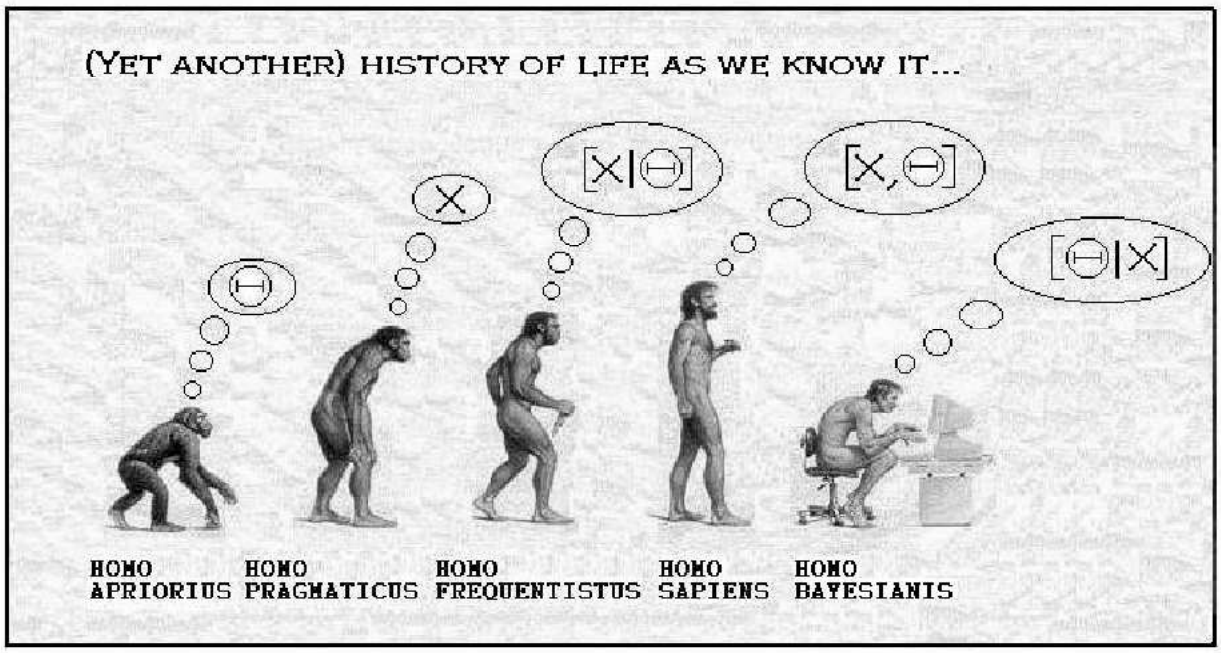
Bayes-tilastotiede 1, 5 op

Arto Luoma[1]
Matematiikan ja tilastotieteen laitos
PL 35, (MaD)
40014 Jyväskylän yliopisto
Syksy 2019



[1]Luentorunko perustuu Antti Penttisen aiempaan luentodiasarjaan.

Sisältö

Kirjallisuutta

Pääasiallinen lähde

- Gelman Andrew, Carlin John B., Stern Hal S., Dunson, David B., Vehtari Aki and Rubin Donald B (2013). Bayesian Data Analysis, Third Edition, CRC Press.

Muuta kirjallisuutta ja viitteitä

- Albert (2007) J. Bayesian computation with R. Springer.
- Marin and Robert (2014). Bayesian Essentials with R, Second Edition. Springer.
- Gelman Andrew and Hill Jennifer (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
- Lee Peter M. (2012). Bayesian Statistics: An Introduction. Fourth Edition, Wiley.
- de Finetti, Bruno (2017). Theory of Probability: A critical introductory treatment. Chichester: John Wiley & Sons ltd.

Luku 1

Johdanto

Bayes-tilastotiede perustuu Bayesin lauseeseen, jonka alkuperäisen version esitti Thomas Bayes. Hän oli englantilainen antikonformisti, presbyteeri reviisori ja harrastelijamatemaatikko.

Thomas Bayes (1702–1761)

Bayes ei itse julkaissut todennäköisyyteen liittyviä tutkimuksiaan. Kuitenkin hänen ystävänsä Richard Price julkaisi ehdollisia todennäköisyyksiä käsittelevän artikkelin Bayesin muistiinpanojen perusteella [1] tämän kuole-

[1] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances.

man jälkeen.

Bayes-tilastotiede tunnettiin 1800-luvulla ”käänteisten todennäköisyyksien” menetelmänä. Todennäköisyyslaskenta, jota Pascal ja Fermat olivat kehittäneet 1650-luvulta lähtien, kykeni ratkaisemaan ”suoran” ongelman: Jos tehdään n riippumatonta satunnaiskoetta ja kussakin kokeessa onnistumisen todennäköisyys on θ, niin onnistumisten lukumäärä noudattaa binomijaumaa: $X \sim \mathrm{Bin}(n, \theta)$. Bayesin artikkeli ratkaisi käänteisen ongelman: mikä on parametrin θ jakauma, kun tunnetaan satunnaiskokeiden lopputulokset. Bayes oletti, että parametriin θ sisältyvä epävarmuus voitiin kuvata tasajakauman avulla: $\theta \sim \mathrm{Tas}(0, 1)$. Tällöin hän saattoi määrittää havaintoihin perustuvan ehdollisen todennäköisyyden, että θ on välillä $[a, b]$:

$$\mathsf{P}(a < \theta < b \,|\, X = x) = \frac{\int_a^b \binom{n}{x} \theta^x (1-\theta)^{n-x} \mathrm{d}\theta}{\int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \mathrm{d}\theta}.$$

Bayesiläisyyteen on liittynyt vahva filosofinen lataus. Se on jo 1700-luvun lopusta alkaen ollut yksi vaihtoehtoinen lähestymistapa tieteelliseen päättelyyn, myös kiistelty. Edelleenkin tilastotieteilijöiden keskuudessa voidaan havaita jako Bayes-tilastotieteen kannattajiin ja vastustajiin. Väliin mahtuvat ”pragmaattiset” bayesiläiset, jotka soveltavat Bayes-teoriaa mm. yleisenä tapana muotoilla ongelmia ja ratkaista niitä.

Bayes-tilastotieteen käyttöä rajoitti pitkään se, että tilastollisen päättelyn ongelmiin oli suljetun muodon ratkaisuja ainoastaan yksinkertaisissa tapauksissa. Tietokoneiden yleistyminen ja niiden laskukapasiteetin kasvu on poistanut tämän rajoitteen. Erityisesti 1990-luvun alusta alkaen Bayes-tilastotiede on kehittänyt tehokkaita laskennallisia työkaluja, joiden avulla voidaan lähestyä hyvinkin monimutkaisua ongelmia. Tämä on tehnyt lähestymistavasta hyvin suositun, niin että sitä sovelletaan monilla tieteenaloilla.

1.1 Bayesiläisen data-analyysin vaiheet

1. Valitaan todennäköisyysmalli, joka on tutkimuskohteen havaittavien ja ei-havaittavien suureiden yhteisjakauma. Mallin tulisi olla yhteensopiva tutkittavaan ongelmaan liittyvän tiedon ja aineistonkeräysprosessin kanssa.

Philosophical Transactions of the Royal Society, 330–418. (Reprinted with biographical note by G.A. Barnard in *Biometrika* **45**, 293–315, 1958.)

2. Lasketaan ja tulkitaan posteriorijakauma, joka on kiinnostavien, ei-havaittavien suureiden ehdollinen jakauma, kun havaintoaineisto on annettu.

3. Arvioidaan mallin yhteensopivuus havaintoaineiston kanssa ja posteriorijakauman perusteella tehtävien päätelmien järkevyys. Tutkitaan, miten herkkiä tulokset ovat mallin oletuksille. Tarvittaessa muutetaan tai laajennetaan mallia ja toistetaan kolme askelta.

1.2 Keskeisiä käsitteitä

Tilastollisessa päättelyssä pyritään tekemään päätelmiä numeerisen aineiston perusteella ei-havaituista suureista. *Estimoitavat suureet (estimandit)* ovat joko 1) mahdollisesti havaittavia, kuten prosessin tulevat havainnot ja puuttuvat havainnot, tai 2) parametreja. Parametrejä ei voi edes periaatteessa suoraan havaita; ne määrittävät havainnot tuottavaa prosessia. Merkitsemme parametreja kreikkalaisilla kirjaimilla, vektoreita ja skalaareja pienillä roomalaisilla kirjaimilla ja matriiseja isoilla kirjaimilla. Käytämme yleistä merkintää θ kiinnostuksen kohteena olevien parametrien vektorille, merkintää y havaitulle aineistolle ja merkintää $\tilde{y}$ tuntemattomille mutta mahdollisesti havaittaville suureille.

Yleensä tilastollisessa tutkimuksessa kerätään tietoa n havaintoyksiköstä ja havaintoaineisto voidaan esittää vektorina $y = (y_1, ..., y_n)$. Jos kustakin yksiköstä mitataan useampi kuin yksi muuttuja, havainnot y_i ovat vektoreita ja koko havaintojoukko y on matriisi. Yleensä oletetaan, että havainnot y_i ovat *vaihdannaisia (exchangeable)*, mikä tarkoittaa sitä, että havaintojen yhteisjakauma $p(y_1, ..., y_n)$ on riippumaton niiden indeksoinnista. Näin on esimerkiksi silloin, kun havainnot ovat riippumattomia ja samoin jakautuneita *(iid)* kiinteällä parametrin θ arvolla. *Hierarkisissa malleissa eli monitasomalleissa* on eri tasoilla olevia havaintoyksikköjä ja kullakin tasolla voidaan erikseen puhua vaihdannaisuudesta.

On yleistä, että havaintoaineistossa on mukana *selittäviä muuttujia* eli *kovariaatteja* x, joita ei kannata mallintaa satunnaisina. Jos kuitenkin pidetään selittäviä muuttujia satunnaisina, voidaan määritellä, että havaintovektorit $(x, y)_i$ ovat vaihdannaisia, jos niiden yhteisjakauma pysyy muuttumattomana, vaikka vaihdettaisiin indeksien järjestystä mielivaltaisella tavalla.

Merkitsemme yleisesti reunajakauman tiheys- tai todennäköisyysfunktio-

ta $p(.)$:llä ja ehdollista tiheys- tai todennäköisyysfunktiota $p(.|.)$:llä riippumatta siitä, onko kyseessä parametrin vai havaintojen jakauma. Merkintä $N(\mu, \sigma^2)$ viittaa satunnaismuuttujaan ja $N(\theta|\mu, \sigma^2)$ tiheysfunktioon. Jatkossa voimme puhua lyhyesti jakaumasta, kun tarkoitamme jakauman tiheys- tai todennäköisyysfunktiota.

Luku 2

Todennäköisyys epävarmuuden mittana

Bayesiläisessä tilastotieteessä todennäköisyyskäsitteen tulkinta on laajempi kuin klassisessa. Klassinen tilastotiede perustuu yleensä todennäköisyyden frekventistiseen tulkintaan, jonka mukaan tapahtuman todennäköisyys on tapahtuman esiintymisten suhteellinen osuus riippumattomissa toistokokeissa. Esim. kolikon heitossa kruunan todennäköisyys on $1/2$, koska kruuna esiintyy noin puolessa tapauksista pitkässä heittosarjassa. Bayesiläisessä tilastotieteessä todennäköisyyttä käytetään yleisemmin epävarmuuden mittarina ja todennäköisyydestä voidaan puhua sellaistenkin tapahtumien yhteydessä, jotka eivät ole toistettavissa. Esim. voidaan pohtia, mikä on todennäköisyys, että huomenna sataa tai että joku tietty henkilö voittaa presidentinvaalit. Ks. perusteluja todennäköisyyden käytöstä epävarmuuden mittarina Gelmanin kirjasta.

2.1 Todennäköisyyden aksioomat (Kolmogorovin aksioomat)

Todennäköisyyslaskennassa *tapahtumat* E ovat perusjoukon Ω osajoukkoja. Tapahtumille E määriteltävä todennäköisyys $\mathsf{P}(.)$ toteuttaa seuraavat aksioomat:

- $\mathsf{P}(E) \geq 0$, missä E on jokin tapahtuma (perusjoukon osajoukko),
- $\mathsf{P}(\Omega) = 1$, missä Ω on perusjoukko,

- $\mathsf{P}(\cup_n E_n) = \sum_n \mathsf{P}(E_n)$, missä joukot E_n ovat pareittain pistevieraita ja muodostavat numeroituvan kokoelman joukkoja.

