

Curs 3

Cristian Niculescu

1 Dispersia (varianța) variabilelor aleatoare discrete

1.1 Scopurile învățării

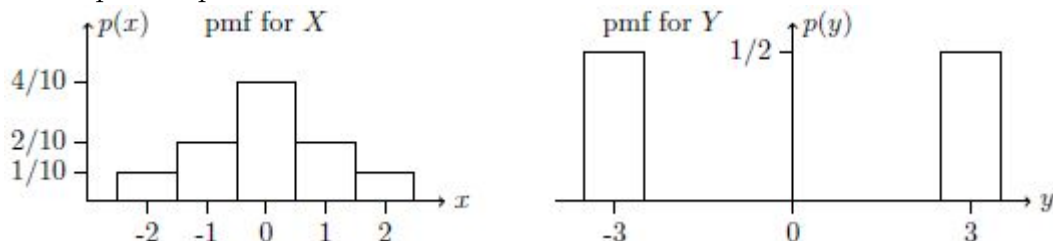
1. Să poată să calculeze dispersia și deviația standard a unei variabile aleatoare discrete.
2. Să înțeleagă că deviația standard este o măsură a scalării sau împrăstierii.
3. Să poată să calculeze dispersia folosind proprietățile de scalare și liniaritate.

1.2 Împrăștierea

Media unei variabile aleatoare este o măsură a [tendinței centrale](#). Dacă ar trebui să rezumăm cu un singur număr o variabilă aleatoare, media ar fi o alegere bună. Totuși, media nu cuprinde toată informația. De exemplu, variabilele aleatoare X și Y de mai jos au ambele media 0, dar masa lor de probabilitate este împrăștiată în jurul mediei complet diferit.

$$X \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 1/10 & 2/10 & 4/10 & 2/10 & 1/10 \end{pmatrix} \quad Y \sim \begin{pmatrix} -3 & 3 \\ 1/2 & 1/2 \end{pmatrix}.$$

Pentru a vedea împrăștierea diferită, reprezentăm pmf-urile. Folosim bare în loc de puncte pentru a da un sens mai bun masei.



Pmf-urile pentru 2 repartiții diferite, ambele cu media 0

1.3 Dispersia și deviația standard

Luând media centrul repartiției unei variabile aleatoare, **dispersia** este o măsură a cât de mult masa de probabilitate este **împrăștiată** în jurul acestui centru.

Definiție. Dacă X este o variabilă aleatoare cu media $E(X) = \mu$, atunci **dispersia** lui X este

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

Deviația standard σ a lui X este

$$\sigma = \sqrt{Var(X)}.$$

Dacă variabila aleatoare relevantă este clară din context, atunci dispersia și deviația standard sunt adesea notate cu σ^2 și σ ("sigma"), la fel cum media este μ ("miu").

Rescriem definiția explicit ca o sumă. Dacă X ia valorile x_1, x_2, \dots, x_n cu pmf p , atunci

$$Var(X) = E((X - \mu)^2) = \sum_{i=1}^n p(x_i)(x_i - \mu)^2.$$

Formula pentru $Var(X)$ este o medie ponderată a pătratelor distanțelor la medie. Prin ridicare la pătrat, ne asigurăm că mediem numai valori nenegative, astfel încât împrăștierea la dreapta mediei să nu se anuleze cu cea de la stânga. Prin folosirea mediei, ponderăm valorile cu probabilitate mai mare mai mult decât valorile cu probabilitate mai mică.

Observații despre unități de măsură:

1. σ are aceleași unități de măsură ca X .
 2. $Var(X)$ are aceleași unități de măsură ca pătratul lui X . Astfel, dacă X este în metri, atunci $Var(X)$ este în metri pătrați.
- Deoarece σ și X au aceleași unități de măsură, deviația standard este o măsură firească a împrăstierii.

Exemplul 1. Calculați media, dispersia și deviația standard ale variabilei aleatoare

$$X \sim \begin{pmatrix} 1 & 3 & 5 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

Răspuns: Întâi calculăm $E(X) = 1 \cdot (1/4) + 3 \cdot (1/4) + 5 \cdot (1/2) = 7/2$. Apoi completăm tabelul cu $(X - 7/2)^2$.

| | | | |
|---------------|------|-----|-----|
| valoarea x | 1 | 3 | 5 |
| $p(x)$ | 1/4 | 1/4 | 1/2 |
| $(x - 7/2)^2$ | 25/4 | 1/4 | 9/4 |

Acum calculul dispersiei este similar celui al mediei:

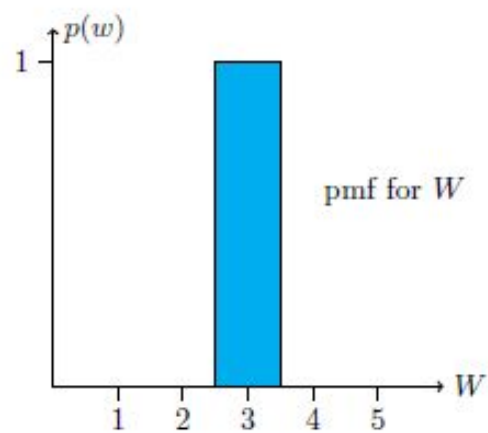
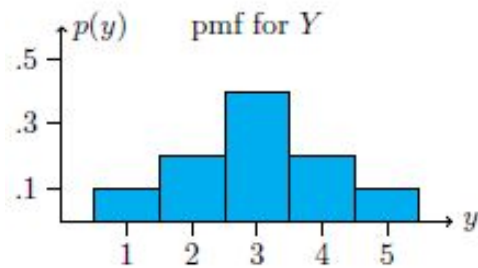
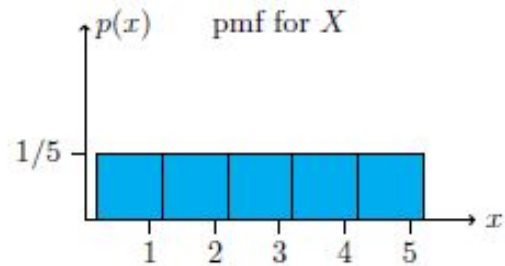
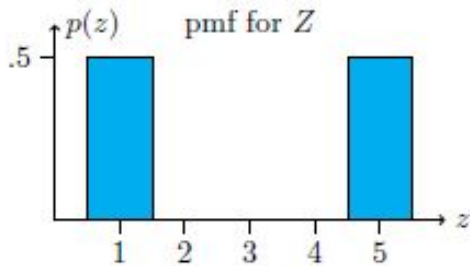
$$\text{Var}(X) = \frac{25}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} + \frac{9}{4} \cdot \frac{1}{2} = \frac{11}{4}.$$

Extrăgând radicalul, avem deviația standard $\sigma = \sqrt{11}/2$.

Exemplul 2. Pentru fiecare variabilă aleatoare X, Y, Z și W reprezentați pmf și calculați media și dispersia.

