MS-A0504 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

5B Lisää Bayes-päättelystä

Jukka Kohonen

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2021–2022
Periodi IV

Sisältö

Miten käyttää posteriorijakaumaa?

Viime luennolla opittiin, miten tuntemattomalle parametrille Θ saadaan havaintojen perusteella uusi jakauma (posteriorijakauma). Mihin jakaumaa voi käyttää?

Kuten mitä tahansa tn-jakaumaa. Tapoja on monta, riippuen siitä

1. mihin kysymykseen halutaan vastaus
2. mitä on (riittävän) helppo laskea.

Tyypillisiä tarkasteltavia asioita:

- posteriorin moodi = piste, jossa sen arvo on suurin — vrt. SU-estimaatti
- posteriorin odotusarvo = tn-painotettu keskiarvo
- posteriorin mediaani = piste, jonka alla 50% tn:stä
- uskottavuusväli, joka sisältää esim. 95% posteriorista
- raportoidaan / piirretään koko posteriorijakauma
- tulevan datan ennustaminen posteriorin perusteella

Näitä kaikkia käsitellään tällä luennolla.

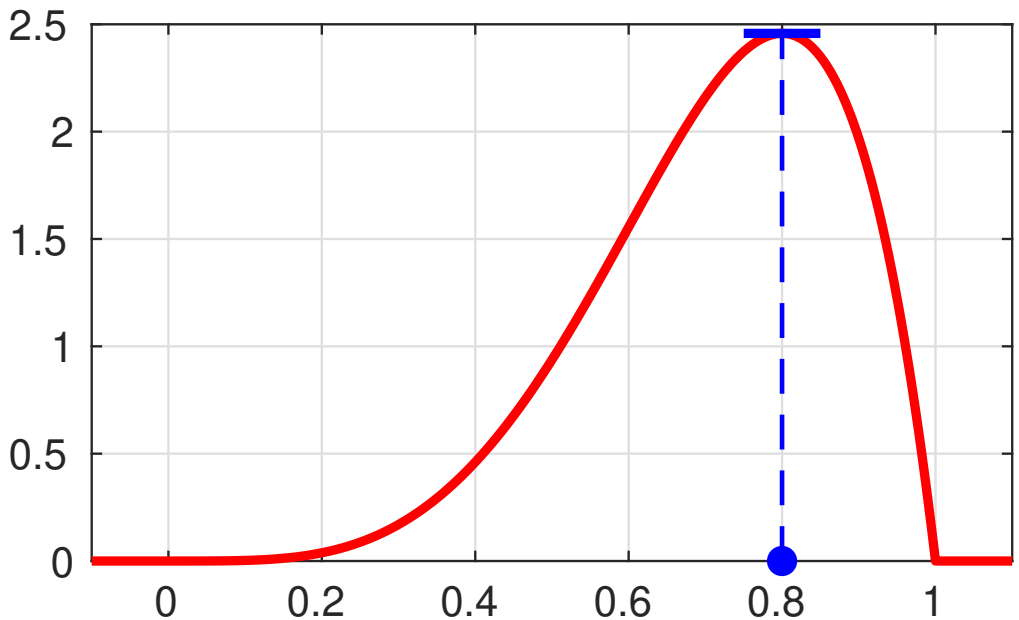
Posteriorin moodi (MAP = **M**ax **A** **P**osteriori estimate)

Tuntematon kolikko, tasapriori, havaittu 4 kruunaa, 1 klaava.
Posteriori on Beta(5,2), jonka tiheys on (kun $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

Funktion maksimikohta löydetään tyypillisesti tutkimalla derivaatan nollakohdat ja välin päätepisteet.

(Normalisointivakio 30 ei vaikuta maksimointiin, joten voisi yhtä hyvin käyttää normalisoimatonta posterioria. Vrt. myös suurimman uskottavuuden estimaatti!)



