0	1/12	0	1/12	0	1/12	0	1/12	0	0	???
$H_{20}$	0	1/20	0	1/20	0	1/20	0	1/20	0	0	???

Trebuie să ne ajustăm convingerile despre ce este posibil sau, mai probabil, suspectăm o greșală accidentală sau deliberată a datelor.

## 2.4 Exemplu: malaria

Iată un exemplu real adaptat din *Statistics, A Bayesian Perspective* de Donald Berry:

Prin anii 1950, oamenii de știință au început să formuleze ipoteza că purtătorii genei falciforme erau mai rezistenți la malarie ca nepurtătorii. Existau probe indirecte pentru această ipoteză. Aceasta ajuta și la explicarea persistenței în populație a unei gene altfel dăunătoare. Într-un experiment oamenii de știință au injectat 30 de voluntari africani cu malarie. 15 dintre voluntari purtau o copie a genei falciforme și ceilalți 15 erau nepurtători. 14 din cei 15 nepurtători și doar 2 din cei 15 purtători au făcut malarie. Susține acest mic eșantion ipoteza că gena falciformă protejează împotriva malariei?

Fie  $S$  un purtător al genei falciforme și  $N$  un nepurtător.  $D+$  indică dezvoltarea malariei și  $D-$  indică nedezvoltarea malariei. Datele pot fi puse într-un tabel.

	$D+$	$D-$	
$S$	2	13	15
$N$	14	1	15
	16	14	30

Înainte să analizăm datele ar trebui să spunem câteva cuvinte despre experiment și proiectarea lui. În primul rând, este clar neetic: pentru a obține ceva informație au infectat 16 oameni cu malarie. Trebuie de asemenea să ne îngrijorăm despre deplasare. Cum au ales subiecții testului? Este posibil ca nepurtătorii să fi fost mai slabii și astfel mai susceptibili la malarie decât purtătorii? Berry arată că este rezonabil să presupunem că o injecție este similară cu o mușcătură de țânțar, dar nu este garantat. Acest ultim punct înseamnă că dacă experimentul arată o relație între celule-seceră și protecție împotriva malariei injectate, trebuie să considerăm ipoteza că protecția împotriva malariei transmisă de țânțari este mai slabă sau inexistentă. În sfârșit, vom formula ipoteza noastră ca ”celulele-seceră protejează împotriva malariei”, dar în realitate tot ce putem spera să spunem dintr-un studiu ca acesta este că ”celula-seceră este corelată cu protecția împotriva malariei”.

**Modelul.** Pentru modelul nostru, fie  $\theta_S$  probabilitatea că un purtător injectat  $S$  face malarie și, analog, fie  $\theta_N$  probabilitatea că un nepurtător injectat  $N$  face malarie. Presupunem independența între toți subiecții experimentului. Cu acest model, verosimilitatea este o funcție de  $\theta_S$  și  $\theta_N$ :

$$P(\text{date}|\theta_S, \theta_N) = c\theta_S^2(1 - \theta_S)^{13}\theta_N^{14}(1 - \theta_N).$$

Ca de obicei lăsăm factorul constant  $c$  ca o literă. (Este produsul a 2 coeficienți binomiali:  $c = C_{15}^2 \cdot C_{15}^{14}$ .)

**Ipoteze.** Fiecare ipoteză constă dintr-o pereche  $(\theta_N, \theta_S)$ . Pentru a păstra lucrurile simple vom considera doar un număr finit de valori pentru aceste probabilități. Am putea considera mult mai multe valori sau chiar un domeniu continuu pentru ipoteze. Presupunem că  $\theta_N$  și  $\theta_S$  pot avea fiecare valorile 0, 0.2, 0.4, 0.6, 0.8 sau 1. Aceasta duce la tabele 2 dimensionale.

Primul este un tabel de ipoteze. Codul de culori indică următoarele:

1. Dreptunghiurile portocaliu deschis de-a lungul diagonalei sunt unde  $\theta_S = \theta_N$ , i.e. celulele-seceră nu contează în niciun fel.
2. Dreptunghiurile roz și roșii de deasupra diagonalei sunt unde  $\theta_N > \theta_S$ , i.e. celulele-seceră dau protecție împotriva malariei.
3. În dreptunghiurile roșii  $\theta_N - \theta_S \geq 0.6$ , i.e. celulele-seceră dau multă protecție.
4. Dreptunghiurile albe de sub diagonală sunt unde  $\theta_S > \theta_N$ , i.e. celulele-

seceră de fapt cresc probabilitatea de a face malarie.

$\theta_N \setminus \theta_S$	0	0.2	0.4	0.6	0.8	1
1	(0,1)	(.2,1)	(.4,1)	(.6,1)	(.8,1)	(1,1)
0.8	(0,.8)	(.2,.8)	(.4,.8)	(.6,.8)	(.8,.8)	(1,.8)
0.6	(0,.6)	(.2,.6)	(.4,.6)	(.6,.6)	(.8,.6)	(1,.6)
0.4	(0,.4)	(.2,.4)	(.4,.4)	(.6,.4)	(.8,.4)	(1,.4)
0.2	(0,.2)	(.2,.2)	(.4,.2)	(.6,.2)	(.8,.2)	(1,.2)
0	(0,0)	(.2,0)	(.4,0)	(.6,0)	(.8,0)	(1,0)

Ipotezele asupra nivelului de protecție dată de  $S$ : roșu = mare; roz = mic; portocaliu = 0; alb = negativ.

Următorul este tabelul verosimilităților. (De fapt am profitat de indiferența noastră la scalare și am scalat toate verosimilitățile cu  $100000/c$  pentru a face tabelul mai prezentabil.) Observăm că, la precizia tabelului, multe verosimilități sunt 0. Codul de culori este același ca în tabelul de ipoteze. Am evidențiat cele mai mari verosimilități cu o margine albastră.

$\theta_N \setminus \theta_S$	0	0.2	0.4	0.6	0.8	1
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	1.93428	0.18381	0.00213	0.00000	0.00000
0.6	0.00000	0.06893	0.00655	0.00008	0.00000	0.00000
0.4	0.00000	0.00035	0.00003	0.00000	0.00000	0.00000
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Verosimilitățile  $p(\text{date}|\theta_S, \theta_N)$  scalate cu  $100000/c$ .

#### 2.4.1 A priori plată

Presupunem că nu avem nicio opinie despre dacă sau în ce măsură celula-seceră protejează împotriva malariei. În acest caz este rezonabil să folosim o a priori plată. Deoarece sunt 36 de ipoteze, fiecare primește o probabilitate a priori de  $1/36$ . Aceasta apare în tabelul de mai jos. Reamintim că fiecare dreptunghi din tabel reprezintă o ipoteză. Deoarece este un tabel de probabilități includem pmf-urile marginale.

$\theta_N \setminus \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.8	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(\theta_S)$	1/6	1/6	1/6	1/6	1/6	1/6	1

A priori plată  $p(\theta_S, \theta_N)$ : fiecare ipoteză (dreptunghi) are aceeași probabilitate

Pentru a calcula a posteriori înmulțim tabelul verosimilităților cu tabelul a priori și normalizăm. Normalizarea ne asigură că suma din întregul tabel este 1.

$\theta_N \setminus \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N   \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88075	0.08370	0.00097	0.00000	0.00000	0.96542
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00002	0.00000	0.00000	0.00000	0.00018
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S   \text{data})$	0.00000	0.91230	0.08670	0.00100	0.00000	0.00000	1.00000

A posteriori la a priori plată:  $p(\theta_S, \theta_N | \text{date})$

Pentru a decide dacă  $S$  dă protecție împotriva malariei, calculăm probabilitățile a posteriori pentru "protecție" și "protecție puternică". Acestea sunt calculate adunând numerele din dreptunghiurile corespunzătoare din tabelul a posteriori.

Protecție:  $P(\theta_N > \theta_S) = \text{suma din roz și roșu} = 0.99995$

Protecție puternică:  $P(\theta_N - \theta_S \geq 0.6) = \text{suma din roșu} = 0.88075$ .

Lucrând de la a priori plată, este efectiv sigur că celula-seceră dă protecție și foarte probabil că dă protecție puternică.

### 2.4.2 A priori informată

Acet experiment nu a fost făcut fără informație a priori. Erau multe dovezi că gena falciformă oferea protecție împotriva malariei. De exemplu, era raportat un mai mare procentaj de purtători care supraviețuiau până la maturitate.

Iată un mod de a construi o a priori informată: Vom rezerva o cantitate rezonabilă de probabilitate pentru ipoteza că  $S$  nu dă protecție. Să zicem că 24% împărțit egal între cele 6 celule portocalii unde  $\theta_N = \theta_S$ . Știm că nu ar trebui să punem nicio probabilitate a priori 0, deci hai să împărțim 6% din probabilitate între cele 15 celule de sub diagonală. Aceasta lasă 70% din probabilitate pentru cele 15 dreptunghiuri roz și roșii de deasupra diagonalei.

$\theta_N \setminus \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	0.04667	0.04667	0.04667	0.04667	0.04667	0.04000	0.27333
0.8	0.04667	0.04667	0.04667	0.04667	0.04000	0.00400	0.23067
0.6	0.04667	0.04667	0.04667	0.04000	0.00400	0.00400	0.18800
0.4	0.04667	0.04667	0.04000	0.00400	0.00400	0.00400	0.14533
0.2	0.04667	0.04000	0.00400	0.00400	0.00400	0.00400	0.10267
0	0.04000	0.00400	0.00400	0.00400	0.00400	0.00400	0.06000
$p(\theta_S)$	0.27333	0.23067	0.18800	0.14533	0.10267	0.06000	1.0

A priori informată  $p(\theta_S, \theta_N)$ : folosește informația a priori că celula seceră protejează.

Apoi completăm pmf a posteriori.

$\theta_N \setminus \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N   \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88076	0.08370	0.00097	0.00000	0.00000	0.96543
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00001	0.00000	0.00000	0.00000	0.00017
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S   \text{data})$	0.00000	0.91231	0.08669	0.00100	0.00000	0.00000	1.00000

A posteriori la a priori informată:  $p(\theta_S, \theta_N | \text{date})$

Calculăm din nou probabilitățile a posteriori ale "protecției" și "protecției puternice".

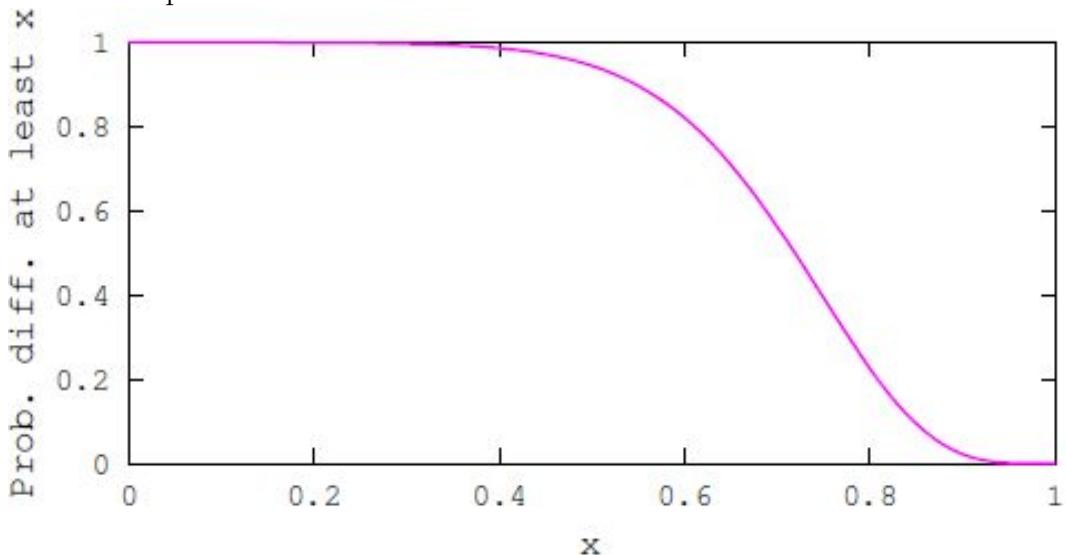
Protectie:  $P(\theta_N > \theta_S) =$  suma din roz și roșu = 0.99996

Protectie puternică:  $P(\theta_N - \theta_S \geq 0.6) =$  suma din roșu = 0.88076.

Observăm că a posteriori informată este aproape identică cu a posteriori din a priori plată.

#### 2.4.3 PDALX

Următoarea reprezentare este bazată pe a priori plată. Pentru fiecare  $x$ , dă probabilitatea ca  $\theta_N - \theta_S \geq x$ . Pentru a o face netedă au fost folosite mult mai multe ipoteze.



Probabilitatea că diferența  $\theta_N - \theta_S$  este cel puțin  $x$  (PDALX).

Observăm că este practic sigur că diferența este cel puțin 0.4.

# Curs 15

Cristian Niculescu

## 1 Intervale de probabilitate

### 1.1 Scopurile învățării

1. Să poată afla intervale de probabilitate fiind date o pmf sau pdf.
2. Să înțeleagă cum intervalele de probabilitate rezumă convingerea în actualizarea Bayesiană.
3. Să poată folosi intervale de probabilitate subiective pentru a construi a priori rezonabile.
4. Să poată construi intervale de probabilitate estimând sistematic cuantilele.

### 1.2 Intervale de probabilitate

Presupunem că avem o pmf  $p(\theta)$  sau pdf  $f(\theta)$  descriind convingerea noastră despre valoarea parametrului necunoscut de interes  $\theta$ .

**Definiție.** Un [interval de  \$p\$ -probabilitate](#) pentru  $\theta$  este un interval  $[a, b]$  cu  $P(a \leq \theta \leq b) = p$ .

**Observații.**

1. În cazul discret cu pmf  $p(\theta)$ , aceasta înseamnă  $\sum_{a \leq \theta_i \leq b} p(\theta_i) = p$ .
2. În cazul continuu cu pdf  $f(\theta)$ , aceasta înseamnă  $\int_a^b f(\theta) d\theta = p$ .
3. Putem spune [interval de 90%-probabilitate](#) pentru interval de 0.9-probabilitate. Intervalele de probabilitate sunt de asemenea numite [intervale credibile](#) spre a le deosebi de intervalele de încredere.

**Exemplul 1.** Între 0.05 și 0.55 cuantilele este un interval de 0.5-probabilitate. Sunt multe intervale de 50% probabilitate, de exemplu intervalul dintre 0.25 și 0.75 cuantilele.

În particular, observăm că intervalul de  $p$ -probabilitate [nu este unic](#).

**Q-notație.** Putem formula intervalele de probabilitate în termeni de **cuantile**. Reamintim că  $s$ -cuantila pentru  $\theta$  este valoarea  $q_s$  cu  $P(\theta \leq q_s) = s$ . Deci, pentru  $s \leq t$ , cantitatea de probabilitate dintre  $s$ -cuantila și  $t$ -cuantila este chiar  $t - s$ . În acești termeni, un interval de  $p$ -probabilitate este orice

interval  $[q_s, q_t]$  cu  $t - s = p$ .

**Exemplul 2.** Avem intervalele de 0.5 probabilitate  $[q_{0.25}, q_{0.75}]$  și  $[q_{0.05}, q_{0.55}]$ .

### Intervale de probabilitate simetrice.

Intervalul  $[q_{0.25}, q_{0.75}]$  este **simetric** deoarece cantitatea de probabilitate rămasă în afara lui, în oricare din cele 2 părți, este aceeași, și anume 0.25. Dacă pdf nu este prea înclinată, intervalul simetric este de obicei o bună alegere implicită.

### Mai multe observații.

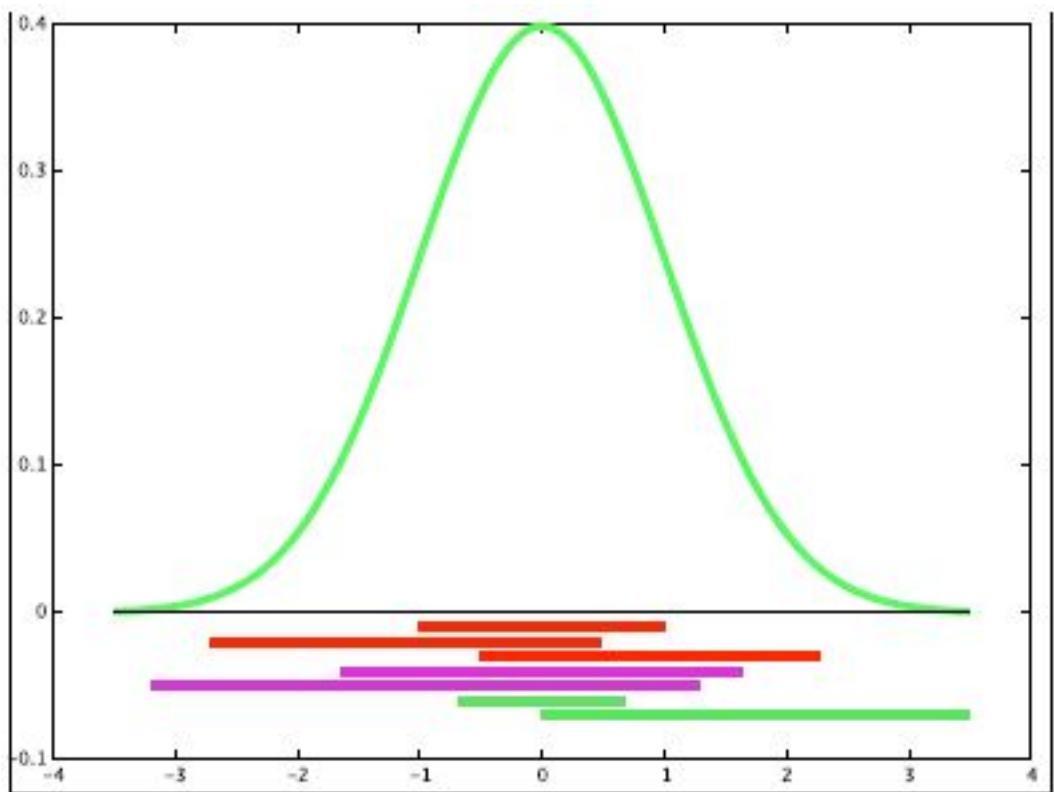
1. Diferite intervale de  $p$ -probabilitate pentru  $\theta$  pot avea lungimi diferite. Putem face lungimea mai mică centrând intervalul sub cea mai înaltă parte a pdf. Un astfel de interval este de obicei o bună alegere deoarece conține cele mai probabile valori.

2. Deoarece lungimea poate varia pentru  $p$  fixat, un  $p$  mai mare nu înseamnă totdeauna o lungime mai mare. Iată ce este adevărat: dacă un interval de  $p_1$ -probabilitate este inclus într-un interval de  $p_2$ -probabilitate, atunci  $p_1 \leq p_2$ .

**Intervale de probabilitate pentru o repartiție normală.** Figura arată un număr de intervale de probabilitate pentru normala standard.

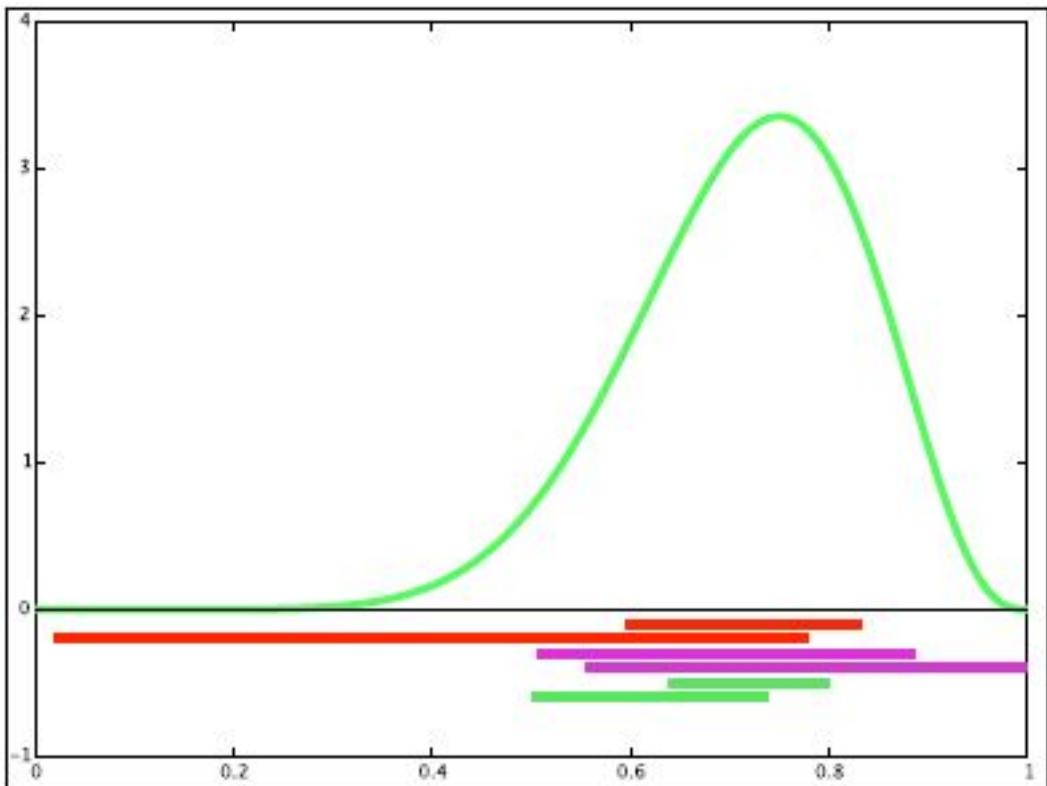
1. Toate barele roșii cuprind un interval de 0.68-probabilitate. Observați că cea mai mică bară roșie merge de la  $-1$  la  $1$ . Acest interval este de la a 16-a percentilă la a 84-a percentilă, deci este simetric.

2. Toate barele mov cuprind un interval de 0.9-probabilitate. Ele sunt mai lungi decât barele roșii deoarece cuprind mai multă probabilitate. Observați din nou că cea mai scurtă bară mov corespunde unui interval simetric.



roșu = 0.68, mov = 0.9, verde = 0.5

**Intervale de probabilitate pentru o repartiție beta.** Următoarea figură arată intervale de probabilitate pentru o repartiție beta. Observați că cele 2 bare roșii au lungimi foarte diferite, totuși cuprind aceeași probabilitate  $p = 0.68$ .



## 1.3 Utilizări ale intervalelor de probabilitate

### 1.3.1 Rezumarea și comunicarea convingerilor noastre

Intervalurile de probabilitate sunt un mod intuitiv și eficace de a rezuma și comunica convingerile noastre. Este greu de descris o întreagă funcție  $f(\theta)$  în cuvinte. Dacă funcția nu este dintr-o familie parametrizată, atunci este și mai greu. Chiar și cu o repartiție beta este mai ușor de interpretat "Cred că  $\theta$  este între 0.45 și 0.65 cu 50% probabilitate" decât "Cred că  $\theta$  are o repartiție beta(8, 6)". O excepție de la această regulă de comunicare poate fi repartiția normală, dar numai dacă interlocutorul este familiarizat cu deviația standard. Desigur, ce câștigăm în claritate pierdem în precizie, deoarece funcția conține mai multă informație decât intervalul de probabilitate.

Intervalurile de probabilitate se comportă bine în actualizarea Bayesiană. Dacă actualizăm de la a priori  $f(\theta)$  la a posteriori  $f(\theta|x)$ , atunci intervalul de  $p$ -probabilitate pentru a posteriori va tinde să fie mai scurt decât intervalul de  $p$ -probabilitate pentru a priori. În acest sens, datele ne fac mai siguri.

## 1.4 Construirea unei a priori folosind intervale de probabilitate subiective

Ierarhia de probabilitate sunt de asemenea utile când nu avem o pmf sau pdf la îndemâna. În acest caz, **intervalele de probabilitate subiective** ne dă o metodă de a construi o a priori rezonabilă pentru  $\theta$  "de la 0". Procesul de gândire este să ne punem o serie de întrebări, de exemplu: "care este media lui  $\theta$ ?"; "intervalul de 0.5-probabilitate?"; "intervalul de 0.9-probabilitate?". Apoi construim o a priori care este potrivită cu aceste intervale.

### 1.4.1 Estimarea directă a intervalelor

#### Exemplul 3. Construirea a priori

În 2013 au fost alegeri speciale pentru un loc în congres într-un district din Carolina de Sud. Alegerile au adus în arenă pe republicanul Mark Sanford contra democraticei Elizabeth Colbert Busch. Fie  $\theta$  fracția din populație care l-au favorizat pe Sanford. Scopul nostru în acest exemplu este să construim o a priori subiectivă pentru  $\theta$ . Vom folosi următoarele dovezi a priori: Sanford este un fost parlamentar și guvernator de Carolina de Sud.

El a demisionat cu tam-tam după ce a avut o aventură în Argentina în timp ce pretindea că face o drumeție pe un traseu din Munții Apalași.

În 2013 Sanford a câștigat alegerile primare republicane în fața a 15 opoziționi.

În district, în alegerile prezidențiale, republicanul Romney a învins pe democratul Obama cu 58% la 40%.

Avantajul lui Colbert: Elizabeth Colbert Busch este sora cunoscutului comic Stephen Colbert.

Strategia noastră va fi să ne folosim intuiția pentru a construi unele intervale de probabilitate și apoi să găsim o repartiție beta care se potrivește aproximativ cu aceste intervale. Acestea sunt subiective, deci altcineva poate da un răspuns diferit.

**Pasul 1.** Folosim dovezile a priori pentru a construi intervale de 0.5 și 0.9 probabilitate pentru  $\theta$ .

Vom începe gândindu-ne la intervalul de 90%. Singura dovedă a priori cea mai puternică este 58% la 40% la Romney contra Obama. Date fiind boacăna lui Sanford nu ne așteptăm să câștige mai mult de 58% din voturi. Deci vom pune marginea superioară a intervalului de 0.9 la 0.65. Din cauza boacănei, Sanford ar putea să piardă mult. Deci vom pune marginea inferioară la 0.3.

$$\text{intervalul de 0.9 : } [0.3, 0.65]$$

Pentru intervalul de 0.5 vom muta aceste margini înăuntru. Pare improbabil ca Sanford să obțină mai multe voturi ca Romney, deci putem lăsa 0.25 din

probabilitate ca el să ia peste 57%. Marginea inferioară pare mai greu de prezis. Vom lăsa 0.25 din probabilitate ca el să ia sub 42%.

$$\text{intervalul de } 0.5 : [0.42, 0.57]$$

**Pasul 2.** Folosim intervalele noastre de 0.5 și 0.9 probabilitate pentru a alege o repartiție beta care aproximează aceste intervale. Se folosește funcția din R `pbeta` și câteva încercări pentru a alege beta(11,12). Iată codul din R:

$$a = 11$$

$$b = 12$$

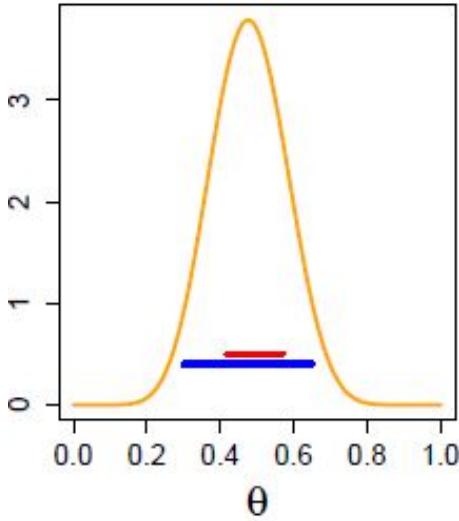
$$\text{pbeta}(0.65, a, b) - \text{pbeta}(0.3, a, b)$$

$$\text{pbeta}(0.57, a, b) - \text{pbeta}(0.42, a, b)$$

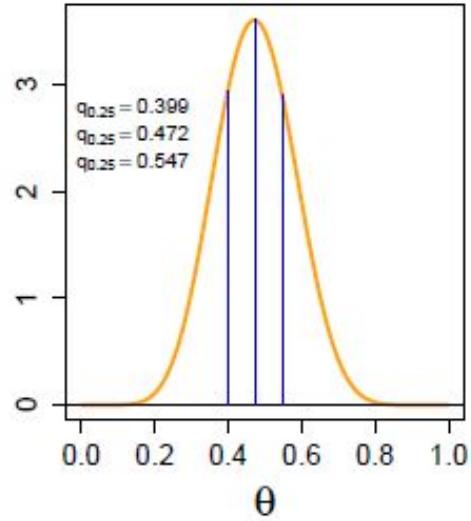
Obținem  $P([0.3, 0.65]) = 0.91$  și  $P([0.42, 0.57]) = 0.52$ . Deci intervalele noastre sunt de fapt intervale de 0.91 și 0.52-probabilitate. Aceasta este destul de aproape de ce am vrut.

În stânga este graficul densității lui beta(11,12). Linia roșie arată intervalul nostru  $[0.42, 0.57]$  și linia albastră arată intervalul nostru  $[0.3, 0.65]$ .

PDF for beta(11,12)



PDF for beta(9.9,11.0)



beta(11, 12) aflată folosind intervale de probabilitate și beta (9.9, 11) aflată folosind cuantilele

#### 1.4.2 Construcția unei a priori prin estimarea cuantilelor

Pentru a construi o a priori estimând quantilele, strategia de bază este a estima întâi mediană, apoi prima și a 4-a quartilă. Apoi alegem o repartiție a priori care se potrivește cu aceste estimări.

**Exemplul 4.** Refață exemplul de alegeri Sanford contra Colbert-Busch

folosind cuantilele.

**Răspuns.** Începem estimând mediana. Ca și mai devreme, singura dovedă a priori cea mai puternică este victoria cu 58% la 40% a lui Romney contra lui Obama. Totuși, dată fiind boacăna lui Sanford și avantajul lui Colbert, vom estima mediana la 0.47. Într-un district care a dat 58 la 40 pentru republicanul Romney este greu de imaginat că votul pentru Sanford va scădea sub 40%. Deci vom estima a 25-a percentilă pentru Sanford la 0.4. Analog, dată fiind boacăna lui, este greu de imaginat că va urca peste 58%, deci vom estima a 75-a percentilă a lui la 0.55.

Se folosește R pentru a căuta printre valorile lui  $a$  și  $b$  cu o zecimală cea mai bună potrivire. Se găsește beta(9.9, 11). Deasupra este o reprezentare a lui beta(9.9, 11) cu quartilele ei reale. În loc de " $q_{0.25} = 0.472$ " se va citi " $q_{0.5} = 0.472$ ". În loc de " $q_{0.25} = 0.547$ " se va citi " $q_{0.75} = 0.547$ ". Acestea se potrivesc cu quartilele dorite destul de bine.

**Notă istorică.** În alegeri Sanford a câștigat 54% din voturi și Busch a câștigat 45.2%. (Sursa: <http://elections.huffingtonpost.com/2013/mark-sanford-vs-elizabeth-colbert-busch-sc1>.)

## 2 Școala frecvenționistă de statistică

### 2.1 Scopurile învățării

1. Să poată să explice diferența dintre abordările frecvenționistă și Bayesiană ale statisticii.
2. Să știe definiția de lucru a statisticii și să poată distinge o statistică de o nestatistică.

### 2.2 Introducere

Pentru mare parte din secolul XX, statistica frecvenționistă a fost școala dominantă. Dacă ați întâlnit vreodată intervale de încredere,  $p$ -valori,  $t$ -teste sau  $\chi^2$ -teste, ați văzut statistică frecvenționistă. Odată cu dezvoltarea calculatoarelor de mare viteza și a datelor mari, metodele Bayesiene devin mai frecvente.

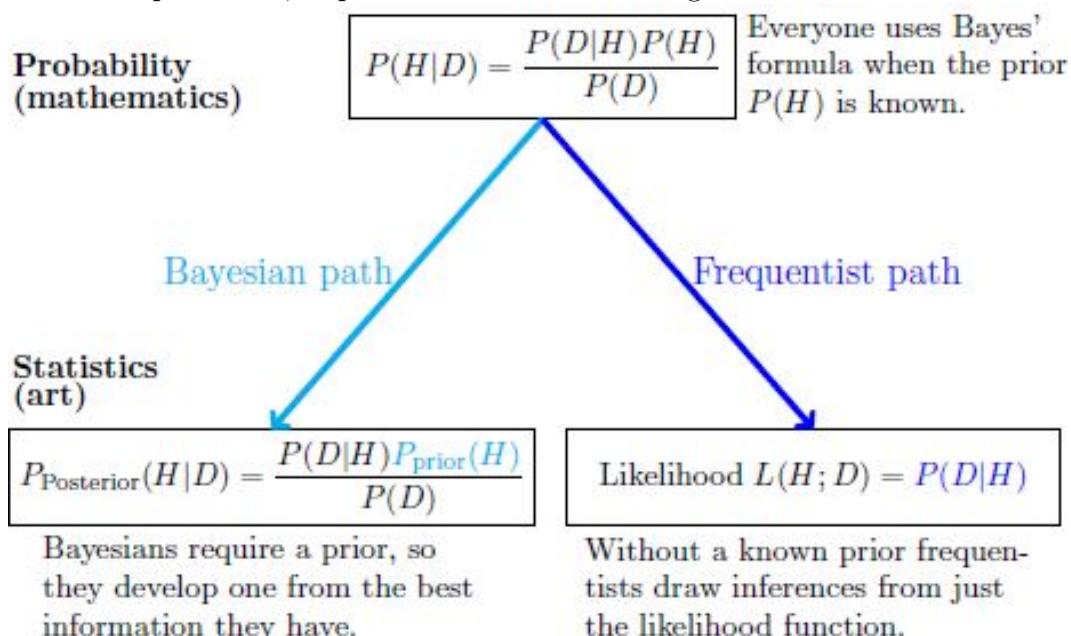
#### 2.2.1 Răspântia

Ambele școli de statistică încep cu probabilitatea. În particular ambele știu și apreciază teorema lui Bayes:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

Când a priori este cunoscută exact toți statisticienii vor folosi această formulă. Pentru deducția Bayesiană luăm  $H$  o ipoteză și  $D$  niște date. Date fiind o a priori și un model de verosimilitate, teorema lui Bayes este o rețetă completă pentru actualizarea convingerilor noastre în fața noilor date. Aceasta funcționează perfect când a priori a fost cunoscută perfect.

În practică de obicei nu există o a priori universal acceptată - persoane diferite vor avea **convingeri a priori** diferite - dar tot ne-ar plăcea să facem deducții utile din date. Bayesienii și frecvenționistii au abordări fundamentale diferite la această provocare, după cum este rezumat în figura următoare.



Motivele pentru această împărțire sunt atât practice (ușurința implementării și calculului) cât și filozofice (subiectivitate versus obiectivitate și natura probabilității).

### 2.2.2 Ce este probabilitatea?

Principala diferență filozofică privește înțelesul probabilității. Termenul **frecvenționist** se referă la idea că probabilitățile reprezintă frecvențe pe termen lung ale experimentelor aleatoare repetabile. De exemplu, "o monedă are probabilitatea 1/2 a aversului" înseamnă că frecvența relativă a aversurilor (numărul de aversuri supra numărul de aruncări) tinde la 1/2 când numărul de aruncări tinde la  $\infty$ . Aceasta înseamnă că frecvenționistii găsesc fără sens specificarea unei repartiții de probabilitate pentru un parametru cu o valoare fixată. În timp ce Bayesienii folosesc probabilitatea pentru a descrie cunoașterea lor incompletă a unui parametru fixat, frecvenționistii resping folosirea proba-

bilității pentru a cuantifica gradul de convingere în ipoteză.

**Exemplul 1.** Presupunem că avem o monedă cu probabilitate necunoscută  $\theta$  a aversului. Valoarea lui  $\theta$  poate fi necunoscută, dar este o valoare fixată. Astfel, pentru frecvenționist nu poate exista o pdf a priori  $f(\theta)$ . Prin comparație, Bayesianul poate fi de acord că  $\theta$  are o valoare fixată, dar interpretează  $f(\theta)$  ca reprezentând **incertitudinea** despre acea valoare. Atât Bayesianul cât și frecvenționistul sunt de acord cu  $p(\text{avers}|\theta) = \theta$ , deoarece frecvența pe termen lung a aversurilor dat fiind  $\theta$  este  $\theta$ .

Pe scurt, Bayesienii pun repartiții de probabilitate pe orice (ipoteze și date), în timp ce frecvenționiștii pun repartiții de probabilitate pe date (aleatoare, repetabile, experimentale) cunoscând o ipoteză. Pentru frecvenționist, când are de-a face cu date dintr-o repartitione necunoscută doar verosimilitatea are sens. A priori și a posteriori n-au.

