Curs 6

Cristian Niculescu

1 Probleme recapitulative

1.1 Numărare şi probabilităţi

1) a) În câte moduri putem aranja literele cuvântului STATISTICS? (De exemplu, SSSTTTIIAC este un aranjament).
b) Dacă toate aranjamentele sunt egal posibile, care este probabilitatea ca "I"-urile să fie unul după altul?

1.2 Probabilitate condiţionată şi teorema lui Bayes

2) Corupt de puterea lui, juriul showului tv *Următorul matematician de top al Americii* a luat mită de la unii participanţi. 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 participanţi au mituit juriul. Aceiaşi participanţi mituiesc juriul în ambele runde, i.e., dacă un participant dă mită în prima rundă, dă mită şi în a 2-a (şi 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?

1.3 Independenţă

3) Aruncăm un zar corect cu 20 de feţe, numerotate de la 1 la 20. Determinaţi dacă următoarele perechi de evenimente sunt independente.

a) "Număr par" şi "Număr ≤ 10".
b) "Număr par" şi "Număr prim".

1.4 Medie şi dispersie

4) Fie $X \sim \begin{pmatrix} -1 & 0 & 1 \\ 1/8 & 1/4 & 5/8 \end{pmatrix}$.
a) Calculaţi $E(X)$.
b) Daţi pmf pentru $Y = X^2$ şi folosiţi-o pentru a calcula $E(Y)$.
c) În loc de aceasta, calculaţi $E(X^2)$ direct dintr-un tabel extins.
d) Calculaţi Var(X).
5) Calculaţi media şi dispersia unei variabile aleatoare Bernoulli(p).
6) Presupunem că 100 de oameni îşi aruncă pălăria într-o cutie şi apoi aleg aleator câte o pălărie din cutie. Care este media numărului de oameni care îşi iau înapoi pălăria proprie?
Indicaţie: Exprimaţi numărul de oameni care îşi iau înapoi propria pălărie ca o sumă de variabile aleatoare a căror medie este uşor de calculat.

1.5 Funcţii masă de probabilitate, funcţii densitate de probabilitate şi funcţii de repartiţie

7) a) Presupunem că X are funcţia densitate de probabilitate $f_X(x) = \lambda e^{-\lambda x}$ pentru $x \geq 0$. Calculaţi cdf, $F_X(x)$.
b) Dacă $Y = X^2$, calculaţi pdf şi cdf ale lui Y.
8) Presupunem că aruncăm un zar corect de 100 de ori (independent) şi primim 3 dolari de fiecare dată când dăm 6. Fie X_1, numărul de dolari primiţi în aruncările de la 1 la 25.
Fie X_2, numărul de dolari primiţi în aruncările de la 26 la 50.
Fie X_3, numărul de dolari primiţi în aruncările de la 51 la 75.
Fie X_4, numărul de dolari primiţi în aruncările de la 76 la 100.
Fie $X = X_1 + X_2 + X_3 + X_4$, numărul de dolari primiţi în 100 de aruncări.
a) Care este funcţia masă de probabilitate a lui X?
b) Care este media şi dispersia lui X?
c) Fie $Y = 4X_1$. (Deci, în loc să aruncăm de 100 de ori, aruncăm doar de 25 de ori şi înmulţim câştigurile cu 4.)
i) Care sunt media şi dispersia lui Y?
ii) Cum sunt media şi dispersia lui Y în comparaţie cu ale lui X? (I.e., sunt mai mari, mai mici, sau egale?) Explicaţi pe scurt de ce aceasta are sens.

1.6 Probabilitate comună, covarianţă, corelaţie

9) (**Puzzle aritmetic**) Pmf-urile comune şi marginale ale lui X şi Y sunt parţial date în următoarul tabel.

$X \backslash Y$	1	2	3	
1	1/6	0	...	1/3
2	...	1/4	...	1/3
3	...	...	1/4	...
	1/6	1/3	...	1

a) Completaţi tabelul.
b) Sunt X şi Y independente?
10) **Covarianţă şi independenţă**
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) Completaţi următorul tabel dând funcţia de frecvenţă comună pentru X şi Y. Includeţi probabilităţile marginale.