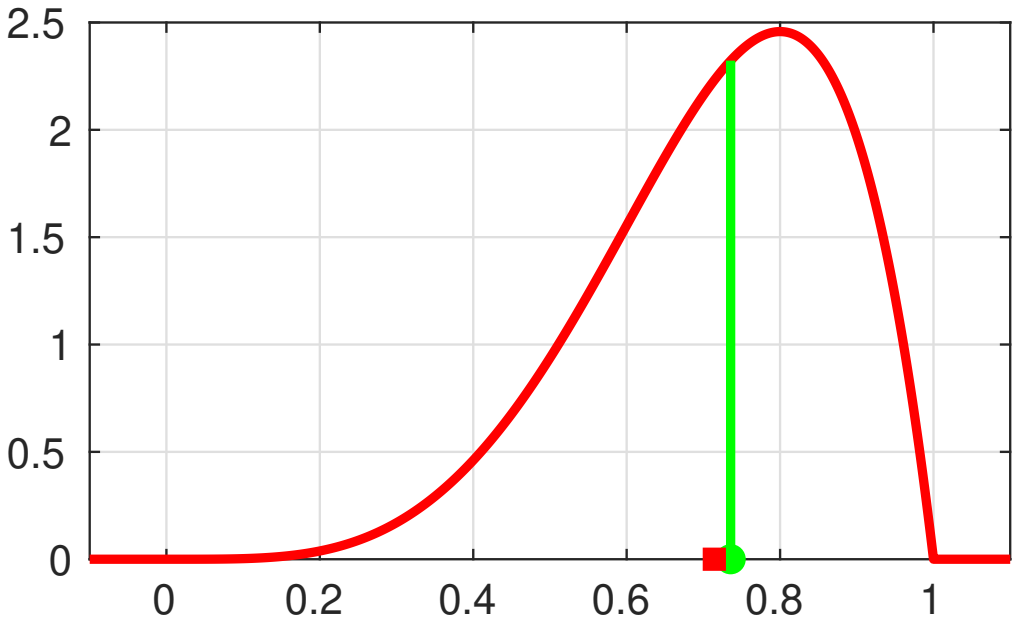
Posteriorimoodi = MAP-estimaatti = 0.8

Posteriorin odotusarvo ja mediaani

Tuntematon kolikko, tasapriori, havaittu 4 kruunaa, 1 klaava.
Posteriori on Beta(5,2), jonka tiheys on (kun $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

Odotusarvo voidaan laskea integraalina tuttuun tapaan.
Mediaanin voi etsiä ratkaisemalla missä kertymä=0.5. (R: `qbeta`)



