Curs 15

Cristian Niculescu

1 Intervale de probabilitate

1.1 Scopurile învăţării

1. Să poată afla intervale de probabilitate fiind date o pmf sau pdf.
2. Să înţeleagă cum intervalele de probabilitate rezumă convingerea în actualizarea Bayesiană.
3. Să poată folosi intervale de probabilitate subiective pentru a construi a priori rezonabile.
4. Să poată construi intervale de probabilitate estimând sistematic cuantilele.

1.2 Intervale de probabilitate

Presupunem că avem o pmf $p(\theta)$ sau pdf $f(\theta)$ descriind convingerea noastră despre valoarea parametrului necunoscut de interes θ.
Definiţie. Un interval de p-probabilitate pentru θ este un interval $[a, b]$ cu $P(a \leq \theta \leq b) = p$.
Observaţii.
1. În cazul discret cu pmf $p(\theta)$, aceasta înseamnă $\sum_{a \leq \theta_i \leq b} p(\theta_i) = p$.
2. În cazul continuu cu pdf $f(\theta)$, aceasta înseamnă $\int_a^b f(\theta)d\theta = p$.
3. Putem spune interval de 90%-probabilitate pentru interval de 0.9-probabilitate.

Intervalele de probabilitate sunt de asemenea numite intervale credibile spre a le deosebi de intervalele de încredere.
Exemplul 1. Între 0.05 şi 0.55 cuantilele este un interval de 0.5-probabilitate. Sunt multe intervale de 50% probabilitate, de exemplu intervalul dintre 0.25 şi 0.75 cuantilele.
În particular, observăm că intervalul de p-probabilitate nu este unic.
Q-notaţie. Putem formula intervalele de probabilitate în termeni de **cuantile**. Reamintim că s-cuantila pentru θ este valoarea q_s cu $P(\theta \leq q_s) = s$. Deci, pentru $s \leq t$, cantitatea de probabilitate dintre s-cuantila şi t-cuantila este chiar $t - s$. În aceşti termeni, un interval de p-probabilitate este orice

interval $[q_s, q_t]$ cu $t - s = p$.

Exemplul 2. Avem intervalele de 0.5 probabilitate $[q_{0.25}, q_{0.75}]$ şi $[q_{0.05}, q_{0.55}]$.

Intervale de probabilitate simetrice.

Intervalul $[q_{0.25}, q_{0.75}]$ este simetric deoarece cantitatea de probabilitate rămasă în afara lui, în oricare din cele 2 părţi, este aceeaşi, şi anume 0.25. Dacă pdf nu este prea înclinată, intervalul simetric este de obicei o bună alegere implicită.

Mai multe observaţii.