### 2.3 Definiția de lucru a statisticii

**Statistică.** O **statistică** este orice poate fi calculat din date. Uneori, pentru a fi mai precisi, vom spune că o statistică este o **regulă** pentru a calcula ceva din date și **valoarea** statisticii este ce este calculat. Aceasta poate include calculul verosimilităților unde facem ipoteze asupra valorilor parametrului modelului. Dar nu include ceva care cere să știm adevărata valoare a parametrului cu valoare necunoscută a modelului.

- Exemple.**
1. Media datelor este o statistică. Este o regulă care spune că știind datele  $x_1, \dots, x_n$  calculăm  $\frac{x_1 + \dots + x_n}{n}$ .
  2. Maximul datelor este o statistică. Este o regulă care spune să alegem valoarea maximă a datelor  $x_1, \dots, x_n$ .
  3. Presupunem  $x \sim N(\mu, 9)$ , unde  $\mu$  este necunoscută. Atunci verosimilitatea

$$p(x|\mu = 7) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-7)^2}{18}}$$

este o statistică. Totuși, distanța de la  $x$  la adevărata medie  $\mu$  nu este o statistică deoarece nu putem să o calculăm fără să ști pe  $\mu$ .

**Statistică punctuală.** O **statistică punctuală** este o singură valoare calculată din date. De exemplu, media și maximul sunt ambele statistici punctuale. Estimarea de verosimilitate maximă este de asemenea o statistică punctuală deoarece este calculată direct din date pe baza unui model de verosimilitate.

**Statistică interval.** O **statistică interval** este un interval calculat din date. De exemplu domeniul de la minimul lui  $x_1, \dots, x_n$  la maximul lui  $x_1, \dots, x_n$  este o statistică interval, de exemplu datele 0.5, 1, 0.2, 3, 5 au domeniul [0.2, 5].

**Statistică mulțime.** O **statistică mulțime** este o mulțime calculată din

date.

**Exemplu.** Presupunem că avem 5 zaruri: cu 4, 6, 8, 12 și 20 de fețe. Alegem aleator unul și îl aruncăm. Valoarea aruncării este data. Multimea zarurilor pentru care această valoare este posibilă este o statistică mulțime. De exemplu, dacă aruncarea este 10, atunci valoarea acestei statistici mulțime este  $\{12, 20\}$ . Dacă aruncarea este 7, atunci această statistică mulțime are valoarea  $\{8, 12, 20\}$ .

O statistică este o variabilă aleatoare deoarece este calculată din date aleatoare. De exemplu, dacă datele provin din  $N(\mu, \sigma^2)$ , atunci media a  $n$  date are repartiția  $N(\mu, \sigma^2/n)$ .

**Repartiția de selecție.** Repartiția de probabilitate a unei statistici este numită [repartiția de selecție](#) a ei.

**Estimare punctuală.** Putem folosi statisticile pentru a face o [estimare punctuală](#) a parametrului  $\theta$ . De exemplu, dacă parametrul  $\theta$  reprezintă adevărata medie, atunci media datelor  $\bar{x}$  este o estimare punctuală a lui  $\theta$ .

# Curs 16

Cristian Niculescu

## 1 Testarea semnificației ipotezei 0 I

### 1.1 Scopurile învățării

1. Să știe definițiile termenilor de testare a semnificației: NHST, ipoteză 0, ipoteză alternativă, ipoteză simplă, ipoteză compusă, nivel de semnificație, putere.
2. Să poată face un test de semnificație pentru date Bernoulli și binomiale.
3. Să poată calcula o  $p$ -valoare pentru o ipoteză normală și să o utilizeze într-un test de semnificație.

### 1.2 Introducere

Statistica frecvenționistă este adesea aplicată în cadrul testării semnificației ipotezei 0 (NHST). Paradigma **Neyman-Pearson** se concentrează pe o ipoteză numită **ipoteza 0**. Sunt și alte paradigmă pentru testarea ipotezelor, dar Neyman-Pearson este cea mai uzuală. Spus simplu, această metodă întrebă dacă datele sunt în afara regiunii unde ne-am așteptă să le vedem sub ipoteza 0. Dacă este așa, respingem ipoteza 0 în favoarea unei a 2-a ipoteze, numită ipoteza alternativă.

Calculele făcute aici necesită toate funcția de verosimilitate. Sunt 2 diferențe majore între ce vom face aici și ce am făcut la actualizarea Bayesiană:

1. Dovezile datelor vor fi considerate doar prin funcția de verosimilitate, nu vor mai fi ponderate de convingerile noastre a priori.
2. Vom avea nevoie de o noțiune de date extreme, de exemplu 95 de aversuri din 100 de aruncări ale unei monede sau o muscă efemeră care trăiește o lună.

#### 1.2.1 Exemple motivante

**Exemplul 1.** Presupunem că vrem să decidem dacă o monedă este corectă. Dacă obținem 85 de aversuri din 100 de aruncări, ati considera că moneda este probabil incorectă? Dar la 60 de aversuri? Sau la 52? Majoritatea

oamenilor ar ghici că 85 de aversuri este o dovardă puternică pentru faptul că moneda este incorectă în timp ce 52 de aversuri nu este deloc dovardă. 60 de aversuri este mai puțin clar. Testarea semnificației ipotezei 0 (NHST) este o abordare frecvenționistă pentru a gândi cantitativ despre aceste întrebări.

**Exemplul 2.** Presupunem că vrem să comparăm un tratament medical nou cu un placebo sau cu standardul curent de îngrijire. Ce fel de dovezi v-ar convinge că noul tratament este mai bun decât placebo sau standardul curent? Din nou, NHST este un cadru cantitativ pentru a răspunde acestei întrebări.

### 1.3 Testarea semnificației

Listăm ingredientele pentru NHST.

#### 1.3.1 Ingrediente

$H_0$ : **ipoteza 0**. Aceasta este presupunerea implicită pentru modelul care generează datele.

$H_A$ : **ipoteza alternativă**. Dacă respingem ipoteza 0, acceptăm această alternativă ca cea mai bună explicație pentru date.

$X$ : **statistica testului**. O calculăm din date.

**Repartiția 0**: repartiția de probabilitate a lui  $X$  presupunând  $H_0$ .

**Regiunea de respingere**: dacă  $X$  este în regiunea de respingere, respingem  $H_0$  în favoarea lui  $H_A$ .

**Regiunea de nerespingere**: complementara regiunii de respingere. Dacă  $X$  este în această regiune, nu respingem  $H_0$ . Spunem "nu respingem" mai degrabă decât "acceptăm" deoarece cel mai bun lucru pe care-l putem spune este că datele nu susțin respingerea lui  $H_0$ .

Ipoteza 0  $H_0$  și ipoteza alternativă  $H_A$  joacă roluri diferite. Respingerem  $H_0$  doar dacă avem destule dovezi împotriva ei.

### 1.4 Terminologia NHST

În această secțiune vom folosi un exemplu extins pentru a introduce și explora terminologia utilizată în testarea semnificației ipotezei 0 (NHST).

**Exemplul 3.** Pentru a testa dacă o monedă este corectă o aruncăm de 10 ori. Dacă obținem un număr neașteptat de mare sau de mic de aversuri, vom suspecta că moneda nu este corectă. Precizăm aceasta în limbajul NHST setând ingredientele. Fie  $\theta$  probabilitatea aversului la aruncarea monedei.

1. Ipoteza 0:  $H_0 = \text{"moneda este corectă"}$ , i.e.  $\theta = 0.5$ .
2. Ipoteza alternativă:  $H_A = \text{"moneda nu este corectă"}$ , i.e.  $\theta \neq 0.5$ .

3. Statistica testului:  $X$  = numărul de aversuri în 10 aruncări.
4. Repartiția 0: Aceasta este funcția de probabilitate bazată pe ipoteza 0

$$p(x|\theta = 0.5) \sim \text{binomială}(10, 0.5).$$

Iată tabelul de probabilitate pentru repartiția 0:

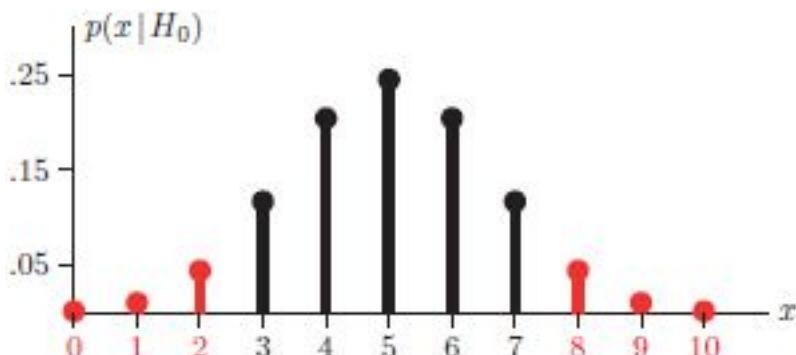
$x$	0	1	2	3	4	5	6	7	8	9	10
$p(x H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

5. Regiunea de respingere: sub ipoteza 0 ne așteptăm să obținem circa 5 aversuri în 10 aruncări. Vom respinge  $H_0$  dacă numărul de aversuri este mult mai mic sau mai mare ca 5. Fie regiunea de respingere  $\{0, 1, 2, 8, 9, 10\}$ . Adică, dacă numărul de aversuri din 10 aruncări este această regiune, vom respinge ipoteza că moneda este corectă în favoarea ipotezei că nu este corectă.

Putem rezuma toate acestea în graficul și tabelul de probabilitate de mai jos. Regiunea de respingere constă din valorile lui  $x$  în roșu. Probabilitățile corespunzătoare sunt pe fond roșu. Arătăm de asemenea repartiția 0 ca o reprezentare cu valorile lui  $x$  de respingere în roșu.

$x$	0	1	2	3	4	5	6	7	8	9	10
$p(x H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

Regiunea de respingere și probabilitățile 0 ca un tabel pentru exemplul 3.



Regiunea de respingere și probabilitățile 0 ca o reprezentare pentru exemplul 3.

Observații pentru exemplul 3:

1. Ipoteza 0 este **implicit precaută**: nu vom pretinde că moneda este incorrectă dacă nu avem argumente convingătoare.
2. Regiunea de respingere constă din date care sunt **extreme sub ipoteza 0**. Adică, constă din rezultatele care sunt în coada repartiției 0, departe de centrul de probabilitate mare. Cât de departe, depinde de  $\alpha$ , nivelul de semnificație al testului.

3. Dacă obținem 3 aversuri în 10 aruncări, atunci statistica testului este în regiunea de nerespingere. Limbajul științific ușual ar fi să spunem că datele "nu sprijină respingerea ipotezei 0". Chiar dacă avem 5 aversuri, nu vom pretinde că datele demonstrează că ipoteza 0 este adevărată.

**Întrebare:** Dacă avem o monedă corectă, care este probabilitatea că vom decide incorrect că este incorrectă?

**Răspuns:** Ipoteza 0 este că moneda este corectă. Se cere probabilitatea ca datele dintr-o monedă corectă să fie în regiunea de respingere. Adică, probabilitatea că vom obține 0, 1, 2, 8, 9 sau 10 aversuri în 10 aruncări. Aceasta este suma probabilităților în roșu. Adică,

$$P(\text{respingerea lui } H_0 | H_0 \text{ este adevărată}) = 0.11.$$

#### 1.4.1 Ipoteze simple și compuse

**Definiție: ipoteză simplă:** O ipoteză simplă este una pentru care putem specifica complet repartiția ei. O ipoteză simplă tipică este că parametrul de interes ia o anumită valoare.

**Definiție: Ipoteze compuse:** Dacă repartiția nu poate fi specificată complet, spunem că ipoteza este compusă. O ipoteză compusă tipică este că parametrul de interes se află într-un domeniu de valori.

În exemplul 3, ipoteza 0 este că  $\theta = 0.5$ , deci repartiția 0 este binomială(10, 0.5). Deoarece repartiția 0 este complet specificată,  $H_0$  este simplă. Ipoteza alternativă este că  $\theta \neq 0.5$ . În realitate sunt multe ipoteze în una:  $\theta$  ar putea fi 0.51, 0.7, 0.99, etc. Deoarece repartiția alternativă binomială(10,  $\theta$ ) nu este complet specificată,  $H_A$  este compusă.

**Exemplul 4.** Presupunem că avem datele  $x_1, \dots, x_n$ . Presupunem că ipotezele noastre sunt

$H_0$ : datele sunt extrase din  $N(0, 1)$

$H_A$ : datele sunt extrase din  $N(1, 1)$ .

Acestea sunt ambele ipoteze simple - fiecare ipoteză specifică complet o repartiție.

**Exemplul 5. (Ipoteze compuse.)** Acum presupunem că ipotezele noastre sunt

$H_0$ : datele sunt extrase dintr-o repartiție Poisson cu parametru necunoscut.

$H_A$ : datele nu sunt extrase dintr-o repartiție Poisson.

Acestea sunt ambele ipoteze compuse, deoarece ele nu specifică complet repartiția.

**Exemplul 6.** Într-un experiment de perceptie extrasenzorială (ESP), unui subiect i se cere să identifice culorile (pică, cupă, romb, treflă) a 100 de cărți trase (cu înlocuire) dintr-un pachet de cărți de joc. Fie  $T$ , numărul

succeselor. Ipoteza 0 (simplă) că subiectul nu are ESP este dată de

$$H_0 : T \sim \text{binomial}(100, 0.25).$$

Ipoteza alternativă (compusă) că subiectul are ESP este dată de

$$H_A : T \sim \text{binomial}(100, p) \text{ cu } p > 0.25.$$

O altă ipoteză alternativă (compusă) este că se întâmplă ceva în afara şansei pure, i.e. subiectul are ESP sau anti-ESP. Aceasta este dată de

$$H_A : T \sim \text{binomial}(100, p) \text{ cu } p \neq 0.25.$$

Valorile  $p < 0.25$  reprezintă ipotezele că subiectul are un fel de anti-ESP.

#### 1.4.2 Tipuri de eroare

Sunt 2 tipuri de erori pe care le putem face. Putem să respingem incorect ipoteza 0 când ea este adevărată sau putem să eșuăm incorect să-o respingem când este falsă. Acestea sunt numite **erori de tipul I și de tipul II**. Rezumăm aceasta în următorul tabel.

		True state of nature	
		$H_0$	$H_A$
Our decision	Reject $H_0$	Type I error	correct decision
	'Don't reject' $H_0$	correct decision	Type II error

Tipul I: respingere falsă a lui  $H_0$ .

Tipul II: nerespingere falsă ("acceptare" falsă) a lui  $H_0$ .

#### 1.4.3 Nivel de semnificație și putere

Nivelul de semnificație și puterea sunt folosite pentru a cuantifica calitatea testului de semnificație. Ideal, un test de semnificație n-ar face erori. Adică, n-ar respinge  $H_0$  când  $H_0$  era adevărată și ar respinge  $H_0$  în favoarea lui  $H_A$  când  $H_A$  era adevărată. Sunt cu totul 4 probabilități importante corespunzătoare tabelului  $2 \times 2$  de mai sus:

$$\begin{array}{ll} P(\text{respungem } H_0 | H_0) & P(\text{respungem } H_0 | H_A) \\ P(\text{nu respungem } H_0 | H_0) & P(\text{nu respungem } H_0 | H_A). \end{array}$$

Cele 2 probabilități pe care ne concentrăm sunt:

$$\begin{aligned} \text{Nivelul de semnificație} &= P(\text{respungem } H_0 | H_0) \\ &= \text{probabilitatea să respungem incorect } H_0 \\ &= P(\text{eroarea de tipul I}). \end{aligned}$$

$$\begin{aligned}
\text{Puterea} &= \text{probabilitatea să respingem corect } H_0 \\
&= P(\text{respingerem } H_0 | H_A) \\
&= 1 - P(\text{eroarea de tipul II}).
\end{aligned}$$

Ideal, un test al unei ipoteze ar trebui să aibă un nivel de semnificație mic (aproape de 0) și o putere mare (aproape de 1).

### Analogii

1. Gândiți  $H_0$  ca ipoteza "nu se întâmplă nimic demn de remarcat", i.e. "moneda este corectă", "tratamentul nu este mai bun ca placebo" etc. și gândiți  $H_A$  ca opusa: "ceva interesant se întâmplă". Atunci puterea este probabilitatea de a detecta ceva interesant când este prezent și nivelul de semnificație este probabilitatea de a pretinde greșit că ceva interesant a apărut.

2. În SUA, inculpații sunt presupuși nevinovați până sunt demonstrați vinovați dincolo de un orice dubiu rezonabil. Putem formula aceasta în termeni de NHST ca

$H_0$ : acuzatul este nevinovat (implicit)

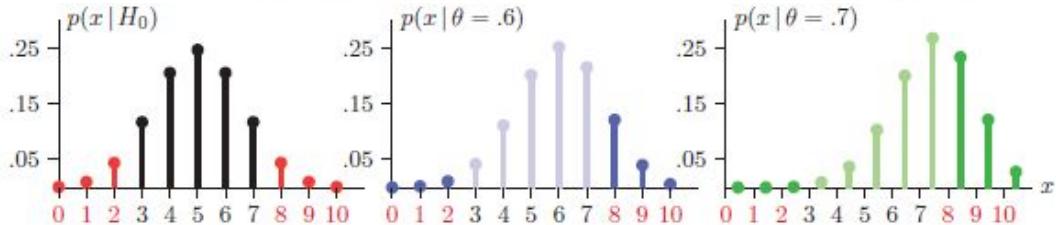
$H_A$ : acuzatul este vinovat.

Nivelul de semnificație este probabilitatea de a găsi vinovată o persoană nevinovată. Puterea este probabilitatea de a găsi corect vinovată o parte vinovată. "Dincolo de orice dubiu rezonabil" înseamnă că ar trebui să cerem ca nivelul de semnificație să fie foarte mic.

### Ipoteze compuse

$H_A$  este compusă în exemplul 3, deci puterea este diferită pentru diferite valori ale lui  $\theta$ . Extindem tabelul anterior de probabilități pentru a include unele valori alternative ale lui  $\theta$ . Facem același lucru cu reprezentările. Ca întotdeauna în jocul NHST, ne uităm la **verosimilități**: probabilitatea datelor cunoscând ipoteza.

$x$	0	1	2	3	4	5	6	7	8	9	10
$H_0 : p(x \theta = 0.5)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
$H_A : p(x \theta = 0.6)$	.000	.002	.011	.042	.111	.201	.251	.215	.121	.040	.006
$H_A : p(x \theta = 0.7)$	.000	.0001	.001	.009	.037	.103	.200	.267	.233	.121	.028



Regiunea de respingere și probabilități 0 și alternative pentru exemplul 3.  
Folosim tabelul de probabilități pentru a calcula nivelul de semnificație și

puterea acestui test.

$$\begin{aligned}
 \text{Nivelul de semnificație} &= \text{probabilitatea să respingem } H_0 \text{ când } H_0 \text{ este adevărată} \\
 &= \text{prob. că stat. testului e în regiunea de respingere când } H_0 \text{ e adev.} \\
 &= \text{prob. că stat. testului e în regiunea de respingere pe linia } H_0 \text{ a tab.} \\
 &= \text{suma valorilor căsuțelor roșii din linia } \theta = 0.5 \\
 &= 0.11.
 \end{aligned}$$

$$\begin{aligned}
 \text{Puterea când } \theta = 0.6 &= \text{probabilitatea să respingem } H_0 \text{ când } \theta = 0.6 \\
 &= \text{prob. că stat. testului e în regiunea de respingere când } \theta = 0.6 \\
 &= \text{prob. că stat. test. e în regiunea de respingere pe linia } \theta = 0.6 \text{ a tab.} \\
 &= \text{suma valorilor căsuțelor albastru închis din linia } \theta = 0.6 \\
 &= 0.18.
 \end{aligned}$$

$$\begin{aligned}
 \text{Puterea când } \theta = 0.7 &= \text{probabilitatea să respingem } H_0 \text{ când } \theta = 0.7 \\
 &= \text{prob. că stat. testului e în regiunea de respingere când } \theta = 0.7 \\
 &= \text{prob. că stat. test. e în regiunea de respingere pe linia } \theta = 0.7 \text{ a tab.} \\
 &= \text{suma valorilor căsuțelor verde închis din linia } \theta = 0.7 \\
 &= 0.384.
 \end{aligned}$$

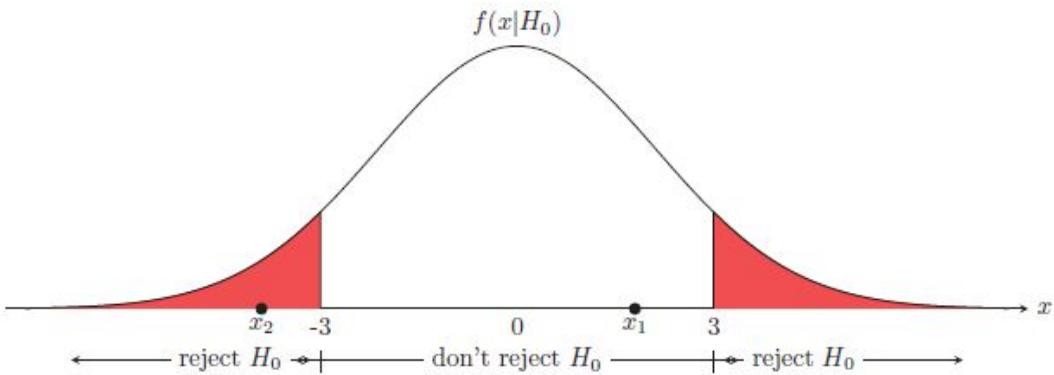
Vedem că puterea este mai mare pentru  $\theta = 0.7$  decât pentru  $\theta = 0.6$ . Aceasta nu este surprinzător deoarece ne așteptăm să fie mai ușor să recunoaștem că o monedă cu  $\theta = 0.7$  este incorrectă decât este să recunoaștem că o monedă cu  $\theta = 0.6$  este incorrectă. Tipic, obținem putere mai mare când ipoteza alternativă este mai departe de ipoteza 0. În exemplul 3, ar fi foarte greu să distingem o monedă corectă de una cu  $\theta = 0.51$ .

#### 1.4.4 Schițe conceptuale

Ilustrăm noțiunile de ipoteză 0, regiune de respingere și putere cu schițe ale pdf-urilor pentru ipotezele 0 și alternative.

##### **Repartiția 0: regiuni de respingere și nerespingere**

Prima diagramă de mai jos ilustrează o repartiție 0 cu regiunile de respingere și nerespingere. Sunt arătate de asemenea 2 posibile statistici ale testului:  $x_1$  și  $x_2$  ("reject" = "respingem", "don't reject" = "nu respingem").



Statistica  $x_1$  a testului este în regiunea de nerespingere. Deci, dacă datele noastre au produs statisica testului  $x_1$ , atunci nu vom respinge ipoteza 0  $H_0$ . Pe de altă parte, statistica testului  $x_2$  este în regiunea de respingere, deci, dacă datele noastre au produs  $x_2$ , atunci vom respinge ipoteza 0 în favoarea ipotezei alternative.

Sunt câteva lucruri de observat în această imagine.

1. Regiunea de respingere constă din valorile departe de centrul repartiției 0.
2. Regiunea de respingere este bilaterală.
3. Ipoteza alternativă nu este menționată. Respingem sau nu respingem  $H_0$  bazați numai pe verosimilitatea  $f(x|H_0)$ . Ipoteza alternativă  $H_A$  ar trebui considerată când alegem o regiune de respingere, dar formal nu joacă niciun rol în respingerea sau nerespingerea lui  $H_0$ .
4. Uneori numim regiunea de nerespingere **regiunea de acceptare**. Aceasta este tehnic incorrect deoarece niciodată nu acceptăm cu adevărat ipoteza 0. Sau respingem, sau spunem că datele nu sprijină respingerea lui  $H_0$ . Aceasta este adesea rezumat prin afirmația: **nu putem niciodată demonstra ipoteza 0**.

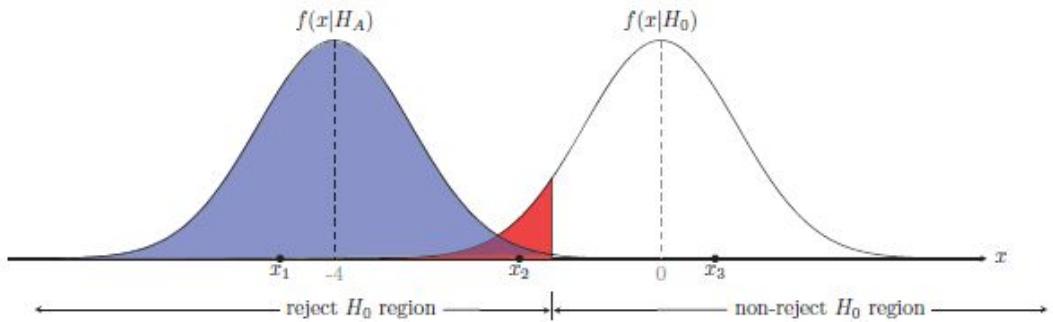
**Teste de putere mare sau mică** Următoarele 2 figuri arată teste de putere mare, respectiv mică.

Aria umbrită de sub  $f(x|H_0)$  reprezintă nivelul de semnificație. Reamintim că nivelul de semnificație este:

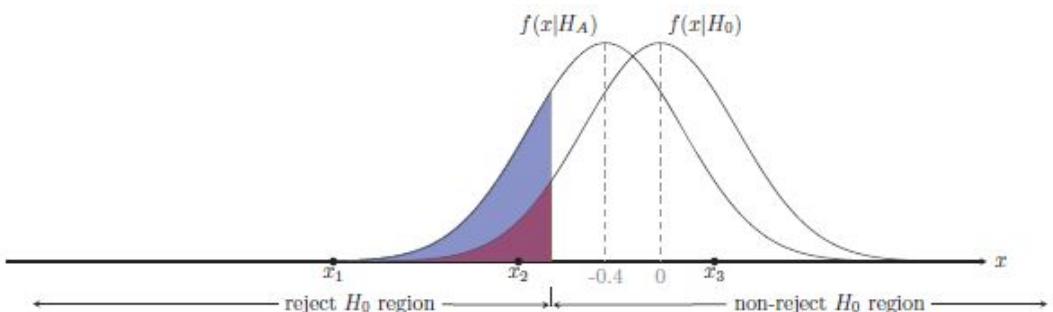
probabilitatea respingerii false a ipotezei 0 când ea este adevărată;

probabilitatea ca statistica testului să cadă în regiunea de respingere chiar dacă  $H_0$  este adevărată.

Analog, regiunea umbrită de sub  $f(x|H_A)$  reprezintă puterea, i.e. probabilitatea că statistica testului este în regiunea de respingere a lui  $H_0$  când  $H_A$  este adevărată. Ambele teste au același nivel de semnificație, dar dacă  $f(x|H_A)$  are o suprapunere considerabilă cu  $f(x|H_0)$ , puterea este mult mai mică.



Test de putere mare



Test de putere mică

În ambele teste, repartițiile  $H_0$  sunt normale standard. Repartiția  $H_0$ , regiunea de respingere și nivelul de semnificație sunt aceleași. (Nivelul de semnificație este aria roșie/mov de sub  $f(x|H_0)$  și de deasupra regiunii de respingere.) În figura de deasupra vedem că mediile celor 2 repartiții sunt la distanță de 4 deviații standard una față de cealaltă. Doarece ariile de sub densități au o foarte mică suprapunere, testul are putere mare. Adică, dacă datele  $x$  sunt extrase din  $H_A$ , vor fi aproape sigur în regiunea de respingere a lui  $H_0$ . De exemplu  $x_3$  ar fi un rezultat foarte surprinzător pentru repartiția  $H_A$ .

În figura de jos vedem că mediile sunt la doar 0.4 deviații standard una de alta. Deoarece ariile de sub densități au multă suprapunere, testul are putere mică. Adică, dacă datele  $x$  sunt extrase din  $H_A$ , este foarte probabil să se afle în regiunea de nerespingere a lui  $H_0$ . De exemplu,  $x_3$  n-ar fi un rezultat foarte surprinzător pentru repartiția  $H_A$ .

Tipic, putem crește puterea unui test crescând numărul de date și, prin aceasta, scăzând dispersia repartițiilor  $0$  și alternativă. În proiectarea experimentului este important să determinăm dinainte numărul de încercări sau de subiecți de care avem nevoie pentru a atinge o putere dorită.

**Exemplul 7.** Presupunem că un medicament pentru o boală este comparat cu un placebo. Alegem ipotezele  $0$  și alternativă astfel:

$H_0$  = medicamentul nu este mai bun ca placebo;

$H_A$  = medicamentul este mai bun ca placebo.

Puterea testului ipotezei este probabilitatea ca testul să concluzioneze că medicamentul este mai bun, dacă este într-adevăr mai bun. Nivelul de semnificație este probabilitatea ca testul să concluzioneze că medicamentul este mai bun, când de fapt nu este mai bun.

## 1.5 Proiectarea unui test al ipotezei

Formal, tot ce necesită un test al ipotezei este  $H_0, H_A$ , o statistică a testului și o regiune de respingere. În practică proiectarea este adesea făcută folosind pașii următori:

### 1. Alege ipoteza 0 $H_0$ .

Adesea alegem  $H_0$  să fie simplă. Sau, adesea alegem  $H_0$  să fie cea mai simplă și cea mai prudentă explicație, i.e. medicamentul nu are efect, fără ESP, moneda este corectă.

### 2. Decide dacă $H_A$ este unilaterală sau bilaterală.

În exemplul 3 am vrut să stim dacă moneda a fost incorectă. O monedă incorectă poate fi deplasată pentru sau împotriva aversului, deci  $H_A : \theta \neq 0.5$  este o ipoteză bilaterală. Dacă ne pasă doar dacă moneda este deplasată sau nu pentru avers am putea folosi o ipoteză unilaterală  $H_A : \theta > 0.5$ .

### 3. Alege o statistică a testului.

De exemplu, media de selecție, totalul de selecție sau dispersia de selecție. Unele statistici standard sunt  $z, t$  și  $\chi^2$ . Repartițiile care merg cu aceste statistici sunt totdeauna condiționate de ipoteza 0, adică vom calcula verosimilități ca  $f(z|H_0)$ .

### 4. Alege un nivel de semnificație și determină regiunea de respingere.

Vom folosi de obicei  $\alpha$  pentru a nota nivelul de semnificație. Paradigma Neyman-Pearson este de a alege  $\alpha$  înainte. Valori tipice sunt 0.1, 0.05, 0.01. Reamintim că nivelul de semnificație este probabilitatea unei erori de tip I, i.e. respingerea incorectă a ipotezei 0 când ea este adevărată. Valoarea pe care o alegem va depinde de consecințele unei erori de tipul I. Odată ce nivelul de semnificație este ales, putem determina regiunea de respingere în coada (cozile) repartiției 0. În exemplul 3,  $H_A$  este bilaterală, deci regiunea de respingere este împărțită între cele 2 cozile ale repartiției 0. Această repartitione este dată în următorul tabel:

$x$	0	1	2	3	4	5	6	7	8	9	10
$p(x H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

Dacă punem  $\alpha = 0.05$ , atunci regiunea de respingere trebuie să conțină cel

mult 0.05 probabilitate. Pentru o regiune bilaterală, obținem

$$\{0, 1, 9, 10\}.$$

Dacă punem  $\alpha = 0.01$ , regiunea de respingere este

$$\{0, 10\}.$$

Presupunem că schimbăm  $H_A$  în ”moneda este deplasată în favoarea aversului”. Acum avem o ipoteză unilaterală  $\theta > 0.5$ . Regiunea noastră de respingere va fi acum în coada dreaptă, deoarece nu vrem să respingem  $H_0$  în favoarea lui  $H_A$  dacă obținem un număr mic de aversuri. Acum, dacă  $\alpha = 0.05$ , regiunea de respingere este domeniul unilateral

$$\{9, 10\}.$$

Dacă punem  $\alpha = 0.01$ , regiunea de respingere este

$$\{10\}.$$

### 5. Determină puterea (puterile)

După cum am văzut în exemplul 3, odată ce regiunea de respingere este stabilită, putem determina puterea testului la diverse valori ale ipotezei alternative.

**Exemplul 8. (Consecințele semnificației.)** Dacă  $\alpha = 0.1$  aşteptăm o rată a erorii de tipul I de 10%. Adică, aşteptăm să respingem ipoteza 0 în 10% din acele experimente unde ipoteza  $H_0$  este adevărată. Faptul dacă 0.1 este un nivel de semnificație rezonabil depinde de deciziile care vor fi făcute folosindu-l.

De exemplu, dacă faci un experiment pentru a determina dacă ciocolata ta are mai mult de 72% cacao, atunci o rată a erorii de tipul I de 10% este probabil în regulă. Adică, a considera fals că o ciocolată 72% are mai mult de 72%, este probabil acceptabil. Pe de altă parte, dacă laboratorul tău criministic identifică amprente pentru un proces de crimă, atunci o rată de 10% a erorii de tipul I, i.e. a pretinde greșit că amprentele găsite la locul crimei au aparținut cuiva care în realitate este nevinovat, este categoric inacceptabilă. **Semnificația pentru o ipoteză 0 compusă.** Dacă  $H_0$  este compusă, atunci  $P(\text{eroarea de tipul I})$  depinde de care membru al lui  $H_0$  este adevărat. În acest caz nivelul de semnificație este definit ca maximul acestor probabilități.

## 1.6 Valori critice

**Valorile critice** sunt ca cuantilele cu excepția faptului că se referă la probabilitatea de la dreapta valorii, în loc de cea de la stânga.

**Exemplul 9.** Folosiți R pentru a afla valoarea critică 0.05 pentru repartiția normală standard.

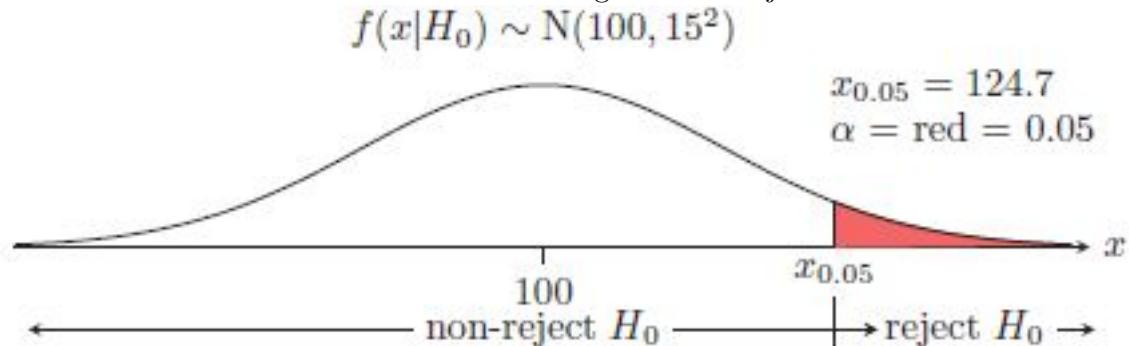
**Răspuns.** Notăm această valoare critică cu  $z_{0.05}$ . Valoarea critică  $z_{0.05}$  este chiar 0.95 cuantila, i. e. are 5% probabilitate la dreapta ei și de aceea 95% probabilitate la stânga. O calculăm cu funcția R qnorm: `qnorm(0.95, 0, 1)` sau `qnorm(0.95)` și obținem 1.644854.

Într-un test de semnificație tipic, regiunea de respingere constă din una sau ambele cozi ale repartiției 0. Valoarea care marchează începutul regiunii de respingere este o **valoare critică**.

**Exemplul 10. Valori critice și regiuni de respingere.** Presupunem că statis-tica  $x$  a testului nostru are repartiția  $N(100, 15^2)$ , i. e.

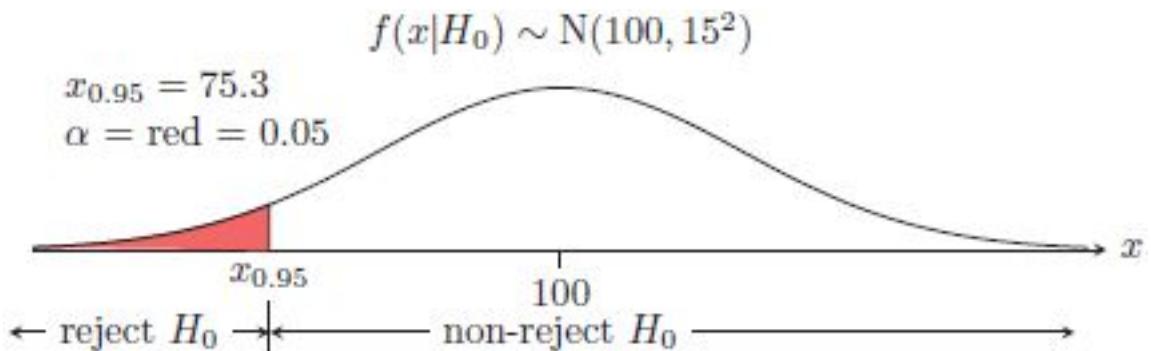
$f(x|H_0) \sim N(100, 15^2)$ . Presupunem de asemenea că regiunea noastră de respingere este în dreapta și avem un nivel de semnificație de 0.05. Aflați valoarea critică și schițați repartiția 0 și regiunea de respingere.

