

Curs 6

Cristian Niculescu

1 Media, dispersia și deviația standard pentru variabile aleatoare continue

1.1 Scopurile învățării

1. Să poată să calculeze și să interpreteze media, dispersia și deviația standard pentru variabile aleatoare continue.
2. Să poată să calculeze și să interpreteze quantilele pentru variabile aleatoare continue sau discrete.

1.2 Introducere

Până acum am studiat media, deviația standard și dispersia pentru variabile aleatoare discrete. Aceste statistici de rezumat au același înțeles pentru variabile aleatoare continue:

- Media $\mu = E(X)$ este o măsură a tendinței centrale.
- Deviația standard σ este o măsură a împrăstierii.
- Dispersia $\sigma^2 = Var(X)$ este pătratul deviației standard.

Pentru a trece de la discret la continuu, pur și simplu înlocuim sumele din formule cu integrale. Un alt tip de [statistică de rezumat](#) sunt [quantilele](#). 0.5 quantila unei repartiții se mai numește [mediana](#) sau a 50-a percentilă.

1.3 Media unei variabile aleatoare continue

Definiție: Fie X o variabilă aleatoare continuă cu domeniul de valori $[a, b]$ (cu convenția că, dacă $a = -\infty$ sau $b = \infty$, intervalul este deschis în acel capăt) și funcția densitate de probabilitate $f(x)$. [Media](#) lui X este

$$E(X) = \int_a^b x f(x) dx.$$

Să vedem cum se compară aceasta cu formula pentru o variabilă aleatoare discretă:

$$E(X) = \sum_{i=1}^n x_i p(x_i).$$

Formula discretă spune să luăm o sumă ponderată a valorilor x_i ale lui X , unde ponderile sunt probabilitățile $p(x_i)$. Reamintim că $f(x)$ este o **densitate** de probabilitate. Unitățile ei de măsură sunt prob/(unitatea de măsură a lui X). Deci $f(x)dx$ reprezintă probabilitatea că X este într-un domeniu de valori infinitesimal de lățime dx în jurul lui x . Astfel, putem interpreta formula pentru $E(X)$ ca o integrală ponderată de valorile x ale lui X , unde ponderile sunt probabilitățile $f(x)dx$.

1.3.1 Exemple

Exemplul 1. Fie $X \sim U(0, 1)$. Aflați $E(X)$.

Răspuns: X are domeniul de valori $[0, 1]$ și densitatea $f(x) = 1$. De aceea,

$$E(X) = \int_0^1 x f(x) dx = \int_0^1 x dx = \left. \frac{x^2}{2} \right|_0^1 = \frac{1}{2}.$$

Media este la mijlocul domeniului de valori.

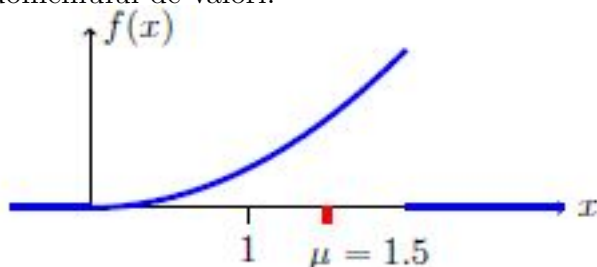
Exemplul 2. Fie X cu domeniul de valori $[0, 2]$ și densitatea $f(x) = \frac{3}{8}x^2$. Aflați $E(X)$.

Răspuns:

$$E(X) = \int_0^2 x f(x) dx = \int_0^2 \frac{3}{8} x^3 dx = \left. \frac{3x^4}{32} \right|_0^2 = \frac{3}{2}.$$

Are sens că X are media în jumătatea dreaptă a domeniului lui de valori?

Răspuns: Da. Deoarece densitatea de probabilitate crește când x crește în domeniul de valori, media lui X ar trebui să fie în jumătatea dreaptă a domeniului de valori.



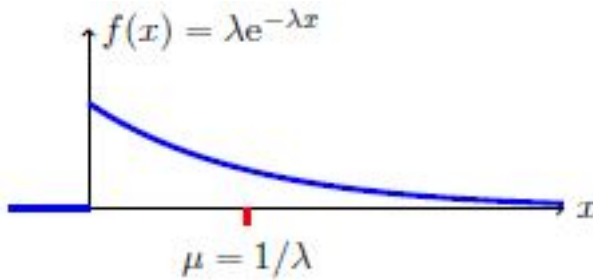
μ este "tras" la dreapta mijlocului 1 deoarece este mai multă masă la dreapta.

Exemplul 3. Fie $X \sim \exp(\lambda)$. Aflați $E(X)$.

Răspuns: Domeniul de valori al lui X este $[0, \infty)$ și pdf este $f(x) = \lambda e^{-\lambda x}$.
Deci

$$\begin{aligned} E(X) &= \int_0^\infty x f(x) dx = \int_0^\infty \lambda x e^{-\lambda x} dx = \int_0^\infty x (-e^{-\lambda x})' dx \\ &= -x e^{-\lambda x} \Big|_0^\infty + \int_0^\infty x' e^{-\lambda x} dx = 0 + \int_0^\infty e^{-\lambda x} dx = -\frac{e^{-\lambda x}}{\lambda} \Big|_0^\infty = \frac{1}{\lambda}. \end{aligned}$$

Am folosit integrarea prin părți și faptul că $\lim_{x \rightarrow \infty} x e^{-\lambda x} = \lim_{x \rightarrow \infty} e^{-\lambda x} = 0$.



Media unei variabile aleatoare exponențiale

Exemplul 4. Fie $Z \sim N(0, 1)$. Aflați $E(Z)$.

Răspuns: Domeniul de valori al lui Z este \mathbb{R} și pdf este $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$.
Avem $E(Z) = \int_{-\infty}^\infty z \phi(z) dz$. Arătăm că integrala converge, i.e. media chiar există (unele variabile aleatoare nu au medie).

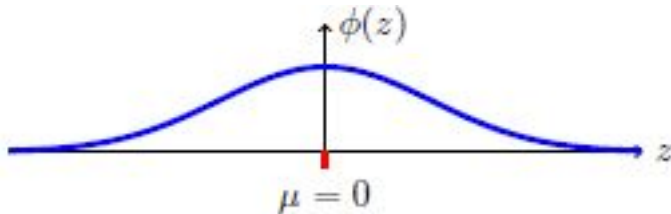
$$\int_0^\infty z \phi(z) dz = \frac{1}{\sqrt{2\pi}} \int_0^\infty z e^{-z^2/2} dz.$$

Substituția $u = z^2/2$ dă $du = z dz$. Deci integrala devine

$$\frac{1}{\sqrt{2\pi}} \int_0^\infty z e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-u} du = -\frac{1}{\sqrt{2\pi}} e^{-u} \Big|_0^\infty = \frac{1}{\sqrt{2\pi}}.$$

Analog, $\int_{-\infty}^0 z \phi(z) dz = -\frac{1}{\sqrt{2\pi}}$.

Deci $E(Z) = \int_{-\infty}^\infty z \phi(z) dz = \int_{-\infty}^0 z \phi(z) dz + \int_0^\infty z \phi(z) dz = -\frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}} = 0$.



Repartiția normală standard este simetrică și are media 0.