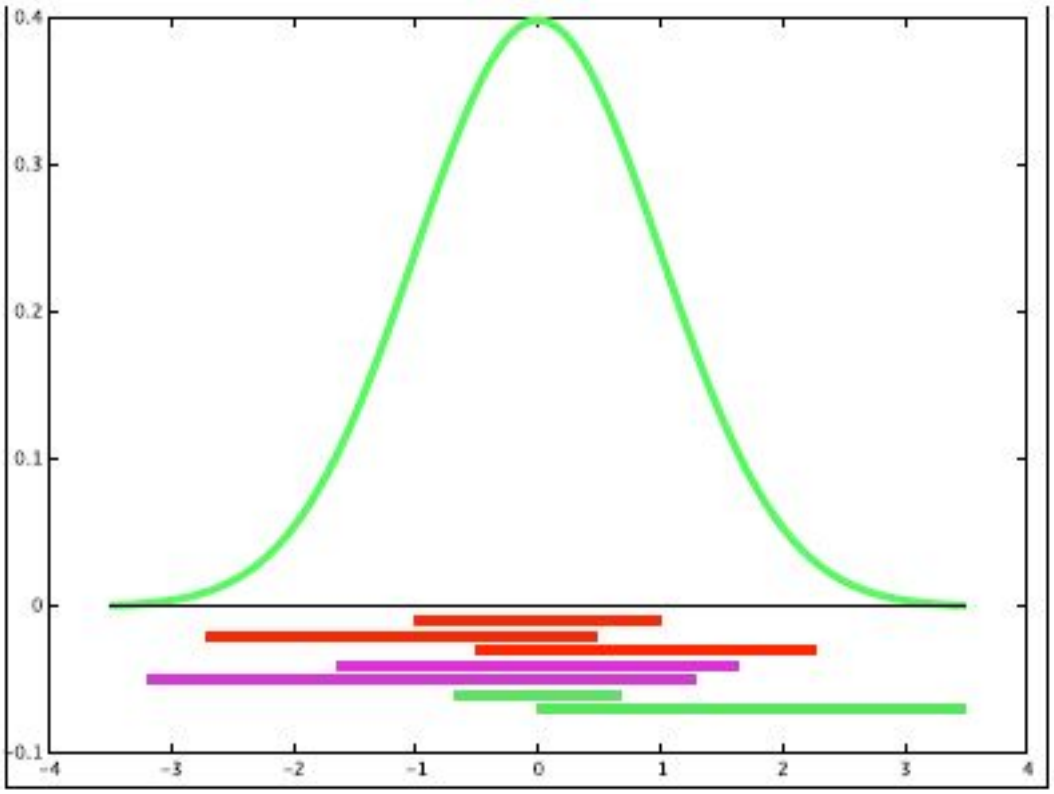
1. Diferite intervale de p-probabilitate pentru θ pot avea lungimi diferite. Putem face lungimea mai mică centrând intervalul sub cea mai înaltă parte a pdf. Un astfel de interval este de obicei o bună alegere deoarece conţine cele mai probabile valori.

2. Deoarece lungimea poate varia pentru p fixat, un p mai mare nu înseamnă totdeauna o lungime mai mare. Iată ce este adevărat: dacă un interval de p_1-probabilitate este inclus într-un interval de p_2-probabilitate, atunci $p_1 \leq p_2$.

Intervale de probabilitate pentru o repartiţie normală. Figura arată un număr de intervale de probabilitate pentru normala standard.

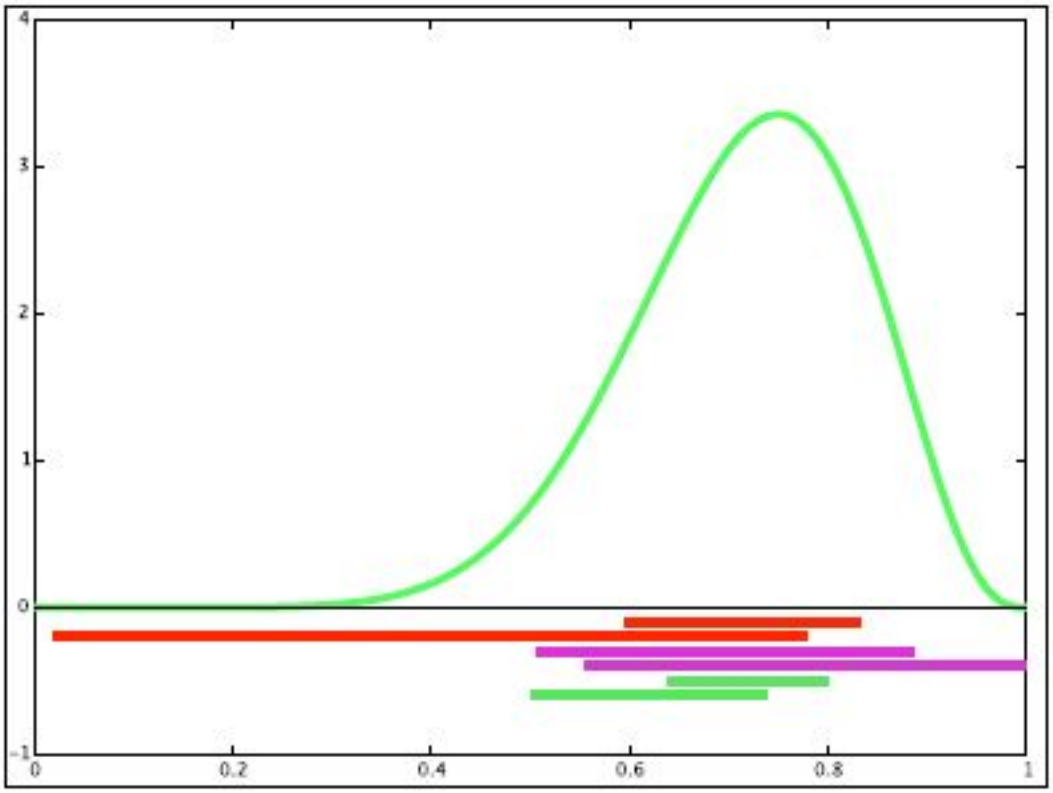
1. Toate barele roşii cuprind un interval de 0.68-probabilitate. Observaţi că cea mai mică bară roşie merge de la -1 la 1. Acest interval este de la a 16-a percentilă la a 84-a percentilă, deci este simetric.

