Inferență statistică în ML

Cap 1. Introducere. Probabilitate. Probabilități condiționate. Valoare așteptată.

March 3, 2019

Scop

Cursul își propune introducerea noțiunilor fundamentale ce permit abordarea datelor într-un mod științific.

Se vor introduce uneltele necesare pentru analiza datelor, pornind de la datele stocate în baza de date până la realizarea de grafice interactive.

Se prezintă fundamentele inferenței statistice: scopul, modelele și presupunerile inferenței statistice. Ca rezultat, se deprind abilitățile de a realiza raționamente de inferență în diverse situații și pentru ML în special.

Inferența statistică

- **Inferența statistică**: procesul de generare de concluzii despre o populație pornind de la un sample (noisy)
- se generează concluzii noi pornind de la date
- sistem formal de inferență
- exemplu 1: câștigătorul alegerilor următoare, pornind de la sample-ul votanților actuali
  - noisy sample: unii nu vor vota, alții se vor răzgândi, alții vor minți
  - eliminăm incertitudinea
- exemplu 2: probabilitatea ca mâine să plouă este 70% - asociază o probabilitate unui eveniment

Paradigme inferențiale

- există două abordări pentru inferența statistică:
  1. abordarea frecventistă
  2. abordarea bayesiană
- abordarea frecventistă este fundația pentru raționamentele de inferență statistică

- abordarea probabilității din punctul de vedere al analizei numărului de ori al obținerii capului la aruncarea unei monede ideale (Head = cap, Tail = stema)
- experiment repetat, într-un anumit procent de cazuri ceva se întâmplă - parametru al populației
- https://en.wikipedia.org/wiki/Fair_coin

Exemple de probleme de cauzalitate

- cu ajutorul inferenței statistice se pot construi raționamente care să ilustreze cauzalitatea, și nu doar asocierea
  1. provoacă fumatul cancer?
  2. o campanie de reclamă determină creșterea traficului web?
  3. un tratament nou este chiar util pentru vindecarea bolii respective?
- cursul se concentrează pe construirea de modele frecventiste și testarea asumpțiilor
- ne concentrăm mai mult pe concepte
- quiz-uri, teme de casă

Noțiunea de probabilitate

- **probabilitatea** asociază un număr între 0 și 1 unui eveniment, pentru a da un sens noțiunii de "șansă" de producere a acelui eveniment
- probabilitatea modelează ceea ce par a fi fenomene întâmplătoare (aleatoare)
- scopul final este de a folosi un model probabilist, pentru a trage concluzii pornind de la un sample al unei populații
- există un set de reguli pe care noțiunea de probabilitate le îndeplinește

Probabilitate (2)

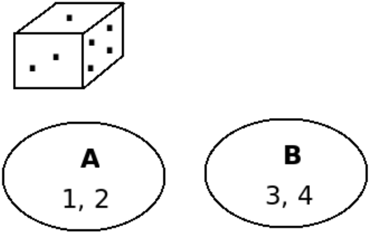
- pentru un experiment aleator, de exemplu aruncarea cu zarul, o măsură a probabilității este o cantitate asociată **populației** care sumarizează caracterul aleator
- probabilitatea este asociată intrinsec populației, este o proprietate a ei, pe care dorim să o estimăm
- probabilitatea ia fiecare eveniment posibil care se produce în urma experimentului și îi asociază o valoare numerică

Aruncarea cu zarul

- aruncarea cu zarul se numește **experiment**; faptul că iese fața 1 este un **eveniment**
- evenimentului "a ieșit fața 1" îi asociem o probabilitate cuprinsă între 0 și 1 (1/6 în cazul unui zar corect)

- exemplu de evenimente: număr par - $\{1, 3, 5\}$ sau număr impar - $\{2, 4, 6\}$
- probabilitatea evenimentului sigur (să iasă un număr la aruncarea cu zarul) este 1
- probabilitatea evenimentului imposibil (acum plouă cu bani) este 0
- evenimentul compus din două evenimente mutual exclusive (fie iese par fie iese impar) este suma probabilităților celor două evenimente

