Probabilitate și statistică în Data Science

Partea III. Distribuții

npsimid

Ce ne așteaptă?

1. Ce este distribuția
2. Distribuția uniformă
3. Distribuția binomială
4. Distribuția Poisson
5. Distribuția normală
6. Scorul z

Notații

$P(E)$ – ***probabilitatea evenimentului E***

n_E – ***numarul de evenimente***

$P(x{:}\,n,p)$ – ***functia masa de probabilitate binomiala a succeselor x din n cu probabilitate p***

$\binom{n}{x}$ – ***combinatii din n luate cate x***

λ – ***parametru lambda egal cu valoarea medie a populatiei***

$P(x)$ – ***functia masa de probabilitate Poisson***

$P(X{:}\, x < N)$ – ***functia masa cumulativa Poisson***

$f(x)$ – ***functia densitatii probabilitatii***

z – ***scorul z***

p – ***valoarea percentilei***

1. Ce este distribuția

Noțiune de distribuție

- Distribuția – descrie toate rezultatele posibile ale unei variabile
- Distribuția discretă – rezultatele posibile ale variabilei aparțin unui set discret de valori
- Distribuția continuă – rezultatele posibile ale variabilei aparțin unui interval continuu de valori
- În distribuția discretă suma tuturor probabilităților individuale trebuie să fie 1
- În distribuția continuă aria suprafeței de sub curba probabilității trebuie să fie 1

Tipuri de distribuție

- **Grupul distribuțiilor discrete include:**
  - **Distribuția uniformă**
  - **Distribuția binomială**
  - **Distribuția Poisson**
  - **etc**
- **Probabilitatea distribuției discrete se mai numește funcția masă de probabilitate**
- **Grupul distribuțiilor continue include:**
  - **Distribuția normală**
  - **Distribuția exponențială**
  - **Distribuția Beta**
  - **etc**
- **Probabilitatea distribuției discrete se mai numește funcția densitate de probabilitate**

2. Distribuția uniformă

Esența distribuției uniforme

- **Distribuția uniformă este distribuția discretă pentru care probabilitățile tuturor rezultatelor sunt egal răspândite pe întreg spațiu fundamental**

- **Probabilitatea unui eveniment elementar este egală cu valoarea inversă numărului total de evenimente elementare**

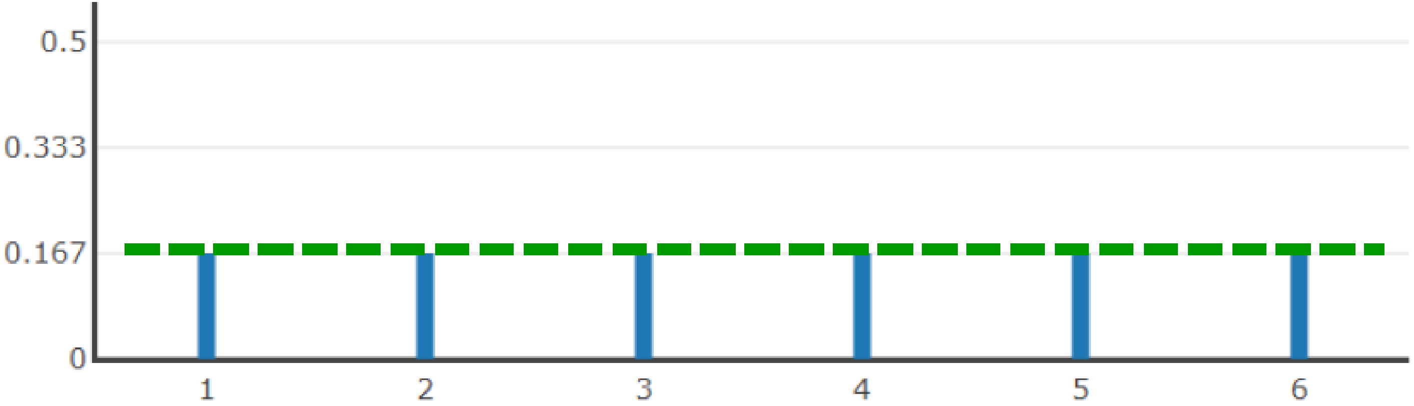
$$P(E) = \frac{1}{n_E}$$

Exemplu distribuție uniformă

- Aruncarea unui zar cu cifre poate avea 6 rezultate posibile de probabilitate egală.
- Rezultatul aruncării poate fi un număr de la 1 la 6 nicidecum 1,5
- Probabilitatea apariției unui dintre cele 6 numere se determină:

$$P(E) = \frac{1}{6} = 0{,}167$$

- Graficul distribuție la aruncarea zarului:



3. Distribuția binomială

Procese Bernoulli

- **Binomială presupune existența a două rezultate reciproc exclusive: succes și eșec**
- **Procesul Bernoulli este un experiment aliator în care pot exista doar 2 rezultate posibile: succes sau eșec**
- **O serie din *n* procese Bernoulli va avea o distribuție binară atâta timp cât:**
  - **probabilitatea *p* a succeselor este constantă**
  - **procesele sunt independente unul de altul**

Funcția masă de probabilitate binomială

- Se dă probabilitatea de observarea a *x* succese în *n* procese

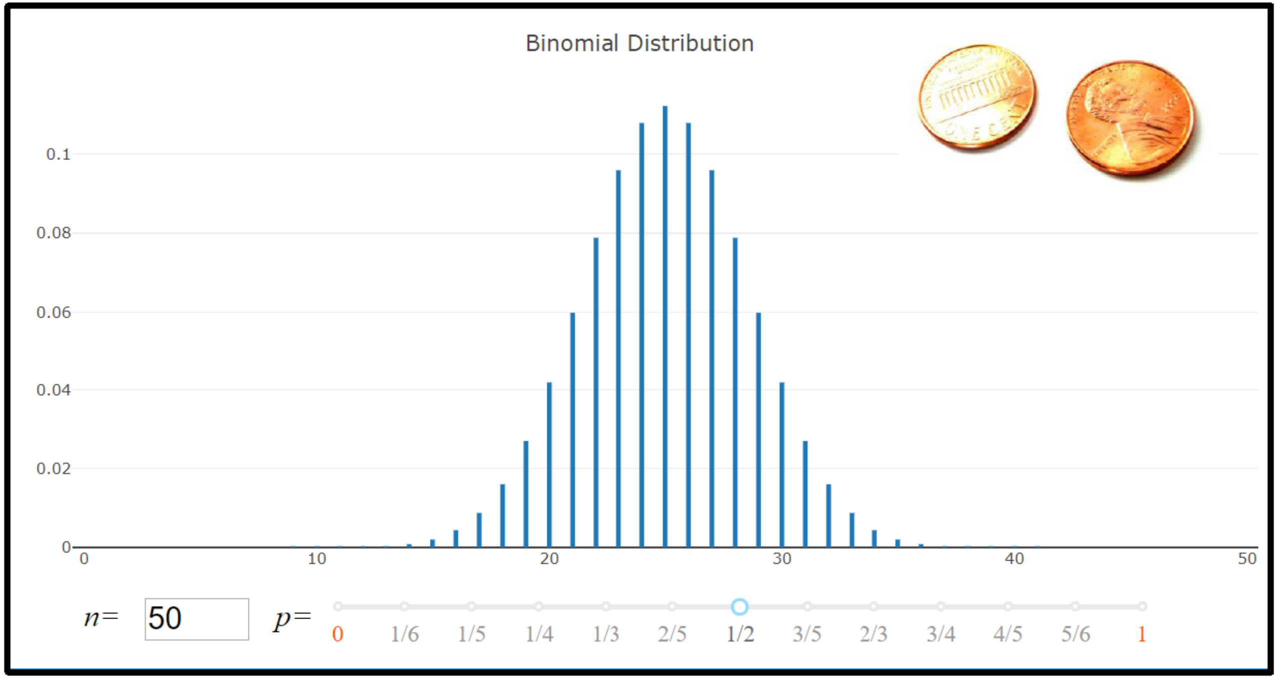
- Probabilitatea succesului la realizarea unui proces se consideră *p* și avea valoare constantă pentru toate procesele

