Curs 14

Cristian Niculescu

1 A priori conjugate: Beta şi normală

1.1 Scopurile învăţării

1. Să înţeleagă beneficiile a priori conjugate.
2. Să poată să actualizeze o a priori beta dată fiind o verosimilitate Bernoulli, binomială sau geometrică.
3. Să înţeleagă şi să poată folosi formula pentru actualizarea unei a priori normale fiind dată o verosimilitate normală cu dispersie cunoscută.

1.2 Introducere şi definiţie

Cu o a priori conjugată, a posteriori este de acelaşi tip, de exemplu pentru verosimilitate binomială, a priori beta devine o a posteriori beta. A priori conjugate sunt utile deoarece ele reduc actualizarea Bayesiană la modificarea parametrilor repartiţiei a priori (aşa numiţii hiperparametri) în locul calculului de integrale.
Ne vom concentra pe 2 exemple importante de a priori conjugate: beta şi normală. O listă mult mai cuprinzătoare este în tabelele din `http://en.wikipedia.org/wiki/Conjugate_prior_distribution`.
Definiţie. Presupunem că avem date cu funcţia de verosimilitate $f(x|\theta)$ depinzând de un parametru θ. Mai presupunem că repartiţia a priori pentru θ este una dintr-o familie de repartiţii parametrizate. Dacă repartiţia a posteriori pentru θ este în aceeaşi familie ca repartiţia a priori, spunem că a priori este o a priori conjugată pentru verosimilitate.

1.3 Repartiţia beta

Arătăm că repartiţia beta este o a priori conjugată pentru verosimilităţi binomiale, Bernoulli şi geometrice.

1.3.1 Verosimilitate binomială

Am văzut că repartiţia beta este o a priori conjugată pentru repartiţia binomială. Aceasta înseamnă că dacă funcţia de verosimilitate este binomială şi repartiţia a priori este beta, atunci repartiţia a posteriori este tot beta.
Mai concret, presupunem că funcţia de verosimilitate are o repartiţie binomială(N, θ), unde N este cunoscut şi θ este parametrul (necunoscut) de interes. Avem de asemenea că data x este un întreg între 0 şi N. Atunci, pentru o a priori beta avem următorul tabel:

hypothesis	data	prior	likelihood	posterior
θ	x	beta(a, b)	binomial(N, θ)	beta$(a + x, b + N - x)$
θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$c_2\theta^x(1-\theta)^{N-x}$	$c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

Tabelul este simplificat scriind coeficienţii de normalizare ca c_1, c_2 şi c_3.

$$c_1 = \frac{(a+b-1)!}{(a-1)!(b-1)!},\ c_2 = C_N^x = \frac{N!}{x!(N-x)!},\ c_3 = \frac{(a+b+N-1)!}{(a+x-1)!(b+N-x-1)!}.$$

1.3.2 Verosimilitate Bernoulli

Repartiţia beta este o a priori conjugată pentru repartiţia Bernoulli. Acesta este de fapt un caz special al repartiţiei binomiale, deoarece Bernoulli(θ) este aceeaşi ca binomiala$(1, \theta)$. În tabelul de mai jos, arătăm actualizările corespunzând succesului $(x = 1)$ şi eşecului $(x = 0)$ pe linii separate.

hypothesis	data	prior	likelihood	posterior
θ	x	beta(a, b)	Bernoulli(θ)	beta$(a + 1, b)$ or beta$(a, b + 1)$
θ	$x = 1$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	θ	$c_3\theta^a(1-\theta)^{b-1}$
θ	$x = 0$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$1-\theta$	$c_3\theta^{a-1}(1-\theta)^b$

Constantele c_1 şi c_3 au aceleaşi formule ca în cazul precedent (al verosimilităţii binomiale) cu $N = 1$.

