Calcul numeric

Metode Bayesiene

Paul Irofti
Andrei Pătrașcu
Cristian Rusu

Departmentul de Informatică

Facultatea de Matematică și Informatică

Universitatea din București

Email: prenume.nume@fmi.unibuc.ro

Cuprins

- rezultate probabilistice
    - clasificare
    - regresie
- gândirea probabilistică
- gândirea probabilistică pentru analiza datelor
- avantaje/dezvantaje ale metodei Bayesiene

Rezultate probabilistice

- sunt situații în care vrem să facem o predicție și să oferim și un interval de încredere (cât de încrezători suntem că rezultatul este corect)

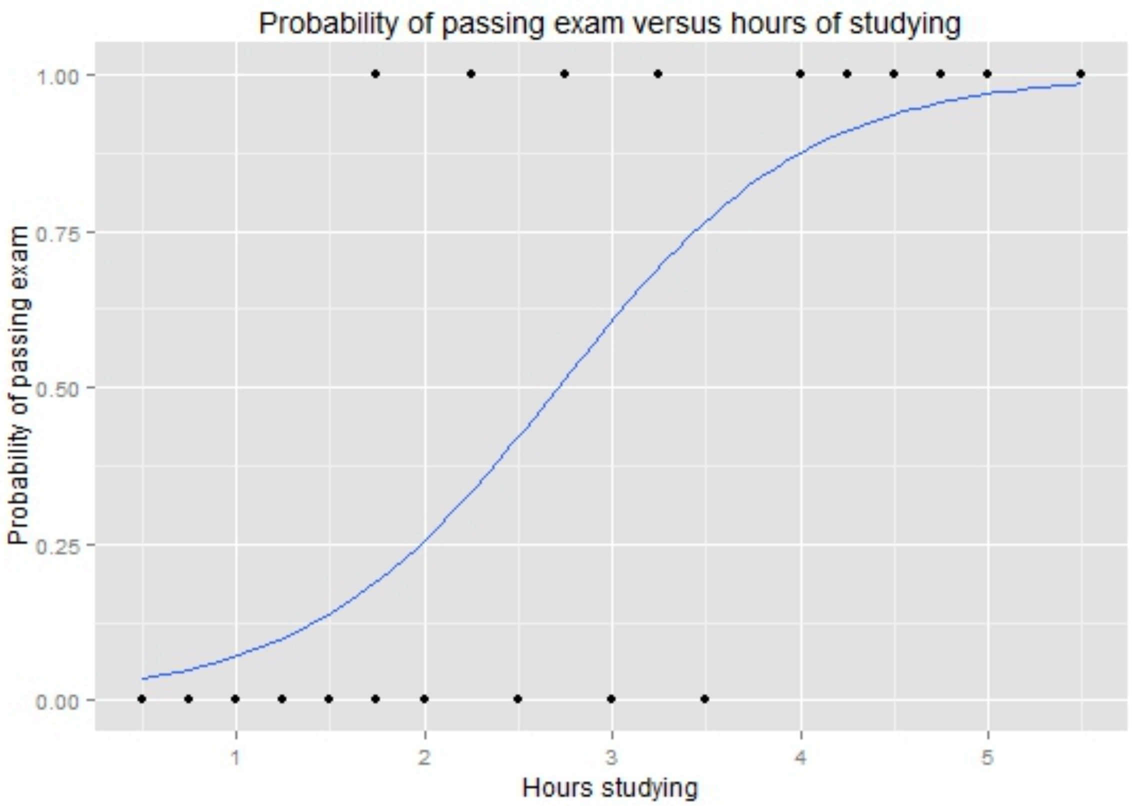
- de exemplu:
    - într-o problemă de clasificare cats/dogs vrem să spunem: suntem 90% siguri că imaginea clasificată este un câine

    - într-o problemă de regresie în care estimăm prețul de vânzare a unui apartament vrem să spunem: prețul este $100\,000 \pm 5000$

Rezultate probabilistice

- clasificare
  - aici lucrurile sunt relative simple
  - funcția sigmoid $f(x) = \dfrac{1}{1 + e^{-x}}$
  - rezultatul poate fi interpretat