Odotusarvo $= 5/7 \approx 0.7143$

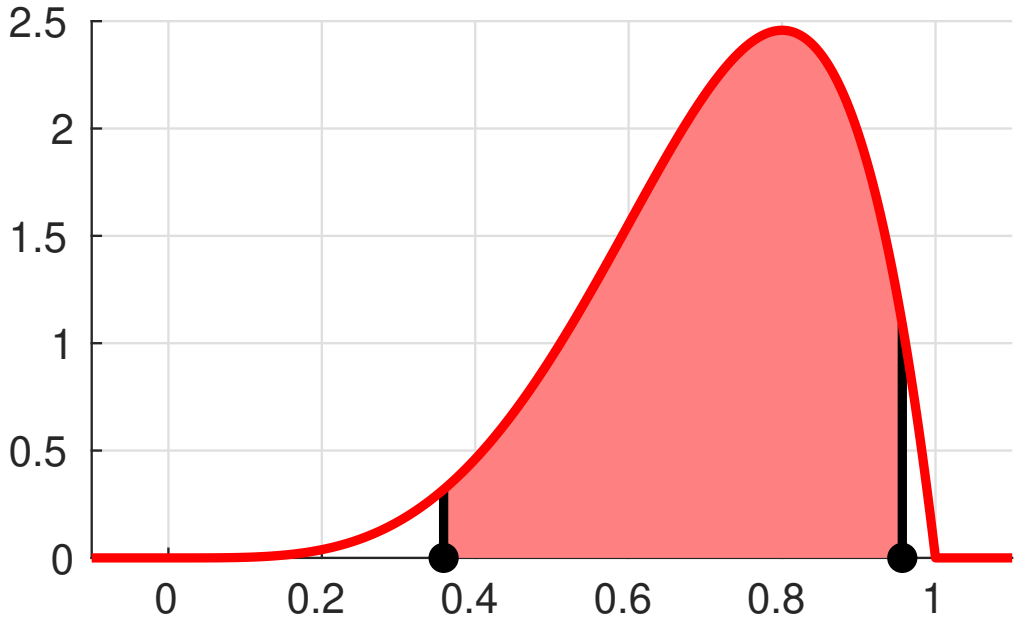
Mediaani ≈ 0.7356

Uskottavuusväli engl. credible interval

Tuntematon kolikko, tasapriori, havaittu 4 kruunaa, 1 klaava.
Posteriori on Beta(5,2), jonka tiheys on (kun $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

Etsi pisteet joissa kertymä=0.025 ja kertymä=0.975. (R: `qbeta`)
$\Rightarrow \Theta$ on niiden välillä tn:llä 95%.



$$\mathbb{P}(0.3588 \leq \Theta \leq 0.9567 \mid \vec{X} = \vec{x}) = 95\%$$

Tulevan datan ennustaminen

- Posteriorijakauma sisältää parhaan tietämyksemme siitä, millainen arvo Θ:lla voi olla (priorin ja datan perusteella).
- Yleensä posteriorijakauma ei keskity yhteen pisteeseen. Toisin sanoen se ilmaisee aidosti epävarmuutemme parametrin arvosta. Ei ole ehkä perusteltua "leikkiä", että tiedämme parametrin arvon olevan juuri (jonkun) piste-estimaatin kohdalla.
- Toki posteriorijakauma kapenee sitä mukaa kuin dataa saadaan lisää.

Kysymys. Nähtyämme viisi havaintoa $\vec{x} = (1, 1, 1, 1, 0)$ meillä on posteriori $\Theta \sim \text{Beta}(5, 2)$. Mitä voimme sanoa seuraavasta havainnosta?

Vastaus. Laskemme sille (posteriori)ennustejakauman käyttäen tn-laskennan osituskaavaa.

Ennustejakauma (kolikkoesimerkki)

Viisi havaintoa $\vec{x}$, ennustetaan kuudes havainto Y. Osituskaavan (additiivisuuden) perusteella

$$f_{Y|\vec{X}}(y \mid \vec{x}) = \int f(y \mid \theta) f(\theta \mid \vec{x}) d\theta.$$

Tulkinta: Eri θ-arvot antavat erilaisia ennusteita Y:lle. Osituskaava ottaa nämä kaikki mahdollisuudet huomioon ja painottaa niitä Θ:n posterioritiheyden mukaisesti.

Tämä on Y:n ennustejakauma, joka perustuu kaikkeen siihen mitä tiedämme Θ:sta (priorin ja datan perusteella).

Huom. Emme **valinneet** ennustustehtävää varten vain yhtä θ:n arvoa (esim. todennäköisintä).

Huom. Emme edes "hylänneet" melko epätodennäköisiä arvoja (jakauman häntiä), ne ovat mukana tarkastelussa.

Ennustejakauma (kolikkoesimerkki)

Viisi havaintoa $\vec{x}$, ennustetaan kuudes havainto Y.

- stokastinen malli on kuten aiemminkin $\mathbb{P}(Y = 1 \mid \Theta = \theta) = \theta$.
- Θ:n posteriorijakauma on Beta(5,2).

Joten lasketaan

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid \vec{X} = \vec{x}) &= f_{Y|\vec{X}}(1 \mid \vec{x}) \\
&= \int f_{Y|\Theta}(1 \mid \theta)\, f_{\Theta|\vec{X}}(\theta \mid \vec{x}) d\theta \\
&= \int_0^1 \theta \cdot 30\theta^4(1-\theta) d\theta \\
&= 30 \int_0^1 (\theta^5 - \theta^6) d\theta \\
&= 30 \cdot \left(\frac{1}{6} - \frac{1}{7}\right) \approx \mathbf{0.7143}.
\end{aligned}$$

Yhden havainnon ennusteessa tn onkin sama kuin Θ:n posteriorin odotusarvo. Mutta ei innostuta liikaa...