- Funcția masă de probabilitate binomială va avea forma:

$$P(x: n, p) = \binom{n}{x} (p)^x (1-p)^{(n-x)}$$

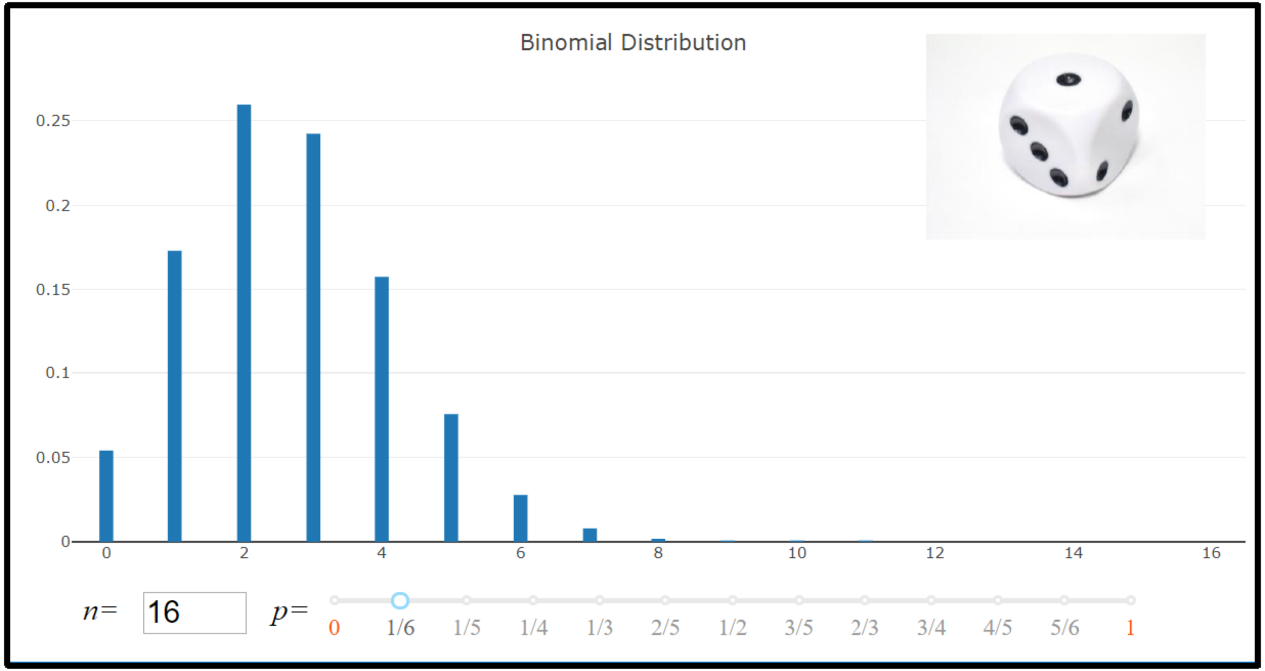
Grafic al distribuției binomiale (1)

- **Distribuția obținerii coroanei la aruncare de 50 de ori a monedei dacă probabilitatea apariție coroanei la o aruncare este 1/2**



Grafic al distribuției binomiale (2)

- Distribuția obținerii cifrei 6 la aruncare de 16 de ori a zarului cu probabilitatea apariție cifrei 6 la o aruncare este 1/6



Problemă distribuția binomială

- **Dacă se aruncă zarul de 16 ori, care este probabilitate că cifra 5 va apărea de 3 ori?**

$$x = 3 \qquad n = 16 \qquad p = \frac{1}{6}$$

$$P(x\colon n, p) = \binom{n}{x} (p)^x (1-p)^{(n-x)} =$$

$$= \left(\frac{n!}{x!\,(n-x)!}\right) (p)^x (1-p)^{(n-x)}$$

$$P\left(3\colon 16, \frac{1}{2}\right) = \left(\frac{16!}{3!\,(16-3)!}\right) \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^{(16-3)} = 0{,}242$$

Distribuția binomială în Python și Excel

- **Dacă se aruncă zarul de 16 ori, care este probabilitate că cifra 5 va apărea de 3 ori?**

- **Soluția problemei în Python:**

```
from scipy.stats import binom
p=binom.pmf(3,16,1/6)
print(p)
```

- **Soluția problemei în Excel:**

```
=BINOM.DIST(3,16,1/6,FALSE)
```

4. Distribuția Poisson

Esența distribuției Poisson

- Distribuția Poisson ia în considerarea numărul de succese pe o anumită unitate continuă la realizarea mai multor procese Bernoulli:

- Drept unitate continuă de cele mai multe ori se consideră timpul dar poate și alte mărimi, de exemplu lungimea etc.

- Deosebirea dintre distribuția binomială și distribuția Poisson constă în faptul că prima consideră numărul de succese în funcție de numărul de încercări pe când a doua consideră numărul de succese într-o unitate de timp.