**Răspuns.** Notația folosită pentru valoarea critică cu coada dreaptă conținând 0.05 probabilitate este  $x_{0.05}$ . Valoarea critică  $x_{0.05}$  este chiar 0.95 cuantila, i. e. are 5% probabilitate la dreapta ei și de aceea 95% probabilitate la stânga. O calculăm cu funcția R qnorm: `qnorm(0.95, 100, 15)` și obținem 124.6728  $\approx$  124.7. Aceasta este arătată în figura de mai jos.



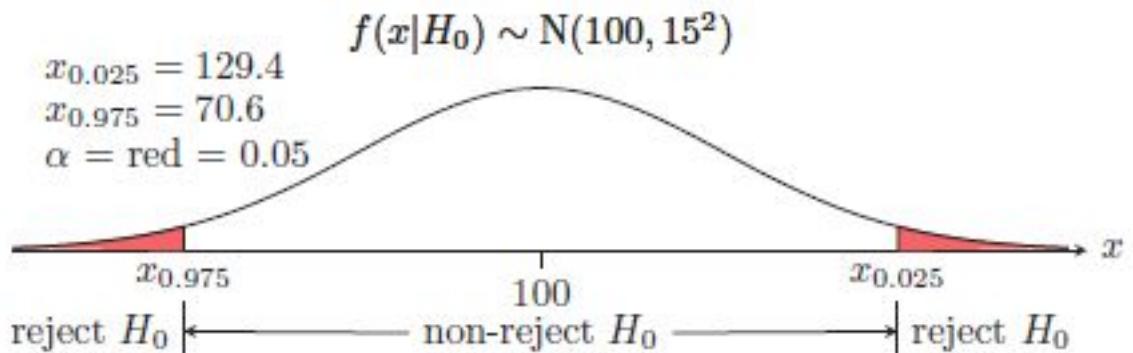
**Exemplul 11. Valori critice și regiuni de respingere.** Repetați exemplul precedent pentru o regiune de respingere în stânga cu nivelul de semnificație 0.05.

**Răspuns.** În acest caz valoarea critică are 0.05 probabilitate la stânga și de aceea 0.95 probabilitate la dreapta. Deci o notăm cu  $x_{0.95}$ . Deoarece este 0.05 cuantila, o calculăm în R cu `qnorm(0.05, 100, 15)` și obținem 75.3272  $\approx$  75.3.



**Exemplul 12. Valori critice.** Repetați exemplul anterior pentru o regiune de respingere bilaterală. Puneți jumătate din semnificație în fiecare coadă.

**Răspuns.** Pentru a avea un total de semnificație de 0.05 punem 0.025 în fiecare coadă. Adică, coada stângă începe la  $x_{0.975} = q_{0.025}$  și coada dreaptă începe la  $x_{0.025} = q_{0.975}$ . Calculăm aceste valori cu `qnorm(0.025, 100, 15)` și `qnorm(0.975, 100, 15)`. Valorile aproximative sunt arătate în figura de mai jos.



## 1.7 Valori $p$

În practică, oamenii adesea specifică nivelul de semnificație și fac testul de semnificație folosind **valori  $p$** .

Dacă valoarea  $p$  este mai mică decât nivelul de semnificație  $\alpha$ , atunci respingem  $H_0$ . Altfel nu respingem  $H_0$ .

**Definiție.**  $p$ -valoarea este probabilitatea, presupunând ipoteza 0, de a vedea datele **cel puțin la fel de extreme ca datele experimentale**. Ce înseamnă "cel puțin la fel de extreme" depinde de proiectarea experimentului.

Ilustrăm definiția și utilizarea valorilor  $p$  cu un exemplu unilateral. Acest exemplu introduce de asemenea **testul  $z$** . Toate acestea înseamnă că statistica testului nostru este normală standard (sau aproximativ normală standard).

### **Exemplul 13. testul $z$ pentru ipoteze normale**

IQ-ul este repartizat normal în populație conform unei repartiții  $N(100, 15^2)$ . Presupunem că cei mai mulți dintre studenții FMI au IQ-ul peste medie, deci vom formula următoarele ipoteze:

$$H_0 = \text{IQ-urile studenților FMI sunt repartizate identic cu cele ale populației generale} \\ = \text{IQ-urile studenților din FMI au o repartiție } N(100, 15^2).$$

$$H_A = \text{IQ-urile studenților FMI tend să fie mai mari decât cele ale populației generale} \\ = \text{media IQ-urilor studenților din FMI este mai mare ca 100.}$$

Observați că  $H_A$  este unilaterală.

Presupunem că testăm 9 studenți și aflăm că au un IQ mediu de  $\bar{x} = 112$ .

Putem respinge  $H_0$  la nivelul de semnificație  $\alpha = 0.05$ ?

**Răspuns.** Pentru a calcula  $p$ , întâi standardizăm datele: Sub ipoteza 0,  $\bar{x} \sim N(100, 15^2/9)$  și de aceea

$$z = \frac{\bar{x} - 100}{15/\sqrt{9}} = \frac{36}{15} = 2.4 \sim N(0, 1).$$

Adică, repartiția 0 pentru  $z$  este normală standard. Numim  $z$  o **statistică  $z$** , o vom utiliza ca statistică testului nostru.

Pentru o ipoteză alternativă la dreapta fraza "datele cel puțin la fel de extreme" este o coadă unilaterală la dreapta lui  $z$ . Valoarea  $p$  este atunci

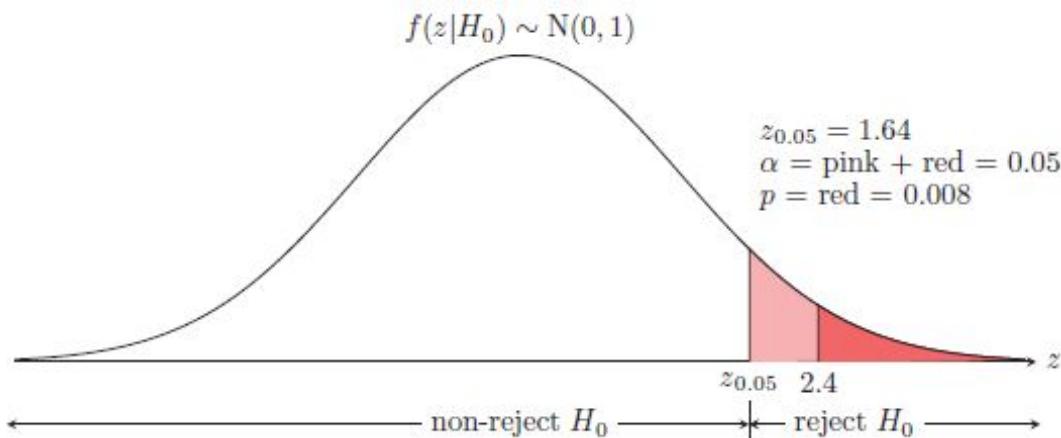
$$p = P(Z \geq 2.4) = 1 - \text{pnorm}(2.4, 0, 1) = 0.008197536.$$

Deoarece  $p \leq \alpha$  respingem ipoteza 0. Formulăm concluzia noastră astfel: Respingem ipoteza 0 în favoarea ipotezei că studenții FMI au IQ-uri mai mari în medie. Am făcut aceasta la nivelul de semnificație 0.05 cu o valoare  $p$  de 0.008.

Observații: 1. Media  $\bar{x} = 112$  este aleatoare: dacă facem experimentul din nou am putea obține o valoare diferită pentru  $\bar{x}$ .

2. Am fi putut folosi statistică  $\bar{x}$  direct. Standardizarea este preferabilă deoarece, cu practica, vom simți bine sensul diferențelor valorii  $z$ .

Justificarea respingerii lui  $H_0$  când  $p \leq \alpha$  este dată în următoarea figură.



În acest exemplu  $\alpha = 0.05$ ,  $z_{0.05} = 1.64$  și regiunea de respingere este domeniul de la dreapta lui  $z_{0.05}$ . De asemenea,  $z = 2.4$  și valoarea  $p$  este probabilitatea de la dreapta lui  $z$ . Figura ilustrează că

- $z = 2.4$  este în regiunea de respingere;
- aceasta este același lucru cu  $z$  este la dreapta lui  $z_{0.05}$ ;
- aceasta este același lucru cu probabilitatea de la dreapta lui  $z$  este mai mică decât 0.05,
- ceea ce înseamnă  $p < 0.05$ .

## 1.8 Mai multe exemple

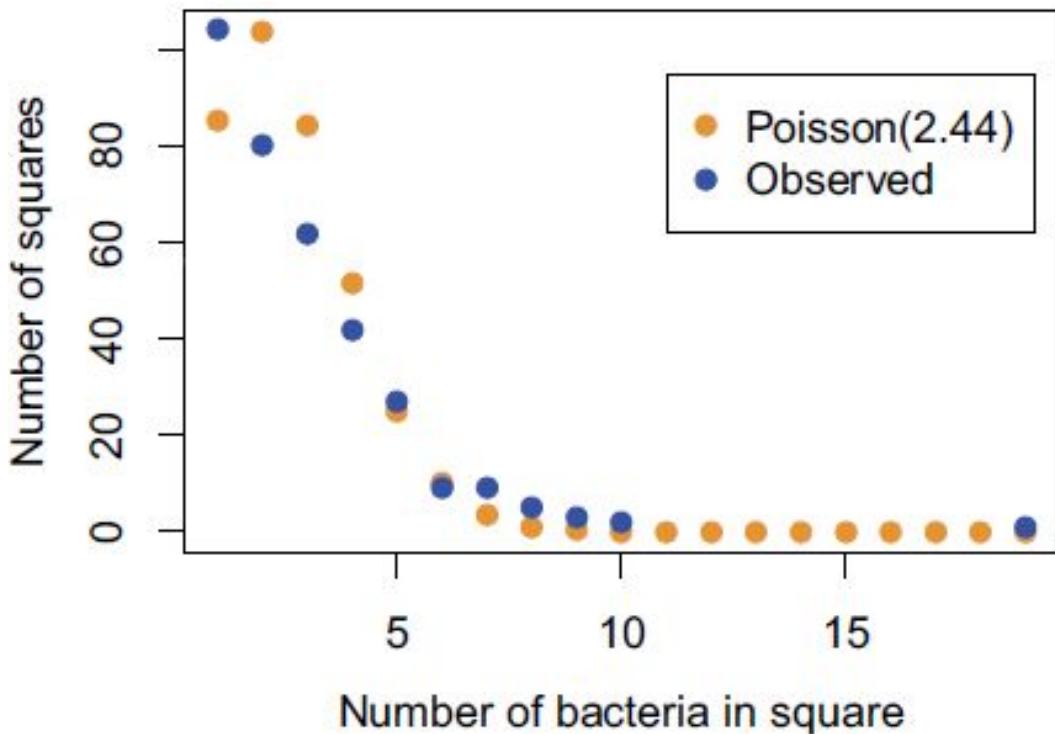
Testarea ipotezelor este larg utilizată în statistica deductivă.

**Exemplul 14.** Statistica  $\chi^2$  și bunătatea potrivirii.

Pentru a testa nivelul contaminării bacteriene, a fost vărsat lapte peste o rețea cu 400 de pătrate. Cantitatea de bacterii din fiecare pătrat a fost numărată. Rezumăm datele în tabelul de mai jos. Ultima linie a tabelului dă numărul de pătrate diferite care aveau o cantitate dată de bacterii.

Amount of bacteria	0	1	2	3	4	5	6	7	8	9	10	19
Number of squares	56	104	80	62	42	27	9	9	5	3	2	1

Cantitatea medie de bacterii pe pătrat este 2.44. Deoarece repartiția Poisson( $\lambda$ ) este folosită pentru a modela numărările evenimentelor relativ rare și parametrul  $\lambda$  este media repartiției, decidem să vedem dacă aceste date ar putea veni dintr-o repartiție Poisson. Pentru a face asta, întâi comparăm grafic frecvențele observate cu cele așteptate din Poisson(2.44).



Facem un test al ipotezei cu

$H_0$ : datele vin dintr-o repartiție Poisson(2.44).

$H_A$ : datele vin dintr-o repartiție diferită.

Folosim o statistică  $\chi^2$ , numită aşa deoarece ea are aproximativ o repartiție  $\chi^2$ . Pentru a calcula  $X^2$  întâi combinăm ultimele câteva celule în tabel astfel încât numărul așteptat minim este în jur de 5 (o regulă a degetului mare generală în acest joc.)

Numărul așteptat de pătrate cu o anumită cantitate de bacterii vine din considerarea a 400 de încercări dintr-o repartiție Poisson(2.44), de exemplu, cu  $l = 2.44$  numărul așteptat de pătrate cu 3 bacterii este  $400 \cdot e^{-l} \frac{l^3}{3!} = 84.4$ .

Statistica  $\chi^2$  este  $\sum \frac{(O_i - E_i)^2}{E_i}$ , unde  $O_i$  este numărul observat și  $E_i$  este numărul așteptat.

Number per square	0	1	2	3	4	5	6	> 6
Observed	56	104	80	62	42	27	9	20
Expected	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
Component of $X^2$	12.8	4.2	5.5	6.0	1.7	0.14	0.15	44.5

Adunând obținem  $X^2 = 74.99$ .

Doarece media (2.44) și numărul total de încercări (400) sunt fixate, cele 8 celule au doar 6 grade de libertate. Deci, presupunând  $H_0$ , statistica noastră  $X^2$  are (aproximativ) o repartiție  $\chi_6^2$ . Folosind această repartiție,

$P(X^2 > 74.99) = 0$ . Astfel, respingem decisiv ipoteza 0 în favoarea ipotezei alternative că repartitia nu este Poisson(2.44).

Pentru a analiza mai departe, privim la componentele individuale ale lui  $X^2$ . Sunt contribuții mari în coada distribuției, deci acolo potrivirea nu merge.

**Exemplul 15.** Testul  $t$  al lui Student.

Presupunem că vrem să comparăm un tratament medical pentru creșterea speranței de viață cu un placebo. Dăm la  $n$  oameni tratamentul și la  $m$  oameni placebo. Fie  $X_1, \dots, X_n$  numerele de ani pe care oamenii îi trăiesc după primirea tratamentului. Analog, fie  $Y_1, \dots, Y_m$  numerele de ani pe care oamenii îi trăiesc după primirea placebo. Fie  $\bar{X}$  și  $\bar{Y}$  mediile de selecție. Vrem să stim dacă diferența dintre  $\bar{X}$  și  $\bar{Y}$  este semnificativă statistic. Formulăm aceasta ca un test al ipotezei. Fie  $\mu_X$  și  $\mu_Y$  mediile (necunoscute).

$$H_0 : \mu_X = \mu_Y, \quad H_A : \mu_X \neq \mu_Y.$$

Cu anumite presupuneri și o formulă adecvată pentru eroarea standard unică  $s_p$  statistică testului  $t = \frac{\bar{X} - \bar{Y}}{s_p}$  are o repartiție  $t$  cu  $n + m - 2$  grade de libertate. Deci regiunea noastră de respingere este determinată de un prag  $t_0$  cu  $P(t > t_0) = \alpha$ .

## 2 Testarea semnificației ipotezei 0 II

### 2.1 Scopurile învățării

1. Să poată să listeze pașii comuni tuturor testelor semnificației ipotezei 0.
2. Să poată defini și calcula probabilitatea erorilor de tipul I și II.
3. Să poată să caute și să aplice  $t$ -teste pentru unul sau 2 eșantioane.

### 2.2 Introducere

Fiecare test face unele presupuneri despre date - adesea că sunt provenite dintr-o repartiție normală. Toate testele urmează același tipar. Doar calculul statisticii testului și tipul repartiției 0 se schimbă.

### 2.3 Configurarea și folosirea unui test de semnificație

Există o mulțime standard de pași care se fac pentru a configura și a utiliza un test al semnificației ipotezei 0.

1. Proiectează un experiment pentru a colecta datele și a alege o statistică  $x$  a testului pentru a fi calculată din date. Cerința cheie aici este să stim repartiția 0  $f(x|H_0)$ . Pentru a calcula puterea, trebuie să cunoaștem și

repartiția alternativă  $f(x|H_A)$ .

2. Decide dacă testul este unilateral sau bilateral bazat pe  $H_A$  și forma repartiției 0.
3. Alege un nivel de semnificație  $\alpha$  pentru respingerea ipotezei 0. Dacă se poate, calculează puterea corespunzătoare a testului.
4. Folosește experimentul pentru a colecta datele  $x_1, x_2, \dots, x_n$ .
5. Calculează statistica testului  $x$ .
6. Calculează valoarea  $p$  corespunzătoare lui  $x$  folosind repartiția 0.
7. Dacă  $p < \alpha$ , respinge ipoteza 0 în favoarea ipotezei alternative.

### Observații.

1. În loc de a alege un nivel de semnificație, putem alege o regiune de respingere și să respingem  $H_0$  dacă  $x$  este în această regiune. Nivelul de semnificație corespunzător este atunci probabilitatea ca  $x$  să fie în regiunea de respingere.
2. Ipoteza 0 este adesea numită "ipoteza precaută". Cu cât punem nivelul de semnificație mai mic, cu atât mai multe "dovezi" vor fi necesare înainte de a respinge ipoteza noastră precaută în favoarea unei alternative. Este o practică standard a publica valoarea  $p$  însăși, astfel încât alții să poată trage concluziile lor proprii.
3. **Un punct cheie de confuzie:** Un nivel de semnificație de 0.05 nu înseamnă că testul greșește doar în 5% din cazuri. Înseamnă că dacă ipoteza 0 este adevărată, atunci probabilitatea ca testul să o respingă din greșală este 5%. Puterea testului măsoară acuratețea testului când ipoteza alternativă este adevărată. Anume, puterea testului este probabilitatea de a respinge ipoteza 0 dacă ipoteza alternativă este adevărată. De aceea probabilitatea de a eșua fals să respingem ipoteza 0 este 1 minus puterea.

**Erori.** Putem rezuma aceste 2 tipuri de erori și probabilitățile lor după cum urmează:

Eroarea de tipul I = respingerea lui  $H_0$  când  $H_0$  este adevărată.

Eroarea de tipul II = nerespingerea lui  $H_0$  când  $H_A$  este adevărată.

$$\begin{aligned} P(\text{eroarea de tipul I}) &= \text{probabilitatea de a respinge fals } H_0 \\ &= P(\text{statistica testului este în regiunea de respingere} | H_0) \\ &= \text{nivelul de semnificație al testului} \\ P(\text{eroarea de tipul II}) &= \text{probabilitatea de a nu respinge fals } H_0 \\ &= P(\text{statistica testului este în regiunea de acceptare} | H_A) \\ &= 1 - \text{putere}. \end{aligned}$$

**Analogii utile.** În termeni de testare medicală pentru o boală, o eroare de tipul I este un rezultat fals pozitiv și o eroare de tipul II este un rezultat fals negativ. În termenii unui proces juridic, o eroare de tipul I este condamnarea

unui nevinovat și eroarea de tipul II este achitarea unui vinovat.

## 2.4 Înțelegerea unui test de semnificație

Întrebări de pus:

1. Cum au colectat datele? Care este schema experimentului?
2. Care sunt ipotezele 0 și alternativă?
3. Ce tip de test de semnificație a fost folosit?

Se potrivesc datele lor cu criteriile necesare pentru a utiliza acest tip de test?  
Cât de robust este testul la deviații de la aceste criterii?

4. De exemplu, unele teste compară 2 grupe de date presupunând că grupele sunt provenite din repartiții care au aceeași dispersie. Acest fapt trebuie verificat înainte de aplicarea testului. Adesea verificarea este făcută folosind alt test de semnificație proiectat pentru a compara dispersiile a 2 grupuri de date.

5. Cum este calculată valoarea  $p$ ?

Un test de semnificație vine cu o statistică a testului și o repartiție 0. În cele mai multe teste valoarea  $p$  este

$$p = P(\text{datele cel puțin la fel de extreme ca ce am obținut} | H_0)$$

Ce înseamnă "datele cel puțin la fel de extreme ca datele pe care le-am văzut"? I.e., este testul unilateral sau bilateral?

6. Care este nivelul de semnificație pentru acest test? Dacă  $p < \alpha$ , atunci experimentatorul va respinge  $H_0$  în favoarea  $H_A$ .

## 2.5 Teste $t$

Multe teste de semnificație presupun că datele provin dintr-o repartiție normală, deci înainte de a folosi un astfel de test trebuie să examinăm datele pentru a vedea dacă ipoteza normalității este rezonabilă. Reprezentarea unei histogramme este un start bun. Ca și testul  $z$ , testele  $t$  cu o selecție și cu 2 selecții pe care le considerăm mai jos pleacă de la această presupunere de normalitate.

### 2.5.1 Test $z$

Recapitulăm testul  $z$ .

Date: presupunem  $x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ , unde  $\mu$  este necunoscută și  $\sigma$  este cunoscută.

Ipoteza 0:  $\mu = \mu_0$  pentru o anumită valoare specifică  $\mu_0$ .

Statistica testului:  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \text{media standardizată.}$

Repartiția 0:  $f(z|H_0)$  este pdf a  $Z \sim N(0, 1)$ .

Valoare  $p$  unilaterală (partea dreaptă):  $p = P(Z > z|H_0)$ .

Valoare  $p$  unilaterală (partea stângă):  $p = P(Z < z|H_0)$ .

Valoare  $p$  bilaterală:  $p = P(|Z| > |z||H_0)$ .

**Exemplul 1.** Presupunem că avem datele care au o repartiție normală de medie necunoscută  $\mu$  și dispersie cunoscută 4. Fie ipoteza 0:  $\mu = 0$ . Fie ipoteza alternativă  $H_A : \mu > 0$ . Presupunem că obținem următoarele date:

$$1, 2, 3, 6, -1.$$

La un nivel de semnificație de  $\alpha = 0.05$  ar trebui să respingem ipoteza 0?

**Răspuns.** Sunt 5 date cu media  $\bar{x} = 2.2$ . Deoarece avem date normale cu o dispersie cunoscută ar trebui să folosim un  $z$  test. Statistica noastră  $z$  este

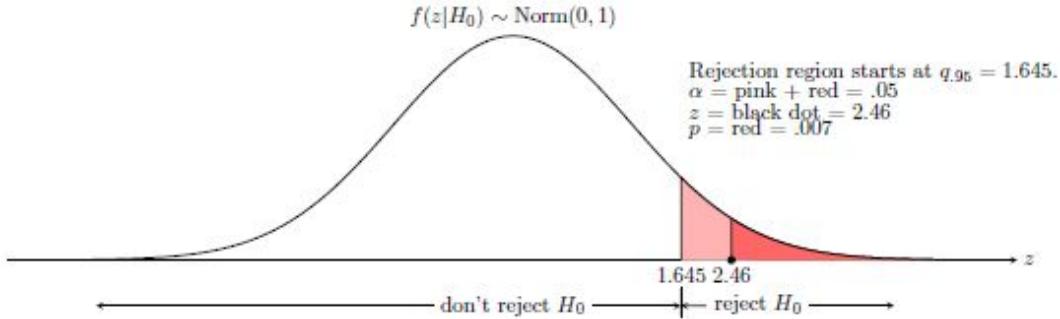
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.2 - 0}{2/\sqrt{5}} = 2.46.$$

Testul nostru este unilateral deoarece ipoteza alternativă este unilaterală. Deci (folosind R) valoarea  $p$  a noastră este

$$p = P(Z > z) = P(Z > 2.46) = 1 - \text{pnorm}(2.46) \approx 0.007.$$

Deoarece  $p < 0.05$ , respingem ipoteza 0 în favoarea ipotezei alternative  $\mu > 0$ .

Putem vizualiza testul astfel:



### 2.5.2 Repartiția $t$ a lui Student

”Student” este pseudonimul utilizat de William Gosset care a descris primul acest test și această repartiție. Vezi [http://en.wikipedia.org/wiki/Student's\\_t-test](http://en.wikipedia.org/wiki/Student's_t-test).

Repartiția  $t$  este simetrică și are formă de clopot ca repartiția normală. Are un parametru  $df$  care stă pentru grade de libertate. Pentru  $df$  mic repartiția  $t$  are mai multă probabilitate în cozile ei ca repartiția normală standard. Când  $df$  crește,  $t(df)$  devine din ce în ce mai asemănătoare cu repartiția normală

standard.

Iată o aplicație care arată  $t(df)$  și o compară cu repartiția normală standard: <http://mathlets.org/mathlets/t-distribution>.

Ca de obicei în R, funcțiile `pt`, `dt`, `qt`, `rt` corespund la cdf, pdf, cuantile, respectiv generarea unui eşantion aleator dintr-o repartiție  $t$ . Reamintim că puteți consulta ?`dt` în RStudio pentru a vedea fișierul de ajutor care specifică parametrii lui `dt`. De exemplu, `pt(1.65, 3)` calculează probabilitatea ca  $x$  să fie mai mic sau egal cu 1.65 dat fiind că  $x$  provine dintr-o repartiție  $t$  cu 3 grade de libertate, i.e.  $P(x \leq 1.65)$  dat fiind că  $x \sim t(3)$ .

### 2.5.3 Testul $t$ cu o selecție

Pentru testul  $z$ , am presupus că dispersia repartiției datelor este cunoscută. Totuși, adesea nu știm  $\sigma$  și trebuie să o estimăm din date. În aceste cazuri, folosim un test  $t$  cu o selecție în locul unui test  $z$  și media **Studentizată** în locul mediei standardizate.

Date: presupunem  $x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ , unde atât  $\mu$  cât și  $\sigma$  sunt necunoscute.

Ipoteza 0:  $\mu = \mu_0$  pentru o anumită valoare specifică  $\mu_0$ .

Statistica testului:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

unde

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Aici  $t$  este numită **media Studentizată** și  $s^2$  este numită **dispersia de selecție**. Ultima este o estimare a adevărării dispersiei  $\sigma^2$ .

Repartiția 0:  $f(t|H_0)$  este pdf pentru  $T \sim t(n-1)$ , repartiția  $t$  cu  $n-1$  grade de libertate.\*

Valoare  $p$  unilaterală (partea dreaptă):  $p = P(T > t|H_0)$ .

Valoare  $p$  unilaterală (partea stângă):  $p = P(T < t|H_0)$ .

Valoare  $p$  bilaterală:  $p = P(|T| > |t|)$ .

\*Există o teoremă (nu o presupunere) că dacă datele sunt normale cu media  $\mu_0$ , atunci media Studentizată are o repartiție  $t$ . Puteți vedea demonstrația în [http://en.wikipedia.org/wiki/Student%27s\\_t-distribution#Derivation](http://en.wikipedia.org/wiki/Student%27s_t-distribution#Derivation).

**Exemplul 2.** Presupunem acum că în exemplul precedent dispersia este necunoscută. Adică, avem date care au o repartiție normală cu medie necunoscută  $\mu$  și dispersie necunoscută  $\sigma^2$ . Presupunem că obținem aceleasi date:

$$1, 2, 3, 6, -1.$$

Ca mai sus, fie  $H_0 : \mu = 0$  și  $H_A : \mu > 0$ . La nivelul de semnificație de  $\alpha = 0.05$  ar trebui să respingem ipoteza 0?

**Răspuns.** Sunt 5 date cu media  $\bar{x} = 2.2$ . Deoarece avem date normale cu medie și dispersie necunoscute ar trebui să folosim un  $t$  test pentru o selecție. Calculând dispersia de selecție obținem

$$s^2 = \frac{1}{4}((1 - 2.2)^2 + (2 - 2.2)^2 + (3 - 2.2)^2 + (6 - 2.2)^2 + (-1 - 2.2)^2) = 6.7.$$

(Puteam folosi și comenziile în R

`x=c(1,2,3,6,-1)`

`var(x).`)

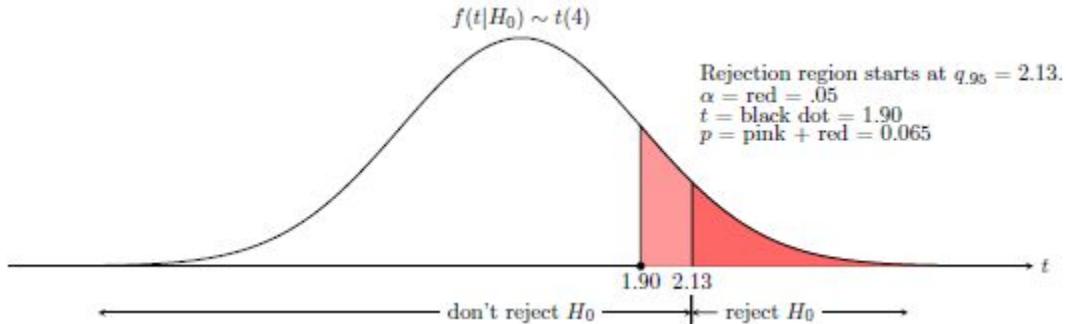
Statistica noastră  $t$  este

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.2 - 0}{\sqrt{6.7}/\sqrt{5}} = 1.900511.$$

Testul nostru este unilateral deoarece ipoteza alternativă este unilaterală. Deci (folosind R), valoarea  $p$  este

$$p = P(T > t) = P(T > 1.900511) = 1 - pt(1.900511, 4) = 0.06508106.$$

Deoarece  $p > 0.05$ , nu respingem ipoteza 0. Putem vizualiza testul astfel:



#### 2.5.4 Testul $t$ pentru 2 selecții cu dispersii egale

Considerăm în continuare cazul comparării mediilor a 2 selecții. De exemplu, putem fi interesați în compararea eficiențelor medii ale 2 tratamente medicale.

Date: Presupunem că avem 2 mulțimi de date provenite din repartiții normale

$$\begin{aligned} x_1, x_2, \dots, x_n &\sim N(\mu_1, \sigma^2) \\ y_1, y_2, \dots, y_m &\sim N(\mu_2, \sigma^2), \end{aligned}$$

unde mediile  $\mu_1$  și  $\mu_2$  și dispersia  $\sigma^2$  sunt toate necunoscute. Presupunem că cele 2 repartiții au **aceeași dispersie**. Sunt  $n$  date în primul grup și  $m$  date în al 2-lea.

Ipoteza 0:  $\mu_1 = \mu_2$  (valorile lui  $\mu_1$  și  $\mu_2$  nu sunt specificate).

Statistica testului:

$$t = \frac{\bar{x} - \bar{y}}{s_p},$$

unde  $s_p^2$  este **dispersia combinată**

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} \left( \frac{1}{n} + \frac{1}{m} \right).$$

Aici  $s_x^2$  și  $s_y^2$  sunt dispersiile de selecție ale  $x_i$ -urilor și respectiv  $y_j$ -urilor.

Repartiția 0:  $f(t|H_0)$  este pdf pentru  $T \sim t(n+m-2)$ .

Valoare  $p$  unilaterală (partea dreaptă):  $p = P(T > t|H_0)$ .

Valoare  $p$  unilaterală (partea stângă):  $p = P(T < t|H_0)$ .

Valoare  $p$  bilaterală:  $p = P(|T| > |t|)$ .

**Observația 1.** Unii autori folosesc o notație diferită. Ei definesc dispersia combinată ca

$$s_{p\text{-alți-autori}}^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

și ce am numit dispersia combinată ei arată că este dispersia estimată a lui  $\bar{x} - \bar{y}$ . Adică,

$$s_p^2 = s_{p\text{-alți-autori}}^2 \cdot \left( \frac{1}{n} + \frac{1}{m} \right) \approx s_{\bar{x}-\bar{y}}^2.$$

**Observația 2.** Există o versiune a  $t$ -testului cu 2 selecții care permite celor 2 grupe să aibă dispersii diferite. În acest caz statistica testului este mai complicată, dar R o tratează la fel de ușor.

**Exemplul 3.** Următoarele date vin dintr-un studiu real în care 1408 femei au fost interne la o maternitate pentru (i) motive medicale sau (ii) prin internare de urgență. Durata sarcinii este măsurată în săptămăni complete de la începutul ultimului ciclu. Putem rezuma datele astfel:

Medicale: 775 de observații cu  $\bar{x}_M = 39.08$  și  $s_M^2 = 7.77$ .

Urgențe: 633 de observații cu  $\bar{x}_E = 39.6$  și  $s_E^2 = 4.95$ .

Configurați și folosiți un test  $t$  cu 2 selecții pentru a investiga dacă durata medie diferă pentru cele 2 grupuri.

Ce presupuneri ați făcut?

**Răspuns.** Dispersia combinată pentru aceste date este

$$s_p^2 = \frac{774 \cdot 7.77 + 632 \cdot 4.95}{1406} \left( \frac{1}{775} + \frac{1}{633} \right) = 0.01866256.$$

Statistica  $t$  pentru repartiția 0 este

$$\frac{\bar{x}_M - \bar{x}_E}{s_P} = -3.806429.$$

Avem 1406 grade de libertate. Folosind R pentru a calcula valoarea  $p$  bilaterală, obținem

$$p = P(|T| > |t|) = 2 * \text{pt}(-3.806429, 1406) = 0.000147048.$$

$p$  este foarte mică, mult mai mică decât  $\alpha = 0.05$  sau  $\alpha = 0.01$ . De aceea respingem ipoteza 0 în favoarea alternativei că există o diferență între durele medii.

În loc să calculăm valoarea  $p$  bilaterală exact folosind o repartiție  $t$  am fi putut observa că, cu 1406 de grade de libertate, repartiția  $t$  este în esență normală standard și 3.806429 este aproape 4 deviații standard. Deci

$$p = P(|T| \geq 3.806429) \approx P(|Z| \geq 3.806429) < 0.001.$$

Am presupus că datele au fost normale și că cele 2 grupe au avut dispersii egale. Data fiind diferența mare dintre dispersiile de selecție, această presupunere nu poate fi garantată.

De fapt, sunt alte teste de semnificație care testează dacă datele sunt aproximativ normale și dacă cele 2 grupe au aceeași dispersie. În practică se pot aplica acestea întâi pentru a determina dacă un test  $t$  este potrivit în primul rând.

[http://en.wikipedia.org/wiki/Normality\\_test](http://en.wikipedia.org/wiki/Normality_test)

[http://en.wikipedia.org/wiki/F-test\\_of\\_equality\\_of\\_variances](http://en.wikipedia.org/wiki/F-test_of_equality_of_variances)

# Curs 17

Cristian Niculescu

## 1 Testarea semnificației ipotezei 0 III

### 1.1 Scopurile învățării

1. Fiind date ipotezele și datele, să poată identifica un test de semnificație potrivit dintr-o listă de teste uzuale.
2. Fiind date ipotezele, datele și un test de semnificație sugerat, să știe cum să caute detaliiile și să aplice testul.

### 1.2 Introducere

Toate testele de semnificație au același model de bază în proiectarea și implementarea lor.

#### Proiectarea unui test al semnificației ipotezei 0 (NHST):

Specificăm ipoteza 0 și ipotezele alternative.

Alegem o statistică a testului ale cărei repartiție 0 și repartiție alternativă (repartiții alternative) sunt cunoscute.

Specificăm o regiune de respingere. Cel mai adesea acest lucru este făcut implicit specificând un nivel de semnificație  $\alpha$  și o metodă pentru calculul  $p$ -valorilor bazată pe cozile repartiției 0.

Calculăm puterea folosind repartiția alternativă (repartițiile alternative).

#### Folosirea unui NHST:

Colectăm datele și calculăm statistica testului.

Verificăm dacă statistica testului este în regiunea de respingere. Cel mai adesea acest lucru este făcut implicit verificând dacă  $p < \alpha$ . Dacă este așa, "respingem ipoteza 0 în favoarea ipotezei alternative". Altfel concluzionăm "datele nu sprijină respingerea ipotezei 0".

Observați formularea prudentă: când eșuăm în a respinge  $H_0$ , nu concluzionăm că  $H_0$  este adevărată. Eșecul respingerii poate avea alte cauze. De exemplu, poate nu avem destule date pentru a distinge clar  $H_0$  și  $H_A$ , în timp ce mai multe date ar indica respingerea lui  $H_0$ .

