

Probabilități și Statistică - Curs 9

Mai, 2018

- 1 Variabile aleatoare continue
 - Variabile aleatoare continue
 - Distribuții continue remarcabile
- 2 Teoremele fundamentale
 - Legea numerelor mari
 - Inegalitățile lui Markov și a lui Cebâșev revăzute
 - Teorema lui Cebâșev
 - Legea numerelor mari
 - Teorema limită centrală
 - Aproximarea normală a distribuției binomiale
- 3 Simulare
 - Simularea variabilelor aleatoare
 - Aplicații ale LNM și TLC
- 4 Bibliography

- În cazul în care $|\Omega| \geq |\mathbb{R}|$ (i.e., cardinalul lui Ω este cel puțin continuu), evenimentele aleatoare se definesc diferit față de cazul discret.
- Diferența principală constă în aceea că nu orice submulțime $A \subseteq \Omega$ este în mod necesar eveniment aleator: familia evenimentelor aleatoare este o σ -algebră $\mathcal{A} \subseteq \mathcal{P}(\Omega)$:
 - $\emptyset, \Omega \in \mathcal{A}$;
 - dacă $A_1, A_2 \in \mathcal{A}$, atunci $A_1 \cap A_2 \in \mathcal{A}$;
 - dacă $(A_n)_{n \geq 1} \subseteq \mathcal{A}$, atunci $\bigcup_{n \geq 1} A_n \in \mathcal{A}$.
- Iar funcția de probabilitate este definită numai pe submulțimile din \mathcal{A} (cu axiomele cunoscute):
$$P : \mathcal{A} \rightarrow [0, 1].$$

- O funcție $X : \Omega \rightarrow \mathbb{R}$ este numită **variabilă aleatoare** dacă pentru orice J interval din $\overline{\mathbb{R}}$, $X^{-1}(J) \in \mathcal{A}$.

- O variabilă aleatoare $X : \Omega \rightarrow \mathbb{R}$ se numește **continuă** dacă are funcția de repartiție continuă (*câteodată, această definiție se referă la toate cazurile când $X(\Omega)$ este de cardinal continuu*).

- Distribuția (repartiția) unei astfel de variabile poate fi dată prin **funcția de repartiție**:

$$F : \mathbb{R} \rightarrow [0, 1], F(a) = P(X \leq a),$$

- sau prin **funcția de densitate (de masă)**, $f : \mathbb{R} \rightarrow [0, +\infty)$, astfel încât funcția de repartiție F poate fi descrisă astfel:

$$F(a) = P(X \leq a) = \int_{-\infty}^a f(t) dt.$$

- Orice funcție $f : \mathbb{R} \rightarrow [0, +\infty)$ cu $\int_{-\infty}^{\infty} f(t) dt = 1$ este funcția de densitate pentru o anumită avariabilă aleatoare continuă (sau, mai simplu, distribuție continuă).
- Folosind funcția de densitate putem calcula (dacă integralele există) media și dispersia:

$$M(X) = \int_{-\infty}^{+\infty} tf(t) dt \text{ și } D^2(X) = \int_{-\infty}^{+\infty} [t - M(X)]^2 f(t) dt.$$

- Probabilitățile asociate unei variabile aleatoare continue se calculează astfel:

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(t) dt,$$

adică aria aflată sub graficul funcției f între punctele $t = a$ și $t = b$.

- Dacă F este continuă, $P(X = a) = F(a) - F(a^-) = 0$ și $P(a \leq X < b) = P(a \leq X \leq b) = P(a < X \leq b)$.

- Pentru o variabilă aleatoare $X: \Omega \rightarrow \mathbb{R}$, dată, operația de **standardizare** constă în următoarea transformare a variabilei X :

$$Y = \frac{X - \mathbb{E}[X]}{StDev[X]}.$$

- Noua variabilă este "standard" pentru că are

$$\mathbb{E}[Y] = 0 \text{ și } Var[Y] = 1.$$

Exemplul 1. Durata vieții în ani a unei anumite componente electronice este o variabilă aleatoare continuă, cu funcția de densitate

$$f(x) = \begin{cases} \frac{k}{x^4}, & x \geq 1 \\ 0, & x < 1 \end{cases}$$

Determinați k , funcția de repartiție și probabilitatea ca viața unei astfel de componente să depășească 2 ani.