- (i) $X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$;
- (ii) $Y \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1/10 & 1/5 & 2/5 & 1/5 & 1/10 \end{pmatrix}$;
- (iii) $Z \sim \begin{pmatrix} 1 & 5 \\ 1/2 & 1/2 \end{pmatrix}$;
- (iv) $W \sim \begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

Răspuns: Fiecare variabilă aleatoare are media 3, dar probabilitatea este împrăștiată diferit. În reprezentările de mai jos, le ordonăm de la cea mai mare la cea mai mică dispersie: Z, X, Y, W .



Calculăm dispersia fiecărei variabile. Toate au media $\mu = 3$. Deoarece dispersia este definită ca o medie, o putem calcula folosind tabele.

| | | | | | |
|---------------|-----|-----|-----|-----|-----|
| valoarea x | 1 | 2 | 3 | 4 | 5 |
| pmf $p(x)$ | 1/5 | 1/5 | 1/5 | 1/5 | 1/5 |
| $(X - \mu)^2$ | 4 | 1 | 0 | 1 | 4 |

$$Var(X) = E((X - \mu)^2) = \frac{4}{5} + \frac{1}{5} + 0 + \frac{1}{5} + \frac{4}{5} = 2.$$

| | | | | | |
|---------------|------|-----|-----|-----|------|
| valoarea y | 1 | 2 | 3 | 4 | 5 |
| pmf $p(y)$ | 1/10 | 1/5 | 2/5 | 1/5 | 1/10 |
| $(Y - \mu)^2$ | 4 | 1 | 0 | 1 | 4 |

$$Var(Y) = E((Y - \mu)^2) = \frac{2}{5} + \frac{1}{5} + 0 + \frac{1}{5} + \frac{2}{5} = 1.2.$$

| | | |
|---------------|-----|-----|
| valoarea z | 1 | 5 |
| pmf $p(y)$ | 1/2 | 1/2 |
| $(Z - \mu)^2$ | 4 | 4 |

$$Var(Z) = E((Z - \mu)^2) = 2 + 2 = 4.$$

| | |
|---------------|---|
| valoarea w | 3 |
| pmf $p(y)$ | 1 |
| $(W - \mu)^2$ | 0 |

$$Var(W) = E((W - \mu)^2) = 0. \text{ Observăm că } W \text{ nu variază, de aceea are dispersia } 0.$$

1.3.1 Dispersia unei variabile aleatoare Bernoulli(p)

Dacă $X \sim \text{Bernoulli}(p)$, atunci

$$Var(X) = p(1 - p).$$

Demonstrație: Știm că $E(X) = p$. Calculăm $Var(X)$ folosind un tabel.

| | | |
|------------------|-------------|-------------|
| valorile lui X | 0 | 1 |
| pmf $p(x)$ | $1 - p$ | p |
| $(X - \mu)^2$ | $(0 - p)^2$ | $(1 - p)^2$ |

$$Var(X) = (1 - p)p^2 + p(1 - p)^2 = p(1 - p)(p + 1 - p) = p(1 - p).$$

Gândiți: Pentru ce valoare a lui p Bernoulli(p) are cea mai mare dispersie? Încercați să răspundeți aplicând inegalitatea mediilor sau aflând maximul funcției de gradul 2 $f(p) = p - p^2$ pe $(0, 1)$.

1.3.2 Variabile aleatoare discrete independente

Până acum am folosit noțiunea de variabile aleatoare independente fără să o definim riguros. De exemplu, o repartiție binomială este sumă de probe Bernoulli [independente](#). Sigur, avem un sens intuitiv despre ce înseamnă independența pentru probe experimentale. Avem de asemenea sensul probabilistic că variabilele aleatoare X și Y sunt independente când cunoașterea

valorii lui X nu ne dă informații despre valoarea lui Y .

Definiție: Variabilele aleatoare discrete X și Y sunt **independente** \iff

$$P(X = a, Y = b) = P(X = a)P(Y = b), \forall a, b.$$

1.3.3 Proprietățile dispersiei

Cele mai utile 3 proprietăți pentru calculul dispersiei sunt:

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

Pentru proprietatea 1, observăm cu atenție cerința că X și Y sunt independente.

Proprietatea 3 dă o formulă pentru $Var(X)$ care este adesea mai ușor de utilizat în calcule.

Exemplul 3. Presupunem că X și Y sunt independente și $Var(X) = 3$ și $Var(Y) = 5$. Aflați:

- (i) $Var(X + Y)$,
- (ii) $Var(3X + 4)$,
- (iii) $Var(X + X)$,
- (iv) $Var(X + 3Y)$.

Răspuns: Pentru a calcula aceste dispersii, folosim proprietățile 1 și 2.

- (i) Deoarece X și Y sunt **independente**, $Var(X + Y) = Var(X) + Var(Y) = 8$.
- (ii) Folosind proprietatea 2, $Var(3X + 4) = 9Var(X) = 27$.
- (iii) Nu vă lăsați păcăliți! Proprietatea 1 nu poate fi aplicată deoarece sigur X nu este independentă de ea însăși. Putem folosi proprietatea 2: $Var(X + X) = Var(2X) = 4Var(X) = 12$. (Observație: dacă din greșală am fi folosit proprietatea 1, am fi răspuns greșit 6.)
- (iv) Folosim ambele proprietăți 1 și 2.

$$Var(X + 3Y) = Var(X) + Var(3Y) = 3 + 9 \cdot 5 = 48.$$

Exemplul 4. Folosiți proprietatea 3 pentru a calcula dispersia lui $X \sim \text{Bernoulli}(p)$.

Răspuns: Din tabelul

| | | | |
|--------|--|---------|-----|
| X | | 0 | 1 |
| $p(x)$ | | $1 - p$ | p |
| X^2 | | 0 | 1 |

avem $E(X^2) = p$. Deci proprietatea 3 dă

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p).$$

Rezultatul coincide cu cel din calculul anterior.

Exemplul 5. Refaceți exemplul 1 folosind proprietatea 3.

Răspuns: Din tabelul

| | | | | |
|--------|--|-----|-----|-----|
| X | | 1 | 3 | 5 |
| $p(x)$ | | 1/4 | 1/4 | 1/2 |
| X^2 | | 1 | 9 | 25 |

avem $E(X)=7/2$ și

$$E(X^2) = 1 \cdot \frac{1}{4} + 9 \cdot \frac{1}{4} + 25 \cdot \frac{1}{2} = \frac{60}{4} = 15.$$

Deci $Var(X) = 15 - (7/2)^2 = 11/4$, ca în exemplul 1.

1.3.4 Dispersia pentru binomial(n, p)

Presupunem $X \sim \text{binomial}(n, p)$. Deoarece X este sumă de n variabile Bernoulli(p) *independente* și fiecare variabilă Bernoulli are dispersia $p(1-p)$, avem

$$X \sim \text{binomial}(n, p) \Rightarrow Var(X) = np(1-p).$$

1.3.5 Demonstrațiile proprietăților 2 și 3

Demonstrația proprietății 2: Rezultă din proprietățile lui $E(X)$.

