Pachetul continuousRV

I. Membrii echipei

1. Vasiliu Diana-Elena – leader
2. Grigore Ioana
3. Gușuleac Bogdan
4. Cibotaru Matei-Ștefan

II. Prezentare generală

Din lipsa unui pachet asemănător pachetului R numit ”discreteRV” (care permite lucrul cu variabile aleatoare discrete), am fost inspirați să creem acest proiect (implementând pachetul R ”continuousRV”, care să permită lucrul cu variabile aleatoare continue).

Pachetul implementează câteva funcții care ajută la calcularea unor rezultate probabilistice pe baza unor date modelate de variabile aleatoare continue. Câteva dintre acestea sunt: verificarea dacă o funcție dată este sau nu functie densitate de probabilitate, calculul mediei, dispersiei și a momentelor unei variabile aleatoare, calculul covariantei și a coeficientului de corelație Pearson și altele.

III. Aspecte teoretice extra

Câteva dintre funcțiile extra (nefolosite în cadrul laboratorului) folosite sunt:

1) *Vectorize()* — funcția *integrate()* primește ca parametru o funcție vectorizată. Acest lucru se rezolvă prin folosirea funcției *Vectorize()*

2) *tryCatch()* — echivalent cu un bloc *try-catch* din alte limbaje de programare. Creează multiple ramuri pe care intră execuția programului – în prima ramură se află codul principal; dacă apar erori sau warning-uri, se intră pe blocul funcției corespunzătoare

3) *sapply()* — asemănătoare cu funcția *lapply()*, care primește ca parametru o listă L și o funcție f și aplică funcția f fiecărui element din lista L. În plus față de *lapply()*, *sapply()* încearcă să simplifice rezultatul

4) *pracma* — pachet pentru folosirea funcției *integral()*

5) *GoFKernel* — pachet pentru folosirea funcției *inverse()*

IV. Software

Proiectul a fost realizat folosind aplicația R Studio, un IDE pentru limbajul de programare R, precum și pachetele extra *pracma* și *GoFKernel*.

V. Explicarea codului

Cerința 1. Se cere determinarea unei constante de normalizare k pentru o funcție f specificată de utilizator.

Funcția: *normConst()*

Parametri: *f* – funcția care trebuie normalizată
domain – domeniul de definiție al funcției f

Această problemă presupune găsirea unei constante k astfel încât funcția g(x) = k * f(x) să fie o densitate de probabilitate. Pentru a fi densitate de probabilitate, această funcție trebuie să fie pozitivă pe tot intervalul $(-\infty, \infty)$ și

$$\int_{-\infty}^{\infty} k * f(x)dx = 1$$

Pentru aflarea constantei k, am integrat funcția f pe intervalul $(-\infty, \infty)$ și am împărțit 1 la această valoare. În cazul în care integrala este egală cu 0 sau funcția nu este pozitivă pe tot domeniul, atunci funcția nu se poate normaliza și este afișat un mesaj de eroare.

Cerința 2. Se cere verificarea dacă o funcție dată este densitate de probabilitate.

Funcția: *isPDF()*

Parametri: *f* – funcția care trebuie verificată
domain – domeniul de definiție al funcției f

Condițiile ca o funcție să fie densitate de probabilitate sunt ca ea să fie pozitivă pe tot intervalul și

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Inițial, este verificat dacă funcția este definită și pozitivă pe tot domeniul. În caz afirmativ, se calculează integrala, care este comparată apoi cu valoarea 1.

Cerința 3. Se cere crearea unui obiect de tip variabilă aleatoare continuă pornind de la o densitate de probabilitate introdusă de utilizator.

Funcție: *valc()*

Parametri: *f* – funcția densitate de probabilitate

Inițial, este verificată funcția f de proprietățile densității de probabilitate. În caz afirmativ, este creată o variabilă aleatoare continuă. Altfel, un mesaj de eroare este afișat.

Cerința 4. Se cere reprezentarea grafică a densității și a funcției de repartiție pentru diferite valori ale parametrilor.

