Curs 10

Cristian Niculescu

1 Alegerea a priori

1.1 Scopurile învăţării

1. Să înveţe că alegerea a priori afectează a posteriori.
2. Să vadă că o a priori prea rigidă poate face dificilă folosirea datelor.
3. Să vadă că mai multe date scad dependenţa a posteriori de a priori.
4. Să poată face o alegere rezonabilă a a priori, bazată pe înţelegerea a priori a sistemului considerat.

1.2 Introducere

Până acum ni s-a dat totdeauna o pdf sau pmf a priori. În acest caz, deducţia statistică din date este în esenţă o aplicaţie a teoremei lui Bayes. Când a priori este cunoscută, nu sunt controverse despre cum trebuie procedat. Arta statisticii începe când a priori nu este cunoscută cu siguranţă. Sunt 2 şcoli principale despre cum să procedăm în acest caz: Bayesiană şi frecvenţionistă. Acum urmăm abordarea Bayesiană. Vom învăţa şi abordarea frecvenţionistă. Reaminitim că fiind cunoscute datele D şi ipoteza H am folosit teorema lui Bayes pentru a scrie

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

$$\text{a posteriori} \propto \text{verosimilitate} \cdot \text{a priori}$$

Bayesiană: Bayesienii fac deducţii folosind a posteriori $P(H|D)$ şi de aceea au nevoie totdeauna de o a priori $P(H)$. Dacă a priori nu este cunoscută cu certitudine, Bayesianul trebuie să încerce să facă o alegere rezonabilă. Sunt multe feluri de a face asta şi oameni rezonabili pot face alegeri diferite. În general este o practică bună justificarea alegerii şi explorarea unui domeniu de a priori pentru a vedea dacă toate indică aceeaşi concluzie.
Frecvenţionistă: Foarte scurt, frecvenţioniştii nu încearcă să creeze o a

priori. În schimb, ei fac deducţii folosind verosimilitatea $P(D|H)$.
2 beneficii ale abordării Bayesiene:
1. Probabilitatea a posteriori $P(H|D)$ pentru ipoteză date fiind dovezile este de obicei exact ce am vrea să ştim. Bayesianul poate spune ceva ca "parametrul de interes are probabilitatea 0.95 de a fi între 0.49 şi 0.51".
2. Presupunerile care se iau în considerare pentru alegerea a priori pot fi clar precizate.
Mai multe date bune: Totdeauna mai multe date bune permit concluzii mai puternice şi scad influenţa a priori. Accentul ar trebui să fie atât pe date bune (calitate) cât şi pe mai multe date (cantitate).

1.3 Exemplu: zaruri

Presupunem că avem un sertar plin de zaruri, fiecare dintre acestea având 4, 6, 8, 12 sau 20 de feţe. De această dată nu ştim câte zaruri de fiecare tip sunt în sertar. Un zar este ales la întâmplare din sertar şi aruncat de 5 ori. Rezultatele sunt în ordine 4, 2, 4, 7 şi 5.

1.3.1 A priori uniformă

Presupunem că nu avem idee care poate fi repartiţia zarurilor din sertar. În acest caz este rezonabil să folosim o a priori plată. Iată tabelul de actualizare pentru probabilităţile a posteriori care rezultă din actualizarea după fiecare aruncare. Pentru a încăpea toate coloanele, am eliminat a posteriori nenormalizate.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	1/5	1/4	0.370	1/4	0.542	1/4	0.682	0	0.000	0	0.000
H_6	1/5	1/6	0.247	1/6	0.241	1/6	0.202	0	0.000	1/6	0.000
H_8	1/5	1/8	0.185	1/8	0.135	1/8	0.085	1/8	0.818	1/8	0.876
H_{12}	1/5	1/12	0.123	1/12	0.060	1/12	0.025	1/12	0.161	1/12	0.115
H_{20}	1/5	1/20	0.074	1/20	0.022	1/20	0.005	1/20	0.021	1/20	0.009

Cunoscând datele, a posteriori finală este puternic ponderată spre ipoteza $H8$ că a fost ales un zar cu 8 feţe.