Fie $\mu = E(X)$. Atunci $E(aX + b) = a\mu + b$ și

$$\begin{aligned} Var(aX + b) &= E((aX + b - (a\mu + b))^2) = E((aX - a\mu)^2) = E(a^2(X - \mu)^2) \\ &= a^2 E((X - \mu)^2) = a^2 Var(X). \end{aligned}$$

Demonstrația proprietății 3: Folosim proprietățile lui $E(X)$. Reamintim că μ este o constantă și că $E(X) = \mu$.

$$\begin{aligned} E((X - \mu)^2) &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - E(X)^2, \text{ q.e.d.} \end{aligned}$$

1.4 Tabele de repartiții și proprietăți

| Repartiția | valorile lui X | pmf $p(x)$ | media $E(X)$ | dispersia $Var(X)$ |
|--------------------|------------------|----------------------------------|-----------------|--------------------|
| Bernoulli(p) | 0,1 | $p(0) = 1 - p, p(1) = p$ | p | $p(1 - p)$ |
| Binomial(n, p) | $0, 1, \dots, n$ | $p(k) = C_n^k p^k (1 - p)^{n-k}$ | np | $np(1 - p)$ |
| Uniform(n) | $1, 2, \dots, n$ | $p(k) = \frac{1}{n}$ | $\frac{n+1}{2}$ | $\frac{n^2-1}{12}$ |
| Geometric(p) | $0, 1, 2, \dots$ | $p(k) = p(1 - p)^k$ | $\frac{1-p}{p}$ | $\frac{1-p}{p^2}$ |

Fie X o variabilă aleatoare discretă cu valorile x_1, x_2, \dots și pmf $p(x_j)$.

| Media: | Dispersia: |
|--|--|
| Sinonime: | varianță |
| Notății: $E(X), \mu, m$ | $Var(X), \sigma^2$ |
| Definiție: $E(X) = \sum_j x_j p(x_j)$ | $E((X - \mu)^2) = \sum_j p(x_j)(x_j - \mu)^2$ |
| Scalare, translată: $E(aX + b) = aE(X) + b$ | $Var(aX + b) = a^2 Var(X)$ |
| Aditivitate: $E(X + Y) = E(X) + E(Y)$ | $X, Y \text{ ind.} \Rightarrow Var(X + Y) = Var(X) + Var(Y)$ |
| Funcții de X : $E(h(X)) = \sum_j h(x_j)p(x_j)$ | |
| Formulă alternativă: | $Var(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$ |

2 Variabile aleatoare continue

2.1 Scopurile învățării

1. Să știe definiția unei variabile aleatoare continue.
2. Să știe definiția funcției densitate de probabilitate (pdf) și funcției de distribuție cumulativă (cdf).
3. Să poată să explice de ce folosim densitatea de probabilitate pentru variabilele aleatoare continue.

2.2 Introducere

Variabilele aleatoare atribuie un număr fiecărui rezultat posibil dintr-un spațiu al probelor. În timp ce variabilele aleatoare discrete iau o mulțime discretă de valori posibile, variabilele aleatoare continue au o mulțime continuă de valori.

Computațional, pentru a merge de la discret la continuu, pur și simplu înlocuim sumele cu integrale.

Exemplul 1. Deoarece timpul este continuu, cantitatea de timp cu care Ion vine mai devreme (sau întârzie) la curs este o variabilă aleatoare continuă.

Presupunem că măsurăm cât de devreme sosește Ion la curs în fiecare zi (în minute). Adică, rezultatul unei probe din experimentul nostru este un timp

în minute. Presupunem că sunt fluctuații aleatoare în timpul exact în care el apare. Deoarece în principiu Ion ar putea sosi, să zicem, cu 3.43 minute mai devreme, sau cu 2.7 minute mai târziu (corespunzând rezultatului -2.7), sau la orice alt timp, spațiul probelor constă din toate numerele reale. Deci variabila aleatoare care dă rezultatul are ea însăși un **domeniu continuu** de valori posibile.

2.3 Încălzire de analiză matematică

Dacă f este o funcție integrabilă și $f(x) \geq 0, \forall x \in [a, b]$, atunci

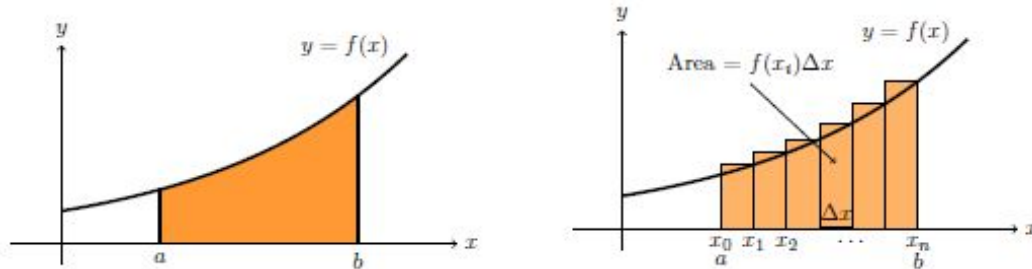
1. $\int_a^b f(x)dx = \text{aria de sub curba } y = f(x)$.

2. $\int_a^b f(x)dx = \text{"suma din } f(x)dx\text{"}$.

Conexiunea dintre cele 2 este:

$$\text{aria} \approx \text{suma ariilor dreptunghiurilor} = f(x_1)\Delta x + f(x_2)\Delta x + \dots + f(x_n)\Delta x = \sum_{i=1}^n f(x_i)\Delta x.$$

Pe măsură ce lungimea Δx a intervalelor devine mai mică, aproximarea devine mai bună.



Aria este aproximativ suma ariilor dreptunghiurilor

Observație: Interesul nostru în integrale vine în primul rând din interpretarea lor ca "sumă" și într-o mai mică măsură din interpretarea lor ca arie.

2.4 Variabile aleatoare continue și funcții densitate de probabilitate

O variabilă aleatoare continuă ia un **domeniu de valori** a cărui lungime poate fi finită sau infinită. Iată câteva exemple de domenii de valori: $[0, 1]$, $[0, \infty)$, \mathbb{R} , $[a, b]$.

Definiție: O variabilă aleatoare X este **continuuă** $\iff \exists$ o funcție $f(x)$ astfel încât $\forall c \leq d$ avem

$$P(c \leq X \leq d) = \int_c^d f(x)dx. \quad (1)$$

Funcția f este numită **funcția densitate de probabilitate (pdf)**.

Pdf satisface totdeauna următoarele proprietăți:

1. $f(x) \geq 0$ (f este nenegativă).
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ (Aceasta este echivalentă cu: $P(-\infty < X < \infty) = 1$).

Funcția densitate de probabilitate f a unei variabile aleatoare continue este analoaga funcției masă de probabilitate p a unei variabile aleatoare discrete.