X / Y	-2	-1	0	1	2	total
0						
1						
4						
total						

b) Aflaţi $E(X)$ şi $E(Y)$.
c) Arătaţi că X şi Y nu sunt independente.
d) Arătaţi $Cov(X, Y) = 0$.
Acesta este un exemplu de variabile aleatoare necorelate dar dependente. Motivul pentru care aceasta se poate întâmpla este că doar dependenţa liniară este măsurată de corelaţie. În acest caz, X şi Y nu sunt legate liniar.
11) **Repartiţii comune continue.** Presupunem că X şi Y sunt variabile aleatoare continue cu funcţia de densitate comună $f(x, y) = x+y$ pe pătratul unitate $[0, 1] \times [0, 1]$.
a) Fie $F(x, y)$ cdf comună. Calculaţi $F(1, 1)$. Calculaţi $F(x, y)$.
b) Calculaţi densităţile marginale pentru X şi Y.
c) Sunt X şi Y independente?
d) Calculaţi $E(X), E(Y), E(X^2 + Y^2), Cov(X, Y)$.

1.7 Legea numerelor mari, teorema limită centrală

12) Presupunem că $X_1, ..., X_{100}$ sunt i.i.d. cu media 1/5 şi dispersia 1/9. Utilizaţi teorema limită centrală pentru a estima $P\left(\sum X_i < 30\right)$.
13) (Mai mult cu teorema limită centrală)
IQ-ul mediu al unei populaţii este 100 cu deviaţia standard 15 (prin definiţie, IQ-ul este normalizat, deci aşa este şi aici). Care este probabilitatea ca un grup de 100 de oameni selectat aleator să aibă un IQ mediu peste 115?

Tabel standard normal al probabilităţilor cozii stângi.
$\Phi(z) = P(Z \leq z)$ pentru $N(0,1)$. *(Folosiţi interpolarea pentru a estima valorile 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

2 Soluţiile problemelor recapitulative

2.1 Numărare şi probabilitate

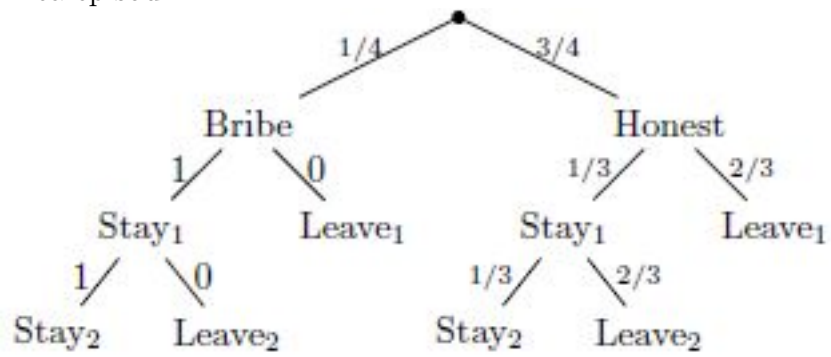
1) a) Creăm un aranjament în etape şi calculăm numărul de posibilităţi la fiecare etapă:
Etapa 1: alegem 3 din cele 10 locuri pentru a pune S-urile: C_{10}^3.
Etapa 2: alegem 3 din cele 7 locuri rămase pentru a pune T-urile: C_7^3.
Etapa 3: alegem 2 din cele 4 locuri rămase pentru a pune I-urile: C_4^2.
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 poziţiile 1 şi 2, poziţiile 2 şi 3,..., poziţiile 9 şi 10.
Deci probabilitatea ca I-urile să fie adiacente este $9/45 = 1/5 = 0.2$.

2.2 Probabilitate condiţionată şi teorema lui Bayes

2) Facem un arbore. Stay_1 înseamnă că participantului i s-a permis să rămână în joc după primul episod şi Stay_2 că i s-a permis să rămână în joc după al 2-lea episod.



Evenimentele relevante sunt:
B = "participantul mituieşte juriul"
H = "participantul este cinstit (nu mituieşte 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 probabilităţii 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) $\Longrightarrow$ $P(S_1) = 1/2$. 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 $P(L_2|S_1) = \frac{1/6}{1/2} = \frac{1}{3}$.