1.3.3 Verosimilitate geometrică

Repartiţia geometrică(θ) descrie probabilitatea a x eşecuri înaintea primului succes, unde probabilitatea succesului în fiecare încercare independentă este θ. Pmf corespunzătoare este $p(x) = \theta(1-\theta)^x$.
Acum presupunem că avem o dată x şi ipoteza noastră θ este că x este extrasă dintr-o repartiţie geometrică(θ). Din tabel vedem că repartiţia beta este o a priori conjugată pentru o verosimilitate geometrică:

ipoteza	data	a priori	verosimilitatea	a posteriori
θ	x	beta(a, b)	geometrică(θ)	beta$(a + 1, b + x)$
θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$\theta(1-\theta)^x$	$c_3\theta^a(1-\theta)^{b+x-1}$

Exemplul 1. În timp ce călătoreau prin Regatul Ciupercilor, Mario şi Luigi

au găsit nişte monede neobişnuite. Ei au căzut de acord asupra unei a priori $f(\theta) \sim \text{beta}(5,5)$ pentru probabilitatea aversului, dar nu au fost de acord ce experiment să facă pentru a investiga θ.
a) Mario decide să arunce o monedă de 5 ori. El obţine un avers în 5 aruncări.
b) Luigi decide să arunce o monedă până la primul avers. El obţine 4 reversuri înaintea primului avers.
Arătaţi că Mario şi Luigi ajung la aceeaşi a posteriori pentru θ şi calculaţi această a posteriori.

Răspuns. Tabelul lui Mario:

ipoteza	data	a priori	verosimilitatea	a posteriori
θ	$x=1$	beta$(5,5)$	binomială(θ)	???
θ	$x=1$	$c_1\theta^4(1-\theta)^4$	$C_5^1\theta(1-\theta)^4$	$c_3\theta^5(1-\theta)^8$

Tabelul lui Luigi:

ipoteza	data	a priori	verosimilitatea	a posteriori
θ	$x=4$	beta$(5,5)$	geometrică(θ)	???
θ	$x=4$	$c_1\theta^4(1-\theta)^4$	$\theta(1-\theta)^4$	$c_3\theta^5(1-\theta)^8$

Atât a posteriori a lui Mario cât şi a lui Luigi au forma unei repartiţii beta$(6,9)$. Factorul de normalizare este acelaşi în ambele cazuri deoarece este determinat cerând ca probabilitatea totală să fie 1.

1.4 Normala generează normală

Repartiţia normală este a priori conjugată cu ea însăşi. În particular, dacă funcţia de verosimilitate este normală cu dispersie cunoscută, atunci o a priori normală dă o a posteriori normală. Acum atât ipotezele şi datele sunt continue.
Presupunem că avem o măsurare $x \sim N(\theta, \sigma^2)$, unde dispersia σ^2 este cunoscută. Adică, media θ este parametrul nostru necunoscut de interes şi ştim că verosimilitatea vine dintr-o repartiţie normală cu dispersia σ^2. Dacă alegem o pdf a priori normală

$$f(\theta) \sim N(\mu_{\text{prior}}, \sigma^2_{\text{prior}}),$$

atunci pdf a posteriori este de asemenea normală: $f(\theta|x) \sim N(\mu_{\text{post}}, \sigma^2_{\text{post}})$, unde

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{x}{\sigma^2}, \quad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2}. \tag{1}$$

Următoarea formă a acestor formule este mai uşor de citit şi arată că μ_{post} este o medie ponderată între μ_{prior} şi data x.

$$a = \frac{1}{\sigma^2_{\text{prior}}}, \quad b = \frac{1}{\sigma^2}, \quad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b}, \quad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{2}$$

Cu aceste formule în minte, putem exprima actualizarea prin tabelul:

hypothesis	data	prior	likelihood	posterior
θ	x	$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$	$f(x\|\theta) \sim \mathrm{N}(\theta, \sigma^2)$	$f(\theta\|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$
θ	x	$c_1 \exp\left(\frac{-(\theta-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}\right)$	$c_2 \exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right)$	$c_3 \exp\left(\frac{-(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\right)$

Demonstraţia formulelor generale se face analog ca în următorul exemplu numeric.