1.3.2 Alte a priori

Pentru a vedea cât de mult a posteriori de mai sus depinde de alegerea noastră a a priori, încercăm alte a priori. Presupunem că avem un motiv de a crede că sunt de 10 ori mai multe zaruri cu 20 de feţe în sertar decât de fiecare alt tip. Tabelul devine:

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.071	1/4	0.222	1/4	0.453	1/4	0.650	0	0.000	0	0.000
H_6	0.071	1/6	0.148	1/6	0.202	1/6	0.193	0	0.000	1/6	0.000
H_8	0.071	1/8	0.111	1/8	0.113	1/8	0.081	1/8	0.688	1/8	0.810
H_{12}	0.071	1/12	0.074	1/12	0.050	1/12	0.024	1/12	0.136	1/12	0.107
H_{20}	0.714	1/20	0.444	1/20	0.181	1/20	0.052	1/20	0.176	1/20	0.083

Chiar şi aici a posteriori finală este puternic ponderată spre ipoteza H_8.
Dar dacă zarurile cu 20 de feţe sunt de 100 de ori mai probabile decât fiecare din celelalte?

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.0096	1/4	0.044	1/4	0.172	1/4	0.443	0	0.000	0	0.000
H_6	0.0096	1/6	0.030	1/6	0.077	1/6	0.131	0	0.000	1/6	0.000
H_8	0.0096	1/8	0.022	1/8	0.043	1/8	0.055	1/8	0.266	1/8	0.464
H_{12}	0.0096	1/12	0.015	1/12	0.019	1/12	0.016	1/12	0.053	1/12	0.061
H_{20}	0.9615	1/20	0.889	1/20	0.689	1/20	0.354	1/20	0.681	1/20	0.475

Cu o astfel de convingere a priori puternică în zarurile cu 20 de feţe, a posteriori finală dă o mare pondere teoriei că datele sunt dintr-un zar cu 20 de feţe, chiar dacă este extrem de improbabil ca un zar cu 20 de feţe să producă un maxim de 7 în 5 aruncări. A posteriori dă acum şanse aproximativ egale ca să fi fost ales un zar cu 8 feţe versus un zar cu 20 de feţe.

1.3.3 A priori rigide

Disonanţă cognitivă uşoară. O convingere a priori prea rigidă poate copleşi orice cantitate de date. Presupunem că suntem convinşi că zarul trebuie să fie cu 20 de feţe. Deci punem a priori a noastră $P(H_{20}) = 1$ cu celelalte 4 ipoteze având probabilitatea 0. Iată ce se întâmplă în tabelul de actualizare.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0	1/4	0	1/4	0	1/4	0	0	0	0	0
H_6	0	1/6	0	1/6	0	1/6	0	0	0	1/6	0
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	1/8	0
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	1/12	0
H_{20}	1	1/20	1	1/20	1	1/20	1	1/20	1	1/20	1

Indiferent care sunt datele, o ipoteză cu probabilitatea a priori 0 va avea probabilitatea a posteriori 0. În acest caz nu vom scăpa niciodată de ipoteza H_{20}, cu toate că putem experimenta o uşoară disonanţă cognitivă.
Disonanţă cognitivă severă. A priori rigide pot de asemenea duce la absurdităţi. Presupunem că suntem convinşi că zarul trebuie să fie cu 4 feţe. Deci punem $P(H_4) = 1$ şi celelalte probabilităţi a priori 0. Cu datele cunoscute, la a 4-a aruncare intrăm în impas. 7 nu poate veni de la un zar cu 4 feţe. Totuşi, aceasta este singura ipoteză pe care o permitem. A posteriori a

noastră nenormalizată este o coloană de zerouri care nu poate fi normalizată.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	unnorm. post_4	post_4
H_4	1	1/4	1	1/4	1	1/4	1	0	0	???
H_6	0	1/6	0	1/6	0	1/6	0	0	0	???
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	???
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	???
H_{20}	0	1/20	0	1/20	0	1/20	0	1/20	0	???