1.3.2 Proprietățile lui $E(X)$

Proprietățile lui $E(X)$ pentru variabile aleatoare continue sunt aceleași ca pentru cele discrete:

1. Dacă X și Y sunt variabile aleatoare continue pe un spațiu al probelor Ω , atunci

$$E(X + Y) = E(X) + E(Y). \quad (\text{liniaritate I})$$

2. Dacă a și b sunt constante, atunci

$$E(aX + b) = aE(X) + b. \quad (\text{liniaritate II})$$

Exemplul 5. Verificați că pentru $X \sim N(\mu, \sigma^2)$ avem $E(X) = \mu$.

Răspuns: În exemplul 4 am arătat că pentru Z normală standard, $E(Z) = 0$. Am putea mima calculul de acolo pentru a arăta că $E(X) = \mu$. În loc de aceasta folosim proprietățile de liniaritate ale lui $E(X)$. La manipularea variabilelor aleatoare am arătat că, dacă $X \sim N(\mu, \sigma^2)$, o putem **standardiza**:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

De aici avem $X = \sigma Z + \mu$. Liniaritatea mediei dă

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \sigma \cdot 0 + \mu = \mu.$$

1.3.3 Media funcțiilor de X

Aceasta merge exact ca în cazul discret. Dacă h este o funcție continuă, atunci $Y = h(X)$ este o variabilă aleatoare și

$$E(Y) = E(h(X)) = \int_{-\infty}^{\infty} h(x)f_X(x)dx.$$

Exemplul 6. Fie $X \sim \exp(\lambda)$. Aflați $E(X^2)$.

Răspuns. Integrăm prin părți de 2 ori:

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2 f(x) dx = \int_0^{\infty} \lambda x^2 e^{-\lambda x} dx = \int_0^{\infty} x^2 (-e^{-\lambda x})' dx \\ &= -x^2 e^{-\lambda x} \Big|_0^{\infty} + \int_0^{\infty} 2x e^{-\lambda x} dx = 0 + \int_0^{\infty} 2x \left(-\frac{e^{-\lambda x}}{\lambda} \right)' dx \\ &= -2x \frac{e^{-\lambda x}}{\lambda} \Big|_0^{\infty} + 2 \int_0^{\infty} \frac{e^{-\lambda x}}{\lambda} dx = 0 - 2 \frac{e^{-\lambda x}}{\lambda^2} \Big|_0^{\infty} = \frac{2}{\lambda^2}. \end{aligned}$$

1.4 Dispersia

Definiția dispersiei este identică cu cea pentru variabile aleatoare discrete.

Definiție: Fie X o variabilă aleatoare continuă cu media μ . Dispersia lui X este

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

1.4.1 Proprietăți ale dispersiei

Acestea sunt exact aceleași ca în cazul discret.

1. Dacă X și Y sunt **independente**, atunci $Var(X + Y) = Var(X) + Var(Y)$.
2. Pentru constantele a și b , $Var(aX + b) = a^2 Var(X)$.
3. **Teoremă:** $Var(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$.

Proprietatea 3 dă o formulă pentru $Var(X)$ care este adesea mai ușor de folosit la calcule. Demonstrațiile proprietăților 2 și 3 sunt analoge celor din cazul discret.

Exemplul 7. Fie $X \sim U(0, 1)$. Aflați $Var(X)$ și σ_X .

Răspuns: În exemplul 1 am aflat $\mu = 1/2$.

$$\begin{aligned} Var(X) &= E((X - \mu)^2) = \int_0^1 (x - 1/2)^2 dx = \frac{(x - 1/2)^3}{3} \Big|_0^1 = \frac{1}{12}. \\ \sigma_X &= \sqrt{Var(X)} = \frac{1}{\sqrt{12}} = \frac{1}{2\sqrt{3}} = \frac{\sqrt{3}}{6}. \end{aligned}$$

Exemplul 8. Fie $X \sim exp(\lambda)$. Aflați $Var(X)$ și σ_X .

Răspuns: În exemplele 3 și 6 am calculat:

$$E(X) = \frac{1}{\lambda} \text{ și } E(X^2) = \frac{2}{\lambda^2}.$$

Deci, din proprietatea 3,

$$Var(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \text{ și } \sigma_X = \sqrt{Var(X)} = \frac{1}{\lambda}.$$

Puteam să calculăm și direct din $Var(X) = \int_0^\infty (x - 1/\lambda)^2 \lambda e^{-\lambda x} dx$.

Exemplul 9. Fie $Z \sim N(0, 1)$. Arătați că $Var(Z) = 1$.

Răspuns: Deoarece $E(Z) = 0$, avem

$$\begin{aligned} Var(Z) &= E(Z^2) - E(Z)^2 = E(Z^2) - 0^2 = E(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z(-e^{-z^2/2})' dz = \frac{1}{\sqrt{2\pi}} \left(-ze^{-z^2/2} \Big|_{-\infty}^{\infty} \right) + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz. \end{aligned}$$

Primul termen este 0 deoarece exponențiala tinde la 0 mai repede decât z tinde la $\pm\infty$. Al 2-lea termen este 1 deoarece este exact integrala probabilității totale a pdf $\phi(z)$ pentru $N(0, 1)$. Deci $Var(Z) = 1$.

Exemplul 10. Fie $X \sim N(\mu, \sigma^2)$. Arătați că $Var(X) = \sigma^2$.

Răspuns: Făcând schimbarea de variabilă $z = (x - \mu)/\sigma$, avem

$$\begin{aligned} Var(X) = E((X - \mu)^2) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \sigma^2. \end{aligned}$$

Ultima integrală a fost calculată la exemplul 9.

1.5 Quantile

Definiție: **Mediana** lui X este valoarea x pentru care $P(X \leq x) = 0.5$, i.e. valoarea lui x astfel încât $P(X \leq x) = P(X \geq x)$. Cu alte cuvinte, X are probabilitate egală de a fi deasupra sau sub mediană și de aceea fiecare probabilitate este 1/2. În termeni de cdf $F(x) = P(X \leq x)$, putem defini echivalent mediana ca valoarea lui x satisfăcând $F(x) = 0.5$.

Gândiți: Care este mediana lui Z ?

Răspuns: Din simetrie, mediana este 0.

Exemplul 11. Aflați mediana lui $X \sim \exp(\lambda)$.

Răspuns: Cdf a lui X este $F(x) = 1 - e^{-\lambda x}$ pe $[0, \infty)$. Mediana este valoarea lui x pentru care $F(x) = 1 - e^{-\lambda x} = 0.5$. Rezolvând ecuația obținem $x = (\ln 2)/\lambda$.

În acest caz mediana este mai mică decât media $\mu = 1/\lambda$.

Definiție: A **p -a quantilă** a lui X este valoarea q_p astfel încât $P(X \leq q_p) = p$.

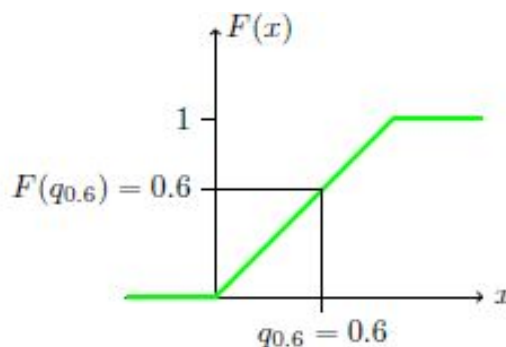
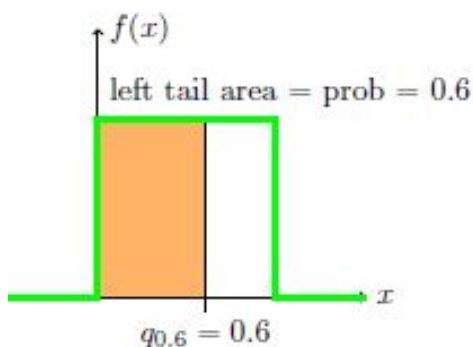
Observații. 1. Cu această notație mediana este $q_{0.5}$.