ca o probabilitate de a aparține unei clase



https://en.wikipedia.org/wiki/Logistic_regression

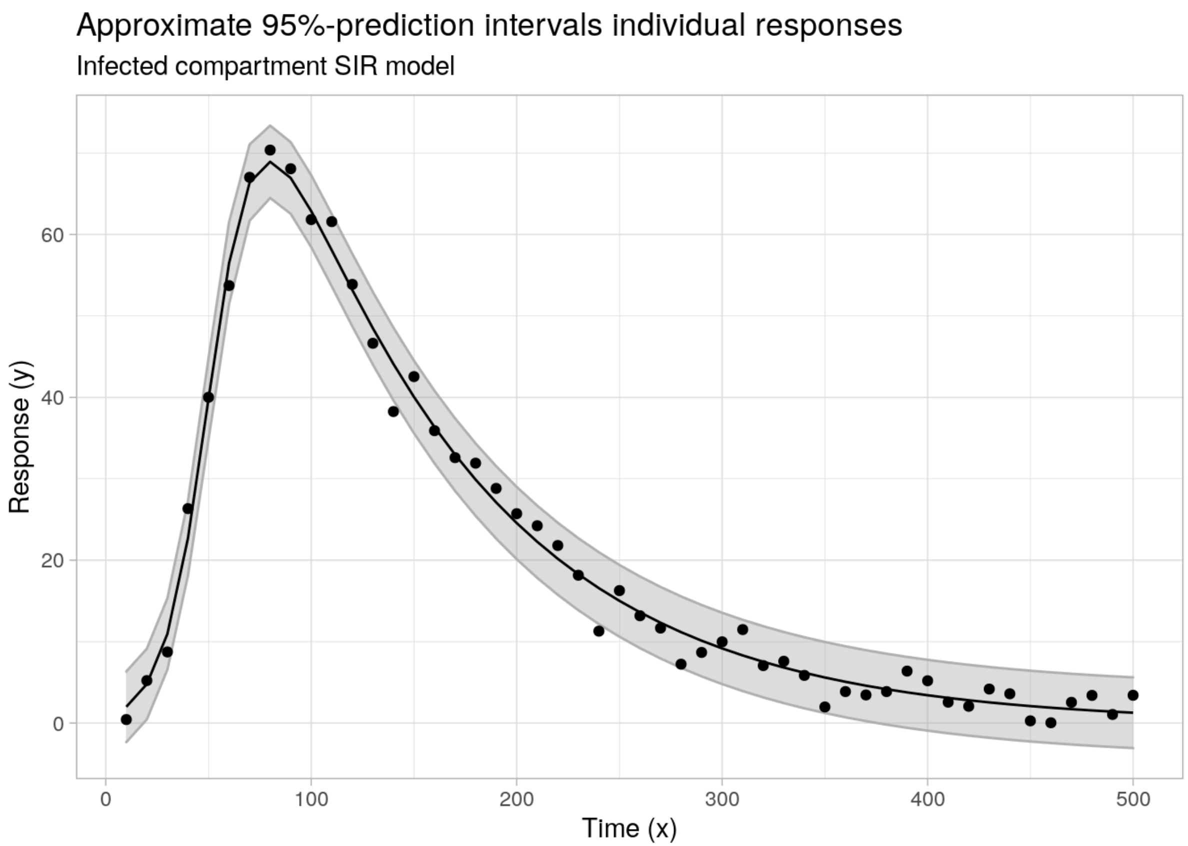
Rezultate probabilistice

- regresie
  - modelul este media
  - dar raportăm și deviația standard

  - deviația standard oferă gradul

de încredere pe care îl avem în

media raportată de model

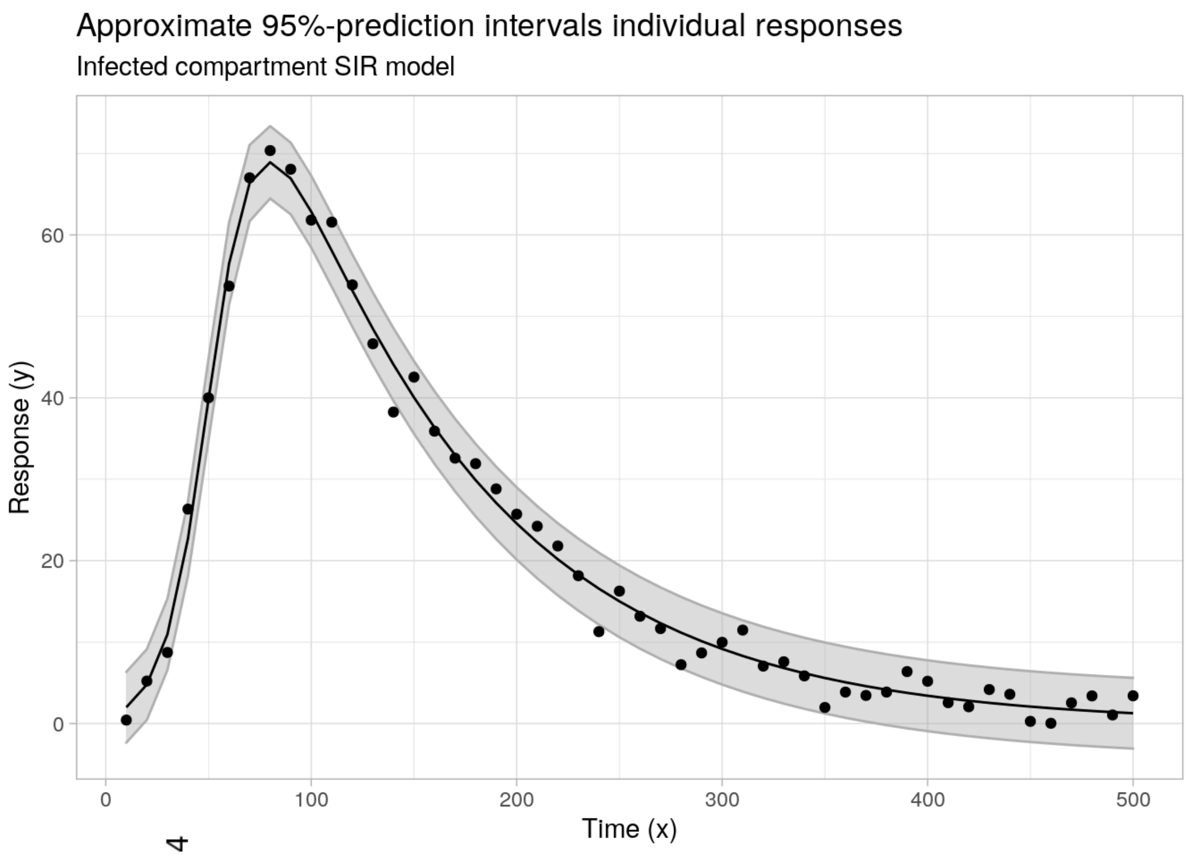


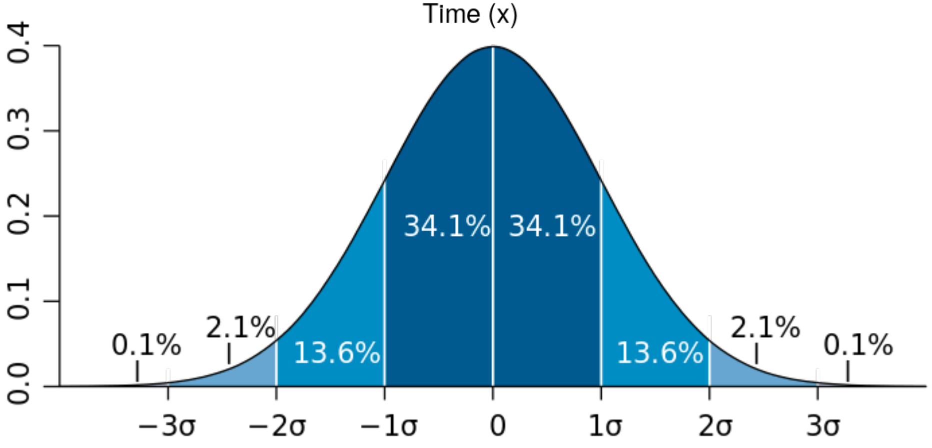
https://www.r-bloggers.com/2021/07/asymptotic-confidence-intervals-for-nls-regression-in-r/

Rezultate probabilistice

- regresie
  - modelul este media
  - dar raportăm și deviația standard

  - deviația standard oferă gradul de încredere pe care îl avem în media raportată de model





https://www.r-bloggers.com/2021/07/asymptotic-confidence-intervals-for-nls-regression-in-r/

Gândirea probabilistică

- credem ceva (cu o anumită probabilitate)
- observăm ceva (un eveniment se întâmplă)
- actualizăm ce credem (actualizăm probabilitatea)

Gândirea probabilistică

- exemplu:
  - facem un test pentru o boală care are probabilitatea de apariție în populație de 0.1%
  - testul este corect în 99% din cazuri în care pacientul este bolnav
  - testul greșește în 1% din cazurile în care pacientul nu este bolnav

  - primim un test pozitiv, care e probabilitatea că pacientul are boala?