2. Toate barele mov cuprind un interval de 0.9-probabilitate. Ele sunt mai lungi decât barele roşii deoarece cuprind mai multă probabilitate. Observaţi din nou că cea mai scurtă bară mov corespunde unui interval simetric.



roşu = 0.68, mov = 0.9, verde = 0.5

Intervale de probabilitate pentru o repartiţie beta. Următoarea figură arată intervale de probabilitate pentru o repartiţie beta. Observaţi că cele 2 bare roşii au lungimi foarte diferite, totuşi cuprind aceeaşi probabilitate $p = 0.68$.



1.3 Utilizări ale intervalelor de probabilitate

1.3.1 Rezumarea şi comunicarea convingerilor noastre

Intervalele de probabilitate sunt un mod intuitiv şi eficace de a rezuma şi comunica convingerile noastre. Este greu de descris o întreagă funcţie $f(\theta)$ în cuvinte. Dacă funcţia nu este dintr-o familie parametrizată, atunci este şi mai greu. Chiar şi cu o repartiţie beta este mai uşor de interpretat "Cred că θ este între 0.45 şi 0.65 cu 50% probabilitate" decât "Cred că θ are o repartiţie beta$(8, 6)$". O excepţie de la această regulă de comunicare poate fi repartiţia normală, dar numai dacă interlocutorul este familiarizat cu deviaţia standard. Desigur, ce câştigăm în claritate pierdem în precizie, deoarece funcţia conţine mai multă informaţie decât intervalul de probabilitate.
Intervalele de probabilitate se comportă bine în actualizarea Bayesiană. Dacă actualizăm de la a priori $f(\theta)$ la a posteriori $f(\theta|x)$, atunci intervalul de p-probabilitate pentru a posteriori va tinde să fie mai scurt decât intervalul de p-probabilitate pentru a priori. În acest sens, datele ne fac mai siguri.

1.4 Construirea unei a priori folosind intervale de probabilitate subiective

Intervalele de probabilitate sunt de asemenea utile când nu avem o pmf sau pdf la îndemână. În acest caz, intervalele de probabilitate subiective ne dau o metodă de a construi o a priori rezonabilă pentru θ "de la 0". Procesul de gândire este să ne punem o serie de întrebări, de exemplu: "care este media lui θ?"; "intervalul de 0.5-probabilitate?"; "intervalul de 0.9-probabilitate?". Apoi construim o a priori care este potrivită cu aceste intervale.