### 1.3 Parametrii populației și statistici de selecție

**Exemplul 1.** Dacă alegem aleator 10 bărbați dintr-o populație și le măsurăm înălțimile, spunem că **am selectat înălțimile** din populație. În acest caz **media de selecție**, să spunem  $\bar{x}$ , este media înălțimilor selectate. Este o statistică și îi știm explicit valoarea. Pe de altă parte, adevărată înălțime medie a populației, să spunem  $\mu$ , este necunoscută și putem doar să estimăm valoarea ei. Numim  $\mu$  **parametru al populației**.

Principalul scop al testării semnificației este folosirea statisticilor de selecție pentru a trage concluzii despre parametrii de selecție. De exemplu, putem testa dacă înălțimea medie a bărbaților dintr-o populație dată este mai mare ca 1.78 m.

### 1.4 O galerie a testelor de semnificație uzuale legate de repartitia normală

Toate aceste teste presupun **date normale**.

Repartițiile 0 pentru aceste teste sunt **toate legate de repartitia normală** prin formule explicite. Argumentele sunt accesibile oricui știe analiză matematică și este interesat de înțelegerea lor. Dat fiind numele oricărei repartiții, putem căuta online detaliiile construcției și proprietăților ei. De asemenea putem folosi R pentru a explora repartitia numeric și grafic.

Când analizăm datele cu oricare din aceste teste, un lucru de importanță cheie este să verificăm că presupunerile sunt adevărate sau cel puțin aproximativ adevărate. De exemplu, nu ar trebui să folosim un test care presupune că datele sunt normale până nu am verificat că datele sunt aproximativ normale.

#### 1.4.1 Testul $z$

Utilizare: Testăm dacă media populației este egală cu o medie ipotetică.

Date:  $x_1, x_2, \dots, x_n$ .

Presupuneri: Datele sunt selecții normale independente:

$$x_i \sim N(\mu, \sigma^2) \text{ unde } \mu \text{ este necunoscută, dar } \sigma \text{ este cunoscută.}$$

$H_0$ : Pentru un  $\mu_0$  specificat,  $\mu = \mu_0$ .

$H_A$ :

Bilaterală:  $\mu \neq \mu_0$ ;

unilaterală mai mare:  $\mu > \mu_0$ ;

unilaterală mai mică:  $\mu < \mu_0$ .

Statistica testului:  $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ .

Repartiția 0:  $f(z|H_0)$  este pdf a lui  $Z \sim N(0, 1)$ .  
valoarea  $p$ :

$$\begin{array}{lll} \text{Bilaterală:} & p = P(|Z| > |z|) & = 2 * (1 - \text{pnorm}(\text{abs}(z), 0, 1)); \\ \text{unilaterală mai mare:} & p = P(Z > z) & = 1 - \text{pnorm}(z, 0, 1); \\ \text{unilaterală mai mică:} & p = P(Z < z) & = \text{pnorm}(z, 0, 1). \end{array}$$

Codul R: Nu pare a fi o singură funcție R pentru a folosi un tezt  $z$ . Desigur, se poate folosi R pentru a calcula  $z$  și valoarea  $p$ .

**Exemplul 2.** Vezi exemplul 13 din cursul 16.

#### 1.4.2 Testul $t$ de o selecție pentru medie

Utilizare: Testăm dacă media populației este egală cu o medie ipotetică.

Date:  $x_1, x_2, \dots, x_n$ .

Presupuneri: Datele sunt selecții normale independente:

$$x_i \sim N(\mu, \sigma^2) \text{ unde atât } \mu \text{ cât și } \sigma \text{ sunt necunoscute.}$$

$H_0$ : Pentru un  $\mu_0$  specificat,  $\mu = \mu_0$ .

$H_A$ :

$$\begin{array}{l} \text{Bilaterală: } \mu \neq \mu_0; \\ \text{unilaterală mai mare: } \mu > \mu_0; \\ \text{unilaterală mai mică: } \mu < \mu_0. \end{array}$$

Statistica testului:  $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ ,

$$\text{unde } s^2 \text{ este dispersia de selecție: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Repartiția 0:  $f(t|H_0)$  este pdf a lui  $T \sim t(n-1)$  (repartiția  $t$  a lui Student cu  $n-1$  grade de libertate).

Valoarea  $p$ :

$$\begin{array}{lll} \text{Bilaterală:} & p = P(|T| > |t|) & = 2 * (1 - \text{pt}(\text{abs}(t), n-1)); \\ \text{unilaterală mai mare:} & p = P(T > t) & = 1 - \text{pt}(t, n-1); \\ \text{unilaterală mai mică:} & p = P(T < t) & = \text{pt}(t, n-1). \end{array}$$

Exemplu de cod R: Pentru datele  $x = 1, 3, 5, 7, 2$  putem face un test  $t$  de o selecție cu  $H_0: \mu = 2.5$  folosind comenziile în R:

```
x=c(1,3,5,7,2)
t.test(x, mu=2.5)
```

Acestea vor da câteva informații, inclusiv media datelor, valoarea  $t$  și valoarea  $p$  bilaterală. Vezi ajutorul pentru această funcție pentru alte setări ale argumentelor.

**Exemplul 3.** Vezi exemplul 2 din partea a 2-a a cursului 16.

#### 1.4.3 Testul $t$ cu 2 selecții pentru compararea mediilor

##### Cazul dispersiilor egale

Începem descriind testul [presupunând dispersii egale](#).

Utilizare: Testăm dacă mediile populației din 2 populații diferă printr-o cantitate ipotetică.

Date:  $x_1, x_2, \dots, x_n$  și  $y_1, y_2, \dots, y_m$ .

Presupuneri: Ambele grupe de date sunt selecții normale independente:

$$\begin{aligned}x_i &\sim N(\mu_x, \sigma^2), \\y_j &\sim N(\mu_y, \sigma^2),\end{aligned}$$

unde atât  $\mu_x$  cât și  $\mu_y$  sunt necunoscute și posibil diferite. Dispersia  $\sigma^2$  este necunoscută, dar aceeași pentru ambele grupe.

$H_0$ : Pentru un  $\mu_0$  specificat:  $\mu_x - \mu_y = \mu_0$ .

$H_A$ :

Bilaterală:	$\mu_x - \mu_y \neq \mu_0$ ;
unilaterală mai mare:	$\mu_x - \mu_y > \mu_0$ ;
unilaterală mai mică:	$\mu_x - \mu_y < \mu_0$ .

Statistica testului:  $t = \frac{\bar{x} - \bar{y} - \mu_0}{s_p}$ , unde  $s_x^2$  și  $s_y^2$  sunt dispersiile de selecție ale datelor  $x$ , respectiv  $y$  și  $s_p^2$  este (uneori numită) dispersia de selecție combinată:

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} \left( \frac{1}{n} + \frac{1}{m} \right) \text{ și } df = n+m-2.$$

Repartiția 0:  $f(t|H_0)$  este pdf a lui  $T \sim t(df)$ , repartiția  $t$  cu  $df = n+m-2$  grade de libertate.

Valoarea  $p$ :

Bilaterală:	$p = P( T  >  t ) = 2 * (1 - pt(abs(t), df))$ ;
unilaterală mai mare:	$p = P(T > t) = 1 - pt(t, df)$ ;
unilaterală mai mică:	$p = P(T < t) = pt(t, df)$ .

Codul R: Funcția din R `t.test` va rula un test  $t$  cu 2 selecții cu dispersii egale setând argumentul `var.equal=TRUE`.

**Exemplul 4.** Vezi exemplul 3 din a 2-a parte a cursului 16.

**Observații.** 1. Cel mai adesea testul este făcut cu  $\mu_0 = 0$ . Adică, ipoteza 0 este că **mediile sunt egale**, i.e.  $\mu_x - \mu_y = 0$ .

2. Dacă datele  $x$  și  $y$  au aceeași lungime,  $n$ , atunci formula pentru  $s_p^2$  devine mai simplă:

$$s_p^2 = \frac{s_x^2 + s_y^2}{n}.$$

### Cazul dispersiilor inegale

Există o formă a testului  $t$  pentru cazul **când dispersiile nu sunt presupuse egale**. Uneori este numit **testul t al lui Welch**.

Acesta arată exact la fel cu excepția unor mici schimbări în presupunerii și în formula dispersiei combinate:

Utilizare: Testăm dacă mediile populației din 2 populații diferă printr-o cantitate ipotetică.

Date:  $x_1, x_2, \dots, x_n$  și  $y_1, y_2, \dots, y_m$ .

Presupunerii: Ambele grupe de date sunt selecții normale independente:

$$\begin{aligned} x_i &\sim N(\mu_x, \sigma_x^2), \\ y_j &\sim N(\mu_y, \sigma_y^2), \end{aligned}$$

unde atât  $\mu_x$  cât și  $\mu_y$  sunt necunoscute și posibil diferite. Dispersiile  $\sigma_x^2$  și  $\sigma_y^2$  sunt necunoscute și **nu sunt presupuse a fi egale**.

$H_0, H_A$ : Exact aceeași ca în cazul dispersiilor egale.

Statistica testului:  $t = \frac{\bar{x} - \bar{y} - \mu_0}{s_p}$ , unde  $s_x^2$  și  $s_y^2$  sunt dispersiile de selecție ale datelor  $x$ , respectiv  $y$  și  $s_p^2$  este (uneori numită) dispersia de selecție combinată:

$$s_p^2 = \frac{s_x^2}{n} + \frac{s_y^2}{m} \text{ și } df = \frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/n)^2/(n-1) + (s_y^2/m)^2/(m-1)}.$$

Repartiția 0:  $f(t|H_0)$  este pdf a lui  $T \sim t(df)$ , repartiția  $t$  cu  $df$  grade de libertate.

Valoarea  $p$ : Exact aceeași ca în cazul dispersiilor egale.

Codul R: Funcția din R `t.test` va rula un test  $t$  în acest caz setând argumentul `var.equal=FALSE`.

### Testul $t$ cu 2 selecții în perechi

Când datele vin火 în perechi  $(x_i, y_i)$ , putem folosi **testul t cu 2 selecții în perechi** (după de verificăm că presupunerile sunt valide!)

**Exemplul 5.** Pentru a măsura eficiența unei medicații de scădere coles-terolului, putem testa fiecare subiect înainte și după tratamentul cu medicamentul. Deci pentru fiecare subiect avem o pereche de măsurători:  $x_i$  = nivelul de colesterol înainte de tratament și  $y_i$  = nivelul de colesterol după

tratament.

**Exemplul 6.** Pentru a măsura eficiența unui tratament pentru cancer putem pune în pereche fiecare subiect care a primit tratamentul cu unul care nu l-a primit. În acest caz am vrea să punem în perechi subiecții care sunt similari în termeni de stadiu al bolii, vârstă, sex, etc.

Utilizare: Testăm dacă diferența medie dintre valorile perechilor dintr-o populație este egală cu o valoare ipotetică.

Date:  $x_1, x_2, \dots, x_n$  și  $y_1, y_2, \dots, y_n$  trebuie să aibă aceeași lungime.

Presupuneri: Diferențele  $w_i = x_i - y_i$  sunt independente dintr-o repartitie normală  $N(\mu, \sigma^2)$ , unde  $\mu$  și  $\sigma$  sunt necunoscute.

**OBSERVAȚIE:** Acesta este chiar un test  $t$  de o selecție folosind  $w_i$ .

$H_0$ : Pentru un  $\mu_0$  specificat,  $\mu = \mu_0$ .

$H_A$ :

- Bilaterală:  $\mu \neq \mu_0$ ;
- unilaterală mai mare:  $\mu > \mu_0$ ;
- unilaterală mai mică:  $\mu < \mu_0$ .

Statistica testului:  $t = \frac{\bar{w} - \mu_0}{s/\sqrt{n}}$ ,

unde  $s^2$  este dispersia de selecție:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (w_i - \bar{w})^2$ .

Repartiția 0:  $f(t|H_0)$  este pdf a lui  $T \sim t(n-1)$  (repartiția  $t$  a lui Student cu  $n-1$  grade de libertate).

Valoarea  $p$ :

$$\begin{array}{lll} \text{Bilaterală:} & p = P(|T| > |t|) & = 2 * (1 - pt(\text{abs}(t), n-1)); \\ \text{unilaterală mai mare:} & p = P(T > t) & = 1 - pt(t, n-1); \\ \text{unilaterală mai mică:} & p = P(T < t) & = pt(t, n-1). \end{array}$$

Cod R: Funcția `t.test` din R va face un test cu 2 selecții în perechi dacă punem argumentul `paired=TRUE`. Putem de asemenea să facem un test  $t$  de o selecție pentru  $x - y$ .

**Exemplul 7.** Acest exemplu este din John Rice, *Mathematical Statistics and Data Analysis*, 2nd edition, p. 412.

Pentru a studia efectul fumatului asupra agregării trombocitelor, P.H. Levine (An acute effect of cigarette smoking on platelet function, *Circulation*, 48, 619-623, 1973) a luat sânge de la 11 subiecți înainte și după ce au fumat o cără și a măsurat gradul în care trombocitele s-au agregat. Iată datele:

Before	25	25	27	44	30	67	53	53	52	60	28
After	27	29	37	56	46	82	57	80	61	59	43
Difference	2	4	10	12	16	15	4	27	9	-1	15

Ipoteza 0 este că fumatul nu a avut efect asupra agregării trombocitelor, i.e. diferența ar trebui să aibă media  $\mu_0 = 0$ . Iată codul R pentru un test  $t$  cu 2 selecții pentru a testa această ipoteză:

```
before.cig=c(25,25,27,44,30,67,53,53,52,60,28)
after.cig=c(27,29,37,56,46,82,57,80,61,59,43)
t.test(after.cig,before.cig,mu=0,paired=TRUE)
```

Iată rezultatul:

Paired t-test

```
data: after.cig and before.cig
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
4.91431 15.63114
sample estimates:
mean of the differences
10.27273
Obțineam aceleași rezultate cu testul  $t$  de o selecție:
t.test(after.cig-before.cig,mu=0)
```

#### 1.4.4 ANOVA unidirecțională (Testul $F$ pentru medii egale)

Utilizare: Testăm dacă mediile populațiilor din  $n$  grupe coincid.

Date ( $n$  grupe,  $m$  date din fiecare grup)

$$\begin{aligned} &x_{11}, \quad x_{12}, \quad \dots, \quad x_{1m} \\ &x_{21}, \quad x_{22}, \quad \dots, \quad x_{2m} \\ &\dots \\ &x_{n1}, \quad x_{n2}, \quad \dots, \quad x_{nm} \end{aligned}$$

Presupuneri: Datele pentru fiecare grup sunt independente normale din repartiții cu medii (posibil) diferite, dar [aceeași dispersie](#):

$$\begin{aligned} x_{1j} &\sim N(\mu_1, \sigma^2) \\ x_{2j} &\sim N(\mu_2, \sigma^2) \\ &\dots \\ x_{nj} &\sim N(\mu_n, \sigma^2) \end{aligned}$$

Mediile grupurilor  $\mu_i$  sunt necunoscute și posibil diferite. Dispersia  $\sigma$  este cunoscută, dar aceeași pentru toate grupurile.

$H_0$ : Toate mediile sunt identice  $\mu_1 = \mu_2 = \dots = \mu_n$ .

$H_A$ : Nu toate mediile sunt egale.

Statistica testului:  $w = \frac{MS_B}{MS_W}$ , unde

$$\begin{aligned}\bar{x}_i &= \text{media grupului } i \\ &= \frac{x_{i1} + x_{i2} + \dots + x_{im}}{m}.\end{aligned}$$

$\bar{x}$  = media tuturor datelor.

$$\begin{aligned}s_i^2 &= \text{dispersia de selecție a grupului } i \\ &= \frac{1}{m-1} \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2.\end{aligned}$$

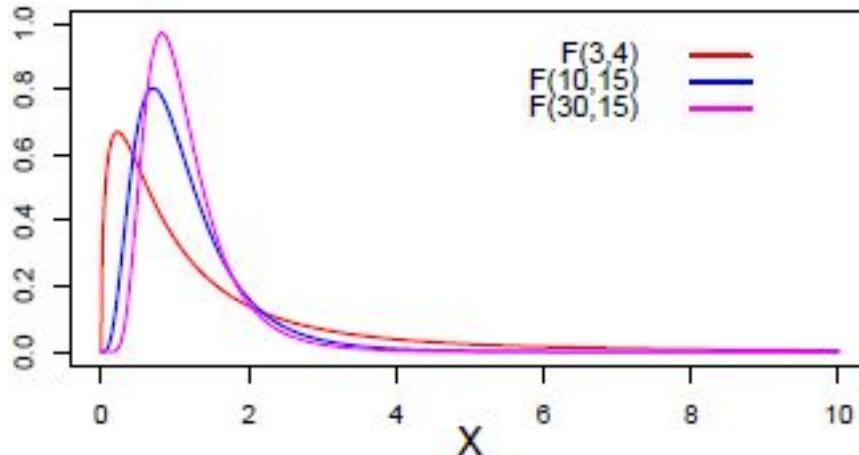
$$\begin{aligned}MS_B &= \text{dispersia între grupuri} \\ &= m \cdot \text{dispersia de selecție a mediilor grupurilor} \\ &= \frac{m}{n-1} \sum_{i=1}^n (\bar{x}_i - \bar{x})^2.\end{aligned}$$

$$\begin{aligned}MS_W &= \text{media dispersiilor grupurilor} \\ &= \text{media de selecție a } s_1^2, \dots, s_n^2 \\ &= \frac{s_1^2 + s_2^2 + \dots + s_n^2}{n}.\end{aligned}$$

Ideea: Dacă  $\mu_i$  sunt toate egale, acest raport ar trebui să fie aproape de 1. Dacă ele nu sunt egale, atunci  $MS_B$  ar trebui să fie mai mare, în timp ce  $MS_W$  ar trebui să rămână cam aceeași, deci  $w$  ar trebui să fie mai mare.

Repartiția 0:  $f(w|H_0)$  este pdf a lui  $W \sim F(n-1, n(m-1))$ .

Aceasta este repartiția  $F$  cu  $n-1$  și  $n(m-1)$  grade de libertate. Câteva repartiții  $F$  sunt reprezentate astfel:



Valoarea  $p$ :  $p = P(W > w) = 1 - \text{pf}(w, n-1, n*(m-1))$ .

**Observații.** 1. ANOVA testează dacă toate mediile sunt egale. Nu testează

dacă unele din medii sunt egale.

2. Există un test unde dispersiile nu sunt presupuse egale.
3. Există un test unde grupurile nu au toate același număr de date.
4. R are o funcție `aov()` pentru teste ANOVA. Vezi:

[https://personality-project.org/r/r\\_guide/r\\_anova.html#oneway](https://personality-project.org/r/r_guide/r_anova.html#oneway)  
<http://en.wikipedia.org/wiki/F-test>.

**Exemplul 8.** Tabelul arată nivelul de durere perceput de pacienți (pe o scară de la 1 la 6) după 3 proceduri medicale diferite.

$T_1$	$T_2$	$T_3$
2	3	2
4	4	1
1	6	3
5	1	3
3	4	5

(1) Faceți un test  $F$  care să compare mediile acestor tratamente.

(2) Bazându-vă pe test, ce puteți concluziona despre tratamente?

**Răspuns.** Folosind codul de mai jos, statistica  $F$  este 0.325 și valoarea  $p$  este 0.729. La orice nivel rezonabil de semnificație nu respingem ipoteza 0 că nivelul mediu de durere este același pentru toate 3 tratamentele.

Nu este rezonabil să concluzionăm că ipoteza 0 este adevărată. Cu doar 5 date pe procedură ne poate lipsi puterea de a distinge mediile diferite.

#### Codul R pentru test

```
# DATE ----
T1=c(2,4,1,5,3)
T2=c(3,4,6,1,4)
T3=c(2,1,3,3,5)
procedure=c(rep('T1',length(T1)),rep('T2',length(T2)),rep('T3',length(T3)))
pain=c(T1,T2,T3)
data.pain=data.frame(procedure,pain)
aov.data=aov(pain~procedure,data=data.pain) # face analiza dispersiei
print(summary(aov.data))# arată tabelul rezumatului
```

Rezumatul arată o valoare  $p$  (arătată ca  $\text{Pr}(>F)$ ) de 0.729. De aceea nu respingem ipoteza 0 că toate 3 mediile sunt egale.

#### 1.4.5 Testul $\chi^2$ pentru bunătatea potrivirii

Acesta este un test despre cât de bine o repartiție de probabilitate ipotetică se potrivește cu datele. Statistica testului este numită **statistică  $\chi^2$**  și repartitia

0 asociată statisticii  $\chi^2$  este **repartiția  $\chi^2$** . Ea este notată cu  $\chi^2(df)$ , unde parametrul  $df$  este numărul de **grade de libertate**.

Presupunem că avem o funcție masă de probabilitate necunoscută dată de următorul tabel.

<b>Outcomes</b>	$\omega_1$	$\omega_2$	...	$\omega_n$
<b>Probabilities</b>	$p_1$	$p_2$	...	$p_n$

În testul  $\chi^2$  pentru bunătatea potrivirii ipoteza este o mulțime de valori pentru probabilități. Tipic, ipoteza este că probabilitățile au o repartiție cunoscută cu anumiți parametri, de exemplu binomială, Poisson, multinomială. Apoi testul încearcă să determine dacă această mulțime de probabilități ar fi putut genera rezonabil datele colectate.

Utilizare: Testăm dacă date discrete se potrivesc cu o anumită funcție masă de probabilitate finită.

Date: Câte un număr observat  $O_i$  pentru fiecare rezultat posibil  $\omega_i$ .

Presupuneri: Niciuna.

$H_0$ : Datele provin dintr-o anumită repartiție discretă.

$H_A$ : Datele provin dintr-o repartiție diferită.

Statistica testului: Datele constau din numerele observate  $O_i$  pentru fiecare  $\omega_i$ . Din tabelul de probabilitate al ipotezei 0 obținem o mulțime de numere așteptate  $E_i$ . Sunt 2 statistici pe care le putem folosi:

$$\text{Statistica raportului de verosimilitate } G = 2 \sum O_i \ln \left( \frac{O_i}{E_i} \right)$$

$$\text{Statistica } \chi^2 \text{ a lui Pearson } X^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

Există o teoremă că sub ipoteza 0  $X^2 \approx G$  și ambele sunt aproximativ  $\chi^2$ . Înainte de calculatoare,  $X^2$  era folosită deoarece era mai ușor de calculat. Acum, este mai bine să folosim  $G$  deși încă veți vedea  $X^2$  folosită destul de des.

Grade de libertate  $df$ : Pentru testele  $\chi^2$  numărul gradelor de libertate poate fi un pic complicat. În acest caz  $df = n - 1$ . Este calculat ca numărul de celule care pot fi completate liber în concordanță cu statisticile folosite pentru a calcula numerele așteptate din celule presupunând  $H_0$ .

Repartiția 0: Presupunând  $H_0$ , ambele statistici au (aproximativ) o repartiție  $\chi^2$  cu  $df$  grade de libertate. Adică atât  $f(G|H_0)$  cât și  $f(X^2|H_0)$  au aceeași pdf ca  $Y \sim \chi^2(df)$ .

Valoarea  $p$ :

$$\begin{aligned} p &= P(Y > G) &= 1 - \text{pchisq}(G, df) \\ p &= P(Y > X^2) &= 1 - \text{pchisq}(X^2, df) \end{aligned}$$

Codul R: Funcția R `chisq.test` poate fi folosită pentru a face calcule pentru un test  $\chi^2$  folosind  $X^2$ . Pentru  $G$ , trebuie să facem calculele ”cu mâna” (ajutați de R) sau să găsim un pachet care are o funcție. (Probabil va fi numită `likelihood.test` sau `G.test`.)

**Observație.** Când statistica raportului de verosimilitate  $G$  este folosită, testul este de asemenea numit `test G` sau `testul raportului de verosimilitate`.

**Exemplul 9.** [Primul exemplu  \$\chi^2\$](#) . Presupunem că avem un experiment care produce date numerice. Pentru acest experiment rezultatele posibile sunt 0, 1, 2, 3, 4, 5 sau mai mult. Facem 51 de încercări și numărăm frecvența fiecarui rezultat, obținând următoarele date:

Outcomes	0	1	2	3	4	$\geq 5$
Observed counts	3	10	15	13	7	3

Presupunem că ipoteza noastră  $H_0$  este că datele provin din 51 de încercări ale repartiției binomiale(8,0.5) și ipoteza noastră alternativă  $H_A$  este că datele provin dintr-o altă repartitione.

- Faceți un tabel al numerelor observate și așteptate.
- Calculați statistica raportului de verosimilitate  $G$  și statistica  $\chi^2$  a lui Pearson  $X^2$ .
- Calculați numărul de grade de libertate ale repartiției 0.
- Calculați valorile  $p$  corespunzătoare lui  $G$  și  $X^2$ .

**Răspuns.** 1. Presupunând  $H_0$ , datele provin dintr-o repartitione binomială(8,0.5).

Avem 51 de observații, deci numărul așteptat pentru fiecare rezultat este 51 · probabilitatea lui. Folosind R se obține:

Rezultate	0	1	2	3	4	$\geq 5$
Numere observate	3	10	15	13	7	3
Probabilități $H_0$	0.00390625	0.03125	0.109375	0.21875	0.2734375	0.3632812
Numere așteptate	0.1992188	1.59375	5.578125	11.15625	13.9453125	18.52734

- Folosind formulele de mai sus calculăm  $X^2 = 116.4056$  și  $G = 66.08122$ .
- Singura statistică folosită în calculul numerelor așteptate a fost numărul total de observații 51. Deci, numărul de grade de libertate este 5, i.e. putem pune 5 dintre numerele din celule liber și ultimul este determinat de cerința ca totalul să fie 51.
- Valorile  $p$  sunt  $p_G = 1 - \text{pchisq}(G, 5)$  și  $p_{X^2} = 1 - \text{pschisq}(X^2, 5)$ . Ambele valori  $p$  sunt efectiv 0. Pentru aproape orice nivel de semnificație respingem  $H_0$  în favoarea  $H_A$ .

**Exemplul 10.** ([Grade de libertate](#).) Presupunem că avem aceleași date ca în exemplul precedent, dar ipoteza 0 a noastră este că datele vin din încercări independente ale repartiției binomiale(8,  $\theta$ ), unde  $\theta$  poate fi oricât între 0 și 1.  $H_A$  este că datele vin dintr-o altă repartitione. În acest caz trebuie să estimăm

$\theta$  din date, de exemplu folosind MLE. În total am calculat 2 valori din date: numărul total de date și estimarea lui  $\theta$ . Deci, numărul de grade de libertate este  $6 - 2 = 4$ .

**Exemplul 11. Experimentele genetice ale lui Mendel.** (Adaptat din Rice, *Mathematical Statistics and Data Analysis*, 2nd ed., exemplul C, p. 314).

În unul din experimentele lui pe mazăre, Mendel a încrucișat 556 masculi netezi, galbeni cu femele verzi zbârcite. Presupunând că genele netede și zbârcite apar cu frecvență egală ne-am așteptă ca  $1/4$  din populația de mazăre să aibă 2 gene netede ( $SS$ ),  $1/4$  să aibă 2 gene zbârcite ( $ss$ ) și  $1/2$  rămasă să fie heterozigoți  $Ss$ . De asemenea așteptăm aceste fracții pentru gene galbene ( $Y$ ) și verzi ( $y$ ). Dacă genele de culoare și netezime sunt moștenite independent și cele netede și galbene sunt ambele dominante ne-am aștepta la următorul tabel de frecvențe pentru fenotipuri.

	Yellow	Green	
Smooth	$9/16$	$3/16$	$3/4$
Wrinkled	$3/16$	$1/16$	$1/4$
	$3/4$	$1/4$	1

Tabelul de probabilitate pentru ipoteza 0

Deci, din 556 de încrucișări, numărul așteptat de păstăi netede galbene este  $556 \cdot 9/16 = 312.75$ . Analog pentru celelalte posibilități. Iată un tabel care dă numerele observate și așteptate din experimentele lui Mendel.

	Observed count	Expected count
Smooth yellow	315	312.75
Smooth green	108	104.25
Wrinkled yellow	102	104.25
Wrinkled green	31	34.75

Ipoteza 0 este că numerele observate sunt selecții aleatoare repartizate conform tabelului de frecvențe de mai sus. Folosim numerele să calculăm statisticile noastre.

Statistica raportului de verosimilitate este

$$\begin{aligned}
G &= 2 \sum O_i \ln \left( \frac{O_i}{E_i} \right) \\
&= 2 \left( 315 \ln \left( \frac{315}{312.75} \right) + 108 \ln \left( \frac{108}{104.25} \right) + 102 \ln \left( \frac{102}{104.25} \right) + 31 \ln \left( \frac{31}{34.75} \right) \right) \\
&= 0.6184391.
\end{aligned}$$

Statistica  $\chi^2$  a lui Pearson este

$$\begin{aligned}
X^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\
&= \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(102 - 104.25)^2}{104.25} + \frac{(31 - 34.75)^2}{34.75} \\
&= 0.6043165.
\end{aligned}$$

Cele 2 statistici sunt foarte apropiate. Aceasta este cazul de obicei. În general statistica raportului de verosimilitate este mai robustă și ar trebui preferată. Numărul de grade de libertate este 3, deoarece sunt 4 cantități observate și o relație între ele: suma lor este 556. Deci, sub ipoteza 0,  $G$  are o repartiție  $\chi^2(3)$ . Folosind R pentru a calcula valoarea  $p$  obținem

$$p = 1 - \text{pchisq}(0.6184391, 3) = 0.8921985.$$

Nu vom respinge ipoteza 0 la orice nivel rezonabil de semnificație.

Valoarea  $p$  folosind statistica lui Pearson este 0.8954435 - aproape identică.

#### 1.4.6 Testul $\chi^2$ pentru omogenitate

Acesta este un test pentru a vedea dacă unele seturi independente de date provin din aceeași repartiție. (Înțelesul omogenității în acest caz este că toate repartițiile sunt aceleași.)

Utilizare: Testăm dacă  $m$  mulțimi independente de date discrete provin din aceeași repartiție.

Rezultate:  $\omega_1, \omega_2, \dots, \omega_n$  sunt rezultate posibile. Acestea sunt aceleași pentru fiecare set de date.

Date: Presupunem  $m$  seturi independente de date dând numere pentru fiecare rezultat posibil. Adică, pentru setul de date  $i$  avem un număr observat  $O_{ij}$  pentru fiecare rezultat posibil  $\omega_j$ .

Presupuneri: Niciuna.

$H_0$ : Fiecare set de date provine din aceeași repartiție. (Nu specificăm care este această repartiție.)

$H_A$ : Seturile de date nu provin din aceeași repartiție.

Statistica testului: Vezi exemplul de mai jos. Sunt  $mn$  celule conținând numerele pentru fiecare rezultat pentru fiecare set de date. Folosind Repartiția 0 putem estima numerele așteptate pentru fiecare din seturile de date. Statisticile  $X^2$  și  $G$  sunt calculate exact ca mai sus.

Numărul de grade de libertate  $df = (m - 1)(n - 1)$ . (Vezi exemplul de mai jos.)

Repartiția 0:  $\chi^2(df)$ . Valorile  $p$  sunt calculate ca în testul  $\chi^2$  pentru bunătatea potrivirii.

Codul R: Funcția R `chisq.test` poate fi folosită pentru a face calcule pentru un test  $\chi^2$  folosind  $X^2$ . Pentru  $G$ , trebuie să facem calculele "cu mâna" (ajutați de R) sau să găsim un pachet care are o funcție. (Probabil va fi numită `likelihood.test` sau `G.test`.)

**Exemplul 12.** Cineva pretinde că a găsit o piesă de teatru pierdută de mult a lui William Shakespeare. Vă cere să testați dacă piesa a fost sau nu scrisă de Shakespeare.

Mergeți la <http://www.opensourceshakespeare.org> și alegeti aleator 12 pagini din *Regele Lear* și numărați folosirea cuvintelor uzuale. Faceți același lucru pentru "piesă pierdută de mult". Obțineți următorul tabel.

Word	a	an	this	that
<i>King Lear</i>	150	30	30	90
<i>Long lost work</i>	90	20	10	80

Folosind aceste date, faceți un test de semnificație pentru pretenția că piesa pierdută de mult este de William Shakespeare. Folosiți un nivel de semnificație de 0.1.

**Răspuns.**  $H_0$ : Pentru cele 4 cuvinte numărate piesa pierdută de mult are aceleasi frecvențe relative ca numerele luate din *Regele Lear*.

Totalul cuvintelor numărate în ambele piese este 500, deci estimarea de verosimilitate maximă a frecvențelor relative presupunând  $H_0$  este numărul total pentru fiecare cuvânt împărțit la 500.

Word	a	an	this	that	Total count
<i>King Lear</i>	150	30	30	90	300
<i>Long lost work</i>	90	20	10	80	200
totals	240	50	40	170	500
rel. frequencies under $H_0$	240/500	50/500	40/500	170/500	500/500

Acum, numerele așteptate pentru fiecare piesă sub  $H_0$  sunt numerele totale pentru acea piesă înmulțite cu frecvențele relative din tabelul de mai sus. Următorul tabel dă numerele: (observate, așteptate) pentru fiecare piesă. Pe

ultima linie, în loc de ”249” se va citi ”240”.

Word	a	an	this	that	Totals
King Lear	(150, 144)	(30, 30)	(30, 24)	(90, 102)	(300, 300)
Long lost work	(90, 96)	(20, 20)	(10, 16)	(80, 68)	(200, 200)
Totals	(249, 240)	(50, 50)	(40, 40)	(170, 170)	(500, 500)

Statistica  $\chi^2$  este

$$\begin{aligned} X^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{6^2}{144} + \frac{0^2}{30} + \frac{6^2}{24} + \frac{12^2}{102} + \frac{6^2}{96} + \frac{0^2}{20} + \frac{6^2}{16} + \frac{12^2}{68} \\ &\approx 7.904412. \end{aligned}$$

Sunt 8 celule și toate numerele marginale sunt fixate deoarece a fost nevoie de ele pentru a determina numerele așteptate. Am putea să punem liber valorile din 3 celule din tabel, de exemplu cele 3 celule albastre, apoi restul celulelor sunt determinate pentru a face corecte totalele marginale. Astfel,  $df = 3$ . (Sau ne putem reaminti că  $df = (m - 1)(n - 1) = 3 \cdot 1 = 3$ , unde  $m$  este numărul de coloane și  $n$  este numărul de linii.)

Folosind R găsim  $1 - pchisq(7.904412, 3) = 0.04802909$ . Deoarece aceasta este mai mică decât nivelul nostru de semnificație de 0.1, respingem ipoteza 0 că frecvențele relative ale cuvintelor sunt aceleași în ambele piese.

Dacă facem încă plus presupunerea că toate piesele lui Shakespeare au frecvențe similare ale cuvintelor (ceea ce este ceva ce putem verifica) concluzionăm că piesa pierdută probabil nu este a lui Shakespeare.

## 2 Comparație între deducțiile frecvenționistă și Bayesiană

### 2.1 Scopul învățării

Să poată să explice diferența dintre valoare  $p$  și probabilitatea a posteriori unui doctor.

### 2.2 Introducere

Am aflat despre cele 2 școli de deducție statistică: Bayesiană și frecvenționistă. Ambele abordări permit evaluarea dovezilor despre ipoteze concurente.

## 2.3 Formula lui Bayes ca piatră de încercare

Formula lui Bayes este o propoziție abstractă despre probabilități condiționate ale evenimentelor:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Am început deducția Bayesiană reinterpretând evenimentele din formula lui Bayes:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}.$$

Acum  $\mathcal{H}$  este o ipoteză și  $\mathcal{D}$  sunt datele care pot da dovezi pentru sau împotriva lui  $\mathcal{H}$ . Fiecare termen din formula lui Bayes are un nume și un rol. A priori  $P(\mathcal{H})$  este probabilitatea ca  $\mathcal{H}$  să fie adevărată înainte de considerarea datelor.

A posteriori  $P(\mathcal{H}|\mathcal{D})$  este probabilitatea ca  $\mathcal{H}$  să fie adevărată după considerarea datelor.

Verosimilitatea  $P(\mathcal{D}|\mathcal{H})$  este dovada despre  $\mathcal{H}$  furnizată de datele  $\mathcal{D}$ .

$P(\mathcal{D})$  este probabilitatea totală a datelor luând în considerare toate ipotezele posibile.

Dacă a priori și verosimilitatea sunt cunoscute pentru toate ipotezele, atunci formula lui Bayes calculează a posteriori exact. Acesta a fost cazul când am aruncat un zar selectat aleator dintr-o urnă [al cărei conținut îl știam](#). Numim aceasta logica deductivă a teoriei probabilităților și ea dă un mod direct de a compara ipotezele, a trage concluzii și a lua decizii.