2. În termeni de cdf: $F(q_p) = p$.

În raport cu pdf f , quantila q_p este valoarea astfel încât sub graficul lui f este o arie de p la stânga lui q_p și o arie de $1 - p$ la dreapta lui q_p . În exemplele de mai jos, observați cum putem reprezenta quantila grafic folosind aria de sub graficul pdf sau înălțimea cdf.

Exemplul 12. Aflați 0.6 quantila pentru $X \sim U(0, 1)$.

Răspuns. Cdf pentru X este $F(x) = x$ pe domeniul de valori $[0, 1]$. Din $F(q_{0.6}) = 0.6 \Rightarrow q_{0.6} = 0.6$.

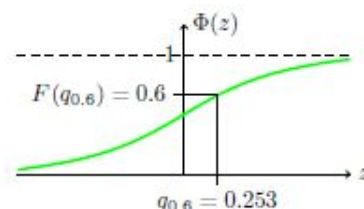
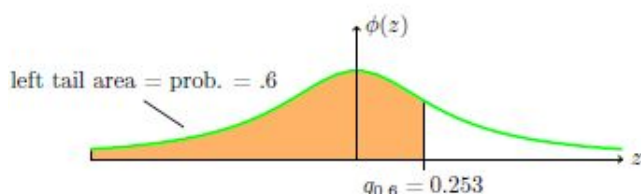


$q_{0.6}$: aria cozii stângi = 0.6 $\Leftrightarrow F(q_{0.6}) = 0.6$

Exemplul 13. Aflați 0.6 quantila repartiției normale standard.

Răspuns. Nu avem o formulă pentru cdf, deci vom folosi ”funcția quantilă” din R, `qnorm`.

$$q_{0.6} = \text{qnorm}(0.6, 0, 1) = 0.2533471$$



$q_{0.6}$: aria cozii stângi = 0.6 $\Leftrightarrow F(q_{0.6}) = 0.6$

Quantilele dau o măsură utilă a **locăției** pentru o variabilă aleatoare.

1.5.1 Percentile, decile, quartile

Pentru confort, quantilele sunt adesea descrise în termeni de percentile, decile sau quartile. A 60-a **percentilă** este 0.6 quantila. De exemplu, sunteți în a 60-a percentilă pentru înălțime dacă sunteți mai înalt(ă) decât 60% din populație, i.e. **probabilitatea** să fiți mai înalt(ă) decât o persoană aleasă aleator este 60%.

De asemenea, **decilele** reprezintă pași de 1/10. A 3-a decilă este 0.3 cuantila.

Quartilele sunt în pași de 1/4. A 3-a quartilă este 0.75 quantila și a 75-a percentilă.

2 Teorema limită centrală și legea numerelor mari

2.1 Scopurile învățării

1. Să înțeleagă enunțul legii numerelor mari.
2. Să înțeleagă enunțul teoremei limită centrală.
3. Să poată să utilizeze teorema limită centrală pentru a aproxima probabilități ale mediilor și sumelor de variabile aleatoare independente identic distribuite.

2.2 Introducere

Media multor măsurări ale aceleiași cantități necunoscute tinde să dea o estimare mai bună decât o singură măsurare. Intuitiv, aceasta este deoarece eroarea aleatoare a fiecărei măsurări se reduce în medie. 2 moduri de a face precisă această intuiție sunt legea numerelor mari (LoLN) și teorema limită centrală (CLT).

Scurt, atât legea numerelor mari cât și teorema limită centrală sunt despre multe date independente din aceeași repartiție. LoLN ne spune 2 lucruri:

1. Media multor date independente este (cu mare probabilitate) aproape de media repartiției de bază.
2. Histograma densității multor date independente este (cu mare probabilitate) aproape de graficul densității repartiției de bază.

Pentru a fi absolut corecți matematic trebuie să facem aceste afirmații mai precise, dar, așa cum au fost făcute, sunt un bun mod de a gândi despre legea numerelor mari.

Teorema limită centrală spune că suma sau media multor copii independente ale unei variabile aleatoare este aproximativ o variabilă aleatoare normală. CLT continuă dând valori precise pentru media și deviația standard ale variabilei normale.

Adesea în practică n nu trebuie să fie foarte mare. Valorile $n > 30$ sunt adesea suficiente.

2.2.1 Este mai multă experimentare decât matematică

Matematica LoLN spune că media unui lot de date independente dintr-o variabilă aleatoare se va apropia aproape sigur de media variabilei. Matematica [nu ne spune](#) dacă instrumentul sau experimentul produce date care merită să li se facă media. De exemplu, dacă dispozitivul de măsurare este defect sau

prost calibrat, atunci media multor măsurări va fi o estimare foarte precisă a unui lucru greșit! Acesta este un exemplu de [eroare sistematică](#) sau [deplasare a datelor](#), în contrast cu [eroarea aleatoare](#) controlată de legea numerelor mari.

2.3 Legea numerelor mari

Presupunem că X_1, X_2, \dots, X_n sunt variabile aleatoare independente cu aceeași repartiție. În acest caz, spunem că X_i sunt [independente și identic distribuite](#) sau [i.i.d.](#) În particular, X_i au aceeași medie μ și deviație standard σ . Fie \bar{X}_n media lui X_1, X_2, \dots, X_n :

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i.$$

\bar{X}_n este ea însăși o variabilă aleatoare. Legea numerelor mari și teorema limită centrală ne spun despre valoarea, respectiv, repartiția lui \bar{X}_n .

LoLN: Când n crește, probabilitatea că \bar{X}_n este aproape de μ tinde la 1.

CLT: Când n crește, repartiția lui \bar{X}_n converge la repartiția normală $N(\mu, \sigma^2/n)$.

Exemplul 1. Medii ale variabilelor aleatoare Bernoulli

Presupunem că fiecare X_i este o aruncare independentă a unei monede corecte, deci $X_i \sim \text{Bernoulli}(0.5)$ și $\mu = 0.5$. Atunci \bar{X}_n este proporția de aversuri în n aruncări și ne așteptăm că această proporție este aproape de 0.5 pentru n mare. Acest fapt nu este garantat; de exemplu, putem obține 1000 de aversuri în 1000 de aruncări, deși probabilitatea ca aceasta să aibă loc este foarte mică.

[Cu mare probabilitate](#) media de selecție \bar{X}_n este aproape de media 0.5 pentru n mare. Vom verifica asta făcând unele calcule în R.