2.3 Independenţă

3) E = "număr par" $= \{2, 4, ..., 20\}$.
L = "număr ≤ 10" $= \{1, 2, ..., 10\}$.
B = "număr prim" $= \{2, 3, 5, 7, 11, 13, 17, 19\}$.
a) $P(E) = 10/20 = 1/2,\ P(E|L) = 5/10 = 1/2 \Longrightarrow P(E) = P(E|L) \Longrightarrow$ E şi L sunt independente.
b) $P(E) = 10/20 = 1/2,\ P(E|B) = 1/8 \Longrightarrow P(E) \neq P(E|B) \Longrightarrow$ E şi B nu sunt independente.

2.4 Media şi 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
probabilităţile	$1/8$	$1/4$	$5/8$
X^2	1	0	1

$Y \sim \begin{pmatrix} 0 & 1 \\ 1/4 & 3/4 \end{pmatrix} \Longrightarrow E(Y) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{3}{4} = \frac{3}{4}$.
c) Formula schimbării de variabile spune să folosim ultima linie a tabelului extins de la b): $E\left(X^2\right) = 1 \cdot \frac{1}{8} + 0 \cdot \frac{1}{4} + 1 \cdot \frac{5}{8} = \frac{3}{4}$ (la fel ca la b)).
d) $Var(X) = E\left(X^2\right) - E(X)^2 = \frac{3}{4} - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$.
5) Facem un tabel:

X	0	1
probabilităţile	$1-p$	p
X^2	0	1

Din tabel, $E(X) = 0 \cdot (1-p) + 1 \cdot p = p$.
Deoarece X şi X^2 au acelaşi tabel, $E\left(X^2\right) = E(X) = p$.
De aceea, $Var(X) = E\left(X^2\right) - E(X)^2 = p - p^2 = p(1-p)$.

6) Fie X numărul de persoane care îşi iau înapoi propria pălărie.
Urmând indicaţia: fie $X_j = 1$ dacă persoana j îşi ia inapoi pălăria proprie şi $X_j = 0$ 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 aceeaşi probabilitate orice pălărie, avem $P(X_j = 1) = 1/100$.
Astfel, $X_j \sim$ Bernoulli$(1/100) \implies E(X_j) = 1/100 \implies E(X) = 1$.

2.5 Funcţii masă de probabilitate, funcţii densitate de probabilitate şi funcţii de repartiţie

7) a) Cdf a lui X este

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = \left(-e^{-\lambda t}\right)\Big|_0^x = 1 - e^{-\lambda x}, \forall x > 0.$$

b) Pentru $y > 0$ 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) a) Fie T numărul de apariţii ale lui 6 în 100 de aruncări.
Ştim că $T \sim$ Binomial$(100, 1/6)$.
Deoarece primim 3 dolari de fiecare dată când dăm 6, avem $X = 3T$. Deci, putem scrie

$$P(X = 3k) = C_{100}^k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{100-k}, \text{ pentru } k = 0, 1, 2, ..., 100.$$

Sau putem scrie

$$P(X = x) = C_{100}^{x/3} \left(\frac{1}{6}\right)^{x/3} \left(\frac{5}{6}\right)^{100-x/3}, \text{ pentru } x = 0, 3, 6, ..., 300.$$

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 apariţii ale lui 6 în primele 25 de aruncări. Deci, $X_1 = 3T_1$ şi $Y = 12T_1$.
$T_1 \sim$ Binomial$(25, 1/6)$, deci

$$E(Y) = E(12T_1) = 12E(T_1) = 12 \cdot 25 \cdot \frac{1}{6} = 50$$

şi

$$Var(Y) = Var(12T_1) = 144Var(T_1) = 144 \cdot 25 \cdot \frac{1}{6} \cdot \frac{5}{6} = 500.$$

ii) Mediile sunt aceleaşi din liniaritate deoarece X şi Y sunt ambele $300\times$ o 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.

2.6 Probabilitate comună, covarianţă, corelaţie

9) (**Puzzle aritmetic**) a) Suma probabilităţilor marginale este 1, deci cele 2 probabilităţi marginale lipsă sunt $P(X = 3) = 1/3$, $P(Y = 3) = 1/2$. Suma probabilităţilor 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:

$X \backslash Y$	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) **Covarianţă şi independenţă**
a)