Soluție. Trebuie ca $f(t) \geq 0, \forall t \in \mathbb{R}$ și $\int_{-\infty}^{\infty} f(t) dt = 1$, deci

$$k \geq 0 \text{ și } 1 = \int_1^{\infty} \frac{k}{t^4} dt = \left[-\frac{k}{3t^3} \right]_1^{\infty} = \frac{k}{3}, \text{ de unde } k = 3.$$

Funcția de repartiție este $F(x) = \int_{-\infty}^x f(t) dt$, deci

$$F(x) = \begin{cases} 1 - \frac{1}{x^3}, & x \geq 1 \\ 0, & x < 1 \end{cases}$$

Fie X durata de viață a acestei componente electronice, probabilitatea ca durata de viață să fie cel puțin 2 ani este $P(X \geq 2) = 1 - P(X < 2) = 1 - F(2) = 1/8$ (deoarece F e continuă).

Exemplul 2. Fie X o variabilă aleatoare continuă cu următoarea funcție de densitate:

$$f(x) = \begin{cases} \alpha x, & 0 \leq x \leq 2 \\ 0, & \text{altfel} \end{cases}$$

Determinați α , funcția de repartiție, media și dispersia lui X .

Exemplul 3. Timpul (în minute) necesar unui anumit sistem să repornească este o variabilă aleatoare continuă cu densitatea

$$f(t) = \begin{cases} C(10 - x)^2, & 0 < x < 10 \\ 0, & \text{altfel} \end{cases}$$

Calculați C și probabilitatea ca timpul de repornire să fie între 1 și 2 minute.

Exemplul 4. Durata de viață (în ani) a unui tip de HD este o variabilă aleatoare continuă cu densitatea

$$f(t) = \begin{cases} K - \frac{x}{50}, & 0 < x < 10 \\ 0, & \text{altfel} \end{cases}$$

Calculați K , probabilitatea ca o eroare hardware să apară în primii 5 ani și durata medie de viață a acestui HD.

Distribuția uniformă. Se notează cu $U(a, b)$ și are funcția de densitate

$$f(t) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & x \in [a, b] \\ 0, & x > b \end{cases}$$

Dacă $X : U(a, b)$, atunci $\mathbb{E}[X] = \frac{a+b}{2}$ and $Var[X] = \frac{(b-a)^2}{12}$.

$U(0, 1)$ se numește **distribuția uniformă standard**.

Distribuția exponențială. Se notează cu $Exp(\lambda)$ și are funcția de densitate ($\lambda > 0$)

$$f(t) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$

Pentru $X : Exp(\lambda)$, $\mathbb{E}[X] = \frac{1}{\lambda}$, $Var[X] = \frac{1}{\lambda^2}$.

Distribuția exponențială este utilizată pentru a modela timpul de așteptare, timpul între două sosiri, durata de viață hardware etc; într-o secvență de evenimente rare timpul dintre două astfel de evenimente este distribuit exponențial.

Distribuția exponențială nu are memorie (trecerea a x_0 minute nu are relevanță): chiar dacă $X > x$, când timpul total de așteptare depășește x , timpul rămas are o distribuție exponențială: $P(X > x + \Delta x | X > x) = P(X > \Delta x)$ (de ce?).

Distribuția gaussiană (normală). Se notează cu $N(\mu, \sigma^2)$ și are funcția de densitate

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}}.$$

Dacă $X : N(\mu, \sigma^2)$, atunci $\mathbb{E}[X] = \mu$ and $\text{Var}[X] = \sigma^2$.

Distribuția $N(0, 1)$ este numită *distribuția normală standard*.

Valorile unei variabile normale distribuite au următoarea împrăștiere (simetric în jurul mediei): %68 se găsesc în intervalul $[\mu - \sigma, \mu + \sigma]$, %95 în $[\mu - 2\sigma, \mu + 2\sigma]$, iar %99.7 aparțin intervalului $[\mu - 3\sigma, \mu + 3\sigma]$.