2.2 Ehdollinen todennäköisyys

Bayesiläinen tilastotiede perustuu ehdollisten todennäköisyyksien laskentaan. Tapahtuman E todennäköisyys ehdolla H määritellään $\mathsf{P}(E|H) = \mathsf{P}(E \cap H)/\mathsf{P}(H)$.

Tapahtuma H edustaa taustatietoa tai joskus hypoteesia. Ehdollinen todenäköisyys $\mathsf{P}(E|H)$ mittaa epävarmuutta tiedon H valossa. Sitä voidaan tulkita seuraavasti:

- $\mathsf{P}(E|H) = 1$, jos olet varma, että E tapahtuu,
- $\mathsf{P}(E|H) = 0$, jos olet varma, että E ei tapahdu,
- $\mathsf{P}(E|H) = p\,, \quad 0 < p < 1$, jos E:hen liittyy epävarmuutta (mutta **ei** välttämättä satunnaisuutta),
- $\mathsf{P}(E|H) > \mathsf{P}(F|H)$, jos E:n epävarmuus on pienempi kuin F:n.

Kahdella tarkastelijalla voi olla kuitenkin eri käsitys epävarmuudesta, koska heidän taustatietonsa poikkeavat (eri H). Ehdollinen todennäköisyys $\mathsf{P}(E|H)$ muuttuu, kun informaatio H muuttuu. Siksi Bayes-teoria perustuu *subjektiivisiin todennäköisyyksiin*[1].

Seuraavat ominaisuudet seuraavat aksioomista ja ehdollisen todennäköisyyden määritelmästä:

- $\mathsf{P}(E|H) \geq 0$ kaikilla E ja H,
- $\mathsf{P}(H|H) = 1$ kaikilla H,
- $\mathsf{P}(E \cup F|H) = \mathsf{P}(E|H) + \mathsf{P}(F|H)$, kun E ja F ovat toisensa poissulkevia,
- $\mathsf{P}(E \cap F|H) = \mathsf{P}(F|H)\mathsf{P}(E|F,H)$ kaikilla E,F ja H.

[1] Näin tulkitsevat Bayes-todennäköisyyden *subjektivistit*, joita edustaa de Finetti (ks. kirjallisuusviite). *Objektivistit* tulkitsevat todennäköisyyden logiikan laajennuksena. Heidän mukaansa eri tarkastelijat päätyvät samaan lopputulokseen, jos heillä on tarkalleen samat tiedot käytössään.

2.3 Vedonlyönti ja subjektiivinen todennäköisyys

Subjektiivisen todennäköisyys voidaan määrittää vedonlyöntisuhteen avulla. Veto E:stä vedonlyöntisuhteella $\omega : 1$ (at odds) ja panoksella (stake) M tarkoittaa:

- jos E ei toteudu, menetät M;
- jos E toteutuu, voitat $\omega \cdot M$.

Jo uskot vahvasti E:hen, hyväksyt vedon pienellä vedonlyöntisuhteella $\omega(H)$ (ω riippuu H:sta). Olkoon $\tilde{\omega}(H)$ *reilun pelin* vedonlyöntisuhde. Reilussa pelissä

$$\mathsf{P}(E|H) \cdot \tilde{\omega}(H)M + \mathsf{P}(\bar{E}|H) \cdot (-M) = 0\,,$$

missä $\bar{E} = \Omega \setminus E$ on E:n komplementtitapahtuma. E:n epävarmuudeksi saadaan todennäköisyys

$$\mathsf{P}(E|H) = \frac{1}{1 + \tilde{\omega}(H)}\,.$$

Näin siis voit periaatteessa määrittää henkilökohtaisen subjektiivisen todennäköisyytesi sen perusteella, minkä arvioit reilun pelin vedonlyöntisuhteeksi.

Huomaa, että tilastotieteessä ja todennäköisyyslaskennassa vedonlyöntisuhde määritellään hiukan eri tavalla. Vedonlyöntisuhde E:n puolesta on todennäköisyyksien suhde $\mathsf{P}(E|H)/\mathsf{P}(\bar{E}|H)$, ja E:tä vastaan se on $\mathsf{P}(\bar{E}|H)/\mathsf{P}(E|H)$.

2.4 Todennäköisyyden päivittyminen

Voit määritellä E:hen liittyvän epävarmuuden (ehdollisena) todennäköisyytenä $\mathsf{P}(E|H)$. Kun hankit uutta tietoa F, voit "oppia" siitä. Nyt E:hen liittyvä epävarmuus **ei** välttämättä enää ole $\mathsf{P}(E|H)$ vaan $\mathsf{P}(E|F,H)$.

Esim. 1 Heitetään pinssiä, jonka toinen puoli (kuvapuoli) on kupera. Olkoon tulos 0, jos kuvapuoli on ylöspäin, ja 1 jos se on alaspäin. Merk. E:llä tapahtumaa "1". Ennakkokäsitys epävarmuudesta olkoon $\mathsf{P}(E|H) = \frac{1}{2}$.

Olkoon F koesarja, jossa on 10 heittoa:

0	0	0	0	0	1	1	0	0	1

Tämän jälkeen epävarmuus on $\mathsf{P}(E|F,H) < \frac{1}{2}$.

Tilastotiede on "taitolaji": miten voidaan oppia havainnoista – ts. miten voidaan pienentää johonkin asiaan liittyvää epävarmuutta. *Bayes-tilastotiede* tarjoaa oppimiseen todennäköisyyslaskentaan perustuvan kehikon.

2.5 Riippumattomuus (independence)

Määr. Tapahtumat E ja F ovat riippumattomia, jos $\mathsf{P}(E \cap F) = \mathsf{P}(E) \cdot \mathsf{P}(F)$.

Riippumattomuus on yhtäpitävää sen kanssa, että

- $\mathsf{P}(E|F) = \mathsf{P}(E) = \mathsf{P}(E|\bar{F})$ tai
- $\mathsf{P}(F|E) = \mathsf{P}(F)$.

Tämä voidaan ymmärtää niin, että E ja F ovat riippumattomia, jos toisen (E tai F) tunteminen ei muuta käsitystä toisen epävarmuudesta. F:stä ei opita, kun pohditaan E:n epävarmuutta.

Vastaavasti määritellään ehdollinen riippumattomuus: E ja F ovat riippumattomia ehdolla H, jos

$$\mathsf{P}(E \cap F|H) = \mathsf{P}(E|H) \cdot \mathsf{P}(F|H).$$

Riippumattomuus ei kuitenkaan ole aivan mutkaton asia, mikä voidaan ymmärtää seuraavasta esimerkistä:

Esim. 2 Laatikossa on r punaista ja k valkoista palloa. Tehdään otanta takaisinsijoittaen. Tarkastellaan tapahtumia

E = "punainen 1. vedolla",

F = "punainen 2. vedolla".

Tapaus 1: Punaisten suhde $\theta = r/(r+k)$ tunnetaan. Tällöin $\mathsf{P}(F|E) = \theta = \mathsf{P}(F)$, joten E ja F ovat riippumattomia. Näin ajatellaan "klassisessa" tilastotieteessä.

Tapaus 2: Punaisten osuutta **ei** tunneta. Tällöin E ja F eivät ole riippumattomia, vaikka otannat ovat. Syy on se, että E:stä saadaan informaatiota punaisten osuudesta ja se vaikuttaa F:n ehdolliseen todennäköisyyteen. Bayes-tilastotieteessä sanotaan, että tapahtuvat ovat *vaihdannaisia*.

Vaikka θ **ei** ole tunnettu, voimme kirjoittaa siihen perustuvat ehdolliset todennäköisyydet (ajattelemalla, että θ on satunnaismuuttuja):

$$\mathsf{P}(F|E,\theta) = \mathsf{P}(F|\theta) = \theta.$$

Siis E ja F ovat *ehdollisesti riippumattomia ehdolla* θ.

Ehdollistaminen tuntemattomien parametrien suhteen (tässä tapauksessa θ:n suhteen) on oleellinen osa (bayesiläistä) tilastollista mallinnusta.

2.6 Bayesin kaava

Olkoon $H_1, H_2, \ldots$ joukon Ω *ositus*, ts. $H_i \cap H_j = \emptyset$, $i \neq j$, ja $\cup_n H_n = \Omega$. Silloin on voimassa (todista) *kokonaistodennäköisyyden* kaava

$$\mathsf{P}(E) = \sum_n \mathsf{P}(E|H_n)\,\mathsf{P}(H_n)\,.$$

Tästä kaavasta ja yhtälöstä

$$\mathsf{P}(H_n \cap E) = \mathsf{P}(H_n)\mathsf{P}(E|H_n)$$

seuraa *Bayesin kaava*

$$\mathsf{P}(H_n|E) = \frac{\mathsf{P}(H_n)\mathsf{P}(E|H_n)}{\sum_k \mathsf{P}(H_k)\mathsf{P}(E|H_k)}.$$

Bayesin kaava voidaan esittää myös suppeammassa muodossa

$$\mathsf{P}(H_n|E) \propto \mathsf{P}(H_n)\mathsf{P}(E|H_n),$$

josta on tiputettu pois suhteellisuuskerroin $1/\mathsf{P}(E)$.

Huomaa, että tarkasteltaessa vedonlyöntisuhteita suhteellisuuskerroin supistuu pois:

$$\underbrace{\frac{\mathsf{P}(H_j|E)}{\mathsf{P}(H_k|E)}}_{\text{posteriorivedonlyöntisuhde}} = \underbrace{\frac{\mathsf{P}(H_j)}{\mathsf{P}(H_k)}}_{\text{priorivedonlyöntisuhde}} \times \underbrace{\frac{\mathsf{P}(E|H_j)}{\mathsf{P}(E|H_k)}}_{\text{uskottavuussuhde}}.$$

Yleisemmässä tapauksessa, kun E edustaa havaintoaineistoa ja H_i:t eri malleja (tai hypoteesejä), posteriorivedonlyöntisuhteita voidaan käyttää mallien vertailussa (tai hypoteesien testauksessa).

Esim. 3 Oletetaan, että tytön syntymän todennäköisyys on $\frac{1}{2}$. Tarkastellaan todennäköisyyttä, että kaksoset ovat tyttöjä. Onko tämä todennäköisyys $\frac{1}{4}$? Vastaus on, että **ei ole**:

Kaksoset voivat olla *identtiset* (monotsygoottiset), merk. M, tai *ei-identtiset* (ditsygoottiset), merk. D. Identtiset kaksoset ovat aina samaa sukupuolta (P poika, T tyttö). Saamme ehdolliset todennäköisyydet

$$\begin{aligned} \mathsf{P}(TT|M) = \mathsf{P}(PP|M) &= \frac{1}{2}; \qquad \mathsf{P}(TP|M) = 0; \\ \mathsf{P}(TT|D) = \mathsf{P}(PP|D) &= \frac{1}{4}; \qquad \mathsf{P}(TP|D) = \frac{1}{2}. \end{aligned}$$

Lähtien näistä tuloksista saamme kokonaistodennäköisyyden kaavan avulla tyttökaksosten todennäköisyydeksi

$$\begin{aligned} \mathsf{P}(TT) &= \mathsf{P}(TT|M)\mathsf{P}(M) + \mathsf{P}(TT|D)\mathsf{P}(D) \\ &= \frac{1}{2}\mathsf{P}(M) + \frac{1}{4}[1 - \mathsf{P}(M)] \\ &= \frac{1}{4}[\mathsf{P}(M) + 1]. \end{aligned}$$

Tästä seuraa, että $\mathsf{P}(TT) > \frac{1}{4}$, koska $\mathsf{P}(M) > 0$.

Sivutuotteena saadaan yhtälö

$$\mathsf{P}(M) = 4\mathsf{P}(TT) - 1\,.$$

Tämän avulla voidaan estimoida identtisten kaksosten osuus populaatiossa, koska $\mathsf{P}(TT)$ voidaan estimoida väestörekisteristä.

Luku 3

Mallit

Tilastollisella mallilla tarkoitetaan yleensä havaintoaineistoa kuvaavaa todennäköisyysmallia, jonka parametrit ovat tuntemattomia. Puhtaasta matemaattisesta mallista tilastollien malli eroaa tyypillisesti siinä, että malliin sisältyvät suureet eivät riipu toisistaan täysin deterministisesti vaan mukana on satunnaisuutta.

Parametrit voivat olla skalaareja tai vektoreita. Jälkimäisessä tapauksessa voidaan joko ajatella niin, että mallissa on useita parametreja tai että sillä on yksi parametri, joka on vektori. Parametrien mahdollisesti saamien arvojen joukkoa kutsutaan *parametriavaruudeksi*, merk. Θ. Ns. epäparametrisissa tilastollisissa malleissa Θ on ääretönulotteinen.