Y \ X	-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 repartiţiile 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 intersecţiei nu este produsul probabilităţilor:
$P(X = -2, Y = 0) = 0 \neq 1/25 = P(X = -2)P(Y = 0)$.
Deoarece acestea nu sunt egale, X şi Y nu sunt independente. (Este evident că X^2 nu este independent de X.)
d) Folosind tabelul din a) şi 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) **Repartiţii comune continue**
a) $F(a,b) = P(X \leq a, Y \leq b) = \int_0^a \int_0^b (x+y)dydx = \int_0^a \left(xy + \frac{y^2}{2}\right)\Big|_{y=0}^{y=b} dx$
$= \int_0^a \left(xb + \frac{b^2}{2}\right) dx = \left(\frac{x^2}{2}b + \frac{b^2}{2}x\right)\Big|_0^a = \frac{a^2b+ab^2}{2}$.
Deci $F(x,y) = \frac{x^2y+xy^2}{2}$ pe $[0,1] \times [0,1]$ şi $F(1,1) = 1$.
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 şi Y sunt independente, verificăm dacă densitatea comună este produsul densităţilor marginale.
$f(x,y) = x + y$, $f_X(x) \cdot f_Y(y) = \left(x + \frac{1}{2}\right)\left(y + \frac{1}{2}\right)$.
Deoarece acestea nu sunt egale, X şi Y nu sunt independente.
d) $E(X) = \int_0^1 \int_0^1 x(x+y)dydx = \int_0^1 \left[\left(x^2y + x\frac{y^2}{2}\right)\Big|_{y=0}^{y=1}\right] dx = \int_0^1 \left(x^2 + \frac{x}{2}\right) dx$
$= \left(\frac{x^3}{3} + \frac{x^2}{4}\right)\Big|_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$.

(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, $E(Y) = \frac{7}{12}$.
$E\left(X^2+Y^2\right) = \int_0^1\int_0^1 \left(x^2+y^2\right)(x+y)dydx = \int_0^1\int_0^1 \left(x^3+x^2y+xy^2+y^3\right)dydx$
$= \int_0^1 \left[\left(x^3y+\frac{x^2y^2}{2}+\frac{xy^3}{3}+\frac{y^4}{4}\right)\Big|_{y=0}^{y=1}\right]dx = \int_0^1\left(x^3+\frac{x^2}{2}+\frac{x}{3}+\frac{1}{4}\right)dx$
$= \left(\frac{x^4}{4}+\frac{x^3}{6}+\frac{x^2}{6}+\frac{x}{4}\right)\Big|_0^1 = \frac{1}{4}+\frac{1}{6}+\frac{1}{6}+\frac{1}{4} = \frac{5}{6}$.
$E(XY) = \int_0^1\int_0^1 xy(x+y)dydx = \int_0^1\int_0^1(x^2y+xy^2)dydx = \int_0^1\left[\left(\frac{x^2y^2}{2}+\frac{xy^3}{3}\right)\Big|_{y=0}^{y=1}\right]dx$
$= \int_0^1\left(\frac{x^2}{2}+\frac{x}{3}\right)dx = \left(\frac{x^3}{6}+\frac{x^2}{6}\right)\Big|_0^1 = \frac{1}{6}+\frac{1}{6} = \frac{1}{3}$.
$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \frac{7}{12}\cdot\frac{7}{12} = -\frac{1}{144}$.

2.7 Legea numerelor mari, teorema limită centrală

12) Standardizăm:

$$\begin{aligned} P\left(\sum_i X_i < 30\right) &= P\left(\frac{\frac{1}{n}\sum X_i - \mu}{\sigma/\sqrt{n}} < \frac{30/n-\mu}{\sigma/\sqrt{n}}\right) \\ &\approx P\left(Z < \frac{30/100-1/5}{1/30}\right) \text{ (din teorema limită centrală)} \\ &= P(Z<3) \\ &\approx 0.9987 \text{ (din tabelul probabilităţilor normale sau din R cu } \texttt{pnorm(3)}\text{)} \end{aligned}$$

13) (Mai mult cu teorema limită centrală)
Fie X_j IQ-ul persoanei j. Avem $E(X_j) = 100$ şi $\sigma_{X_j} = 15$.
Fie $\overline{X}$ media IQ-urilor a 100 de persoane selectate aleator. Atunci avem

$$E(\overline{X}) = 100 \text{ şi } \sigma_{\overline{X}} = 15/\sqrt{100} = 1.5.$$

Problema cere $P(\overline{X} > 115)$. Standardizând, obţinem
$P(\overline{X} > 115) \approx P(Z > 10) = 0$ (din R, cu `1-pnorm(10)`).