Gândirea probabilistică

- H = ipoteza că avem boala (*the hypothesis*)
- E = evenimentul că vedem testul pozitiv (*the event*)

- P(E) = probabilitatea că testul este pozitiv
- P(H) = probabilitatea că suntem bolnavi (înainte de orice test/eveniment)

- P(H | E) = probabilitatea că avem boala dacă testul este pozitiv
- P(E | H) = probabilitatea că testul este pozitiv dacă chiar suntem bolnavi

Gândirea probabilistică

- P(E) = probabilitatea că testul este pozitiv
- P(H) = probabilitatea că suntem bolnavi (înainte de test) 0.1%

- P(H | E) = probabilitatea că avem boala dacă testul este pozitiv
- P(E | H) = probabilitatea că testul este pozitiv dacă suntem bolnavi 99%

Gândirea probabilistică

- P(E) = probabilitatea că testul este pozitiv
- P(H) = probabilitatea că suntem bolnavi (înainte de test) 0.1%

- P(H | E) = probabilitatea că avem boala dacă testul este pozitiv
- P(E | H) = probabilitatea că testul este pozitiv dacă suntem bolnavi 99%

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} = \frac{P(E|H) \times P(H)}{P(H) \times P(E|H) + P(\neg H) \times P(E|\neg H)}$$

Gândirea probabilistică

- P(E) = probabilitatea că testul este pozitiv
- P(H) = probabilitatea că suntem bolnavi (înainte de test) 0.1%

- P(H | E) = probabilitatea că avem boala dacă testul este pozitiv
- P(E | H) = probabilitatea că testul este pozitiv dacă suntem bolnavi 99%

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(H) \times P(E|H) + P(\neg H) \times P(E|\neg H)} = \frac{0.99 \times 0.001}{0.001 \times 0.99 + 0.999 \times 0.01} \approx 9\,\%$$

Gândirea probabilistică

- P(E) = probabilitatea că testul este pozitiv
- P(H) = probabilitatea că suntem bolnavi (înainte de test) 0.1%

- P(H | E) = probabilitatea că avem boala dacă testul este pozitiv
- P(E | H) = probabilitatea că testul este pozitiv dacă suntem bolnavi 99%

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(H)\times P(E|H) + P(\neg H)\times P(E|\neg H)} = \frac{0.99 \times 0.001}{0.001\times 0.99 + 0.999\times 0.01} \approx 9\,\%$$

pare un rezultat foarte prost (nu sunt deloc siguri că avem boala) … ce putem face mai departe?

Gândirea probabilistică

- P(E) = probabilitatea că testul este pozitiv
- P(H) = probabilitatea că suntem bolnavi (înainte de test) 9%

- P(H | E) = probabilitatea că avem boala dacă testul este pozitiv
- P(E | H) = probabilitatea că testul este pozitiv dacă suntem bolnavi 99%

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(H) \times P(E|H) + P(\neg H) \times P(E|\neg H)} = \frac{0.99 \times 0.09}{0.09 \times 0.99 + 0.91 \times 0.01} \approx 91\,\%$$

mai facem un test, dar de data aceasta folosim rezultatul anterior

Gândirea probabilistică

Teorema lui Bayes

- P(E) = probabilitatea că testul este pozitiv
- P(H) = probabilitatea că suntem bolnavi (înainte de test) apriori

- P(H | E) = probabilitatea că avem boala dacă testul este pozitiv aposteriori
- P(E | H) = probabilitatea că testul este pozitiv dacă suntem bolnavi

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(H) \times P(E|H) + P(\neg H) \times P(E|\neg H)}$$

Gândirea probabilistică

- credem ceva (cu o anumită probabilitate)

- observăm ceva (un eveniment se întâmplă)

- actualizăm ce credem (actualizăm probabilitatea)

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \propto P(E|H) \times P(H)$$

Gândirea probabilistică pentru analiza datelor

- avem nevoie de 3 ingrediente:
  - date (multe)
  - un model generativ (cum se schimbă lucrurile când vedem date)
  - model apriori (ce credem înainte să vedem datele)

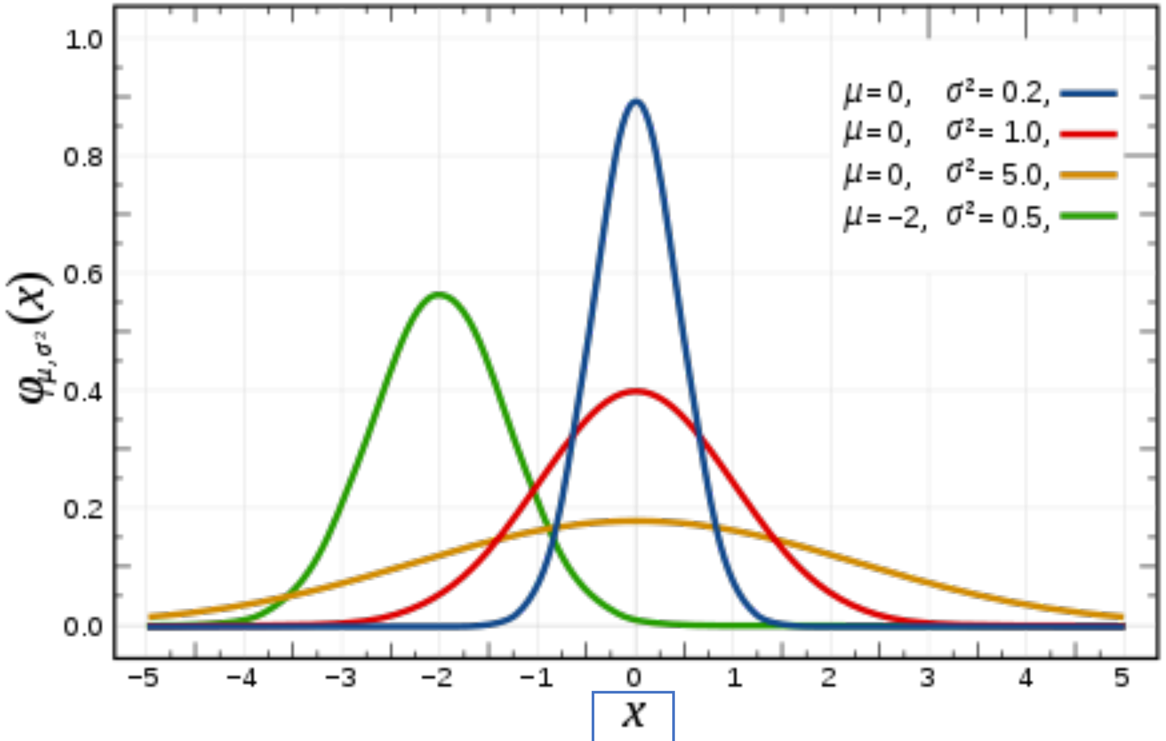
Gândirea probabilistică pentru analiza datelor