- Distribuția normală are un rol central în teoria probabilităților și statistică pentru cel puțin două motive.
- Drept o consecință a Teoremei Limită Centrale (TLC - vezi mai jos) sumele și/sau mediile variabilelor independente și identic distribuite au ca aproximație o distribuție normală.
- De-a lungul timpului s-a observat că distribuția normală este un model potrivit pentru foarte multe variabile: temperatura, greutatea, înălțimea și chiar notele studenților.
- Distribuția normală a fost utilizată implicit de către de Moivre pentru aproximarea distribuției binomiale și ulterior de către Laplace și Gauss (în mod explicit).

Distribuția Student (sau t). Este notată cu $t(r)$ și are funcția de densitate

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{-\frac{r+1}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

unde $\Gamma(y) = \int_0^{+\infty} x^{y-1} e^{-x} dx$. Dacă $X : t(r)$, atunci $\mathbb{E}[X] = 0$ and $\text{Var}[X] = \frac{r}{r-2}$.

Cu cât este mai mare numărul de grade de libertate cu atât distribuția Student seamănă mai mult cu cea normală standard.

Distribuția Gamma. Se notează cu $\Gamma(\alpha, \lambda)$ și are funcția de densitate

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases},$$

unde $\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x} dx$. α este forma iar λ este rata (sau frecvența) distribuției. Pentru $X : \Gamma(\alpha, \lambda)$, avem $\mathbb{E}[X] = \frac{\alpha}{\lambda}$ and $Var[X] = \frac{\alpha}{\lambda^2}$.

Să presupunem că avem un proces care constă în α pași independenți și fiecare astfel de pas necesită un timp egal cu $Exp(\lambda)$, atunci durata totală urmează o distribuție Gamma. Astfel, distribuția Gamma este o sumă de α variabile independente identic repartizate exponențial.

Proposition 1

Fie $X \geq 0$ o variabilă aleatoare. Dacă $a > 0$, atunci

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

proof:

$$\begin{aligned} \mathbb{E}[X] &= \int_0^{+\infty} tf(t)dt = \int_0^a tf(t)dt + \int_a^{+\infty} tf(t)dt \geq \\ &\int_a^{+\infty} tf(t)dt \geq a \int_a^{+\infty} f(t)dt = aP(X \geq a). \end{aligned}$$



Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Proposition 2

Fie X o variabilă aleatoare cu media μ și dispersia σ^2 .

Atunci

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

proof: Considerăm variabila $Y = (X - \mu)^2$ și $a = k^2$ în inegalitatea lui Markov

$$P(|X - \mu| \geq k) = P[(X - \mu)^2 \geq k^2] \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}.$$



Theorem 1.1

Fie $(X_n)_{n \geq 1}$ un șir de variabile aleatoare independente având dispersii finite, uniform mărginite, i. e. $\text{Var}[X_n] \leq c$, pentru orice $n \geq 1$. Atunci

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \right| < \epsilon \right) = 1.$$

proof: Știm că

$$M \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \text{ și } D^2 \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] < \frac{c}{n}.$$

Folosind inegalitatea lui Cebâșev pentru variabila $\frac{1}{n} \sum_{i=1}^n X_i$ obținem

$$1 \geq P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \right| < \epsilon \right) \geq 1 - \frac{D^2 \left[\frac{1}{n} \sum_{i=1}^n X_i \right]}{\epsilon^2} \geq 1 - \frac{c}{n\epsilon^2}.$$

Trecând la limită,

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \right| < \epsilon \right) = 1. \blacksquare$$

- Legea numerelor mari spune că pe măsură ce crește numărul de variabile independente, identic distribuite, media lor de selecție se apropie de media lor comună.

Theorem 1.2

(Legea slabă numerelor mari , legea lui Khintchine) Fie $(X_n)_{n \geq 1}$ un șir de variabile aleatoare independente și identic distribuite cu media μ și dispersia σ^2 . Atunci

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| < \epsilon \right) = 1 \text{ sau } \lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right) = 0$$

proof: O consecință a teoremei 1.1: $\frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu$.



Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Theorem 1.3

(Legea tare numerelor mari) Fie $(X_n)_{n \geq 1}$ un șir de variabile aleatoare independente și identic distribuite cu media μ și dispersia σ^2 . Atunci

$$P \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mu \right) = 1.$$

proof: Fiind mai complicată este omisă.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

- Bernoulli este primul care a demonstrat legea slabă numerelor mari dar doar pentru distribuții Bernoulli.
- Să presupunem că avem o experiență aleatoare și un eveniment aleator asociat A cu $P(A) = p$.
- Repetăm în mod independent experiența și considerăm următorul șir de variabile aleatoare : $X_i = 1$ dacă A se produce la a i -a repetare și 0 altfel.
- Variabilele sunt independente și distribuite Bernoulli cu media p .

- Legea numerelor mari spune că, cu probabilitate 1,

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow p.$$

- $\sum_{i=1}^n X_i$ este numărul de realizări ale evenimentului A în n repetări ale experienței.
- Altfel spus, conform legii numerelor mari, A apare cu frecvența p .

- James Bernoulli A demonstrat legea slabă a numerelor mari în 1700; Poisson i-a generalizat rezultatul în 1800.
- Cebâșev a descoperit inegalitatea care-i poartă numele în 1866, iar Markov a extins rezultatul lui Bernoulli la variabile aleatoare dependente.
- În 1909 Émile Borel a demonstrat teorema care astăzi este cunoscută sub numele de legea tare a numerelor mari (care generalizează o dată în plus teorema lui Bernoulli).
- În 1926 Kolmogorov a obținut o condiție mai generală, suficientă pentru ca un șir de variabile aleatoare independente să respecte legea numerelor mari. Condiția este

$$\sum_{n \geq 1} \frac{\text{Var}[X_n]}{n^2} < +\infty.$$

Theorem 2.1

(Teorema limită centrală, Lindeberg-Lévy) Fie $(X_n)_{n \geq 1}$ un șir de variabile aleatoare independente și identic distribuite cu media μ și dispersia σ^2 . Atunci

$$\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0, 1) \text{ sau}$$

$$\lim_{n \rightarrow \infty} P \left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq a \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a \exp(-t^2/2) dt.$$

- Teorema limită centrală permite estimarea unor probabilități asociate sumelor de variabile (independente și identic distribuite).
- Pe de altă parte, teorema explică de ce atât de multe procese (din științele sociale, biologie, psihologie etc) urmează o lege normală.
- în esența ei teorema limită centrală spune că, pentru eșantioane suficient de mari ($n \geq 30$), variabila

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma/\sqrt{n}}$$

urmează o lege normală standard, $N(0, 1)$.

- Teorema limită centrală are loc chiar și pentru variabile dependente, dacă au corelația foarte scăzută.

Proposition 3

Fie α_n numărul de apariții ale unui eveniment A în n repetări independente ale unei experiențe aleatoare. Dacă $f_n = \frac{\alpha_n}{n}$ este frecvența relativă a apariției lui A , atunci șirul $(f_n)_{n \geq 1}$ converge în probabilitate la $p = P(A)$.

proof: $\alpha_n = nf_n$ este o variabilă distribuită binomial, astfel $\mathbb{E}[\alpha_n] = np$ și $\text{Var}[\alpha_n] = np(1-p)$. Mai mult,

$$\begin{aligned} P(|f_n - p| < \epsilon) &= P(|\alpha_n - np| < n\epsilon) = P(|\alpha_n - \mathbb{E}[\alpha_n]| < n\epsilon) \geq \\ &\geq 1 - \frac{p(1-p)}{n\epsilon^2}. \end{aligned}$$

Evident, trecând la limită, $\lim_{n \rightarrow \infty} P(|f_n - p| < \epsilon) = 1$, pentru orice $\epsilon > 0$. ■

Aproximarea normală a distribuției binomiale

- Fie X_n un șir de variabile Bernoulli(p) independente.
- $X = \sum_{i=1}^n X_i$ are o distribuție binomială, $B(n, p)$.
- Folosind teorema limită centrală obținem teorema de Moivre-Laplace care spune că, pentru n suficient de mare, variabila
$$Y = \frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}[X]}} = \frac{X - np}{\sqrt{np(1-p)}}$$
este o variabilă normală standard ($N(0, 1)$).
- Aproximarea este bună pentru $np(1-p) \geq 10$.