Funcția: *plots()*

Parametri: *t* – intervalul pe care trebuie reprezentate repartițiile
norm, binom, exp, unif – indicatori ai tipurilor de repartiții ce trebuie reprezentate grafic
pdf, cdf – indicatori ai tipurilor de funcții ce trebuie reprezentate grafic

Pentru a ști ce grafic (pdf sau cdf) și al cărei repartiții să genereze graficul, funcția primește ca parametri niște valori de tip TRUE/FALSE, pentru fiecare dintre cazurile dorite de utilizator. De exemplu, dacă se dorește generarea exemplelor de grafice pentru pdf-ul unei repartiții normale, se vor seta parametrii `pdf=TRUE` și `norm=TRUE`. În funcție de acești parametri, vor fi afișate graficele pentru repartițiile cerute, iar parametrii repartițiilor vor fi afișati in legendă.

Cerința 5. Se cere calculul mediei, dispersiei și momentelor inițiale și centrate până la ordin 4.

Funcții: *E()* – media
Var() – dispersia
imoment() – momentul initial
cmoment() – momentul centrat

Parametri:

- funcțiile E() și Var(): *pdf* – funcția densitate de probabilitate a variabilei aleatoare
  … – argumente suplimentare
- funcțiile imoment() și cmoment(): *pdf* – funcția densitate de probabilitate a variabilei aleatoare
  ord – ordinul momentului care trebuie calculat
  mean – media variabilei aleatoare (necesar pentru *cmoment()*)

Media este calculată după formula

$$E(X) = \int_{-\infty}^{\infty} x * f(x)dx$$

unde f(x) este funcția densitate de probabilitate a variabilei aleatoare continue X. Funcția din pachetul R calculează inițial o funcție auxiliară, h(x) = x * f(x), pe care apoi o integrează pe $(-\infty, \infty)$.

Varianța unei variabile aleatoare continue este calculată după formula

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

care este echivalentă cu formula momentului centrat de ordin 2.

Momentul inițial de ordin r este calculat după formula

$$\mu_r = \int_{-\infty}^{\infty} (x - m)^r * f(x)dx$$

unde m este media lui X, iar momentul centrat de ordin r este calculat după formula

$$m_r = \int_{-\infty}^{\infty} x^r * f(x)dx$$

Toate aceste formule au fost implementate ca atare în R, folosind funcția *integrate()* din pachetul *stats*.

Cerința 6. Se cere calculul mediei și dispersiei unei variabile aleatoare h(X), unde X are o repartiție cunoscută, iar g este o funcție continuă precizată de utilizator.

Funcții: *mean_h()* – media variabilei aleatoare h(X)
var_h() – dispersia variabilei aleatoare h(X)

Parametri: *h* – funcția h aplicată variabilei aleatoare X
pdf – funcția densitate de probabilitate a variabilei aleatoare X

Media unei variabile aleatoare h(X) este calculată după formula

$$E(X) = \int_{-\infty}^{\infty} h(x) * f(x)dx$$

iar dispersia este calculată după formula clasică

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

în pachetul continuousRV fiind folosită, în acest caz, prima variantă, din motive de simplitate: funcția *mean_h()* calculează prima parte, funcția E(X) calculează a doua parte, iar rezultatele sunt puse împreună în funcția *var_h()*.

Cerința 9. Se cere generarea a n valori dintr-o repartiție de variabile aleatoare continue, unde n și tipul repartiției sunt specificate de utilizator.

Funcție: *generate()*

Parametri: fără parametri

Funcția cere de la tastatură un număr n de valor de generat și tipul repartiției dorite și generează cele n valori ale repartiției și o histogramă, pentru vizualizare.

Cerința 10. Se cere calculul covarianței și al coeficientului de corelație pentru două variabile aleatoare continue.

Funcții: *Cov()* – covarianța
Cor() – coeficientul de corelație

Parametri: *fcommon* – funcția densitate de probabilitate comună
interval_x – domeniul de valori al lui X
interval_y – domeniul de valori al lui Y

Pentru această problemă, este nevoie de densitatea comună celor două variabile aleatoare, întrucât covarianța are formula

$$Cov(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big) =$$

$$\int_c^d \int_a^b xy * f(x,y)dxdy - \mu_X\mu_Y$$

unde $\mu_X = E(X)$, $\mu_Y = E(Y)$, X ia valori în intervalul [a,b], Y ia valori în intervalul [c,d], iar f(x,y) este densitatea comună pentru X și Y.