- un model generativ

alegem niște parametrii pentru modelul nostru

model generativ

date generate



-3.1, -2.3, -2.1, -1.7, -2.5, -1.9 …

https://en.wikipedia.org/wiki/Normal_distribution

Gândirea probabilistică pentru analiza datelor

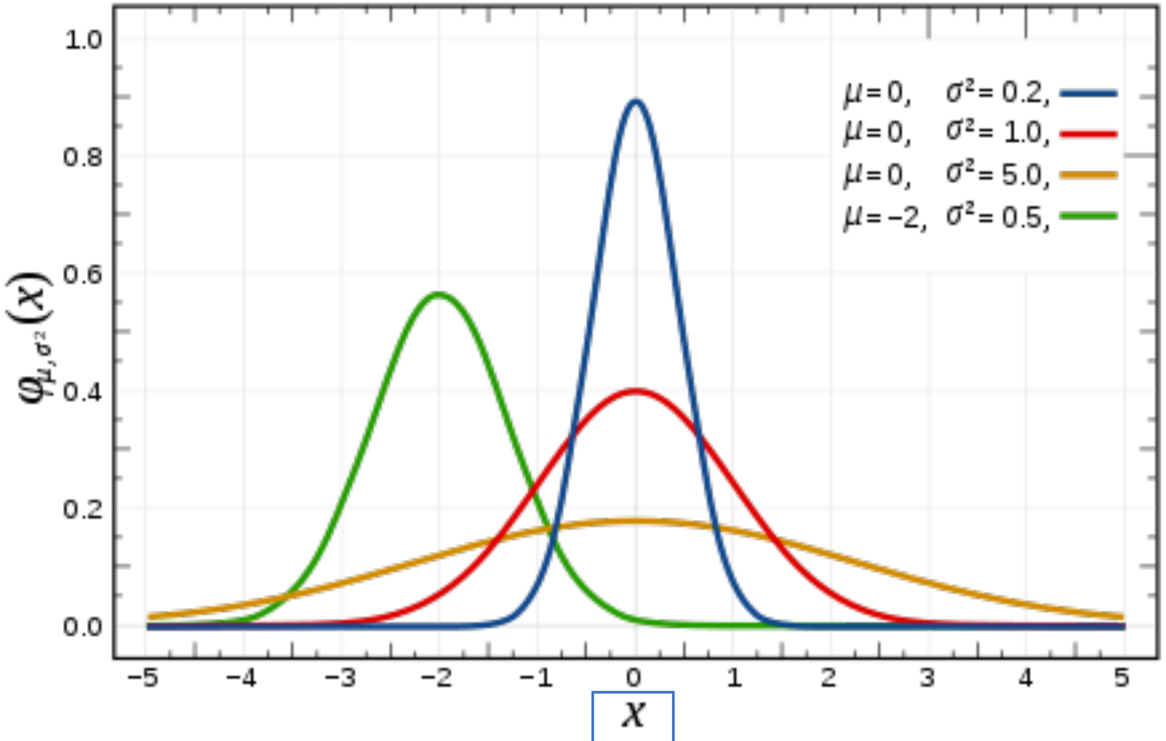
- un model generativ

alegem niște parametrii pentru modelul nostru

model generativ

date generate

asta e ușor, dacă știm parametrii μ și σ



-3.1, -2.3, -2.1, -1.7, -2.5, -1.9 …

https://en.wikipedia.org/wiki/Normal_distribution

Gândirea probabilistică pentru analiza datelor

-3.1, -2.3, -2.1, -1.7, -2.5, -1.9 …

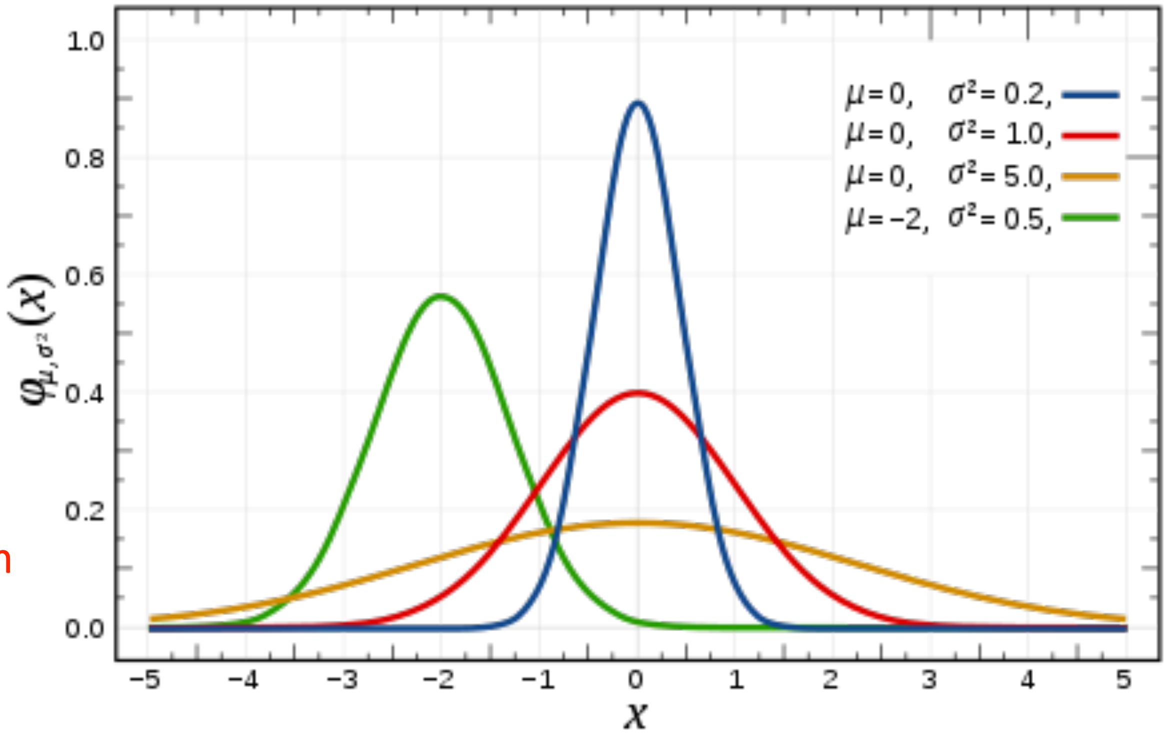
- un model generativ

date generate

model generativ

alegem niște parametrii pentru modelul nostru

în analiza datelor, avem datele dar nu avem parametrii distribuției care i-au generat (este problema inversă)



https://en.wikipedia.org/wiki/Normal_distribution

Gândirea probabilistică pentru analiza datelor

- exemplu: o firmă de marketing vrea să știe cât de eficientă va fi noua campanie de reclame pe care a dezvoltat-o și alege 16 persoane aleatoare (din populație) cărora le arată reclame iar la sfârșit îi întreabă dacă ar cumpăra sau nu noul produs. 6 persoane declară “da”