În cele mai multe experimente, probabilitățile a priori ale ipotezelor nu sunt cunoscute. În acest caz, recurgem la arta deducției statistice: ori inventăm o a priori (Bayesianii), ori facem tot ce putem folosind doar verosimilitatea (frecvenționiștii).

Școala Bayesiană modelează incertitudinea printr-o repartiție de probabilitate peste ipoteze. Abilitatea cuiva de a face deducții depinde de gradul de încredere în a priori aleasă și robustețea constatărilor pentru a alterna repartițiile a priori poate fi relevantă și importantă.

Școala frecvenționistă folosește doar repartiții condiționate ale datelor cunoscând ipoteze specifice. Presupunerea este că o ipoteză (parametru specificând repartiția condiționată a datelor) este adevărată și că datele observate provin din acea repartiție. În particular, abordarea frecvenționistă nu depinde de o a priori subiectivă care poate varia de la un investigator la altul.

Acste 2 școli pot fi comparate mai departe după cum urmează:

### Deduția Bayesiană

folosește probabilități atât pentru ipoteze cât și pentru date;  
depinde de a priori și verosimilitatea datelor observate;

cere să se știe sau să se construiască o ”a priori subiectivă”; a dominat practica statistică înainte de secolul 20; poate fi intens calculată din cauza integrării peste mulți parametri.

### **Deduția frecvenționistă (NHST)**

nu folosește niciodată probabilitatea unei ipoteze (nu a priori sau a posteriori);

deinde de verosimilitatea  $P(\mathcal{D}|\mathcal{H})$  atât pentru datele observate cât și pentru cele neobservate;

nu cere o a priori;

practica statistică dominantă de-a lungul secolului 20;

tinde să fie mai puțin intensă din punctul de vedere al calculelor.

Măsuri frecvenționiste ca valori  $p$  sau intervale de încredere continuă să domine cercetarea, în special în științele vieții. Totuși, în era actuală a computerelor puternice și datelor mari, metodele Bayesiene au avut o renaștere enormă în domenii ca învățarea automată și genetica. Acum sunt un număr de studii clinice mari, în curs de desfășurare folosind protocoale Bayesiene, ceea ce ar fi fost greu de imaginat cu o generație în urmă.

În timp ce împărțirile profesionale rămân, consensul care se formează printre statisticienii de top este că cea mai eficace abordare a problemelor complexe adesea folosește cele mai bune perspective din ambele școli lucrând împreună.

## **2.4 Critici și apărări**

### **2.4.1 Critica deducției Bayesiene**

1. Principala critică a deducției Bayesiene este că o a priori este subiectivă. Nu există o singură metodă de a alege a priori, deci persoane diferite pot produce a priori diferite și de aceea pot ajunge la a posteriori și concluzii diferite.
2. Mai departe, sunt obiecții filozofice la atribuirea de probabilități ipotezelor, deoarece ipotezele nu sunt rezultate ale experimentelor repetabile în care se poate măsura frecvența pe termen lung. Mai degrabă, o ipoteză este ori adevărată ori falsă, indiferent dacă se știe care este cazul. O monedă este ori corectă, ori incorectă; tratamentul 1 este ori mai bun, ori mai rău ca tratamentul 2; soarele va răsări sau nu va răsări mâine.

### **2.4.2 Apărarea deducției Bayesiene**

1. Probabilitatea ipotezelor este exact ce ne trebuie pentru a lua decizii. Când doctorul îmi spune că un test a ieșit pozitiv vreau să știu care este probabilitatea că aceasta înseamnă că sunt bolnav. Adică, vreau să știu probabilitatea ipotezei ”Sunt bolnav”.

2. Folosirea teoremei lui Bayes este riguroasă logic. Odată ce avem a priori toate calculele noastre au certitudinea logicii deductive.
3. Încercând diferite a priori, putem vedea cât de sensibile sunt rezultatele noastre de alegerea a priori.
4. Este ușor de comunicat un rezultat încadrat în termeni de probabilitățiile ipotezelor.
5. Chiar dacă a priori pot fi subiective, se pot specifica presupunerile folosite pentru a ajunge la ele, care permit altor oameni să le folosească sau să încerce alte a priori.
6. Dovezile derivate din date sunt independente de noțiuni despre "datele mai extreme", care depind de schema experimentală exactă.
7. Datele pot fi folosite pe măsură ce vin. Nu există o cerință ca fiecare contingent să fie planificat înainte.

#### **2.4.3 Critica deducției frecvenționiste**

1. Este ad-hoc și nu are forță logicii deductive. Noțiuni ca "date mai extreme" nu sunt bine definite. Valoarea  $p$  depinde de schema experimentală exactă.
2. Experimentele trebuie complet specificate înainte. Aceasta poate conduce la rezultate aparent paradoxale. Vezi "istoria voltmetrului" în:  
[http://en.wikipedia.org/wiki/Likelihood\\_principle](http://en.wikipedia.org/wiki/Likelihood_principle).
3. Valoarea  $p$  și nivelul de semnificație sunt notoriu predispuse la interpretare greșită. Un nivel de semnificație de 0.05 înseamnă că probabilitatea unei erori de tipul I este 5%. Adică, **dacă ipoteza 0 este adevărată**, atunci în 5% din timp va fi respinsă din cauza caracterului aleatoriu. Mulți oameni gândesc eronat că o valoare  $p$  de 0.05 înseamnă că probabilitatea ipotezei 0 este 5%. Ați putea argumenta că aceasta nu este o critică a deducției frecvenționiste ci, mai degrabă, o critică a ignoranței populare. Totuși, subtilitatea ideilor contribuie sigur la problemă.

#### **2.4.4 Apărarea deducției frecvenționiste**

1. Este obiectivă: toți statisticienii vor fi de acord cu valoarea  $p$ . Orice individ poate apoi decide dacă valoarea  $p$  garantează respingerea ipotezei 0.
2. Testarea ipotezei folosind testarea semnificației frecvenționistă este aplicată în analiza statistică a investigațiilor științifice, evaluând puterea dovezilor împotriva unei ipoteze 0 cu date. Interpretarea rezultatelor este lăsată utilizatorului testelor. Diferiți utilizatori pot aplica diferite nivele de semnificație pentru a determina semnificația statistică. Statistica frecvenționistă nu pretinde că dă un mod de a alege un nivel de semnificație; mai degrabă descrie explicit

compromisul între erorile de tipul I și de tipul II.

3. Proiectarea experimentului frecvenționistă cere o descriere atentă a experimentului și metodelor de analiză înaintea startului. Aceasta ajută controlul pentru părtinirea experimentatorului.
4. Abordarea frecvenționistă a fost folosită de peste 100 de ani și am văzut progres științific extraordinar. Cu toate că frecvenționistul însuși n-ar pune o probabilitate pe încrederea că metodele frecvenționiste sunt valoroase, n-ar trebui ca această istorie să dea Bayesianului o încredere a priori puternică în utilitatea metodelor frecvenționiste?

## 2.5 Fiți atenți la $p$ -urile voastre

Facem un test  $t$  cu 2 selecții pentru medii egale, cu  $\alpha = 0.05$  și obținem o valoare  $p$  de 0.04. Care sunt șansele ca cele 2 selecții să provină din repartiții cu aceeași medie?

- (a) 19/1 (b) 1/19 (c) 1/20 (d) 1/24 (e) necunoscute.

**Răspuns.** (e) necunoscute. Metodele frecvenționiste dau doar probabilități ale statisticilor condiționate de ipoteze. Ele nu dau probabilități ale ipotezelor.

## 2.6 Reguli de oprire

Când derulăm o serie de încercări, avem nevoie de o regulă despre când să ne oprim. 2 reguli uzuale sunt:

1. Facem exact  $n$  încercări și ne oprim.
2. Facem încercări până vedem un rezultat și apoi ne oprim.

În acest exemplu vom considera 2 experimente de aruncarea monedei.

Experimentul 1: Aruncăm moneda exact de 6 ori și raportăm numărul aversurilor.

Experimentul 2: Aruncăm moneda până apare primul revers și raportăm numărul aversurilor.

Jon este îngrijorat că moneda lui este deplasată spre aversuri, deci, înainte să-o utilizeze la cursuri, el îi testează corectitudinea. El face un experiment și raportează lui Jerry că secvența lui de aruncări a fost  $HHHHHT$ . Dar Jerry nu este atent și uită ce experiment a făcut Jon să producă datele.

### Abordarea frecvenționistă

Deoarece a uitat ce experiment a făcut Jon, Jerry frecvenționistul decide să calculeze valorile  $p$  pentru ambele experimente cunoscând datele lui Jon.

Fie  $\theta$  probabilitatea aversului. Avem ipotezele 0 și alternativă unilaterală

$$H_0 : \theta = 0.5, \quad H_A : \theta > 0.5.$$

Experimentul 1: Repartiția 0 este binomială(6, 0.5) deci valoarea  $p$  unilaterală este probabilitatea a 5 sau 6 aversuri în 6 aruncări. Folosind R obținem

$$p = 1 - \text{pbinom}(4, 6, 0.5) = 0.109375.$$

Experimentul 2: Repartiția 0 este geometrică(0.5) deci, valoarea  $p$  unilaterală este probabilitatea a 5 sau mai multe aversuri înaintea primului revers. Folosind R obținem

$$p = 1 - \text{pgeom}(4, 0.5) = 0.03125.$$

Folosind nivelul de semnificație tipic de 0.05, aceleași date duc la concluzii diferite! Am respinge  $H_0$  în experimentul 2, dar nu în experimentul 1.

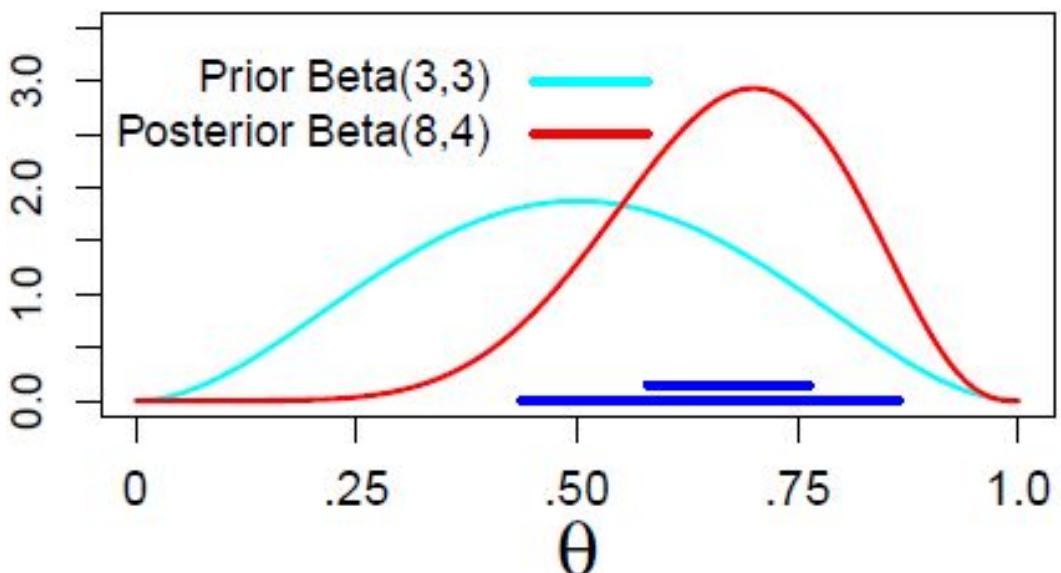
Frecvenționistul nu are nicio problemă cu aceasta. Multimea rezultatelor este diferită pentru experimentele diferite, deci noțiunea de date extreme și de aceea, de valoare  $p$ , este diferită. De exemplu, în experimentul 1 am considera  $THHHHH$  a fi la fel de extrem ca  $HHHHHT$ . În experimentul 2 n-am vedea niciodată  $THHHHH$  deoarece experimentul s-ar termina după primul revers.

### **Abordarea Bayesiană.**

Jerry Bayesianul știe că nu contează care dintre cele 2 experimente a fost făcut de Jon, deoarece funcțiile (coloanele) de verosimilitate binomială și geometrică pentru datele  $HHHHHT$  sunt proporționale. În orice caz, el trebuie să inventeze o a priori și alege Beta(3, 3). Aceasta este o a priori relativ plată concentrată peste intervalul  $0.25 \leq \theta \leq 0.75$ .

Vezi <http://mathlets.org/mathlets/beta-distribution>.

Deoarece repartițiile beta și binomială (sau geometrică) formează o pereche conjugată, actualizarea Bayesiană este simplă. Datele de 5 aversuri și 1 revers dau o repartition a posteriori Beta(8,4). Iată graficul a priori și a posteriori. Liniile albastre de jos sunt intervale de probabilitate 50% și 90% pentru a posteriori.



Repartițiile a priori și a posteriori cu intervale de probabilitate 0.5 și 0.9

Iată calculele relevante din R:

Interval de probabilitate 50% a posteriori:

`qbeta(c(0.25,0.75),8,4) = [0.579529, 0.7635974].`

Interval de probabilitate 90% a posteriori:

`qbeta(c(0.05,0.95),8,4) = [0.4356258, 0.8649245].`

$P(\theta > 0.5 | \text{date}) = 1 - pbeta(0.5, 8, 4) = 0.8867188.$

Plecând de la a priori Beta(3,3), probabilitatea a posteriori ca moneda să fie deplasată spre avers este 0.8867188.

## 2.7 Luarea deciziilor

Destul de des scopul deducției statistice este de a ajuta la luarea deciziilor, de exemplu dacă sau nu să fie supus unei intervenții chirurgicale, cât de mult să investească într-o acțiune, dacă sau nu să mergă să absolvească școala, etc.

În teoria deciziilor statistice, consecințele acțiunilor sunt măsurate printr-o funcție de utilitate. Funcția de utilitate atribuie o pondere fiecărui rezultat posibil; în limbajul probabilităților este pur și simplu o variabilă aleatoare. De exemplu, în investițiile mele aş putea atribui o utilitate de  $d$  rezultatului unui câștig de  $d$  lei pe acțiune (dacă  $d < 0$ , utilitatea mea este negativă). Pe de altă parte, dacă toleranța mea pentru risc este mică, voi atribui mai multă utilitate negativă pentru pierderi decât pentru câștiguri (să zicem,  $-d^2 - 1$  dacă  $d < 0$  și  $d$  dacă  $d \geq 0$ ).

O regulă de decizie combină utilitatea medie cu dovezile pentru fiecare ipoteză cunoscând datele (de exemplu, valori  $p$  sau repartiții a posteriori) într-un cadru statistic formal pentru luarea deciziilor.

În acest cadru, frecvenționistul va considera media utilității dată fiind o ipoteză

$$E(U|\mathcal{H}),$$

unde  $U$  este variabila aleatoare reprezentând utilitatea. Există metode frecvenționiste pentru combinarea utilității medii cu valori  $p$  ale ipotezelor pentru a ghida deciziile.

Bayesianul poate combina  $E(U|\mathcal{H})$  cu a posteriori (sau a priori dacă este înaintea colectării datelor) pentru a crea o regulă de decizie Bayesiană.

În orice cadru, 2 persoane considerând aceeași investiție pot avea funcții de utilitate diferite și lua decizii diferite. De exemplu, o acțiune riscantă (cu un potențial mai mare de a fluctua) va fi mai atrăgătoare în raport cu prima funcție de utilitate de mai sus decât în raport cu a 2-a.

Un rezultat teoretic semnificativ este că pentru orice regulă de decizie există o regulă de decizie Bayesiană care este, într-un sens precis o regulă cel puțin la fel de bună.

# Curs 18

Cristian Niculescu

## 1 Intervale de încredere bazate pe date normale

### 1.1 Scopurile învățării

1. Să poată să determine dacă o expresie definește o statistică de interval validă.
2. Să poată calcula intervale de încredere  $z$  și  $t$  pentru media datelor normale.
3. Să poată calcula intervalul de încredere  $\chi^2$  pentru dispersia datelor normale.
4. Să poată defini nivelul de încredere al unui interval de încredere.
5. Să poată explica relația dintre intervalul de încredere  $z$  (și nivelul de încredere) și regiunea de nerespingere  $z$  (și nivelul de încredere) în NHST.

### 1.2 Introducere

Intervalele de încredere sunt unelte ale statisticii frecvenționiste. Presupunem că avem un model (repartiție de probabilitate) pentru date observate cu un parametru necunoscut. Am văzut cum NHST folosește datele pentru a testa ipoteza că parametrul necunoscut are o valoare particulară.

Am văzut de asemenea cum estimări punctuale ca MLE folosesc datele pentru a da o estimare a parametrului necunoscut. Pe cont propriu, o estimare punctuală ca  $\bar{x} = 2.2$  nu poartă informație despre acuratețea ei; este doar un singur număr, indiferent dacă este bazat pe 10 date sau 1000000 de date.

Din acest motiv, statisticienii măresc estimările punctuale la intervale de încredere. De exemplu, pentru a estima o medie necunoscută  $\mu$  am putea spune că cea mai bună estimare a mediei este  $\bar{x} = 2.2$  cu un interval de 95% încredere  $[1.2, 3.2]$ . Alt mod de a descrie intervalul este  $\bar{x} \pm 1$ .

Luate împreună, lungimea intervalului și nivelul de încredere dau o măsură a puterii dovezilor care sprijină ipoteza că  $\mu$  este aproape de estimarea noastră

$\bar{x}$ . Nivelul de încredere al unui interval este analog nivelului de semnificație al unui NHST. Nu este un accident că adesea vedem nivelul de semnificație  $\alpha = 0.05$  și nivelul de încredere  $0.95 = 1 - \alpha$ .

Intervalele de încredere  $z$  și  $t$  pentru medie și  $\chi^2$  pentru dispersie sunt ușor de calculat. Provocarea cu intervalele de încredere nu este calculul lor, ci mai degrabă interpretarea lor corectă și a ști cum să le folosim în practică.

### 1.3 Statistici interval

Reamintim definiția noastră de lucru: o statistică este orice se poate calcula din date. În particular, formula pentru o statistică nu poate include cantități necunoscute.

**Exemplul 1.** Presupunem că  $x_1, \dots, x_n$  provin din  $N(\mu, \sigma^2)$ , unde  $\mu$  și  $\sigma$  sunt necunoscute.

- (i)  $\bar{x}$  și  $\bar{x} - 5$  sunt statistici.
- (ii)  $\bar{x} - \mu$  nu este o statistică deoarece  $\mu$  este necunoscută.
- (iii) Dacă  $\mu_0$  este o valoare cunoscută, atunci  $\bar{x} - \mu_0$  este o statistică. Acest caz apare când considerăm ipoteza  $H_0: \mu = \mu_0$ . De exemplu, dacă ipoteza 0 este  $\mu = 5$ , atunci statistică  $\bar{x} - \mu_0$  este chiar  $\bar{x} - 5$  de la (i).

Putem juca același joc cu intervale pentru a defini [statistici interval](#).

**Exemplul 2.** Presupunem că  $x_1, \dots, x_n$  provin din  $N(\mu, \sigma^2)$  unde  $\mu$  este necunoscută.

- (i) Intervalul  $[\bar{x} - 2.2, \bar{x} + 2.2] = \bar{x} \pm 2.2$  este o statistică interval.
- (ii) Dacă  $\sigma$  este [cunoscută](#), atunci  $[\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}]$  este o statistică interval.
- (iii) Pe de altă parte, dacă  $\sigma$  este [necunoscută](#), atunci  $[\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}]$  **nu** este o statistică interval.
- (iv) Dacă  $s^2$  este dispersia de selecție, atunci  $[\bar{x} - \frac{2s}{\sqrt{n}}, \bar{x} + \frac{2s}{\sqrt{n}}]$  este o statistică interval deoarece  $s^2$  este calculată din date.

Tehnic, o statistică interval nu este mai mult decât o pereche de statistici punctuale care dau marginile inferioară și superioară ale intervalului.

#### Observații.

1. Intervalul este aleator - noi date aleatoare vor produce un nou interval.
2. Intervalul nu depinde de valoarea unui parametru necunoscut.
3. Pdf pentru unele repartiții este funcție pară, pentru altele nu.

**Exemplul 3.** Pdf-urile pentru  $N(0, 1)$  și pentru  $t$  sunt funcții pare, i.e. graficul este simetric față de axa  $Oy$ , dar pentru  $\chi^2$  nu.

## 1.4 Intervale de încredere $z$ pentru medie

Presupunem că avem date normal repartizate:

$$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$$

### 1.4.1 Definiția intervalelor de încredere $z$ pentru medie

**Definiție.** Presupunem că datele  $x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ , cu medie necunoscută  $\mu$  și dispersie cunoscută  $\sigma^2$ . Intervalul de încredere  $(1 - \alpha)$  pentru  $\mu$  este

$$\left[ \bar{x} - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \right], \quad (1)$$

unde  $z_{\alpha/2}$  este **valoarea critică din dreapta**, pentru care  $P(Z > z_{\alpha/2}) = \alpha/2$ . De exemplu, dacă  $\alpha = 0.05$ , atunci  $z_{\alpha/2} = qnorm(1 - 0.05/2) \approx 1.96$ , deci intervalul de încredere 0.95 (sau 95%) este

$$\left[ \bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right].$$

Următoarea aplicație generază date normale și arată intervalele de încredere  $z$  sau  $t$  pentru medie.

<http://mathlets.org/mathlets/confidence-intervals/>.

**Exemplul 4.** Presupunem că colectăm 100 de date dintr-o repartiție  $N(\mu, 3^2)$  și media de selecție este  $\bar{x} = 12$ . Dați intervalul de încredere 95% pentru  $\mu$ .

**Răspuns.** Folosind formula, intervalul de încredere 95% pentru  $\mu$  este

$$\left[ \bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right] = \left[ 12 - \frac{1.96 \cdot 3}{\sqrt{100}}, 12 + \frac{1.96 \cdot 3}{\sqrt{100}} \right] = [11.412, 12.588].$$

### 1.4.2 Explicarea definiției partea 1: regiuni de respingere

Explicăm definiția (1) plecând de la cunoștințele noastre despre regiuni de respingere/nerespingere. Exprimarea **"regiune de nerespingere"** nu este frumoasă, dar o vom folosi în locul exprimării incorecte **"regiune de acceptare"**.

**Exemplul 5.** Presupunem că  $n = 12$  date provin din  $N(\mu, 5^2)$ , unde  $\mu$  este necunoscută. Faceți un test de semnificație bilateral pentru  $H_0: \mu = 2.71$  folosind statistică  $\bar{x}$  la nivelul de semnificație  $\alpha = 0.05$ . Descrieți regiunile de respingere și nerespingere.

**Răspuns.** Sub ipoteza 0 avem  $\mu = 2.71$ ,  $x_i \sim N(2.71, 5^2)$  și deci

$$\bar{x} \sim N(2.71, 5^2/12),$$

unde  $5^2/12$  este dispersia  $\sigma_{\bar{x}}^2$  a lui  $\bar{x}$ . Știm că semnificația  $\alpha = 0.05$  corespunde la o regiune de respingere în afara a 1.96 deviații standard ale mediei ipotetice. Adică, regiunile de nerespingere și de respingere sunt separate de valorile critice  $2.71 \pm 1.96\sigma_{\bar{x}}$ .

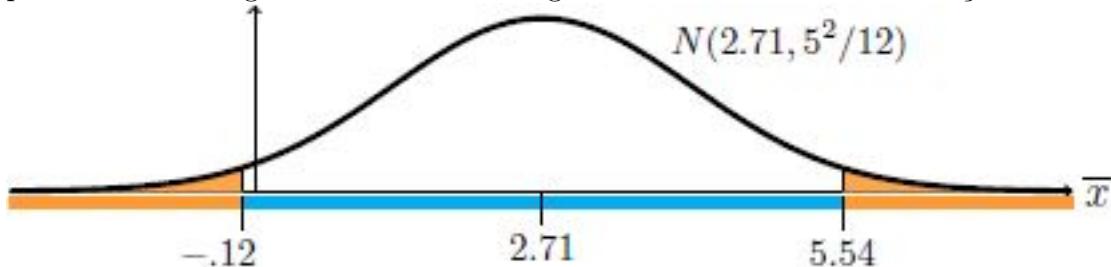
Regiunea de nerespingere:

$$\left[ 2.71 - \frac{1.96 \cdot 5}{\sqrt{12}}, 2.71 + \frac{1.96 \cdot 5}{\sqrt{12}} \right] \approx [-0.12, 5.54].$$

Regiunea de respingere:

$$\left( -\infty, 2.71 - \frac{1.96 \cdot 5}{\sqrt{12}} \right) \cup \left( 2.71 + \frac{1.96 \cdot 5}{\sqrt{12}}, \infty \right) \approx (-\infty, -0.12) \cup (5.54, \infty).$$

Următoarea figură arată regiunile de respingere și nerespingere pentru  $\bar{x}$ . Regiunile reprezintă domenii pentru  $\bar{x}$ , deci sunt reprezentate de bare colorate pe axa  $\bar{x}$ . Aria regiunii colorate de sub grafic este nivelul de semnificație.



Regiunile de respingere (portocalie) și de nerespingere (albastră) pentru  $\bar{x}$ .

**Exemplul 6.** Presupunem că  $n$  date provin din  $N(\mu, \sigma^2)$ , unde  $\mu$  este necunoscută și  $\sigma$  este cunoscută. Faceți un test de semnificație bilateral al lui  $H_0: \mu = \mu_0$  folosind statistică  $\bar{x}$  la nivelul de semnificație  $\alpha = 0.05$ . Descrieți regiunile de respingere și nerespingere.

**Răspuns.** Sub ipoteza 0  $\mu = \mu_0$  avem  $x_i \sim N(\mu_0, \sigma^2)$  și deci

$$\bar{x} \sim N(\mu_0, \sigma^2/n),$$

unde  $\sigma^2/n$  este dispersia  $\sigma_{\bar{x}}^2$  a lui  $\bar{x}$  și  $\mu_0, \sigma$  și  $n$  sunt toate valori cunoscute. Fie  $z_{\alpha/2}$  valoarea critică:  $P(Z > z_{\alpha/2}) = \alpha/2$ . Regiunile de nerespingere și de respingere sunt separate de valorile lui  $\bar{x}$  care sunt la distanță de  $z_{\alpha/2} \cdot \sigma_{\bar{x}}$  de media ipotecă. Deoarece  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ , avem:

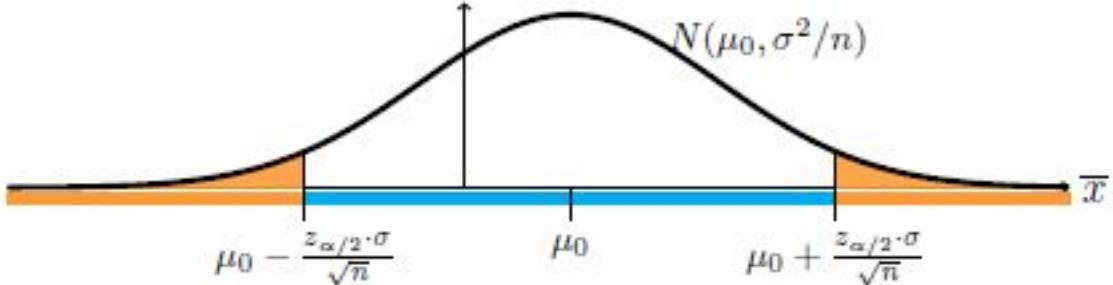
Regiunea de nerespingere:

$$\left[ \mu_0 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \mu_0 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \right]. \quad (2)$$

Regiunea de respingere:

$$\left( -\infty, \mu_0 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \right) \cup \left( \mu_0 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \infty \right).$$

Obținem aceeași figură ca mai sus, cu numerele explicite înlocuite de valori simbolice.



Regiunile de respingere (portocalie) și de nerespingere (albastră) pentru  $\bar{x}$ .

#### 1.4.3 Manipularea intervalelor: pivotarea

**Pivotarea** este ideea că  $\bar{x} \in [\mu_0 - a, \mu_0 + a] \iff \mu_0 \in [\bar{x} - a, \bar{x} + a], \forall a > 0$ .

**Exemplul 7.** Presupunem că avem media de selecție  $\bar{x}$  și media ipotetică  $\mu_0 = 2.71$ . Presupunem de asemenea că repartiția 0 este  $N(\mu_0, 3^2)$ . Atunci, cu un nivel de semnificație de 0.05, avem:

$$\mu_0 + 1.96\sigma = 2.71 + 1.96 \cdot 3 = 2.71 + 5.88 \text{ este valoarea critică } 0.025;$$

$$\mu_0 - 1.96\sigma = 2.71 - 1.96 \cdot 3 = 2.71 - 5.88 \text{ este valoarea critică } 0.975.$$

Regiunea de nerespingere este centrată în  $\mu_0 = 2.71$ . Adică, nu respingem  $H_0$  dacă  $\bar{x}$  este în intervalul

$$[\mu_0 - 1.96\sigma, \mu_0 + 1.96\sigma] = [2.71 - 5.88, 2.71 + 5.88].$$

Intervalul de încredere este centrat în  $\bar{x}$ . Intervalul de încredere 0.95 are aceeași lungime cu regiunea de nerespingere. Este intervalul

$$[\bar{x} - 1.96\sigma, \bar{x} + 1.96\sigma] = [\bar{x} - 5.88, \bar{x} + 5.88].$$

Aici este o simetrie:

$$\bar{x} \in [2.71 - 1.96\sigma, 2.71 + 1.96\sigma] \iff 2.71 \in [\bar{x} - 1.96\sigma, \bar{x} + 1.96\sigma].$$

Această simetrie este numită pivotare. Iată câteva exemple numerice de pivotare:

**Exemplul 8. (i)**  $1.5 \in [0 - 2.3, 0 + 2.3] \implies 0 \in [1.5 - 2.3, 1.5 + 2.3]$ .

(ii) Analog  $1.5 \notin [0 - 1, 0 + 1] \implies 0 \notin [1.5 - 1, 1.5 + 1]$ .

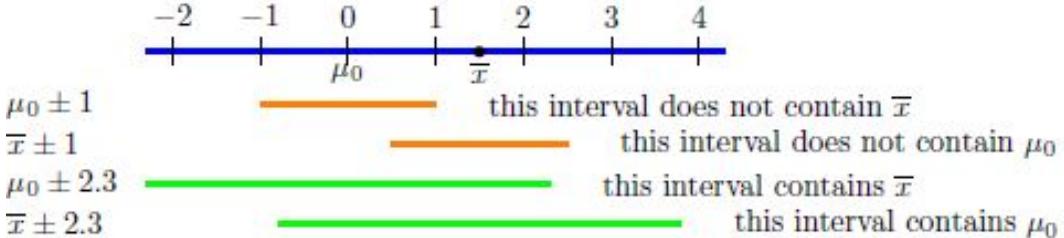
Simetria poate fi mai clară dacă vorbim în termeni de distanțe: afirmația

” $1.5 \in [0 - 2.3, 0 + 2.3]$ ”

spune că distanța de la 1.5 la 0 este cel mult 2.3. Analog, afirmația

” $0 \in [1.5 - 2.3, 1.5 + 2.3]$ ”

spune exact același lucru, i.e. distanța de la 0 la 1.5 este cel mult 2.3.  
Iată o vizualizare a *pivotării* din intervalele din jurul lui  $\mu_0$  în intervalele din jurul lui  $\bar{x}$ .



Distanța dintre  $\bar{x}$  și  $\mu_0$  este 1.5. Deoarece  $1 < 1.5$ ,  $[\mu_0 - 1, \mu_0 + 1]$  nu se întinde destul de departe pentru a conține  $\bar{x}$ . Analog, intervalul  $[\bar{x} - 1, \bar{x} + 1]$  nu se întinde destul de mult pentru a conține  $\mu_0$ . Deoarece  $2.3 > 1.5$ ,  $\bar{x} \in [\mu_0 - 2.3, \mu_0 + 2.3]$  și  $\mu_0 \in [\bar{x} - 2.3, \bar{x} + 2.3]$ .

#### 1.4.4 Explicarea definiției, partea a 2-a: translatarea regiunii de nerespingere într-un interval de încredere

În exemplele precedente avem o ipoteză 0. Dar dacă [nu avem o ipoteză 0?](#) În acest caz, avem estimarea punctuală  $\bar{x}$ , dar încă vrem să folosim datele pentru a estima un interval pentru media necunoscută. Adică, vrem o statistică interval. Aceasta este dată de un interval de încredere.

Folosind numerele din exemplul 5, spunem că la nivelul de semnificație de 0.05, nu respingem  $H_0$  dacă

$$\bar{x} \in \left[ 2.71 - \frac{1.96 \cdot 5}{\sqrt{12}}, 2.71 + \frac{1.96 \cdot 5}{\sqrt{12}} \right]. \quad (3)$$

Roulurile lui  $\bar{x}$  și 2.71 sunt simetrice. Nu respingem dacă

$$2.71 \text{ este în intervalul } \bar{x} \pm \frac{1.96 \cdot 5}{\sqrt{12}}. \quad (4)$$

Am ajuns la scopul nostru al unei statistici interval care estimează media necunoscută. Putem rescrive (4) astfel: la nivelul de semnificație de 0.05 nu respingem dacă

$$2.71 \in \left[ \bar{x} - \frac{1.96 \cdot 5}{\sqrt{12}}, \bar{x} + \frac{1.96 \cdot 5}{\sqrt{12}} \right]. \quad (5)$$

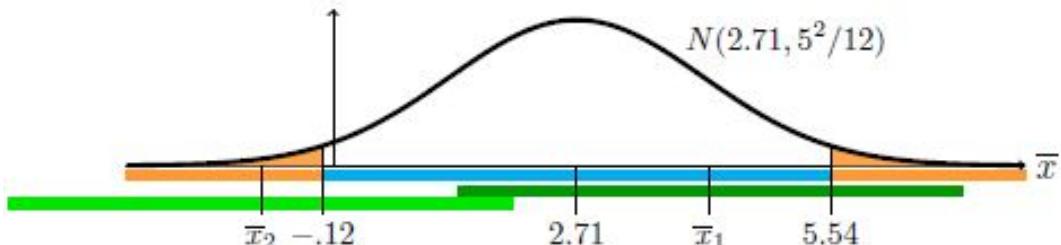
Astfel, diferite valori ale lui  $\bar{x}$  generează diferite intervale.

Intervalul din (5) este exact [intervalul de încredere](#) definit în (1).

Facem observații despre intervalul de încredere:

1. Depinde doar de  $\bar{x}$ , deci este o statistică.

2. Nivelul de semnificație  $\alpha = 0.05$  înseamnă că, **presupunând că ipoteza 0 că  $\mu = 2.71$  este adevărată**, datele aleatoare ne vor duce la respingerea ipotezei 0 în 5% din timp (o eroare de tipul I).
3. Din nou, **presupunând că  $\mu = 2.71$** , atunci 5% din timp intervalul de încredere nu va conține 2.71 și reciproc, 95% din timp va conține 2.71.
- Următoarea figură ilustrează cum nu respingem  $H_0$  dacă intervalul de încredere centrat în  $\bar{x}$  conține  $\mu_0$  și respingem  $H_0$  dacă intervalul de încredere nu conține  $\mu_0$ .
1. Am plecat cu figura de la exemplul 5, care arată repartitia 0 pentru  $\mu_0 = 2.71$  și regiunile de respingere și nerespingere.
  2. Am adăugat 2 valori posibile ale statisticii  $\bar{x}$ ,  $\bar{x}_1$  și  $\bar{x}_2$  și intervalele lor de încredere. Observați că lungimea fiecărui interval este exact aceeași ca lungimea regiunii de nerespingere deoarece ambele folosesc  $\pm \frac{1.96 \cdot 5}{\sqrt{12}}$ . Prima valoare,  $\bar{x}_1$  este în regiunea de nerespingere și intervalul ei de încredere conține  $\mu_0 = 2.71$ . Aceasta ilustrează că **nerespingerea lui  $H_0$  corespunde faptului că intervalul de încredere conține pe  $\mu_0$** .
  - A 2-a valoare,  $\bar{x}_2$ , este în regiunea de respingere și intervalul ei de încredere nu conține pe  $\mu_0$ . Aceasta ilustrează că **respingerea lui  $H_0$  corespunde faptului că intervalul de încredere nu conține pe  $\mu_0$** .



Regiunea de nerespingere (albastru) și cele 2 intervale de încredere (verde).

Alegerea noastră pentru ipoteza 0  $\mu = 2.71$  a fost complet arbitrară. Dacă înlocuim  $\mu = 2.71$  cu altă ipoteză  $\mu = \mu_0$ , atunci intervalul (5) va fi același. Numim intervalul (5) un **interval de încredere 95%** deoarece, **presupunând  $\mu = \mu_0$** , în medie, va conține  $\mu_0$  în 95% din experimentele aleatoare.