Ennustejakauma (ennustetaan pitemmälle)

Viisi havaintoa $\vec{x}$, ennustetaan ovatko seuraavat kolme heittoa kruunia eli $\vec{Y} = (1,1,1)$.

- stokastinen malli $\mathbb{P}(Y_i = 1 \,|\, \Theta = \theta) = \theta$.
- Θ:n posteriorijakauma on Beta(5,2).

Joten lasketaan

$$\begin{aligned}
\mathbb{P}(\vec{Y} = (1,1,1) \,|\, \vec{X} = \vec{x}) &= f_{\vec{Y}|\vec{X}}(1,1,1 \,|\, \vec{x}) \\
&= \int f_{\vec{Y}|\Theta}(1,1,1 \,|\, \theta)\; f_{\Theta|\vec{X}}(\theta \,|\, \vec{x}) d\theta \\
&= \int_0^1 \theta^3 \cdot 30\theta^4(1-\theta) d\theta \\
&= 30 \int_0^1 (\theta^7 - \theta^8) d\theta \\
&= 30 \cdot \left(\frac{1}{8} - \frac{1}{9} \right) \approx \mathbf{0.4167}.
\end{aligned}$$

Tulos **ei ole** $0.7143^3 \approx 0.3645$ vaan isompi. Huom. kuution paikka.

Ennustejakaumat — epävarmuuden vaikutus

Kun ennusteessa aidosti huomioidaan Θ:n arvoon liittyvä epävarmuus, tämä näkyy ennusteissa.

Yksinkertaisuuden vuoksi tässä kolikkoesimerkki diskreetillä Θ-parametrilla.

Vertaa seuraavia kahta tilannetta:

- Malli A: Tiedämme kolikon olevan varmasti reilu ($\Theta = 0.5$).
- Malli B: Kolikon Θ-parametri saa arvot $0, 0.5, 1$ todennäköisyyksin $0.01, 0.98, 0.01$.

Yhden heittotuloksen ennustamisessa molemmat mallit antavat saman tuloksen (tulos on kruuna tn:llä 0.5).

Sadan heittotuloksen ennustamisessa mallit antavat täysin erilaiset ennusteet.

Ennustejakaumat — epävarmuuden vaikutus

Mikä on tn, että seuraavat 100 heittoa ovat kaikki kruunia?

Malli A: Tiedämme kolikon olevan varmasti reilu ($\Theta = 0.5$).

- Kruunien määrä seur. 100 heitolla on Bin(100, 0.5)-jakautunut
- 100 kruunaa tn:llä $1/2^{100} \approx 8 \cdot 10^{-31}$

Malli B: Kolikon Θ-parametri saa arvot 0, 0.5, 1 todennäköisyyksin $0.01, 0.98, 0.01$.

- Osituskaavan mukaisesti, tn saada 100 kruunaa on

$$(0.01 \cdot 0) + (0.98 \cdot 8 \cdot 10^{-31}) + (0.01 \cdot 1) \approx 0.01$$

Huom: Jos malli B vastaa tekemiämme havaintoja, niin kaikki seuraavat menettelyt antaisivat vääriä ennusteita:

- pelkän moodin $\Theta = 0.5$ käyttäminen ennustamiseen
- pelkän odotusarvon $\Theta = 0.5$ käyttäminen ennustamiseen
- 5% häntien poistaminen (poistaisi molemmat äärikohdat)

Epävarmuuden pitäminen mukana laskuissa kannattaa, koska tulokset ovat oikeammat!

Sisältö

Monen vaihtoehdon toistokoe

Aiemmin käsitellyssä binaarimallissa data oli $\{0, 1\}$-arvoista ja mallin ainoa parametri Θ kuvasi ykkösen todennäköisyyttä. (Θ:lle voitiin ottaa erilaisia priorijakaumia.)

Entä jono satunnaismuuttujia, joilla on $k > 2$ mahdollista arvoa?

Esim.