Bayes-tilastotiede eroaa klassisesta tilastotieteestä siinä, että kaikkia tuntemattomia parametrejä käsitellään satunnaismuuttujina ja niihin liittyvää epävarmuutta mallinnetaan todenäköisyysjakaumilla. Bayesiläinen tilastollinen malli on siinä mielessä laajempi kuin klassinen, että se sisältää myös parametrien jakauman havaintojen jakauman lisäksi. Parametrien satunnaisuudesta seuraa se, että klassisen tilastotieteen oletus havaintojen *riippumattomuudesta* korvautuu niiden *vaihdannaisuudella*.

Esim. 1. Rahanheitto. Tarkastellaan rahanheittoa eikä tiedetä, onko raha symmetrinen. Olkoon E tapahtuma ”tulos on kruuna”, ja merkitään kruunan todennäköisyyttä (suhteellista frekvenssiä äärettömän pitkässä koesarjassa) parametrilla θ. Tällöin voidaan asettaa tilastollinen malli: heittojen tulokset ovat riippumattomia, kun θ tunnetaan, ja jokaisella heitolla kruunan todennäköisyys on θ.

Jos parametria θ käsitellään satunnaismuuttujana, malli voidaan esittää

ehdollisena todennäköisyytenä

$$\mathsf{P}(E|\theta) = \theta.$$

Havainnot (heittojen tulokset) ovat riippumattomia ehdolla θ. Sen sijaan ei-ehdolliset havainnot eivät ole riippumattomia vaan *vaihdannaisia*. Tässä on ero verrattuna klassiseen tilastotieteeseen, jossa havainnot oletetaan riippumattomiksi.

Tarkastellaan yksityiskohtaisemmin, miten epävarmuus kehittyy heittosarjassa. Olkoon E "ensimmäinen heitto on kruuna". Malli on

$$\mathsf{P}(E|\theta) = \theta.$$

Silloin

$$\mathsf{P}(E) = \int g(\theta)\mathsf{P}(E|\theta)\mathrm{d}\theta = \mathsf{E}(\theta)$$

missä $g(\theta)$ on θ:n todennäköisyysjakauma.

Olkoon F "toinen heitto on kruuna". Silloin $\mathsf{P}(F|E,\theta) = \mathsf{P}(F|\theta) = \theta$ ja $\mathsf{P}(E \cap F|\theta) = P(E|\theta)P(F|\theta) = \theta^2$. Jälkimmäisestä seuraa, että

$$\mathsf{P}(E \cap F) = \mathsf{E}[\mathsf{P}(E \cap F|\theta)] = \mathsf{E}(\theta^2).$$

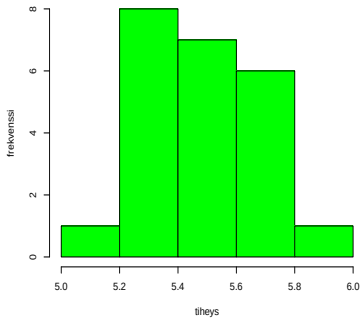
Nyt voimme osoittaa, että $\mathsf{P}(F|E) \geq P(F)$:

$$\begin{aligned}\mathsf{P}(F|E) =& \frac{\mathsf{P}(E \cap F)}{\mathsf{P}(E)} = \frac{\mathsf{E}(\theta^2)}{\mathsf{E}(\theta)} = \frac{\mathsf{Var}(\theta) + \mathsf{E}(\theta)^2}{\mathsf{E}(\theta)} = \frac{\mathsf{Var}(\theta)}{\mathsf{E}(\theta)} + \mathsf{E}(\theta) \\ =& \frac{\mathsf{Var}(\theta)}{\mathsf{E}(\theta)} + \mathsf{P}(F) \geq \mathsf{P}(F).\end{aligned}$$

Yhtäsuuruus on voimassa jos ja vain jos $\mathsf{Var}(\theta) = 0$. Siis E ja F ovat riippumattomia vain, jos θ on ei-satunnainen.

Esim. 2. Normaalimalli, Cavendishin aineisto. Englantilainen fyysikko Henry Cavendish (1731–1810) pyrki määrittämään maapallon tiheyden seuraavan aineiston perusteella:

5.36	5.29	5.58	5.65	5.57
5.53	5.62	5.29	5.44	5.34
5.79	5.10	5.27	5.39	5.42
5.47	5.63	5.34	5.46	5.3
5.78	5.68	5.85		



Malli kytkee parametrit ja havainnon informaation toisiinsa. Yksinkertainen malli on seuraava (olettaen mittaustarkkuuden tunnetuksi):

- havainnot $y_1, \ldots, y_n$ ovat riippumattomia ehdolla θ,
- $y_i|\theta \sim \mathrm{N}(\theta, 0.04)$.

Mallin oletuksista seuraa, että havaintojen $y_1, \ldots, y_n$ yhteisjakauma ehdolla θ on

$$p(y_1, \ldots, y_n|\theta) = \left(\frac{1}{2\pi \cdot 0.04}\right)^{\frac{n}{2}} e^{-\frac{1}{2 \cdot 0.04}\sum_{i=1}^{n}(y_i-\theta)^2}.$$

Kun jakaumaa $p(y_1, \ldots, y_n|\theta)$ ajatellaan θ:n funktiona kiinteillä arvoilla $y_1, \ldots, y_n$, sitä kutsutaan *uskottavuusfunktioksi*.

Huomaa, että malliin sisältyy subjektiivisia oletuksia. Uskot havaintojen ehdolliseen riippumattomuuteen, normaalisuuteen ja esitettyyn varianssiin. Mallia voidaan yleistää niin, että uskotaan ainoastaan ehdolliseen riippumattomuuteen (ehdolla mallin parametrit) ja havaintojen normaalisuuteen, jolloin

$$y_i|\theta, \sigma^2 \sim \mathrm{N}(\theta, \sigma^2)$$

ja havainnot ovat riippumattomia ehdolla (θ, σ^2).

Aiemmin määriteltiin, että havainnot $y_1, \ldots, y_n$ ovat *vaihdannaisia* (exchangeable), jos niiden yhteisjakauma $p(y_1, \ldots, y_n)$ pysyy samana, vaikka havaintojen järjestystä permutoidaan. Esimerkissä ehdollinen riippumattomuus ja havaintojen sama jakauma takaavat vaihdannaisuuden. Kuitenkaan havainnot $y_1, \ldots, y_n$ eivät ole riippumattomia reunajakauman $p(y_1, \ldots, y_n)$ suhteen.

Luku 4

Priori, posteriori ja ennustejakaumat

Bayes-tilastotieteessä johtopäätökset perustuvat posteriorijakaumaan

$$p(\theta \,|\, \text{aineisto}) \,.$$

Tarkastellaan ensin, miten posteriorijakaumaa käytetään tilastollisessa päättelyssä. Sen jälkeen tutkitaan, miten posteriorijakauma muodostetaan.

4.1 Posteriori päättelyn välineenä

Oletetaan, että y on havaittu aineisto, josta ei tässä vaiheessa tehdä rajoittavia oletuksia. Olkoon θ tuntematon suure. Se on usein parametri(vektori), mutta se voi myös olla puuttuva havainto(vektori) tai vaikkapa latentti muuttuja. Bayes-tilastotieteessä ei näiden eri tyyppisten tuntemattomien suureiden välillä ole periaatteellista eroa.

Oletetaan lisäksi, että käytössä on θ:n *posteriorijakauma* $p(\theta|y)$. Se on tuntemattoman suureen θ ehdollinen todennäköisyysjakauma ehdolla havainto y. Posteriorijakaumaa voidaan käyttää päättelyn välineenä seuraavilla tavoilla:

(a) **Kuvaaja**. Voidaan piirtää posteriorijakauman kuvaaja, kun $\dim(\theta) \leq 2$. Jos $\theta = (\theta_1, \ldots, \theta_k)$, niin usein käytetään reunaposteriorijakaumien

$$p(\theta_i|y) = \int \cdots \int p(\theta|y) \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{i-1} \mathrm{d}\theta_{i+1} \cdots \mathrm{d}\theta_k$$

kuvaajia, $i = 1, \ldots, k$. Nämä esittävät osittaista informaatiota posteriorista, mutta eivät ilmaise parametrien välistä (posteriorista) riippuvuutta. Voidaan myös kuvata kahden parametrin θ_i ja θ_j kaksiulotteista reunaposterioria.

(b) **Tunnusluvut**. Graafisen esityksen sijasta posteriorijakaumaa voidaan kuvata tunnuslukujen avulla. Usein käytetään (yksiulotteiselle θ:lle) kuvausta

odotusarvo	hajonta	10 % raja	mediaani	90 % raja
$\mathsf{E}(\theta\|y)$	$\sqrt{Var(\theta\|y)}$	$\theta_{0.1}^{post}$	$\theta_{0.5}^{post}$	$\theta_{0.9}^{post}$

missä θ_α^{post} on posteriorijakauman α-kvantiili: $\mathsf{P}(\theta \leq \theta_\alpha^{post}|y) = \alpha$.

(c) **Väliestimointi,** (kun $\dim(\theta) = 1$). Määritetään sellaiset reaaliluvut a ja b, että

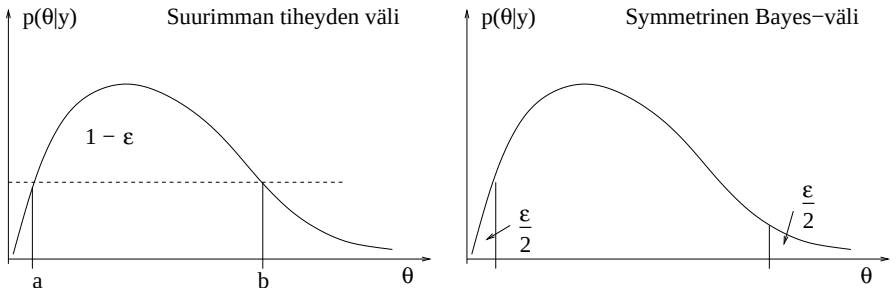
$$P(a \leq \theta \leq b|y) = 1 - \epsilon.$$

Tätä kutsutaan *posteriorivälikisi (posterior interval)* ja toisinaan *bayesiläiseksi luottamusväliksi (credible interval).*

Ratkaisu ei ole yksikäsitteinen. Tavallisimmat muodostamisperiaatteet ovat

- **lyhin väli;**
- **symmetrinen väli**, $P(\theta < a|y) = \frac{\epsilon}{2}$ ja $P(\theta > b|y) = \frac{\epsilon}{2}$;
- **suurimman tiheyden väli** (HDI–highest density interval), moniulotteisessa tapauksessa HDR (region): jakauman tiheysfunktio on välin sisällä suurempi kuin missään välin ulkopuolisessa pisteessä. Tämä on tyypillinen bayesiläinen vaihtoehto!

  Huomaa, että HDI-väliä ei välttämättä ole olemassa yksiulotteisessa tapauksessa vaan HDR voi koostua useammasta osavälistä, jos posteriorijakauma ei ole yksihuippuinen. Jos HDI-väli on olemassa, se on samalla lyhin väli.



(d) **Hypoteesien testaus.** Tarkastellaan hypoteesia H : $\theta > 0$. Hypoteesin posterioritodennäköisyys on

$$\begin{aligned} \mathsf{P}(\mathrm{H}|y) &= \mathsf{P}(\theta > 0|y) = \int_0^\infty p(\theta|y)\mathrm{d}\theta \\ &= \text{ tn, että hypoteesi on tosi ehdolla aineisto } y. \end{aligned}$$

Huomaa, että tämä on todellakin hypoteesin todennäköisyys. Sen sijaan klassinen p-arvo **ei ole** nollahypoteesin todennäköisyys!

(e) **Piste-estimointi**. Joskus halutaan tiivistää posteriorijakauman informaatio yhteen arvoon, θ:n Bayes-estimaattiin. Vaihtoehtoja ovat

$\tilde{\theta} = \operatorname{argmax} p(\theta|x)$ posteriorimoodi;

$\hat{\theta}$, posteriorimediaani, jakauman $p(\theta|y)$ mediaani;

$\bar{\theta} = \mathsf{E}(\theta|y)$ posterioriodotusarvo.

Näistä ensimmäinen on analoginen suurimman uskottavuuden estimaattorin kanssa. Sitä sanotaan MAP (Maximum A Posterior estimator) -estimaattoriksi.

(f) **Kiinnostavat todennäköisyydet.** Posteriorista voidaan laskea todennäköisyyksiä

$$\mathsf{P}(\theta \in A \,|\, y),$$

missä A on tutkijan määräämä tapahtuma, usein hypoteesi.