Iată 2 diferențe importante:

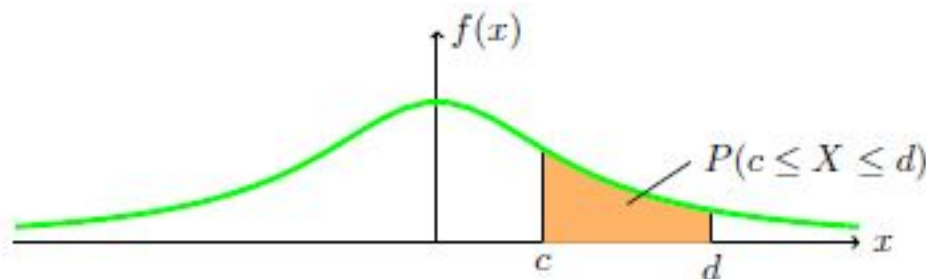
1. Spre deosebire de p , pdf f nu este o probabilitate. Trebuie s-o integrăm pentru a obține probabilitate.
2. Deoarece f nu este o probabilitate, nu există restricția $f(x) \leq 1$.

Observație: În proprietatea 2, am integrat de la $-\infty$ la ∞ deoarece n-am știut domeniul de valori luate de X . Formal, aceasta are sens deoarece definim $f(x) = 0$ în afara domeniului de valori al lui X . În practică, vom integra între marginile date de domeniul de valori al lui X .

2.4.1 Vederea grafică a probabilității

Dacă facem graficul pdf a unei variabile aleatoare continue X , atunci

$$P(c \leq X \leq d) = \text{aria de sub graficul pdf dintre } c \text{ și } d.$$



Gândiți: Cât este aria totală de sub pdf f ?

2.4.2 Termenii "masă de probabilitate" și "densitate de probabilitate"

De ce folosim termenii masă și densitate pentru a descrie pmf și pdf? Care este diferența dintre cele 2? Răspunsul simplu este că acești termeni sunt complet analogi masei și densității din fizică.

Masa ca o sumă:

Dacă masele m_1, m_2, m_3 și m_4 sunt puse pe o linie în pozițiile x_1, x_2, x_3 și x_4 , atunci masa totală este $m_1 + m_2 + m_3 + m_4$.

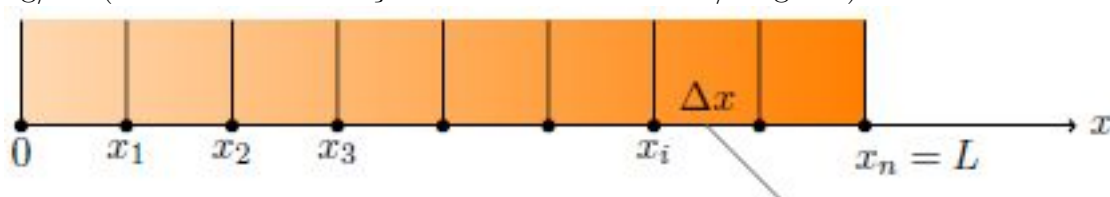


Putem defini o "funcție de masă" p cu $p(x_j) = m_j$ pentru $j = 1, 2, 3, 4$ și $p(x) = 0$ altfel. Cu această notație, masa totală este $p(x_1) + p(x_2) + p(x_3) + p(x_4)$.

Funcția masă de probabilitate se comportă în exact același mod, cu excepția faptului că are dimensiunea probabilității în loc de cea a masei.

Masa ca o integrală a densității:

Presupunem că avem o tijă de lungime L metri cu densitate variabilă $f(x)$ kg/m. (Observăm că unitățile de măsură sunt masă/lungime.)



masa celei de-a i -a

bucăți $\approx f(x_i)\Delta x$

Dacă densitatea variază continuu, trebuie să aflăm masa totală a tijei prin integrare:

$$\text{masa totală} = \int_0^L f(x)dx.$$

Această formulă provine din împărțirea tijei în bucăți mici și "adunarea" masei fiecărei bucăți. Adică:

$$\text{masa totală} \approx \sum_{i=1}^n f(x_i)\Delta x.$$

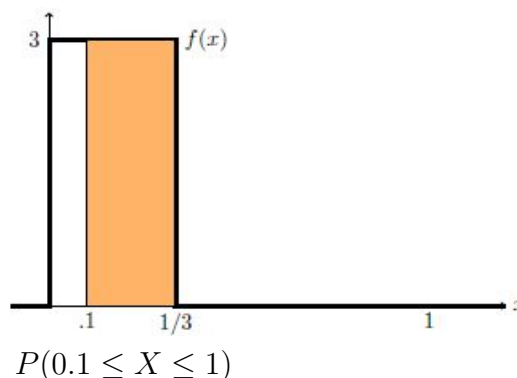
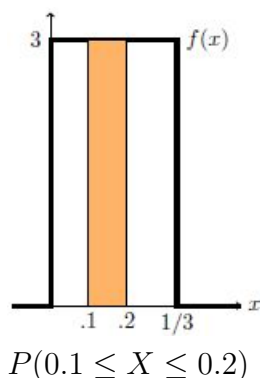
La limită când Δx tinde la 0, suma tinde la integrală.

Funcția densitate de probabilitate se comportă exact la fel, exceptând faptul că are ca unități de măsură probabilitatea/(unitatea lui x) în loc de kg/m. Într-adevăr, relația (1) este exact analoaga integralei de mai sus pentru masa totală.

Pentru ambele variabile aleatoare, discretă și continuă, media este **centrul masei** sau punctul de echilibru.

Exemplul 2. Presupunem că X are pdf $f(x) = 3$ pe $[0, 1/3]$ (aceasta înseamnă că $f(x) = 0$ în afara lui $[0, 1/3]$). Reprezentați pdf și calculați $P(0.1 \leq X \leq 0.2)$ și $P(0.1 \leq X \leq 1)$.

Răspuns:



$$P(0.1 \leq X \leq 0.2) = \int_{0.1}^{0.2} f(x)dx = \int_{0.1}^{0.2} 3dx = 3x|_{0.1}^{0.2} = 0.6 - 0.3 = 0.3.$$

Sau putem afla probabilitatea geometric:

$$P(0.1 \leq X \leq 0.2) = \text{aria dreptunghiului} = 3 \cdot 0.1 = 0.3.$$

Deoarece $P(0.1 \leq X \leq 1)$ este aria de sub $f(x)$ de la 0.1 doar până la $1/3$, avem $P(0.1 \leq X \leq 1) = 3(1/3 - 0.1) = 3 \cdot (7/30) = 0.7$.

Gândiți: În exemplul anterior $f(x)$ ia valori mai mari ca 1. De ce aceasta nu încalcă regula că probabilitățile sunt totdeauna între 0 și 1?

Observație asupra notației. Putem defini o variabilă aleatoare dându-i domeniul de valori și funcția densitate de probabilitate. De exemplu, putem spune: fie X o variabilă aleatoare cu domeniul de valori $[0, 1]$ și pdf $f(x) = 2x$. Implicit, aceasta înseamnă că X are densitatea de probabilitate 0 în afara domeniului de valori dat. Dacă am fi vrut să fim absolut riguroși, am fi scris explicit și că $f(x) = 0$ în afara lui $[0, 1]$, dar în practică aceasta nu este necesar.

Exemplul 3. Fie X o variabilă aleatoare cu domeniul de valori $[0, 1]$ și pdf $f(x) = Cx^2$. Cât este C ?