Aproximarea normală a distribuției binomiale

- Un alt mod de a vedea acest rezultat este următorul: când k este aproape de np

$$\binom{n}{k} p^k (1-p)^{n-k} \sim \frac{\exp - \frac{(k-np)^2}{2np(1-p)}}{\sqrt{2\pi np(1-p)}}.$$

- Considerăm următorul exemplu: fie X numărul de apariții ale stemei în 40 de aruncări ale unei monede.

- Cât este $P(X = 20)$?

$$\begin{aligned} P(X = 20) &= P(19.5 \leq X \leq 20.5) = \\ &= P\left(\frac{19.5 - 20}{\sqrt{10}} \leq \frac{X - 20}{\sqrt{10}} \leq \frac{20.5 - 20}{\sqrt{10}}\right) = P\left(-0.16 \leq \frac{X - 20}{\sqrt{10}} \leq 0.16\right) \\ &\sim \Phi(0.16) - \Phi(-0.16) = 0.1272, \end{aligned}$$

unde $\Phi(\cdot)$ este funcția de repartiție a variabilei $N(0, 1)$.

• Corecția continuă este o ajustare care se face ori de câte ori o distribuție discretă este aproximată printr-una continuă.

• $P(X = 10) = P(9.5 \leq X \leq 10.5)$, $P(X > 15) = P(X \geq 15.5)$, $P(X < 13) = P(X \leq 12.5)$.

Generarea de numere aleatoare uniform distribuite

- Când vorbim despre numere aleatoare ne gândim cel mai adesea la numere aleatoare uniform distribuite.
- Există două tipuri de variabile aleatoare uniforme: discretă și continuă.
- De exemplu, pentru a alege un număr întreg aleator uniform distribuit între 1 și n (câteodată între 0 și $(n - 1)$) trebuie să generăm o valoare a unei variabile aleatoare discrete uniforme U_n .
- Pe de altă parte, dacă dorim să alegem un număr aleator uniform din intervalul $[0, 1]$ trebuie să generăm o valoare a unei variabile continue uniforme $U_{[0,1]}$.
- În general, *a simula o anumită variabilă aleatoare înseamnă a genera valori care urmează acea distribuție.*

- Aproape orice limbaj de programare conține câte un generator de numere aleatoare uniforme (discrete și continue); noi vom utiliza generatoarele din R care acoperă și alte distribuții în afară de cele uniforme.
- Trecem în revistă comenzile R pentru distribuțiile discrete sau continue uzuale.
- Funcțiile care încep cu p , q , d și r returnează funcție de repartiție (sau CDF - cumulative distribution function), inversa CDF, function de densitate de probabilitate (PDF), respectiv o valoare a unei variabile aleatoare având distribuția specificată.
- Pentru a genera doar valori discrete uniforme se poate utiliza funcția *sample()*.

| Distribuția | Comenzi | | | |
|--------------------|-----------------|-----------------|-----------------|-----------------|
| Binomial | <i>pbinom()</i> | <i>qbinom()</i> | <i>dbinom()</i> | <i>rbinom()</i> |
| Geometric | <i>pgeom()</i> | <i>qgeom()</i> | <i>dgeom()</i> | <i>rgeom()</i> |
| Poisson | <i>ppois()</i> | <i>qpois()</i> | <i>dpois()</i> | <i>rpois()</i> |
| Uniform | <i>punif()</i> | <i>qunif()</i> | <i>dunif()</i> | <i>runif()</i> |
| Exponential | <i>pexp()</i> | <i>qexp()</i> | <i>dexp()</i> | <i>rexp()</i> |
| Normal | <i>pnorm()</i> | <i>qnorm()</i> | <i>dnorm()</i> | <i>rnorm()</i> |
| Student | <i>pt()</i> | <i>qt()</i> | <i>dt()</i> | <i>rt()</i> |
| Gamma | <i>pgamma()</i> | <i>qgamma()</i> | <i>dgamma()</i> | <i>rgamma()</i> |

- Detalii se pot afla folosind *help(nume)* în R sau Rstudio.

- Pentru a simula o variabilă aleatoare discretă este nevoie de repartiția ei.