#### 1.4.5 Explicarea definiției partea a 3-a: translatarea unei regiuni de nerespingere generale într-un interval de încredere

Valorile specifice ale lui  $\sigma$  și  $n$  din exemplul precedent pot fi înlocuite prin litere. În acest fel putem trece exemplul 6 prin aceiași pași ca exemplul 5. În cuvinte, relația (2) și figura corespunzătoare spune că nu respingem dacă

$$\bar{x} \in \left[ \mu_0 - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \mu_0 + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \right].$$

Aceasta este echivalent cu a spune că nu respingem dacă

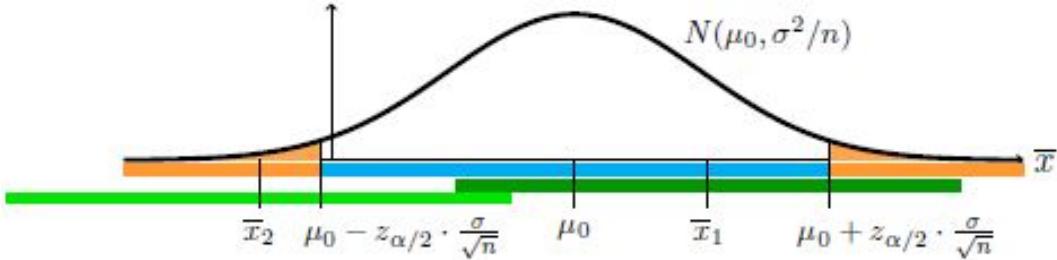
$$\mu_0 \text{ este în intervalul } \bar{x} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}. \quad (6)$$

Putem rescrie relația (6) astfel: la nivelul de semnificație  $\alpha$  nu respingem dacă

$$\mu_0 \in \left[ \bar{x} - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \right]. \quad (7)$$

Numim intervalul din (7) un **interval de încredere**  $(1 - \alpha)$  deoarece, presupunând  $\mu = \mu_0$ , în medie va conține  $\mu_0$  în fracția de  $(1 - \alpha)$  din experimentele aleatoare.

Figura următoare ilustrează că  $\mu_0$  este în intervalul de încredere  $(1 - \alpha)$  centrat în  $\bar{x} \iff \bar{x}$  este în regiunea de nerespingere (la nivelul de semnificație  $\alpha$ ) pentru  $H_0: \mu = \mu_0$ .



$\bar{x}_1$  este în regiunea de nerespingere pentru  $H_0: \mu = \mu_0 \iff$  intervalul de încredere centrat în  $\bar{x}_1$  conține pe  $\mu_0$ .

#### 1.4.6 Exemplu de calcul

**Exemplul 9.** Presupunem că datele 2.5, 5.5, 8.5, 11.5 provin dintr-o repartiție  $N(\mu, 10^2)$  cu medie necunoscută  $\mu$ .

- Calculați estimarea punctuală  $\bar{x}$  pentru  $\mu$  și intervalele de încredere 95%, 80% și 50% corespunzătoare.
- Considerăm ipoteza  $H_0: \mu = 1$ . Ati respinge  $H_0$  la  $\alpha = 0.05?$   $\alpha = 0.2?$   $\alpha = 0.5?$  Faceți aceasta în 2 moduri: întâi verificând dacă valoarea ipotetică a lui  $\mu$  este în intervalul de încredere relevant și apoi construind o regiune de respingere.

**Răspuns.** a)  $\bar{x} = \frac{2.5+5.5+8.5+11.5}{4} = 7$ . Punctele critice sunt

$$z_{0.025} = \text{qnorm}(0.975) \approx 1.96, z_{0.1} = \text{qnorm}(0.9) \approx 1.28, z_{0.25} = \text{qnorm}(0.75) \approx 0.67.$$

Deoarece  $n = 4$ , avem  $\bar{x} \sim N(\mu, 10^2/4)$ , i.e.  $\sigma_{\bar{x}} = 5$ . Deci avem:

$$\text{interval de încredere } 95\% = [\bar{x} - z_{0.025}\sigma_{\bar{x}}, \bar{x} + z_{0.025}\sigma_{\bar{x}}] \approx [7 - 1.96 \cdot 5, 7 + 1.96 \cdot 5] = [-2.8, 16.8].$$

$$\text{interval de încredere } 80\% = [\bar{x} - z_{0.1}\sigma_{\bar{x}}, \bar{x} + z_{0.1}\sigma_{\bar{x}}] \approx [7 - 1.28 \cdot 5, 7 + 1.28 \cdot 5] = [0.6, 13.4].$$

$$\text{interval de încredere } 50\% = [\bar{x} - z_{0.25}\sigma_{\bar{x}}, \bar{x} + z_{0.25}\sigma_{\bar{x}}] \approx [7 - 0.67 \cdot 5, 7 + 0.67 \cdot 5] = [3.65, 10.35].$$

Fiecare dintre aceste intervale este o estimare interval pentru  $\mu$ . Dacă nivelul de încredere este mai mare, atunci intervalul corespunzător este mai lung.

b) Deoarece  $\mu = 1$  este în intervalele de încredere 95% și 80%, nu respingem ipoteza 0 la nivelele  $\alpha = 0.05$  sau  $\alpha = 0.2$ . Deoarece  $\mu = 1$  nu este în intervalul de încredere 50%, respingem  $H_0$  la nivelul  $\alpha = 0.5$ .

Construim regiunile de respingere folosind aceleași valori critice ca la a). Diferența este că regiunile de respingere sunt complementare de intervale centrate în valoarea ipotetică pentru  $\mu$ :  $\mu_0 = 1$  și intervalele de încredere sunt centrate în  $\bar{x}$ . Iată regiunile de respingere:

$$\alpha = 0.05 \implies (-\infty, \mu_0 - z_{0.025}\sigma_{\bar{x}}) \cup (\mu_0 + z_{0.025}\sigma_{\bar{x}}, \infty) \approx (-\infty, -8.8) \cup (10.8, \infty).$$

$$\alpha = 0.2 \implies (-\infty, \mu_0 - z_{0.1}\sigma_{\bar{x}}) \cup (\mu_0 + z_{0.1}\sigma_{\bar{x}}, \infty) \approx (-\infty, -5.4) \cup (7.4, \infty).$$

$$\alpha = 0.5 \implies (-\infty, \mu_0 - z_{0.25}\sigma_{\bar{x}}) \cup (\mu_0 + z_{0.25}\sigma_{\bar{x}}, \infty) \approx (-\infty, -2.35) \cup (4.35, \infty).$$

Pentru a face NHST trebuie să verificăm dacă  $\bar{x} = 7$  se află sau nu în regiunea de respingere.

$$\alpha = 0.05 : 7 \notin (-\infty, -8.8) \cup (10.8, \infty).$$

Nu respingem ipoteza că  $\mu = 1$  la nivelul de semnificație de 0.05.

$$\alpha = 0.2 : 7 \notin (-\infty, -5.4) \cup (7.4, \infty).$$

Nu respingem ipoteza că  $\mu = 1$  la nivelul de semnificație de 0.2.

$$\alpha = 0.5 : 7 \in (-\infty, -2.35) \cup (4.35, \infty).$$

Respingem ipoteza că  $\mu = 1$  la nivelul de semnificație de 0.5.

Obținem aceleași răspunsuri cu orice metodă.

## 1.5 Intervale de încredere $t$ pentru medie

În acest cadru  $\sigma$  este necunoscută, deci trebuie să facem următoarele înlocuiri:

1. Folosim  $s_{\bar{x}} = \frac{s}{\sqrt{n}}$  în locul lui  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ . Aici  $s$  este dispersia de selecție pe care am folosit-o mai înainte la teste  $t$ .

2. Folosim valori critice  $t$  în locul valorilor critice  $z$ .

### 1.5.1 Definiția intervalelor de încredere $t$ pentru medie

**Definiție.** Presupunem că  $x_1, \dots, x_n \sim N(\mu, \sigma^2)$ , unde valorile mediei  $\mu$  și deviației standard  $\sigma$  sunt ambele necunoscute. Intervalul de încredere  $(1 - \alpha)$  pentru  $\mu$  este

$$\left[ \bar{x} - \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}, \bar{x} + \frac{t_{\alpha/2} \cdot s}{\sqrt{n}} \right], \quad (8)$$

unde  $t_{\alpha/2}$  este valoarea critică din dreapta a. î.  $P(T > t_{\alpha/2}) = \alpha/2$  pentru  $T \sim t(n - 1)$  și  $s^2$  este dispersia de selecție a datelor.

### 1.5.2 Construirea intervalelor de încredere $t$

Presupunem că  $n$  date provin din  $N(\mu, \sigma^2)$ , unde  $\mu$  și  $\sigma^2$  sunt necunoscute. Sub ipoteza 0  $\mu = \mu_0$ , avem  $x_i \sim N(\mu_0, \sigma^2)$ . Deci media Studentizată are o repartiție Student  $t$  cu  $n - 1$  grade de libertate:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n - 1).$$

Fie  $t_{\alpha/2}$  valoarea critică a. î.  $P(T > t_{\alpha/2}) = \alpha/2$  pentru  $T \sim t(n - 1)$ . Știm din testul  $t$  de o selecție că regiunea de nerespingere este dată de

$$|t| \leq t_{\alpha/2}.$$

Folosind definiția statisticii  $t$  pentru a scrie regiunea de respingere în termeni de  $\bar{x}$ , obținem: la nivelul de semnificație  $\alpha$  nu respingem dacă

$$\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \leq t_{\alpha/2} \iff |\bar{x} - \mu_0| \leq t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

Geometric, partea dreaptă spune că nu respingem dacă

$$\mu_0 \text{ este la cel mult } t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \text{ de } \bar{x}.$$

Aceasta este echivalent cu a spune că nu respingem dacă

$$\mu_0 \in \left[ \bar{x} - \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}, \bar{x} + \frac{t_{\alpha/2} \cdot s}{\sqrt{n}} \right].$$

Acest interval este intervalul de încredere definit în (8).

**Exemplul 10.** Presupunem că datele 2.5, 5.5, 8.5, 11.5 provin dintr-o repartiție  $N(\mu, \sigma^2)$  cu  $\mu$  și  $\sigma^2$  ambele necunoscute.

Dați estimările interval pentru  $\mu$  aflând intervalele de încredere 95%, 80% și 50%.

**Răspuns.** În R:

`x=c(2.5,5.5,8.5,11.5)`

$\bar{x} = \text{mean}(x) = 7$

$s^2 = \text{var}(x) = 15$ .

Punctele critice sunt

$t_{0.025} = \text{qt}(0.975, 3) = 3.182446$ ,

$$t_{0.1} = \text{qt}(0.9, 3) = 1.637744,$$

$$t_{0.25} = \text{qt}(0.75, 3) = 0.7648923.$$

$$\text{interval de încredere } 95\% = [\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}] = [0.8372198, 13.16278].$$

$$\text{interval de încredere } 80\% = [\bar{x} - t_{0.1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{0.1} \cdot \frac{s}{\sqrt{n}}] = [3.828522, 10.17148].$$

$$\text{interval de încredere } 50\% = [\bar{x} - t_{0.25} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{0.25} \cdot \frac{s}{\sqrt{n}}] = [5.518792, 8.481208].$$

Toate aceste intervale de încredere dau estimări interval pentru valoarea  $\mu$ . Din nou, dacă nivelul de încredere este mai mare, atunci intervalul corespunzător este mai lung.

## 1.6 Intervale de încredere $\chi^2$ pentru dispersie

**Definiție.** Presupunem că datele  $x_1, \dots, x_n$  provin din  $N(\mu, \sigma^2)$  cu media  $\mu$  și deviația standard  $\sigma$  ambele necunoscute. Intervalul de încredere  $(1 - \alpha)$  pentru dispersia  $\sigma^2$  este

$$\left[ \frac{(n-1)s^2}{c_{\alpha/2}}, \frac{(n-1)s^2}{c_{1-\alpha/2}} \right]. \quad (9)$$

Aici  $c_{\alpha/2}$  este **valoarea critică din dreapta** a. i.  $P(X^2 > c_{\alpha/2}) = \alpha/2$  pentru  $X^2 \sim \chi^2(n-1)$  și  $s^2$  este dispersia de selecție a datelor.

Deducerea acestui interval este aproape identică cu deducerile anterioare, acum plecând de la testul  $\chi^2$  pentru dispersie. Faptul de bază de care avem nevoie este că, pentru date provenite din  $N(\mu, \sigma^2)$  cu  $\sigma$  cunoscută, statistica

$$\frac{(n-1)s^2}{\sigma^2}$$

are o repartiție  $\chi^2$  cu  $n-1$  grade de libertate. Deci, dată fiind ipoteza 0  $H_0$ :  $\sigma = \sigma_0$ , statistica testului este  $(n-1)s^2/\sigma_0^2$  și regiunea de nerespingere la nivelul de semnificație  $\alpha$  este

$$c_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma_0^2} \leq c_{\alpha/2}.$$

Aceasta este echivalentă cu

$$\frac{(n-1)s^2}{c_{1-\alpha/2}} \geq \sigma_0^2 \geq \frac{(n-1)s^2}{c_{\alpha/2}}.$$

Aceasta spune că nu respingem dacă

$$\sigma_0^2 \in \left[ \frac{(n-1)s^2}{c_{\alpha/2}}, \frac{(n-1)s^2}{c_{1-\alpha/2}} \right].$$

Acesta este intervalul nostru de încredere  $(1 - \alpha)$ .

## 2 Intervale de încredere: 3 puncte de vedere

### 2.1 Scopurile învățării

1. Să poată produce intervale de încredere  $z$ ,  $t$  și  $\chi^2$  pe baza statisticilor standardizate corespunzătoare.
2. Să poată să utilizeze un test de ipoteză pentru a construi intervalul de încredere pentru un parametru necunoscut.
3. Să refuze să răsundă la întrebarări de genul ”dat fiind un interval de încredere, care este probabilitatea sau şansele ca el să conțină adevărata valoare a parametrului necunoscut?”

### 2.2 Introducere

Abordarea noastră a intervalelor de încredere a fost o combinație între statisticile standardizate și testarea ipotezelor. Azi vom considera fiecare dintre aceste perspective separat și vom introduce și un al 3-lea punct de vedere formal. Fiecare dă propria sa perspectivă.

1. **Statistica standardizată.** Cele mai multe intervale de încredere sunt bazate pe statistică standardizată cu repartiții cunoscute ca  $z$ ,  $t$  sau  $\chi^2$ . Aceasta dă un mod direct de a construi și interpreta intervalele de încredere ca o estimare punctuală plus sau minus o eroare.

2. **Testarea ipotezei.** Intervale de încredere pot fi de asemenea construite din teste de ipoteză. În cazurile unde nu avem o statistică standardizată această metodă încă funcționează. Este în acord cu abordarea statisticii standardizate în cazurile unde se aplică ambele.

Această perspectivă leagă noțiunile de nivel de semnificație  $\alpha$  pentru testarea ipotezei și nivel de încredere  $(1 - \alpha)$  pentru intervale de încredere; în ambele cazuri  $\alpha$  este probabilitatea de a face o eroare de ”tipul 1”. Aceasta dă o înțelegere pentru folosirea cuvântului ”încredere”. Această perspectivă ajută de asemenea la evidențierea naturii frecvenționiste a intervalelor de încredere.

3. **Formal.** Definiția formală a intervalelor de încredere este precisă și generală. Într-un sens matematic dă o perspectivă în funcționarea interioară a intervalelor de încredere. Totuși, deoarece este atât de generală, uneori duce la intervale fără proprietăți utile.

### 2.3 Intervale de încredere via statistici standardizate

Strategia de aici este în esență aceeași ca aceea anterioară. Presupunând date normale, avem ceea ce numim statistică standardizată, ca media standardizată, media Studentizată și dispersia standardizată. Aceste statistică

au repartiții bine cunoscute care depind de valorile ipotetice ale lui  $\mu$  și  $\sigma$ . Apoi folosim calcul algebric pentru a produce intervale de încredere pentru  $\mu$  și  $\sigma$ .

Ideea subiacentă intervalelor de încredere este în esență simplă: plecăm cu o statistică standardizată (de exemplu  $z$ ,  $t$  sau  $\chi^2$ ) și folosim ceva calcul algebric pentru a obține un interval care depinde doar de date și parametrii **cunoscuți**.

### 2.3.1 Intervale de încredere $z$ pentru $\mu$ : date normale cu $\sigma$ cunoscută

Intervalele de încredere  $z$  pentru media datelor normale sunt bazate pe **media standardizată**, adică pe statistica  $z$ . Plecăm cu  $n$  date independente normale

$$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2).$$

Presupunem că  $\mu$  este parametrul necunoscut de interes și  $\sigma$  este cunoscută. Știm că media standardizată este normală standard:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Pentru valoarea critică normală standard  $z_{\alpha/2}$  avem:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

De aceea,

$$P\left(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} | \mu\right) = 1 - \alpha.$$

Un pic de algebră pune aceasta în forma unui interval în jurul lui  $\mu$ :

$$P\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} | \mu\right) = 1 - \alpha.$$

Putem sublinia că intervalul depinde doar de statistica  $\bar{x}$ ,  $n$  și valoarea cunoscută  $\sigma$  scriind aceasta ca

$$P\left(\mu \in \left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] | \mu\right) = 1 - \alpha.$$

Acesta este intervalul de încredere  $z$   $(1 - \alpha)$  pentru  $\mu$ . Adesea îl scriem folosind prescurtarea

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Gândiți-l ca  $\bar{x} \pm$  eroarea.

### 2.3.2 Intervale de încredere $t$ pentru $\mu$ : date normale cu $\mu$ și $\sigma$ necunoscute

Intervalele de încredere  $t$  pentru media datelor normale sunt bazate pe [media Studentizată](#), adică pe statistica  $t$ . Plecăm cu  $n$  date independente normale

$$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2).$$

Presupunem că  $\mu$  este parametrul necunoscut de interes și  $\sigma$  este necunoscută. Știm că media Studentizată are o repartiție Student  $t$  cu  $n - 1$  grade de libertate:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n - 1),$$

unde  $s^2$  este dispersia de selecție.

Acum, tot ce avem de făcut este să înlocuim media standardizată cu [media Studentizată](#) și aceeași logică pe care am utilizat-o pentru  $z$  ne dă intervalul de încredere  $t$ : plecăm cu

$$P\left(-t_{\alpha/2} \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2} | \mu\right) = 1 - \alpha.$$

Un pic de algebră pune aceasta în forma unui interval în jurul lui  $\mu$ :

$$P\left(\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} | \mu\right) = 1 - \alpha.$$

Putem sublinia că intervalul depinde doar de statisticile  $\bar{x}$  și  $s$  și de  $n$  scriind aceasta ca

$$P\left(\mu \in \left[\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] | \mu\right) = 1 - \alpha.$$

Acesta este intervalul de încredere  $t$   $(1 - \alpha)$  pentru  $\mu$ . Adesea îl scriem folosind prescurtarea

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

Gândiți-l ca  $\bar{x} \pm$  eroarea.

### 2.3.3 Intervale de încredere $\chi^2$ pentru $\sigma^2$ : date normale cu $\mu$ și $\sigma$ necunoscute

Intervalele de încredere  $\chi^2$  pentru dispersia datelor normale sunt bazate pe [dispersia standardizată](#), adică pe statistica  $\chi^2$ .

Urmăm aceeași logică ca mai sus pentru a obține un interval de încredere  $\chi^2$

pentru  $\sigma^2$ .

Presupunem că avem  $n$  date normale independente:  $x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ .

Presupunem că  $\mu$  și  $\sigma$  sunt ambele necunoscute. Dispersia standardizată este

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

Ştim că statistica  $X^2$  are o repartiție  $\chi^2$  cu  $n-1$  grade de libertate.

Pentru  $Z$  și  $t$  am folosit simetria repartițiilor pentru a înlocui  $z_{1-\alpha/2}$  cu  $-z_{\alpha/2}$  și  $t_{1-\alpha/2}$  cu  $-t_{\alpha/2}$ . Deoarece repartiția  $\chi^2$  nu este simetrică trebuie să fim expliciti despre valorile critice atât la stânga cât și la dreapta. Adică,

$$P(c_{1-\alpha/2} \leq X^2 \leq c_{\alpha/2}) = 1 - \alpha,$$

unde  $c_{\alpha/2}$  și  $c_{1-\alpha/2}$  sunt valorile critice din [coada dreaptă](#). Deci

$$P\left(c_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq c_{\alpha/2} | \sigma\right) = 1 - \alpha.$$

Un pic de algebră pune aceasta în forma unui interval în jurul lui  $\sigma^2$ :

$$P\left(\frac{(n-1)s^2}{c_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{c_{1-\alpha/2}} | \sigma\right) = 1 - \alpha.$$

Putem sublinia că intervalul depinde doar de statistica  $s^2$  și de  $n$  scriind aceasta ca

$$P\left(\sigma^2 \in \left[\frac{(n-1)s^2}{c_{\alpha/2}}, \frac{(n-1)s^2}{c_{1-\alpha/2}}\right] | \sigma\right) = 1 - \alpha.$$

Acesta este intervalul de încredere  $\chi^2(1-\alpha)$  pentru  $\sigma^2$ .

## 2.4 Intervale de încredere via testarea ipotezei

Presupunem că avem date provenite dintr-o repartiție cu un parametru  $\theta$  a cărui valoare este necunoscută. Un test de semnificație pentru valoarea  $\theta$  are următoarea descriere scurtă:

1. Fixăm ipoteza 0  $H_0: \theta = \theta_0$  pentru o valoare specială  $\theta_0$ , de exemplu adesea avem  $H_0: \theta = 0$ .
2. Folosim datele pentru a calcula valoarea unei statistici a testului, s-o numim  $x$ .
3. Dacă  $x$  este destul de departe în coada repartiției 0 (repartiția [presupunând](#) ipoteza 0), atunci respingem  $H_0$ .

În cazul unde nu este o valoare specială pentru a fi testată, putem dori încă

să estimăm  $\theta$ . Aceasta este reversul testului de semnificație: în loc să vedem dacă ar trebui să respingem o anumită valoare a lui  $\theta$  deoarece nu se potrivește cu datele, mai degrabă vrem să găsim un domeniu de valori ale lui  $\theta$  care, într-un anumit sens, se potrivesc cu datele.

Aceasta ne dă următoarele definiții:

**Definiție.** Fiind dată o valoare  $x$  a statisticii testului, intervalul de încredere  $(1-\alpha)$  conține toate valorile  $\theta_0$  care nu sunt respinse (la nivelul de semnificație  $\alpha$ ) când aceste valori sunt ipoteza 0.

**Definiție.** O eroare de tipul 1 CI apare când intervalul de încredere nu conține adevarata valoare a lui  $\theta$ .

Pentru un interval de încredere  $(1 - \alpha)$  rata erorii de tipul 1 CI este  $\alpha$ .

**Exemplul 1.** Iată un exemplu care leagă intervalele de încredere și testele ipotezei. Presupunem că datele  $x$  provin dintr-o repartiție binomială  $(12, \theta)$  cu  $\theta$  necunoscută. Fie  $\alpha = 0.1$  și creăm intervalul de încredere  $(1 - \alpha) = 90\%$  pentru fiecare valoare posibilă a lui  $x$ .

Strategia noastră este să considerăm pe rând câte o valoare posibilă a lui  $\theta$  și să alegem regiunile de respingere pentru testul de semnificație cu  $\alpha = 0.1$ .

Odată ce acest lucru este făcut, vom ști, pentru orice valoare a lui  $x$ , ce valori ale lui  $\theta$  nu sunt respinse, i.e. intervalul de încredere asociat cu  $x$ .

Pentru a începe, facem un tabel de verosimilitate pentru binomiala  $(12, \theta)$  (tabelul 1). Fiecare linie arată probabilitățile  $p(x|\theta)$  pentru o valoare a lui  $\theta$ .

Pentru a păstra mărimea rezonabilă arătăm  $\theta$  doar în creșteri de 0.1.

$\theta \setminus x$	0	1	2	3	4	5	6	7	8	9	10	11	12
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28	
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

**Tabelul 1.** Tabelul de verosimilitate pentru binomiala  $(12, \theta)$

Tabelele 2-4 de mai jos arată regiunea de respingere (în portocaliu) și regiunea de nerespingere (în albastru) pentru diferite valori ale lui  $\theta$ . Pentru a sublinia natura linie cu linie a procesului, tabelul 2 arată aceste regiuni pentru  $\theta = 1$ , apoi tabelul 3 adaugă regiunile pentru  $\theta = 0.9$  și tabelul 4 le arată pentru toate valorile lui  $\theta$ .

Imediat după tabele explicăm detaliat cum au fost alese regiunile de respingere/nerespingere. "significance" = "semnificație".

$\theta \setminus x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28		
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07	
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01	
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00	
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00	
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

**Tabelul 2.** Tabelul de verosimilitate pentru binomiala(12,  $\theta$ ) cu regiunile de respingere/nerespingere pentru  $\theta = 1$

$\theta \setminus x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28		0.026
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07	
0.7	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01	
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02	0.00	
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00	0.00	
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00	0.00	
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00	0.00	
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

**Tabelul 3.** Tabelul de verosimilitate cu regiunile de respingere/nerespingere arătate pentru  $\theta = 1$  și 0.9

$\theta \setminus x$	0	1	2	3	4	5	6	7	8	9	10	11	12	significance
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.000
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.23	0.38	0.28		0.026
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	0.13	0.24	0.28	0.21	0.07	
0.7	0.00	0.00	0.00	0.01	0.03	0.08	0.16	0.23	0.24	0.17	0.07	0.01		0.052
0.6	0.00	0.00	0.00	0.01	0.04	0.10	0.18	0.23	0.21	0.14	0.06	0.02		0.077
0.5	0.00	0.00	0.02	0.05	0.12	0.19	0.23	0.19	0.12	0.05	0.02	0.00		0.092
0.4	0.00	0.02	0.06	0.14	0.21	0.23	0.18	0.10	0.04	0.01	0.00	0.00		0.077
0.3	0.01	0.07	0.17	0.24	0.23	0.16	0.08	0.03	0.01	0.00	0.00	0.00		0.052
0.2	0.07	0.21	0.28	0.24	0.13	0.05	0.02	0.00	0.00	0.00	0.00	0.00		0.073
0.1	0.28	0.38	0.23	0.09	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.026
0.0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

**Tabelul 4.** Tabelul de verosimilitate cu regiunile de respingere/nerespingere

**Alegerea regiunilor de respingere și nerespingere în tabele** Prima problemă cu care ne confruntăm este cum să alegem exact regiunea de respingere. Am folosit 2 reguli:

- Probabilitatea totală a regiunii de respingere, i.e. semnificația, trebuie să fie mai mică sau egală cu 0.1. (Deoarece avem o repartiție discretă este imposibil să facem semnificația exact 0.1.)
- Construim regiunea de respingere alegând pe rând valorile lui  $x$ , tot-

deauna luând valoarea nefolosită cu cea mai mică probabilitate. Ne oprim când valoarea următoare ar face semnificația mai mare ca 0.1.

Sunt și alte moduri de a alege regiunea de respingere din care ar putea rezulta mici diferențe.

Tabelul 2 arată regiunile de respingere (portocalie) și nerespinger (albastră) pentru  $\theta = 1$ . Acesta este un caz special deoarece cele mai multe probabilități de pe această linie sunt 0. Ne mutăm la tabelul următor și parcurgem procesul pentru acesta.

În tabelul 3, parcurgem pașii pentru a afla regiunile pentru  $\theta = 0.9$ .

Cea mai mică probabilitate este când  $x = 0$  și aceasta este mai mică decât 0.1, deci  $x = 0$  este în regiunea de respingere.

Următoarea cea mai mică probabilitate este când  $x = 1$  și suma celor 2 probabilități (pentru  $x = 0$  și  $x = 1$ ) este mai mică decât 0.1, deci  $x = 1$  este în regiunea de respingere.

Continuăm cu  $x = 2, 3, \dots, 8$ . În acest punct probabilitatea totală din regiunea de respingere este 0.026.

Următoarea cea mai mică probabilitate este când  $x = 9$ . Adunarea acestei probabilități (0.09) la 0.026 ar face probabilitatea totală peste 0.1. Deci lăsăm  $x = 9$  în afara regiunii de respingere și oprim procesul.

Observăm 3 lucruri pentru linia  $\theta = 0.9$ :

1. Niciuna din probabilitățile de pe această linie nu este cu adevărat 0, totuși unele dintre ele sunt suficient de mici astfel încât aproximarea lor la 2 zecimale este 0.
2. Arătăm semnificația pentru această valoare a lui  $\theta$  în marginea din dreapta. Mai precis, arătăm nivelul de semnificație al NHST cu ipoteza  $0 < \theta = 0.9$  și regiunea de respingere dată.
3. Regiunea de respingere constă din valori ale lui  $x$ . Când spunem că regiunea de respingere este arătată în portocaliu, în realitate ne referim la faptul că regiunea de respingere constă din valorile lui  $x$  corespunzătoare probabilităților evidențiate cu portocaliu.

**Gândiți:** Întoarceți-vă la linia  $\theta = 1$  și asigurați-vă că înțelegeți de ce regiunea de respingere este  $x \in \{0, 1, \dots, 11\}$  și semnificația este 0.

**Exemplul 2.** Folosind tabelul 4, determinați intervalul de încredere 0.9 când  $x = 8$ .

**Răspuns.** Intervalul de încredere 90% constă din toate acele  $\theta$  care nu ar fi respinse de un test al ipotezei cu  $\alpha = 0.1$  când  $x = 8$ . Privind la tabel, intrările albastre (nerespinse) din coloana  $x = 8$  corespund la  $0.5 \leq \theta \leq 0.8$ ; intervalul de încredere este  $[0.5, 0.8]$ .

**Observație.** Scopul acestui exemplu este de a arăta cum sunt legate intervalele de încredere și testele ipotezei. Deoarece tabelul 4 are doar un număr finit de valori pentru  $\theta$ , răspunsul nostru este aproape, dar nu exact. Pentru

această problemă am putea utiliza R pentru a afla că, corect la 2 zecimale, intervalul de încredere este  $[0.42, 0.84]$ .

**Exemplul 3.** Explicați de ce rata erorii CI de tipul 1 va fi cel mult 0.092, cu condiția ca adevărata valoare a lui  $\theta$  să fie în tabel.

**Răspuns.** Răspunsul scurt este că acesta este maximul semnificațiilor pentru orice  $\theta$  din tabelul 4. Mai pe larg: facem o eroare CI de tipul 1 dacă intervalul de încredere nu conține adevărata valoare a lui  $\theta$ , s-o numim  $\theta_{\text{adevărată}}$ . Aceasta se întâmplă exact când data  $x$  este în regiunea de respingere pentru  $\theta_{\text{adevărată}}$ . Probabilitatea ca aceasta să se întâpte este semnificația pentru  $\theta_{\text{adevărată}}$  și aceasta este cel mult 0.092.

**Observație.** Scopul acestui exemplu este să arate cum nivelul de încredere, rata erorii CI de tipul 1 și semnificația pentru fiecare ipoteză sunt legate. Ca în exemplul precedent, putem folosi R pentru a calcula semnificația pentru mult mai multe valori ale lui  $\theta$ . Când facem asta aflăm că semnificația maximă pentru orice  $\theta$  este 0.1 apărând când  $\theta = 0.04524072$ .

**Observații de sinteză:**

1. Începem cu o statistică de test  $x$ . Intervalul de încredere este aleator deoarece depinde de  $x$ .
2. Pentru fiecare valoare ipotetică a lui  $\theta$  facem un test de semnificație cu nivelul de semnificație  $\alpha$  alegând regiunile de respingere.
3. Pentru o valoare specifică a lui  $x$  intervalul de încredere asociat pentru  $\theta$  constă din toți  $\theta$  care nu au fost respinși pentru acea valoare, i.e. toți  $\theta$  care au  $x$  în regiunile lor de nerespingere.
4. Deoarece repartitia este discretă nu putem atinge totdeauna nivelul exact de semnificație, deci intervalul de încredere al nostru este în realitate un "interval de încredere cel puțin 90%".

**Exemplul 4.** Deschideți aplicația

<http://mathlets.org/mathlets/confidence-intervals/>. Jucați-vă cu aplicația pentru a înțelege natura aleatoare a intervalelor de încredere și sensul încrederii ca (1 - rata erorii CI de tipul 1).

- a) Citiți ajutorul. Este scurt și vă va ajuta să vă orientați în aplicație. Jucați-vă cu diferite setări ale parametrilor pentru a vedea cum afectează ele mărimea intervalelor de încredere.
- b) Stabiliți numărul experimentelor la  $N = 1$ . Apăsați butonul "Run N trials" repetat și vedeți că de fiecare dată când datele sunt generate intervalele de încredere sărăcăză.
- c) Acum puneți nivelul de încredere la  $c = .5$ . Când apăsați butonul "Run N trials" ar trebui să vedeți că aproximativ 50% din intervalele de încredere includ adevărata valoare a lui  $\mu$ . Valorile "Z correct" și "t correct" ar trebui să se schimbe în consecință.
- d) Acum puneți numărul încercărilor la  $N = 100$ , cu  $c = .8$ . Butonul "Run

$N$  trials” va face acum 100 de încercări o dată. Doar ultimul interval de încredere va fi arătat în grafic, dar încercările sunt făcute toate și statistica ”percent correct” va fi actualizată pe baza tuturor celor 100 de încercări. Apăsați butonul ”Run  $N$  trials” repetat. Urmăriți cum ratele corecte încep să convergă la nivelul de încredere. Pentru a converge și mai rapid, puneti  $N = 1000$ .

## 2.5 Viziune formală a intervalelor de încredere

Reamintim: O statistică interval este un interval  $I_x$  calculat din datele  $x$ . Un interval este determinat de marginile lui inferioară și superioară și acestea sunt aleatoare deoarece  $x$  sunt aleatoare.

Presupunem că  $x$  provin dintr-o repartiție cu pdf  $f(x|\theta)$ , unde parametrul  $\theta$  este necunoscut.

**Definiție:** Un interval de încredere  $(1 - \alpha)$  pentru  $\theta$  este o statistică interval  $I_x$  astfel încât

$$P(\theta_0 \in I_x | \theta = \theta_0) = 1 - \alpha$$

pentru toate valorile posibile ale lui  $\theta_0$ .

Această definiție este un mod de a cântări dovezile produse de datele  $x$ .

Nivelul de încredere a unei statistici interval este o probabilitate privind un interval aleator și o valoare ipotetică  $\theta_0$  pentru parametrul necunoscut. Exact, este probabilitatea ca intervalul aleator (calculat din datele aleatoare) să conțină valoarea  $\theta_0$ , **dat fiind că parametrul modelului este cu adevărat  $\theta_0$** . Deoarece adevărată valoare a lui  $\theta$  este necunoscută, statisticianul frecvenționist definește intervalele de încredere 95% astfel încât probabilitatea 0.95 este validă **indiferent care valoare ipotetică a parametrului este de fapt adevărată**.

## 2.6 Comparație cu intervalele de probabilitate Bayesiene

Intervalele de încredere sunt o noțiune frecvenționistă și frecvenționiștii nu atribuie probabilități ipotezelor, de exemplu valorii unui parametru necunoscut. Mai degrabă ei calculează verosimilități; adică, probabilități despre date sau statistici asociate dată fiind o ipoteză (observați condiția  $\theta = \theta_0$  din viziunea formală a intervalelor de încredere). Construcția intervalelor de încredere provine în întregime din tabelul complet de verosimilitate.

În contrast, intervalele de probabilitate a posteriori Bayesiene dau într-adevăr probabilitatea ca valoarea parametrului necunoscut să fie în domeniul raportat. Adăugăm avertismentul ușual că aceasta depinde de alegerea specifică a unei a priori Bayesiene (posibil subiectivă).

Această distincție între cele 2 este subtilă deoarece intervalele de probabilitate a posteriori Bayesiene și intervalele de încredere frecvenționiste împart următoarele proprietăți:

1. Ele pleacă de la un model  $f(x|\theta)$  pentru datele observate  $x$  cu parametrul necunoscut  $\theta$ .
2. Date fiind datele  $x$ , ele dă un interval  $I(x)$  specificând un domeniu de valori pentru  $\theta$ .