Răspuns: Deoarece probabilitatea totală trebuie să fie 1, avem

$$\int_0^1 f(x)dx = 1 \Leftrightarrow \int_0^1 Cx^2dx = 1.$$

Calculând integrala, ultima relație devine

$$((Cx^3)/3)|_0^1 = 1 \Rightarrow C/3 = 1 \Rightarrow C = 3.$$

Observație: Avem nevoie de constanta C de mai sus pentru a [norma](#) densitatea astfel încât probabilitatea totală să fie 1.

Exemplul 4. Fie X variabila aleatoare din exemplul 3. Aflați $P(X \leq 1/2)$.

Răspuns: $P(X \leq 1/2) = \int_0^{1/2} 3x^2 dx = x^3|_0^{1/2} = \frac{1}{8}$.

Gândiți: Pentru acest X (sau orice variabilă aleatoare continuă):

- Cât este $P(a \leq X \leq a)$?
- Cât este $P(X = 0)$?
- $P(X = a) = 0$ înseamnă că X nu poate fi niciodată egal cu a ?

În cuvinte, întrebările de mai sus duc la faptul că probabilitatea ca înălțimea unei persoane aleatoare să fie exact 1.7526 m (la precizie infinită, i.e. fără rotunjiri!) este 0. Totuși, este posibil ca înălțimea cuiva să fie exact 1.7526 m. Deci răspunsurile la întrebările de gândire sunt 0, 0 și Nu.

2.4.3 Funcția de distribuție cumulativă (funcția de repartiție)

Funcția de distribuție cumulativă (cdf) a unei variabile aleatoare continue X este definită în exact același mod ca cdf a unei variabile aleatoare discrete.

$$F(b) = P(X \leq b).$$

Observăm că definiția este despre probabilitate. Când folosim cdf ar trebui să ne gândim mai întâi la ea ca la o probabilitate. Apoi, când o **calculăm** putem folosi

$$F(b) = P(X \leq b) = \int_{-\infty}^b f(x) dx, \text{ unde } f(x) \text{ este pdf a lui } X.$$

Observații:

1. Pentru variabile aleatoare discrete, am definit funcția de distribuție cumulativă, dar nu am avut prea mult ocazia să o folosim. Cdf joacă un rol mult mai proeminent pentru variabile aleatoare continue.
2. Ca mai înainte, am integrat de la $-\infty$ deoarece n-am știut precis domeniul de valori al lui X . Formal, aceasta are sens deoarece $f(x) = 0$ în afara domeniului de valori al lui X . În practică, știm domeniul de valori și integrăm de la începutul lui.
3. În practică spunem adesea " **X are repartiția $F(x)$** " mai degrabă decât " X are funcția de distribuție cumulativă $F(x)$ ".

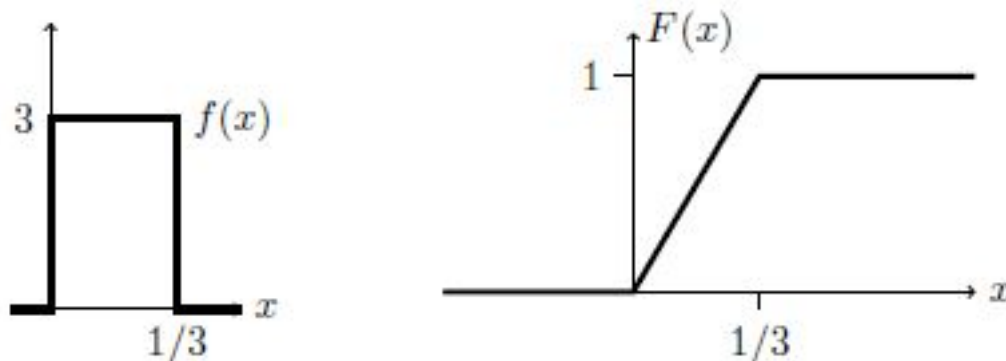
Exemplul 5. Aflați cdf pentru densitatea din exemplul 2.

Răspuns: Pentru $a \in [0, 1/3]$ avem $F(a) = \int_0^a f(x) dx = \int_0^a 3x dx = 3x|_0^a = 3a$.

Deoarece $f(x) = 0$ în afara lui $[0, 1/3]$ avem $F(a) = P(X \leq a) = 0$ pentru $a < 0$ și $F(a) = 1$ pentru $a > 1/3$. Punând toate acestea împreună, avem

$$F(a) = \begin{cases} 0, & \text{dacă } a < 0, \\ 3a, & \text{dacă } 0 \leq a \leq 1/3, \\ 1, & \text{dacă } a > 1/3. \end{cases}$$

Iată graficele lui f și F .



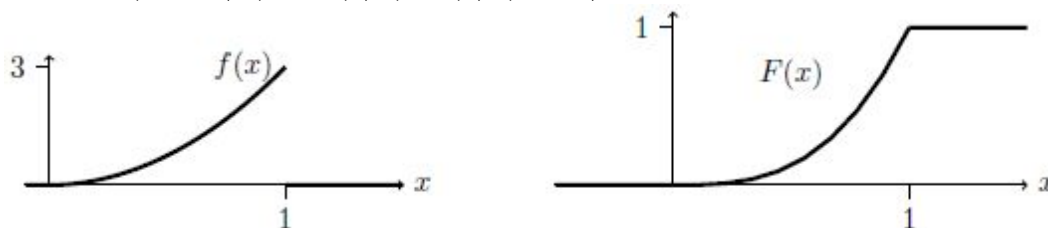
Observați scalele diferite pe axele verticale. Reamintim că axa verticală pentru pdf reprezintă densitatea de probabilitate, iar cea pentru cdf reprezintă probabilitatea.

Exemplul 6. Aflați cdf pentru pdf din exemplul 3, $f(x) = 3x^2$ pe $[0, 1]$. Presupunem că X este o variabilă aleatoare cu această repartiție. Aflați $P(X < 1/2)$.

Răspuns: $f(x) = 3x^2$ pe $[0, 1] \Rightarrow F(a) = \int_0^a 3x^2 dx = x^3|_0^a = a^3$ pe $[0, 1]$. De aceea,

$$F(a) = \begin{cases} 0, & \text{dacă } a < 0, \\ a^3, & \text{dacă } 0 \leq a \leq 1, \\ 1, & \text{dacă } a > 1. \end{cases}$$

Astfel, $P(X < 1/2) = F(1/2) = (1/2)^3 = 1/8$. Iată graficele lui f și F :



2.4.4 Proprietăți ale funcției de distribuție cumulative

Iată un rezumat al celor mai importante proprietăți ale funcției de distribuție cumulative (cdf) pentru o variabilă aleatoare continuă:

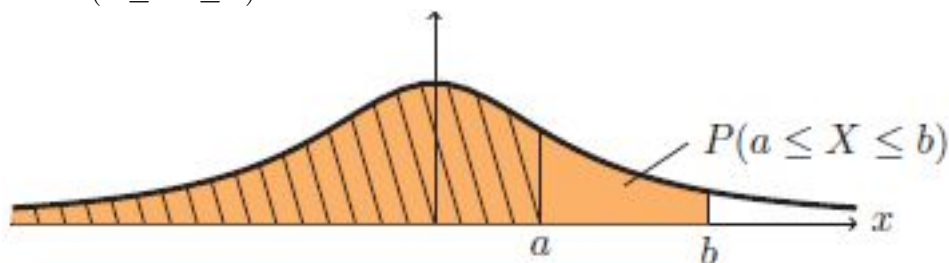
1. (Definiția) $F(x) = P(X \leq x)$.
2. $0 \leq F(x) \leq 1, \forall x \in \mathbb{R}$.
3. F este crescătoare, i.e. dacă $a \leq b$, atunci $F(a) \leq F(b)$.
4. $\lim_{x \rightarrow -\infty} F(x) = 0$ și $\lim_{x \rightarrow \infty} F(x) = 1$.
5. $P(a \leq X \leq b) = F(b) - F(a)$.
6. $F'(x) = f(x), \forall x$ în care F este derivabilă.