Evenimente mutual exclusive



- evenimentele A și B nu pot apărea simultan
- $P(A \cup B) = P(A) + P(B)$

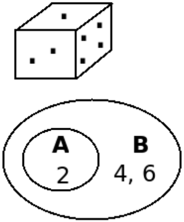
- reguli simple dar suficiente pentru a deduce de aici toate regulile pe care le urmează probabilitățile (Kolmogorov)

Regulile probabilităților

1. probabilitatea evenimentului imposibil este 0
   - ex. la zar, probabilitatea să iasă un număr mai mare ca 6
2. probabilitatea evenimentului sigur este 1
   - ex. probabilitatea să iasă un număr între 1 și 6
3. probabilitatea ca opusul unui eveniment să apară este 1 - (probabilitatea acelui eveniment)
   - ex. probabilitatea să iasă un număr par este 1 - (probabilitatea să iasă un număr impar)
4. probabiliatea ca să apară cel puțin unul din mai multe evenimente mutual exclusive[1] să apară este suma probabilităților lor
5. dacă producerea unui eveniment A implică producerea unui eveniment B, atunci probabilitatea ca A să apară e mai mică decât probabilitatea ca B să apară

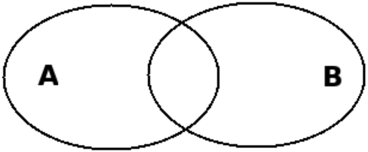
[1] care nu pot să apară simultan

Eveniment dependent



- producerea evenimentului A implică producerea lui B (dacă A s-a produs, atunci sigur și B s-a produs pentru că A este parte din B)
- ex. dacă a apărut evenimentul "fața 2", atunci sigur a apărut și eventimentul "față număr par"
- A este un subset al lui B
- $P(B) > P(A)$ pentru exemplul cu zarul

Evenimente ne-independente



- cele două evenimente nu sunt mutual exclusive
- suma probabilităților lor include intersecția de două ori
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- probabilitățile nu se pot aduna dacă evenimentele asociate lor au o intersecție nevidă

6. pentru orice două evenimente, probabilitatea ca cel puțin unul dintre ele să apară este suma probabilităților lor minus intersecția

Exemplu de evenimente cu intersecție nevidă

- Un studiu relevă faptul că 3% din populația adultă are apnee în somn. Un alt studiu afirmă că 10% din populația adultă manifestă "sindromul piciorului neliniștit" în somn. Putem afirma că 13% din populația adultă are cel puțin una din probleme cu somnul de acest gen?

Exemplu de evenimente cu intersecție nevidă

- Un studiu relevă faptul că 3% din populația adultă are apnee în somn. Un alt studiu afirmă că 10% din populația adultă manifestă "sindromul piciorului neliniștit" în somn. Putem afirma că 13% din populația adultă are cel puțin una din probleme cu somnul de acest gen?
- Nu, pentru că este posibil ca o anumită fracție din populație să aibă ambele tipuri de probleme, și atunci fracția din populație care manifestă cel puțin una din probleme să fie mult mai mică (cel puțin cât?)

Densități de probabilitate

- Axiomele (regulile) probabilităților sunt utile pentru înțelegerea regulilor pe care probabilitățile le urmează
- Totuși avem nevoie de modele cu ajutorul cărora să definim rezultatele numerice ale experimentelor
- ne interesează densities și mass functions pentru variabile aleatoare
  - **probability mass function** se referă la probabilități asociate evenimentelor discrete (aruncarea zarului)
  - **probability density function** se referă la probabilitatea asociată variabilelor continue (evenimentul "cantitatea de apă de ploaie dintr-o zi")
- ce se estimează nu sunt apariții punctuale în date, ci proprietăți ale populației din care observăm un sample (un număr limitat de indivizi)

Variabile aleatoare

- o **variabilă aleatoare** este rezultatul numeric al unui experiment
- ele pot fi **discrete** sau **continue** (exemple: culoarea părului)
- pentru variabile discrete, asignăm probabilități valorilor pe care le pot lua
- pentru variabile continue, asignăm probabilități intervalelor de valori pe care acestea le pot lua
- exmple de variabile discrete: evenimentul fată de monedă, sau faţă de zar, numărul de vizitatori zilnici ai unui site
- exemple de variabile continue: BMI, IQ