3. Ele vin cu un număr (să zicem 0.95) care este probabilitatea a ceva.

În practică, mulți oameni interpretează greșit intervalele de încredere ca intervale de probabilitate Bayesiene, uitând că frecvenționistii nu pun **nicio-dată** probabilități peste ipoteze (aceasta este analoga greșelii că valoarea  $p$  în NHST este probabilitatea că  $H_0$  este falsă). Dauna acestei interpretări greșite este cumva atenuată de faptul că, fiind date destule date și o a priori rezonabilă, intervalele Bayesian și frecvenționist adesea se dovedesc a fi destul de similare.

Pentru un exemplu amuzant ilustrând cât de diferite pot fi, vezi primul răspuns de aici (care implică fursecuri cu ciocolată!):

<http://stats.stackexchange.com/questions/2272/whats-the-difference-between-a->

Acest exemplu folosește definiții formale și este în realitate despre multimi de încredere în locul intervalelor de încredere.

# Curs 19

Cristian Niculescu

## 1 Intervale de încredere pentru media datelor nenormale

### 1.1 Scopurile învățării

1. Să poată deduce formula pentru intervale de încredere normale conservatoare pentru proporția  $\theta$  din date Bernoulli.
2. Să poată calcula intervale de încredere 95% cu regula degetului mare pentru proporția  $\theta$  a unei repartiții Bernoulli.
3. Să poată calcula intervale de încredere de selecție mare pentru media unei repartiții generale.

### 1.2 Introducere

Până acum, ne-am concentrat pe construirea intervalelor de încredere pentru date provenite dintr-o repartitie normală. Acum vom învăța despre intervale de încredere pentru medie când datele nu sunt neapărat normale.

Întâi vom studia estimarea probabilității  $\theta$  de succes când datele provin din o repartitie Bernoulli( $\theta$ ) - reamintim că  $\theta$  este de asemenea media repartitiei Bernoulli.

Apoi vom considera cazul unei selecții mari dintr-o repartitie necunoscută; în acest caz putem apela la teorema limită centrală pentru a justifica intervalele de încredere  $z$ .

### 1.3 Datele Bernoulli și votarea

O utilizare obișnuită a intervalelor de încredere este pentru estimarea proporției  $\theta$  dintr-o repartitie Bernoulli( $\theta$ ). De exemplu, presupunem că vrem să folosim un sondaj politic pentru a estima proporția din populație care susține candidatul A, sau echivalent, probabilitatea  $\theta$  ca o persoană aleatoare să susțină candidatul A. În acest caz avem o regulă a degetului mare care ne permite să calculăm repede un interval de încredere.

### 1.3.1 Intervale de încredere normale conservatoare

Presupunem că avem datele i.i.d.  $x_1, x_2, \dots, x_n$  toate provenind dintr-o repartiție Bernoulli( $\theta$ ). Atunci un interval de încredere  $(1 - \alpha)$  **normal conservator** pentru  $\theta$  este dat de

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}. \quad (1)$$

Demonstrația dată mai jos folosește teorema limită centrală și observația că  $\sigma = \sqrt{\theta(1 - \theta)} \leq 1/2$ .

Veți vedea de asemenea în deducerea de mai jos că această formulă este conservatoare, dând un interval de încredere "cel puțin  $(1 - \alpha)$ ".

**Exemplul 1.** Un sondaj întreabă 196 de persoane dacă preferă candidatul A candidatului B și află că 120 preferă A și 76 preferă B. Aflați intervalul de încredere normal conservator pentru  $\theta$ , proporția din populație care preferă A.

**Răspuns.** Avem  $\bar{x} = 120/196 \approx 0.612$ ,  $\alpha = 0.05$  și  $z_{0.025} \approx 1.96$ . Formula spune că un interval de încredere 95% este

$$I \approx 0.612 \pm \frac{1.96}{2 \cdot 14} = 0.612 \pm 0.07 = [0.542, 0.682].$$

### 1.3.2 Demonstrația formulei (1)

Demonstrația formulei (1) se va baza pe următoarea lemă.

**Lemă.** Deviația standard a unei repartiții Bernoulli( $\theta$ ) este cel mult 0.5.

**Demonstrația lemei.** Notăm această deviație standard cu  $\sigma_\theta$  pentru a sublinia dependența ei de  $\theta$ . Atunci dispersia este  $\sigma_\theta^2 = \theta(1 - \theta)$ . Este ușor de văzut folosind analiza matematică sau proprietățile funcției de gradul 2 sau făcând graficul parabolei că maximul apare când  $\theta = 1/2$ . De aceea dispersia maximă este  $1/4$ , ceea ce implică faptul că deviația standard  $\sigma$  este cel mult  $\sqrt{1/4} = 1/2$ .

**Demonstrația formulei (1).** Demonstrația se bazează pe teorema limită centrală care spune că (pentru  $n$  mare) repartiția lui  $\bar{x}$  este aproximativ normală cu media  $\theta$  și deviația standard  $\sigma_\theta/\sqrt{n}$ . Pentru date normale avem intervalul de încredere  $z(1 - \alpha)$

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma_\theta}{\sqrt{n}}.$$

Acum trucul este să înlocuim  $\sigma_\theta$  cu  $\frac{1}{2}$ : deoarece  $\sigma_\theta \leq \frac{1}{2}$ , intervalul rezultat în jurul lui  $\bar{x}$

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}$$

este cel puțin la fel de lung ca intervalul folosind  $\pm\sigma_\theta/\sqrt{n}$ . Un interval mai lung este mai probabil să conțină adevărata valoare a lui  $\theta$ , deci avem un interval de încredere  $(1 - \alpha)$  ”conservator” pentru  $\theta$ .

Din nou, numim acesta **conservator deoarece  $\frac{1}{2\sqrt{n}}$  supraestimează deviația standard a lui  $\bar{x}$** , rezultând un interval mai lung decât este necesar pentru a atinge un nivel de încredere de  $(1 - \alpha)$ .

### 1.3.3 Cum sunt raportate sondajele politice

Sondajele politice sunt adesea raportate ca o valoare cu o marjă de eroare. De exemplu puteți auzi

52% favorizează candidatul A cu o marjă de eroare de  $\pm 5\%$ .

Sensul precis al acestui fapt este dacă  $\theta$  este proporția din populație care sprijină pe A, atunci estimarea punctuală pentru  $\theta$  este 52% și intervalul de încredere 95% este  $52\% \pm 5\%$ . Observați că reporterii sondajelor la știri nu menționează încrederea 95%. Doar va trebui să știi că asta este ceea ce fac sondajele.

#### Intervalul de încredere cu regula degetului mare 95%.

Reamintim că intervalul de încredere normal conservator  $(1 - \alpha)$  este

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{1}{2\sqrt{n}}.$$

Dacă folosim aproximarea standard  $z_{0.025} \approx 2$  (în loc de 1.96) obținem **intervalul de încredere 95% cu regula degetului mare** pentru  $\theta$ :

$$\bar{x} \pm \frac{1}{\sqrt{n}}.$$

**Exemplul 2.** Votare. Presupunem că în curând vor fi alegeri locale între candidatul A și candidatul B. Presupunem că fracția din populația cu drept de vot care-l susține pe A este  $\theta$ .

2 organizații de sondaj întreabă votanții pe cine preferă.

1. Firma *Rapid și primul* sondează 40 de votanți aleatori și află că 22 sprijină pe A.

2. Firma *Rapid, dar precaut* sondează 400 de votanți aleatori și află că 190 sprijină pe A.

Aflați estimările punctuale și intervalele de încredere cu regula degetului mare 95%. Explicați cum statisticile reflectă intuiția că sondajul a 400 de votanți este mai exact.

**Răspuns.** Pentru sondajul 1 avem

Estimare punctuală:  $\bar{x} = 22/40 = 0.55$ .

Interval de încredere:  $\bar{x} \pm \frac{1}{\sqrt{n}} = 0.55 \pm \frac{1}{\sqrt{400}} = 0.55 \pm 0.16 = 55\% \pm 16\%$ .

Pentru sondajul 2 avem

Estimare punctuală:  $\bar{x} = 190/400 = 0.475$ .

Interval de încredere:  $\bar{x} \pm \frac{1}{\sqrt{n}} = 0.475 \pm \frac{1}{\sqrt{400}} = 0.475 \pm 0.05 = 47.5\% \pm 5\%$ .

Precizia mai mare a sondajului cu 400 de votanți este reflectată de marginea mai mică a erorii, i.e. 5% pentru sondajul cu 400 de votanți vs. 16% pentru sondajul cu 40 de votanți.

### Intervale de încredere pentru proporția binomială

Sunt multe metode de a produce intervale de încredere pentru proporția  $p$  a unei repartiții binomiale( $n, p$ ). Pentru un număr de abordări uzuale, vezi:

[http://en.wikipedia.org/wiki/Binomial\\_proportion\\_confidence\\_interval](http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval).

## 1.4 Intervale de încredere de selecție mare

Un scop tipic în statistică este să estimăm media unei repartiții. Când datele au o repartitie normală am putea folosi intervale de încredere bazate pe statistici standardizate pentru a estima media.

Dar presupunem că datele  $x_1, x_2, \dots, x_n$  provin dintr-o repartitie cu pmf sau pdf  $f(x)$  care poate să nu fie normală sau chiar parametrică. Dacă repartitia are medie și dispersie finite și dacă  $n$  este suficient de mare, atunci următoarea versiune a teoremei limită centrală ne arată că încă putem folosi o statistică standardizată.

**Teorema limită centrală.** Pentru  $n$  mare, repartitia de selecție a mediei Studentizate este aproximativ normală standard:  $\frac{\bar{x} - \mu}{s/\sqrt{n}} \approx N(0, 1)$ .

Deci, pentru  $n$  mare, intervalul de încredere  $(1 - \alpha)$  pentru  $\mu$  este aproximativ

$$\left[ \bar{x} - \frac{s}{\sqrt{n}} \cdot z_{\alpha/2}, \bar{x} + \frac{s}{\sqrt{n}} \cdot z_{\alpha/2} \right],$$

unde  $z_{\alpha/2}$  este valoarea critică  $\alpha/2$  pentru  $N(0, 1)$ . Aceasta este numit **intervalul de încredere de selecție mare**.

### Exemplul 3. Cât de mare trebuie să fie $n$ ?

Reamintim că o eroare CI de tipul 1 apare când intervalul de încredere nu conține adevărată valoare a parametrului, în acest caz media. Să numim valoarea  $(1 - \alpha)$  nivelul de încredere *nominal*. Spunem nominal deoarece dacă  $n$  nu este mare, nu ar trebui să ne așteptăm ca adevărată rată a erorii CI de tipul 1 să fie  $\alpha$ .

Putem face simulări numerice pentru a approxima adevăratul nivel de încredere. Ne așteptăm că, atunci când  $n$  devine mai mare, adevăratul nivel de încredere al intervalului de încredere de dispersie mare să conveargă la valoarea nominală.

Facem astfel de simulări pentru  $x$  provenite de la repartiția exponențială  $\exp(1)$  (care este departe de normală). Pentru câteva valori ale lui  $n$  și nivelul de încredere nominal  $c$  facem 100000 de încercări. Fiecare încercare constă din următorii pași:

1. Extragem  $n$  date din  $\exp(1)$ .
2. Calculăm media de selecție  $\bar{x}$  și deviația standard de selecție  $s$ .
3. Construim intervalul de încredere  $c$  de selecție mare:  $\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ .
4. Verificăm pentru o eroare CI de tipul 1, i.e. vedem dacă adevărata medie  $\mu = 1$  nu este în interval.

Cu 100000 de încercări, nivelul de încredere empiric ar trebui să aproximeze bine adevăraturul nivel.

Pentru comparație facem aceleași teste pe date provenite dintr-o repartiție normală standard. Iată rezultatele:

$n$	nominal conf. $1 - \alpha$	simulated conf.
20	0.95	0.905
20	0.90	0.856
20	0.80	0.762
50	0.95	0.930
50	0.90	0.879
50	0.80	0.784
100	0.95	0.938
100	0.90	0.889
100	0.80	0.792
400	0.95	0.947
400	0.90	0.897
400	0.80	0.798

Simulări pentru  $\exp(1)$

$n$	nominal conf. $1 - \alpha$	simulated conf.
20	0.95	0.936
20	0.90	0.885
20	0.80	0.785
50	0.95	0.944
50	0.90	0.894
50	0.80	0.796
100	0.95	0.947
100	0.90	0.896
100	0.80	0.797
400	0.95	0.949
400	0.90	0.898
400	0.80	0.798

Simulări pentru  $N(0, 1)$

Pentru repartiția  $\exp(1)$  vedem că pentru  $n = 20$  încrederea simulată a intervalului de încredere de selecție mare este mai mică decât încrederea nominală  $(1 - \alpha)$ . Dar pentru  $n = 100$  încrederea simulată și încrederea nominală sunt foarte aproape. Deci, pentru  $\exp(1)$ ,  $n$  undeva între 50 și 100 este destul de mare pentru cele mai multe scopuri.

**Gândiți.** Pentru  $n = 20$  de ce încrederea simulată pentru repartiția  $N(0, 1)$  este mai mică decât încrederea nominală?

Aceasta este deoarece am folosit  $z_{\alpha/2}$  în loc de  $t_{\alpha/2}$ . Pentru  $n$  mare acestea sunt foarte aproape, dar pentru  $n = 20$  există o diferență sesizabilă, de exemplu  $z_{0.025} \approx 1.96$  și  $t_{0.025} \approx 2.09$ .

## 2 Intervale de încredere bootstrap

### 2.1 Scopurile învățării

1. Să poată construi și selecta din repartiția empirică a datelor.
2. Să poată explica principiul bootstrap.
3. Să poată proiecta și folosi un bootstrap empiric pentru a calcula intervale de încredere.
4. Să poată proiecta și folosi un bootstrap parametric pentru a calcula intervale de încredere.

### 2.2 Introducere

Bootstrap-ul empiric este o tehnică statistică popularizată de Bradley Efron în 1979. Cu toate că este remarcabil de simplu de aplicat, bootstrap-ul n-ar fi posibil fără puterea de calcul modernă. Ideea cheie este să facem calcule asupra datelor pentru a estima variația statisticilor care sunt calculate din aceleasi date. "Bootstrap" = "curea de cizmă". (O căutare pe google a "by ones own bootstrap" vă va da o etimologie a acestei metafore.) Astfel de tehnici existau încă dinainte de 1979, dar Efron a lărgit aplicabilitatea lor și a demonstrat modul de aplicare al bootstrap-ului folosind efectiv calculatoarele. El a inventat de asemenea termenul "bootstrap".

Aplicația noastră principală a bootstrap-ului va fi estimarea variației estimărilor punctuale; adică, estimarea intervalelor de încredere.

**Exemplul 1.** Presupunem că avem datele

$$x_1, x_2, \dots, x_n.$$

Dacă am ști că datele provin din  $N(\mu, \sigma^2)$  cu medie necunoscută  $\mu$  și dispersie cunoscută  $\sigma^2$ , atunci am văzut că

$$\left[ \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right],$$

este un interval de încredere 95% pentru  $\mu$ .

Acum presupunem că datele provin dintr-o repartiție complet necunoscută. Pentru a avea un nume, vom numi această repartiție  $F$  și media ei (necunoscută)  $\mu$ . Putem folosi încă media de selecție  $\bar{x}$  ca o [estimare punctuală](#) a lui  $\mu$ . Dar cum putem afla un interval de încredere pentru  $\mu$  în jurul lui  $\bar{x}$ ? Răspunsul nostru va fi să folosim bootstrap-ul!

De fapt, bootstrap-ul tratează alte statistici la fel de ușor cum tratează media. De exemplu: mediana sau alte percentile. Acestea sunt statistici unde,

chiar pentru repartițiile normale, poate fi dificil să calculăm un interval de încredere doar din teorie.

## 2.3 Selectarea

În statistică, a **selecta** dintr-o mulțime însenmă a alege elemente din acea mulțime. Într-o selecție aleatoare, elementele sunt alese aleator. Sunt 2 metode uzuale pentru selectare aleatoare.

### Selectare fără înlocuire

Presupunem că tragem 10 cărți de joc dintr-un pachet de 52 de cărți fără a pune una din cărți înapoi în pachet între trageri. Aceasta este numită **selectare fără înlocuire** sau **selectare aleatoare simplă**. Cu această metodă de selectare selecția noastră de 10 cărți nu va avea cărți dupăcat.

### Selectare cu înlocuire

Acum presupunem că tragem 10 cărți aleator din pachet, dar după fiecare extragere punem cartea la loc în pachet și amestecăm cărțile. Aceasta este numită **selectare cu înlocuire**. Cu această metodă, selecția de 10 cărți poate avea duplicate. Este chiar posibil să tragem 6 de cupă de 10 ori.

**Gândiți:** Care este probabilitatea să tragem 6 de cupă de 10 ori la rând?

**Exemplul 2.** Putem vedea aruncarea repetată a unui zar cu 8 fețe ca selectarea cu înlocuire din mulțimea  $\{1, 2, 3, 4, 5, 6, 7, 8\}$ . Deoarece fiecare număr este egal probabil, spunem că selectăm uniform din date. Aici este o subtilitate: fiecare dată este egal probabilă, dar dacă sunt valori repetitive în date, acele valori vor avea o probabilitate mai mare de a fi alese.

**Observație.** În practică, dacă luăm un număr mic dintr-o mulțime foarte mare, atunci nu contează dacă selectăm cu sau fără înlocuire. De exemplu, dacă selecțăm aleator 400 din cei 300 de milioane de locuitori ai SUA, atunci este atât de improbabil că aceeași persoană va fi aleasă de 2 ori încât nu este o diferență reală între selectarea cu sau fără înlocuire.

## 2.4 Repartitia empirică a datelor

Repartitia empirică a datelor este pur și simplu repartitia pe care o vedeti în date.

**Exemplul 3.** Presupunem că aruncăm un zar cu 8 fețe de 10 ori și obținem următoarele date, scrise în ordine crescătoare:

$$1, 1, 2, 3, 3, 3, 3, 4, 7, 7.$$

Imaginați-vă că scriem aceste valori pe 10 bucățele de hârtie, le punem într-o pălărie și tragem una aleator. Atunci, de exemplu, probabilitatea de a trage

un 3 este  $4/10$  și probabilitatea de a trage un 4 este  $1/10$ . Repartiția empirică completă poate fi pusă într-un tabel de probabilitate ("value" = "valoarea").

value $x$	1	2	3	4	7
$p(x)$	$2/10$	$1/10$	$4/10$	$1/10$	$2/10$

**Notatie.** Dacă notăm adevărata repartiție din care provin datele cu  $F$ , atunci vom nota repartiția empirică a datelor cu  $F^*$ . Dacă avem destule date, atunci legea numerelor mari ne spune că  $F^*$  ar trebui să fie o bună aproximare a lui  $F$ .

**Exemplul 4.** În exemplul cu zaruri de mai sus, repartițiile adevărată și empirică sunt:

value $x$	1	2	3	4	5	5	7	8
true $p(x)$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$
empirical $p(x)$	$2/10$	$1/10$	$4/10$	$1/10$	0	0	$2/10$	0

Adevărata repartiție  $F$  și repartiția empirică  $F^*$  a unui zar cu 8 fețe.

Deoarece  $F^*$  este obținută strict din date, o numim **repartiția empirică** a datelor. O vom numi de asemenea **repartiția de reselectare**. Totdeauna știm  $F^*$  explicit. În particular media lui  $F^*$  este chiar media de selecție  $\bar{x}$ .

## 2.5 Reselectarea

Bootstrap-ul empiric începe prin reselectarea datelor. Continuăm exemplul cu zaruri de mai sus.

**Exemplul 5.** Presupunem că avem 10 date, date în ordine crescătoare:

$$1, 1, 2, 3, 3, 3, 3, 4, 7, 7.$$

Vedem aceasta ca o **selecție** luată dintr-o repartiție subiacentă. A **reselecta** este a selecta cu înlocuire din repartiția empirică, de exemplu a pune aceste 10 numere într-o pălărie și a trage unul aleator. Apoi puneți numărul înapoi în pălărie și trageți iar. Trageți atâtea numere ca mărimea dorită a reselectiei. Pentru a ajunge un pic mai aproape de a implementa aceasta pe un calculator, reformulăm aceasta în modul următor. Notăm cele 10 date cu  $x_1, x_2, \dots, x_{10}$ . A **reselecta** este a trage un număr  $j$  din repartiția uniformă pe  $\{1, 2, \dots, 10\}$  și a lua  $x_j$  ca valoarea noastră reselectionată. În acest caz am putea face aceasta aruncând un zar cu 10 fețe. De exemplu, dacă aruncăm un 6, atunci valoarea noastră reselectată este 3, al 6-lea element din lista noastră.

Dacă vrem o mulțime de date reselectate de mărime 5, atunci vom arunca zarul cu 10 fețe de 5 ori și vom alege elementele corespunzătoare din lista

datelor. Dacă cele 5 aruncări sunt

$$5, 3, 6, 6, 1,$$

atunci reselecția este

$$3, 2, 3, 3, 1.$$

**Observații.** **1.** Deoarece selectăm cu înlocuire, aceeași dată poate apărea de mai multe ori când reselecțăm.

**2.** De asemenea, deoarece selectăm cu înlocuire, putem avea o mulțime de date reselectate de orice mărime vrem, de exemplu am putea reselecța de 1000 de ori.

Desigur, în practică se folosește un pachet software ca R pentru a face reselecțarea.

### 2.5.1 Notația cu steluță

Dacă avem o selecție de mărime  $n$

$$x_1, x_2, \dots, x_n,$$

atunci notăm o **reselecție de mărime  $m$**  adăugând o steluță simbolurilor

$$x_1^*, x_2^*, \dots, x_m^*.$$

Similar, la fel cum  $\bar{x}$  este media datelor originale, scriem  $\bar{x}^*$  pentru **media datelor reselecționate**.

## 2.6 Bootstrap-ul empiric

Presupunem că avem  $n$  date

$$x_1, x_2, \dots, x_n,$$

provenite dintr-o repartiție  $F$ . O **selecție bootstrap empirică** este o reselecție de **aceeași mărime  $n$** :

$$x_1^*, x_2^*, \dots, x_n^*.$$

Ar trebui să gândești ultima ca o selecție de mărime  $n$  extrasă din repartiția empirică  $F^*$ . Pentru orice statistică  $v$  calculată din datele selecției originale, putem defini o statistică  $v^*$  cu aceeași formulă, dar calculată folosind datele reselectate în locul celor originale. Cu această notație putem formula principiul bootstrap.

### 2.6.1 Principiul bootstrap

Schema bootstrap este după cum urmează:

1.  $x_1, x_2, \dots, x_n$  este o selecție de date provenite dintr-o repartiție  $F$ .
2. O statistică  $u$  este calculată din selecție.
3.  $F^*$  este repartiția empirică a datelor (repartiția de reselectare).
4.  $x_1^*, x_2^*, \dots, x_n^*$  este o reselecție a datelor **de aceeași mărime** ca selecția originală.
5.  $u^*$  este statistica calculată din reselecție.

Atunci **principiul bootstrap** spune că

1.  $F^* \approx F$ .
2. Variația lui  $u$  este bine aproximată de variația lui  $u^*$ .

Interesul nostru real este în punctul 2: putem aproxima variația lui  $u$  prin variația lui  $u^*$ . Vom exploata aceasta pentru a estima mărimea intervalor de încredere.

### 2.6.2 De ce reselecția are aceeași mărime ca selecția originală?

Aceasta este direct: variația statisticii  $u$  va depinde de mărimea selecției. Dacă vrem să aproximăm această variație trebuie să folosim reselecții de aceeași mărime.

### 2.6.3 Exemplu didactic de un interval de încredere bootstrap empiric

**Exemplul 6.** **Exemplu didactic.** Începem cu o mulțime inventată de date care este destul de mică pentru a arăta fiecare pas explicit. Datele selecției sunt

$$30, 37, 36, 43, 42, 43, 43, 46, 41, 42.$$

Problema: Estimați media  $\mu$  a repartiției subiacente și dați un interval de încredere bootstrap 80%.

**Răspuns.** Media de selecție este  $\bar{x}=40.3$ . Folosim aceasta ca o estimare a adevăratei medii  $\mu$  a repartiției subiacente. Ca în exemplul 1, pentru a face intervalul de încredere trebuie să stim cât de mult repartiția lui  $\bar{x}$  variază în jurul lui  $\mu$ . Adică, ne-ar plăcea să stim repartiția lui

$$\delta = \bar{x} - \mu.$$

Dacă am ști această repartiție, am putea afla  $\delta_{.1}$  și  $\delta_{.9}$ , valorile critice 0.1 și 0.9 ale lui  $\delta$ . Atunci am avea

$$P(\delta_{.9} \leq \bar{x} - \mu \leq \delta_{.1} | \mu) = 0.8 \iff P(\bar{x} - \delta_{.9} \geq \mu \geq \bar{x} - \delta_{.1} | \mu) = 0.8,$$

cea ce dă intervalul de încredere 80%

$$[\bar{x} - \delta_{.1}, \bar{x} - \delta_{.9}].$$

Ca întotdeauna la intervalele de încredere, ne-am grăbit să subliniem că probabilitățile calculate mai sus sunt probabilitățile privind statistica  $\bar{x}$  **dat fiind că adevărata medie este  $\mu$** .

Principiul bootstrap dă o abordare practică a estimării distribuției lui  $\delta = \bar{x} - \mu$ . Spune că o putem aproxima prin repartiția lui

$$\delta^* = \bar{x}^* - \bar{x},$$

unde  $\bar{x}^*$  este media unei selecții bootstrap empirice.

Iată cheia frumoasă pentru aceasta: deoarece  $\delta^*$  este calculată reselectând datele originale, putem simula pe calculator  $\delta^*$  de câte ori vrem. Prin urmare, din legea numerelor mari, putem estima repartiția lui  $\delta^*$  cu mare precizie.

Acum să ne întoarcem la datele de selecție cu 10 puncte. Am folosit R pentru a genera 20 de selecții bootstrap, fiecare de mărime 10. Fiecare dintre cele 20 de coloane ale următorului tabel este o selecție bootstrap.

43	36	46	30	43	43	43	37	42	42	43	37	36	42	43	43	42	43	42	43
43	41	37	37	43	43	46	36	41	43	43	42	41	43	46	36	43	43	43	42
42	43	37	43	46	37	36	41	36	43	41	36	37	30	46	46	42	36	36	43
37	42	43	41	41	42	36	42	42	43	42	43	41	43	36	43	43	41	42	46
42	36	43	43	42	37	42	42	42	46	30	43	36	43	43	42	37	36	42	30
36	36	42	42	36	36	43	41	30	42	37	43	41	41	43	43	42	46	43	37
43	37	41	43	41	42	43	46	46	36	43	42	43	30	41	46	43	46	30	43
41	42	30	42	37	43	43	42	43	43	46	43	30	42	30	42	30	43	43	42
46	42	42	43	41	42	30	37	30	42	43	42	43	37	37	42	43	43	46	
42	43	43	41	42	36	43	30	37	43	42	43	41	36	37	41	43	42	43	43

Apoi calculăm  $\delta^* = \bar{x}^* - \bar{x}$  pentru fiecare selecție bootstrap (i.e. fiecare coloană) și le sortăm de la cea mai mică la cea mai mare:

$$-1.6, -1.4, -1.4, -0.9, -0.5, -0.2, -0.1, 0.1, 0.1, 0.2, 0.2, 0.4, 0.4, 0.4, 0.7, 0.9, 1.1, 1.2, 1.2, 1.6, 1.6, 2.$$

Vom aproxima valorile critice  $\delta_{.1}$  și  $\delta_{.9}$  prin  $\delta_{.1}^*$  și  $\delta_{.9}^*$ . Deoarece  $\delta_{.1}^*$  este la a 90-a percentilă, alegem al 18-lea element din listă, i.e. 1.6. Analog, deoarece  $\delta_{.9}^*$  este la a 10-a percentilă, alegem al 2-lea element din listă, i.e. -1.4.

De aceea intervalul nostru de încredere 80% bootstrap pentru  $\mu$  este

$$[\bar{x} - \delta_{.1}, \bar{x} - \delta_{.9}] = [40.3 - 1.6, 40.3 + 1.4] = [38.7, 41.7].$$

În acest exemplu am generat doar 20 de selecții bootstrap pentru ca să încapă pe pagină. Folosind R, am putea genera cel puțin 10000 de selecții bootstrap pentru a obține o estimare foarte precisă pentru  $\delta_{.1}^*$  și  $\delta_{.9}^*$ .

## 2.6.4 Justificarea pentru principiul bootstrap

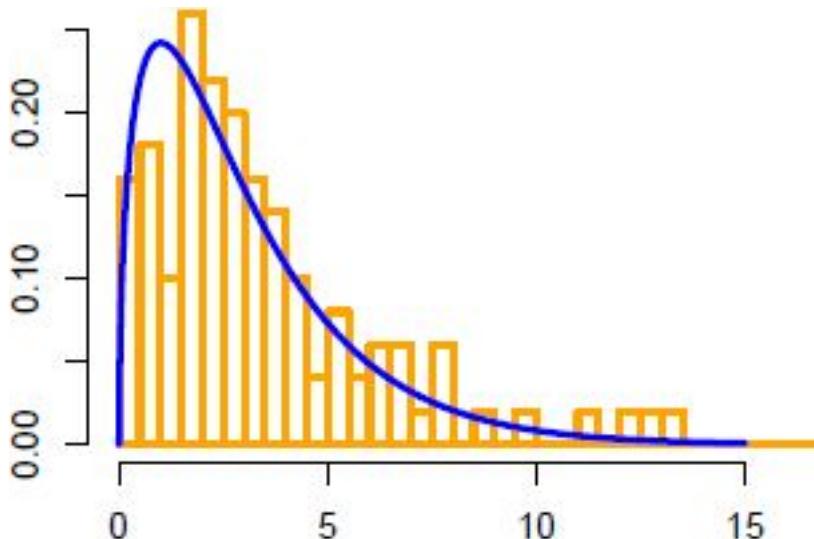
Bootstrap-ul este remarcabil pentru că reselectarea ne dă o estimare decentă a cum poate varia estimarea punctuală. Bootstrap-ul este bazat pe legea numerelor mari, care spune, pe scurt, că având destule date, repartiția empirică va fi o bună aproximare a adevărăratei repartiții. Vizual, spune că histograma datelor ar aproxima densitatea adevărăratei repartiții.

Reselectarea nu poate îmbunătăți estimarea noastră punctuală. De exemplu, dacă estimăm media  $\mu$  prin  $\bar{x}$ , atunci în bootstrap calculăm  $\bar{x}^*$  pentru multe reselectări ale datelor. Dacă luăm media tuturor  $\bar{x}^*$ , ne-am aștepta să fie foarte aproape de  $\bar{x}$ . Aceasta nu ne-ar spune nimic nou despre adevărata valoare a lui  $\mu$ .

Chiar cu o cantitate corectă de date, potrivirea dintre repartițiile adevărată și empirică nu este perfectă, deci va fi o eroare în estimarea mediei (sau oricarei alte valori). Dar cantitatea de variație a estimărilor este mult mai puțin sensibilă la diferențe între densitate și histogramă. Atât timp cât sunt rezonabil de aproape, atât repartiția adevărată, cât și cea empirică vor avea cantități similare de variație. Deci, în general principiul bootstrap este mai robust când aproximăm repartiția variației relative decât când aproximăm repartiții absolute.

Repartiția (peste diferite seturi de date experimentale) a lui  $\bar{x}$  este "centrată" în  $\mu$  și repartiția lui  $\bar{x}^*$  este centrată în  $\bar{x}$ . Dacă este o separare semnificativă între  $\bar{x}$  și  $\mu$ , atunci aceste 2 repartiții vor difera de asemenea semnificativ. Pe de altă parte, repartiția lui  $\delta = \bar{x} - \mu$  descrie variația lui  $\bar{x}$  în jurul centrului lui. Analog, repartiția lui  $\delta^* = \bar{x}^* - \bar{x}$  descrie variația lui  $\bar{x}^*$  în jurul centrului lui. Deci chiar dacă cele 2 centre sunt destul de diferite, cele 2 variații în jurul centrelor pot fi aproximativ egale.

Figura de mai jos ilustrează cum repartiția empirică aproximează adevărata repartiție. Pentru a face figura au fost generate 100 de valori aleatoare dintr-o repartiție  $\chi^2$  cu 3 grade de libertate. Figura arată pdf-ul adevărăratei repartiții ca o linie albastră și histograma repartiției empirice cu portocaliu.



Repartițiile adevărată și empirică sunt aproximativ egale.

## 2.7 Alte statistici

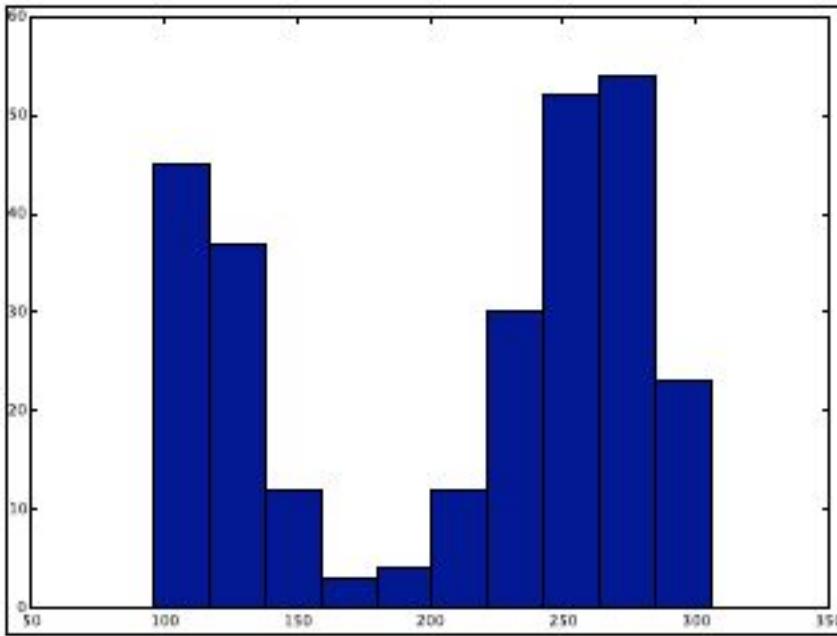
Până acum am evitat intervalle de încredere pentru mediană și alte statistici deoarece repartițiile lor de selecție sunt greu de descris teoretic. Bootstrap-ul nu are această problemă. De fapt, pentru a trata mediana, tot ce avem de făcut este să schimbăm "mean" în "median" în codul R de la exemplul 6.

### **Exemplul 7. Old Faithful: intervale de încredere pentru mediană**

Old Faithful este un gheizer în parcul național Yellowstone din Wyoming:

[http://en.wikipedia.org/wiki/Old\\_Faithful](http://en.wikipedia.org/wiki/Old_Faithful).

Există un set de date care dă duratele a 272 de erupții consecutive. Iată histograma datelor.



**Întrebare.** Estimați lungimea mediană a unei erupții și dați un interval de încredere de 90% pentru mediană.

**Răspuns.** Facem un rezumat al pașilor necesari pentru a răspunde întrebării.

1. Datele:  $x_1, \dots, x_{272}$ .
2. Mediana datelor:  $x_{\text{median}} = 240$ .
3. Află mediana  $x_{\text{median}}^*$  a unei selecții bootstrap  $x_1^*, \dots, x_{272}^*$ . Repetă de 1000 de ori.
4. Calculează diferențele bootstrap

$$\delta^* = x_{\text{median}}^* - x_{\text{median}}.$$

Pune aceste 1000 de valori în ordine crescătoare și alege valorile critice .95 și .05, i.e. a 50-a și a 950-a valoare. Numește aceste valori  $\delta_{.95}^*$  și  $\delta_{.05}^*$ .

5. Principiul bootstrap spune că putem folosi  $\delta_{.95}^*$  și  $\delta_{.05}^*$  ca estimări pentru  $\delta_{.95}$  și  $\delta_{.05}$ . Deci intervalul nostru de încredere bootstrap 90% pentru mediană este

$$[x_{\text{median}} - \delta_{.05}^*, x_{\text{median}} - \delta_{.95}^*].$$

CI (intervalul de încredere) 90% bootstrap aflat pentru datele Old Faithful a fost [235,250]. Deoarece s-au folosit 1000 de selecții bootstrap, o nouă simulare plecând de la aceleași date de selecție va produce un interval similar. Dacă în pasul 3 creștem numărul de selecții bootstrap la 10000, atunci intervalele produse prin simulare vor varia chiar mai puțin. O strategie uzuală este să creștem numărul de selecții bootstrap până simulările rezultate produc intervale care variază mai puțin decât un nivel acceptabil.

**Exemplul 8.** Folosind datele Old Faithful, estimați  $P(|\bar{x} - \mu| > 5|\mu|)$ .

Răspuns. Procedăm exact ca în exemplul precedent folosind media în locul medianei.