3 Introducere în statistică

3.1 Scopurile învăţării

1. Să ştie cele 3 "faze" care se suprapun ale practicii statistice.
2. Să ştie ce se înţelege prin termenul *statistică*.

3.2 Introducere în statistică

Statistica se ocupă cu date. Vorbind în general, scopul statisticii este de a face inferenţe (deducţii) bazate pe date. Putem împărţi acest proces în 3 faze: colectarea datelor, descrierea datelor şi analiza datelor. Aceasta se potriveşte cu paradigma metodei ştiinţifice. Facem ipoteze despre ce este adevărat, colectăm date în experimente, descriem rezultatele şi apoi inferăm din rezultate puterea dovezilor care privesc ipotezele noastre.

3.2.1 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).

3.2.2 Statistică descriptivă

Datele neprelucrate iau adesea forma unei masive liste, matrice sau baze de date de etichete şi numere. Pentru a da sens datelor, putem calcula statistici de rezumat ca media, mediana şi domeniile intercuartile. Putem de asemenea să vizualizăm datele folosind dispozitive grafice ca histogramele, reprezentări ale împrăştierii şi cdf empirică. Aceste metode sunt utile atât pentru comunicarea cât şi pentru explorarea datelor pentru a obţine o perspectivă în structura lor, de exemplu pentru a recunoaşte dacă urmează o repartiţie de probabilitate cunoscută.

3.2.3 Statistică deductivă (inferenţială)

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 repartiţie normală. (Observăm că aceasta este totdeauna o aproximare deoarece ştim că eroarea trebuie să aibă o margine în timp ce repartiţia normală are domeniul $\mathbb{R}$.) Putem apoi folosi datele pentru a produce dovezi pentru sau împotriva acestei ipoteze. De exemplu, presupunând că lungimea gestaţiei are o repartiţie $N(\mu, \sigma^2)$, folosim datele lungimilor gestaţiei să zicem a 500 de sarcini pentru a trage concluzii despre valorile parametrilor μ şi σ. Similar, putem modela rezultatul alegerii a 2 candidaţi printr-o repartiţie Bernoulli(p) şi utiliza datele

sondajului pentru a trage concluzii despre valoarea lui p.
Rareori putem face afirmaţii 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 afirmaţii probabiliste. Din nefericire, media şi publicul larg înţeleg greşit sensul probabilist al afirmaţiilor statistice. De fapt, cercetătorii înşişi fac adesea aceeaşi eroare.
Exemplul 1. Pentru a studia eficacitatea unui nou tratament pentru cancer, pacienţii sunt recrutaţi şi apoi împărţiţi într-un grup experimental şi un grup de control. Grupului experimental i se dă noul tratament şi grupul de control primeşte standardul curent de îngrijire. Datele colectate de la pacienţi pot include informaţie demografică, istorie medicală, stadiul iniţial al cancerului, progresul cancerului în timp, costul tratamentului şi efectul tratamentului asupra mărimii tumorii, ratele de remisie, longevitatea şi calitatea vieţii. Datele vor fi folosite pentru a face deducţii despre eficacitatea noului tratament comparativ cu standardul curent de îngrijire.
Obsevăm că acest studiu va trece prin toate cele 3 faze descrise mai sus. Proiectarea experimentului trebuie să specifice mărimea studiului, cine va fi eligibil să se alăture, cum vor fi alese grupurile experimental şi de control, cum vor fi administrate tratamentele, dacă subiecţii sau doctorii ştiu sau nu cine ce tratament primeşte şi precis ce date se colectează, printre alte lucruri. Odată ce datele sunt colectate, trebuie descrise şi analizate pentru a determina dacă ele susţin ipoteza că noul tratament este mai (sau mai puţin) eficient decât cel(e) curent(e) şi cu cât. Aceste concluzii statistice vor fi formulate ca afirmaţii precise implicând probabilităţi.
După cum s-a observat mai sus, interpretarea greşită a înţelesului afirmaţiilor statistice este o sursă obişnuită 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 naştere (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 morţii subite infantile. Cu toate că a fost achitată în 2003, a suferit în timpul şi după închisoare de probleme psihiatrice foarte serioase şi a murit în 2007 de intoxicaţie cu alcool. Vezi `http://en.wikipedia.org/wiki/Sally_Clark`.
O discuţie despre cazul Sally Clark şi alte exemple de intuiţie statistică proastă este la adresa `http://www.youtube.com/watch?v=kLmzxmRcUTo`.