PMF

- Probability Mass Function e o funcție, ce dă valoarea probabilității pentru cazul în care variabila aleatoare ia o anumită valoare
- această funcție $P(\cdot)$ trebuie să satisfacă:
    - trebuie să fie mai mare sau egală cu 0
    - suma valorilor probabilităților pe care variabila aleatoare poate să le ia trebuie să aibă suma 1

    - problema zarului:
      $P(1) = 1/6 = P(X = 1)$
      $P(2) = 1/6 = P(X = 2)$
      $\ldots$
      $P(6) = 1/6 = P(X = 6)$

      $$P(X = 1) + P(X = 2) + \cdots + P(X = 6) = 1$$

PMF pentru coin flip - distribuția Bernoulli

X = 0 reprezintă tails iar X = 1 reprezintă heads[2]
$p(x) = (1/2)^x (1/2)^{1-x}$, unde x este o variabilă[3], x = 0, 1

$p(0) = (1/2)^0 (1/2)^{1-0} = 1/2$
$p(1) = (1/2)^1 (1/2)^{1-1} = 1/2$

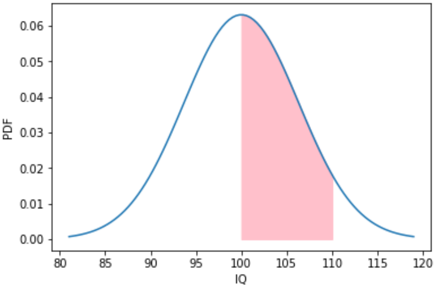
- pentru o monedă trucată, avem:
  $p(x) = \theta^x (1 - \theta)^{1-x}$, for x = 0, 1

  $p(0) = \theta^0 (1 - \theta)^{1-0} = 1 - \theta$
  $p(1) = \theta^1 (1 - \theta)^{1-1} = \theta$
- exemplu, acest $p(\cdot)$ poate descrie prevalența hipertensiunii într-o populație; dar nu cunoaștem θ, putem folosi datele pentru a-l estima[4]

[2] litera mare X reprezintă o valoare potențială, nerealizată, a variabilei aleatoare
[3] placeholder folosit pentru a plasa valoarea asociată evenimentului realizat
[4] tehnica se numește Maximum Likelihood Estimation, alegem θ astfel ca datele să se potrivească cel mai bine

PDF

- o funcție a densității de probabilitate este o funcție asociată cu o variabilă aleatoare continuă
- ca și PMF, va satisface anumite proprietăți:
  1. mai mare sau egală cu zero peste tot (în domeniul de definiție)
  2. aria totală de sub funcție trebuie să fie 1 (Area Under Curve, AuC)
- ariile de sub PDFs corespund cu probabilitățile pentru acea variabilă aleatoare



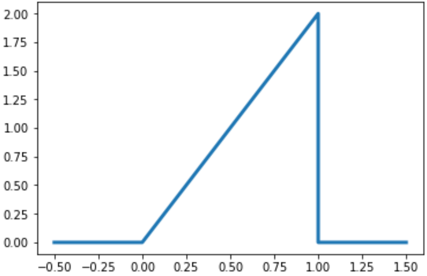
- probabilitatea ca o persoană să aibă IQ între 100 și 110
- PDF e un atribut al populației coeficientului de inteligență, nu o afirmație despre date
- vom folosi datele să deducem însușiri ale populației

PDF (2)

- pentru PDF, probabilitatea ca variabila aleatoae să ia EXACT o anumită valoare (ex. 100, în exemplul de mai sus), e ZERO
- a nu se confunda cu PMF !
- densitatea de probabilitate normală este denumită și densitate de probabilitate gaussiană, sau “bell-shaped curve”, sau “clopot”, sau “pălărie”
- denumită și distribuție normală, ea este caracterizată de două valori: media μ și dispersia (sau deviația standard) σ
- în exemplul anterior, $\mu = 100$ iar $\sigma = 6$ (aproximativ)

