CURSUL 14 (10 ianuarie 2022)

Exp: (O margine superioară în inegalitatea lui Chebyshev)

Fie X o v.a., $X \in [a,b]$ de medie μ și variantă σ^2. Atunci din inegalitatea lui Chebyshev avem că

$$P(|X-\mu| \geq c) \leq \sigma^2/c^2, \ \forall c>0$$

Dacă nu cunoaștem σ^2 atunci:

$$P(|X-\mu| \geq c) \leq \frac{(b-a)^2}{4c^2}$$

Vrem să arătăm că

$$\sigma^2 \leq \frac{(b-a)^2}{4}$$

Am văzut că

$$E[(X-\gamma)^2] \geq 0, \ \forall \gamma$$
$$E[X^2] - 2E[X]\gamma + \gamma^2 \geq 0, \ \forall \gamma$$

funcția de gr 2 în γ: $\gamma^2 - 2E[X]\gamma + E[X^2]$ este minimă pt $\gamma = E[X]$

Adică $\sigma^2 = E[(X-E[X])^2] \leq E[(X-\gamma)^2], \ \forall \gamma$

Luăm $\gamma = \frac{a+b}{2}$ și avem că

$$\sigma^2 \leq E\left[\left(X - \frac{a+b}{2}\right)^2\right] = E\left[(X-a)(X-b) + \frac{(b-a)^2}{4}\right]$$

deci

$$\sigma^2 \leq E[(X-a)(X-b)] + \frac{(b-a)^2}{4} \leq \frac{(b-a)^2}{4}$$

pt că $X \in [a,b]$ prin urmare $(X-a)(X-b) \leq 0$ de unde $E[(X-a)(X-b)] \leq 0$.

Teoreme limită

Legea Nr. Mari și Teorema Limită Centrală

① Legea Numerelor Mari

Def: (Convergența aproape sigură: a.s.) opțional

Fie $(X_n)_{n \geq 1}$ un șir de v.a. (independenți și identic repartizate) (i.i.d.) și X o v.a.

Spunem că șirul $(X_n)_{n\geq 1}$ converge aproape sigur la X, notăm $X_n \xrightarrow{a.s.} X$, dacă

$$P(\lim_n X_n = X) = 1$$

Ce înseamnă?

Ev. $\{\omega \in \Omega \mid \lim X_n(\omega) = X(\omega)\}$ este un eveniment de proba. 1. Altfel spus avem convergență punctuală (în sensul convergenței de la funcții) în toate punctele din Ω cu excepția eventual a unei mulțimi de probabilitate $=0$.

Def (Convergența în probabilitate: $\xrightarrow{P}$)

Fie $(X_n)_n$ un șir de v.a. și X o v.a.

Spunem că șirul $(X_n)_n$ converge în probabilitate la X, notăm $X_n \xrightarrow{P} X$, dacă $\forall \varepsilon > 0$ avem

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

Ce înseamnă?

$\forall \varepsilon > 0$ și $\forall \delta > 0$, $\exists n_0$ a.î.

$$P(|X_n - X| \geq \varepsilon) \leq \delta, \ \forall n \geq n_0$$

Dacă ne referim la ε ca fiind nivelul de acuratețe al aproximării și δ ca fiind nivelul de încredere atunci definiția converg. în probabilitate: pt orice nivel de acuratețe și încredere dat, X_n va fi egal cu X, între aceste nivele de acuratețe și încredere, dat fiind că n este suficient de mare.

Exp: $(X_n)_n$ un șir de v.a repartizate $\mathcal{U}([0,1])$, independente. Fie $Y_n = \min\{X_1, X_2, \dots X_n\}$

Vrem să arătăm că $Y_n \xrightarrow{P} 0$

Fie $\varepsilon > 0$ și calculăm

$$P(|Y_n - 0| \geq \varepsilon) = P(|Y_n| \geq \varepsilon) \overset{(Y_n \geq 0)}{=} P(Y_n \geq \varepsilon)$$
$$= P(X_1 \geq \varepsilon, X_2 \geq \varepsilon, \dots X_n \geq \varepsilon)$$
$$\overset{indep}{=} P(X_1 \geq \varepsilon) P(X_2 \geq \varepsilon) \cdot \ldots \cdot P(X_n \geq \varepsilon)$$
$$= (1 - P(X_1 < \varepsilon))(1 - P(X_2 < \varepsilon)) \dots (1 - P(X_n < \varepsilon))$$
$$\overset{X_i \sim \mathcal{U}([0,1])}{=} (1-\varepsilon)\cdot(1-\varepsilon)\cdot(1-\varepsilon) \dots (1-\varepsilon)$$
$$= (1-\varepsilon)^n$$