1.4.1 Estimarea directă a intervalelor

Exemplul 3. Construirea a priori
În 2013 au fost alegeri speciale pentru un loc în congres într-un district din Carolina de Sud. Alegerile au adus în arenă pe republicanul Mark Sanford contra democratei Elizabeth Colbert Busch. Fie θ fracţia din populaţie care l-au favorizat pe Sanford. Scopul nostru în acest exemplu este să construim o a priori subiectivă pentru θ. Vom folosi următoarele dovezi a priori:
Sanford este un fost parlamentar şi guvernator de Carolina de Sud.
El a demisionat cu tam-tam după ce a avut o aventură în Argentina în timp ce pretindea că face o drumeţie pe un traseu din Munţii Apalaşi.
În 2013 Sanford a câştigat alegerile primare republicane în faţa a 15 oponenţi.
În district, în alegerile prezidenţiale, republicanul Romney a învins pe democratul Obama cu 58% la 40%.
Avantajul lui Colbert: Elizabeth Colbert Busch este sora cunoscutului comic Stephen Colbert.
Strategia noastră va fi să ne folosim intuiţia pentru a construi unele intervale de probabilitate şi apoi să găsim o repartiţie beta care se potriveşte aproximativ cu aceste intervale. Acestea sunt subiective, deci altcineva poate da un răspuns diferit.
Pasul 1. Folosim dovezile a priori pentru a construi intervale de 0.5 şi 0.9 probabilitate pentru θ.
Vom începe gândindu-ne la intervalul de 90%. Singura dovadă a priori cea mai puternică este 58% la 40% la Romney contra Obama. Dată fiind boacăna lui Sanford nu ne aşteptăm să câştige mai mult de 58% din voturi. Deci vom pune marginea superioară a intervalului de 0.9 la 0.65. Din cauza boacănei, Sanford ar putea să piardă mult. Deci vom pune marginea inferioară la 0.3.

$$\text{intervalul de } 0.9 : [0.3, 0.65]$$

Pentru intervalul de 0.5 vom muta aceste margini înăuntru. Pare improbabil ca Sanford să obţină mai multe voturi ca Romney, deci putem lăsa 0.25 din

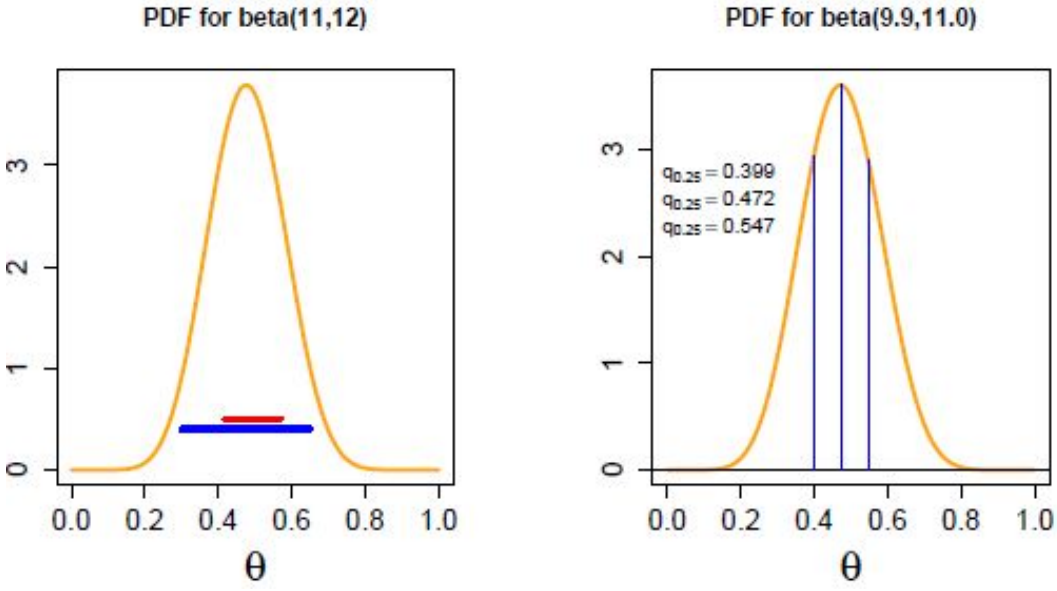
probabilitate ca el să ia peste 57%. Marginea inferioară pare mai greu de prezis. Vom lăsa 0.25 din probabilitate ca el să ia sub 42%.