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_k & \dots \\ p_1 & p_2 & \dots & p_k & \dots \end{pmatrix}$$

- X se poate simula astfel: generăm o valoare aleatoare uniformă (continuă) U și returnăm x_i dacă $\sum_{j=1}^{i-1} p_j \leq U < \sum_{j=1}^i p_j$.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Exemplul 1. (LNM - Problema acului lui Buffon) Problema (formulaă în 1733 și rezolvată în 1777 de naturalistul și matematicianul francez Comte de Buffon) cere să se determine probabilitatea ca un ac de lungime l să intersecteze o dreaptă când este aruncat pe o suprafață plană pe care sunt desenate (o infinitate de) drepte la distanța $2d$.

Soluție. Vom presupune că lungimea acului este mai mică decât distanța dintre drepte (cea mai ușoară variantă de analizat); există două variabile care determină poziția relativă a acului față de cea mai apropiată dreaptă: unghiul, x , pe care îl face acul cu direcția liniilor și distanța de la mijlocul acului la această dreaptă, y . Acul va intersecta cea mai apropiată linie dacă și numai $y \leq l/2 \sin x$, oricare ar fi $x \in [0, \pi]$.

Probabilități și Statistică

Probabilități și Statistică

Probabilități și Statistică

Toate situațiile posibile sunt complet descrise de perechile $(x, y) \in [0, \pi] \times [0, d]$, iar cazurile favorabile sunt perechile care dau puncte aflate sub graficul funcției $f : [0, \pi] \rightarrow \mathbb{R}$, $f(x) = l/2 \sin x$.

Astfel, probabilitatea este

$$\frac{\int_0^{\pi} f(x) dx}{\pi \cdot d} = \frac{1}{\pi d} \int_0^{\pi} \frac{l}{2} \sin x dx = \frac{l}{2\pi d} [-\cos x]_0^{\pi} = \frac{l}{\pi d}.$$

Pentru $l = d = 1$, adică atunci când acul are lungimea egală cu jumătate din distanța dintre drepte, probabilitatea este $1/\pi$.

Introducem experiența aleatoare care constă în a arunca acul și notăm cu X variabila Bernoulli care are valoarea 1 dacă și numai dacă acul intersectează o dreaptă; probabilitatea de succes și media variabilei X au valoarea $1/\pi$.

Dacă repetăm în mod independent experimentul de n ori vom obține un eșantion de dimensiune n , $(X_i)_{i=1,n}$. Datorită Legii Numerelor Mari $\bar{x}_n \rightarrow 1/\pi$, astfel, pentru valori mari ale lui n ,

$$\bar{x}_n = \frac{\text{numărul de succese}}{n} \approx \frac{1}{\pi}.$$

Acest tip de relație a fost utilizat pentru a obține aproximări experimentale ale lui π . Mai mulți "aruncători" de ace au efectuat deja acest experiment.

Exemplul 2. (*Verificarea LNM*) Considerăm o repartiție X cu media μ și dispersia σ^2 și un șir de n variabile aleatoare independente și identic distribuite cu X , X_i , $i = \overline{1, n}$. Legea numerelor mari spune că (într-un anumit sens, probabilistic) media de selecție converge la medie:

$$\bar{x}_n \rightarrow \mu \text{ as } n \rightarrow \infty$$

Verificăm această lege utilizând o distribuție Poisson cu diverși parametri λ (pentru o astfel de distribuție $\mu = \lambda$).

| λ | 2 | 3 | 4 | 6 | 8 | 12 | 15 |
|-------------|-------|-------|-------|-------|-------|--------|--------|
| \bar{x}_n | 1.955 | 2.977 | 4.003 | 6.027 | 8.018 | 12.093 | 14.925 |

Se observă că mediile de selecție obținute ($n = 5000$) sunt foarte apropiate de mediile corespunzătoare. (Eșantioanele au fost obținute folosind $rpois(n, \lambda)$.)

Dacă repetăm testul anterior cu distribuția Gamma pentru diverse valori ale parametrilor (α, λ) (media este $\mu = \alpha/\lambda$) obținem

| | | | | | | | | |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|
| α | 2 | 2 | 3 | 4 | 6 | 6 | 6 | 12 |
| λ | 1.5 | 2 | 2 | 3 | 5 | 4 | 8 | 4 |
| \bar{x}_n | 1.361 | 1.009 | 1.489 | 1.345 | 1.204 | 1.501 | 0.752 | 2.973 |
| μ | 1.333 | 1.000 | 1.500 | 1.333 | 1.200 | 1.500 | 0.750 | 3.000 |

Mediile de selecție obținute ($n = 5000$) sunt foarte apropiate de mediile corespunzătoare. (Eșantioanele au fost obținute folosind $rpois(n, \lambda)$.)