- Painotetun nopan heittoja (3,6,6,2,6,1,3,4,6,6)
- Puoluekantoja otoksessa (A,B,A,A,C,B,A,A,C,C)
- DNA-sekvenssi jossa neljää "kirjainta" GTCTACCAG...
- Tekstiä kirjainjonona kielen kirjainfrekvenssien mukaan
  T, h, e, väli, q, u, i, c, k
- Tekstiä sanajonona, jokainen sana valitaan tietystä jakaumasta (the, quick, brown, fox, jumped, over, the, lazy, dog)

Dataa voidaan tarkastella jonona luokka-arvoisia muuttujia tai luokkien esiintyvyyksien muodostamana numeerisena vektorina.

Multinomimalli

- n riippumatonta satunnaismuuttujaa $\vec{X} = (X_1, X_2, \ldots, X_n)$.
- Kullakin X_i sama diskreetti jakauma k:sta vaihtoehdosta
- Jakaumalla on k kpl todennäköisyysparametreja $\vec{p} = (p_1, p_2, \ldots, p_k)$
- Jos näitä pidetään satunnaismuuttujina, muodostuu satunnaisvektori $\vec{P} = (P_1, P_2, \ldots, P_k)$

Jo opitut menetelmät toimivat (kunhan yksityiskohdat selvitetään)

- Parametrivektorille oletetaan priorijakauma $f_{\vec{P}}(\vec{p})$
- Datalle oletetaan stokastinen malli $f(\vec{x} \,|\, \vec{p})$ (uskottavuus)
- Havaintojen perusteella lasketaan posteriori $f(\vec{p} \,|\, \vec{x})$

Multinomimallin stokastinen malli — 3 luokkaa

Väestössä on kolmen puolueen A, B, C kannattajia osuuksin $\vec{p} = (p, q, r) = (0.5, 0.3, 0.2)$.

Poimitaan satunnaisesti $n = 10$ henkilöä. Kullakin poimitulla henkilöllä on em. todennäköisyydet $\vec{p}$ kuulua em. puolueiden kannattajiin.

Tarkastellaan kahta kysymystä:

- Millaisia (järjestettyjä) henkilöjonoja saatamme havaita?
  esim. AAAABBBBCC
- Millaisia lukumäärävektoreita saatamme havaita?
  esim. $(4, 4, 2)$

Esimerkiksi

- $\mathbb{P}(\text{AAAAAAAAAA}) = p^{10} \approx 0.000977$ Pieni
- $\mathbb{P}(\text{AAAABBBBCC}) = p^4 q^4 r^2 \approx 0.000020$ Pienempi!?

Multinomimallin stokastinen malli — 3 luokkaa

Kombinatoriikasta tiedämme, että on $3^{10} = 59049$ erilaista 10 kirjaimen jonoa (3 vaihtoehdolla). Luetellaan ne ja ryhmitellään sen kirjainten A,B,C lukumäärien mukaan. Muistetaan $(p, q, r) = (0.5, 0.3, 0.2)$.

jono	kirjainmäärät	$\mathbb{P}$(jono)	
AAAAAAAAAA	$(10, 0, 0)$	$p^{10} = 0.000977$	} 1 kpl
...			
AAAABBBBCC	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	} 3150 kpl
BBCAABBAAC	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
AABCCAABBB	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
...			
CCBBBBAAAA	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
...			
CCCCCCCCCC	$(0, 0, 10)$	$r^{10} = 0.0000001$	} 1 kpl

$$\mathbb{P}(\text{lukumäärät 4,4,2}) = 3150 \times 0.000020 = 0.0638 \approx 6.4\%$$

$$\mathbb{P}(\text{lukumäärät 10,0,0}) = 1 \times 0.000977 \approx 0.1\%$$

Välihuomio — multinomikertoimet

Mistä tuli luku 3150 edellisellä kalvolla?
Kyseessä on ns. multinomikerroin, joka ilmaisee, miten monella tavalla voidaan järjestää 4 A-kirjainta, 4 B-kirjainta ja 2 C-kirjainta.

Voidaan päätellä seuraavasti tuloperiaatteen ja binomikertoimen avulla:

- 10 paikkaa, valitaan 4:ään A-kirjain: $\binom{10}{4} = 210$ tapaa
- Jäljellä 6 paikkaa, valitaan 4:ään B-kirjain: $\binom{6}{4} = 15$ tapaa
- Jäljellä 2 paikkaa, niihin menee C-kirjain.

