# Probabilităţi şi Statistică - Curs 9

Mai, 2018

- În cazul în care $|\Omega| \geqslant |\mathbb{R}|$ (i.e., cardinalul lui $\Omega$ este cel puţin continuu), evenimentele aleatoare se definesc diferit faţă de cazul discret.
- Diferenţa principală constă în aceea că nu orice submulţime $A \subseteq \Omega$ este în mod necesar eveniment aleator: familia evenimentelor aleatoare este o $\sigma$-algebră $\mathcal{A} \subseteq \mathcal{P}(\Omega)$:
  - $\varnothing, \Omega \in \mathcal{A}$;
  - dacă $A_1, A_2 \in \mathcal{A}$, atunci $A_1 \cap A_2 \in \mathcal{A}$;
  - dacă $(A_n)_{n \geqslant 1} \subseteq \mathcal{A}$, atunci $\bigcup_{n \geqslant 1} A_n \in \mathcal{A}$.
- Iar funcţia de probabilitate este definită numai pe submulţimile din $\mathcal{A}$ (cu axiomele cunoscute):

$$P : \mathcal{A} \to [0, 1].$$

- O funcţie $X : \Omega \to \mathbb{R}$ este numită **variabilă aleatoare** dacă

$$\text{pentru orice } J \text{ interval din } \overline{\mathbb{R}},\ X^{-1}(J) \in \mathcal{A}.$$

- O variabilă aleatoare $X : \Omega \to \mathbb{R}$ se numeşte **continuă** dacă are funcţia de repartiţie continuă (*câteodată, această definiţie se referă la toate cazurile când $X(\Omega)$ este de cardinal continuu*).
- Distribuţia (repartiţia) unei astfel de variabile poate fi dată prin **funcţia de repartiţie**:

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

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

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

- Orice funcţie $f:\mathbb{R}\to[0,+\infty)$ cu $\int\limits_{-\infty}^{\infty} f(t)\,dt = 1$ este funcţia de densitate pentru o anumită avariabilă aleatoare continuă (sau, mai simplu, distribuţie continuă).
- Folosind funcţia de densitate putem calcula (dacă integralele există) media şi dispersia:

$$M(X) = \int\limits_{-\infty}^{+\infty} tf(t)\,dt \text{ şi } D^2(X) = \int\limits_{-\infty}^{+\infty} [t - M(X)]^2 f(t)\,dt.$$

- Probabilităţile asociate unei variabile aleatoare continue se calculează astfel:

$$P(a < X \leqslant b) = F(b) - F(a) = \int\limits_{a}^{b} f(t)\,dt,$$

adică aria aflată sub graficul funcţiei $f$ între punctele $t = a$ şi $t = b$.

- Dacă $F$ este continuă, $P(X = a) = F(a) = 0$ şi $P(a \leqslant X < b) = P(a \leqslant X \leqslant b) = P(a < X < b)$.
- Pentru o variabilă aleatoare $X : \Omega \to \mathbb{R}$, dată, operaţia de **standardizare** constă în următoarea transformare a variabilei $X$:

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

- Noua variabilă este "standard" pentru că are

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

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

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

Determinaţi $k$, funcţia de repartiţie şi probabilitatea ca viaţa unei astfel de componente sa depăşească 2 ani.

*Soluţie.* Trebuie ca $f(t) \geqslant 0, \forall t \in \mathbb{R}$ şi $\int\limits_{-\infty}^{\infty} f(t)\,dt = 1$, deci

$k \geqslant 0$ şi $1 = \int\limits_{1}^{\infty} \frac{k}{t^4}\,dt = \left[-\frac{k}{3t^3}\right]_1^{\infty} = \frac{k}{3}$, de unde $k = 3$.

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

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

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

*Exemplul 2.* Fie $X$ o variabilă aleatoare continuă cu următoarea funcţie de densitate:

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

Determinaţi $\alpha$, funcţia de repartiţie, media şi dispersia lui $X$.

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

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

Calculaţi $C$ şi probabilitatea ca timpul de repornire să fie între 1 şi 2 minute.

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

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

Calculaţi $K$, probabilitatea ca o eroare hardware să apară în primii 5 ani şi durata medie de viaţă a acestui HD.

*Distribuţia uniformă.* Se notează cu $U(a, b)$ şi are funcţia de densitate

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

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

$U(0, 1)$ se numeşte *distribuţia uniformă standard*.