La început, vom considera probabilitatea de a fi în 0.1 a mediei. Putem exprima această probabilitate ca

$$P(|\bar{X}_n - 0.5| \leq 0.1) = P(0.4 \leq \bar{X}_n \leq 0.6).$$

Legea numerelor mari spune că această probabilitate tinde la 1 când numărul de aruncări n devine mare. Codul R produce următoarele valori pentru $P(0.4 \leq \bar{X}_n \leq 0.6)$:

```
n = 10 : pbinom(6, 10, 0.5) - pbinom(3, 10, 0.5) = 0.65625
n = 50 : pbinom(30, 50, 0.5) - pbinom(19, 50, 0.5) = 0.8810795
n = 100 : pbinom(60, 100, 0.5) - pbinom(39, 100, 0.5) = 0.9647998
n = 500 : pbinom(300, 500, 0.5) - pbinom(199, 500, 0.5) = 0.9999941
n = 1000 : pbinom(600, 1000, 0.5) - pbinom(399, 1000, 0.5) = 1
```

După cum s-a prezis de LoLN, probabilitatea tinde la 1 când n crește. Refacem aceste calcule pentru a vedea probabilitatea de a fi în 0.01 a mediei. Codul R produce următoarele valori pentru $P(0.49 \leq \bar{X}_n \leq 0.51)$:

```

n = 10 : pbinom(5, 10, 0.5) - pbinom(4, 10, 0.5) = 0.2460937
n = 100 : pbinom(51, 100, 0.5) - pbinom(48, 100, 0.5) = 0.2356466
n = 1000 : pbinom(510, 1000, 0.5) - pbinom(489, 1000, 0.5) = 0.49334
n = 10000 : pbinom(5100, 10000, 0.5) - pbinom(4899, 10000, 0.5) = 0.9555742

```

Din nou, vedem probabilitatea de a fi aproape de medie tinzând la 1 când n crește. Deoarece $0.01 < 0.1$ este nevoie de valori mai mari ale lui n pentru a crește probabilitatea să se apropie de 1.

Convergența probabilității la 1 este LoLN în acțiune! Astfel, conform LoLN, [cu probabilitate mare](#), media unui mare număr de date independente din aceeași repartiție va fi foarte aproape de media repartiției.

2.3.1 Enunțul formal al legii numerelor mari

Teoremă (Legea numerelor mari): Presupunem că $X_1, X_2, \dots, X_n, \dots$ sunt variabile aleatoare i.i.d. cu media μ și dispersia σ^2 . Pentru fiecare n , fie \bar{X}_n media primelor n variabile. Atunci pentru orice $a > 0$, avem

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < a) = 1.$$

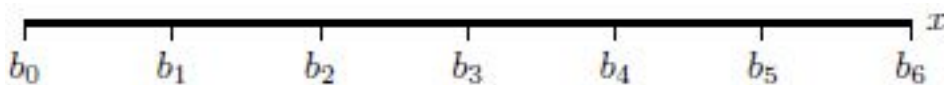
Aceasta spune precis că atunci când n crește, probabilitatea de a fi în a a mediei tinde la 1. Ne gândim la a ca la o mică toleranță a erorii de la adevărata medie μ . În exemplul nostru, dacă vrem ca probabilitatea să fie cel puțin $p = 0.99999$ ca proporția de aversuri \bar{X}_n este în $a = 0.1$ a lui $\mu = 0.5$, atunci $n > N = 500$ va fi suficient de mare. Dacă descreștem toleranța a și/sau creștem probabilitatea p , atunci N trebuie să fie mai mare.

2.4 Histograme

Putem rezuma date multiple x_1, \dots, x_n ale unei variabile aleatoare într-o [histogramă](#). Aici vrem să construim histogramele cu grijă astfel încât ele să semene cu aria de sub pdf.

Instrucțiunile pas cu pas pentru a construi o histogramă de densitate sunt următoarele.

1. Alegem un interval de pe dreapta reală și-l împărțim în m intervale, cu capetele b_0, b_1, \dots, b_m . De obicei aceste intervale au lungimi egale, deci presupunem asta de la început.



Fiecare din intervale este numit **coș** (în engleză bin). De exemplu, în figura de mai sus, primul coș este $[b_0, b_1]$ și ultimul coș este $[b_5, b_6]$. Fiecare coș are o **lățime a coșului**, de exemplu $b_1 - b_0$ este lățimea primului coș. De obicei toate coșurile au aceeași lățime, numită lățimea coșului histogramei.

2. Plasăm fiecare x_i în coșul care-i conține valoarea. Dacă x_i este pe frontiera a 2 coșuri, îl punem în coșul din stânga (aceasta este modul implicit în R, deși poate fi schimbat).

3. Pentru a desena o **histogramă de frecvențe**: punem un dreptunghi vertical deasupra fiecărui coș. **Înălțimea** dreptunghiului ar trebui să fie egală cu numărul de x_i -uri din coș.

4. Pentru a desena o **histogramă de densitate**: punem un dreptunghi vertical deasupra fiecărui coș. **Aria** dreptunghiului ar trebui să fie egală cu fracția din toate datele care sunt în coș.

Observații:

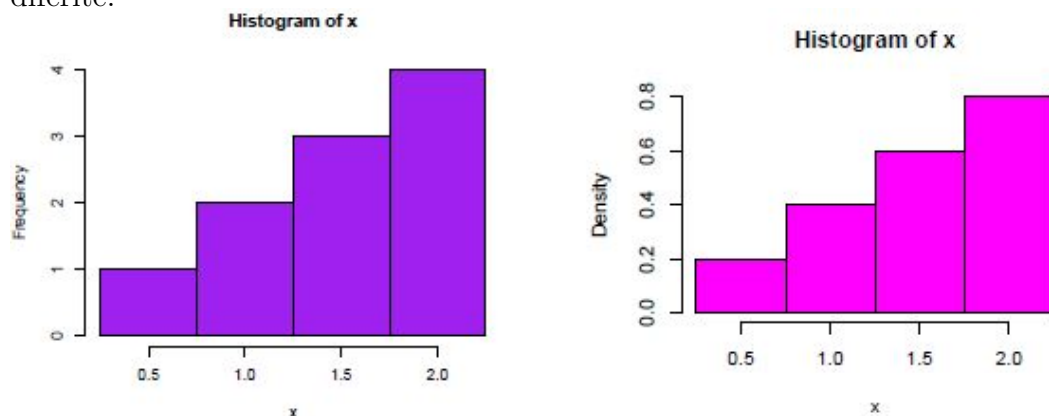
1. Când toate coșurile au aceeași lățime, dreptunghiurile histogramei de frecvențe au aria proporțională cu numărul. Deci histograma de densitate rezultă simplu din împărțirea înălțimii fiecărui dreptunghi la aria totală a histogramei de frecvențe. **Ignorând scala verticală, cele 2 histograme arată identic.**

2. **Atenție:** dacă lățimile coșurilor diferă, histogramele de frecvență și de densitate pot arăta foarte diferit.

În general, preferăm histograma de densitate deoarece scala ei verticală este aceeași cu cea a pdf.

Exemple. Iată câteva exemple de histograme, toate cu datele $[0.5, 1, 1, 1.5, 1.5, 1.5, 2, 2, 2, 2]$.