Exemplul 3. (*TLC - de Moivre-Laplace*) Mărimea ideală a anului I la un colegiu particular este de 150 studenți. Conducerea colegiului, știind din experiență, că, în medie, doar 30% din elevii care trec examenul de admitere vor urma cursurile, aprobă 450 de locuri pentru admitere. Calculați probabilitatea ca cel puțin 151 de elevi admiși să rămână și să urmeze cursurile colegiului.

Soluție. Fie X numărul de elevi admiși care urmează cursurile colegiului; vom presupune că fiecare elev admis va urma independent cursurile colegiului. Atunci $X : B(450, 0.3)$ și

$$P(X > 150) = P(X \geq 150.5) = P\left(\frac{X - np}{\sqrt{np(1-p)}} \geq \frac{150.5 - np}{\sqrt{np(1-p)}}\right) =$$

$$= P\left(\frac{X - 135}{\sqrt{81}} \geq \frac{15.5}{\sqrt{81}}\right) = P(Z \geq 1.722)$$

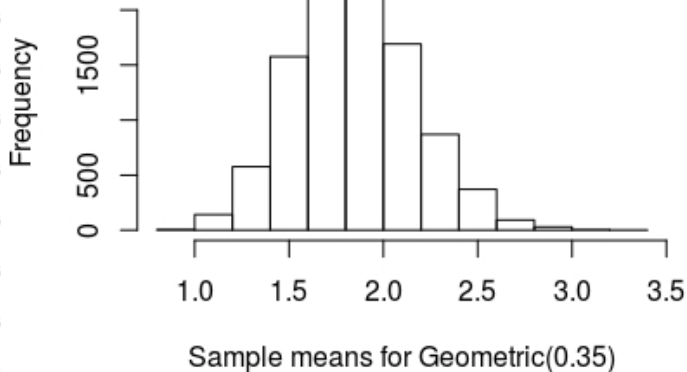
unde $Z : N(0, 1)$. Astfel $P(X \geq 150) \approx 1 - \text{pnorm}(1.722) = 0.0425$.

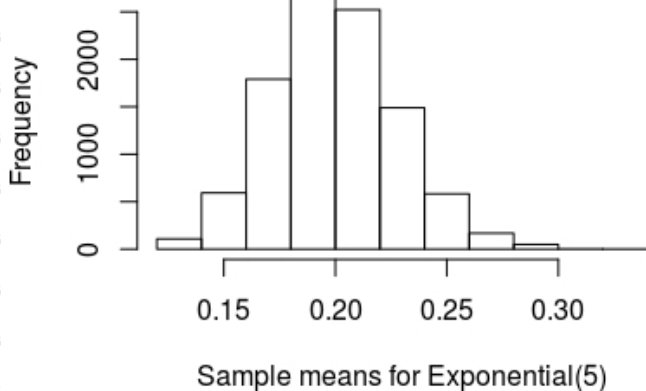
Exemplul 4. (TLC) O populație de muncitori are media greutateii 167 și deviația standard 27. Dacă se alege un eșantion de 36 muncitori, care este probabilitatea ca media de selecție să fie cuprinsă între 163 și 170?

Soluție. Fie \bar{x}_n media de selecție, din TLC, $\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$ urmează cu aproximație o distribuție normală standard, astfel

$$\begin{aligned} P(163 \leq \bar{x}_n \leq 170) &= P\left(\frac{163 - 167}{4.5} \leq \frac{\bar{x}_n - 167}{4.5} \leq \frac{170 - 167}{4.5}\right) = \\ &= P\left(-0.888 \leq \frac{\bar{x}_n - 167}{4.5} \leq 0.888\right) \approx P(-0.888 \leq Z \leq 0.888) = \\ &= pnorm(0.888) - pnorm(-0.888) = 2 \cdot pnorm(0.888) - 1 = 0.625 \end{aligned}$$