Trebuie să ne ajustăm convingerile despre ce este posibil sau, mai probabil, suspectăm o greşeală accidentală sau deliberată a datelor.

1.4 Exemplu: malaria

Iată un exemplu real adaptat din *Statistics, A Bayesian Perspective* de Donald Berry:
Prin anii 1950, oamenii de ştiinţă au început să formuleze ipoteza că purtătorii genei falciforme erau mai rezistenţi la malarie ca nepurtătorii. Existau probe indirecte pentru această ipoteză. Aceasta ajuta şi la explicarea persistenţei în populaţie a unei gene altfel dăunătoare. Într-un experiment oamenii de ştiinţă au injectat 30 de voluntari africani cu malarie. 15 dintre voluntari purtau o copie a genei falciforme şi ceilalţi 15 erau nepurtători. 14 din cei 15 nepurtători şi doar 2 din cei 15 purtători au făcut malarie. Susţine acest mic eşantion ipoteza că gena falciformă protejează împotriva malariei?
Fie S un purtător al genei falciforme şi N un nepurtător. $D+$ indică dezvoltarea malariei şi $D-$ indică nedezvoltarea malariei. Datele pot fi puse într-un tabel.

	$D+$	$D-$	
S	2	13	15
N	14	1	15
	16	14	30

Înainte să analizăm datele ar trebui să spunem câteva cuvinte despre experiment şi proiectarea lui. În primul rând, este clar neetic: pentru a obţine ceva informaţie au infectat 16 oameni cu malarie. Trebuie de asemenea să ne îngrijorăm despre deplasare. Cum au ales subiecţii testului? Este posibil ca nepurtătorii să fi fost mai slabi şi astfel mai susceptibili la malarie decât purtătorii? Berry arată că este rezonabil să presupunem că o injecţie este similară cu o muşcătură de ţânţar, dar nu este garantat. Acest ultim punct înseamnă că dacă experimentul arată o relaţie între celule-seceră şi protecţie împotriva malariei injectate, trebuie să considerăm ipoteza că protecţia împotriva malariei transmisă de ţânţari este mai slabă sau inexis-

tentă. În sfârşit, vom formula ipoteza noastră ca "celulele-seceră protejează împotriva malariei", dar în realitate tot ce putem spera să spunem dintr-un studiu ca acesta este că "celula-seceră este corelată cu protecţia împotriva malariei".

Modelul. Pentru modelul nostru, fie θ_S probabilitatea că un purtător injectat S face malarie şi, analog, fie θ_N probabilitatea că un nepurtător injectat N face malarie. Presupunem independenţa între toţi subiecţii experimentului. Cu acest model, verosimilitatea este o funcţie de θ_S şi θ_N:

$$P(\text{date}|\theta_S, \theta_N) = c\theta_S^2(1-\theta_S)^{13}\theta_N^{14}(1-\theta_N).$$

Ca de obicei lăsăm factorul constant c ca o literă. (Este produsul a 2 coeficienţi binomiali: $c = C_{15}^2 \cdot C_{15}^{14}$.)

Ipoteze. Fiecare ipoteză constă dintr-o pereche (θ_N, θ_S). Pentru a păstra lucrurile simple vom considera doar un număr finit de valori pentru aceste probabilităţi. Am putea considera mult mai multe valori sau chiar un domeniu continuu pentru ipoteze. Presupunem că θ_N şi θ_S pot avea fiecare valorile 0, 0.2, 0.4, 0.6, 0.8 sau 1. Aceasta duce la tabele 2 dimensionale.