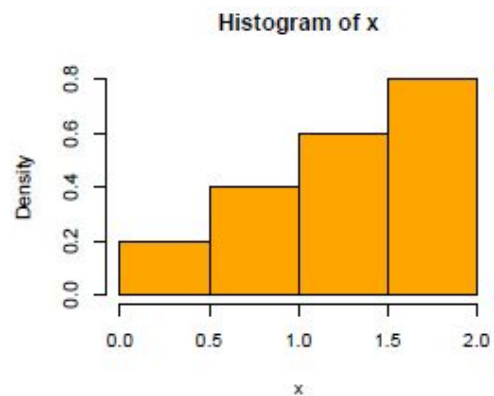
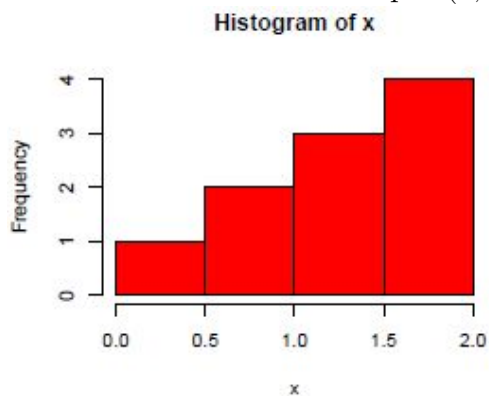
1. Aici reprezentările frecvenței și densității arată la fel, dar au scale verticale diferite.



Coșuri centrate în 0.5, 1, 1.5, 2, i.e. lățime 0.5, margini la 0.25, 0.75, 1.25,

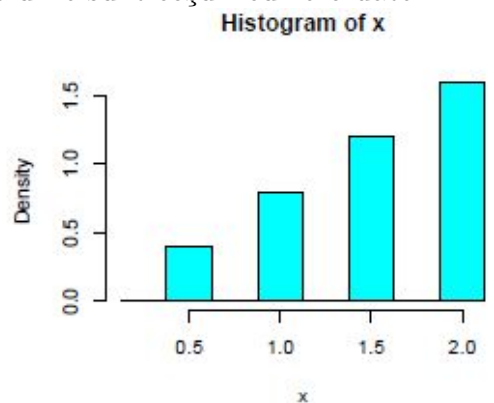
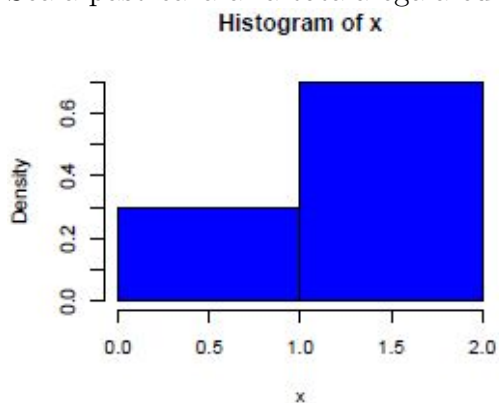
1.75, 2.25.

2. Aici valorile sunt toate pe frontierele coșurilor și sunt puse în coșul din stânga. Adică, coșurile sunt **închise la dreapta**, de exemplu primul coș este intervalul închis la dreapta $(0,0.5]$.



Marginile coșurilor la 0, 0.5, 1, 1.5, 2.

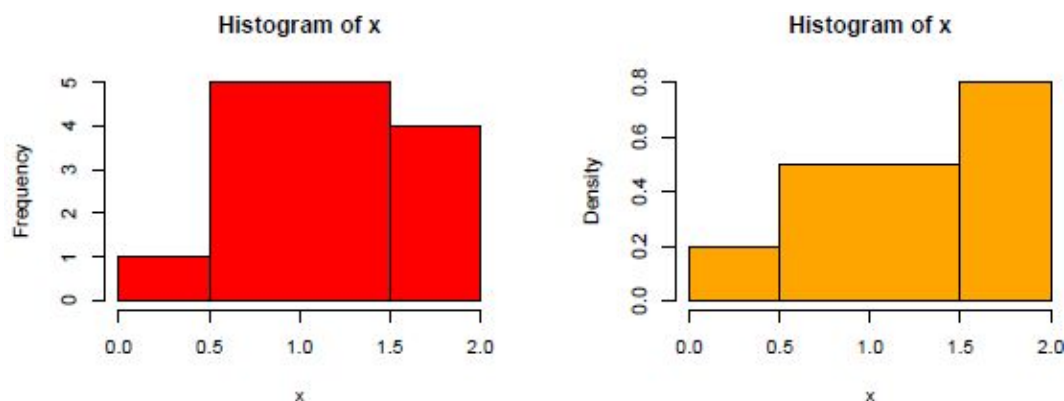
3. Aici sunt histograme de densitate bazate pe diferite lățimi de coșuri. Scala păstrează aria totală egală cu 1. Golurile sunt coșuri cu zero date.



Stânga: coșuri late.

Dreapta: coșuri înguste.

4. Aici folosim lățimi diferite de coșuri, așa că histogramele de frecvență și densitate arată diferit.



Nu vă lăsați păcăliți! Acestea sunt bazate pe aceleași date.

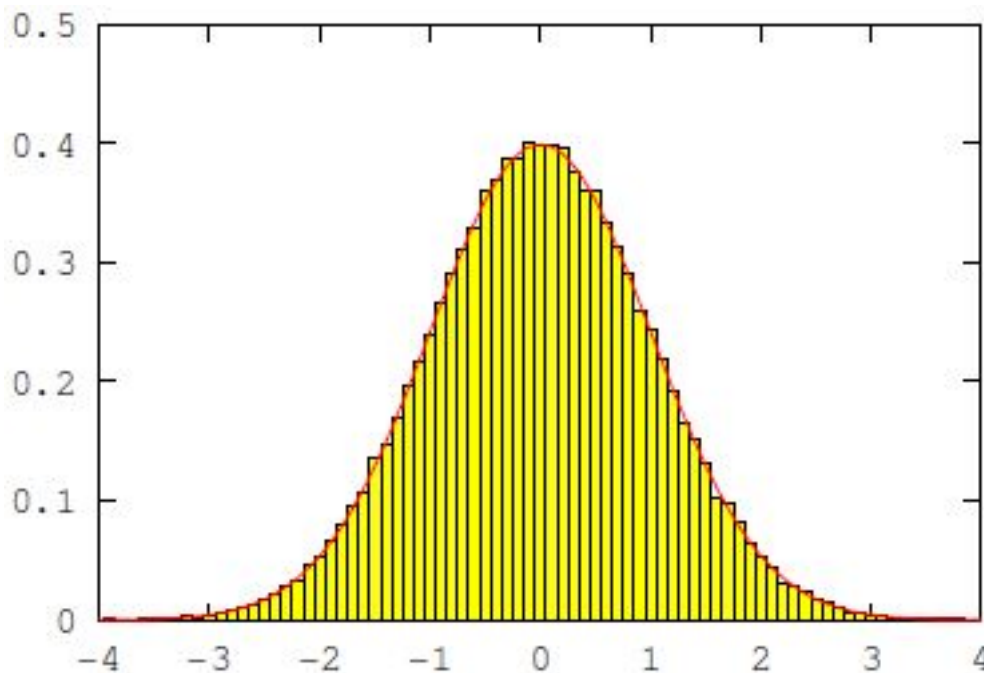
Histograma de densitate este alegerea mai bună la lățimi diferite de coșuri. De fapt, R se va plânge dacă încercăm să facem o histogramă de frecvențe cu lățimi diferite de coșuri. Comparați histograma de frecvențe cu lățimi diferite de coșuri cu toate celelalte histogramme pe care le-am desenat pentru aceste date. Arată clar diferit. Ce s-a întâmplat este că prin combinarea datelor din coșurile $(0.5, 1]$ și $(1, 1.5]$ într-un coș $(0.5, 1.5]$ am făcut mai mare înălțimea ambelor coșuri mai mici.