- avem nevoie de 3 ingrediente:
  - date (multe)
  - un model generativ (cum se schimbă lucrurile când vedem date)
  - model apriori (ce credem înainte să vedem datele)

Gândirea probabilistică pentru analiza datelor

- exemplu: o firmă de marketing vrea să știe cât de eficientă va fi noua campanie de reclame pe care a dezvoltat-o și alege 16 persoane aleatoare (din populație) cărora le arată reclame iar la sfârșit îi întreabă dacă ar cumpăra sau nu noul produs. 6 persoane declară "da"

- avem nevoie de 3 ingrediente:
  - date (multe)
  - un model generativ (cum se schimbă lucrurile când vedem date)
  - model apriori (ce credem înainte să vedem datele)

Gândirea probabilistică pentru analiza datelor

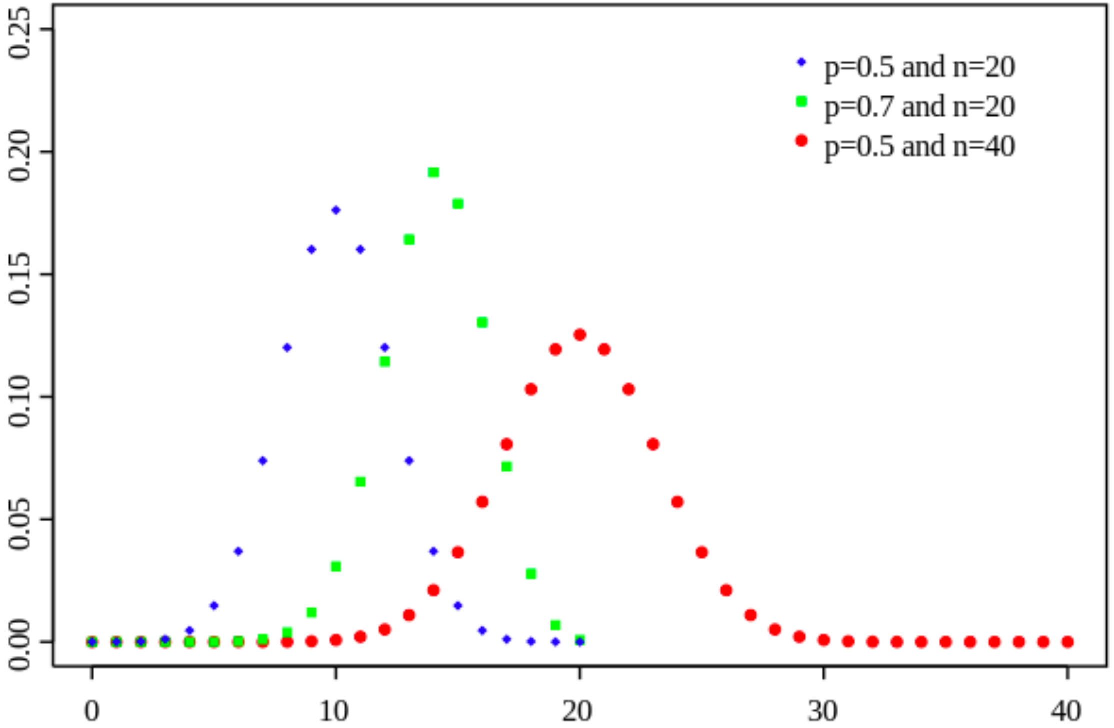
- exemplu: o firmă de marketing vrea să știe cât de eficientă va fi noua campanie de reclame pe care a dezvoltat-o și alege 16 persoane aleatoare (din populație) cărora le arată reclame iar la sfârșit îi întreabă dacă ar cumpăra sau nu noul produs. 6 persoane declară “da”

- avem nevoie de 3 ingrediente:
  - date (multe)
  - un model generativ (cum se schimbă lucrurile când vedem date)
  - model apriori (ce credem înainte să vedem datele)

Gândirea probabilistică pentru analiza datelor

- exemplu: o firmă de marketing vrea să știe cât de eficientă va fi noua campanie de reclame pe care a dezvoltat-o și alege 16 persoane aleatoare (din populație) cărora le arată reclame iar la sfârșit îi întreabă dacă ar cumpăra sau nu noul produs. 6 persoane declară “da”

- avem nevoie de 3 ingrediente:
  - un model generativ “da” / “nu”
  - distribuție Binomială $B(n, p)$
  - la noi $n = 16, p = 6/16 = 0.375$



https://en.wikipedia.org/wiki/Binomial_distribution

Gândirea probabilistică pentru analiza datelor

- exemplu: o firmă de marketing vrea să știe cât de eficientă va fi noua campanie de reclame pe care a dezvoltat-o și alege 16 persoane aleatoare (din populație) cărora le arată reclame iar la sfârșit îi întreabă dacă ar cumpăra sau nu noul produs. 6 persoane declară "da"

- avem nevoie de 3 ingrediente:
  - date (multe)
  - un model generativ (cum se schimbă lucrurile când vedem date)
  - model apriori (ce credem înainte să vedem datele)

aici nu e clar, putem presupune că nu știm nimic, sau ne putem baza pe campanii similare de marketing din trecut

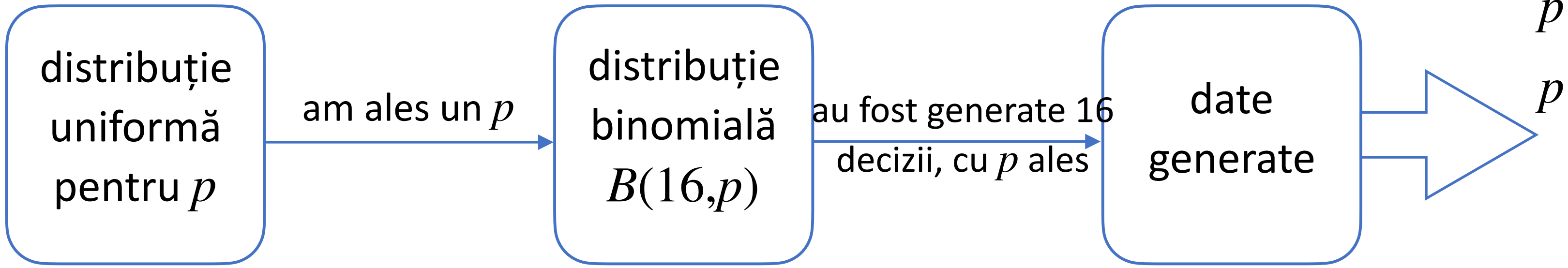
Gândirea probabilistică pentru analiza datelor

- exemplu: o firmă de marketing vrea să știe cât de eficientă va fi noua campanie de reclame pe care a dezvoltat-o și alege 16 persoane aleatoare (din populație) cărora le arată reclame iar la sfârșit îi întreabă dacă ar cumpăra sau nu noul produs. 6 persoane declară “da”