Proprietățile 2, 3 și 4 sunt identice cu cele pentru repartiții discrete. Graficele din exemplele anterioare le ilustrează.

Proprietatea 5 poate fi demonstrată:

$$\begin{aligned} \int_{-\infty}^b f(x)dx &= \int_{-\infty}^a f(x)dx + \int_a^b f(x)dx \\ \Leftrightarrow \int_a^b f(x)dx &= \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx \\ \Leftrightarrow P(a \leq X \leq b) &= F(b) - F(a). \end{aligned}$$

Proprietatea 5 poate fi de asemenea văzută geometric. Regiunea colorată de mai jos reprezintă $F(b)$ și regiunea hașurată reprezintă $F(a)$. Diferența lor este $P(a \leq X \leq b)$.



Proprietatea 6 este o teoremă fundamentală a analizei matematice.

2.4.5 Densitatea de probabilitate ca o tablă de darts

Ne gândim la prelevarea de valori ale unei variabile aleatoare continue ca la aruncarea de darts-uri într-o tablă de darts. Considerăm regiunea de sub graficul pdf ca o tablă de darts. Împărțim regiunea în pătrate mici de aceeași mărime și presupunem că, atunci când aruncăm un dart, este egal probabil ca el să aterizeze în orice pătrat. Probabilitatea ca dartul să aterizeze într-o regiune dată este fracția din totalul ariei de sub curbă preluată de regiune.

Deoarece totalul ariei este 1, această fracție este chiar aria regiunii. Dacă X reprezintă coordonata x a dartzului, atunci probabilitatea ca dartzul să aterizeze cu coordonata x între a și b este chiar

$$P(a \leq X \leq b) = \text{aria de sub } f(x) \text{ dintre } a \text{ și } b = \int_a^b f(x)dx.$$

3 Exemple de variabile aleatoare continue

3.1 Scopurile învățării

1. Să poată da exemple de ceea ce modelează repartițiile uniformă, exponențială și normală.
2. Să poată da domeniile de valori și pdf-urile repartițiilor uniformă, exponențială și normală.

3.2 Introducere

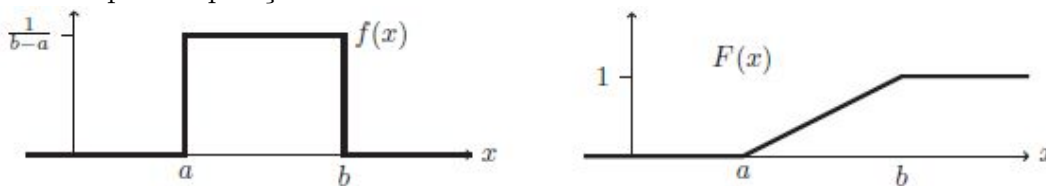
Introducem câteva repartiții continue fundamentale. Pentru fiecare repartiție dăm domeniul de valori, pdf, cdf și o scurtă descriere a situațiilor pe care le modelează. Toate aceste repartiții depind de parametri, pe care îi specificăm.

Cu toate că o abordăm spre final, repartiția normală este cea mai importantă din cele definite aici. Când dăm funcția de repartiție (cdf, prescurtat "repartiția") se subînțelege, dacă este cazul, că la stânga intervalului pe care este dată este 0, iar la dreapta este 1.

3.3 Repartiția uniformă

1. Parametri: $a < b$.
2. Domeniu de valori: $[a, b]$.
3. Notăție: $\text{uniform}(a, b)$ sau $U(a, b)$.
4. Densitate: $f(x) = \frac{1}{b-a}$ pentru $a \leq x \leq b$.
5. Repartiția: $F(x) = (x - a)/(b - a)$ pentru $a \leq x \leq b$.
6. Modele: Toate rezultatele din domeniul de valori au probabilitate egală (mai precis, toate rezultatele au aceeași densitate de probabilitate).

Grafice pentru pdf și cdf:



Exemple. 1. Presupunem că avem o ruletă marcată în milimetri. Dacă măsurăm (până la cel mai apropiat marcaj) lungimea unor articole care au aproximativ 1 metru, eroarea de rotunjire va fi repartizată uniform între -0.5 și 0.5 milimetri.

2. Multe jocuri de tablă folosesc săgeți care se învârt pentru a introduce hazardul. Când este învârtită, săgeata se oprește la un unghi care este repartizat uniform între 0 și 2π radiani.

3. În cele mai multe generatoare de numere pseudo-aleatoare, generatorul de bază simulează o repartiție uniformă și toate celelalte repartiții sunt construite transformând generatorul de bază.

Teorema de universalitate a repartiției uniforme (Paul Levy)

X variabilă aleatoare

$Y \sim U(0, 1)$

$F_X^{-1} : (0, 1) \rightarrow \mathbb{R}$, $F_X^{-1}(y) = \inf \{x \in \mathbb{R} | F_X(x) \geq y\}$, $\forall y \in (0, 1)$ - funcția cuantilă (inversa generalizată) asociată lui F_X

$\implies X$ și $F_X^{-1}(Y)$ sunt repartizate la fel.

Observație. Acest rezultat joacă un rol important în simulare.

El ne permite să generăm observații independente din orice repartiție, dacă știm funcția cuantilă.

x_1, \dots, x_n , observații independente repartizate $U(0, 1)$ generate prin `runif(n, 0, 1)`.

$F_X^{-1}(x_1), \dots, F_X^{-1}(x_n)$ sunt observații independente repartizate la fel ca X .

Observație. Dacă $X \sim U(a, b)$, atunci următoarele comenzi din R dau valorile funcției de repartiție, densității și funcției cuantilă ale lui X :

$F_X(x) = \text{punif}(x, a, b)$.

$f_X(x) = \text{dunif}(x, a, b)$.

$F_X^{-1}(x) = \text{qunif}(x, a, b)$.