Primul este un tabel de ipoteze. Codul de culori indică următoarele:

1. Dreptunghiurile portocaliu deschis de-a lungul diagonalei sunt unde $\theta_S = \theta_N$, i.e. celulele-seceră nu contează în niciun fel.
2. Dreptunghiurile roz şi roşii de deasupra diagonalei sunt unde $\theta_N > \theta_S$, i.e. celulele-seceră dau protecţie împotriva malariei.
3. În dreptunghiurile roşii $\theta_N - \theta_S \geq 0.6$, i.e. celulele-seceră dau multă protecţie.
4. Dreptunghiurile albe de sub diagonală sunt unde $\theta_S > \theta_N$, i.e. celulele-seceră de fapt cresc probabilitatea de a face malarie.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	(0,1)	(.2,1)	(.4,1)	(.6,1)	(.8,1)	(1,1)
0.8	(0,.8)	(.2,.8)	(.4,.8)	(.6,.8)	(.8,.8)	(1,.8)
0.6	(0,.6)	(.2,.6)	(.4,.6)	(.6,.6)	(.8,.6)	(1,.6)
0.4	(0,.4)	(.2,.4)	(.4,.4)	(.6,.4)	(.8,.4)	(1,.4)
0.2	(0,.2)	(.2,.2)	(.4,.2)	(.6,.2)	(.8,.2)	(1,.2)
0	(0,0)	(.2,0)	(.4,0)	(.6,0)	(.8,0)	(1,0)

Ipotezele asupra nivelului de protecţie dată de S: roşu = mare; roz = mic; portocaliu = 0; alb = negativ.

Următorul este tabelul verosimilităţilor. (De fapt am profitat de indiferenţa

noastră la scalare şi am scalat toate verosimilităţile cu $100000/c$ pentru a face tabelul mai prezentabil.) Observăm că, la precizia tabelului, multe verosimilităţi sunt 0. Codul de culori este acelaşi ca în tabelul de ipoteze. Am evidenţiat cele mai mari verosimilităţi cu o margine albastră.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	1.93428	0.18381	0.00213	0.00000	0.00000
0.6	0.00000	0.06893	0.00655	0.00008	0.00000	0.00000
0.4	0.00000	0.00035	0.00003	0.00000	0.00000	0.00000
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Verosimilităţile $p(\text{date}|\theta_S, \theta_N)$ scalate cu $100000/c$.

1.4.1 A priori plată

Presupunem că nu avem nicio opinie despre dacă sau în ce măsură celula-seceră protejează împotriva malariei. În acest caz este rezonabil să folosim o a priori plată. Deoarece sunt 36 de ipoteze, fiecare primeşte o probabilitate a priori de 1/36. Aceasta apare în tabelul de mai jos. Reamintim că fiecare dreptunghi din tabel reprezintă o ipoteză. Deoarece este un tabel de probabilităţi includem pmf-urile marginale.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.8	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(\theta_S)$	1/6	1/6	1/6	1/6	1/6	1/6	1

A priori plată $p(\theta_S, \theta_N)$: fiecare ipoteză (dreptunghi) are aceeaşi probabilitate

Pentru a calcula a posteriori înmulţim tabelul verosimilităţilor cu tabelul a

priori şi normalizăm. Normalizarea ne asigură că suma din întregul tabel este 1.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88075	0.08370	0.00097	0.00000	0.00000	0.96542
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00002	0.00000	0.00000	0.00000	0.00018
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91230	0.08670	0.00100	0.00000	0.00000	1.00000

A posteriori la a priori plată: $p(\theta_S, \theta_N | \text{date})$
Pentru a decide dacă S dă protecţie împotriva malariei, calculăm probabilităţile a posteriori pentru "protecţie" şi "protecţie puternică". Acestea sunt calculate adunând numerele din dreptunghiurile corespunzătoare din tabelul a posteriori.

$$\text{Protecţie: } P(\theta_N > \theta_S) = \text{ suma din roz şi roşu } = 0.99995$$
$$\text{Protecţie puternică: } P(\theta_N - \theta_S \geq 0.6) = \text{ suma din roşu } = 0.88075.$$

Lucrând de la a priori plată, este efectiv sigur că celula-seceră dă protecţie şi foarte probabil că dă protecţie puternică.

