Proiect la Statistică

Grupele 311, 312, 322

-partea I- 15p.

I. Se dau numerele reale a,b,c astfel încât $a < b \le c$ și $b \ge 0$. Se generează aleator numere din intervalul (a,b) până când suma lor depășește valoarea c.

1) Determinați, construind un algoritm în R care efectuează experimentul de mai sus de 10^9 ori, o aproximare a numărului k ce reprezintă de câte ori este necesar, *în medie,* a se extrage aleator numerele respective până când suma lor depășeste valoarea c.
2) Determinați valoarea exactă a lui k și comparați cu valoarea obținută în urma simulării. Se poate determina în ce fel depinde eroarea de aproximare de numerele a,b,c?

Indicație: Începeți prin a lucra cu un caz particular(alegeți a,b,c într-o manieră convenabilă) și apoi generalizați soluția.

II. Se consideră o activitate care presupune parcurgerea secvențială a n etape. Timpul necesar finalizării etapei i de către o persoană A este o variabilă aleatoare $T_i \sim Exp(\lambda_i)$. După finalizarea etapei i, A va trece în etapa $i+1$ cu probabilitatea α_i sau va opri lucrul cu probabilitatea $1-\alpha_i$. Fie T timpul total petrecut de persoana A în realizarea activității respective.

1) Construiți un algoritm în R care simulează 10^6 valori pentru v.a. T și în baza acestora aproximați $E(T)$. Reprezentați grafic într-o manieră adecvată valorile obținute pentru T. Ce puteți spune despre repartiția lui T?
2) Calculați valoarea exactă a lui $E(T)$ și comparați cu valoarea obținută prin simulare.
3) În baza simulărilor de la 1) aproximați probabilitatea ca persoana A să finalizeze activitatea.
4) În baza simulărilor de la 1) aproximați probabilitatea ca persoana A să finalizeze activitatea într-un timp mai mic sau egal cu σ.
5) În baza simulărilor de la 1) determinați timpul minim și respectiv timpul maxim în care persoana A finalizează activitatea și reprezentați grafic timpii de finalizare a activității din fiecare simulare. Ce puteți spune despre repartiția acestor timpi de finalizare a activității?
6) În baza simulărilor de la 1) aproximați probabilitatea ca persoana A să se oprească din lucru înainte de etapa k, unde $1 < k \le n$. Reprezentați grafic probabilitățile obținute într-o manieră corespunzătoare. Ce puteți spune despre repartiția probabilităților obținute?

-partea a II-a - 15p.

III. Folosind funcţia ***check.convergence*** din pachetul R ***ConvergenceConcepts*** verificaţi dacă pentru următoarele exemple sunt verificate convergenţa în lege(în distribuţie), în probabilitate şi respectiv convergenţa aproape sigură. Interpretaţi şi comentaţi rezultatele obţinute.

1) Fie $X_1, X_2, \dots X_n$ v.a. i.i.d $X_i \sim Beta(\frac{1}{n}, \frac{1}{n})$ şi $X \sim Binomial(1, \frac{1}{2})$. Verificaţi dacă $X_n \xrightarrow{D} X$. Dar în cazul în care $X_i \sim Beta(\frac{a}{n}, \frac{b}{n})$, cu $a>0, b>0$?
2) Fie $X_1, X_2, \dots X_n$ v.a. i.i.d uniform distribuite pe mulţimea de valori $\{\frac{1}{n}, \frac{2}{n} \dots 1\}$ şi $X \sim Unif(0,1)$. Verificaţi dacă $X_n \xrightarrow{D} X$. Dar $X_n \xrightarrow{P} X$?
3) Fie $X_1, X_2, \dots X_n$ v.a. i.i.d . Notăm cu m şi respectiv M infimumul şi respectiv supremumul mulţimii valorilor pe care le poate lua X_1.
(i.e. $P(m \le X \le M)=1$, $P(X_1<a)>0$ şi $P(X_1>b)>0$ pentru orice $a>m$ şi respectiv $b<M$).
Verificaţi că $\min\{X_1, X_2 \dots X_n\} \xrightarrow{a.s.} m$ şi că $\max\{X_1, X_2 \dots X_n\} \xrightarrow{a.s.} M$.

Notă: Pentru utilizarea pachetului ***ConvergenceConcepts*** folosiţi documentaţia primită la laborator.

IV. **Def.** O variabilă aleatoare X face parte din ***familia exponențială s-dimensională*** dacă densitatea/funcția de masă poate fi scrisă sub forma:

$$p(x;\theta) = h(x) \cdot \exp\left(\sum_{i=1}^{s} \eta_i(\theta) \cdot T_i(x) - A(\theta)\right),$$

unde η_i şi A sunt funcții reale de $\theta = (\theta_1, \theta_2, \dots \theta_s)$, T_i reprezintă statistici suficiente, iar h este o funcție pozitivă de x.