$$\text{intervalul de } 0.5 : [0.42, 0.57]$$

Pasul 2. Folosim intervalele noastre de 0.5 şi 0.9 probabilitate pentru a alege o repartiţie beta care aproximează aceste intervale. Se foloseşte funcţia din R `pbeta` şi câteva încercări pentru a alege beta(11,12). Iată codul din R:

```
a = 11
b = 12
pbeta(0.65, a, b) - pbeta(0.3, a, b)
pbeta(0.57, a, b) - pbeta(0.42, a, b)
```

Obţinem $P([0.3, 0.65]) = 0.91$ şi $P([0.42, 0.57]) = 0.52$. Deci intervalele noastre sunt de fapt intervale de 0.91 şi 0.52-probabilitate. Aceasta este destul de aproape de ce am vrut.
În stânga este graficul densităţii lui beta(11,12). Linia roşie arată intervalul nostru [0.42, 0.57] şi linia albastră arată intervalul nostru [0.3, 0.65].



beta$(11, 12)$ aflată folosind intervale de probabilitate şi beta $(9.9, 11)$ aflată folosind cuantilele

1.4.2 Construcţia unei a priori prin estimarea cuantilelor

Pentru a construi o a priori estimând quantilele, strategia de bază este a estima întâi mediana, apoi prima şi a 4-a cuartilă. Apoi alegem o repartiţie a priori care se potriveşte cu aceste estimări.

Exemplul 4. Refaceţi exemplul de alegeri Sanford contra Colbert-Busch

folosind cuantilele.
Răspuns. Începem estimând mediana. Ca şi mai devreme, singura dovadă a priori cea mai puternică este victoria cu 58% la 40% a lui Romney contra lui Obama. Totuşi, dată fiind boacăna lui Sanford şi avantajul lui Colbert, vom estima mediana la 0.47. Într-un district care a dat 58 la 40 pentru republicanul Romney este greu de imaginat că votul pentru Sanford va scădea sub 40%. Deci vom estima a 25-a percentilă pentru Sanford la 0.4. Analog, dată fiind boacăna lui, este greu de imaginat că va urca peste 58%, deci vom estima a 75-a percentilă a lui la 0.55.
Se foloseşte R pentru a căuta printre valorile lui a şi b cu o zecimală cea mai bună potrivire. Se găseşte beta(9.9, 11). Deasupra este o reprezentare a lui beta(9.9, 11) cu quartilele ei reale. În loc de "$q_{0.25} = 0.472$" se va citi "$q_{0.5} = 0.472$". În loc de "$q_{0.25} = 0.547$" se va citi "$q_{0.75} = 0.547$". Acestea se potrivesc cu cuartilele dorite destul de bine.
Notă istorică. În alegeri Sanford a câştigat 54% din voturi şi Busch a câştigat 45.2%. (Sursa: `http://elections.huffingtonpost.com/2013/mark-sanford-vs-elizabeth-colbert-busch-sc1`.)

2 Şcoala frecvenţionistă de statistică

2.1 Scopurile învăţării

1. Să poată să explice diferenţa dintre abordările frecvenţionistă şi Bayesiană ale statisticii.
2. Să ştie definiţia de lucru a statisticii şi să poată distinge o statistică de o nestatistică.

