Proiect

Grupele 243 și 244

Notă: Raportul poate fi scris în *Microsoft Word* sau LaTeX(pentru ușurință recomand folosirea pachetului *rmarkdown* din *R* - mai multe informații găsiți pe site la secțiunea *Link-uri utile*). Toate simulările, figurile și codurile folosite trebuie incluse în raport (acesta trebuie să conţină comentarii şi concluzii, acolo unde sunt cerute). Se va folosi doar limbajul *R* (scripturile trebuie să fie comentate).

1 Problema 1

Considerăm următoarele distribuții: $\mathcal{B}(n, p)$, Pois(λ), Exp(λ), $\mathcal{N}(\mu, \sigma^2)$

1. Generați $N = 1000$ de realizări independente din fiecare repartiție și calculați media și varianța eșantionului.
2. Ilustrați grafic funcțiile de masă, respectiv funcțiile de densitate pentru fiecare din repartițiile din enunțul problemei. Considerați câte 5 seturi de parametrii diferiți pentru fiecare repartiție și suprapuneți graficele pe aceeași figură pentru fiecare rapetiție. Adăugați și legenda.
3. Pentru seturile de parametrii de la punctul anterior trasați funcțiile de repartiție pentru fiecare repartiție (tot suprapuse) și adăugați legenda corespunzătoare.

Scopul următoarelor subpuncte este de a evalua acuratețea unor aproximări ale funcției de repartiție a binomialei $\mathcal{B}(n, p)$, notată în cele ce urmează cu $F_{n,p}(k) = \mathbb{P}(X \leq k)$. Vom compara următoarele aproximări:

a) *Aproximarea Poisson*

$$F_{n,p}(k) \approx F_\lambda(k) = \sum_{x=0}^{k} e^{-\lambda} \frac{\lambda^x}{x!}, \quad \lambda = np$$

b) *Aproximarea Normală* (rezultată din Teorema Limită Centrală)

$$F_{n,p}(k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right)$$

c) *Aproximarea Normală cu factor de corecție*

$$F_{n,p}(k) \approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

d) *Aproximarea Camp-Paulson* (A se vedea (Lesch and Jeske 2009))

$$F_{n,p}(k) \approx \Phi\left(\frac{c - \mu}{\sigma}\right)$$

pentru $c = (1-b)r^{\frac{1}{3}}$, $\mu = 1 - a$ și $\sigma^2 = a + br^{\frac{2}{3}}$ unde $a = \frac{1}{9(n-k)}$, $b = \frac{1}{9(k+1)}$ și respectiv $r = \frac{[(k+1)(1-p)]}{[p(n-k)]}$.

Se cere:

4. Pentru fiecare $n \in \{25, 50, 100\}$ și fiecare $p \in \{0.05, 0.1\}$ să se afișeze un tabel cu șase coloane (`k`, `Binomiala`, `Poison`, `Normala`, `Normala Corecție`, `Camp-Paulson`) în care să apară aproximările de mai sus pentru funcția de repartiție și de masă a binomialei, pentru $k \in \{1, 2, \dots, 10\}$.

5. Pentru a cuantifica acuratețea aproximărilor de mai sus vom folosi ca metrică, *eroarea maximală absolută* dintre două funcții de repartiții F și H (numită și distanța Kolmogorov) dată de formula

$$d_K(F(k), H(k)) = \max_k |F(k) - H(k)|\,.$$

Pentru fiecare $n \in \{25, 50, 100\}$ ilustrați pe același grafic erorile maximale absolute (folosind diferite culori și simboluri pentru puncte) dintre funcția de repartiție binomială și cele patru aproximări de mai sus considerând $0.01 \leq p \leq 0.5$. Ce observați ?

Următoarele subpuncte fac referire la aproximarea funcției de repartiție a binomialei $\mathcal{B}(n, p)$ prin intermediul repartiției normale-asimetrice (skew-normal) (Chang et al. 2008). Spunem că o variabilă aleatoare X este repartizată normal-asimetric de parametrii μ (locație), σ (scală) și λ (coeficient de asimetrie) și notăm $X \sim SN(\mu, \sigma, \lambda)$ dacă are densitatea de repartiție

$$f(x) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right),$$

unde ϕ și Φ reprezintă densitatea respectiv funcția de repartiție a normalei standard.