Exemplul 2. Presupunem că avem a priori $\theta \sim N(4, 8)$ şi verosimilitatea $x \sim N(\theta, 5)$. Presupunem de asemenea că avem o măsurare $x_1 = 3$. Arătaţi că repartiţia a posteriori este normală.

Răspuns.

a priori: $f(\theta) = c_1 e^{-(\theta-4)^2/16}$; verosimilitatea: $f(x_1|\theta) = c_2 e^{-(x_1-\theta)^2/10} = c_2 e^{-(3-\theta)^2/10}$.

Înmulţim a priori cu verosimilitatea pentru a obţine a posteriori:

$$f(\theta|x_1) = c_1 c_2 e^{-(\theta-4)^2/16} e^{-(3-\theta)^2/10} = c_1 c_2 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right).$$

Completăm pătratul din exponent

$$\begin{aligned} -\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} &= -\frac{5(\theta-4)^2 + 8(3-\theta)^2}{80} \\ &= -\frac{13\theta^2 - 88\theta + 152}{80} \\ &= -\frac{\theta^2 - \frac{88}{13}\theta + \frac{152}{13}}{80/13} \\ &= -\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}. \end{aligned}$$

De aceea, a posteriori este

$$f(\theta|x_1) = c_1 c_2 e^{-\frac{(\theta-44/13)^2 + 152/13 - (44/13)^2}{80/13}} = c_3 e^{-\frac{(\theta-44/13)^2}{80/13}}.$$

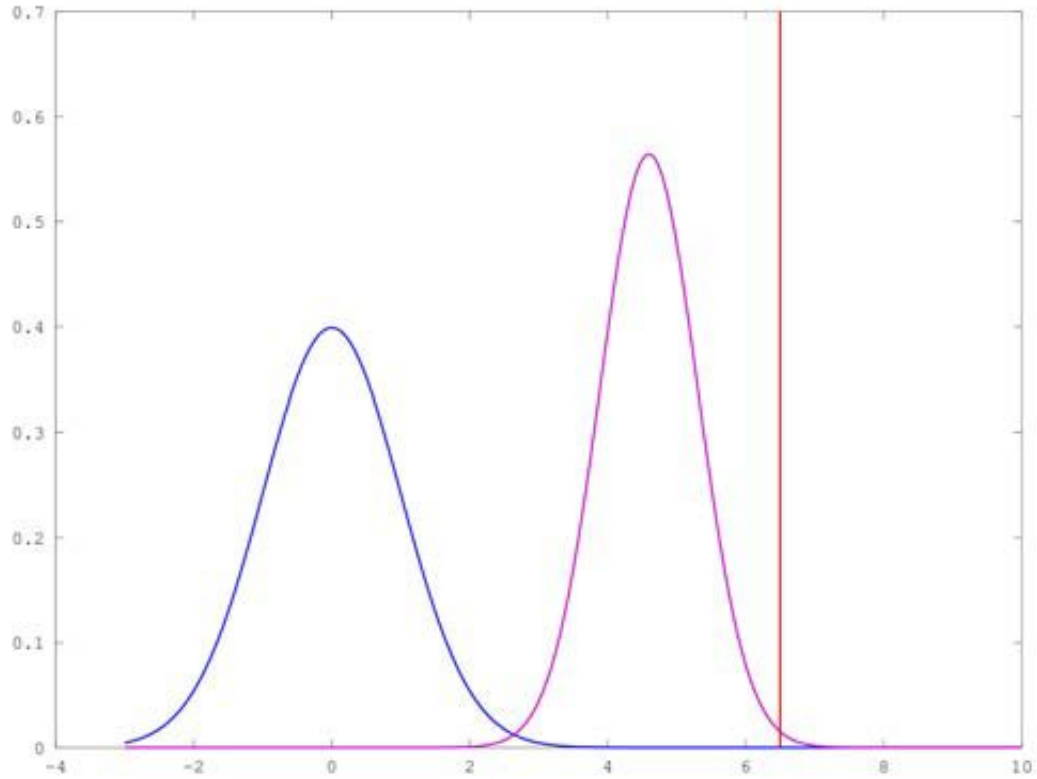
Aceasta are forma pdf pentru $N(44/13, 40/13)$, q.e.d.
Verificăm aceasta cu formulele (2).