Todennäköisyyksien laskeminen on tyypillistä Bayes-menetelmälle. Yleinen sääntö on se, että posteriorijakaumalle voidaan tehdä kaikkea sitä, mikä tn-jakaumalle on sallittua.

4.2 Priorijakauma

Jotta voisimme laskea posteriorijakauman, tarvitsemme 1) mallin havainnoille ja 2) priorijakauman tuntemattomalle parametrille. Bayesin kaavassa mallia havainnoille edustaa *otantajakauma*[1] $p(y|\theta)$. Funktiota $p(y|\theta)$ kutsutaan *uskottavuusfunktioksi*, kun sitä tarkastellaan θ:n funktiona. (Tällöin käytetään myös merkintää $L(\theta; y)$.) *Priorijakauma* $p(\theta)$ on θ:n ei-ehdollinen jakauma, ja sen avulla voidaan kuvata (subjektiivista) ennakkokäsitystä parametrin arvosta.

Nämä kaksi asiaa yhdistetään *Bayesin kaavan*[2] avulla:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)\mathrm{d}\theta}.$$

Siis posteriori voidaan laskea normeeraamalla priorijakaman tiheysfunktion ja uskottavuusfunktion tulo jakaumaksi.

"Klassinen" tilastotiede perustuu paljolti uskottavuuteen $L(\theta; y) = p(y|\theta)$, kun taas Bayes-tilastotiede posteriorijakaumaan $p(\theta|y)$. Jälkimmäinen on todennäköisyysjakauma mitä uskottavuus taas **ei** ole.

Posteriori on normeerausta vaille priori×uskottavuus

$$p(\theta|y) \propto p(\theta)\, p(y|\theta),$$

missä merkintä $\propto$ tarkoittaa: 'suoraan verrannollinen'.

Jos priori ja uskottavuusfunktio tunnetaan, niin posteriori tunnetaan normeerausta vaille. Laskennalliset ongelmat liittyvät posteriorin normeerukseen, ts. integraalin $\int p(\theta)\, p(y|\theta)\, \mathrm{d}\theta$ laskemiseen. Toisaalta posteriorisuhde

$$\frac{p(\theta_0|y)}{p(\theta_1|y)} = \frac{p(\theta_0)}{p(\theta_1)} \times \frac{p(y|\theta_0)}{p(y|\theta_1)}\,.$$

voidaan määrittää ilman kyseisen integraalin laskemista.

Huomaa, että Bayes-analyysi perustuu kahteen informaatiolähteeseen koskien tuntematonta suuretta θ. Nämä ovat: 1) ennakkokäsitys $p(\theta)$:n kautta ja 2) aineisto $p(y|\theta)$:n kautta.

[1] Tarkemmin: otantajakauman tiheysfunktio tai todennäköisyysfunktio.

[2] Tämä versio Bayesin kaavasta soveltuu jatkuva-arvoiselle parametrille θ. Kappaleessa 2.6 esitettiin diskreetille priori-informaatiolle soveltuva kaava.

Esim. 1. Normaalimalli. Olkoot havainnot $y = (y_1, \ldots, y_n)$ riippumattomia ehdolla θ ja normaalijakautuneita $y_i|\theta \sim \mathrm{N}(\theta, v)$, missä varianssi v tunnettu.

Oletetaan vielä, että priori $p(\theta)$ on normaalinen, ts.

$$\theta \sim \mathrm{N}(m_0, w_0)\,,$$

missä m_0 ja w_0 ovat tunnettuja.

Havaintoihin perustuva *uskottavuusfunktio* on

$$\begin{aligned} p(y|\theta) &= p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \\ &\propto \exp\left\{-\frac{1}{2v}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \end{aligned}$$

ehdollisen riippumattomuuden perusteella.

Edelleen *priori* on

$$p(\theta) = \frac{1}{\sqrt{2\pi w_0}} e^{-\frac{1}{2w_0}(\theta-m_0)^2} \propto e^{-\frac{1}{2w_0}(\theta-m_0)^2}\,.$$

Posteriori voidaan kirjoittaa muodossa

$$p(\theta|y) \propto e^{-\frac{1}{2w_0}(\theta-m_0)^2} \times e^{-\frac{1}{2v}\sum_{i=1}^{n}(y_i-\theta)^2} = e^{-\frac{1}{2}Q}\,,$$

missä

$$\begin{aligned} Q &= \left(\frac{1}{w_0} + \frac{n}{v}\right)(\theta - m_1)^2 + \text{ vakio}, \\ m_1 &= \left(\frac{m_0}{w_0} + \frac{n\bar{y}}{v}\right) / \left(\frac{1}{w_0} + \frac{n}{v}\right). \end{aligned}$$

Laskeminen perustuu neliöön täydentämiseen.

Posteriorijakauman tiheysfunktion muodosta näemme, että se on normaalijakauma $\theta|y \sim \mathrm{N}(m_1, w_1)$:

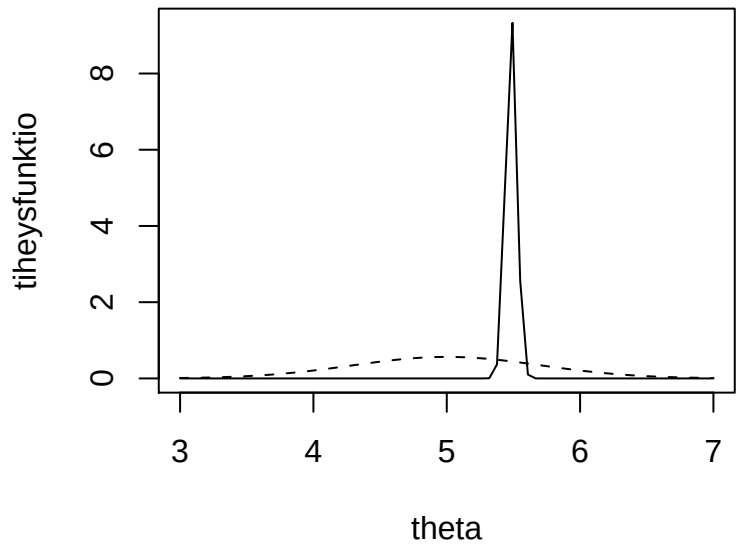
$$\begin{aligned} p(\theta|y) &\propto \exp\left\{-\frac{1}{2w_1}(\theta - m_1)^2\right\}, \\ m_1 &= \left(\frac{1}{w_0} + \frac{n}{v}\right)^{-1}\left(\frac{m_0}{w_0} + \frac{n\bar{y}}{v}\right), \\ w_1 &= \left(\frac{1}{w_0} + \frac{n}{v}\right)^{-1}. \end{aligned}$$

Jos priori ja posteriori kuuluvat samaan jakaumaperheeseen, niin puhutaan *konjugaattiperheistä*. Normaalipriori keskiarvoparametrille (kun varianssi on tunnettu) on normaalimallille ns. konjugaattipriori.

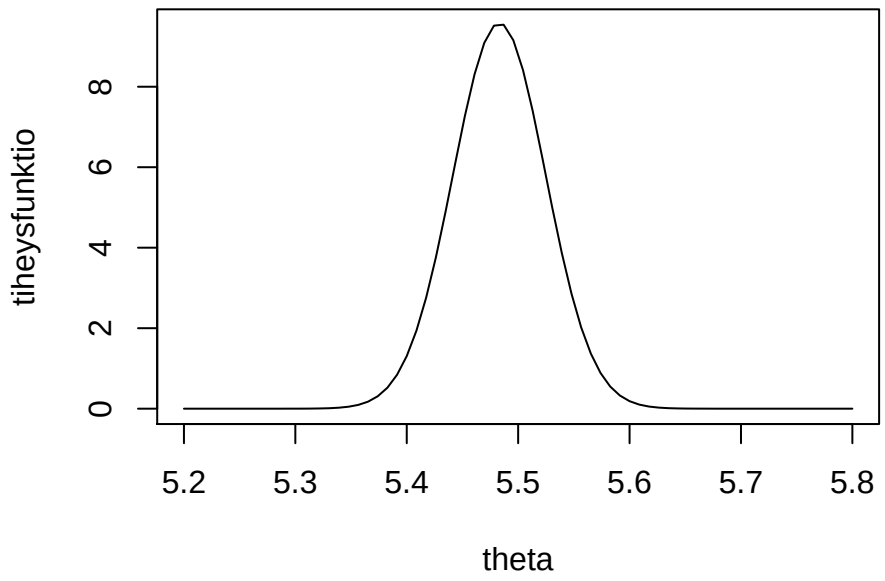
Esim. 2. Cavendishin aineisto (jatk.)

- *Priori*: varmasti $\theta > 1$ (veden tiheys = 1), ennakkoestimaatti $\theta = 5$. Olkoon $\theta \sim \mathrm{N}(5, 0.5)$.
- *Uskottavuus*: Normaali, varianssioletus 0.04 (vakio, mittaustarkkuus).
- *Posteriori*: $\theta|y \sim \mathrm{N}(m_1, w_1)$, missä
  $m_1 = \left(\frac{5}{0.5} + \frac{23 \cdot 5.485}{0.04}\right) / \left(\frac{1}{0.5} + \frac{23}{0.04}\right) = 5.483,$
  $w_1 = \left(\frac{1}{0.5} + \frac{23}{0.04}\right)^{-1} = 0.00173.$

Kuvio: Priori (lattea katkoviiva), posteriori (yhtenäinen viiva). Priori kuvaa θ:aan liittyvää epävarmuutta ennen mittauksia. Aineistosta opitaan ja posteriori kuvaa tätä epävarmuutta sen jälkeen, kun havainnot ovat käytössä.



Posteriorijakauma on näyttää siis seuraavalta:



Jakaumasta voidaan laskea 90%:n todennäköisyysväli maapallon tiheydelle: se on $(5.415, 5.552)$.

Kerrataan vielä tehdyt *oletukset*, jotka ovat johtaneet tähän tulokseen:

1. Havainnot normaaliset, varianssi tunnettu vakio.
2. Havainnot ehdollisesti riippumattomat ehdolla θ (vaihdannaisia).
3. Priori on normaali, melko epäinformatiivinen.

Edellä tarkasteltu posteriorin johto nojautuu vahvasti *konjugaattisuuteen*. Kun siitä luovutaan, asiat mutkistuvat. Posteriorijakaumaa ei tällöin voida välttämättä esittää suljetussa muodossa. Moderni tapa laskea posteriori yleisesti on käyttää simulointia, erityisesti Markov Chain Monte Carlo (MCMC) -menetelmää. Tähän palataan Bayes-tilastotiede 2 -kurssilla.

4.3 Bayesin kaavan peräkkäiskäyttö

Oletetaan että priori on $p(\theta)$ ja uskottavuus $p(y|\theta)$. Olkoon havainto y_1 ja siihen perustuva posteriori $p(\theta|y_1)$.

Tehdään uusi havainto y_2, joka on riippumaton y_1:stä ehdolla θ. Molempiin havaintoihin perustuva posteriori on

$$\begin{aligned} p(\theta|y_1, y_2) &\propto p(\theta)\, p(y_1, y_2|\theta) \\ &= p(\theta)\, p(y_1|\theta)\, p(y_2|\theta) \\ &\propto p(\theta|y_1)\, p(y_2|\theta). \end{aligned}$$

Voidaan siis ajatella, että posteriori $p(\theta|y_1, y_2)$ saadaan kahdessa vaiheessa:

1. Havainto y_1 opettaa prioria $p(\theta)$, tuloksena on posteriori $p(\theta|y_1)$.
2. y_2 opettaa prioria $p(\theta|y_1)$, tuloksena posteriori $p(\theta|y_1, y_2)$.

Edellinen toimii myös käänteisessä järjestyksessä niin, että $p(\theta|y_2)$ opetetetaan havainnolla y_1 (riippumatta siitä, missä järjestyksessä havainnot on tehty).

4.4 Havainnon prioriennustejakauma

Priori- ja posteriorijakaumat ovat tn-jakaumia parametriavaruudessa Θ. Sen sijaan otantajakauma $p(y|\theta)$ on todennäköisyysjakauma *havaintoavaruudessa*, merk. $\mathcal{Y}$. Bayes-tilastotieteessä tarkastellaan kahta muuta avaruudelle $\mathcal{Y}$ määriteltyä jakaumaa: priori- ja posterioriennustejakaumaa. Tässä "havainto" y voi olla skalaari, vektori tai yleisemmin havaintoaineisto.

Havainnon y reunajakaumaa

$$p(y) = \int p(y, \theta)\mathrm{d}\theta = \int p(y|\theta)p(\theta)\mathrm{d}\theta.$$

kutsutaan *prioriennustejakaumaksi*.