3.2.4 Ce este *o* statistică?

Dăm o definiţie simplă al cărei înţeles este cel mai bine elucidat de exemple.
Definiţie. 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 (şi felului în care îl aruncăm) care poate fi estimată folosind datele. O astfel de estimare este dată de statistica "numărul de 6-uri/1000".
Exemplul 4. Presupunem că tratăm un grup de pacienţi cu cancer cu o nouă procedură şi colectăm datele despre cât de mult supravieţuiesc după tratament. Din date putem calcula media timpului de supravieţuire al pacienţilor din grup. Putem folosi această statistică ca o estimare a timpului de supravieţuire pentru viitorii pacienţi cu cancer care urmează noua procedură.
Exemplul 5. Presupunem că întrebăm 1000 de rezidenţi dacă sprijină sau nu legalizarea marijuanei în Massachusetts. Proporţia din cei 1000 care sprijină propunerea este o statistică. Proporţia din toţi rezidenţii din Massachusetts care sprijină propunerea *nu* este o statistică deoarece nu i-am întrebat pe fiecare (observaţi cuvântul "colectate" din definiţie). Mai degrabă, sperăm să tragem o concluzie statistică despre proporţia la nivel de stat, bazată pe datele din eşantionul nostru aleator.
Folosim 2 tipuri generale de statistici:
1. Statistici punctuale: o singură valoare calculată din date, ca media de selecţie $\overline{x}_n$ sau deviaţia standard de selecţie s_n.
2. Statistici de interval: un interval $[a, b]$ calculat din date. Acesta este efectiv doar o pereche de statistici punctuale şi e adesea prezentat în forma $\overline{x} \pm s$.

3.3 Importanţa teoremei lui Bayes

Teorema lui Bayes este foarte importantă pentru statistica inferenţială. Reamintim că teorema lui Bayes ne permite să "inversăm" probabilităţile condiţionate. Adică, dacă H şi D sunt evenimente, atunci teorema lui Bayes spune

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

În experimentele ştiinţifice pornim cu o ipoteză şi colectăm date pentru a testa ipoteza. Adesea $H :=$ "ipoteza noastră e adevărată" şi D sunt datele

colectate. Cu aceste cuvinte, teorema lui Bayes spune

$$P(\text{ipoteza este adevărată} \mid \text{date}) = \frac{P(\text{date} \mid \text{ipoteza este adevărată})P(\text{ipoteza este adevărată})}{P(\text{date})}$$

Membrul stâng este probabilitatea că ipoteza noastră este adevărată date fiind datele colectate. Aceasta este precis ce am vrea să ştim. Când toate probabilităţile din dreapta sunt cunoscute exact, putem calcula probabilitatea din stânga exact. Din păcate, în practică rareori ştim valorile exacte ale tuturor probabilităţilor din dreapta. Statisticienii au dezvoltat un număr de moduri de a face faţă acestei lipse de cunoştinţe şi de a face totuşi inferenţe utile.

Exemplul 6. Test pentru boală

Presupunem că un test pentru o boală are 1% rată fals pozitivă şi 1% rată fals negativă. Presupunem de asemenea că rata bolii în populaţie este 0.002. În sfârşit, presupunem că o persoană selectată aleator face testul şi iese pozitiv. În limbaj de ipoteză şi date avem:
Ipoteza: $H =$ "persoana are boala".
Date: $D =$ "testul a fost pozitiv".
Ce vrem să ştim: $P(H|D) = P($ persoana are boala $\mid$ un test pozitiv $)$.
În acest exemplu toate probabilităţile din dreapta sunt cunoscute, deci putem folosi teorema lui Bayes pentru a calcula ce vrem să ştim.

$$\begin{aligned} P(\text{ipoteză} \mid \text{date}) &= P(\text{persoana are boala} \mid \text{un test pozitiv}) \\ &= P(H|D) \\ &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{0.99 \cdot 0.002}{0.99 \cdot 0.002 + 0.01 \cdot 0.998} \\ &= 0.1655518. \end{aligned}$$

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.