Funcția $A(\theta)$ se numește ***constantă de log-normalizare***(rolul ei este acela de a face ca $p(x;\theta)$ să îndeplinească acele condiții necesare pentru a fi o funcție de densitate de probabilitate/funcție de masă, după caz).

Ex. Fie $X \sim Norm(\mu, \sigma^2)$.

Vom demonstra ca aceasta face parte din familia exponențială 2-dimensională:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \ln(\sigma)\right)$$

Identificăm următoarele relații:

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$\eta_1(\theta_1,\theta_2) = \frac{\mu}{\sigma^2},\ \eta_2(\theta_1,\theta_2) = -\frac{1}{2\sigma^2},\ \theta_1 = \mu,\ \theta_2 = \sigma^2$$

$$A(\theta) = \frac{\mu^2}{2\sigma^2} + \ln(\sigma)$$

$T_1(x) = x,\ T_2(x) = x^2$

Așadar, repartiția normală(cu parametrii media μ și dispersia σ^2) face parte din familia exponențială 2-dimensională.

1) Verificați dacă următoarele repartiții fac parte din familia exponențială:
   a) Binom(3,p)
   b) Binom(n,p)
   c) Geom(p)
   d) Pois(λ)
   e) Gamma(α, β)
   f) Beta(α, β)
   g) $\chi^2(\nu)$

2) Ilustrați grafic în R densitățile/funcțiile de masă(după caz) ale repartițiilor de mai sus pentru 4 parametri particulari, la alegere, în cadrul aceluiași sistem de axe ortogonale(fiecare repartiție va avea însă reprezentări grafice distincte de celelalte repartiții).
3) Construiti funcția de log-verosimilitate(logL) pentru familia exponențială și demonstrați că aceasta este ***concavă***(construiți matricea hessiană și arătați că este negativ definită).
4) Particularizați forma funcției de log-verosimilitate de la 2) pentru repartițiile de la 1) ce fac parte din clasa exponențială cu un parametru. Reprezentați grafic(în **R**) aceste funcții și găsiți punctul lor de maxim(folosiți funcția ***optimise***).
5) Particularizați forma funcției de log-verosimilitate de la 2) pentru repartițiile de la 1) ce fac parte din clasa exponențială cu doi parametri. Fixați, pe rând, unul din parametri(alegeți o valoare particulară după cum doriți), reprezentați grafic(în **R**) aceste funcții în raport cu celălalt parametru și găsiți punctul lor de maxim(folosiți funcția ***optimise***).
6) Calculați MIRC pentru familia exponențială cu un parametru și particularizați valoarea acesteia pentru repartițiile de la 1) ce fac parte din familia exponențială cu un parametru.
7) Construiți în R o funcție care afișează MIRC pentru o repartiție selectată din 8 disponibile(alegeți voi aceste repartiții).
8) Pentru familia exponențială cu un parametru și respectiv cu doi parametri calculați estimatorul de verosimilitate maximă și respectiv estimatorul dat de metoda momentelor. Ce legatură există între aceștia?
9) Ilustrați în R, pentru un eșantion de dimensiune 1000 generat de voi în prealabil -*pentru toate repartițiile de la 1) ce fac parte din familia exponențială*- faptul că estimațiile obținute în baza celor 2 metode de estimare(metoda verosimilității maxime și metoda momentelor) pentru eșantionul respectiv sunt foarte apropiate de valoarea adevarată a parametrului de interes.

Precizări importante

1) Fiecare echipa va trimite prin liderul echipei un singur e-mail la adresa simona.cojocea@fmi.unibuc.ro ce va conține o arhivă cu 2 componente: codul R(comentat!) și documentația proiectului, în format .docx sau .pdf.
2) Documentația proiectului trebuie să conțină, pe prima pagină, numele membrilor echipei, liderul echipei și grupa din care face parte fiecare.
   Pentru fiecare exercițiu în parte documentația trebuie să conțină:
   - **Calculele matematice solicitate**(pot fi tehnoredactate sau scrise de mână și scanate, însă dacă apelați la ultima variantă vă rog să vă asigurați că ați scris **citeț,** fără ștersături sau zone scanate deficitar)
   - **Codul R comentat**
   - **Graficele realizate**(acolo unde au fost cerute)
   - **Comentariile și concluziile voastre**
3) Toate exercițiile sunt obligatorii. Dacă însă, odată cu rezolvarea lor, identificați și rezolvați unele cerințe suplimentare care sunt ***relevante*** pentru subiectul respectiv puteți obține un bonus de până la 10 p, fără ca nota finală de laborator să poată depăși 50 p.
4) Termenul de trimitere a proiectului este **2 februarie 2022 ora 22:00**.