Tämä on siis eri asia kuin havainnon jakauma "klassisessa" tilastotieteessä, jossa havainnon jakaumalla tarkoitetaan ehdollista jakaumaa $p(y|\theta)$. Havainnon prioriennustejakauma ei käytä aineistoa lainkaan: se sisältää vaihtelun, joka tulee prioritiedosta. Se on "ennuste" tulevalle havainnolle, kun uskottavuus ja priori on valittu (mutta havaintoa ei ole tehty). Jakaumaa $p(y)$ voidaan käyttää mm. tutkittaessa priorijakauman järkevyyttä.

Esim. 3. (Cavendishin aineisto, jatk.) Olkoon $\theta \sim \mathrm{N}(5, 0.5)$ ja $\tilde{y}|\theta \sim \mathrm{N}(\theta, 0.04)$ (yksi havainto mallista). Määritettävä $\tilde{y}$:n jakauma $p(\tilde{y})$.

Oikotie, jolla vältetään integrointi: Kirjoitetaan

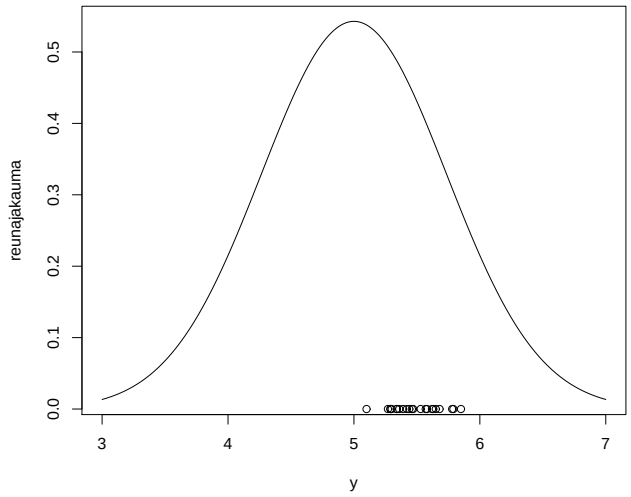
$$\tilde{y} = (\tilde{y} - \theta) + \theta\,.$$

Nyt $\tilde{y} - \theta|\theta \sim \mathrm{N}(0, 0.04)$ ja $\theta \sim \mathrm{N}(5, 0.5)$. Edelleen $\tilde{y} - \theta$ ja θ ovat riippumattomia. [3] Tällöin $\tilde{y}$ on kahden normaalijakautuneen satunnaismuuttujan summana normaalijakautunut, ja

$$\begin{aligned}\mathsf{E}(\tilde{y}) &= \mathsf{E}(\tilde{y} - \theta) + \mathsf{E}(\theta) = 0 + 5 = 5,\\ \mathsf{Var}(\tilde{y}) &= \mathsf{Var}(\tilde{y} - \theta) + \mathsf{Var}(\theta) = 0.04 + 0.5 = 0.54.\end{aligned}$$

Siis $\tilde{y} \sim \mathrm{N}(5, 0.54)$.

Kuviossa on esitetty prioriennustejakauma ja havaintoaineisto pienin palloin.



[3] Merkintä $\tilde{y} - \theta|\theta$ tarkoittaa erotuksen $\tilde{y} - \theta$ ehdollista jakaumaa ehdolla θ. Tämä jakauma ei riipu θ:sta, ts. $p(\tilde{z}|\theta) = p(\tilde{z})$, missä $\tilde{z} = \tilde{y} - \theta$. Siis $\tilde{y} - \theta$ ja θ ovat riippumattomia.

Reunajakauman odotusarvo ja varianssi voidaan myös laskea "ehdollistamistekniikalla" (laske), joka perustuu yhtälöihin

$$\begin{aligned} \mathsf{E}(y) &= \mathsf{E}_\theta \mathsf{E}(y|\theta), \\ \mathsf{Var}(y) &= \mathsf{E}_\theta \mathsf{Var}(y|\theta) + \mathsf{Var}_\theta \mathsf{E}(y|\theta). \end{aligned}$$

4.5 Posterioriennustejakauma

Oletetaan, että aineistoon y perustuva posteriori on $p(\theta|y)$. Se summaa θ:aa koskevan epävarmuuden havainnon y tekemisen jälkeen.

Tehdään uusi havainto $\tilde{y}$, joka on riippumaton y:stä ehdolla θ. Tällöin *posterioriennustejakauma* $\tilde{y}$:lle on $\tilde{y}$:n ehdollinen jakauma ehdolla y, merk. $p(\tilde{y}|y)$. Se on sekoitusjakauma jakaumasta $p(\tilde{y}|\theta)$, missä θ:n painoina on posteriorijakauman $p(\theta|y)$ arvot:

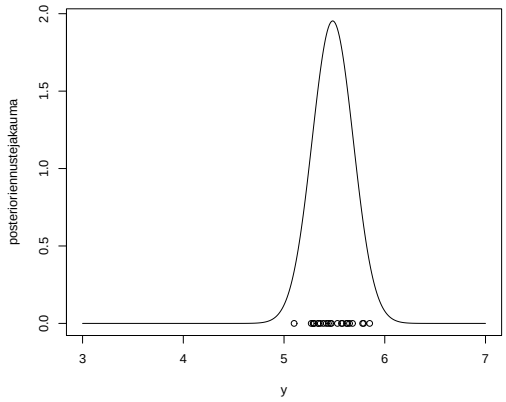
$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta, y) p(\theta|y) \mathrm{d}\theta \\ &= \int p(\tilde{y}|\theta) p(\theta|y) \mathrm{d}\theta. \end{aligned}$$

Esim. 4. Cavendishin aineisto, jatk. Olkoon edelleen $\theta \sim \mathrm{N}(5, 0.5)$, $y|\theta \sim \mathrm{N}(\theta, 0.04)$. Määritettävä jakauma $p(\tilde{y}|y)$. Käytetään samaa trikkiä kuin edellisessä esimerkissä:

$$\tilde{y} = (\tilde{y} - \theta) + \theta \,.$$

Nyt $\tilde{y} - \theta|\theta \sim \mathrm{N}(0, 0.04)$ ja $\theta|y \sim \mathrm{N}(5.483, 0.0017)$. Edelleen $\tilde{y} - \theta$ ja θ ovat riippumattomia.

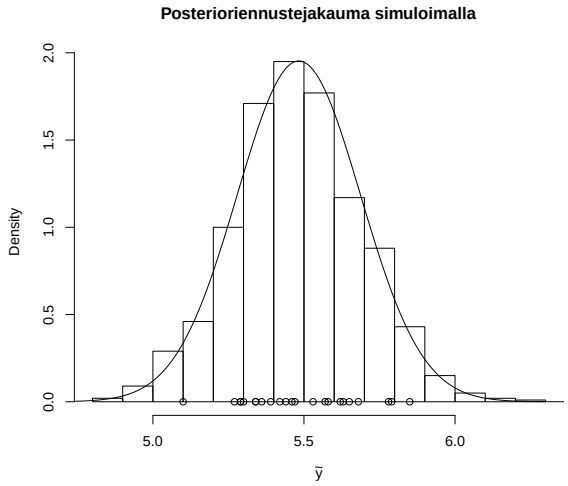
Tästä seuraa, että $\tilde{y}|y \sim \mathrm{N}(5.483, 0.0417)$. Kuvioon on myös piirretty x -akselille havaittu aineisto.



Esim. 5. Simulointi (Cavendish jatk.) Posterioriennustejakauma on myös helppo määrittää simuloimalla, jos pystytään simuloimaan posteriorijakaumaa. Ensin simuloidaan θ (theta) posteriorijakumasta $p(\theta|y)$ ja sitten $\tilde{y}$ (yenn) jakaumasta $p(\tilde{y}|\theta)$ käyttäen simuloitua arvoa θ.

```
n.sims <- 1000
theta <- rnorm(n.sims,5.483,sqrt(0.0017))
yenn <- rnorm(n.sims,theta,0.2)
```

Kuviossa on simulaatioiden histogrammin lisäksi teoreettinen posteriorijakauman käyrä ja alkuperäiset havainnot.



Luku 5

Yksiparametrisia malleja

Seuraavaksi käsittelemme yleisimpiä ja hyödyllisimpiä yksiparametrisia malleja. Näitä ovat normaalijakauma (joko odotusarvo tai varianssi tuntematon), binomijakauma, Poisson-jakauma ja eksponenttijakauma.

5.1 Normaalinen otos, odotusarvo tuntematon

Normaalisen otoksen posteriorijakauma, kun odotusarvo on tuntematon, johdettiin kappaleessa 4.2, esimerkissä 1. Oletimme, että meillä on otos $y = (y_1, ..., y_n)$ ja

- $y_i|\theta \sim \mathrm{N}(\theta, v)$, riippumattomia ehdolla θ,
- varianssi v tunnettu,
- $\theta \sim \mathrm{N}(m_0, w_0)$ (priori).

Tällöin posteriori on $\theta|y \sim \mathrm{N}(m_1, w_1)$, missä

$$\begin{aligned} m_1 &= \frac{\frac{m_0}{w_0} + \frac{n\bar{y}}{v}}{\frac{1}{w_0} + \frac{n}{v}}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \\ w_1 &= \left(\frac{1}{w_0} + \frac{n}{v}\right)^{-1}. \end{aligned}$$

Priorijakauma ja posteriorijakauma ovat molemmat normaalisia. Tämän vuoksi sanomme, että normaalinen priorijakauma on *konjugaattinen* normaalille otantajakaumalle.

Siirryttäessä priorista posterioriin *havainto päivittää normaalijakauman keskiarvoparametria ja varianssia*: Posterioriodotusarvo m_1 voidaan kirjoittaa muodossa

$$m_1 = \frac{w_1}{w_0} \cdot m_0 + \frac{w_1}{v/n} \cdot \bar{y}\,,$$

joten m_1 on prioriodotusarvon m_0 ja otoskeskiarvon $\bar{y}$ *painotettu keskiarvo*. Painot kuvaavat prioritiedon ja havainnon suhteellista voimakkuutta. Tämä näkyy siitä, että painojen suhde on

$$\frac{\frac{w_1}{v/n}}{\frac{w_1}{w_0}} = \frac{w_0}{v/n},$$

joka on priorivarianssin w_0 ja otoskeskiarvon varianssin v/n suhde.

Posterioriodotusarvo on otoskeskiarvon *kutistus* (shrinkage) kohti prioriodotusarvoa. Prioritieto muuttuu epäoleelliseksi, kun otoskoko lähestyy ääretöntä, sillä

$$m_1 \to \bar{y}, \text{ kun } n \to \infty\,.$$

Lisäksi, kun prioritieto on heikkoa ja havainnon informaatio suuri (otoskoko n suuri), niin likimain

$$\theta | y_1, \ldots, y_n \sim \mathrm{N}(\bar{y}, \frac{v}{n})\,.$$

Tästä seuraa esimerkiksi se, että 95%:n HDI on likimäärin

$$\bar{y} \pm 1.96\sqrt{\frac{v}{n}}\,.$$

Numeerisesti 95 %:n HDI on tässä tapauksessa sama kuin "klassinen" 95 %:n luottamusväli – mutta vain numeerisesti, sillä tulkinta on eri!

Bayes-tilastotieteessä usein käytetään *heikkoa prioria* ilmaisemaan tietämättömyyttä. Esimerkkejä heikosta priorista ovat tasajakauma ja normaalijakauma suurella varianssilla:

$$\begin{aligned} p(\theta) &\propto \kappa \quad \text{(vakio)}, \\ \theta &\sim \mathrm{N}(0, \sigma^2),\ \sigma^2 \text{ suuri}. \end{aligned}$$

Näistä prioreista ensimmäinen, tasajakauma äärettömän pitkän välin yli, on ns. *epäaito* priori (ei integroidu). Kuitenkin posteriori *voi olla* aito (kuten tässä tapauksessa):

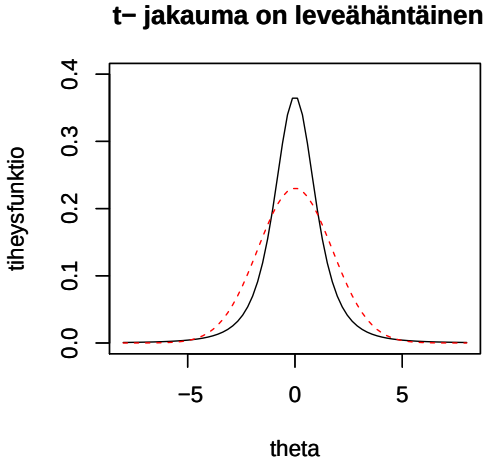
$$\begin{aligned} p(\theta|y) &= \frac{\kappa p(y|\theta)}{\int \kappa p(y|\theta)\mathrm{d}\theta} \\ &\propto \exp\left\{-\frac{1}{2v}\sum_{i=1}^{n}(y_i-\theta)^2\right\}. \end{aligned}$$

Tämä on aito priori, koska integraali

$$\int \exp\left\{-\frac{1}{2v}\sum_{i=1}^{n}(y_i-\theta)^2\right\}\mathrm{d}\theta$$

suppenee.