Funcția masă de probabilitate Poisson

- **Determinarea funcției masei de probabilitate Poisson începe cu o valoarea medie așteptată**

$$E(X) = \mu$$

- **Această valoare se va considera mărimea “lambda” - λ**

$$\lambda = \frac{nr.\_apariti i}{interval\_timp} = \mu$$

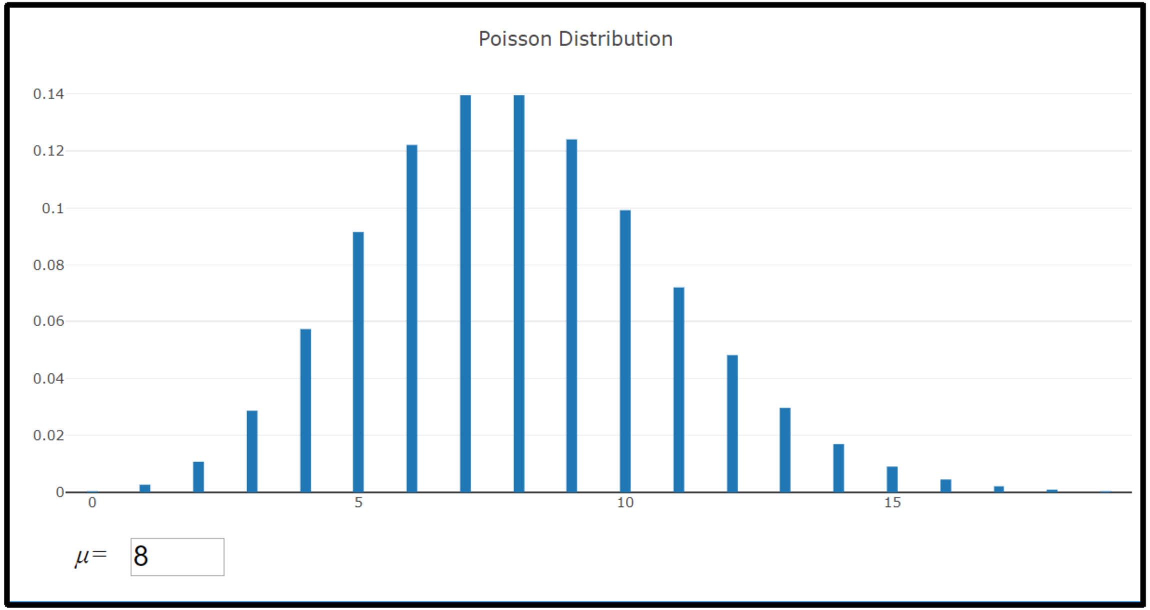
- **Funcția masei de probabilitate Poisson are forma:**

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$e = 2{,}71828\ldots$$

Grafic al distribuției Poisson

- **Distribuția numărului de mașini ce circulă timp de un minut printr-o intersecția dacă în mediu prin aceea intersecție circulă 8 mașini pe minut**



Problemă 1 distribuția Poisson

- **De obicei, timp de un minut printr-o intersecție dată circulă 8 mașini. Care este probabilitatea că această intersecție va fi circulată de exact 4 mașini timp de un minut.**

$$x = 4 \qquad \lambda = 8$$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(4) = \frac{8^4 \times 2{,}71828^{-8}}{4!} = \frac{4086 \times \left(\frac{1}{2980{,}95}\right)}{24} = 0{,}0572$$

Funcția masă cumulativă Poisson

- Funcția masei cumulative reprezintă suma funcțiilor maselor de probabilitate la îndeplinirea unor condiții

- Exemplu - probabilitatea apariției a mai puțin de 4 mașini în intersecție

$$P(X: x < 4) = \sum_{i=0}^{3} \frac{\lambda^i e^{-\lambda}}{i!} = \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!}$$

- Exemplu - probabilitatea apariției a cel puțin 4 mașini în intersecție

$$P(X: x \geq 4) = 1 - P(X: x < 4)$$

Problemă 2 distribuția Poisson

- **De obicei, timp de un minut printr-o intersecție dată circulă 8 mașini. Care este probabilitatea că această intersecție va fi circulată de cel mult 2 mașini timp de un minut.**

$$P(X: x < 3) = \sum_{i=0}^{2} \frac{\lambda^i e^{-\lambda}}{i!} = \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} =$$

$$= \frac{8^0 \times 2{,}71828^{-8}}{0!} + \frac{8^1 \times 2{,}71828^{-8}}{1!} + \frac{8^2 \times 2{,}71828^{-8}}{2!} =$$

$$= \frac{1 \times \left(\frac{1}{2980{,}95}\right)}{1} + \frac{8 \times \left(\frac{1}{2980{,}95}\right)}{1} + \frac{64 \times \left(\frac{1}{2980{,}95}\right)}{2} = 0{,}0137$$