*Distribuţia exponenţială.* Se notează cu $Exp(\lambda)$ şi are funcţia de densitate ($\lambda > 0$)

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

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

*Distribuţia exponenţială este utilizată pentru a modela timpul de aşteptare, timpul între două sosiri, durata de viaţă hardware etc; într-o secvenţă de evenimente rare timpul dintre două astfel de evenimente este distribuit exponenţial.*

Distribuţia exponenţială nu are memorie (trecerea a $x_0$ minute nu are relevanţă): chiar dacă $X > x$, când timpul total de aşeptare depăşeşte $x$, timpul rămas are o distribuţie exponenţială: $P(X > x + \Delta x | X > x) = P(X > \Delta x)$ (de ce?).

*Distribuţia gaussiană (normală).* Se notează cu $N(\mu, \sigma^2)$ şi are funcţia de densitate

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

Dacă $X : N(\mu, \sigma^2)$, atunci $\mathbb{E}[X] = \mu$ and $Var[X] = \sigma^2$.
Distribuţia $N(0, 1)$ este numită *distribuţia normală standard* .
Valorile unei variabile normal distribuite au următoarea împrăştiere (simetric în jurul mediei): %68 se găsesc în intervalul $[\mu - \sigma, \mu + \sigma]$, %95 în $[\mu - 2\sigma, \mu + 2\sigma]$, iar %99.7 aparţin intervalului $[\mu - 3\sigma, \mu + 3\sigma]$.

- Distribuţia normală are un rol central în teoria probabilităţilor şi statistică pentru cel puţin două motive.
- Drept o consecinţă a Teoremei Limită Centrale (TLC - vezi mai jos) sumele şi/sau mediile variabilelor independente şi identic distribuite au cu aproximaţie o distribuţie normală.
- De-a lungul timpului s-a observat că distribuţia normală este un model potrivit pentru foarte multe variabile: temperatura, greutatea, înălţimea şi chiar notele studenţilor.
- Distribuţia normală a fost utilizată implicit de către de Moivre pentru aproximarea distribuţiei binomiale şi ulterior de către Laplace şi Gauss (în mod explicit).

*Distribuţia Student (sau t).* Este notată cu $t(r)$ şi are funcţia de densitate

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

unde $\Gamma(y) = \int\limits_0^{+\infty} x^{y-1} e^{-x}\, dx$. Dacă $X : t(r)$,atunci $\mathbb{E}[X] = 0$ and $Var[X] = \dfrac{r}{r-2}$.

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

*Distribuţia Gamma.* Se notează cu $\Gamma(\alpha, \lambda)$ şi are funcţia de densitate

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

unde $\Gamma(t) = \int\limits_0^{+\infty} x^{t-1} e^{-x} \, dx$. $\alpha$ este *forma* iar $\lambda$ este rata (sau frecvenţa) disribuţiei. Pentru $X : \Gamma(\alpha, \lambda)$, avem $\mathbb{E}[X] = \dfrac{\alpha}{\lambda}$ and $Var[X] = \dfrac{\alpha}{\lambda^2}$.

*Să presupunem că avem un proces care constă în $\alpha$ paşi independenţi şi fiecare astfel de pas necesită un timp egal cu $Exp(\lambda)$, atunci durata totală urmează o distribuţie Gamma.*

Astfel, distribuţia *Gamma* este o sumă de $\alpha$ variabile independente identic repartizate exponenţial.

**Proposition 1**

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

$$P(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}.$$

**proof:**

$$\mathbb{E}[X] = \int_0^{+\infty} tf(t)\,dt = \int_0^a tf(t)\,dt + \int_a^{+\infty} tf(t)\,dt \geqslant$$

$$\int_a^{+\infty} tf(t)\,dt \geqslant a\int_a^{+\infty} f(t)\,dt = aP(X \geqslant a).$$

∎

**Proposition 2**

*Fie $X$ o variabilă aleatoare cu media $\mu$ şi dispersia $\sigma^2$. Atunci*

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

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

$$P(|X-\mu| \geqslant k) = P\left[(X-\mu)^2 \geqslant k^2\right] \leqslant \frac{\mathbb{E}\left[(X-\mu)^2\right]}{k^2} = \frac{\sigma^2}{k^2}.$$

■

**Theorem 1.1**

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

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

**proof:** Ştim că

$$M\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \text{ şi}$$

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

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

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

Trecând la limită,

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

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

**Theorem 1.2**

*(Legea slabă numerelor mari , legea lui Khintchine) Fie* $(X_n)_{n\geqslant 1}$ *un şir de variabile aleatoare independente şi identic distribuite cu media* $\mu$ *şi dispersia* $\sigma^2$*. Atunci*