2.4.1 Legea numerelor mari și histogrammele

Legea numerelor mari are o consecință importantă pentru histogramme de densitate.

LoLN pentru histogramme: Cu probabilitate mare histograma de densitate a unui mare număr de date dintr-o repartiție este o bună aproximare a graficului pdf f a repartiției.

Ilustrăm aceasta generând o histogramă de densitate cu lățimea coșului 0.1 din 10000 de date dintr-o repartiție normală standard. Histograma de densitate urmărește foarte aproape graficul pdf normală standard ϕ .



Histograma de densitate a 10000 de date dintr-o repartiție normală standard, cu graficul lui ϕ în roșu.

2.5 Teorema limită centrală

2.5.1 Standardizare

Fiind dată o variabilă aleatoare X cu media μ și deviația standard σ , definim [standardizarea](#) lui X ca noua variabilă aleatoare

$$Z = \frac{X - \mu}{\sigma}.$$

Z are media 0 și deviația standard 1. Dacă X are o repartiție normală, atunci standardizarea lui X are repartiția normală standard Z cu media 0 și dispersia 1. Aceasta explică termenul ”standardizare” și notația Z de mai sus.

2.5.2 Enunțul teoremei limită centrală

Presupunem că $X_1, X_2, \dots, X_n, \dots$ sunt variabile aleatoare i.i.d., fiecare având media μ și deviația standard σ . Pentru fiecare n notăm cu S_n suma și cu \bar{X}_n

media lui X_1, \dots, X_n .

$$S_n = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i$$

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{S_n}{n}.$$

Proprietățile mediei și dispersiei arată că

$$E(S_n) = n\mu, \text{ } Var(S_n) = n\sigma^2, \text{ } \sigma_{S_n} = \sigma\sqrt{n}$$

$$E(\bar{X}_n) = \mu, \text{ } Var(\bar{X}_n) = \frac{\sigma^2}{n}, \text{ } \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}.$$

Deoarece S_n este multiplu al lui \bar{X}_n , S_n și \bar{X}_n au aceeași standardizare

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Teorema limită centrală: Pentru n mare,

$$\bar{X}_n \approx N(\mu, \sigma^2/n), \text{ } S_n \approx N(n\mu, n\sigma^2), \text{ } Z_n \approx N(0, 1).$$

Observații: 1. În cuvinte: \bar{X}_n este aproximativ o repartiție normală cu aceeași medie ca X , dar o dispersie mai mică.

2. S_n este aproximativ normală.

3. \bar{X}_n și S_n standardizate sunt aproximativ normale standard.

Teorema limită centrală ne permite să aproximăm o sumă sau medie de variabile aleatoare i.i.d. printr-o variabilă aleatoare normală. Ea este extrem de folositoare deoarece de obicei este ușor să facem calcule cu repartiția normală.

Un enunț precis al CLT este: cdf-urile lui Z_n converg la Φ :

$$\lim_{n \rightarrow \infty} F_{Z_n}(z) = \Phi(z).$$

2.5.3 Probabilități normale standard

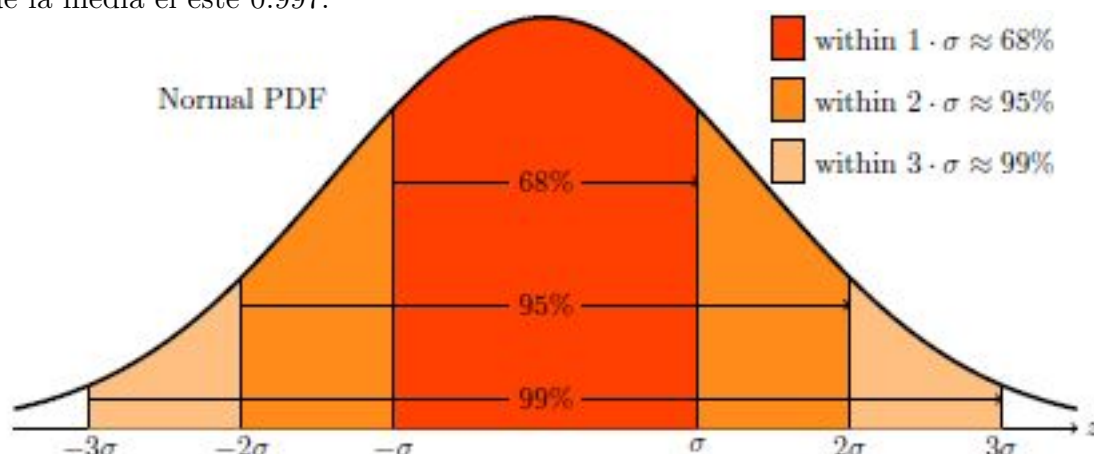
Pentru a aplica CLT vrem să avem la îndemână unele probabilități normale standard. Reamintim: dacă $Z \sim N(0, 1)$, atunci cu rotunjire avem:

1. $P(|Z| < 1) = 0.68$.
2. $P(|Z| < 2) = 0.95$; mai precis, $P(|Z| < 1.96) \approx 0.95$.
3. $P(|Z| < 3) = 0.997$.

Aceste numere sunt ușor de calculat în R folosind `pnorm`. Oricum, ele sunt

demne de reamintit ca regula degetului mare.

1. Probabilitatea că o variabilă aleatoare normală este într-o deviație standard de la media ei este 0.68. (Adică, $X \sim N(\mu, \sigma^2) \Rightarrow P(|X - \mu| < \sigma) \approx 0.68$.)
2. Probabilitatea că o variabilă aleatoare normală este în 2 deviații standard de la media ei este 0.95.
3. Probabilitatea că o variabilă aleatoare normală este în 3 deviații standard de la media ei este 0.997.



Consecințe:

1. $P(Z < 1) \approx 0.84$.
2. $P(Z < 2) \approx 0.977$.
3. $P(Z < 3) \approx 0.999$.

Demonstrație: 1. Știm că $P(|Z| < 1) = 0.68$. Probabilitatea rămasă de 0.32 este în 2 regiuni $Z > 1$ și $Z < -1$. Aceste regiuni sunt numite **coada dreaptă**, respectiv **coada stângă**. Din simetrie, fiecare coadă are 0.16. Deci,

$$P(Z < 1) = P(|Z| < 1) + P(\text{coada stângă}) \approx 0.68 + 0.16 = 0.84.$$

Celelalte 2 consecințe se fac similar.

2.5.4 Aplicații ale CLT

Exemplul 2. Aruncăm o monedă corectă de 100 de ori. Estimați probabilitatea a mai mult de 55 de aversuri.

Răspuns: Fie X_j rezultatul celei de-a j -a aruncări, deci $X_j = 1$ pentru avers și $X_j = 0$ pentru revers. Numărul total de aversuri este

$$S = X_1 + X_2 + \dots + X_{100}.$$

Știm că $E(X_j) = 0.5$ și $Var(X_j) = 1/4$. Deoarece $n = 100$, avem

$$E(S) = 50, \quad Var(S) = 25, \quad \text{și } \sigma_S = 5.$$

Teorema limită centrală spune că standardizarea lui S este aproximativ $N(0, 1)$. Se cere estimarea $P(S > 55)$. Standardizând și folosind CLT obținem

$$P(S > 55) = P\left(\frac{S - 50}{5} > \frac{55 - 50}{5}\right) \approx P(Z > 1) \approx 0.16.$$

Aici Z este o variabilă aleatoare normală și $P(Z > 1) = 1 - P(Z < 1) \approx 1 - 0.84 = 0.16$.