1.4.2 A priori informată

Acest experiment nu a fost făcut fără informaţie a priori. Erau multe dovezi că gena falciformă oferea protecţie împotriva malariei. De exemplu, era raportat un mai mare procentaj de purtători care supravieţuiau până la maturitate.
Iată un mod de a construi o a priori informată: Vom rezerva o cantitate rezonabilă de probabilitate pentru ipoteza că S nu dă protecţie. Să zicem că 24% împărţit egal între cele 6 celule portocalii unde $\theta_N = \theta_S$. Ştim că nu ar trebui să punem nicio probabilitate a priori 0, deci hai să împărţim 6% din probabilitate între cele 15 celule de sub diagonală. Aceasta lasă 70% din probabilitate pentru cele 15 dreptunghiuri roz şi roşii de deasupra diagonalei.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	0.04667	0.04667	0.04667	0.04667	0.04667	0.04000	0.27333
0.8	0.04667	0.04667	0.04667	0.04667	0.04000	0.00400	0.23067
0.6	0.04667	0.04667	0.04667	0.04000	0.00400	0.00400	0.18800
0.4	0.04667	0.04667	0.04000	0.00400	0.00400	0.00400	0.14533
0.2	0.04667	0.04000	0.00400	0.00400	0.00400	0.00400	0.10267
0	0.04000	0.00400	0.00400	0.00400	0.00400	0.00400	0.06000
$p(\theta_S)$	0.27333	0.23067	0.18800	0.14533	0.10267	0.06000	1.0

A priori informată $p(\theta_S, \theta_N)$: foloseşte informaţia a priori că celula seceră protejează.
Apoi completăm pmf a posteriori.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88076	0.08370	0.00097	0.00000	0.00000	0.96543
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00001	0.00000	0.00000	0.00000	0.00017
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91231	0.08669	0.00100	0.00000	0.00000	1.00000

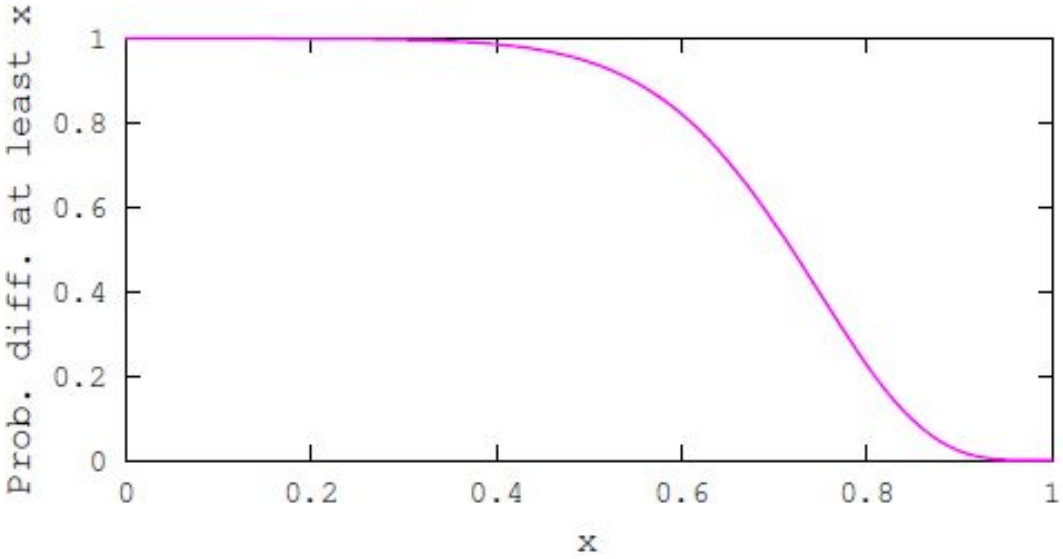
A posteriori la a priori informată: $p(\theta_S, \theta_N | \text{date})$
Calculăm din nou probabilităţile a posteriori ale "protecţiei" şi "protecţiei puternice".

$$\text{Protecţie: } P(\theta_N > \theta_S) = \text{ suma din roz şi roşu } = 0.99996$$
$$\text{Protecţie puternică: } P(\theta_N - \theta_S \geq 0.6) = \text{ suma din roşu } = 0.88076.$$

Observăm că a posteriori informată este aproape identică cu a posteriori din a priori plată.