Pentru $\varepsilon > 0$, $1-\varepsilon < 1$ astfel

$$\lim_{n\to\infty} (1-\varepsilon)^n = 0$$

de unde

$$\lim_{n\to\infty} P(|Y_n - 0| \geq \varepsilon) = 0 \Rightarrow Y_n \xrightarrow{P} 0$$

Def: Numim eșantion de volum n din populația Q un șir de v.a $X_1, X_2, \dots X_n$ indep. și identic repartizate cu repartiția comună Q. $(P \circ X_i^{-1} = Q)$.

Se numește media eșantionului (media empirică) v.a. $\overline{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$

Dacă $X_1, X_2, \dots X_n$ sunt v.a. i.i.d de medie μ și variantă $\sigma^2 < \infty$ atunci

$$E[\overline{X}_n] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{1}{n}(\underbrace{E[X_1]}_{\mu} + \dots + \underbrace{E[X_n]}_{\mu})$$
$$= \frac{1}{n} n\mu = \mu$$

iar

$$Var(\overline{X}_n) = Var\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} Var(X_1 + \dots + X_n)$$
$$\overset{indep}{=} \frac{1}{n^2}(\underbrace{Var(X_1)}_{\sigma^2} + \dots + \underbrace{Var(X_n)}_{\sigma^2}) = \frac{1}{n^2}(n\sigma^2) =$$
$$= \sigma^2/n$$

T (Legea Numerelor Mari versiunea slabă) LNM

Fie $(X_n)_n$ un șir de v.a iid cu media $\mu < \infty$

Atunci

$$\boxed{\overline{X}_n \xrightarrow{P} \mu = E[X_1]}$$

Dem: Pt orice $\varepsilon > 0$

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \xrightarrow[n\to\infty]{} 0$$

Fie $\varepsilon > 0$, atunci din ineg lui Chebyshev:

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{Var(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \xrightarrow[n\to\infty]{} 0$$

Obs: Repartiția lui $\overline{X}_n$ este concentrată în jurul lui μ pt valori mari ale lui n

$$P([\mu - \varepsilon, \mu + \varepsilon] \ni \overline{X}_n) \xrightarrow[n\to\infty]{} 1$$

T (Legea Numerelor Mari - versiunea tare)

Fie $(X_n)_n$ un șir de v.a iid cu media $\mu < \infty$

Atunci

$$\boxed{\overline{X}_n \xrightarrow{a.s.} \mu = E[X_1]}$$

i.e.

$P(\overline{X}_n \to \mu) = 1$ adică $\Omega_0 = \{\omega \in \Omega \mid \overline{X}_n(\omega) \to \mu\}$

atunci $P(\Omega_1) = 1$

Exp: (Legătura dintre probabilitate și frecvența relativă)

Fie A un ev cu $P(A) = p$. Să considerăm n repetiții independente ale experimentului și să notăm cu $X_i = \begin{cases} 1, & \omega \in A \\ 0, & \text{altfel} \end{cases}$ (în experimentul i ev A s-a realizat)

$\overline{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ - frecvența relativă de realizare a ev A în cele n repetiții ale experimentului

$E[X_1] = p$ $(X_i \sim B(p))$

Din LNM: $\forall \varepsilon > 0$

$$P(|\overline{X}_n - \mu| > \varepsilon) = P(|\overline{X}_n - p| \geq \varepsilon) \xrightarrow[n\to\infty]{} 0$$

Interpretarea frecvenționistă:

proba. unui ev (p) $\simeq$ frecv. relativă a ev într-un nr. mare de exp.

Exp: (Campanie de vot)

La alegerile prezidențiale avem mai mulți candidați și fie p procentul alegătorilor din populația cu drept de vot care votează pt candidatul A. Întrebăm n persoane "alese la întâmplare" cu cine au votat și înregistrăm frecvența relativă raportată la candidatul A.

"ales la întâmplare" - am ales uniform și independent

Putem interpreta răspunsul fiecărei persoane ca fiind v.a de tip Bernoulli $X_i \sim B(p)$, X_i sunt independente.