6. Pentru 5 seturi de parametrii diferiți ilustrați grafic densitățile repartiției normale-asimetrice, suprapunând graficele pe aceeași figură. Adăugați și legenda.

e) *Aproximarea prin normala asimetrică* este dată de

$$F_{n,p}(k) \approx \Psi_\lambda\left(\frac{k + 0.5 - \mu}{\sigma}\right)$$

unde $\Psi_\lambda(x) = \int_{-\infty}^{x} 2\phi(t)\Phi(\lambda t)\,dt$ este funcția de repartiție a normalei-asimetrice standard ($SN(0, 1, \lambda)$). Parametrii (μ, σ, λ) se obțin din următoarele relații:

- având date (n, p), $\lambda = sign(1 - 2p)\sqrt{\lambda^2}$ unde λ^2 este soluția ecuației

$$\frac{\left(1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}\right)^3}{\frac{2}{\pi}\left(\frac{4}{\pi} - 1\right)^2\left(\frac{\lambda^2}{1+\lambda^2}\right)^3} = \frac{np(1-p)}{(1-2p)^2}$$

- pentru λ determinat anterior,

$$\sigma^2 = \frac{np(1-p)}{1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}}$$

- pentru λ și σ determinate mai sus,

$$\mu = np - \sigma\sqrt{\frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}}$$

7. Folosind funcția `uniroot()` din `R` pentru determinarea soluțiilor unei ecuații reale (`help(uniroot)`), creați o funcție care să primească parametrii (n, p) ca valori de intrare și să întoarcă valorile parametrilor (μ, σ, λ) conform relațiilor de mai sus. Ilustrați grafic pentru $n = 25$ și $p \in \{0.05, 0.1\}$ repartiția binomială $\mathcal{B}(n, p)$ (ca o funcție cu bare) peste care suprapuneți densitatea normalei-asimetrice corespunzătoare (cu parametrii determinați corespunzător).

8. Pentru fiecare $n \in \{25, 50, 100\}$ și fiecare $p \in \{0.05, 0.1\}$ să se afișeze un tabel cu trei coloane (`k`, `Binomiala`, `Normala Asimetrică`) în care să apară aproximărea de mai sus pentru funcția de repartiție și de masă a binomialei. Pentru calculul funcției de repartiție a normalei-asimetrice standard, $\Psi_\lambda(x)$, folosiți funcția `integrate()` din `R` (`help(integrate)`). Evaluați grafic acuratețea aproximării conform punctului 5.

2 Problema 2

a) Construiţi funcţia `fgam` care implementează proprietăţile integralei *gamma* după cum urmează:

- $\Gamma(1) = 1$
- $\Gamma(n) = (n-1)!$ pentru orice n natural nenul.
- $\Gamma(1/2) = \sqrt{\pi}$
- $\Gamma(a) = (a-1)\Gamma(a-1)$ pentru orice $a > 1$

Pentru $a < 1$ folosiţi funcţia `integrate` din `R` şi calculaţi valoarea integralei.

b) Construiţi funcţia `fbet` care implementează proprietăţile integralei *beta* după cum urmează:

- $\beta(a, b) = \beta(b, a)$
- $\beta(a, b) = \frac{\pi}{\sin(a\pi)}$ dacă $a + b = 1$ ($a > 0$ şi $b > 0$)
- $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ (apelaţi funcţia ‘fgam‘ construită anterior)

c) Fie X şi Y două variabile aleatoare independente repartizate $X \sim \text{Gamma}(a, b)$ şi $Y \sim \text{Beta}(a, b)$. Construiţi funcţiile `fprobgammanr`, `fprobbetanr` şi respectiv `fprobnr` cu *nr* luând valori de la 1 la 9 care primesc ca parametri valorile lui a și b asociate repartiţiilor Gamma şi respectiv Beta şi care calculează următoarele probabilităţi(folosind funcţiile `fgam`, `fbet` şi `optimise`):