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

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

■

**Theorem 1.3**

*(Legea tare numerelor mari) Fie* $(X_n)_{n \geqslant 1}$ *un şir de variabile aleatoare independente şi identic distribuite cu media* $\mu$ *şi dispersia* $\sigma^2$*. Atunci*

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

**proof:** Fiind mai complicată este omisă. ■

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

- Legea numerelor mari spune că, cu probabilitate 1,

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

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

- James Bernoulli A demonstrat legea slabă a numerelor mari în 1700; Poisson i-a generalizat rezultatul în 1800.
- Cebâşev 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 obţinut o condiţie mai generală, suficientă pentru ca un şir de variabile aleatoare independente să respecte legea numerelor mari. Condiţia este

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

**Theorem 2.1**

*(Teorema limită centrală, Lindeberg-Lévy) Fie* $(X_n)_{n\geqslant 1}$ *un şir de variabile aleatoare independente şi identic distribuite cu media* $\mu$ *şi dispersia* $\sigma^2$. *Atunci*

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

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

- Teorema limită centrală permite estimarea unor probabilităţi asociate sumelor de variabile (independente şi identic distribuite).
- Pe de altă parte, teorema explică de ce atât de multe procese (din ştiinţele sociale, biologie, psihologie etc) urmează o lege normală.
- în esenţa ei teorema limită centrală spune că, pentru eşantioane suficient de mari ($n \geqslant 30$), 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 şi pentru variabile dependente, dacă au corelaţia foarte scăzută.

**Proposition 3**

*Fie $\alpha_n$ numărul de apariţii ale unui eveniment $A$ în $n$ repetări independente ale unei experienţe aleatoare. Dacă $f_n = \dfrac{\alpha_n}{n}$ este frecvenţa relativă a apariţiei lui $A$, atunci şirul $(f_n)_{n\geqslant 1}$ converge în probabilitate la $p = P(A)$.*

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

$$P(|f_n - p| < \epsilon) = P(|\alpha_n - np| < n\epsilon) = P(|\alpha_n - \mathbb{E}[\alpha_n]| < n\epsilon) \geqslant$$

$$\geqslant 1 - \frac{p(1-p)}{n\epsilon^2}.$$

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

## Aproximarea normală a distribuţiei binomiale

- Fie $X_n$ un şir de variabile Bernoulli(p) independente.
- $X = \sum_{i=1}^{n} X_i$ are o distribuţie binomială, $B(n, p)$.
- Folosind teorema limită centrală obţinem teorema de Moivre-Laplace care spune că, pentru $n$ suficient de mare, variabila

$$Y = \frac{X - \mathbb{E}[X]}{\sqrt{Var[X]}} = \frac{X - np}{\sqrt{np(1-p)}}$$

  este o variabilă normală standard ($N(0, 1)$).
- Aproximarea este bună pentru $np(1-p) \geqslant 10$.

- 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 apariţii ale stemei în 40 de aruncări ale unei monede.
- Cât este $P(X=20)$?

$$P(X=20) = P(19.5 \leqslant X \leqslant 20.5) =$$

$$= P\left(\frac{19.5-20}{\sqrt{10}} \leqslant \frac{X-20}{\sqrt{10}} \leqslant \frac{20.5-20}{\sqrt{10}}\right) = P\left(-0.16 \leqslant \frac{X-20}{\sqrt{10}} \leqslant 0.1\right.$$

$$\sim \Phi(0.16) - \Phi(-0.16) = 0.1272,$$

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

- Corecţia continuă este o ajustare care se face ori de câte ori o distribuţie discretă este aproximată printr-una continuă.
- $P(X = 10) = P(9.5 \leqslant X \leqslant 10.5)$, $P(X > 15) = P(X \geqslant 15.5)$, $P(X < 13) = P(X \leqslant 12.5)$.

- Când vorbim despre numere aleatoare ne gândim cel mai adesea la numere aleatoare uniform distribuite.
- Există două tipuri de variabile aleatoare uniforme: discretă şi continuă.
- De exemplu, pentru a alege un număr întreg aleator uniform distribuit între 1 şi $n$ (câteodată între 0 şi $(n-1)$) 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 distrubuţie*.

- Aproape orice limbaj de programare conţine câte un generator de numere aleatoare unifome (discrete şi continue); noi vom utiliza generatoarele din R care acoperă şi alte distribuţii în afară de cele uniforme.
- Trecem în revistă comenzile R pentru distribuţiile discrete sau continue uzuale.
- Funcţiile care încep cu $p, q, d$ şi $r$ returnează funcţie de repartiţie (sau CDF - cumulative distribution function), inversa CDF, function de densitate de probabilitate (PDF), respectiv o valoare a unei variable aleatoare având distribuţia specificată.
- Pentru a genera doar valori discrete uniforme se poate utiliza funcţia $sample()$.