Coeficientul de corelație este calculat după formula

$$Cor(X,Y) = \rho = \frac{Cov(X,Y)}{\sqrt{Var(X) * Var(Y)}}$$

Cele două formule au fost implementate în R folosind și o parte din celelalte funcții implementate în cadrul proiectului, cum ar fi *E()* și *marginalPDF()*.

Cerința 11. Se cere construirea densităților marginale pornind de la densitatea comună a două variabile aleatoare continue.

Funcție: *marginalPDF()*

Parametri: *fcommon* – funcția densitate de probabilitate comună
interval_x – domeniul de valori al lui X
interval_y – domeniul de valori al lui Y

Fiind dată densitatea comună $f(x,y)$, cele două densitățile marginale ale lui X, respectiv Y sunt calculate după formulele

$$f_X(x) = \int_c^d f(x,y)dy \qquad\qquad f_Y(y) = \int_a^b f(x,y)dx$$

unde X ia valori în intervalul [a,b] și Y ia valori în intervalul [c,d].

Funcția *marginalPDF()* returnează o listă cu două valori de tip funcție, prima reprezentând $f_X(x)$, iar a doua, $f_Y(y)$.

Cerința 12. Se cere calcularea sumei și a diferenței a două variabile aleatoare continue independente, folosind formula de convoluție.

Funcții: *csum()* – calculează suma
cdif () – calculează diferența

Parametri: *fx, fy* – funcțiile densitate de probabilitate pentru fiecare din cele două variabile aleatoare continue pe care se efectuează calculul

Formula de convoluție este:

$$sum(X,Y) = f_Z(z) = \int_{-\infty}^{\infty} f_X(z - x) f_Y(x)dx$$

$$dif(X,Y) = f_Z(z) = \int_{-\infty}^{\infty} f_X(x - z) f_Y(x)dx$$

Funcțiile *csum()* și *cdif()* implementează ca atare cele două funcții și returnează o funcție care reprezintă funcția densitate de probabilitate a variabilei aleatoare continue $Z = X + Y$, respectiv $Z = X - Y$.

VI. Provocări

Una din principalele provocări a fost înțelegerea profundă a elementelor de sintaxă ale limbajului R, fiind un limbaj nou pentru toți membrii echipei. Cea mai mare parte a timpului a fost investită în căutarea resurselor puse la dispoziție de R pentru calculul elementelor din pachetul continuousRV.

O altă provocare a fost înțelegerea funcției *integrate()* și a faptului că este necesară o funcție vectorizată trimisă ca parametru. Acest lucru este încă parțial neclar pentru unii sau complet neclar pentru alții.

O problemă interesantă a fost la implementarea funcției *marginalPDF(),* unde a fost necesară integrarea unei funcții de 2 variabile în funcție de doar una din ele și returnarea rezultatului ca funcție, precum și selectarea elementelor din lista returnată în cadrul funcțiilor *Cov()* și *Cor()*.

VII. Concluzii

Nivelul de dificultate al proiectului a fost condiderat mediu spre dificil, din cauza întâlnirii cu un limbaj de programare complet nou și nu foarte ușor de înțeles. În plus, teoria de probabilități și statistică nu este ușoară și trebuie ținut cont de multe formule și de multe condiții pentru a implementa o funcție corectă.

În urma realizării proiectului, am învățat mai multe despre limbajul R și am învățat despre cum se creează un pachet. Am asimilat și teoria pentru variabile aleatoare continue, necesară pentru studiul ulterior al statisticii.

VIII. Contribuitori

Vasiliu Diana-Elena – cerințele 1, 2, 10, 11

Grigore Ioana – cerințele 4, 5, 6

Gușuleac Bogdan – cerința 9

Cibotaru Matei-Ștefan – cerințele 3, 12

IX. Bibliografie

rpackage_instructions.pdf
V.a. continue.pdf
https://cran.r-project.org/
https://data-flair.training/blogs/r-arguments-introduction/
https://en.wikipedia.org/wiki/Normalizing_constant
https://en.wikipedia.org/wiki/Quantile_function
http://mazamascience.com/WorkingWithData/?p=912
https://mikmart.rbind.io/2018/02/17/finding-expected-values-of-random-variables/
https://stackoverflow.com/questions/8913603/calculating-double-integrals-in-r-quickly
https://www.statmethods.net/advgraphs/probability.html