1.4.3 PDALX

Următoarea reprezentare este bazată pe a priori plată. Pentru fiecare x, dă probabilitatea ca $\theta_N - \theta_S \geq x$. Pentru a o face netedă au fost folosite mult mai multe ipoteze.



Probabilitatea că diferenţa $\theta_N - \theta_S$ este cel puţin x (PDALX).
Observăm că este practic sigur că diferenţa este cel puţin 0.4.

2 Intervale de probabilitate

2.1 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.

2.2 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.

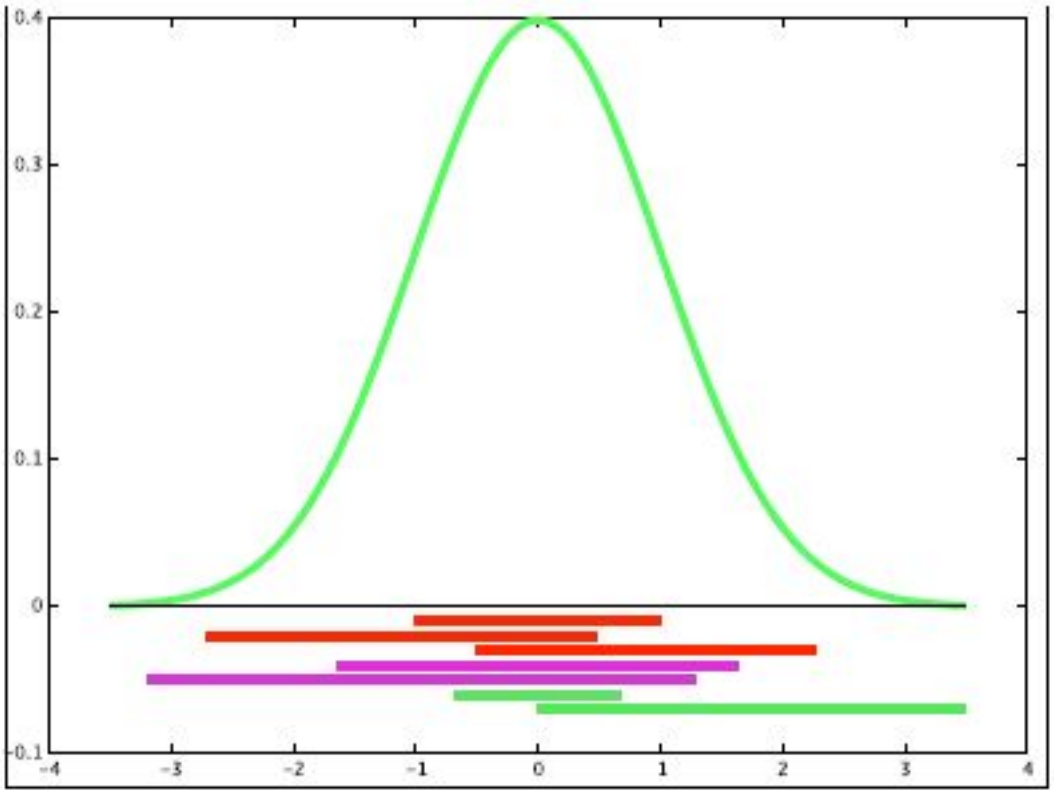
1. Diferite intervale de p-probabilitate pentru θ pot avea lungimi diferite. Putem face lungimea mai mică centrând intervalul sub cea mai înaltă parte a pdf. Un astfel de interval este de obicei o bună alegere deoarece conţine cele mai probabile valori.

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.