Exemplu de PDF

$$f(x) = \begin{cases} 2x & daca\, 0 < x < 1 \\ 0 & altfel \end{cases}$$



- $f(x)$ exprimă proporția de apeluri care este satisfăcută de linia call center într-o zi oarecare
- este o densitate de probabilitate validă? (cele 2 proprietăți)
- care este probabilitatea ca mai puțin din 75% din apeluri să fie satisfăcute? aria:
  $1.5 * 0.75/2 = 0.5625$
  $>>$ stats.beta.cdf(0.75, 2, 1)
  0.5625

CDF și Survival Function

- Cummulative Distribution Function (CDF) pentru o variabilă aleatoare X, este probabilitatea ca variabila aleatoare să fie mai mică sau egală cu valoarea x (mic):

  $F(x) = P(X \leq x)$
- definiția se aplică atât pentru variabile discrete cât și pentru cele continue:

```
>> stats.beta.cdf(0.75, 2, 1)
0.5625
```

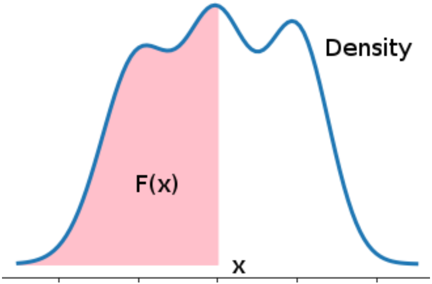
- funcția SciPy beta.cdf() va da probabilitatea ca variablia să fie mai mică decât valoarea dată
- Survival Function: $S(x) = P(X > x) = 1 - F(x)$

```
>> stats.beta.cdf([0.4, 0.5, 0.6], 2, 1)
[0.16 0.25 0.36]
```

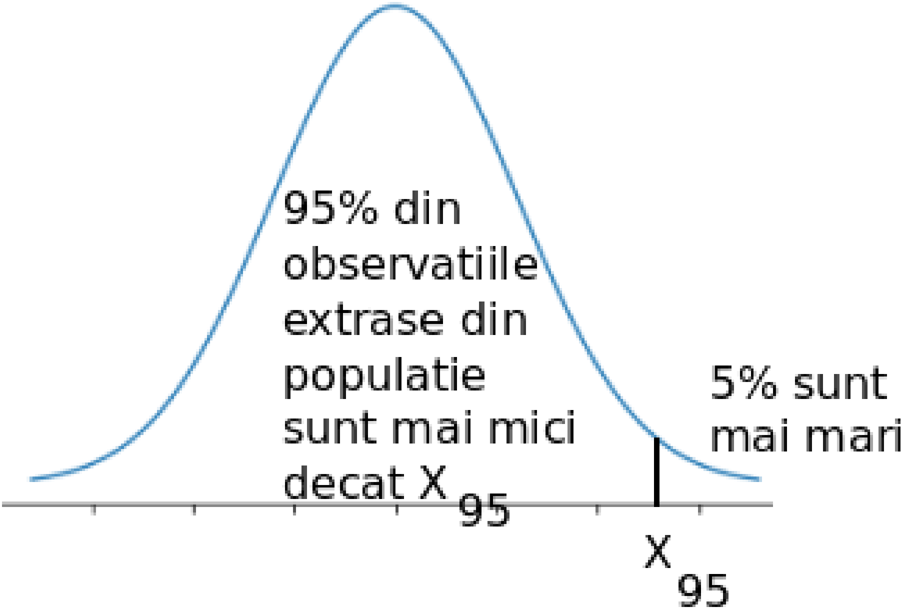
Quantiles

- nota la un examen este quantila 95%, înseamnă că 95% din candidați au avut note mai mici respectiv 5% note mai mari - acestea sunt "sample quantiles"
- quantila α a unei funcții de distribuție F este punctul x_α astfel încât:

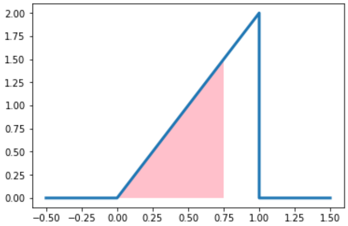
$$F(x_\alpha) = \alpha$$



- se deplasează x până când aria $F(x) = \alpha$
- aria e de fapt CDF (aria totală $= 1$)
- **percentila** e de fapt o quantilă exprimată ca procent
- mediana este percentila 50