3.4 Repartiția exponențială

1. Parametru: $\lambda > 0$.

2. Domeniu de valori: $[0, \infty)$.

3. Notăție: $\text{exponential}(\lambda)$ sau $\text{exp}(\lambda)$.

4. Densitate: $f(x) = \lambda e^{-\lambda x}$ pentru $x \geq 0$.

5. Repartiție: $F(x) = 1 - e^{-\lambda x}$ pentru $x \geq 0$.

6. Repartiția cozii drepte: $P(X > x) = 1 - F(x) = e^{-\lambda x}$.

7. Modele: Timpul de așteptare pentru un proces continuu de schimbare a stării.

Exemple. 1. Dacă ies afară după curs și aștept un taxi, timpul meu de așteptare în minute este repartizat exponențial. În acest caz λ este dat de 1 supra numărul mediu de taxiuri care trec pe minut (în această perioadă de

timp din zilele lucrătoare).

2. Repartiția exponențială modelează timpul de așteptare până când un izotop instabil suferă o descompunere nucleară. În acest caz λ este legat de timpul de înjumătățire al izotopului.

Proprietatea de lipsă a memoriei: Sunt și alte repartiții care modelează timpii de așteptare, dar repartiția exponențială are proprietatea adițională că este lipsită de memorie. Iată ce înseamnă aceasta în contextul exemplului 1. Presupunem că probabilitatea ca un taxi să sosească în primele 5 minute este p . Dacă aștept 5 minute și de fapt nu sosește niciun taxi, atunci probabilitatea ca un taxi să sosească în următoarele 5 minute este tot p .

Din contra, să presupunem că merg la o stație de metrou și aștept următorul tren. Deoarece trenurile sunt coordonate să urmeze un program (de exemplu, cel mult 12 minute între trenuri), dacă aștept 5 minute fără să văd un tren, atunci este o probabilitate mult mai mare ca trenul să sosească în următoarele 5 minute. În particular, timpul de așteptare pentru metrou nu este fără memorie și un model mai bun pentru el ar fi repartiția uniformă pe intervalul $[0,12]$.

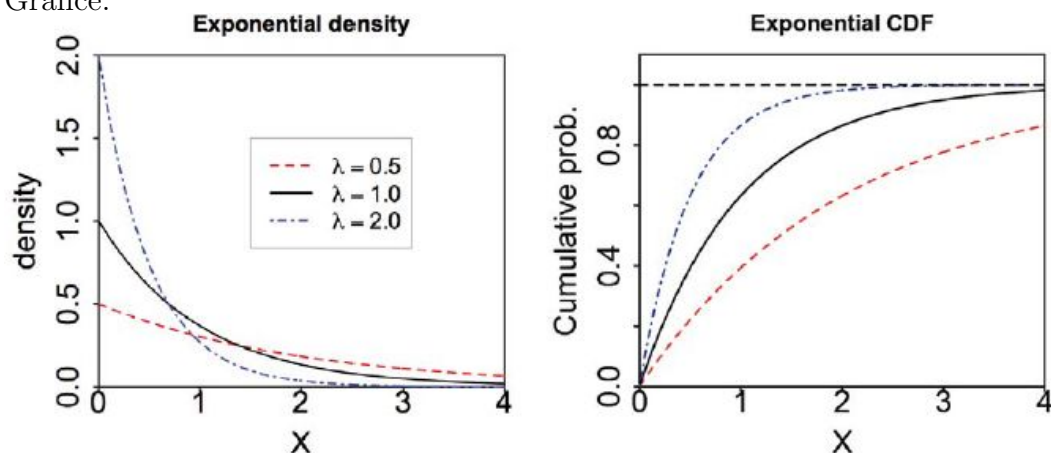
Lipsa de memorie a repartiției exponențiale este analoagă lipsei de memorie a repartiției geometrice (discrete), unde faptul că am aruncat 5 reversuri la rând nu ne dă nicio informație despre următoarele 5 aruncări.

Într-adevăr, repartiția exponențială este precis perechea continuă a repartiției geometrice, care modelează timpul de așteptare pentru un proces discret de schimbare de stări. Mai formal, lipsa de memorie înseamnă că probabilitatea de a aștepta mai mult de t minute este neafectată de faptul că am așteptat deja s minute fără ca evenimentul să se producă. În simboluri, $P(X > s + t | X > s) = P(X > t)$.

Demonstrația lipsei de memorie: Deoarece $(X > s + t) \cap (X > s) = (X > s + t)$, avem

$$P(X > s + t | X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t), \text{ q.e.d.}$$

Grafice:



Observație. Dacă $X \sim \exp(\lambda)$, atunci următoarele comenzi din R dau valorile funcției de repartiție, densității și funcției cuantilă ale lui X :

$$F_X(x) = \text{pexp}(x, \lambda).$$

$$f_X(x) = \text{dexp}(x, \lambda).$$

$$F_X^{-1}(x) = \text{qexp}(x, \lambda).$$

3.5 Repartiția normală

În 1809, Carl Friedrich Gauss a publicat o monografie introducând câteva noțiuni care au devenit fundamentale în statistică: repartiția normală, estimarea de verosimilitate maximă și metoda celor mai mici pătrate. Din acest motiv, repartiția normală mai este numită și repartiția *Gaussiană*. Ea este cea mai importantă repartiție continuă.

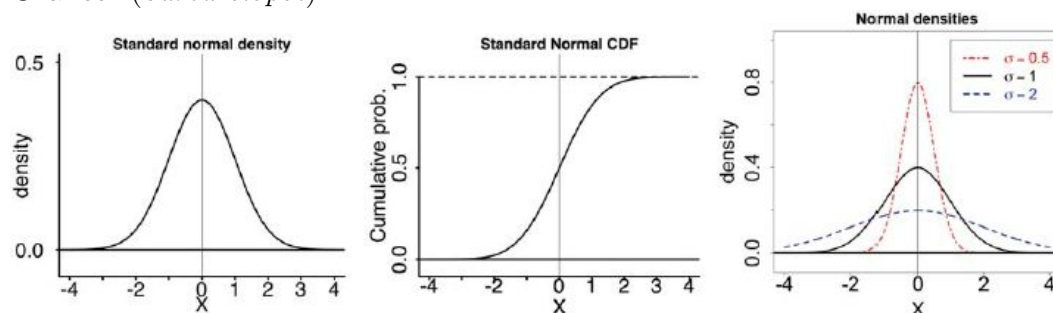
1. Parametri: $\mu \in \mathbb{R}, \sigma > 0$.
2. Domeniul de valori: \mathbb{R} .
3. Notăție: $\text{normal}(\mu, \sigma^2)$ sau $N(\mu, \sigma^2)$.
4. Densitate: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$.
5. Repartiție: $F(x)$ nu are formulă, așa că folosim tabele sau comenzi ca `pnorm` în R pentru a calcula $F(x)$.
6. Modele: măsurarea erorii, inteligenței/abilității, înălțimii, mediilor loturilor de date.

Repartiția normală standard $N(0, 1)$ are media 0 și dispersia 1. Rezervăm Z pentru o variabilă aleatoare normală standard, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ pentru densitatea normală standard și $\Phi(z)$ pentru repartiția normală standard.

Observație: Media și dispersia variabilelor aleatoare continue au aceeași interpretare ca în cazul discret. Repartiția normală $N(\mu, \sigma^2)$ are media μ , dispersia σ^2 și deviația standard σ .