$$\mu_{\text{prior}} = 4,\ \sigma^2_{\text{prior}} = 8,\ \sigma^2 = 5 \implies a = \frac{1}{8},\ b = \frac{1}{5}.$$

De aceea

$$\begin{aligned} \mu_{\text{post}} &= \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{\frac{1}{8}\cdot 4 + \frac{1}{5}\cdot 3}{\frac{1}{8}+\frac{1}{5}} = \frac{44}{13} \approx 3.38, \\ \sigma^2_{\text{post}} &= \frac{1}{a+b} = \frac{1}{\frac{1}{8}+\frac{1}{5}} = \frac{40}{13} \approx 3.08. \end{aligned}$$

Exemplul 3. Presupunem că ştim datele $x \sim N(\theta, \sigma^2)$ şi avem a priori $N(0, 1)$. Obţinem o valoare a datelor $x = 6.5$. Descrieţi schimbările pdf pentru θ în actualizarea de la a priori la a posteriori.
Răspuns. Iată graficul pdf-urilor a priori, a posteriori cu data marcată cu o line roşie.



A priori în albastru, a posteriori în mov, data în roşu.
Media a posteriori va fi o medie ponderată dintre media a priori şi dată. Vârful pdf a posteriori va fi între vârful a priori şi linia roşie. Avem

$$\sigma^2_{\text{post}} = \frac{1}{1/\sigma^2_{\text{prior}} + 1/\sigma^2} = \sigma^2_{\text{prior}} \cdot \frac{\sigma^2}{\sigma^2_{\text{prior}} + \sigma^2} < \sigma^2_{\text{prior}}.$$

Adică a posteriori are dispersie mai mică decât a priori, i.e. data ne face mai siguri despre unde este θ în domeniul său.

1.4.1 Mai mult de o dată

Exemplul 4. Presupunem că avem datele x_1, x_2, x_3. Folosiţi formulele (1) pentru a actualiza succesiv.
Răspuns. Notăm media şi dispersia a priori cu μ_0, respectiv σ_0^2. Mediile şi

dispersiile actualizate vor fi μ_i, respectiv σ_i. Avem succesiv

$$\frac{1}{\sigma_1^2}=\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2};\ \frac{\mu_1}{\sigma_1^2}=\frac{\mu_0}{\sigma_0^2}+\frac{x_1}{\sigma^2}$$
$$\frac{1}{\sigma_2^2}=\frac{1}{\sigma_1^2}+\frac{1}{\sigma^2}=\frac{1}{\sigma_0^2}+\frac{2}{\sigma^2};\ \frac{\mu_2}{\sigma_2^2}=\frac{\mu_1}{\sigma_1^2}+\frac{x_2}{\sigma^2}=\frac{\mu_0}{\sigma_0^2}+\frac{x_1+x_2}{\sigma^2}$$
$$\frac{1}{\sigma_3^2}=\frac{1}{\sigma_2^2}+\frac{1}{\sigma^2}=\frac{1}{\sigma_0^2}+\frac{3}{\sigma^2};\ \frac{\mu_3}{\sigma_3^2}=\frac{\mu_2}{\sigma_2^2}+\frac{x_3}{\sigma^2}=\frac{\mu_0}{\sigma_0^2}+\frac{x_1+x_2+x_3}{\sigma^2}.$$

Exemplul se generalizează la n valori ale datelor $x_1,...,x_n$:

Formule de actualizare normală-normală pentru n date

$$\frac{\mu_{\text{post}}}{\sigma_{\text{post}}^2}=\frac{\mu_{\text{prior}}}{\sigma_{\text{prior}}^2}+\frac{n\overline{x}}{\sigma^2},\ \frac{1}{\sigma_{\text{post}}^2}=\frac{1}{\sigma_{\text{prior}}^2}+\frac{n}{\sigma^2},\ \overline{x}=\frac{x_1+...+x_n}{n}. \tag{3}$$

Din nou dăm o formă mai simplu de citit, arătând că μ_{post} este o medie ponderată între μ_{prior} şi media de selecţie $\overline{x}$:

$$a=\frac{1}{\sigma_{\text{prior}}^2},\ b=\frac{n}{\sigma^2},\ \mu_{\text{post}}=\frac{a\mu_{\text{prior}}+b\overline{x}}{a+b},\ \sigma_{\text{post}}^2=\frac{1}{a+b}. \tag{4}$$

Interpretare: μ_{post} este o medie ponderată între μ_{prior} şi $\overline{x}$. Dacă numărul datelor este mare, atunci ponderea b este mare şi $\overline{x}$ va avea o puternică influenţă asupra a posteriori. Dacă σ_{prior}^2 este mică, atunci ponderea a este mare şi μ_{prior} va avea o puternică influenţă asupra a posteriori. Pentru a rezuma:
1. Multe date au o mare influenţă asupra a posteriori.
2. Siguranţa mare (dispersia mică) în a priori are o mare influenţă asupra a posteriori.

2 Alegerea a priori

2.1 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.

2.2 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).

2.3 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.

2.3.1 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.

2.3.2 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.

2.3.3 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.

2.4 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 inexistentă. 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 $100000/c$.

2.4.1 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.

2.4.2 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

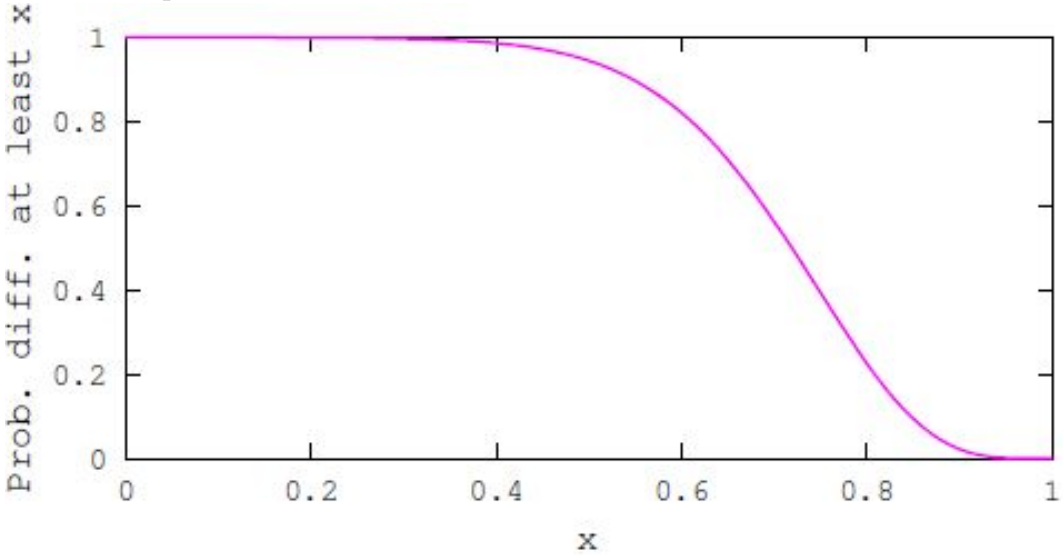
A posteriori la a priori informată: $p(\theta_S, \theta_N | \text{date})$
Calculăm din nou probabilităţile a posteriori ale "protecţiei" şi "protecţiei puternice".

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

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

2.4.3 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.