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

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

- Pentru a simula o variabilă aleatoare discretă este nevoie de repartiţia 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) $U$ şi returnăm $x_i$ dacă $\sum_{j=1}^{i-1} p_j \leqslant U < \sum_{j=1}^{i} p_j$.

*Exemplul 1.* (*LNM - Problema acului lui Buffon*) Problema (formulaă în 1733 şi rezolvată în 1777 de naturalistul şi 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 suprafaţă plană pe care sunt desenate (o infinitate de) drepte la distanţa $2d$.

*Soluţie.* Vom presupune că lungimea acului este mai mică decât distanţa dintre drepte (cea mai uşoară variantă de analizat); există două variabile care determină poziţia relativă a acului faţă de cea mai apropiată dreaptă: unghiul, $x$, pe care îl face acul cu direcţia liniilor şi distanţa de la mijlocul acului la această dreaptă, $y$. Acul va intersecta cea mai apropiată linie dacă şi numai $y \leqslant l/2 \sin x$, oricare ar fi $x \in [0, \pi]$.

Toate situaţiile 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 funcţiei $f : [0, \pi] \to \mathbb{R}$, $f(x) = l/2 \sin x$. Astfel, probabilitatea este

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

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

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

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

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

Acest tip de relaţie a fost utilizat pentru a obţine aproximări experimentale ale lui $\pi$. Mai mulţi "aruncători" de ace au efectuat deja acest experiment.

*Exemplul 2.* (*Verificarea LNM*) Considerăm o repartiţie $X$ cu media $\mu$ şi dispersia $\sigma^2$ şi un şir de $n$ variabile aleatoare independente şi identic distribuite cu $X$, $X_i$, $i = \overline{1, n}$. Legea numerelor mari spune că (într-un anumit sens, probabilistic) media de selecţie converge la medie:

$$\overline{x}_n \to \mu \text{ as } n \to \infty$$

Verificăm această lege utilizând o distribuţie Poisson cu diverşi parametri $\lambda$ (pentru o astfel de distribuţie $\mu = \lambda$).

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

Se observă că mediile de selecţie obţinute ($n = 5000$) sunt foarte apropiate de mediile corespunzătoare. (Eşantioanele au fost obţinute folosind $rpois(n, \lambda)$.)

Dacă repetăm testul anterior cu distribuţia Gamma pentru diverse valori ale parametrilor $(\alpha, \lambda)$ (media este$\mu = \alpha/\lambda$) obţinem

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 2 | 2 | 3 | 4 | 6 | 6 | 6 | 12 |
| $\lambda$ | 1.5 | 2 | 2 | 3 | 5 | 4 | 8 | 4 |
| $\overline{x}_n$ | 1.361 | 1.009 | 1.489 | 1.345 | 1.204 | 1.501 | 0.752 | 2.973 |
| $\mu$ | 1.333 | 1.000 | 1.500 | 1.333 | 1.200 | 1.500 | 0.750 | 3.000 |

Mediile de selecţie obţinute ($n = 5000$) sunt foarte apropiate de mediile corespunzătoare. (Eşantioanele au fost obţinute folosind $rpois(n, \lambda)$.)

*Examplul 3.* (*TLC - de Moivre-Laplace*) Mărimea ideală a anului I la un colegiu particular este de 150 studenţi. Conducerea colegiului, ştiind din experienţă, că, în medie, doar 30% din elevii care trec examenul de admitere vor urma cursurile, aprobă 450 de locuri pentru admitere. Calculaţi probabilitatea ca cel puţin 150 de elevi admişi să rămână şi să urmeze cursurile colegiului.

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

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

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

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

*Examplul 4.* (*TLC*) O populaţie de muncitori are media greutăţii 167 şi deviaţia standard 27. Dacă se alege un eşantion de 36 muncitori, care este probabilitatea ca media de selecţie să fie cuprinsă între 163 şi 170?