Voidaan myös varautua siihen, että priori ei saa olla liian "rajoittava". Tällöin priori on tietyssä mielessä *robusti.* Esimerkiksi voidaan ajatella seuraavaa ennakkotietoa: *Havainnot ovat keskittyneet jonkin kiinnitetyn parametriarvon θ_0 ympärille, mutta suuretkin poikkeamat tästä arvosta voivat tulla kysymykseen.* Tällöin voidaan käyttää paksuhäntäistä prioria kuten t-jakaumaa; esim. $\theta - \theta_0 \sim \mathrm{t}(\nu)$ jollakin vapausasteiden määrällä ν.



5.2 Binomiotos

Havainto y_i on dikotominen, kun se voi saada vain kaksi arvoa, "onnistuminen" ja "epäonnistuminen". Kun onnistumisen todennäköisyys pysyy koko ajan samana toistokokeissa, kyseessä on ns. *Bernoullin koe*. Epäonnistuminen koodataan yleensä luvulla 0 ja onnistuminen luvulla 1. Se, mitä "onnistumisella" tarkoitetaan, riippuu tilanteesta (kruuna, voitto, kuolema, ...).

Yhtä Bernoullin koetta vastaava uskottavuusfunktio on

$$p(y_i|\theta) = \theta^{y_i}\,(1-\theta)^{1-y_i},\; y_i = 0, 1,$$

missä θ on onnistumisen (arvon 1) todennäköisyys. Kun tehdään n toistokoetta, jotka ovat riippumattomia ehdolla θ, uskottavuusfunktio on

$$p(y|\theta) = \prod_{i=1}^{n} \left\{\theta^{y_i}\,(1-\theta)^{1-y_i}\right\} = \theta^s\,(1-\theta)^{n-s},$$

missä $s = \sum_{i=1}^{n} y_i$. Onnistumisten lukumäärä ehdolla θ on binomijakautunut: $s|\theta \sim \text{Bin}(n, \theta)$, $\mathsf{E}(s|\theta) = n\theta$, $\mathsf{Var}(s|\theta) = n\theta(1-\theta)$.

Priorin konstruointi. Tuntematon suure θ voi saada arvoja väliltä $[0, 1]$. Siksi priori on määriteltävä tälle välille. Joustava vaihtoehto on beetajakauma $\text{Beta}(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$, jonka tiheysfunktio on

$$\begin{aligned} p(\theta|\alpha,\beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}, \quad 0 \le \theta \le 1, \\ &\propto \theta^{\alpha-1}\,(1-\theta)^{\beta-1}, \end{aligned}$$

missä $\Gamma(\nu) = \int_0^\infty t^{\nu-1}e^{-t}dt$ on ns. gammafunktio.

Seuraavassa on joitain gamma-funktion ominaisuuksia:

- $\Gamma(\nu) = (\nu-1)\Gamma(\nu-1)$ (rekursiokaava)
- $\Gamma(n) = (n-1)!$, $n = 1, 2, \ldots,$ $\quad \Gamma(1) = 0, \quad \lim_{x\to 0^+}\Gamma(x) = \infty$
- $\Gamma(\nu) \approx \sqrt{2\pi}\,e^{\nu}\;\nu^{\nu-\frac{1}{2}}$ (Stirlingin kaava)

Posteriori. Käytettäessä beetajakaumaa priorina, $\theta \sim \text{Beta}(\alpha, \beta)$, posteriorijakaumaksi saadaan

$$\begin{aligned} p(\theta|y) &\propto \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\cdot\theta^s\,(1-\theta)^{n-s} \\ &\propto \theta^{\alpha+s-1}\,(1-\theta)^{\beta+n-s-1}, \end{aligned}$$

joka on Beta($\alpha + s, \beta + n - s$) -jakauma. (Huomaa, että tyyppiä $p(\theta) \propto \theta^a (1-\theta)^b$, $\theta \in [0,1]$, oleva jakauma on aina beetajakauma, jonka parametrit ovat $\alpha = a + 1$ ja $\beta = b + 1$.)

Beetajakauma on *konjugaattipriori* binomiotokselle, koska myös posteriori on beetajakauma. Otos päivittää jakauman parametrit seuraavasti:

priori	posteriori
α	$\alpha + s$
β	$\beta + n - s$

Beetajakauman ominaisuuksia. Laskemalla voidaan osoittaa (ei tarvitse integroida!), että

$$\begin{aligned} \mathsf{E}(\theta|\alpha,\beta) &= \frac{\alpha}{\alpha+\beta}, \\ \mathsf{Var}(\theta|\alpha,\beta) &= \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}, \\ \text{moodi} &= \frac{\alpha-1}{\alpha+\beta-2}\,. \end{aligned}$$

Edellä olevat pätevät jakaumalle Beta(α, β). Tästä saadaan suoralla sijoituksilla $\alpha \to \alpha + s$ ja $\beta \to \beta + n - s$ posteriorijakauman tunnusluvut, esim.

$$\begin{aligned} \mathsf{E}(\theta|y) &= \frac{\alpha+s}{\alpha+\beta+n}, \\ \text{moodi} &= \frac{\alpha+s-1}{\alpha+\beta+n-2}\,. \end{aligned}$$

Beetajakauman odotusarvo voidaan laskea täydentämällä integrandi (integroitava lauseke) tiheysfunktioksi seuraavalla tavalla:

$$\begin{aligned} & \mathsf{E}(\theta|\alpha,\beta) \\ = & \int_0^1 \theta\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,\mathrm{d}\theta \\ = & \frac{\Gamma(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)}\int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}\theta^{(\alpha+1)-1}(1-\theta)^{\beta-1}\mathrm{d}\theta \\ = & \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \\ = & \frac{\alpha}{\alpha+\beta}\,, \end{aligned}$$

koska 3. rivin integraali on beetajakauman tiheysfunktion integraali yli parametriavaruuden ja se on 1. Esitetty "laskutemppu" on hyödyllinen monissa muissakin vastaavissa tilanteissa ja kannattaa osata.

Vastaavasti voidaan laskea varianssi laskemalla ensin $\mathsf{E}(\theta^2|y)$ ja siitä

$$\mathsf{Var}(\theta|y) = \mathsf{E}(\theta^2|y) - \mathsf{E}(\theta|y)^2 \,.$$

Suurten otosten tapauksessa päädytään samanlaiseen tulokseen kuin normaalisen otoksen tapauksessa. Posterioriodotusarvo on likimain otoskeskiarvo. Nimittäin kun $n \approx \infty$,

$$\mathsf{E}(\theta|y) = \frac{\frac{\alpha}{n} + \frac{s}{n}}{\frac{\alpha}{n} + \frac{\beta}{n} + 1} \approx \frac{s}{n} (= \bar{y}),$$

Päädytään siis "klassiseen" suurimman uskottavuuden estimaattoriin.

Myös keskeistä raja-arvolausetta muistuttava tulos pätee:

$$\left(\frac{\theta - \mathsf{E}(\theta|y)}{\sqrt{\mathsf{Var}(\theta|y)}} \,\middle|\, y \right) \to \mathrm{N}(0,1) \,.$$

Tähän perustuu esim. 95 %:n HDI:n approksimaatio

$$\mathsf{E}(\theta|y) \pm 1.96 \sqrt{\mathsf{Var}(\theta|y)} \,.$$

(Kuitenkin normaalijakauma-approksimaatio on käytännössä tarkempi logit-muunnokselle (vedonlyöntisuhteen logaritmille) $\phi = \log(\theta/(1-\theta))$. Jos $\theta \sim \mathrm{Beta}(\alpha, \beta)$, niin

$$\mathsf{E}(\phi) \approx \log\left(\frac{\alpha - \frac{1}{2}}{\beta - \frac{1}{2}} \right),$$
$$\mathsf{Var}(\phi) \approx \frac{1}{\alpha} + \frac{1}{\beta}.$$

)

Esim. 1. Olkoon tuntematon suure θ kalifornialaisten osuus, jotka kannattavat kuolemanrangaistusta. Oletetaan, että priori on $\theta \sim$ Beta, jolle $\mathsf{E}(\theta) = 0.6$, $\mathsf{sd} = 0.3$. (Usein priori annetaan juuri tässä muodossa – on siis ratkaistava α ja β).

(i) Beetajakauman parametrit ratkaistaan yhtälöparista

$$\begin{aligned} \frac{\alpha}{\alpha+\beta} &= 0.6, \\ \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} &= 0.3^2 \ , \end{aligned}$$

ja saadaan $\alpha = 1$, $\beta = \frac{2}{3}$. Priori on siis

$$p(\theta) \propto (1-\theta)^{-\frac{1}{3}}, \quad 0 \leq \theta \leq 1 \, .$$

(ii) Tehdään kokoa $n = 1000$ oleva otos y. Otokseen tulleista $s = 650$ kannattaa kuolemanrangaistusta. Posteriori on Beta$(651, 350.667)$, ts.

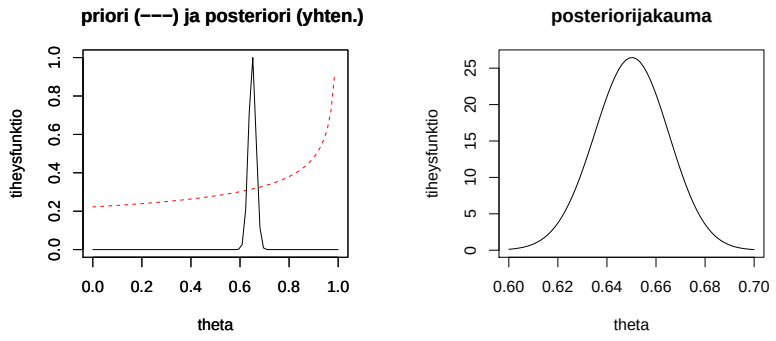
$$p(\theta|y) \propto \theta^{650} \, (1-\theta)^{349.667} \, .$$

Edelleen

$$\begin{aligned} \mathsf{E}(\theta|y) &= \frac{1+650}{1+\frac{2}{3}+1000} = 0.6499, \\ \mathsf{Var}(\theta|y) &= \frac{(1+650)(\frac{2}{3}+1000-650)}{(1+\frac{2}{3}+1000)^2(1+\frac{2}{3}+1000+1)} = 0.00228, \\ \mathsf{sd}(\theta|y) &= 0.0151, \\ \text{HDI}(0.95) &= (0.620, 0.679). \end{aligned}$$

HDI:n laskemisessa on käytetty asymptoottista approksimaatiota $0.6499 \pm 1.96 \cdot 0.0151$.

Kuten seuraavista kuvista näemme, posteriorijakauma on näin suurella otoskoolla lähes normaalinen:



Posterioriennustejakauma. Meillä on siis seuraavanlainen tilanne:

$$\begin{aligned} y_i|\theta &\sim \text{Bin}(1,\theta), \text{ havainnot riippumattomia ehdolla } \theta, \\ \theta &\sim \text{Beta}(\alpha,\beta) \text{ on priori}, \\ \theta|y &\sim \text{Beta}(\alpha+s,\beta+n-s) \text{ on posteriori}. \end{aligned}$$

Olkoon uusi havainto $\tilde{y}$ riippumaton y:stä ehdolla θ. Tehtävänä on määrittää posterioriennustejakauma $p(\tilde{y}|y)$. Koska

$$\begin{aligned} \mathsf{P}(\tilde{y}=1|y) &= \int_0^1 \mathsf{P}(\tilde{y}=1|\theta)\, p(\theta|y)\, \mathrm{d}\theta \\ &= \int_0^1 \theta\, p(\theta|y)\, d\theta \\ &= \frac{\alpha+s}{\alpha+\beta+n}\,,\ s=\sum_{i=1}^n y_i, \end{aligned}$$

voimme päätellä, että posterioriennustejakauma on

$$\tilde{y}|y \sim \text{Bin}\left(1, \frac{\alpha+s}{\alpha+\beta+n}\right).$$

Huomaa, että jos $\alpha=\beta=1$ eli priori on tasajakauma, niin

$$\mathsf{P}(\tilde{y}=1|y) = \frac{s+1}{n+2}.$$

Uuden havainnon jakauma ei siis enää ole ”tavallinen” havainnon jakauma $\text{Bin}(1,\theta)$ vaan jotain aivan muuta. Tämän jakauman myös Thomas Bayes laski.