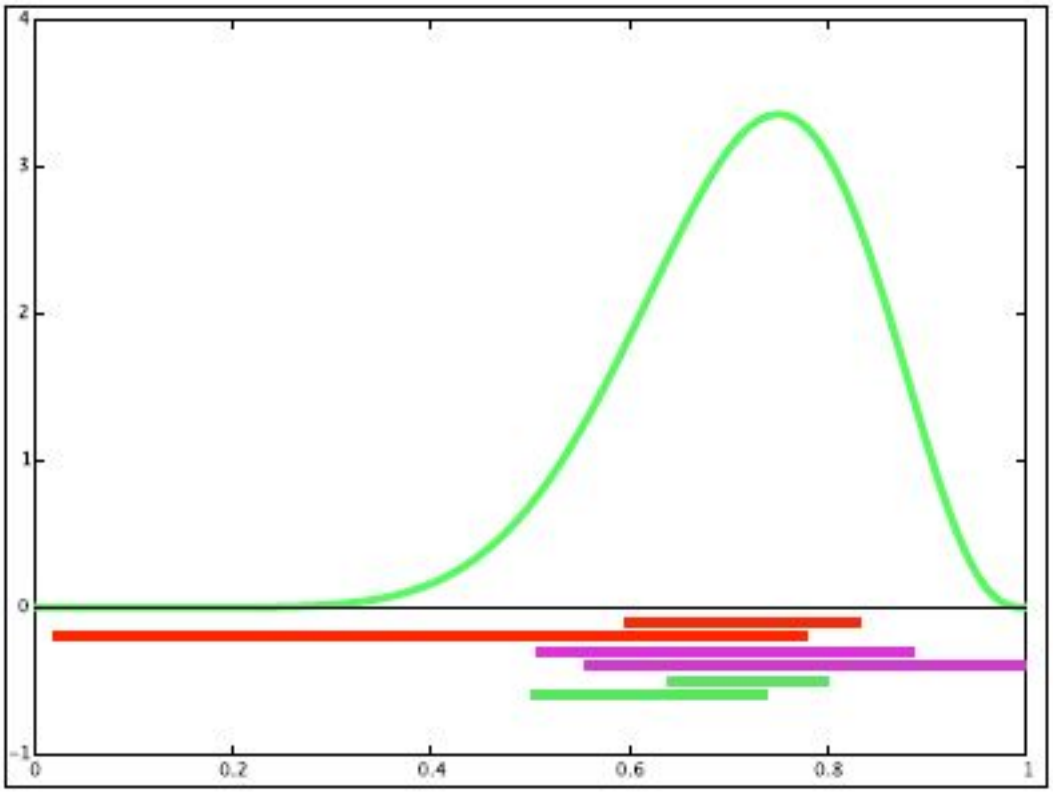
1. Toate barele roşii cuprind un interval de 0.68-probabilitate. Observaţi că cea mai mică bară roşie merge de la -1 la 1. Acest interval este de la a 16-a percentilă la a 84-a percentilă, deci este simetric.

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 $p = 0.68$.



2.3 Utilizări ale intervalelor de probabilitate

2.3.1 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.

2.4 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.