- avem nevoie de 3 ingrediente:
  - date (multe): $n = 16$ persoane, 6 “da”
  - un model generativ: $B(16,p)$, iar p este parametrul
  - model apriori: credem că p este uniform distribuit în [0, 1]

Gândirea probabilistică pentru analiza datelor

- avem nevoie de 3 ingrediente:
  - date (multe): $n = 16$ persoane, 6 “da”
  - un model generativ: $B(16,p)$, iar p este parametrul
  - model apriori: credem că p este uniform distribuit în [0, 1]



$p = 0.21$, "da" s-a întâmplat $5/16$

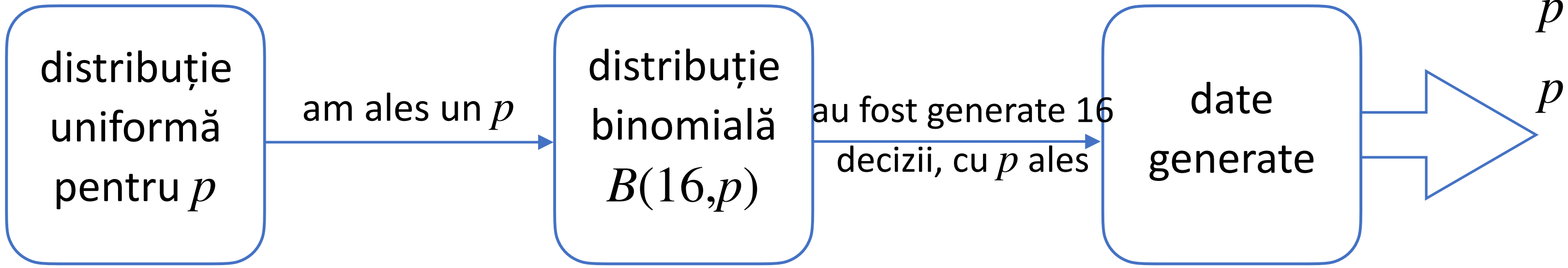
$p = 0.45$, "da" s-a întâmplat $6/16$

$p = 0.21$, "da" s-a întâmplat $6/16$

$p = 0.37$, "da" s-a întâmplat $6/16$

Gândirea probabilistică pentru analiza datelor

- avem nevoie de 3 ingrediente:
  - date (multe): $n = 16$ persoane, 6 “da”
  - un model generativ: $B(16,p)$, iar p este parametrul
  - model apriori: credem că p este uniform distribuit în [0, 1]