95% din
observatiile
extrase din
populatie
sunt mai mici
decat X 95
5% sunt
mai mari
X 95

$$F(x) = P(X \leq x) = \frac{1}{2} Base * Height = \frac{1}{2}(x) * (2x) = x^2$$



- în 50% din zile, 70% din apeluri sau mai puține sunt satisfăcute

- aria hașurată exprimă probabilitatea ca mai puțin de 75% din apeluri să fie satisfăcute
- dorim să aflăm care este proporția de apeluri ce corespunde unei probabilități de 50%

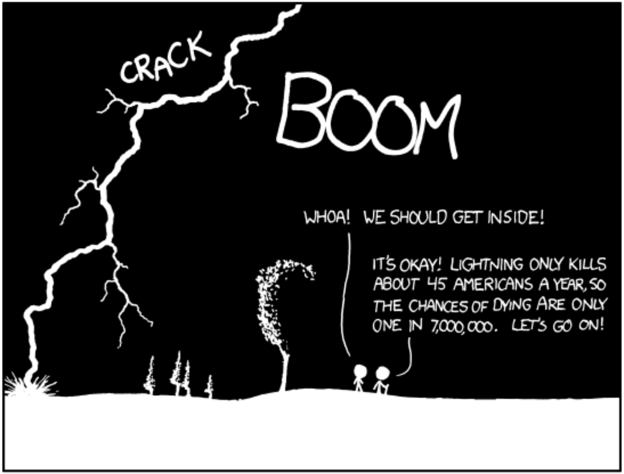
$$0.5 = F(x) = x^2$$

```
>> math.sqrt(0.5)
0.7071067
>> stats.beta.ppf(0.5, 2, 1)
# ppf - percent point function
0.7071067
```

Estimatori

- de obicei, noțiunea de mediană se realiza folosind ordonarea crescătoare a datelor și folosirea elementului de la mijloc
- aici, mediana sample-ului va estima mediana populației
- de exemplu, putem lua mai multe zile, și aflăm proporția de apeluri satisfăcute
- apoi facem mediana acestor proporții, și vom aproxima astfel mediana populației
- mediana populației se numește estimand, iar mediana sample-ului folosit de noi se numește estimator

Conditional risk



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Probabilități condiționate

- la zar, probabilitatea să iasă 1 este $1/6$
- presupunem că știm că fața este număr impar
- folosind această nouă informație, probabilitatea să iasă 1 este .. ?

Probabilități condiționate

- la zar, probabilitatea să iasă 1 este $1/6$
- presupunem că știm că fața este număr impar
- folosind această nouă informație, probabilitatea să iasă 1 este .. ?
- $1/3$

Probabilități condiționate (2)

- fie B un eveniment astfel ca $P(B) > 0$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- dacă A și B sunt independente statistic, atunci:

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

- în acest ultim caz, faptul că B a apărut nu dă informații suplimentare despre A

Probabilități condiționate (3)

- $A = \{1\}$, $B = \{1, 3, 5\}$
- $P(1 \text{ dacă a ieșit un număr impar}) = P(A|B)$
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}$

Regula lui Bayes

- Thomas Bayes, matematician și teolog, Anglia sec. XVIII
- ne permite inversarea rolurilor evenimentelor pentru care vrem să aflăm probabilitățile
- vrem să calculăm $P(B|A)$ dacă avem $P(A|B)$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \qquad \text{respectiv:}$$

$$= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

- se poate privi $P(A)$ ca fiind alcătuită din două evenimente mutual exclusive, $P(A|B)$ respectiv $P(A|B^c)$, când se întâmplă și nu se întâmplă B

Teste și diagnostice

- fie $+$ și - evenimentele ca un rezultat al unui test pentru diagnostic să fie pozitiv sau negativ
- fie D și D^c evenimentul ca subiectul testului să aibă sau nu boala pentru care se testează

$Sensitivity = P(+|D)$, vrem să fie ridicată
$Specificity = P(-|D^C)$, la fel, ridicată pentru ca testul să fie considerat bun

Teste și diagnostice (2)

- suntem interesați, dacă are boala, ca testul să fie pozitiv:

  Positive predictive value $= P(D|+)$

  Negative predictive value $= P(D^C|-)$

- în absența testului:

  Prevalence of disease $= P(D)$

Teste și diagnostice (3)

- considerăm că avem un test HIV cu următoarele valori ipotetice:
  - sensitivity de 99.7%
  - specificity de 98.5%
- în populație, prevalența bolii este de 0.1%
- dacă testul iese pozitiv, care este probabilitatea ca respectivul chiar să aibă boala?