Exemplul 3. Estimați probabilitatea a mai mult de 220 de aversuri în 400 de aruncări.

Răspuns: Este aproape identic cu exemplul precedent. Acum $\mu_S = 200$ și $\sigma_S = 10$ și se cere estimarea $P(S > 220)$. Standardizând și folosind CLT obținem:

$$P(S > 220) = P\left(\frac{S - \mu_S}{\sigma_S} > \frac{220 - 200}{10}\right) \approx P(Z > 2) \approx 0.023.$$

Din nou, $Z \sim N(0, 1)$ și regulile degetului mare arată că $P(Z > 2) \approx 0.023$.

Observație: Cu toate că $55/100 = 220/400$, probabilitatea a mai mult de 55 de aversuri în 100 de aruncări este mai mare decât probabilitatea a mai mult de 220 de aversuri în 400 de aruncări. Aceasta este datorată LoLN și valorii mai mari a lui n în ultimul caz.

Exemplul 4. Estimați probabilitatea unui număr între 40 și 60 de aversuri în 100 de aruncări.

Răspuns: Ca în primul exemplu, $E(S) = 50$, $Var(S) = 25$ și $\sigma_S = 5$. Deci

$$P(40 \leq S \leq 60) = P\left(\frac{40 - 50}{5} \leq \frac{S - 50}{5} \leq \frac{60 - 50}{5}\right) \approx P(-2 \leq Z \leq 2).$$

Putem calcula ultimul membru folosind regula degetului mare. Pentru un răspuns mai precis folosim R:

$$\text{pnorm}(2) - \text{pnorm}(-2) = 0.9544997.$$

Reamintim că am folosit repartiția binomială pentru a calcula un răspuns de 0.9647998. Deci răspunsul nostru aproximativ folosind CLT are o eroare de aproximativ 1%.

Gândiți: Vă așteptați ca metoda CLT să dea o aproximare mai bună sau mai rea pentru $P(200 < S < 300)$ cu $n = 500$?

Verificați răspunsul folosind R.

Exemplul 5. Sondaj. Când se face un sondaj politic, rezultatele sunt adesea raportate ca un număr cu o margine de eroare. De exemplu, $52\% \pm 3\%$ susțin

candidatul A. Regula degetului mare este că dacă sondezi n oameni, atunci marginea de eroare este $\pm \frac{1}{\sqrt{n}}$. Vom vedea acum exact ce înseamnă aceasta și că este o aplicație a teoremei limită centrală.

Presupunem că sunt 2 candidați A și B. Presupunem mai departe că fracția din populație care preferă pe A este p_0 . Adică, dacă întrebi o persoană aleatoare pe cine preferă, probabilitatea să răspundă A este p_0 .

Pentru a face sondajul, un sondator alege aleator n persoane și le întreabă "Susțineți candidatul A sau candidatul B?" Astfel, putem vedea sondajul ca o secvență de n date Bernoulli(p_0) independente, X_1, X_2, \dots, X_n , unde X_i este 1 dacă a i -a persoană preferă pe A și 0 dacă preferă pe B. Frația de oameni sondați că-l preferă pe A este chiar media \bar{X} .

Știm că fiecare $X_i \sim \text{Bernoulli}(p_0)$, deci

$$E(X_i) = p_0 \text{ și } \sigma_{X_i} = \sqrt{p_0(1 - p_0)}.$$

De aceea, teorema limită centrală ne spune că

$$\bar{X} \approx N(p_0, \sigma/\sqrt{n}), \text{ unde } \sigma = \sqrt{p_0(1 - p_0)}.$$

Într-o repartiție normală 95% din probabilitate este în 2 deviații de la medie. Aceasta înseamnă că în 95% din sondajele a n oameni media de selecție \bar{X} va fi în $2\sigma/\sqrt{n}$ de la adevărata medie p_0 . Pasul final este să observăm că pentru orice valoare p_0 avem $\sigma \leq 1/2$. (De exemplu, se poate aplica inegalitatea mediilor.) Aceasta înseamnă că putem spune că în 95% din sondajele a n oameni media de selecție este în $1/\sqrt{n}$ de la adevărata medie. Statisticianul frecvenționist ia intervalul $[\bar{X} - 1/\sqrt{n}, \bar{X} + 1/\sqrt{n}]$ și-l numește intervalul de 95% încredere pentru p_0 .

Un cuvânt de precauție: este tentant și uzual, **dar greșit**, a gândi că este o probabilitate de 95% ca adevărata fracție p_0 să fie în intervalul de încredere. Asest fapt este subtil, dar eroarea este aceeași cu a gândi că ai boala dacă un test 95% precis iese pozitiv. Este adevărat că 95% din oamenii care fac testul primesc rezultatul corect. Nu este necesar adevărat că 95% dintre toate testele pozitive sunt corecte.

2.5.5 De ce folosim CLT?

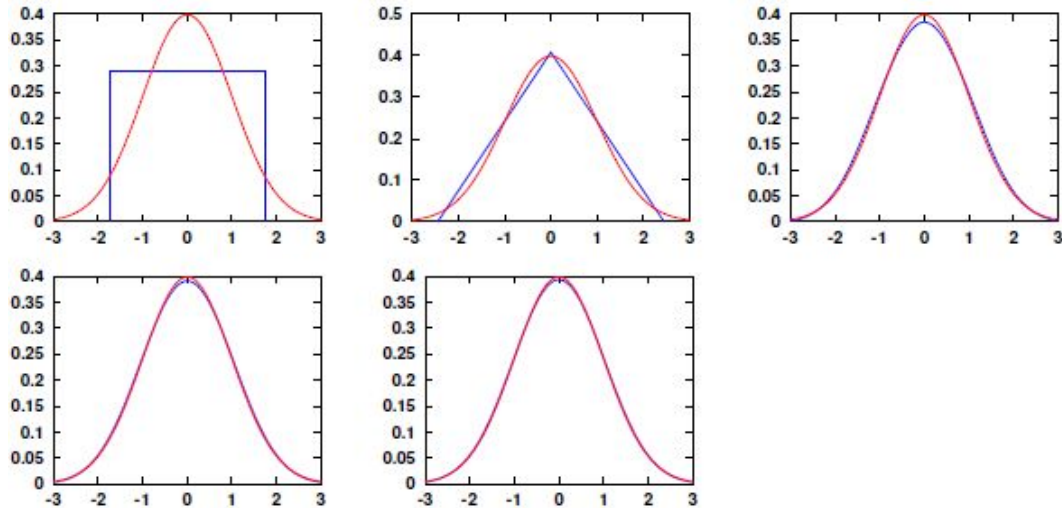
Deoarece probabilitățile din exemplele de mai sus puteau fi calculate folosind repartiția binomială, poate vă întrebați care este rostul aflării unui răspuns aproximativ folosind CLT. De fapt, puteam să calculăm exact aceste probabilități deoarece X_i erau Bernoulli și de aceea suma S era binomială. În general, repartiția lui S nu va fi familiară, de aceea nu vom putea calcula exact probabilitățile pentru S ; se poate întâmpla ca să fie posibil calculul

exact în teorie, dar să fie prea greu de calculat în practică, chiar și pentru un computer. Puterea CLT este că se poate aplica când X_i are aproape orice repartiție.