$p = 0.21$, "da" s-a întâmplat $5/16$ ✗

$p = 0.45$, "da" s-a întâmplat $6/16$ ✓

$p = 0.21$, "da" s-a întâmplat $6/16$ ✓

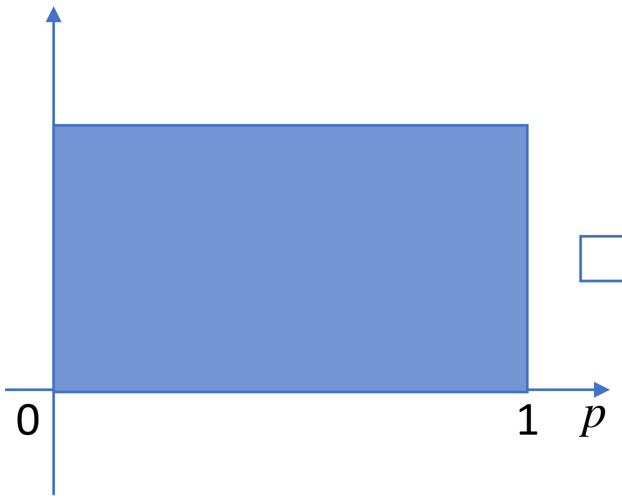
⋮

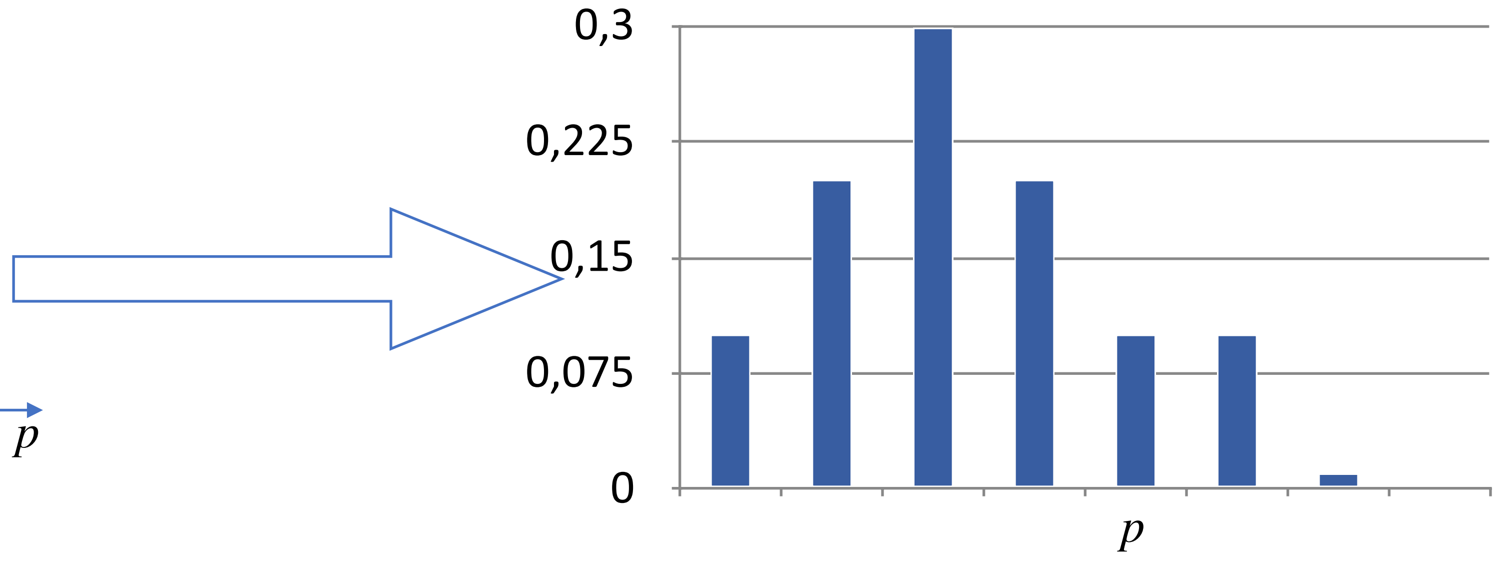
$p = 0.37$, "da" s-a întâmplat $6/16$ ✓

păstrăm doar p-urile pentru care obținem 6/16

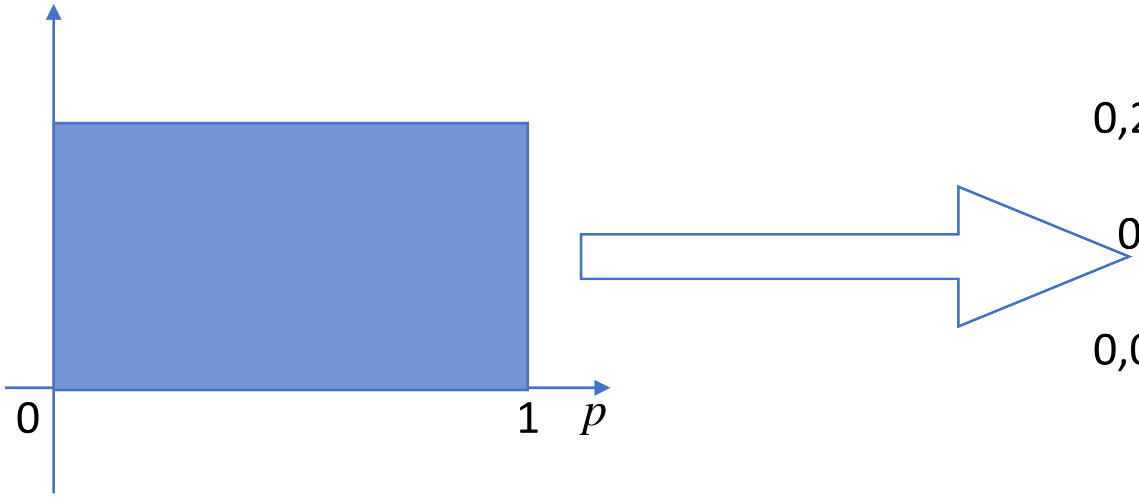
Gândirea probabilistică pentru analiza datelor

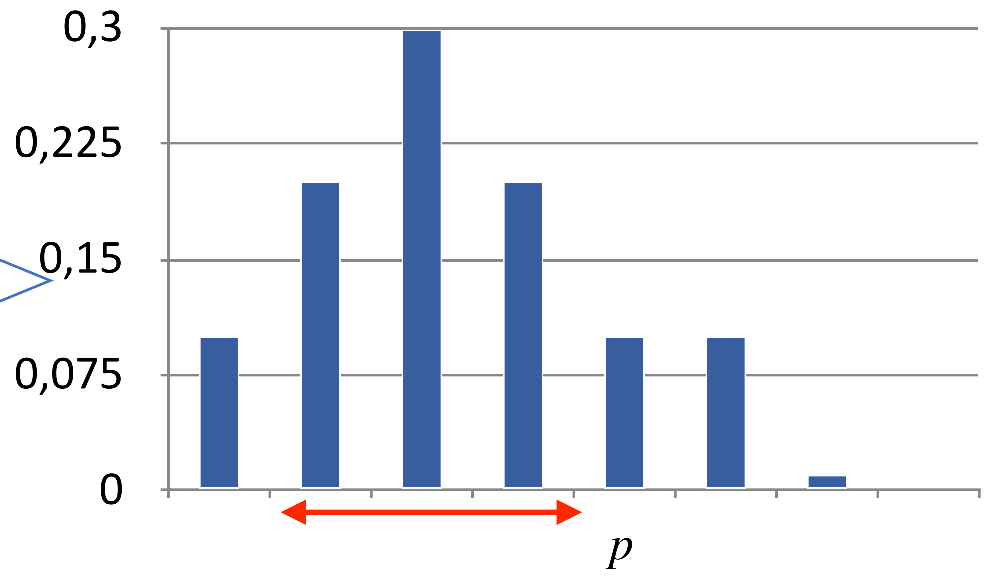
- ce am făcut?
  - am pornit de la o distribuție uniformă pentru p (apriori)
  - am ajuns la o altă distribuție pentru p (aposteriori)
  - ca să ajungem de la una la alta am folosit un model generativ





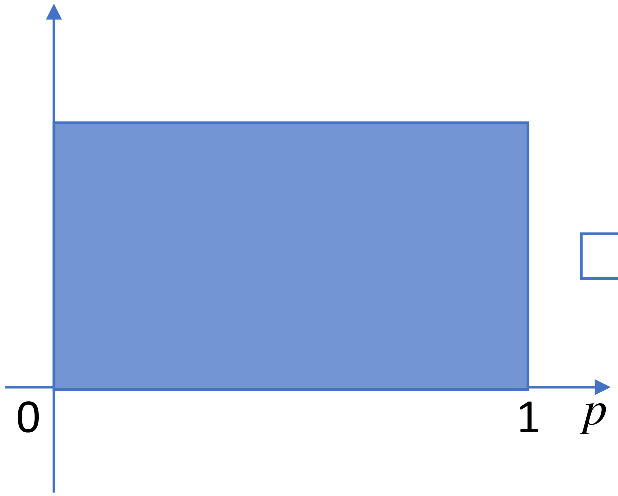
Gândirea probabilistică pentru analiza datelor

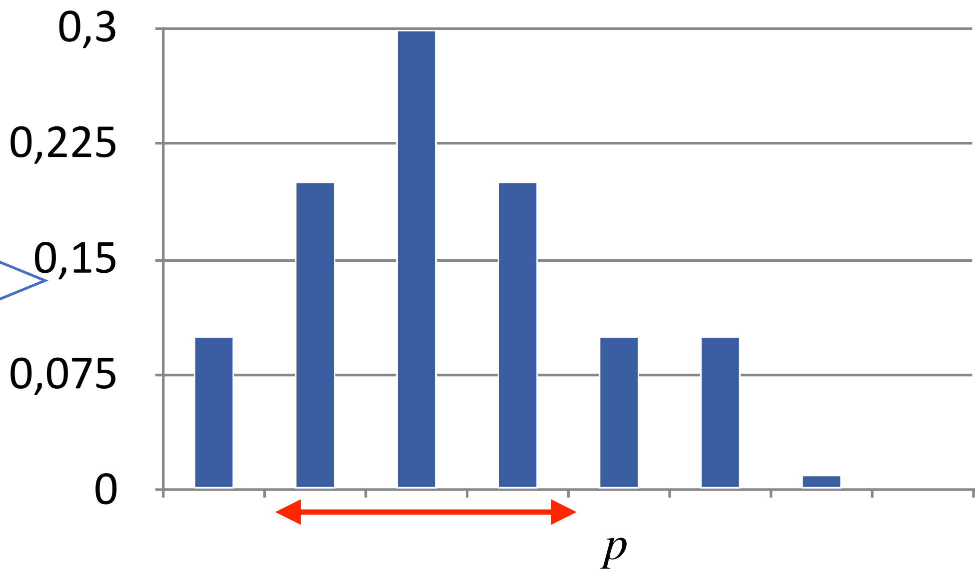




- în dreapta, vârful graficului este atins la $p \approx 0.375$ (maximum likelihood), deci dacă trebuie să dăm o valoare pe asta o returnăm
- dar acum pe dreapta avem o densitate (probability density function)
- deci putem calcul și alte valori: interval de încredere de exemplu