Distribuția Poisson – intervale parțiale

- **Distribuția Poisson presupune că probabilitatea succeselor pe durata unor intervale mai mici este proporțională cu probabilitatea întregului interval**

- **Dacă se cunoaște valoarea așteptată λ pe un interval de o oră atunci valoarea așteptată pe un interval de o minuta se va considera:**

$$\lambda_{1\ minut} = \frac{\lambda_{1\ oră}}{60}$$

Problemă 3 distribuția Poisson

- **De obicei, timp de un minut printr-o intersecție dată circulă 8 mașini. Care este probabilitatea că timp de 10 secunde nu va circula nici o mașină.**

$$x = 4 \qquad \lambda_{1\,minut} = 8$$

$$\lambda_{10\,secunde} = \frac{\lambda_{1\,minut}}{\frac{60s}{10s}} = \frac{8}{6} = 1{,}34$$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(0) = \frac{1{,}34^0 \times 2{,}71828^{-1{,}34}}{0!} = 0{,}2618$$

Distribuția Poisson în Python

- De obicei, timp de un minut printr-o intersecție dată circulă 8 mașini. a) Care este probabilitatea că această intersecție va fi circulată de exact 4 mașini timp de un minut.

```
from scipy.stats import poisson
p=poisson.pmf(4,8)
print(p)
```

- b) Care este probabilitatea că această intersecție va fi circulată de cel mult 2 mașini timp de un minut.

```
from scipy.stats import poisson
p=poisson.cdf(2,8)
print(p)
```

- c) Care este probabilitatea că timp de 10 secunde nu va circula nici o mașină.

```
from scipy.stats import poisson
p=poisson.pmf(0,8/6)
print(p)
```

Distribuția Poisson în Excel

- De obicei, timp de un minut printr-o intersecție dată circulă 8 mașini. a) Care este probabilitatea că această intersecție va fi circulată de exact 4 mașini timp de un minut.

=POISSON.DIST(4,8,FALSE)

- b) Care este probabilitatea că această intersecție va fi circulată de cel mult 2 mașini timp de un minut.

=POISSON.DIST(2,8,TRUE)

- c) Care este probabilitatea că timp de 10 secunde nu va circula nici o mașină.

=POISSON.DIST(0,8/6,FALSE)

5. Distribuția normală

Esența distribuției normale

- Distribuția normală este un tip de distribuție continuă a probabilităților conform căreia cele mai mari probabilități sunt atribuite valorilor apropiate valorii medii și cu depărtarea de valoarea medie probabilitatea scade
- Distribuția normală este foarte des întâlnită în viața reală: masa și înălțimea oamenilor, erorile de măsurare, rezultatele testelor de examinare, etc
- Distribuția normală se mai numește distribuția lui Gauss sau clopotul lui Bell

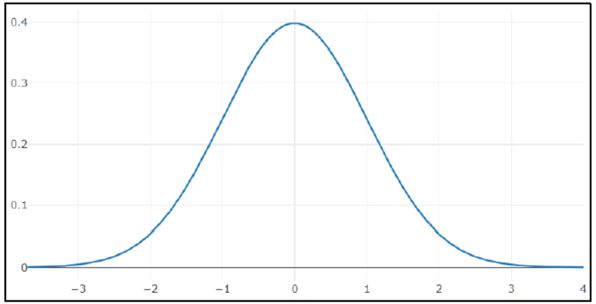
Funcția densității probabilității

- **Dacă setul de date este distribuit normat atunci datele se vor avea probabilități ce pot fi determinate în funcție de media aritmetică și abaterea standard**