2.2 Introducere

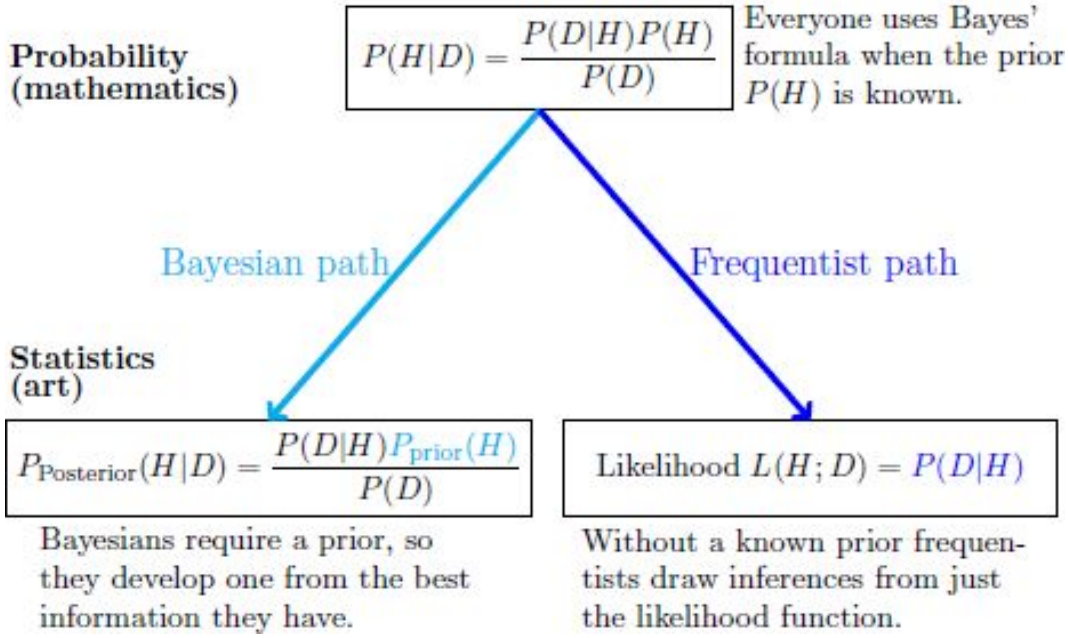
Pentru mare parte din secolul XX, statistica frecvenţioniostă a fost şcoala dominantă. Dacă aţi întâlnit vreodată intervale de încredere, p-valori, t-teste sau χ^2-teste, aţi văzut statistică frecvenţionistă. Odată cu dezvoltarea calculatoarelor de mare viteză şi a datelor mari, metodele Bayesiene devin mai frecvente.

2.2.1 Răspântia

Ambele şcoli de statistică încep cu probabilitatea. În particular ambele ştiu şi apreciază teorema lui Bayes:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

Când a priori este cunoscută exact toţi statisticienii vor folosi această formulă. Pentru deducţia Bayesiană luăm H o ipoteză şi D nişte date. Date fiind o a priori şi un model de verosimilitate, teorema lui Bayes este o reţetă completă pentru actualizarea convingerilor noastre în faţa noilor date. Aceasta funcţionează perfect când a priori a fost cunoscută perfect.
În practică de obicei nu există o a priori universal acceptată - persoane diferite vor avea convingeri a priori diferite - dar tot ne-ar plăcea să facem deducţii utile din date. Bayesienii şi frecvenţioniştii au abordări fundamental diferite la această provocare, după cum este rezumat în figura următoare.



Motivele pentru această împărţire sunt atât practice (uşurinţa implementării şi calculului) cât şi filozofice (subiectivitate versus obiectivitate şi natura probabilităţii).

2.2.2 Ce este probabilitatea?

Principala diferenţă filozofică priveşte înţelesul probabilităţii. Termenul frecvenţionist se referă la ideea că probabilităţile reprezintă frecvenţe pe termen lung ale experimentelor aleatoare repetabile. De exemplu, "o monedă are probabilitatea 1/2 a aversului" înseamnă că frecvenţa relativă a aversurilor (numărul de aversuri supra numărul de aruncări) tinde la 1/2 când numărul de aruncări tinde la ∞. Aceasta înseamnă că frecvenţioniştii găsesc fără sens specificarea unei repartiţii de probabilitate pentru un parametru cu o valoare fixată. În timp ce Bayesienii folosesc probabilitatea pentru a descrie cunoaşterea lor incompletă a unui parametru fixat, frecvenţioniştii resping folosirea proba-