2.4.1 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

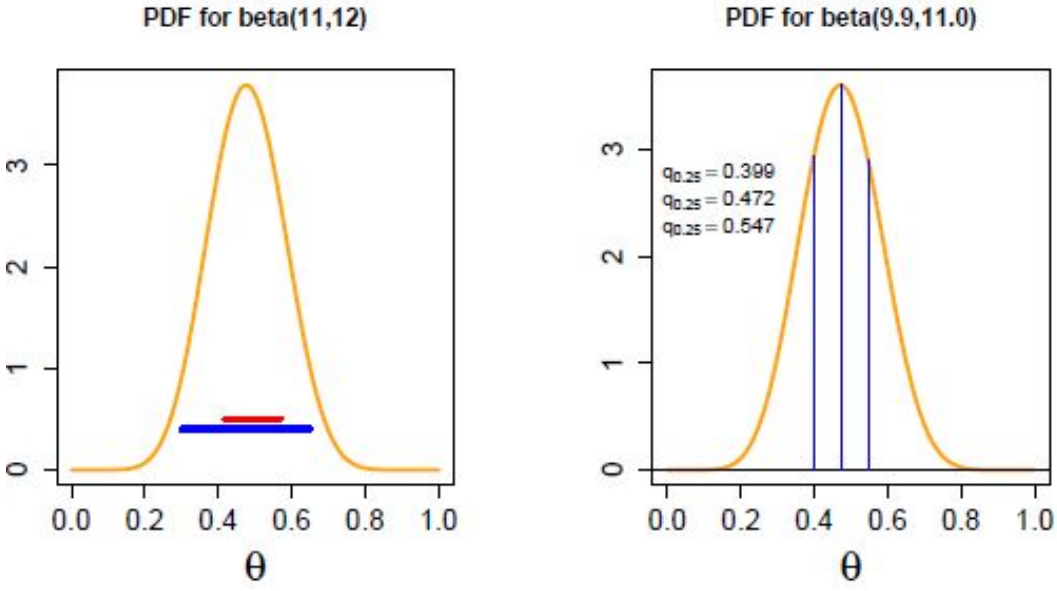
probabilitate ca el să ia peste 57%. Marginea inferioară pare mai greu de prezis. Vom lăsa 0.25 din probabilitate ca el să ia sub 42%.

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

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

2.4.2 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 3-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`.)

3 Şcoala frecvenţionistă de statistică

3.1 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.

3.2 Introducere

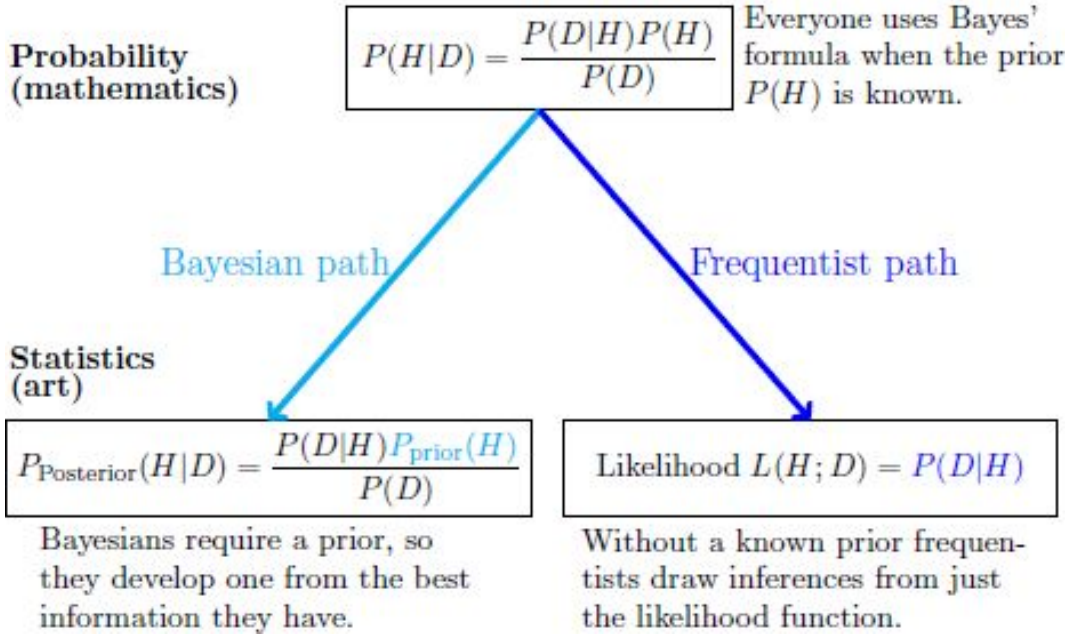
Pentru mare parte din secolul XX, statistica frecvenţioniostă a fost şcoala dominantă. Dacă aţi întâlnit vreodată intervale de încredere, p-valori, t-teste sau χ^2-teste, aţi văzut statistică frecvenţionistă. Odată cu dezvoltarea calculatoarelor de mare viteză şi a datelor mari, metodele Bayesiene devin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

3.2.1 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).

3.2.2 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.

3.3 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.