- **Funcția densității probabilității are forma:**

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

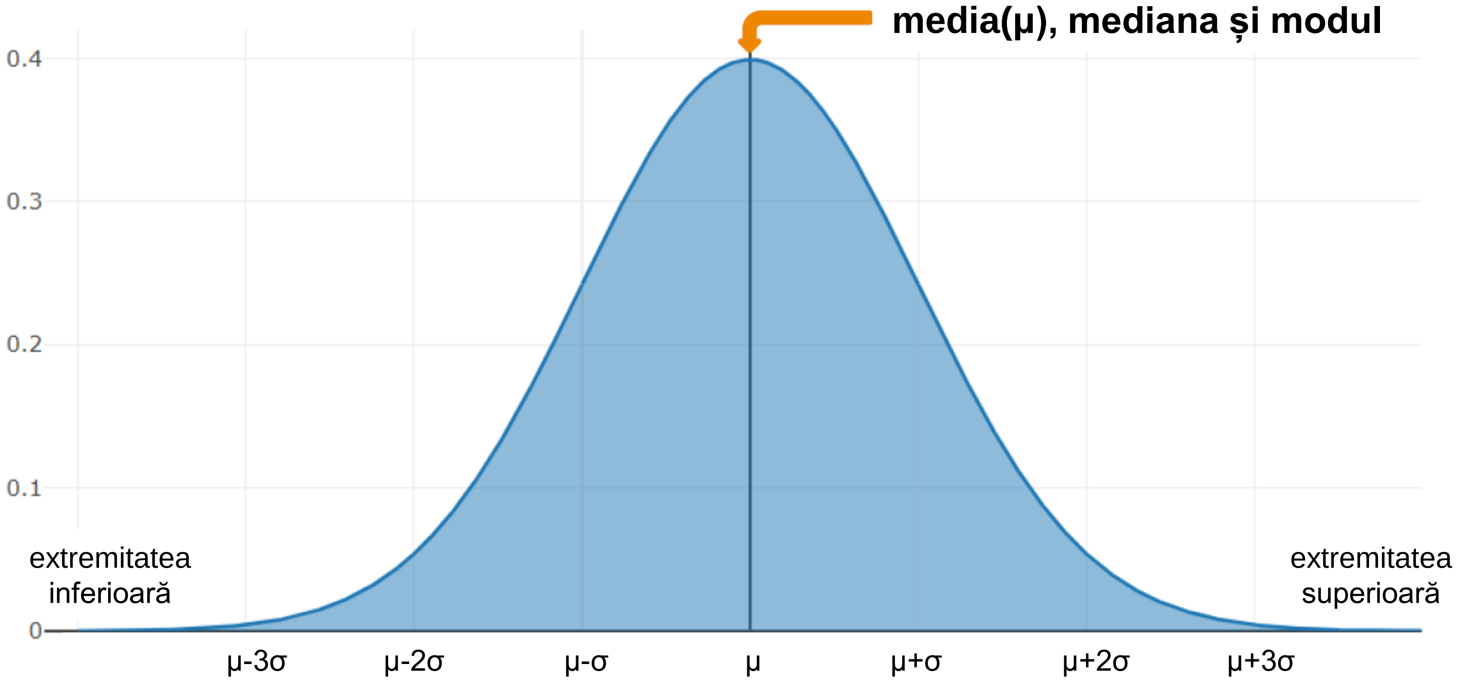
- **Graficul distribuției normale:**



Particularitățile distribuției normale

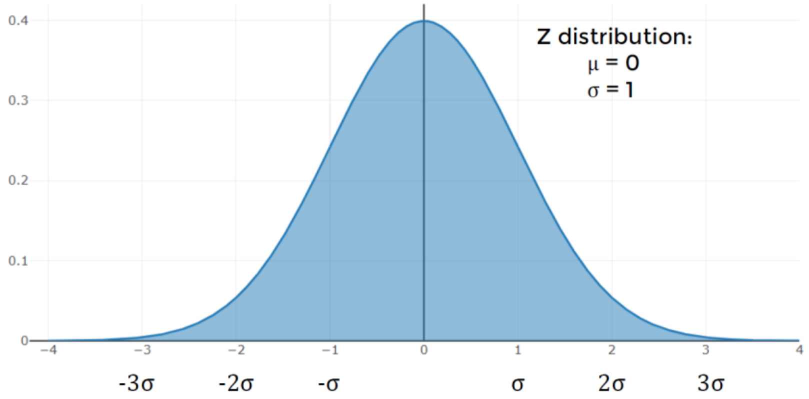
- Aria de sub curba distribuției normate este 1
- Valoarea medie, mediana și modulul au aceleași valori
- Graficul curbei atinge maximul pentru valoarea media și este întotdeauna simetric față de aceasta
- Curba graficului se va apropia de valoarea 0 dar niciodată nu va lua această valoare
- Regiunile graficului unde acesta se apropie de valoarea o se numesc extremități inferioare și superioare
- Probabilitatea unui anumit rezultat este 0 și poate fi determinată doar probabilitatea unui interval de rezultate