Esim 1. (jatkoa) Esimerkissämme uusi havainto $\tilde{y}$ (yhden henkilön mielipide) ehdolla aineisto noudattaa Bernoullin jakaumaa, jolle

$$\mathsf{P}(\tilde{y}|y) = \frac{1+650}{1+\frac{2}{3}+1000} = 0.6499.$$

5.3 Poisson-otos

Poisson-prosessin avulla voidaan mallintaa ajassa tapahtuvaa ilmiötä, josta rekisteröidään tapahtuma-ajat. Merkitään $N(\Delta t)$:llä tapausten lukumäärä aikavälillä Δt ja $|\Delta t|$:lla aikavälin pituutta. Poisson-prosessin oletetaan toteuttavan seuraavat ehdot:

- $\mathsf{P}(N(\Delta t) = 1|\theta) = \theta\,|\Delta t| + o(|\Delta t|)$.
- $\mathsf{P}(N(\Delta t) \geq 2|\theta) = o(|\Delta t|)$.
- Tapahtumat erillisillä väleillä ovat riippumattomia (ehdolla θ).

(Tässä $o(h) : o(h)/h \to 0$, kun $h \to 0$.)

Näiden ehtojen nojalla Poisson-prosessi on malli *harvinaisille* tapahtumille. Se on yleisesti käytetty malli mm. jonoteoriassa ja riskiteoriassa.

Ehdoista seuraa, että

$$\mathsf{P}(N(\Delta t) = k|\theta) = \frac{(\theta|\Delta t|)^k}{k!}\, e^{-\theta|\Delta t|},\ k = 0, 1, 2, \ldots\,,$$

joka on Poissonin jakauman todennäköisyysfunktio, kun parametri on $\theta|\Delta t|$. Siis yhdessä aikayksikössä ($|\Delta t| = 1$) Poisson-tapahtumien lukumäärä noudattaa Poisson(θ)-jakaumaa.

Yleisesti Poissonin jakauman todennäköisyysfunktio on

$$p(y|\theta) = \frac{\theta^y}{y!}e^{-\theta}.$$

Jakauma on siinä mielessä poikkeuksellinen, että odotusarvo ja varianssi ovat yhtä suuret: $\mathsf{E}(y|\theta) = \theta$ ja $\mathsf{Var}(y|\theta) = \theta$.

Konjugaattinen priori Poissonin jakaumalle on gammajakauma Gamma(α, β), missä α on muotoparametri ja β käänteinen asteikkoparametri (intensiteetti). Huomaa, että joskus käytetään asteikkoparametria $1/\beta$ toisena parametrina. Gammajakauman tiheysfunktio on

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, \theta^{\alpha-1}\, e^{-\beta\theta},\ \theta > 0\,,$$

missä

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}\mathrm{d}u$$

on gammafunktio.

Gammajakauman tunnuslukuja ovat

$$\begin{aligned} \mathsf{E}(\theta|\alpha,\beta) &= \frac{\alpha}{\beta}, \\ \mathsf{Var}(\theta|\alpha,\beta) &= \frac{\alpha}{\beta^2}, \\ \text{moodi} &= \frac{\alpha-1}{\beta}. \end{aligned}$$

Odotusarvo voidaan johtaa samalla tavalla kuin beetajakauman tapauksessa:

$$\begin{aligned} \mathsf{E}(\theta|\alpha,\beta) &= \int_0^\infty \theta \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \mathrm{d}\theta \\ &= \frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha)} \int_0^\infty \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} \theta^{(\alpha+1)-1} e^{-\beta\theta} \mathrm{d}\theta \\ &= \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} \cdot 1 \\ &= \frac{\alpha}{\beta}\,. \end{aligned}$$

On hyödyllistä tuntea seuraavanlainen parametrin β ominaisuus: Jos $\phi \sim \text{Gamma}(\alpha, 1)$ ja $\theta = \frac{1}{\beta}\phi$, niin

$$\theta \sim \text{Gamma}(\alpha, \beta)\,.$$

Siten $\frac{1}{\beta}$ on asteikkoparametri.

Posteriorijakauma. Olkoon $y = (y_1, \ldots, y_n)$ Poisson-otos: havainnot $y_i,\ i = 1, ..., n$, ovat riippumattomia ehdolla θ ja $y_i|\theta \sim \text{Poisson}(\theta)$.

Tällöin uskottavuus on

$$p(y|\theta) = \prod_{i=1}^n \left[\frac{\theta^{y_i}}{y_i!}\, e^{-\theta}\right] \propto \theta^s\, e^{-n\theta}\,,$$

missä $s = \sum_{i=1}^n y_i$.

Olkoon lisäksi $\theta \sim \text{Gamma}(\alpha, \beta)$. Tällöin

$$\begin{aligned} p(\theta|y) &\propto p(\theta)\ p(y|\theta) \\ &\propto \theta^{\alpha-1}\, e^{-\beta\theta} \cdot \theta^s\, e^{-n\theta} \\ &\propto \theta^{\alpha+s-1}\, e^{-(\beta+n)\theta}, \end{aligned}$$

joka on Gamma$(\alpha + s, \beta + n)$. Gammajakauman ominaisuuksien perusteella

$$\begin{aligned} \mathsf{E}(\theta|y) &= \frac{\alpha + s}{\beta + n}, \\ \mathsf{Var}(\theta|y) &= \frac{\alpha + s}{(\beta + n)^2}. \end{aligned}$$

Huomaa, että suurilla otoksilla $\mathsf{E}(\theta|y) \approx \frac{s}{n} = \bar{y}$ (su-estimaattori) riippumatta parametrista α ja β:

Reunajakauma eli prioriennustejakauma (yhdelle havainnolle). Oletetaan, että $y|\theta \sim \text{Poisson}(\theta)$ ja $\theta \sim \text{Gamma}(\alpha, \beta)$. Tällöin y:n reunajakauma on (määritelmän mukaisesti)

$$\begin{aligned} p(y) &= \int_0^\infty \frac{\theta^y}{y!} e^{-\theta} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \, \mathrm{d}\theta \\ &= \frac{\Gamma(\alpha+y)\beta^\alpha}{y!(\beta+1)^{\alpha+y}\Gamma(\alpha)} \int_0^\infty \frac{(\beta+1)^{\alpha+y}}{\Gamma(\alpha+y)} \theta^{\alpha+y-1} e^{-(\beta+1)\theta} \, \mathrm{d}\theta \\ &= \frac{\Gamma(\alpha+y)\beta^\alpha}{y!(\beta+1)^{\alpha+y}\Gamma(\alpha)} \\ &= \frac{(\alpha+y-1)(\alpha+y-2)\cdots\alpha\Gamma(\alpha)}{\Gamma(\alpha)y!} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y \\ &= \binom{\alpha+y-1}{y} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y \\ &= \binom{\alpha+y-1}{\alpha-1} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y \quad y = 0, 1, 2, \ldots \end{aligned}$$

Saatu jakauma on *negatiivinen binomijakauma* NegBin(α, β), jolle

$$\begin{aligned} \mathsf{E}(y|\alpha, \beta) &= \frac{\alpha}{\beta}, \\ \mathsf{Var}(y|\alpha, \beta) &= \frac{\alpha}{\beta^2}(\beta + 1). \end{aligned}$$

Negatiivinen binomijakauma siis saadaan sekoituksena Poissonin jakaumista:

$$\text{NegBin}(y|\alpha, \beta) = \int \text{Poisson}(y|\theta) \, \text{Gamma}(\theta|\alpha, \beta) \, \mathrm{d}\theta \, .$$

Jakaumaa käytetään klassisessa tilastotietessä ”ylihajontatilanteissa” kuten heterogeenisten aineistojen mallintamisessa.

Huomaa, että joskus käytetään parametrointia (α, p), missä $p = \frac{\beta}{\beta+1}$ (kuten 1stBayes ja R!).

Posterioriennustejakama. Vastaavasti voidaan johtaa posterioriennustejakauma tulevalle havainnolle $\tilde{y}$, kun on jo havaittu $y_1, \ldots, y_n$. Jakauma $p(\tilde{y}|y)$ on

$$\text{NegBin}(\alpha + s, \beta + n).$$

Esim. 2. Sovellus epidemiologiaan. Oletetaan, että muuttuja y_i kuvaa sairastuneiden lukumäärää alueessa i. Lisäksi tiedetään odotettavissa oleva sairastuneiden määrä x_i alueessa i. Tämä suure saadaan rekisteristä ja se ottaa huomioon väkiluvun, ikä- ja sukupuolijakauman sekä sosioekonomisen statuksen. Epidemiologiassa lasketaan ns. standardoitu sairastavuusluku (SIR), joka on y_i/x_i.

Oletetaan, että

- $y_i|\theta, x_i \sim \text{Poisson}(\theta x_i)$;
- $y_1, \ldots, y_K$ ovat riippumattomia ehdolla θ, x;
- $\theta \sim \text{Gamma}(\alpha, \beta)$.

Tässä θ on tulkittavissa sairastavuusriskiksi.

On helppo havaita, että malli on loglineaarinen malli, sillä

$$\log \mathsf{E}(y_i|\theta, x_i) = \log \theta + \log x_i \,,$$

missä $\log x_i$ on kiinnitetty (ns. offset) muuttuja.

Posteriori on nyt

$$\begin{aligned} p(\theta|y) &\propto \prod_{i=1}^{K} \left\{ (\theta x_i)^{y_i} \, e^{-\theta x_i} \right\} \, \theta^{\alpha-1} \, e^{-\beta\theta} \\ &\propto \theta^{\alpha+\sum_{i=1}^K y_i - 1} \, e^{-(\beta+\sum_{i=1}^K x_i)\theta}, \end{aligned}$$

joka on

$$\text{Gamma}(\alpha + \sum_i y_i, \beta + \sum_i x_i).$$

Kiinnitetään huomiota seuraavaan asiaan: Posteriorin odotusarvo on

$$\mathsf{E}(\theta|y, x) = \frac{\alpha + \sum y_i}{\beta + \sum x_i} \,.$$

Kun tätä verrataan su-estimaattoriin

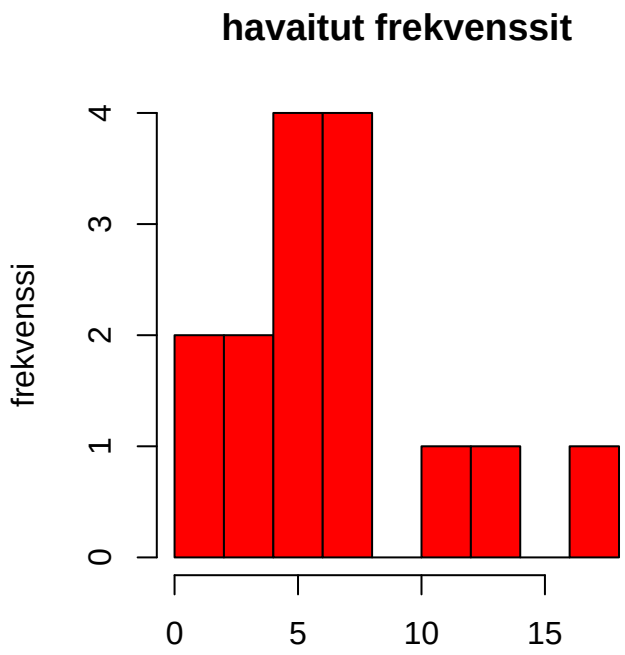
$$\hat{\theta} = \frac{\sum_i y_i}{\sum_i x_i},$$

niin havaitaan, että Bayes-estimaattori ”kutistaa” kohti prioria. (Standardoitu sairastumisluku on su-estimaatti θ:lle Poisson-mallissa, kun θ vaihtelee alueittain.)

Esim. 3. Hirvet. Tutkimusalue koostuu 15:sta 100 km^2:n ruudusta, joista on laskettu lentokoneesta käsin hirvien lukumäärät tietyllä hetkellä. Havaintoaineisto on:

$$5, 7, 7, 12, 2, 14, 7, 8, 5, 6, 18, 6, 4, 1, 4.$$

Ainestosta laskettu keskiarvo on 7.07, varianssi 20.352 ja keskihajonta 4.51.