$X_1, X_2, \dots X_n \sim B(p)$ independente

$$\overline{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Ineg. lui Cebyshev:

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{Var(\overline{X}_n)}{\varepsilon^2}$$

în cazul nostru $\mu = E[X_1] = p$

$$\frac{\sigma^2}{n} = Var(\overline{X}_n) = \frac{Var(X_1)}{n} = \frac{p(1-p)}{n}$$

$$P(|\overline{X}_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2}$$

În realitate nu cunoaștem val lui p dar $p(1-p) \leq 1/4$ astfel

$$P(|\overline{X}_n - p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$$

Eg. $n = 100$, $\varepsilon = 0.1$ atunci

$$P(|\overline{X}_{100} - p| \geq 0.1) \leq \frac{1}{400 \cdot 0.1^2} = 0.25$$

Câte persoane trebuie să intervievăm pt o acuratețe de 0.01 și un grad de încredere de 95% că estimarea noastră $(\overline{X}_n)$ se va afla la 0.01 față de adevărata valoare p.

$$P(|\overline{X}_n - p| \geq 0.01) \leq \frac{1}{4n(0.01)^2} \leq 1 - 0.95 = 0.05$$

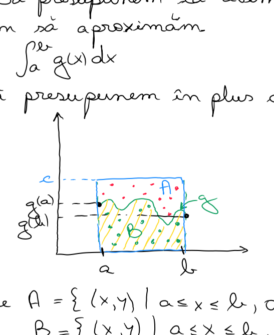
$\Rightarrow n \geq 50\,000$

Exp: (Integrarea de tip Monte-Carlo)

Să presupunem că avem o funcție g și vrem să aproximăm

$$\int_a^b g(x)\,dx$$

Să presupunem în plus că $0 \leq g(x) \leq c$.



Fie $A = \{(x,y) \mid a \leq x \leq b, 0 \leq y \leq c\}$

$B = \{(x,y) \mid a \leq x \leq b, 0 \leq y \leq g(x)\}$

Ideea: Să generăm puncte repartizate uniform pe A și să calculăm proporția punctelor care au căzut în B.

Fie $(X_1,Y_1), (X_2,Y_2), \dots (X_n,Y_n) \sim \mathcal{U}(A)$ (rep uniform pe A)

Cum le generăm?

generăm $\left.\begin{matrix} X \sim \mathcal{U}([a,b]) \\ Y \sim \mathcal{U}([0,c]) \end{matrix}\right\}$ indep $\rightarrow (X,Y) \sim \mathcal{U}(A)$

Densitatea $(X,Y) \sim \mathcal{U}(A)$:

$$f_{(X,Y)}(x,y) = \begin{cases} \frac{1}{\mathcal{A}(A)} = \frac{1}{c(b-a)}, & (x,y) \in A \\ 0, & \text{altfel} \end{cases}$$

Fie $J_j = \begin{cases} 1, & (X_j, Y_j) \in B \\ 0, & \text{altfel} \end{cases}$ atunci $J_1, J_2, \dots J_n$ sunt independente și identic repartizate cu rep comună $B(p)$ unde $p = P(J_j = 1)$

Avem că

$$p = P((X_j, Y_j) \in B) = \frac{\mathcal{A}(B)}{\mathcal{A}(A)} = \frac{\int_a^b g(x)\,dx}{c(b-a)}$$

Aplicăm LNM:

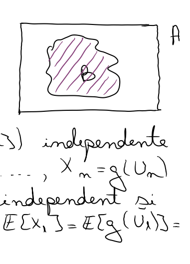
$$\overline{J}_n = \frac{J_1 + J_2 + \dots + J_n}{n} \xrightarrow{P} E[J_1] = p$$

adică

$$\boxed{\int_a^b g(x)\,dx \simeq c(b-a) \frac{1}{n} \sum_{j\geq 1}^{n} J_j}$$

Obs: $P((X,Y) \in B) = \int_B f_{(X,Y)}(x,y)\,dx\,dy =$

$$= \int_B \frac{1}{\mathcal{A}(A)} 1_A(x,y)\,dx\,dy = \frac{\mathcal{A}(B)}{\mathcal{A}(A)}$$



(Metoda alternativă)

Vreau $\int_a^b g(x)\,dx = ?$

Fie $U_1, U_2, \dots U_n \sim \mathcal{U}([a,b])$ independente și fie $X_1 = g(U_1), X_2 = g(U_2), \dots, X_n = g(U_n)$

Atunci șirul $(X_n)_n$ este independent și identic repartizat cu media $E[X_1] = E[g(U_1)] =$

$$= \int g(u) \cdot f_{U_1}(u)\,du$$
$$= \int g(u) \cdot \frac{1}{b-a} 1_{[a,b]}(u)\,du = \int_a^b \frac{1}{b-a} g(u)\,du$$
$$= \frac{1}{b-a} \int_a^b g(u)\,du$$