1) $\mathbb{P}(X < 3)$
2) $\mathbb{P}(2 < X < 5)$
3) $\mathbb{P}(3 < X < 4 \,|\, X > 2)$
4) $\mathbb{P}(Y > 2)$
5) $\mathbb{P}(4 < X < 6)$
6) $\mathbb{P}(0 < X < 1 \,|\, X < 7)$
7) $\mathbb{P}(X + Y < 5)$
8) $\mathbb{P}(X - Y > 0.5)$
9) $\mathbb{P}(X + Y > 3 \,|\, X - Y > 0.5)$

Ilustraţi comportamentul celor 9 funcţii construite prin apelul acestora folosind valori pentru a şi b alese de voi (cel puțin 3).

Observaţie: **Nu** confundaţi funcţiile Γ şi β (adică integralele cu acelaşi nume) cu repartiţiile omonime! Pentru repartiţiile Gamma şi respectiv Beta folosiţi următoarele densităţi:

- dacă $X \sim \text{Gamma}(a, b)$ atunci $f(x) = \frac{1}{b^a\Gamma(a)}x^{a-1}e^{-\frac{x}{b}}$, $x > 0$, $a > 0$, $b > 0$
- dacă $Y \sim \text{Beta}(a, b)$ atunci $f(x) = \frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1}$, $x \in (0, 1)$, $a > 0$, $b > 0$

d) Calculaţi aceleaşi probabilităţi ca la punctul c) folosind funcţiile de sistem din `R` (nu cele construite de voi!) pentru repartiţiile Gamma şi respectiv Beta şi construiţi un tabel în care să centralizaţi rezultatele obţinute (pe prima coloană rezultatele de la punctul c) iar pe a doua rezultatele de la punctul d)). Ce justificare găsiţi pentru diferenţele observate?

3 Problema 3

3. Fie două variabile aleatoare discrete X şi Y cu repartiţiile:

$$X \sim \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad Y \sim \begin{pmatrix} y_1 & y_2 & \cdots & y_m \\ q_1 & q_2 & \cdots & q_m \end{pmatrix}.$$

a) Construiţi o funcţie `frepcomgen` care primeşte ca parametri `m` şi `n` şi care generează un tabel cu repartiţia comună a v.a. X şi Y incompletă, dar într-o formă în care poate fi completată ulterior.

Observaţie: Se cere la a) să generaţi valorile lui X, valorile lui Y şi suficient de multe valori pentru p_i, q_j şi respectiv π_{ij} astfel încât să poată fi determinată repartiţia comună a celor doua v.a.

Notă: În construirea algoritmului puteţi începe de la cazul particular $m = 3$ și $n = 2$. Dacă reuşiţi să oferiţi soluţia doar pentru acest caz particular, dar nu şi pentru cazul general veţi primi punctaj parţial.

b) Construiţi o funcţie `fcomplrepcom` care completează repartiţia comună generată la punctul anterior (pentru cazul particular sau pentru cazul general).

Notă: În cazul în care nu ştiţi să rezolvaţi punctul a) puteţi construi o funcţie care să determine repartiţia comună pornind de la un exemplu discutat la seminar.

c) Având la dispoziţie repartiţia comună a v.a. X şi Y de la punctul b) calculaţi:

- $Cov(3X, 4Y)$
- $\mathbb{P}(0 < X < 5 \,|\, Y > 4)$
- $\mathbb{P}(X > 3, Y < 7)$

d) Pentru exemplul obţinut la punctul b) construiţi două funcţii `fverind` şi respectiv `fvernecor` cu ajutorul cărora să verificaţi dacă variabilele X şi Y sunt:

- independente
- necorelate

Referințe

Chang, Ching-Hui, Jyh-Jiuan Lin, Nabendu Pal, and Miao-Chen Chiang. 2008. "A Note on Improved Approximation of the Binomial Distribution by the Skew-Normal Distribution." *The American Statistician* 62 (2): 167–70. https://doi.org/10.1198/000313008x305359.

Lesch, Scott M., and Daniel R. Jeske. 2009. "Some Suggestions for Teaching About Normal Approximations to Poisson and Binomial Distribution Functions." *The American Statistician* 63 (3): 274–77. https://doi.org/10.1198/tast.2009.08147.