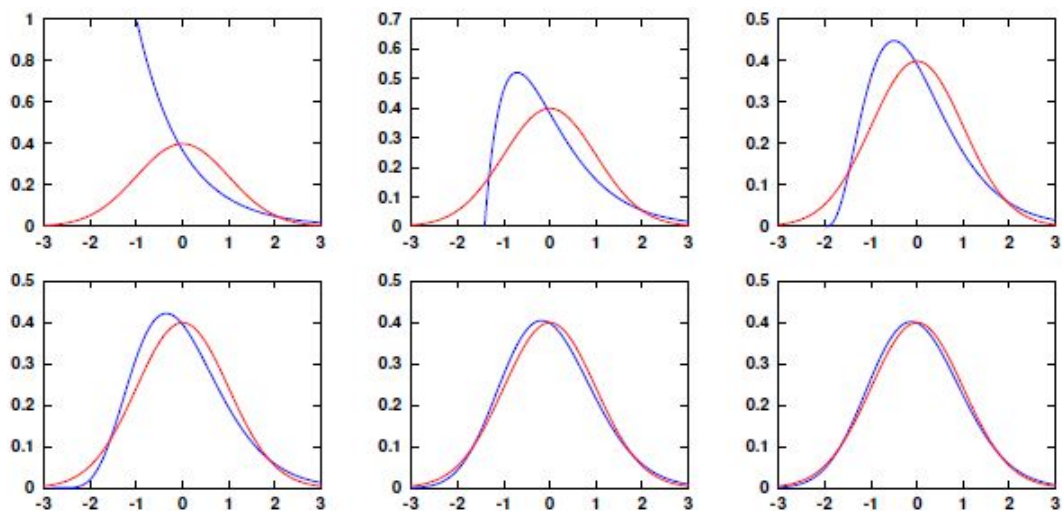
2.5.6 Cât de mare trebuie să fie n pentru a aplica CLT?

Răspuns scurt: adesea, nu așa mare.

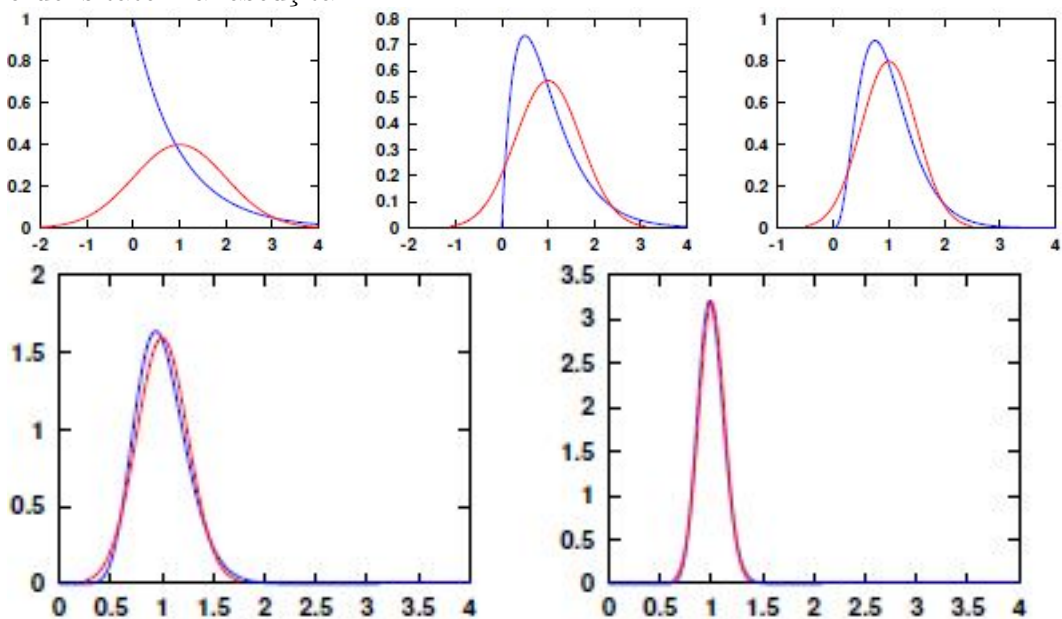
Următoarele secvențe de imagini arată convergența la o repartiție normală. Întâi arătăm media standardizată a n variabile aleatoare **uniforme** i.i.d. cu $n = 1, 2, 4, 8, 12$. Pdf a mediei este cu albastru și pdf normală standard este în roșu. La $n = 12$ potrivirea dintre media standardizată și adevărata normală arată foarte bine.



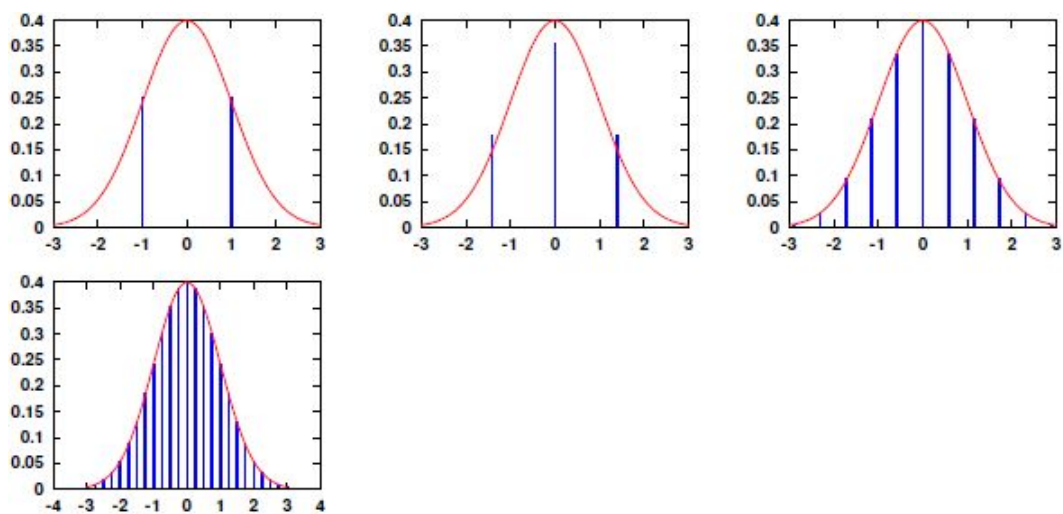
Apoi arătăm media standardizată a n variabile aleatoare **exponențiale** i.i.d. cu $n = 1, 2, 4, 8, 16, 64$. Această densitate asimetrică ia mai mulți termeni pentru a se apropia de densitatea normală.



Apoi arătăm media (nestandardizată) a n variabile aleatoare exponențiale cu $n = 1, 2, 4, 16, 64$. Deviația standard de micșorează când n crește, rezultând o densitate mai ascuțită.



Teorema limită centrală funcționează și pentru variabile aleatoare discrete. Aici este media standardizată a n variabile aleatoare Bernoulli(0.5) i.i.d. cu $n = 1, 2, 12, 64$. Când n crește, media poate lua mai multe valori, ceea ce permite repartiției discrete să "umple" densitatea normală.



La sfârșit arătăm media (nestandardizată) a n variabile aleatoare Bernoulli(0.5), cu $n = 4, 12, 64$. Deviația standard devine mai mică rezultând o densitate mai ascuțită.