Aplicăm LNM șirul $(X_n)_n$:

$$\overline{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{g(U_1) + \dots + g(U_n)}{n} \xrightarrow{P} E[g(U_1)] = \frac{1}{b-a}\int_a^b g(u)\,du$$

Astfel putem aproxima:

$$\int_a^b g(u)\,du \simeq (b-a) \frac{g(U_1) + \dots + g(U_n)}{n}$$

② Teorema limită centrală

$X_1, X_2, \dots$ un șir de v.a iid cu media $\mu < \infty$ atunci din LMN: $\overline{X}_n = \frac{X_1 + \dots + X_n}{n} \xrightarrow{P} \mu$

Definim v.a.

$$Z_n = \frac{(X_1 + \dots + X_n) - E[X_1 + \dots + X_n]}{\sqrt{Var(X_1 + \dots + X_n)}}$$ - variabilă normalizată (z-scor)

Avem că $E[Z_n] = 0$, iar $Var(Z_n) = 1$

$$Z_n = \frac{\overline{X}_n - E[\overline{X}_n]}{\sqrt{\frac{Var(X_1 + \dots + X_n)}{n^2}}} = \frac{\overline{X}_n - \mu}{\sqrt{\frac{n\sigma^2}{n^2}}} = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

$$\boxed{Z_n = \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right)}$$ μ = media lui X_1, σ^2 = var. lui X_1

T (Teorema Limită Centrală)

Fie $X_1, X_2 \dots$ un șir de v.a iid cu media $\mu < \infty$ și varianța $\sigma^2 < \infty$.

Atunci are loc

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z), \ \forall z \in \mathbb{R}$$

unde $Z_n = \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right)$ iar $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\,dx$

(... de rep. a $N(0,1)$)

În practică, dacă notăm cu $S_n = X_1 + X_2 + \dots + X_n$, cu X_i iid de medie μ și variantă $\sigma^2 < \infty$ atunci pt n suficient de mare

$$P(S_n \leq c) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right)$$

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} \simeq Z_n \sim N(0,1)$$

$$\simeq P\left(Z \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right) \text{ unde } Z \sim N(0,1)$$
$$= \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right) \quad (\text{pnorm}(\cdot) = \Phi(\cdot))$$

Astfel pt n suficient de mare

$$\boxed{S_n \simeq N(n\mu, n\sigma^2)}$$

sau

$$\boxed{\overline{X}_n \simeq N(\mu, \sigma^2/n)}$$

Exp: Încărcăm un avion cu 100 de pachete a căror greutate este o v.a rep $\mathcal{U}([5,50])$ Kg. Care este probabilitatea ca greutatea totală să depășească 3000 Kg?

$X_1, X_2, \dots X_{100}$ - sunt greutățile celor 100 de pachete cu $X_i \sim \mathcal{U}([5,50])$ independente

Greutatea totală:

$$S_{100} = X_1 + X_2 + \dots + X_{100}$$

Vrem $P(S_{100} > 3000) = ?$

$$E[S_{100}] = 100\,E[X_1] = 100 \cdot \frac{50+5}{2} = 27.5 \times 100 = 2750 \text{ Kg}$$
$$Var(S_{100}) = 100\,Var(X_1) = 100 \cdot \frac{(50-5)^2}{12}$$

$$P(S_{100} > 3000) = P\left(\frac{S_{100} - E[S_{100}]}{\sqrt{Var(S_{100})}} > \frac{3000 - E[S_{100}]}{\sqrt{Var(S_{100})}}\right)$$

Din TLC $\frac{S_{100} - E[S_{100}]}{\sqrt{Var(S_{100})}} \simeq Z \sim N(0,1)$

$$\simeq P\left(Z > \frac{3000 - 2750}{\sqrt{100 \cdot \frac{45^2}{12}}}\right) = P\left(Z > \frac{250}{\frac{10 \cdot 45}{2\sqrt{3}}}\right)$$
$$= 1 - \Phi\left(\frac{10\sqrt{3}}{9}\right) \quad \left(\text{în R } 1 - \text{pnorm}\left(\frac{10\sqrt{3}}{9}\right)\right)$$
$$\simeq 0.0274$$

Exp: (Campania de votare - cont)

p - prop din populație care votează cu A

$X_1, X_2, \dots X_n \sim B(p)$

$\overline{X}_n$ - frecvența relativă a rasp. ce vizează candidatul A în cei n alegători