*Soluţie.* Fie $\overline{x}_n$ media de selecţie, din TLC, $\dfrac{\overline{x}_n - \mu}{\sigma/\sqrt{n}}$ urmează cu aproximaţie o distribuţie normală standard, astfel

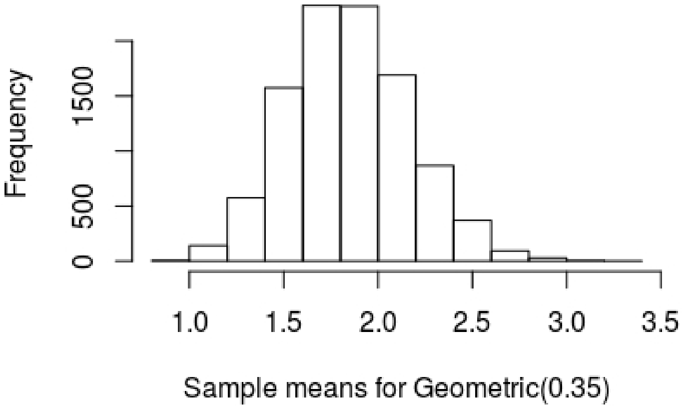
$$P(163 \leqslant \overline{x}_n \leqslant 170) = P\left(\frac{163-167}{4.5} \leqslant \frac{\overline{x}_n - 167}{4.5} \leqslant \frac{170-167}{4.5}\right) =$$

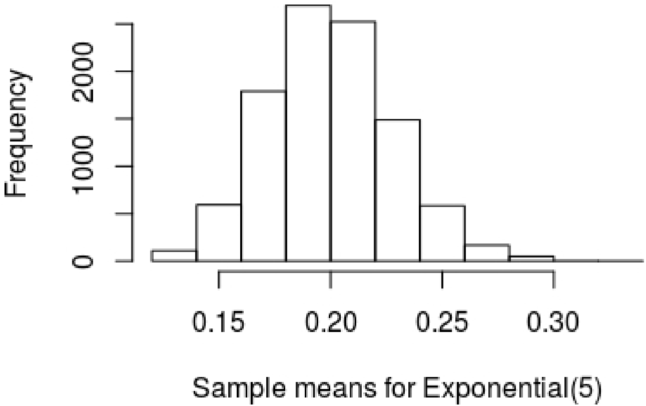
$$= P\left(-0.888 \leqslant \frac{\overline{x}_n - 167}{4.5} \leqslant 0.888\right) \approx P(-0.888 \leqslant Z \leqslant 0.888) =$$

$$= pnorm(0.888) - pnorm(-0.888) = 2 \cdot pnorm(0.888) - 1 = 0.625$$

*Exemplul 5.* (*Verificarea TLC*) Considerăm o distribuţie de probabilitate, $X$, cu media $\mu$ şi dispersia $\sigma^2$ şi un şir format din $n$ variabile aleatoare independente şi identic distribuite $X_i$, $i = \overline{1, n}$. Conform TLC, pentru valori mari ale lui $n$, media de selecţie, $\overline{x}_n$, are o distribuţie normală, $N(\mu, \sigma^2)$.

Dorim să berificăm această afirmaţie şi considerăm $N$ astfel de medii de selecţie şi construim o histogramă. Exemplele de mai jos folosesc distribuţia geometrică $G(0.35)$ şi distribuţia exponenţială $Exp(5)$ ($n = 50$, $N = 10000$).


Frequency
0
500
1500
1.0
1.5
2.0
2.5
3.0
3.5
Sample means for Geometric(0.35)


Frequency
0
1000
2000
0.15
0.20
0.25
0.30
Sample means for Exponential(5)

*Exemplul 6.* (*Verificarea TLC*) Considerăm o distribuţie de probabilitate, $X$, cu media $\mu$ şi dispersia $\sigma^2$ şi un şir format din $n$ variabile aleatoare independente şi identic distribuite $X_i$, $i = \overline{1, n}$. Acest şir poate fi văzut ca un eşantion aleator simplu; dacă $\overline{x}_n$ este media de selecţie, TLC spune că

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

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

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

O methodă pentru a verifica cât de precisă este această aproximare: alegem independent $N$ eşantioane (şiruri) $\left(X_i^k\right)_{i=\overline{1,n}}^{k=\overline{1,N}}$ şi calculăm

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

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

| $z$ | $-1.5$ | $-1.0$ | $-0.5$ | $0$ | $0.5$ | $1.0$ | $1.5$ |
|---|---|---|---|---|---|---|---|
| $P^N(z)$ | 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 $\frac{|P(Z \leqslant z) - P^N(z)|}{P(Z \leqslant z)}$.

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**$(i = \overline{1, N})$

  **if**$(mean(rexp(n, \lambda)) \leqslant c)$

    **j++;**

**return** $j/N;$

# Sfârşit

- Baron, M., *Probability and Statistics for Computer Scientist*, Chapman & Hall/CRC Press, 2013 sau ediţia 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.