Exemplul 5. (Verificarea TLC) Considerăm o distribuție de probabilitate, X , cu media μ și dispersia σ^2 și un șir format din n variabile aleatoare independente și identic distribuite X_i , $i = \overline{1, n}$. Conform TLC, pentru valori mari ale lui n , media de selecție, \bar{x}_n , are o distribuție normală, $N(\mu, \sigma^2)$. Dorim să verificăm această afirmație și considerăm N astfel de medii de selecție și construim o histogramă. Exemplele de mai jos folosesc distribuția geometrică $G(0.35)$ și distribuția exponențială $Exp(5)$ ($n = 50$, $N = 10000$).





Exemplul 6. (Verificarea TLC) Considerăm o distribuție de probabilitate, X , cu media μ și dispersia σ^2 și un șir format din n variabile aleatoare independente și identic distribuite X_i , $i = \overline{1, n}$. Acest șir poate fi văzut ca un eșantion aleator simplu; dacă \bar{x}_n este media de selecție, TLC spune că

$$\lim_{n \rightarrow \infty} P \left[\frac{\bar{x}_n - \mu}{\sigma / \sqrt{n}} \leq z \right] = P(Z \leq z),$$

unde $Z : N(0, 1)$. De obicei, pentru valori mari ale lui n putem face următoarea aproximare

$$P_n(z) = P \left[\frac{\bar{x}_n - \mu}{\sigma / \sqrt{n}} \leq z \right] \approx P(Z \leq z).$$

O metodă pentru a verifica cât de precisă este această aproximare: alegem independent N eșantioane (șiruri) $(X_i^k)_{i=1, n}^{k=1, N}$ și calculăm

$$P^N = \frac{|\{k : \bar{x}_n^k \leq z\sigma / \sqrt{n} + \mu\}|}{N}.$$

Altfel spus, P^N este numărul de eşantioane (dintre cele N) care satisfac inegalitatea $\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \leq z$ supra numărul total de eşantioane. Această statistică ar trebui să aproximeze $P[Z \leq z]$. Pentru distribuția exponențială cu $\lambda = 2$, $n = 50$ și $N = 2000$ rezultatele se găsesc mai jos (un singur eşantion de dimensiune n poate fi obținut cu $\text{rexp}(n, \lambda)$).

| z | -1.5 | -1.0 | -0.5 | 0 | 0.5 | 1.0 | 1.5 |
|------------|-------|-------|-------|-------|-------|-------|-------|
| $P^N(z)$ | 0.055 | 0.154 | 0.313 | 0.509 | 0.723 | 0.831 | 0.931 |
| $Abs.err$ | 16% | 2.5% | 1.6% | 1.8% | 4.6% | 1.8% | 0.2% |
| $pnorm(z)$ | 0.066 | 0.158 | 0.308 | 0.5 | 0.691 | 0.847 | 0.933 |

Eroarea absolută este egală cu $\frac{|P(Z \leq z) - P^N(z)|}{P(Z \leq z)}$.

Pentru a calcula $P^N(z)$ am folosit următorul algoritm:

$\mu \leftarrow 1/\lambda$; // why?

$\sigma \leftarrow 1/\lambda$; // why?

$c \leftarrow z * \sigma / \sqrt{n} + \mu$;

$j \leftarrow 1$;

for($i = 1, N$)

 if($mean(rexp(n, \lambda)) \leq c$)

$j++$;

return j/N ;

Sfârșit

Probabilități și Statistică


Probabilități și Statistică

Probabilități și Statistică


 Baron, M., *Probability and Statistics for Computer Scientist*, Chapman & Hall/CRC Press, 2013 sau ediția electronică <https://ww2.ii.uj.edu.pl/~z1099839/naukowe/RP/rps-michael-byron.pdf>

 Johnosn, J. L., *Probability and Statistics for Computer Science*, Wiley Interscience, 2008.

 Lipschutz, S., *Theory and Problems of Probability*, Schaum's Outline Series, McGraw-Hill, 1965.

 Ross, S. M., *A First Course in Probability*, Prentice Hall, 5th edition, 1998.

 Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.

 Stone, C. J., *A Course in Probability and Statistics*, Duxbury Press, 1996.