Graficul distribuției normale



Distribuția normală standard

- **Distribuția normală standard numită și distribuția Z este un caz particular al distribuției normale pentru care μ=0 și ρ=1**
- **Distribuția normală standard a fost larg studiată și există tabele care oferă zone sub curbă se tinde a transforma oricare altă distribuție normată în acest tip de distribuție prin procedura de standardizare cu ajutorul scorului z**

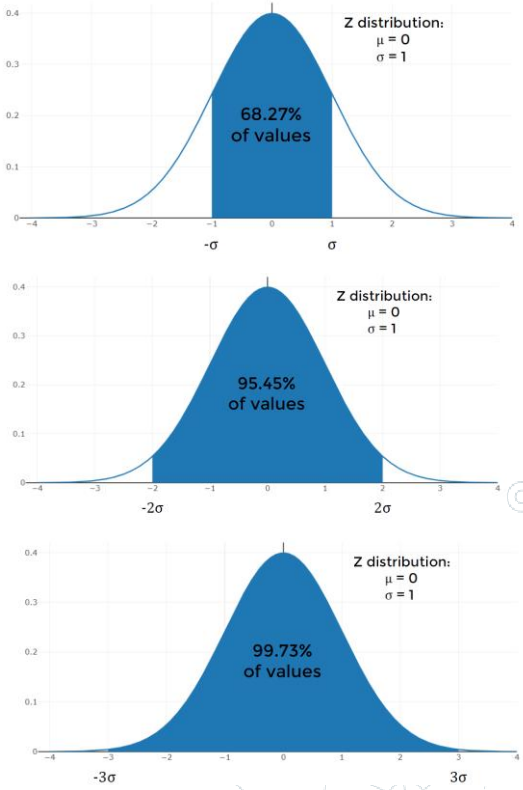


Intervalele definite de valori

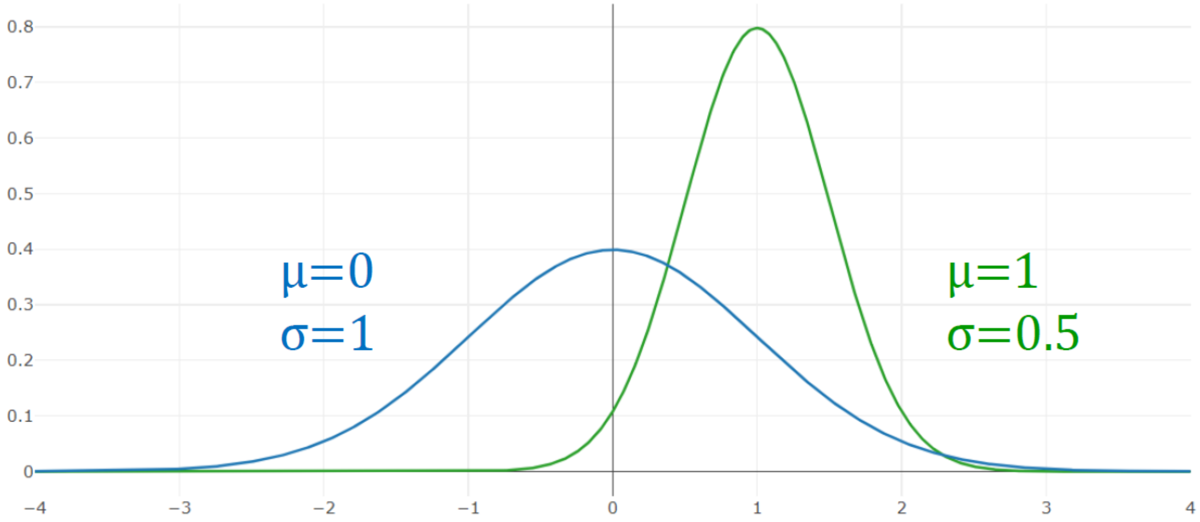
- Există o probabilitate de 68,27% că o valoare să se afle în intervalul –σ...σ

- Există o probabilitate de 95,45% că o valoare să se afle în intervalul –2σ...2σ

- Există o probabilitate de 99,73% că o valoare să se afle în intervalul –3σ...23σ