Gândirea probabilistică pentru analiza datelor

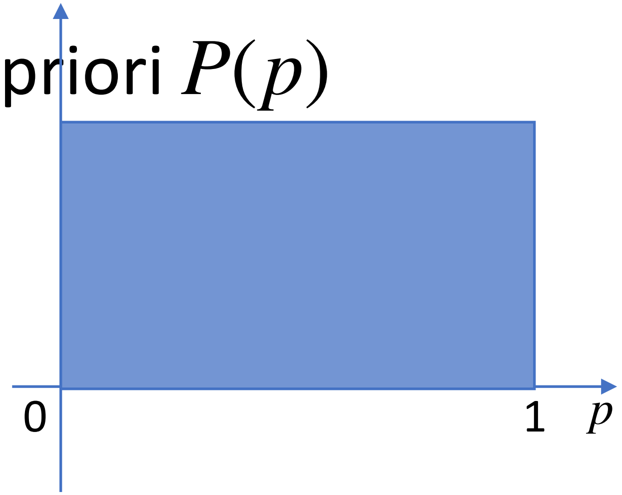




$$P(p = 0.35 \mid 6 \text{ de "da"}) = \frac{P(6 \text{ de "da"} \mid p = 0.35) \times P(p = 0.35)}{\sum_p P(p) \times P(6 \text{ de "da"} \mid p)}$$

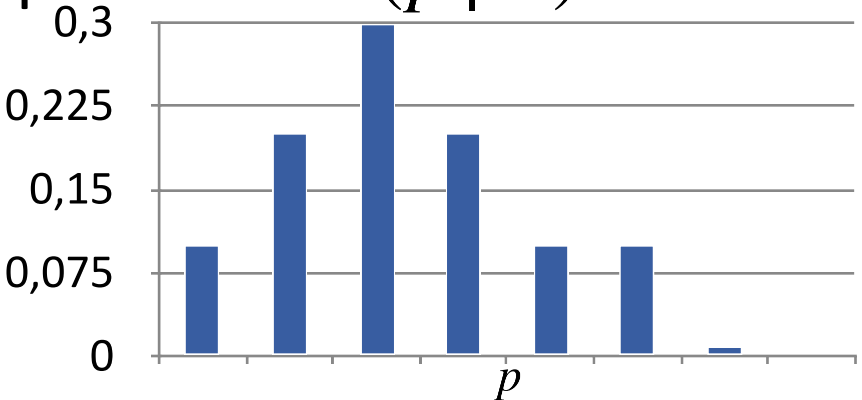
Gândirea probabilistică pentru analiza datelor

- ce am făcut până acum?
  - am pornit de la o densitate de probabilitate apriori $P(p)$



  - am folosit un model generativ $P(D|p)$
  - am ajuns la o densitate de probabilitate aposteriori $P(p|D)$

$$P(p|D) = \frac{P(D|p) \times P(p)}{\sum P(p) \times P(D|p)}$$



Avantaje ale metodei Bayesiene

- putem compara distribuții aposteriorii între ele
    - dacă există două echipe de marketing și a doua echipă reușește să convingă 10 persoane din 35. Este această campanie de marketing mai bună decât prima?

Avantaje ale metodei Bayesiene

- putem adăuga “sfaturile experților” în model
  - cei care lucrează de mult timp la compania de marketing știu că de obicei campaniile lor conving între 10% și 25% dintre persoanele target-ate

Avantaje ale metodei Bayesiene

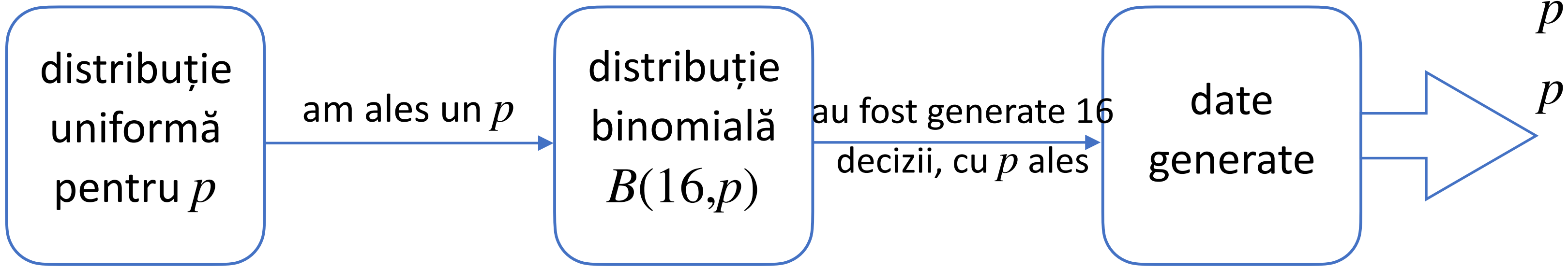
- putem să folosim rezultatul pentru a lua decizii
    - avem două campanii de marketing disponibile (două echipe de marketing) fiecare cu un anumit cost. Care este campania de marketing mai profitabilă?

Avantaje ale metodei Bayesiene

- putem compara distribuții aposteriorii între ele
- putem adăuga “sfaturile experților” în model
- putem să folosim rezultatul pentru a lua decizii
- …

Dezavantaje ale metodei Bayesiene

- este lentă



$p = 0.21$, "da" s-a întâmplat $5/16$

$p = 0.45$, "da" s-a întâmplat $6/16$

$p = 0.21$, "da" s-a întâmplat $6/16$

$p = 0.37$, "da" s-a întâmplat $6/16$

- trebuie generate multe date
- trebuie alese multe valori p

- niște metode moderne care adresează problema asta: MCMC

https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo

Metodei Bayesiene în python

- PyMC3, https://docs.pymc.io/en/v3/
- PyStan, https://pystan.readthedocs.io/en/latest/

- în general, și alte biblioteci generale de machine learning au împlementate și methode Bayesiene
    - scikit-learn, https://scikit-learn.org/stable/
    - TensorFlow, https://www.tensorflow.org/probability/

Rezumat

- prima abordare se numește Approximate Bayesian Computation (ABC)

- metodele moderne (eficiente pentru calcul) se numesc Markov chain Monte Carlo (MCMC) sau Hamiltonian Monte Carlo (HMC)
    - toate aceste metode doresc să aproximeze rezultatul ABC dar cu un timp de calcul mult mai mic decât cel al ABC