  $P(D|+) = ?$

Teste și diagnostice - Bayes

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^C)P(D^C)}$$

- putem exprima totul în funcție de sensitivity $P(+|D)$ și specificity $P(-|D^C)$:

$$= \frac{P(+|D)P(D)}{P(+|D)P(D) + \{1 - P(-|D^C)\}\{1 - P(D))\}}$$

$$= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} = .062$$

Teste și diagnostice - Bayes (2)

- valoarea scăzută a positive predictive value $P(D|+)$ este cauzată de prevalența scăzută a bolii în populație
- pe de altă parte, dacă știm că respectivul este un consumator de droguri, această probabilitate crește (prevalența HIV în rândul consumatorilor de droguri)

$$= \frac{P(+|D)P(D)}{P(+|D)P(D) + \{1 - P(-|D^C)\}\{1 - P(D))\}}$$

$$= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} = .062$$

Likelihood ratios

- vrem să distingem între componenta dependentă de această prevalență și componenta descrisă ca rezultat obiectiv al testului - likelihood ratio
- positive predictive value:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^C)P(D^C)}$$

- formula depinde de specificity, 1 - sensitivity și prevalența bolii (P(D))
- putem exprima cu Teorema Bayes $1 - P(D|C) = P(D^C|+)$, probabilitatea de a nu avea boala dacă testul iese pozitiv:

$$P(D^C|+) = \frac{P(+|D^C)P(D^C)}{P(+|D)P(D) + P(+|D^C)P(D^C)}$$

Likelihood ratios (2)

- prin împărțire directă, obținem:

$$\frac{P(D|+)}{P(D^C|+)} = \frac{P(+|D)P(D)}{P(+|D^C)P(D^C)} = \frac{P(+|D)}{P(+|D^C)} \times \frac{P(D)}{P(D^C)}$$

- prima parte se numește șansa (proporția) bolii dat fiind rezultatul
- ultima parte este șansa (proporția) bolii în lipsa testului
- termenul din mijloc este diagnostic likelihood ratio a rezultatului unui test pozitiv
- post-test likelihood = diagnostic likelihood ratio[5] $\times$ pre-test likelihood

[5] DLR_+

Likelihood ratios (3)

- pentru exemplul inițial:

$$DLR_+ = .997/(1 - .985) = 66$$

- indiferent de proporția pre-test, se înmulțește cu 66 pentru obținerea proporției post-test
- ipoteza bolii este suportată de date de 66 de ori mai puternic decât ipoteza lipsei bolii

Evenimente independente

- evenimentele A și B sunt zise independente dacă: $P(A|B) = P(A)$, unde $P(B) > 0$

  sau

  $P(A \cap B) = P(A)P(B)$

- probabilitățile nu se pot multiplica oricând, ci se multiplică doar probabilitățile evenimentelor independente

Evenimente independente (2)

- exemplu: probabilitatea de a obține de două ori Head?

  A = {Head la prima aruncare}, P(A) = .5

  B = {Head la a doua aruncare}, P(B) = .5

  $A \cap B = \{\text{Heads la ambele aruncări}\}$

  $P(A \cap B) = P(A)P(B) = .5 \times .5 = .25$

Evenimente ne-independente (3)

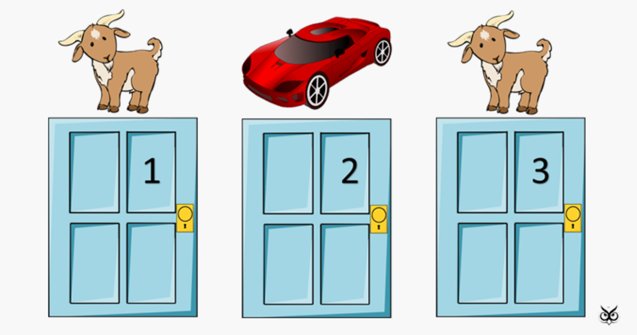
- în 1999, Sally Clark, a fost condamnată pentru crimă împotriva propriilor copii (doi), decedați la vâsta de 2 și 3 luni
- apărarea a pledat pentru "Sudden Infant Death Syndrome"
- `https://understandinguncertainty.org/node/545`