bilităţii pentru a cuantifica gradul de convingere în ipoteză.
Exemplul 1. Presupunem că avem o monedă cu probabilitate necunoscută θ a aversului. Valoarea lui θ poate fi necunoscută, dar este o valoare fixată. Astfel, pentru frecvenţionist nu poate exista o pdf a priori $f(\theta)$. Prin comparaţie, Bayesianul poate fi de acord că θ are o valoare fixată, dar interpretează $f(\theta)$ ca reprezentând incertitudinea despre acea valoare. Atât Bayesianul cât şi frecvenţionistul sunt de acord cu $p(\text{avers}|\theta) = \theta$, deoarece frecvenţa pe termen lung a aversurilor dat fiind θ este θ.
Pe scurt, Bayesienii pun repartiţii de probabilitate pe orice (ipoteze şi date), în timp ce frecvenţioniştii pun repartiţii de probabilitate pe date (aleatoare, repetabile, experimentale) cunoscând o ipoteză. Pentru frecvenţionist, când are de-a face cu date dintr-o repartiţie necunoscută doar verosimilitatea are sens. A priori şi a posteriori n-au.

2.3 Definiţia de lucru a statisticii

Statistica. O statistică este orice poate fi calculat din date. Uneori, pentru a fi mai precişi, vom spune că o statistică este o regulă pentru a calcula ceva din date şi valoarea statisticii este ce este calculat. Aceasta poate include calculul verosimilităţilor unde facem ipoteze asupra valorilor parametrului modelului. Dar nu include ceva care cere să ştim adevărata valoare a parametrului cu valoare necunoscută a modelului.
Exemple. 1. Media datelor este o statistică. Este o regulă care spune că ştiind datele $x_1, ..., x_n$ calculăm $\frac{x_1+...+x_n}{n}$.
2. Maximul datelor este o statistică. Este o regulă care spune să alegem valoarea maximă a datelor $x_1, ..., x_n$.
3. Presupunem $x \sim N(\mu, 9)$, unde μ este necunoscută. Atunci verosimilitatea

$$p(x|\mu = 7) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-7)^2}{18}}$$

este o statistică. Totuşi, distanţa de la x la adevărata medie μ nu este o statistică deoarece nu putem s-o calculăm fără a şti pe μ.
Statistică punctuală. O statistică punctuală este o singură valoare calculată din date. De exemplu, media şi maximul sunt ambele statistici punctuale. Estimarea de verosimilitate maximă este de asemenea o statistică punctuală deoarece este calculată direct din date pe baza unui model de verosimilitate.
Statistică interval. O statistică interval este un interval calculat din date. De exemplu domeniul de la minimul lui $x_1, ..., x_n$ la maximul lui $x_1, ..., x_n$ este o statistică interval, de exemplu datele 0.5, 1, 0.2, 3, 5 au domeniul $[0.2, 5]$.
Statistică mulţime. O statistică mulţime este o mulţime calculată din

date.

Exemplu. Presupunem că avem 5 zaruri: cu 4, 6, 8, 12 şi 20 de feţe. Alegem aleator unul şi îl aruncăm. Valoarea aruncării este data. Mulţimea zarurilor pentru care această valoare este posibilă este o statistică mulţime. De exemplu, dacă aruncarea este 10, atunci valoarea acestei statistici mulţime este $\{12, 20\}$. Dacă aruncarea este 7, atunci această statistică mulţime are valoarea $\{8, 12, 20\}$.

O statistică este o variabilă aleatoare deoarece este calculată din date aleatoare. De exemplu, dacă datele provin din $N(\mu, \sigma^2)$, atunci media a n date are repartiţia $N(\mu, \sigma^2/n)$.

Repartiţia de selecţie. Repartiţia de probabilitate a unei statistici este numită repartiţia de selecţie a ei.

Estimare punctuală. Putem folosi statisticile pentru a face o estimare punctuală a parametrului θ. De exemplu, dacă parametrul θ reprezintă adevărata medie, atunci media datelor $\overline{x}$ este o estimare punctuală a lui θ.