Deosebiri dintre distribuția normată și distribuția normată standard

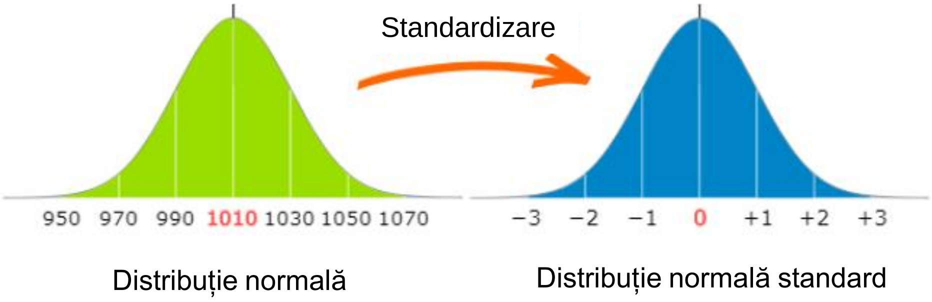


6. Scorul z

Necesitatea scorului z

- **Scopul determinării scorului z este de a relaționa o distribuție normală particulară cu distribuția normală standard**

- **Procedura de relaționarea a unei distribuții normale cu distribuția normală standard se numește standardizare iar necesitatea ei reiese din universalitatea distribuției standard**



Formula scorului z

- Pentru determinarea scorului z a unei valori specifice x al unei distribuții normate particulare se utilizează relația

$$z = \frac{x - \mu}{\sigma}$$

- Scorul z reprezintă numărul de abaterii standard de la media distribuției particulare.

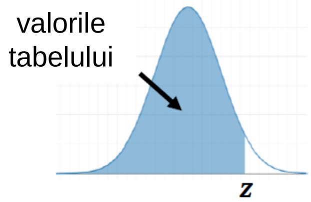
- Scorul z se utilizează pentru determinarea percentilei valorii x

- Percentila reprezintă o cale de a spun “câte procente din date se includ sub această valoare”

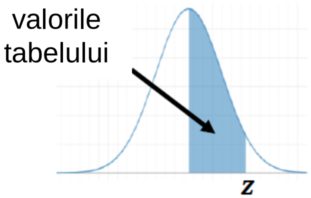
- Pentru determinarea percentilei conform scorului z se utilizează tabelul z a probabilităților normale standarde

Tabele z a probabilităților normale standarde

- **Tabelul z a probabilităților normale standarde conține valorile ariei suprafeței de sub curba plasată la stânga valorii scorului z specificat**
- **În funcție de scopul calculelor tabelele sunt de 2 tipuri:**



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141

Valoarea din tabel se va considera citindu-se cifrele 1 și 2 a scorului z din coloana z și cifra 3 de pe linia z

De exemplu pentru z= 0,25 în primul tabel se citește 0,2 din coloana z și 0,05 de pe linia z iar la intersecție se obține valoarea percentilei p=0,5987

Problemă 1 scorul z

- **Un student a obținut 87 de puncte din 100 la testul de evaluare curentă. Care este procentajul de reușită a acestui student comparativ cu ceilalți dacă la acest test studenții obțin în mediul 75 de puncte cu o abatere standard de 7 puncte.**

$$x = 87 \qquad \mu = 75 \qquad \sigma = 7$$

Se determină scorul z

$$z = \frac{x - \mu}{\sigma} = \frac{87 - 75}{7} = 1{,}71$$

Conform valorii lui z din tabelul z se determină percentila p și respectiv procentajul

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706

$$p = 0{,}9564 \Rightarrow 95{,}64\%$$

Problemă 2 scorul z

- **Un student a obținut o reușită de 96,99% la rezolvarea unui test pentru care în mediu studenții obțin în mediul 75 de puncte cu o abatere standard de 7. Câte puncte a obținut studentul?**

$$p = 96{,}99 \qquad \mu = 75 \qquad \sigma = 7$$

Se determină scorul z utilizând tabelul z

$$z = 1{,}88$$

Se determină valoare lui x din relația

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu = 1{,}88 \times 7 + 75 = 88{,}16 \approx 88$$

Scorul z în Python

- **Care este valoarea percentilei pentru scorul z=0,7**

```
from scipy.stats import stats
z=0.7
p=stats.norm.cdf(z)
print(p)
```

- **Care este valoarea scorului z dacă percentila are valoarea p = 0,95**

```
from scipy.stats import stats
p=0.95
z=stats.norm.ppf(p)
print(z)
```

Scorul z în Excel

- Care este valoarea percentilei pentru scorul z=0,7

=NORMSDIST(0.70)

- Care este valoarea scorului z dacă percentila are valoarea p = 0,95

=NORMSINV(0.95)