Iată câteva grafice ale repartiției normale. Observăm că densitățile au graficele curbe în formă de *clopot*. Mai observăm că pe măsură ce σ crește, ele devin din ce în ce mai împrăștiate.

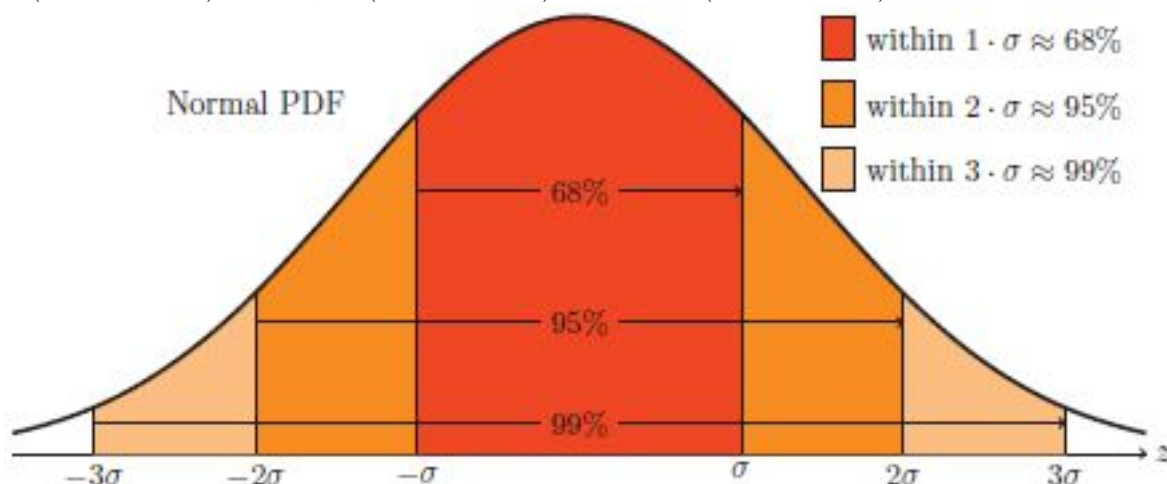
Grafice: (*curba clopot*):



3.5.1 Probabilități normale

Pentru a face aproximații este util să ne reamintim următoarea regulă a degetului mare pentru 3 probabilități aproximative

$P(-1 \leq Z \leq 1) \approx 0.68$, $P(-2 \leq Z \leq 2) \approx 0.95$, $P(-3 \leq Z \leq 3) \approx 0.99$.

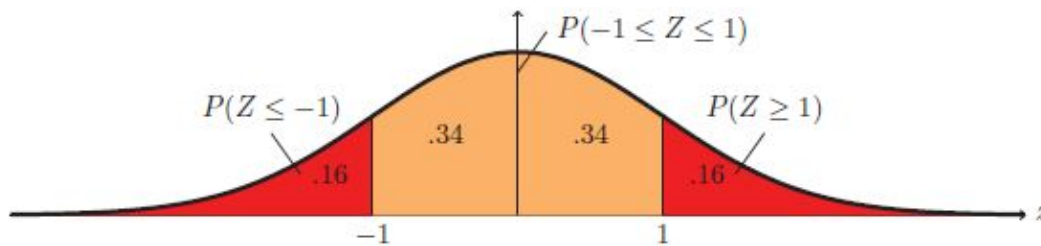


Calculare cu simetrie

Putem folosi simetria graficului densității normale standard față de dreapta $x = 0$ pentru a face unele calcule.

Exemplul 1. Regula degetului mare spune $P(-1 \leq Z \leq 1) \approx 0.68$. Folosiți aceasta pentru a estima $\Phi(1)$.

Răspuns: $\Phi(1) = P(Z \leq 1)$.



În figură, cele 2 cozi (în roșu) au împreună aria $1 - 0.68 = 0.32$. Din simetrie, coada stângă are aria 0.16 (jumătate din 0.32), deci $P(Z \leq 1) \approx 0.16 + 0.68 = 0.84$.

3.5.2 Folosirea lui R pentru a calcula $\Phi(z)$

```
# Folosim funcția R pnorm(x, μ, σ) pentru a calcula F(x) pentru N(μ, σ²).
pnorm(1,0,1)
[1] 0.8413447
pnorm(0,0,1)
[1] 0.5
pnorm(1,0,2)
[1] 0.6914625
pnorm(1,0,1)-pnorm(-1,0,1)
[1] 0.6826895
pnorm(5,0,5)-pnorm(-5,0,5)
[1] 0.6826895
# Desigur, z poate fi un vector de valori:
pnorm(c(-3,-2,-1,0,1,2,3),0,1)
[1] 0.001349898 0.022750132 0.158655254 0.500000000 0.841344746 0.977249868
[7] 0.998650102
```

Observație: Funcția R $\text{pnorm}(x, \mu, \sigma)$ utilizează σ în timp ce notația noastră pentru repartiția normală $N(\mu, \sigma^2)$ folosește σ^2 .

Iată un tabel de valori cu o acuratețe cu mai puține zecimale:

| z: | -2 | -1 | 0 | .3 | .5 | 1 | 2 | 3 |
|-------------|--------|--------|--------|--------|--------|--------|--------|--------|
| $\Phi(z)$: | 0.0228 | 0.1587 | 0.5000 | 0.6179 | 0.6915 | 0.8413 | 0.9772 | 0.9987 |

Exemplul 2. Utilizați R pentru a calcula $P(-1.5 \leq Z \leq 2)$.

Răspuns: $P(-1.5 \leq Z \leq 2) = \Phi(2) - \Phi(-1.5) = \text{pnorm}(2, 0, 1) - \text{pnorm}(-1.5, 0, 1) = 0.9104427$.

Observație. Dacă $X \sim N(\mu, \sigma^2)$, atunci următoarele comenzi din R dau valorile densității și funcției cuantilă ale lui X :

$f_X(x) = \text{dnorm}(x, \mu, \sigma)$.

$F_X^{-1}(x) = \text{qnorm}(x, \mu, \sigma)$.

3.6 Repartiția Pareto

1. Parametri: $m > 0$ și $\alpha > 0$.
2. Domeniul de valori: $[m, \infty)$.
3. Notăție: $\text{Pareto}(m, \alpha)$.
4. Densitate: $f(x) = \frac{\alpha m^\alpha}{x^{\alpha+1}}$.
5. Repartiția: $F(x) = 1 - \frac{m^\alpha}{x^\alpha}$, pentru $x \geq m$.
6. Repartiția coadă: $P(X > x) = m^\alpha/x^\alpha$, pentru $x \geq m$.
7. Modele: Repartiția Pareto modelează o **lege de putere**, unde probabilitatea ca un eveniment să aibă loc variază ca o putere a unui atribut al evenimentului. Multe fenomene urmează o lege de putere, ca mărimea meteoriților, nivelurile veniturilor într-o populație și nivelurile populației în orașe. Vezi http://en.wikipedia.org/wiki/Pareto_distribution#Applications.