Tuloperiaate: $210 \cdot 15 = 3150$ erilaista kirjainjonoa.

Tuloksen voi laskea suoraan multinomikertoimella

$$\binom{10}{4,4,2} = \frac{10!}{4!\,4!\,2!} = 3150.$$

Mahdolliset lukumäärävektorit ja niiden tn:t

$3^{10} = 59049$ eri **havaintojonoa**, mutta 66 eri **lukumäärävektoria**.

(5,3,2)	**0.0851**	(2,5,3)	0.0122	(4,0,6)	0.000840
(6,2,2)	0.0709	(2,4,4)	0.0102	(2,8,0)	0.000738
(6,3,1)	0.0709	(4,6,0)	0.0096	(1,3,6)	0.000726
(4,4,2)	0.0638	(2,6,2)	0.0092	(1,8,1)	0.000590
(5,4,1)	0.0638	(3,2,5)	0.0091	(2,1,7)	0.000346
(5,2,3)	0.0567	(4,1,5)	0.0076	(0,6,4)	0.000245
(4,3,3)	0.0567	(7,0,3)	0.0075	(0,7,3)	0.000210
(7,2,1)	0.0506	(8,0,2)	0.0070	(1,2,7)	0.000207
(4,5,1)	0.0383	(9,1,0)	0.0059	(0,5,5)	0.000196
(3,4,3)	0.0340	(2,3,5)	0.0054	(3,0,7)	0.000192
(7,1,2)	0.0338	(6,0,4)	0.0053	(0,8,2)	0.000118
(6,1,3)	0.0315	(2,7,1)	0.0039	(0,4,6)	0.000109
(3,5,2)	0.0306	(9,0,1)	0.0039	(1,9,0)	0.000098
(4,2,4)	0.0284	(3,7,0)	0.0033	(0,3,7)	0.000041
(6,4,0)	0.0266	(5,0,5)	0.0025	(0,9,1)	0.000039
(7,3,0)	0.0253	(1,6,3)	0.0024	(1,1,8)	0.000035
(3,3,4)	0.0227	(1,5,4)	0.0024	(2,0,8)	0.000029
(8,1,1)	0.0211	(3,1,6)	0.0020	(0,2,8)	0.000010
(5,5,0)	0.0191	(2,2,6)	0.0018	(0,10,0)	0.000006
(5,1,4)	0.0189	(1,4,5)	0.0016	(1,0,9)	0.000003
(8,2,0)	0.0158	(1,7,2)	0.0016	(0,1,9)	0.000002
(3,6,1)	0.0153	(10,0,0)	0.000977	(0,0,10)	0.000000

Multinomimalli — Diskreetti priori

Parametrivektorilla voi tilanteen mukaan olla **erilaisia prioreja**. Yksinkertaisimmillaan parametrivektorilla on vain muutamia (tai vain kaksi) mahdollista arvoa, jolloin priorijakauma on diskreetti. Esim. pussissa kahdenlaisia noppia, 9 reilua ja 1 epäreilu, jonka painotus tunnetaan. Satunnaisesti poimittu noppa on

- tn:llä 0.9 reilu, silmälukujen 1–6 tn:t $\vec{p} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
- tn:llä 0.1 epäreilu, tn:t $\vec{p} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.5)$

Nyt esim. kolmen heiton jonolla $(6, 1, 2)$ parametrivektorin kahdelle arvolle saadaan uskottavuudet $(\frac{1}{6})^3$ ja $0.5 \cdot 0.1 \cdot 0.1$. Posteriorijakauma (samoin vain 2 arvoa) lasketaan tuttuun tapaan posteriori $\propto$ priori $\cdot$ uskottavuus.