1. Datele:  $x_1, \dots, x_{272}$ .
2. Media datelor:  $\bar{x} = 209.27$ .
3. Află media  $\bar{x}^*$  a 1000 de selecții bootstrap empirice:  $x_1^*, \dots, x_{272}^*$ .
4. Calculează diferențele bootstrap

$$\delta^* = \bar{x}^* - \bar{x}.$$

5. Principiul bootstrap spune că putem folosi repartiția lui  $\delta^*$  ca o aproximație pentru repartiția lui  $\delta = \bar{x} - \mu$ . De aici,

$$P(|\bar{x} - \mu| > 5|\mu|) = P(|\delta| > 5|\mu|) \approx P(|\delta^*| > 5).$$

Simularea Bootstrap pentru datele Old Faithful a dat 0.225 pentru această probabilitate.

## 2.8 Bootstrap parametric

Exemplele din secțiunea anterioară au folosit toate bootstrap-ul empiric, care nu face nicio presupunere despre repartiția subiacentă și extrage selecții bootstrap prin reselectarea datelor. În această secțiune vom aborda **bootstrap-ul parametric**. Singura diferență dintre bootstrapul parametric și empiric este sursa selecției bootstrap. Pentru bootstrap-ul parametric, generăm selecția bootstrap dintr-o repartiție parametrizată.

Iată elementele folosirii bootstrap-ului parametric pentru a estima un interval de încredere pentru un parametru:

0. Date:  $x_1, \dots, x_n$  extrase dintr-o repartiție  $F(\theta)$  cu parametru necunoscut  $\theta$ .
1. O statistică  $\hat{\theta}$  care estimează pe  $\theta$ .
2. Selecțiile noastre bootstrap sunt extrase din  $F(\hat{\theta})$ .
3. Pentru fiecare selecție bootstrap

$$x_1^*, \dots, x_n^*$$

calculăm  $\hat{\theta}^*$  și diferența bootstrap  $\delta^* = \hat{\theta}^* - \hat{\theta}$ .

4. Principiul bootstrap spune că repartiția lui  $\delta^*$  aproximează repartiția lui  $\delta = \hat{\theta} - \theta$ .
5. Folosiți diferențele bootstrap pentru a face un interval de încredere bootstrap pentru  $\theta$ .

**Exemplul 9.** Presupunem că datele  $x_1, \dots, x_{300}$  sunt extrase dintr-o repartitie

$\exp(\lambda)$ . Presupunem de asemenea că media datelor  $\bar{x} = 2$ . Estimați  $\lambda$  și dați un interval de încredere bootstrap parametric 95% pentru  $\lambda$ .

**Răspuns.** Cel mai ușor este să explicăm soluția folosind codul R comentat.

```
# Bootstrap parametric
# Sunt 300 de date cu media 2.
# Presupunem că datele sunt exp(lambda).
# PROBLEMA: Calculați un interval de încredere bootstrap parametric
# 95% pentru lambda.
# Se dau numărul datelor și media.
n=300
xbar=2
# MLE pentru lambda este 1/xbar
lambdahat=1/xbar
# Generăm selecțiile bootstrap.
# Fiecare coloană este o selecție bootstrap (de 300 de valori reselecționate).
nboot=1000
# Iată diferența cheie față de bootstrap-ul empiric:
# Extragem selecția bootstrap din exponentiala(lambdahat).
x=rexp(n*nboot,lambdahat)
bootstrapsample=matrix(x,nrow=n,ncol=nboot)
# Calculăm lambdastar bootstrap.
lambdastar=1/colMeans(bootstrapsample)
# Calculăm diferențele.
deltastar=lambdastar-lambdahat
# Aflăm cuantilele 0.05 și 0.95 pentru deltastar.
d=quantile(deltastar, c(0.05,0.95))
# Calculăm intervalul de încredere 95% pentru lambda.
ci=lambdahat-c(d[2],d[1])
# Următoarele linii de cod sunt doar un mod de a forma textul de ieșire.
# R are și alte moduri de a face aceasta.
s=sprintf("Interval de incredere pentru lambda: [% .3f, % .3f]", ci[1],
ci[2])
cat(s)
```

## 2.9 Transcrieri R adnotate

### 2.9.1 Folosirea lui R pentru a genera un interval de încredere bootstrap empiric

Acest cod generează doar 20 de selecții bootstrap. În practica reală ar fi generate mult mai multe selecții bootstrap. Este făcut un interval de încredere pentru medie.

```
# Date pentru exemplul 6
x=c(30,37,36,43,42,43,43,46,41,42)
n=length(x)
# Media de selecție
xbar=mean(x)
nboot=20
# Generăm 20 de selecții bootstrap, i.e. o matrice n x 20 de reselecții aleatoare din x.
tmpdata=sample(x,n*nboot,replace=TRUE)
bootstrapsample=matrix(tmpdata,nrow=n,ncol=nboot)
# Calculăm mediile  $\bar{x}^*$ .
bsmeans=colMeans(bootstrapsample)
# Calculăm  $\delta^*$  pentru fiecare selecție bootstrap.
deltastar=bsmeans-xbar
# Sortăm rezultatele.
sorteddeltastar=sort(deltastar)
# Aflăm valorile critice .1 și .9 ale lui deltastar.
d9=sorteddeltastar[2]
d1=sorteddeltastar[18]
# Calculăm intervalul de încredere 80% pentru medie.
ci=xbar-c(d1,d9)
cat('Intervalul de incredere: ',ci, '\n')
```

# Curs 20

Cristian Niculescu

## 1 Regresia liniară

### 1.1 Scopurile învățării

1. Să poată folosi metoda celor mai mici pătrate pentru a potrivi o dreaptă cu date bivariate.
2. Să poată da o formulă pentru eroarea pătratică totală la potrivirea oricărui tip de curbă cu datele.
3. Să poată spune cuvintele homoscedasticitate și heteroscedasticitate.

### 1.2 Introducere

Presupunem că avem colectate date bivariate  $(x_i, y_i), i = 1, \dots, n$ . Scopul regresiei liniare este modelarea relației dintre  $x$  și  $y$  prin aflarea unei funcții  $y = f(x)$  care dă o potrivire apropiată cu datele. Presupunerile modelării pe care le vom folosi sunt că  $x_i$  nu sunt aleatoare și că  $y_i$  este o funcție de  $x_i$  plus un zgomot aleator. Cu aceste presupuneri,  $x$  este numită **variabilă independentă** sau **predictor** și  $y$  este numită **variabilă dependentă** sau **răspuns**.

**Exemplul 1.** Costul unui timbru de clasa întâi în dolari de-a lungul timpului este dat în lista următoare:

.05 (1963)	.06 (1968)	.08 (1971)	.10 (1974)	.13 (1975)	.15 (1978)	.20 (1981)	.22 (1985)
.25 (1988)	.29 (1991)	.32 (1995)	.33 (1999)	.34 (2001)	.37 (2002)	.39 (2006)	.41 (2007)
.42 (2008)	.44 (2009)	.45 (2012)	.46 (2013)	.49 (2014)			

Folosind codul R:

```
x=c(3,8,11,14,15,18,21,25,28,31,35,39,41,42,46,47,48,49,52,53,54)
```

```
y=c(5,6,8,10,13,15,20,22,25,29,32,33,34,37,39,41,42,44,45,46,49)
```

```
lm(y~x),
```

obținem:

Call:

```
lm(formula = y ~ x)
```

Coefficients:

```
(Intercept) x
```

-0.1324 0.8791.

Am aflat că dreapta care dă ”potrivirea celor mai mici pătrate” cu aceste date (dreapta de regresie) este

$$y = -0.1324 + 0.8791x,$$

unde  $x$  este numărul de ani de la 1960, iar  $y$  este în cenți.

Folosind acest rezultat ”rezem” că în 2021 ( $x = 61$ ), costul unui timbru va fi 53 de cenți (deoarece  $-0.1324 + 0.8791 \cdot 61 = 53.4927$ ).

Folosind codul R (în continuarea celui de mai sus):

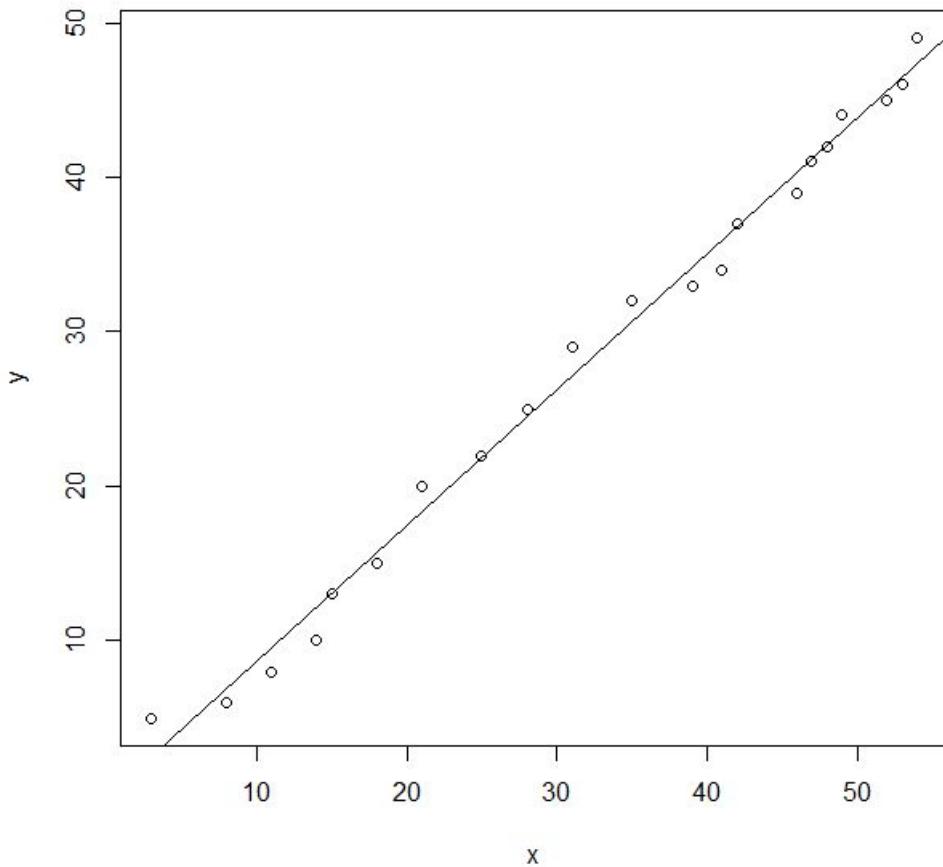
a=-0.1324

b=0.8791

plot(x,y)

abline(a,b)

obținem:



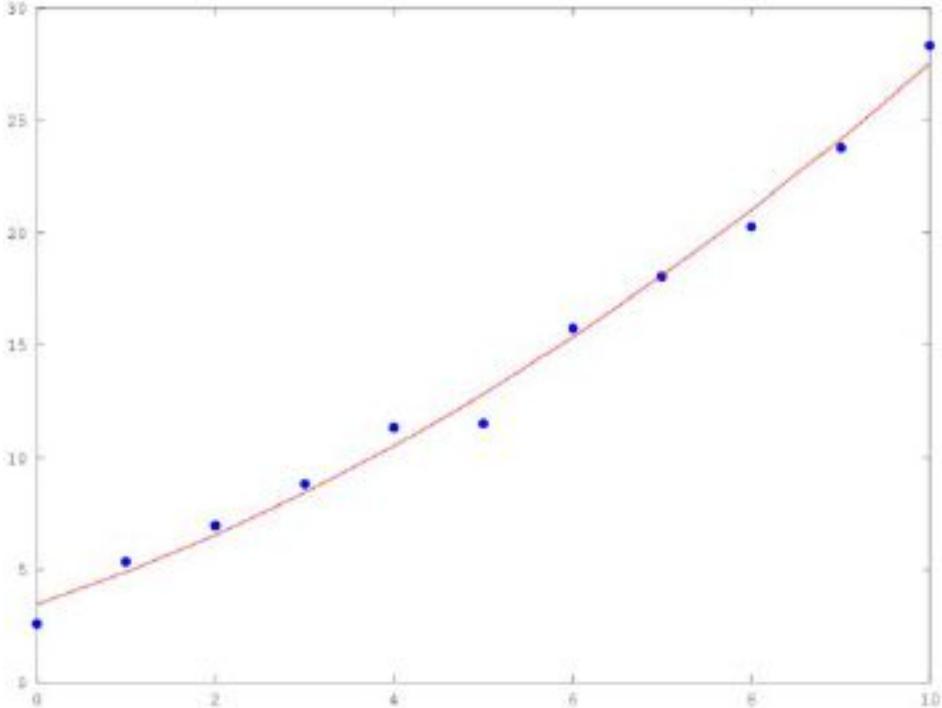
Costul timbrului (cenți) vs. timp (ani din 1960)

Niciuna din date nu se află chiar pe dreaptă. Mai degrabă această dreaptă are "cea mai bună potrivire" în raport cu [toate datele](#), cu o mică eroare pentru fiecare dată.

**Exemplul 2.** Presupunem că avem  $n$  perechi de tată și fi adulții. Fie  $x_i$  și  $y_i$  înălțimile celui de-al  $i$ -lea tată, respectiv fiu. Dreapta celor mai mici pătrate

pentru aceste date poate fi folosită pentru a prezice înălțimea de adult a unui băiat Tânăr din cea a tatălui lui.

**Exemplul 3.** Nu suntem limitați la drepte cu cea mai bună potrivire.  $\forall d \in \mathbb{N}^*$ , metoda celor mai mici pătrate poate fi folosită pentru a afla un polinom de grad  $d$  cu "cea mai bună potrivire cu datele". Iată o figură arătând potrivirea datelor cu o parabolă prin metoda celor mai mici pătrate:



Potrivirea unei parbole,  $y = b_2x^2 + b_1x + b_0$  cu datele

### 1.3 Potrivirea unei drepte folosind cele mai mici pătrate

Presupunem că avem datele  $(x_i, y_i)$  ca mai sus. Scopul este să aflăm dreapta

$$y = \beta_1 x + \beta_0,$$

care "se potrivește cel mai bine" cu datele. Modelul nostru spune că fiecare  $y_i$  este prezis de  $x_i$  până la o eroare  $\epsilon_i$ :

$$y_i = \beta_1 x_i + \beta_0 + \epsilon_i.$$

Deci,

$$\epsilon_i = y_i - \beta_1 x_i - \beta_0.$$

Metoda celor mai mici pătrate află valorile  $\hat{\beta}_0$  și  $\hat{\beta}_1$  ale lui  $\beta_0$  și  $\beta_1$  care minimizează suma pătratelor erorilor:

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)^2.$$

Folosind analiza matematică (detalii în adaos), aflăm

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad (1)$$

unde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_{xx} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

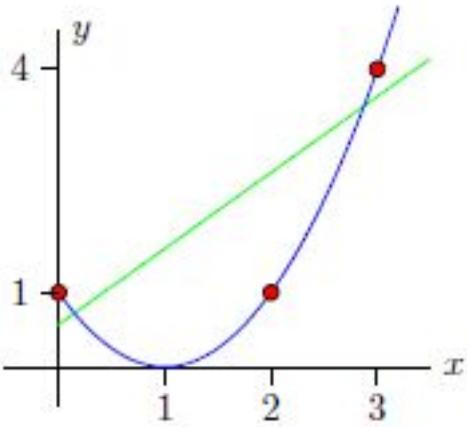
Aici,  $\bar{x}$  este media de selecție a lui  $x$ ,  $\bar{y}$  este media de selecție a lui  $y$ ,  $s_{xx}$  este dispersia de selecție a lui  $x$  și  $s_{xy}$  este covarianța de selecție a lui  $x$  și  $y$ .

**Exemplul 4.** Folosiți cele mai mici pătrate pentru a potrivi cu o dreaptă următoarele date: (0,1), (2,1), (3,4).

**Răspuns.** În cazul nostru,  $(x_1, y_1) = (0, 1)$ ,  $(x_2, y_2) = (2, 1)$  și  $(x_3, y_3) = (3, 4)$ . Deci

$$\begin{aligned} \bar{x} &= \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}(0 + 2 + 3) = \frac{5}{3}, \\ \bar{y} &= \frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}(1 + 1 + 4) = \frac{6}{3} = 2, \\ s_{xx} &= \frac{1}{3-1} \sum_{i=1}^3 (x_i - \bar{x})^2 = \frac{1}{2} \left[ \left(0 - \frac{5}{3}\right)^2 + \left(2 - \frac{5}{3}\right)^2 + \left(3 - \frac{5}{3}\right)^2 \right] = \frac{7}{3}, \\ s_{xy} &= \frac{1}{2} \left[ \left(0 - \frac{5}{3}\right)(1 - 2) + \left(2 - \frac{5}{3}\right)(1 - 2) + \left(3 - \frac{5}{3}\right)(4 - 2) \right] = 2; \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} = \frac{2}{\frac{7}{3}} = \frac{6}{7}; \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 2 - \frac{6}{7} \cdot \frac{5}{3} = 2 - \frac{10}{7} = \frac{4}{7}. \end{aligned}$$

Deci dreapta de regresie a celor mai mici pătrate are ecuația  $y = \frac{4}{7} + \frac{6}{7}x$ . Aceasta este arătată ca dreapta verde din figura următoare.



Potrivirea celor mai mici pătrate a unei drepte (verde) și a unei parbole (albastru)

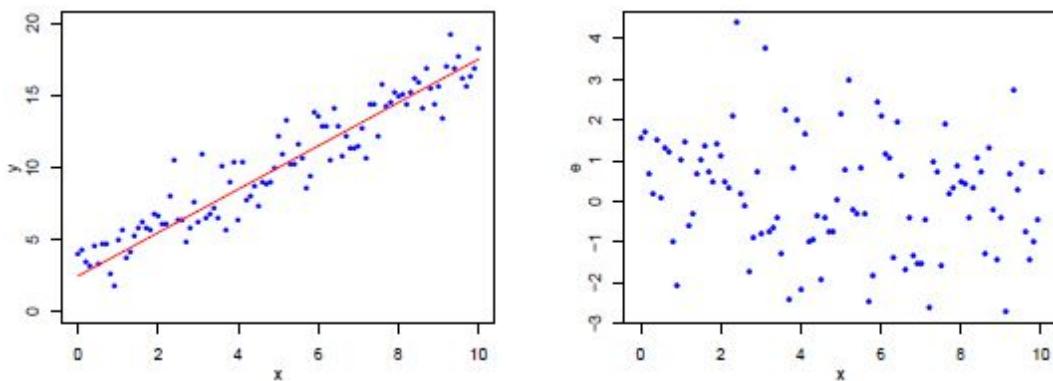
**Regresie liniară simplă:** Este puțin confuz, dar cuvântul "liniară" din "regresie liniară" nu se referă la potrivirea unei drepte. Totuși, cea mai uzuală curbă pentru potrivire este o dreaptă. Potrivirea unei drepte la date bivariate este numită [regresie liniară simplă](#).

### 1.3.1 Reziduuri

Pentru o dreaptă, modelul este

$$y_i = \hat{\beta}_1 x_i + \hat{\beta}_0 + \epsilon_i.$$

Gândim  $\hat{\beta}_1 x_i + \hat{\beta}_0$  ca prezicând sau explicând  $y_i$ . Termenul rămas  $\epsilon_i$  este numit [reziduul](#), pe care-l gândim ca pe un zgomot aleator sau o eroare de măsurare. O verificare vizuală folositoare a modelului de regresie liniară este reprezentarea reziduurilor. Datele ar trebui să fie lângă dreapta de regresie. Reziduurile ar trebui să arate cum de-a lungul domeniului lui  $x$ .

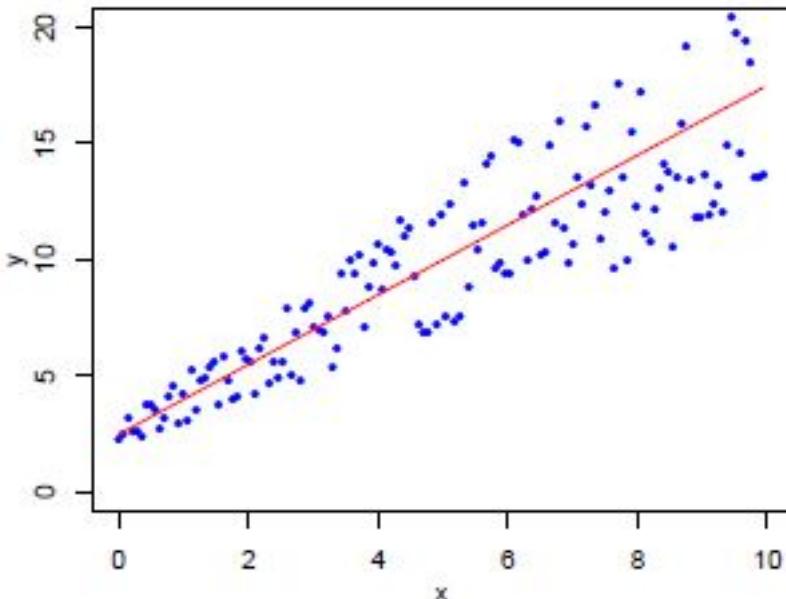


Date cu dreapta de regresie (stânga) și reziduuri (dreapta). Observați homoscedasticitatea.

### 1.3.2 Homoscedasticitatea

O presupunere importantă a modelului de regresie liniară este că reziduurile  $\epsilon_i$  au aceeași dispersie  $\forall i$ . Această presupunere este numită [homoscedasticitate](#). Puteți vedea aceasta în cazul ambelor figuri de mai sus. Datele sunt în banda de lățime fixă în jurul dreptei de regresie și la fiecare  $x$  reziduurile au cam aceeași împrăștiere verticală.

Mai jos este o figură arătând date [heteroscedastice](#). Împrăștierea verticală a datelor crește când  $x$  crește. Înainte de a folosi cele mai mici pătrate pe aceste date ar trebui să transformăm datele pentru a fi homoscedastice.



Date heteroscedastice

## 1.4 Regresie liniară pentru potrivirea polinoamelor

Potrivirea unei drepte la date este numită [regresie liniară simplă](#). Putem de asemenea folosi regresia liniară pentru a potrivi polinoame cu datele. Folosirea cuvântului "liniară" în ambele cazuri poate părea confuză. Aceasta este deoarece cuvântul "liniară" din "regresia liniară" nu se referă la potrivirea unei drepte. Mai degrabă se referă la ecuațiile algebrice liniare pentru parametrii necunoscuți  $\beta_i$ , i.e. fiecare  $\beta_i$  are exponentul 1.

**Exemplul 5.** Luați aceleași date ca în exemplul 4 și folosiți cele mai mici pătrate pentru a afla parabola cu cea mai bună potrivire pentru date.

**Răspuns.** O parabolă are formula  $y = \beta_0 + \beta_1x + \beta_2x^2$ . Eroarea pătratică este

$$S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^3 (y_i - (\beta_0 + \beta_1x_i + \beta_2x_i^2))^2.$$

După substituirea valorilor date pentru  $x_i$  și  $y_i$  putem folosi analiza matematică (egalăm derivatele parțiale în raport cu  $\beta_0, \beta_1, \beta_2$  cu 0, obținând un sistem de 3 ecuații liniare cu 3 necunoscute) pentru a afla tripletul  $(\beta_0, \beta_1, \beta_2)$  care minimizează  $S$ . Sau putem folosi codul R

```
x=c(0,2,3)
y=c(1,1,4)
C=cbind(1,x,x^ 2)
solve(t(C)%*%C,t(C)%*%y).
```

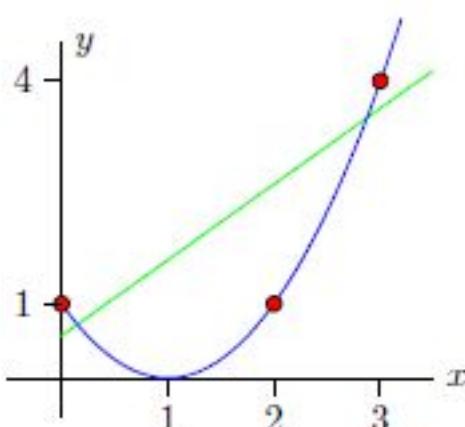
Sau codul R

```
x=c(0,2,3)
y=c(1,1,4)
x1=x
x2=x^ 2
lm(y~x1+x2)
```

Cu aceste date, parabola celor mai mici pătrate are ecuația

$$y = 1 - 2x + x^2.$$

Pentru 3 puncte, potrivirea pătratică este perfectă.



Potrivirea celor mai mici pătrate a unei drepte (verde) și a unei parbole (albastru)

**Exemplul 6.** Perechile  $(x_i, y_i)$  pot da vârsta și mărimea vocabularului a  $n$  copii. Deoarece copiii mici dobândesc cuvinte noi într-un ritm accelerat, putem ghici că un polinom de grad mai mare poate fi cea mai bună potrivire

pentru date.

**Exemplul 7.** (Transformarea datelor). Uneori este necesar să transformăm datele înainte de a folosi regresia liniară. De exemplu, presupunem că relația este exponentială, i.e.  $y = ce^{ax}$ . Atunci

$$\ln(y) = ax + \ln(c).$$

Deci putem folosi regresia liniară simplă pentru a obține un model

$$\ln(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

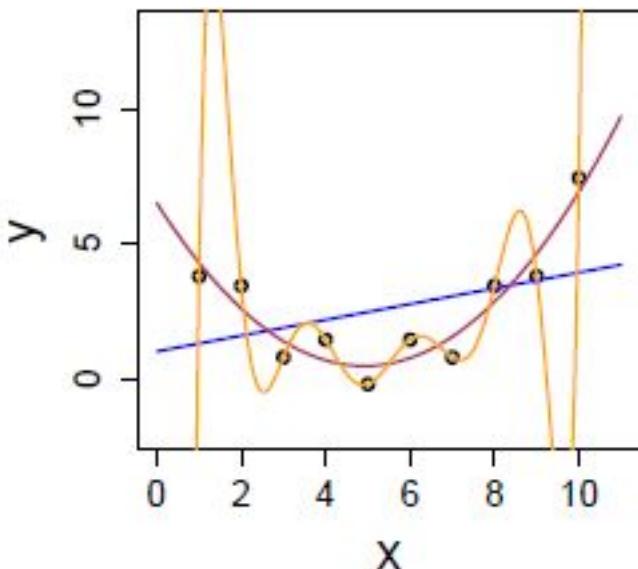
și apoi obținem modelul exponential

$$y_i = e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_i}.$$

#### 1.4.1 Suprapotrivirea

Putem totdeauna obține o potrivire mai bună folosind un polinom de ordin mai mare. De exemplu, date fiind 6 date bivariate (cu  $x_i$  distințe) se poate totdeauna afla un polinom de grad 5 care trece prin toate. Aceasta poate duce la [suprapotrivire](#). Adică, potrivirea zgromotului la fel de bine ca adevărata relație între  $x$  și  $y$ . Un model suprapotrivit va potrivi datele originale mai bine, dar va prezice mai puțin bine  $y$  pentru noi valori ale lui  $x$ . O povocare a modelării statistice este echilibrarea potrivirii modelului cu complexitatea modelului.

**Exemplul 8.** În reprezentarea de mai jos potrivim polinoame de gradul 1, 2 și 9 la 10 date bivariate. Modelul de gradul 2 (maro) dă o potrivire semnificativ mai bună decât modelul de gradul 1 (albastru). Modelul de gradul 9 (portocaliu) dă o potrivire exactă cu datele, dar dintr-o privire am ghici că este suprapotrivit. Adică, nu ne așteptăm că potrivească bine următoarea dată bivariată pe care o vedem. De fapt, datele au fost generate folosind un model pătratic, deci modelul de gradul 2 va tinde să facă cea mai bună potrivire cu date noi.



#### 1.4.2 Funcția R lm

Nu facem regresia liniară cu mâna. Regresia liniară se reduce la rezolvarea sistemelor de ecuații liniare, i.e. la calcul matriceal. Funcția R `lm` poate fi folosită la potrivirea unui polinom de orice grad cu datele. ([lm înseamnă model liniar](#)). De fapt, `lm` poate potrivi multe tipuri de funcții, exceptând polinoamele, după cum puteți explora folosind ajutorul lui R sau google.

## 1.5 Regresie liniară multiplă

Datele nu sunt totdeauna bivariate. Pot fi trivariate sau chiar de o dimensiune mai mare. Presupunem că avem datele de forma

$$(y_i, x_{1i}, x_{2i}, \dots, x_{mi}).$$

Putem analiza aceste date într-o manieră foarte similară cu datele bivariate. Adică, putem folosi cele mai mici pătrate pentru a potrivi modelul

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m.$$

Aici fiecare  $x_j$  este o variabilă predictor și  $y$  este variabila de răspuns. De exemplu, putem fi interesați de cum variază o populație de pești în funcție nivelele măsurate ale câtorva poluanți, sau am vrea să prezicem înălțimea de adult a unui fiu pe baza înălțimilor tatălui și a mamei.

## 1.6 Cele mai mici pătrate ca un model statistic

Modelul de regresie liniară pentru potrivirea unei drepte spune că valoarea  $y_i$  din perechea  $(x_i, y_i)$  este extrasă dintr-o variabilă aleatoare

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

unde termenii de "eroare"  $\varepsilon_i$  sunt variabile aleatoare independente cu media 0 și deviația standard  $\sigma$ . Presupunerea standard este că  $\varepsilon_i$  sunt i.i.d. cu repartiția  $N(0, \sigma^2)$ . În orice caz, media lui  $Y_i$  este dată de:

$$E(Y_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i.$$

Din această perspectivă, metoda celor mai mici pătrate alege valorile lui  $\beta_0$  și  $\beta_1$  care minimizează dispersia de selecție din jurul dreptei.

De fapt, estimarea celor mai mici pătrate  $(\hat{\beta}_0, \hat{\beta}_1)$  coincide cu estimarea de verosimilitate maximă pentru parametrii  $(\beta_0, \beta_1)$ ; adică, dintre toți coeficienții posibili,  $(\hat{\beta}_0, \hat{\beta}_1)$  sunt cei care fac datele observate cele mai probabile.

## 1.7 Regresia la medie

Motivul termenului "regresie" este că variabila de răspuns prezisă  $y$  va tinde să fie "mai aproape" de (i.e. să regreseze la) media ei decât variabila predictor  $x$  este față de media ei. Aici "mai aproape" este în ghilimele deoarece trebuie să controlăm scala (i.e. deviația standard) fiecărei variabile. Modul de a controla scala este mai întâi să standardizăm fiecare variabilă.

$$u_i = \frac{x_i - \bar{x}}{\sqrt{s_{xx}}}, \quad v_i = \frac{y_i - \bar{y}}{\sqrt{s_{yy}}}.$$

Standardizarea schimbă media în 0 și dispersia în 1:

$$\bar{u} = \bar{v} = 0, \quad s_{uu} = s_{vv} = 1.$$

Proprietățile algebrice ale covarianței arată că

$$s_{uv} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \rho,$$

coeficientul de corelație. Astfel, potrivirea celor mai mici pătrate pentru  $v = \beta_0 + \beta_1 u$  are

$$\hat{\beta}_1 = \frac{s_{uv}}{s_{uu}} = \rho \text{ și } \hat{\beta}_0 = \bar{v} - \hat{\beta}_1 \bar{u} = 0.$$

Deci dreapta celor mai mici pătrate este  $v = \rho u$ . Deoarece  $\rho$  este coeficientul de corelație, el este între -1 și 1. Presupunem că este pozitiv și mai mic ca 1 (i.e.,  $x$  și  $y$  sunt pozitiv, dar nu perfect corelate). Atunci formula  $v = \rho u$  înseamnă că, dacă  $u$  este pozitiv, atunci valoarea prezisă a lui  $v$  este mai mică decât  $u$ . Adică,  $v$  este mai aproape de 0 decât  $u$ . Echivalemt,

$$\frac{y - \bar{y}}{\sqrt{s_{yy}}} < \frac{x - \bar{x}}{\sqrt{s_{xx}}},$$

i.e.,  $y$  regresează la  $\bar{y}$ . Standardizarea are grija de scală.

Considerăm cazul extrem al corelației 0 între  $x$  și  $y$ . Atunci, indiferent de valoarea lui  $x$ , valoarea prezisă a lui  $y$  este totdeauna  $\bar{y}$ . Adică,  $y$  a regresat până la media lui.

Dreapta de regresie trece totdeauna prin punctul  $(\bar{x}, \bar{y})$ .

**Exemplul 9.** Regresia la medie este importantă în studiile longitudinale. Rice (*Mathematical Statistics and Data Analysis*) dă următoarul exemplu. Presupunând că li se dau copiilor un test IQ la vârsta de 4 ani și altul la vârsta de 5 ani, ne așteptăm ca rezultatele să fie pozitiv corelate. Analiza de mai sus spune că, în medie, acei copii care au făcut slab la primul test tind să arate îmbunătățire (i.e. regresează la medie) în al 2-lea test. Astfel, o intervenție inutilă poate fi interpretată greșit ca utilă deoarece pare a îmbunătăți scorurile.

**Exemplul 10.** Alt exemplu cu consecințe practice este recompensa și pedeapsa. Imagineați-vă o școală unde performanța înaltă la un examen este recompensată și performanța slabă este pedepsită. Regresia la medie ne spune că (în medie) studenții foarte performanți vor face puțin mai slab la următorul examen și studenții puțin performanți vor face puțin mai bine. O viziune nefisticată a datelor va face să pară că pedeapsa a îmbunătățit performanța și recompensa de fapt a scăzut performanța. Sunt reale consecințe dacă cei cu autoritate acționează după această idee.

## 1.8 Adaos

### 1.8.1 Demonstrația formulei pentru potrivirea celor mai mici pătrate a unei drepte

Cea mai directă demonstrație este cu analiza matematică. Suma erorilor pătrate este

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)^2.$$

Luând derivatele parțiale (și reamintind că  $x_i$  și  $y_i$  sunt date, deci constante)

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} &= \sum_{i=1}^n -2(y_i - \beta_1 x_i - \beta_0) = 0 \\ \frac{\partial S}{\partial \beta_1} &= \sum_{i=1}^n -2x_i(y_i - \beta_1 x_i - \beta_0) = 0.\end{aligned}$$

Se obține următorul sistem de 2 ecuații liniare în necunoscutele  $\beta_0$  și  $\beta_1$ :

$$\begin{aligned}\left(\sum_{i=1}^n x_i\right)\beta_1 + n\beta_0 &= \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i^2\right)\beta_1 + \left(\sum_{i=1}^n x_i\right)\beta_0 &= \sum_{i=1}^n x_i y_i\end{aligned}$$

Rezolvând sistemul obținem formulele (1).

Pentru multe aplicații între discipline vezi:

[http://en.wikipedia.org/wiki/Linear\\_regression#Applications\\_of\\_linear\\_regression](http://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression).

### 1.8.2 Măsurarea potrivirii

Odată ce se calculează coeficienții de regresie, este important să verificăm cât de bine modelul de regresie se potrivește cu datele (i.e., cât de aproape cea mai potrivită dreaptă urmărește datele). O măsură ușuală dar brută a ”bunătății de potrivire” este **coeficientul de determinare**, notat  $R^2$ . Suma totală a pătratelor este dată de:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2.$$

**Suma reziduală a pătratelor** este dată de suma pătratelor reziduurilor. Când potrivim o dreaptă, aceasta este:

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

RSS este portiunea ”neexplicată” a sumei totale a pătratelor, i.e. neexplicată de ecuația de regresie. Diferența TSS–RSS este portiunea ”explicată” a sumei totale a pătratelor. **Coefficientul de determinare**  $R^2$  este raportul dintre portiunea ”explicată” și suma totală a pătratelor:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}.$$

Cu alte cuvinte,  $R^2$  măsoară proporția variabilității datelor care este contabilizată pentru modelul de regresie. O valoare aproape de 1 indică o potrivire bună, în timp ce o valoare aproape de 0 indică o potrivire slabă. În cazul regresiei liniare simple,  $R^2$  este pur și simplu pătratul coeficientului de corelație dintre valorile observate  $y_i$  și valorile prezise  $\hat{y}_i = \beta_0 + \beta_1 x_i$ .

**Exemplul 11.** În exemplul 8 de suprapotrivire, valorile lui  $R^2$  sunt ("degree" = "grad"):

degree	$R^2$
1	0.3968
2	0.9455
9	1.0000

Măsura bunătății potrivirii crește când  $n$  (gradul) crește. Potrivirea este mai bună, dar modelul devine de asemenea mai complex, deoarece este nevoie de mai mulți coeficienți pentru a descrie polinoame de grad mai mare.