Vrem să investigăm

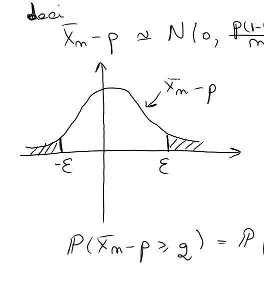
$$P(|\overline{X}_n - p| \geq \varepsilon)$$

Din TLC avem că pt n suficient de mare

$$\overline{X}_n \simeq N(\mu, \sigma^2/n) = N\left(\underset{\mu}{p}, \frac{\overset{\sigma^2}{p(1-p)}}{n}\right)$$

deci

$$\overline{X}_n - p \simeq N\left(0, \frac{p(1-p)}{n}\right)$$



Cum normala este simetrică avem că

$$P(|\overline{X}_n - p| \geq \varepsilon) \simeq 2P(\overline{X}_n - p \geq \varepsilon)$$

$$P(\overline{X}_n - p \geq \varepsilon) = P\Bigg(\underbrace{\frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}}}_{\sim N(0,1)} \geq \frac{\varepsilon}{\sqrt{\frac{p(1-p)}{n}}}\Bigg) \simeq 1 - \Phi\left(\frac{\varepsilon}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

Avem că p este necunoscut dar $p(1-p) \leq 1/4$ astfel

$$\frac{p(1-p)}{n} \leq \frac{1}{4n} \Rightarrow \frac{\varepsilon}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{\varepsilon}{\sqrt{1/4n}} = 2\varepsilon\sqrt{n}$$

Φ este crescătoare avem că

$$1 - \Phi\left(\frac{\varepsilon}{\sqrt{\frac{p(1-p)}{n}}}\right) \leq 1 - \Phi(2\varepsilon\sqrt{n})$$

$$\Rightarrow P(\overline{X}_n - p \geq \varepsilon) \leq 1 - \Phi(2\varepsilon\sqrt{n})$$
$$\Rightarrow P(|\overline{X}_n - p| \geq \varepsilon) \leq 2(1 - \Phi(2\varepsilon\sqrt{n}))$$

Pentru $n = 100$, $\varepsilon = 0.1$ avem

$$P(|\overline{X}_n - p| \geq 0.1) \leq 2(1 - \Phi(2)) = 0.046$$

În mod similar găsim că pt un nivel de acuratețe $\varepsilon = 0.01$ și un grad de încredere de 95%, adică

$$P(|\overline{X}_n - p| \leq 0.01) \geq 0.95$$

găsim că $n \geq 9600$

Exp: (Aproximarea binomialei prin normală)

— Aproximarea de Moivre - Laplace —

Fie $X_1, X_2, \dots \sim B(p)$ indep

$$S_n = X_1 + X_2 + \dots + X_n \sim B(n,p)$$

$E[X_1] = p$

$Var(X_1) = p(1-p)$

Vrem să aproximăm $P(k \leq S_n \leq l) = ?$

Din TLC:

$$P(k \leq S_n \leq l) = P\left(\frac{k - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{l - np}{\sqrt{np(1-p)}}\right)$$

Din TLC avem că

$$\frac{S_n - np}{\sqrt{np(1-p)}} \simeq N(0,1) \text{ deci}$$

$$P(k \leq S_n \leq l) \simeq \Phi\left(\frac{l - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right)$$

Observăm că pt $k = l$ avem

$P(S_n = k) \simeq 0$ ceea ce în general este fals

Sugerăm urm. corecție de continuitate:

$$P(S_n = k) \simeq P(k - 1/2 \leq S_n \leq k + 1/2)$$

Avem că:

$$P(k \leq S_n \leq l) \simeq \Phi\left(\frac{l + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}}\right)$$

- aproximarea normală a binomialei cu factor de corecție

$n = 36$, $p = 0.5$, $S_n \sim B(36, 0.5)$

$$P(S_n \leq 21) = \binom{36}{21} 0.5^{21} \times 0.5^{15} = 0.8785 \quad (\text{pbinom}(\cdot))$$

aproximăm

$\rightarrow \Phi\left(\frac{21 - 36 \cdot 0{,}5}{\sqrt{36 \times 0.5 \times 0.5}}\right) = 0.8413$ (fără factor corecție)

$\rightarrow \Phi\left(\frac{21{,}5 - 36 \times 0{,}5}{\sqrt{36 \times 0{,}5 \times 0{,}5}}\right) = 0.879$ (cu factor corecție)