- prevalența SIDS este de 1 din 8543
- expertul a considerat cele două morți evenimente independente
- șansa ca două să se producă: $(1/8543)^2$ aprox. 1 la 73 milioane
- pe baza acestei argumentații a fost condamnată - eroare judiciară
- evenimentele de acest gen care au un mediu sau componentă genetică comună, nu sunt independente

Variabile aleatoare IID

- variabilele aleatoare sunt zise IID dacă sunt independente și identically distributed
- indepentente: statistic nelegate între ele
- distribuite identic: toate sunt extrase din aceeași distribuție (populație)
- mai multe aruncări de monede sunt IID
- modelul IID este modelul implicit pentru sample-uri extrase aleator

Monty Hall game show



- fiecare ușă ascunde fie mașina, fie o capră
- inițial alegem o ușă
- prezentatorul deschide o ușă cu o capră (întotdeauna)
- prezentatorul ne oferă să schimbăm ușa; o facem?
  `https://www.youtube.com/watch?v=8DMnAAvakh0` The '21' movie fragment

Monty Hall și probabilitățile condiționate

- notăm cu A evenimentul ca mașina să se afle în spatele ușii 1
- notăm cu B evenimentul ca prezentatorul să deschidă ușa 2 (el întotdeauna deschide o ușă în spatele căreia se află o capră)
- probabilitatea ca în spatele ușii 1 să se afle mașina este $P(A) = 1/3$
- pe noi ne interesează ce s-a schimbat cu A în cazul în care se întâmplă B:

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{1}$$

Monty Hall și probabilitățile condiționate (2)

- probabilitatea ca să se întâmple simultan A și B, adică mașina să fie în spatele ușii 1 și el să deschidă ușa 2 (atenție, nu sunt indepentente!), se calculează folosind teorema Bayes:

$$P(A,B) = P(B|A) * P(A) = \frac{1}{2} * \frac{1}{3} = \frac{1}{6} \quad (2)$$

- $P(B|A)$ este probabilitatea ca el să deschidă ușa 2 dacă mașina e în spatele ușii 1 (ceea ce se întâmplă în jumătate din cazuri)

Monty Hall și probabilitățile condiționate (3)

- $P(B)$ se evaluează folosind probabilitățile marginale (condiționate), adică ținem seama de apariția (sau nu) a evenimentului A:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) \tag{3}$$

- prima parte $P(B|A)P(A) = P(A,B) = \frac{1}{6}$ am calculat-o mai sus
- a doua parte este $P(B|A^c)P(A^c) = P(B, A^c)$ și exprimă probabilitatea ca să deschidă ușa 2 și mașina să nu se află în spatele ușii 1

Monty Hall și probabilitățile condiționate (4)

- probabilitatea ca să deschidă ușa 2 și mașina să se afle în spatele ușii 1, $P(B, A^c)$, este dată de două situații:
  - mașina să se afle în spatele ușii 2 și atunci nu deschide ușa;
  - mașina să se afle în spatele ușii 3 și atunci deschide ușa:

$$P(B, A^c) = \frac{1}{3} * 0 + \frac{1}{3} * 1 = \frac{1}{3} \tag{4}$$

- reluînd, avem:

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{2}{6}} = \frac{1}{1+2} = \frac{1}{3} \tag{5}$$

- deci șansele noastre nu s-au îmbunătățit dacă nu schimbăm ușa
- din contra, probabilitatea ca mașina să nu se afle în spatele ușii 1 dacă el deschide ușa 2 este:

$$P(A^c|B) = 1 - P(A|B) = \frac{2}{3} \tag{6}$$

Bibliografie

- https://www.norwegiancreations.com/2018/10/bayes-rule-and-the-monty-hall-problem/
-