Multinomimalli — Jatkuva priori

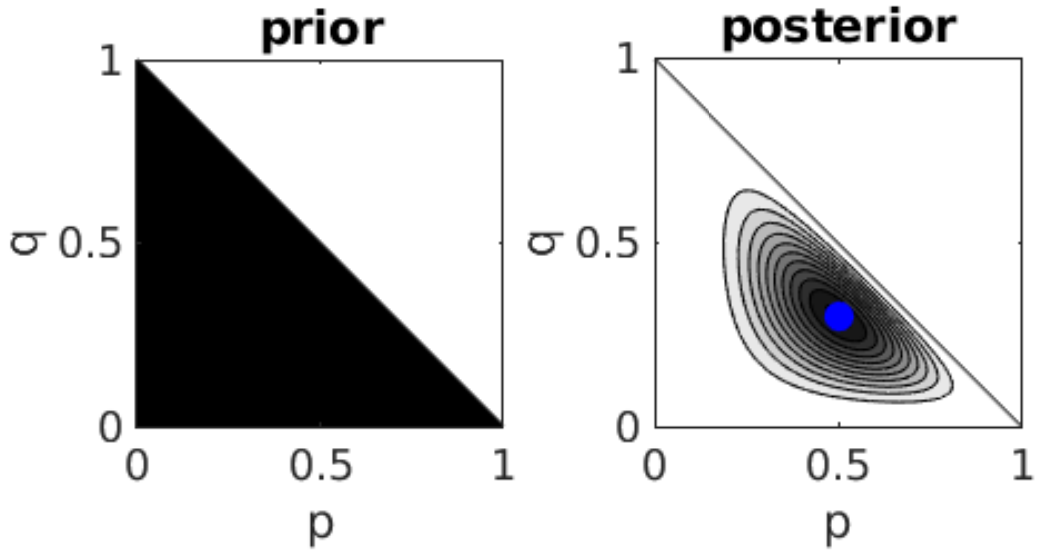
Entä jos parametrit p, q, r voivat saada mitä tahansa reaaliarvoja välillä $[0, 1]$?

- Parametrit (p, q, r) eivät ole täysin vapaita, sillä täytyy olla $p + q + r = 1$.
- Tarkastellaan kahden parametrin vektoria (p, q), jolloin $r = 1 - (p + q)$.
- Vaaditaan $p \geq 0$ ja $q \geq 0$ ja $p + q \leq 1$. Siten (p, q) rajoittuu kolmiomaiseen alueeseen.
- Oletetaan prioriksi tasajakauma kolmioalueessa,

$$f_{P,Q}(p, q) = 2 \qquad \text{jos } p, q \geq 0 \text{ ja } p + q \leq 1.$$

- Nyt meillä on uskottavuus (stokastinen malli) ja priori (parametrivektorin jakauma), joten voidaan laskea posteriori.

Multinomimalli — Jatkuva posteriori



Jos kolmen puolueen kannattajia havaittiin lukumäärät $(5, 3, 2)$, niin satunnaisvektorin (P, Q) posterioritiheys on

$$f(p, q \mid \vec{x}) = c \cdot p^5 q^3 (1 - p - q)^2$$

kolmioalueessa. (c on normalisointivakio)
Posterioritiheydestä voidaan laskea moodi (helposti), odotusarvo (vähän vaikeampi), 95% uskottavuusalue, ennusteita tulevalle datalle jne. Kuvassa posteriorimoodi sinisenä pisteenä.

Sisältö

Priorin valinnasta

Bayes-päättely saattaa vaikuttaa (epäilyttävän?) subjektiiviselta. Jos priorin voi valita vapaasti, niin tokihan voi saada aikaan aivan minkä tahansa haluamansa posteriorin?

- Priori tulee valita rehellisesti, pyrkien kuvaamaan kohtuullisella tarkkuudella, mitä parametrista Θ tiedetään (ennen datan havaitsemista).
- Tasajakauma on usein hyvä priori. Ei aina, jos malli on monimutkainen.
- Varo asettamasta prioritiheyttä nollaksi sellaisille parametriarvoille, jotka voivat ehkä sittenkin toteutua. (Jos priori=0, niin posteriori=0, datasta riippumatta.)
- Datamäärän kasvaessa priorin vaikutus joka tapauksessa vähenee eli "data puhuu puolestaan".
- Kun raportoit tuloksia, raportoi mitä mallia ja prioria käytettiin. Silloin tulokset ovat täysin objektiiviset: kuka tahansa saa ko. lähtöoletuksilla saman posteriorin.

Ensi viikolla jatketaan tilastollisen merkitsevyyden testaamisesta...