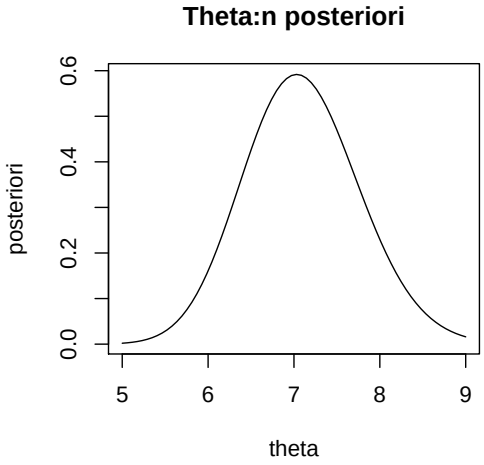


Oletetaan, että $y_i|\theta \sim \text{Poisson}(\theta)$ ovat riippumattomia ehdolla θ. Valitaan prioriksi $\theta \sim \text{Gamma}(4, 0.5)$ (jonka odotusarvoarvo on 8 ja keskihajonta 4).

Tällöin *posteriori* on $\theta|y \sim \text{Gamma}(110, 15.5)$.

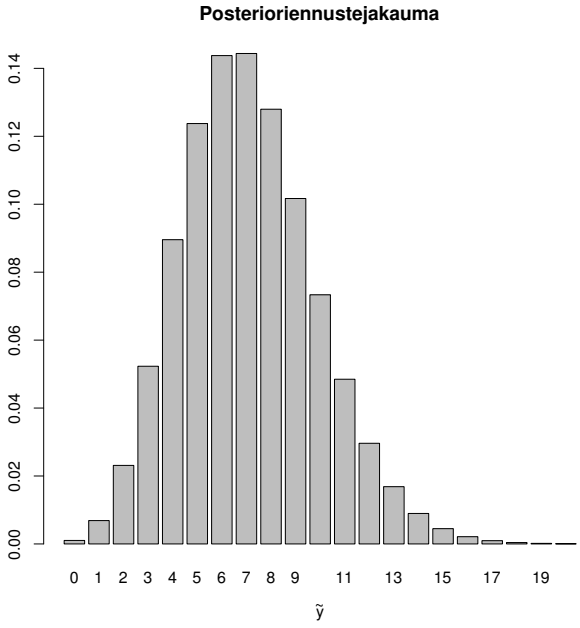
Tunnuslukuja:

odotusarvo	7.0968
varianssi	0.4578
keskihajonta	0.6767
95%:n HDI	$5.7923, 8.4375$



Posterioriennustejakauma (uudelle havainnolle) on NegBin$(110, 15.5)$, jolle

odotusarvo	7.0968
varianssi	7.5546
keskihajonta	2.7486

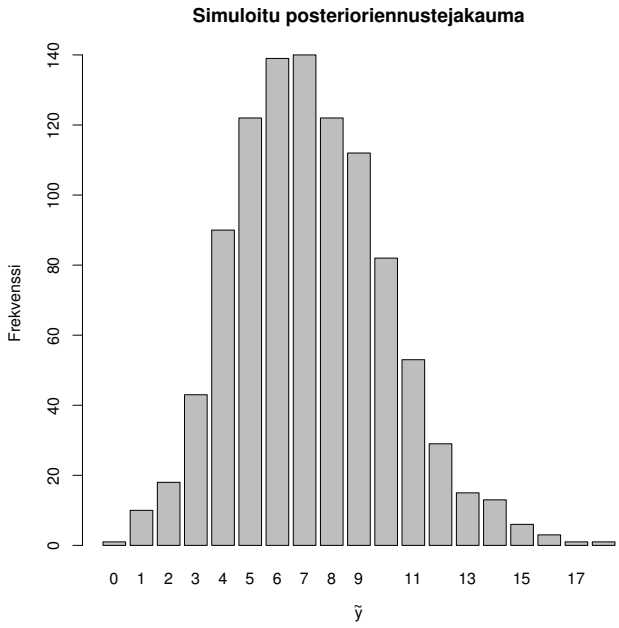


Huomaa, että R ja 1stbayes käyttää negatiivisen binomijakauman parametrisointia $(\alpha, p = \frac{\beta}{\beta+1})$. Jakauman todennäköisyysfunktio on silloin

$$p(y) = \begin{pmatrix} \alpha + y - 1 \\ y \end{pmatrix} p^{\alpha} (1-p)^{y}, \ y = 0, 1, 2, \ldots$$

Posterioriennustejakaumaa voidaan simuloida R:llä joko suoraan negatiivisesta binomijakaumasta tai simuloimalla ensin θ (theta.post) posteriorijakaumasta ja sitten $\tilde{y}$ (y.post) Poisson-jakaumasta ehdolla θ:

```
n.sims <- 1000
#Näin:
y.post <- rnbinom(n.sims,size=110,prob=15.5/(15.5+1))
#Tai näin:
theta.post <- rgamma(n.sims,110,15.5)
y.post <- rpois(n.sims,theta.post)
```



5.4 Otos eksponentiaalisesta mallista

Yksinkertainen malli havainnoille, jotka voivat saada positiivisia reaalilukuarvoja on *eksponettijakauma*. Jakaumaa voidaan käyttää tietyissä tapauksissa elinaikojen mallinnukseen. Jakauman tiheysfunktio on

$$p(y|\theta) = \theta\, e^{-\theta y}\,,\ y > 0\,,\ \theta > 0\,,$$

joka on erikoistapaus gammajakaumasta: $\text{Exp}(\theta) \sim \text{Gamma}(1, \theta)$. Sille pätee

$$\mathsf{E}(y|\theta) = \frac{1}{\theta}\,, \mathsf{Var}(y|\theta) = \frac{1}{\theta^2},$$

ja *eloonjäämisfunktio* (tai välttöfunktio, survival function) on

$$S(t|\theta) = \mathsf{P}(y > t|\theta) = e^{-\theta t},\ t > 0\,.$$

Eksponenttijakaumalla on *unohtavuusominaisuus*: Jokaiselle $t, h > 0$

$$\begin{aligned}\mathsf{P}(y > t + h|y > t, \theta) &= \frac{\mathsf{P}(y > t + h|\theta)}{\mathsf{P}(y > t|\theta)} \\ &= \frac{e^{-\theta(t+h)}}{e^{-\theta t}} \\ &= e^{-\theta h},\end{aligned}$$

joka ei riipu t:stä. Tämä tarkoittaa sitä, että "laite ei vanhene".

Määritellään *vaarafunktio (hazard function)*:

$$h(y|\theta) = \frac{p(y|\theta)}{S(y|\theta)},\ y > 0\,.$$

Tämä voidaan ymmärtää niin, että $h(y|\theta)\,\Delta y$ on likimain yhtä kuin

$$\mathsf{P}(\text{ kuolema väl. } (y, y + \Delta y) \mid \text{elossa hetkellä } y\ , \theta)$$

Eksponenttijakaumalle $h(y|\theta) = \theta$ (vakio). Myös käänteinen on voimassa: jos vaarafunktio on vakio, niin jakauma on eksponentiaalinen. (Vaarafunktio määrittelee yksikäsitteisesti jakauman.)

Jos tapahtumat noudattavat Poisson-prosessia (vrt. kappale 5.3), niin silloin tapahtumien väliset ajat noudattavat eksponenttijakaumaa. Tämä on yksi eksponenttijakauman karakterisointi.

Posteriorijakauma. Olkoot havainnot $y_i|\theta \sim \text{Exp}(\theta)$ ehdollisesti riippumattomia ehdolla θ. Tällöin uskottavuus on

$$\begin{aligned}p(y|\theta) &= \prod_{i=1}^{n} \left\{\theta e^{-\theta y_i}\right\} \\ &= \theta^n\, e^{-\theta s},\end{aligned}$$

missä $s = \sum_{i=1}^{n} y_i$. Voidaan myös helposti osoittaa, että

$$s|\theta \sim \text{Gamma}(n, \theta)\,.$$

Konjugaattinen priori on $\theta \sim \text{Gamma}(\alpha, \beta)$. Käytettäessä tätä prioria, posteriorijakaumaksi saadaan

$$\begin{aligned} p(\theta|y) &\propto \theta^n e^{-\theta s} \cdot \theta^{\alpha-1} e^{-\beta\theta} \\ &\propto \theta^{\alpha+n-1}\, e^{-(\beta+s)\theta}, \end{aligned}$$

joka on $\text{Gamma}(\alpha+n, \beta+s)$. Edelleen

$$\mathsf{E}(\theta|y) = \frac{\alpha+n}{\beta+s}, \quad \mathsf{Var}(\theta|y) = \frac{\alpha+n}{(\beta+s)^2}.$$

Sensurointi elinaikamalleissa. Elinaikatutkimuksessa voi käydä niin, että tutkimus päättyy ennen kuin kaikki koeyksilöt ovat kuolleet (tai komponentit hajonneet.) Voidaan päätyä seuraavanlaiseen *oikealta sensuroituun* otokseen:

- Otoskoko on n.
- Ensimmäiset k havaintoa ovat $y_1, \ldots, y_k\,,\ k \leq n$.
- Lopuista $(n-k)$:sta havainnosta tiedetään, että niille $y_j > y_0$, missä y_0 on tunnettu "sensurointikohta".

Lisäksi oletetaan, että havainnot $y_i|\theta \sim \text{Exp}(\theta)$ ovat riippumattomia ehdolla θ, ja valitaan konjugaattipriori $\theta \sim \text{Gamma}(\alpha, \beta)$.

Havainnoista y_i, $i = k+1, \ldots, n$ tiedetään vain, että $y_i > y_0$ (mikä sekin on informaatiota ja tulee hyödyntää). Tämän tapauksen todennäköisyys on

$$\mathsf{P}(y_i > y_0|\theta) = e^{-\theta y_0} = S(y_0|\theta),$$

joka tulee uskottavuusfunktioon tiheysfunktion tilalle. Sensuroidun otoksen *uskottavuusfunktio* on siis

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{k} \left\{\theta e^{-\theta y_i}\right\} \cdot [S(y_0|\theta)]^{(n-k)} \\ &= \theta^k\, e^{-\theta s_k}\, e^{-(n-k)\theta y_0} \\ &= \theta^k\, e^{-\theta[s_k+(n-k)y_0]}, \end{aligned}$$

missä $s_k = \sum_{i=1}^{k} y_i$.

Posteriorijakauma on nyt

$$\begin{aligned} p(\theta|y) &\propto \theta^k\, e^{-\theta[s_k+(n-k)y_0]}\, \theta^{\alpha-1} e^{-\beta\theta} \\ &= \theta^{\alpha+k-1}\, e^{-[\beta+s_k+(n-k)y_0]\theta}, \end{aligned}$$

joka on

$$\text{Gamma}(\alpha + k, \beta + s_k + (n-k)y_0). \tag{5.1}$$

Sen posteriorikeskiarvo ja -varianssi ovat

$$\begin{aligned} \mathsf{E}(\theta|y) &= \frac{\alpha + k}{\beta + s_k + (n-k)y_0}\,, \\ \mathsf{Var}(\theta|y) &= \frac{\alpha + k}{(\beta + s_k + (n-k)y_0)^2}\,. \end{aligned}$$

Sensuroinnin hallitseminen on erittäin tärkeä elinaikatutkimuksissa – usein aineisto on sensuroitu.

Esim. 4. Sensurointi. Oletetaan, että on tehty seuraavat elinaikahavainnot:

$$1.54, 0.70, 1.23, 0.82, 0.99, 1.33, 0.38, 0.99, 1.97, 1.10, 0.40$$

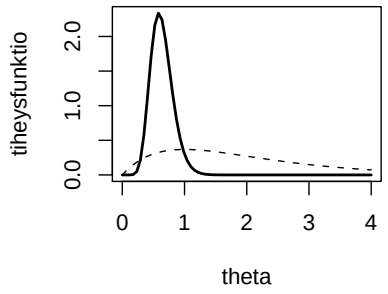
Aineiston tiedetään olevan oikealta sensuroidun arvoon $y_o = 2.0$ ja sensuroitujen havaintojen määrä on 4. Otoksen koko on 15. Oletetaan, että $y_i|\theta \sim \text{Exp}(\theta)$ ja että havainnot ovat ehdollisesti riippumattomia. Oletetaan lisäksi, että priori on $\theta \sim \text{Gamma}(2, 1)$, jolloin prioriodotusarvo on 2 ja priorivarianssi 2 (keskihajonta 1.42).

Tuloksen (5.1) perusteella posteriori on

$$\text{Gamma}(13.00, 20.45),$$

jonka odotusarvo on 0.636, keskihajonta 0.176 ja 95 %:n symmetrinen posterioriväli $(0.338, 1.025)$.

Alla on priorin (katkoviiva) ja posteriorin (yhteinäinen viiva) kuvaajat.



5.5 Normaaliotos, varianssi tuntematon, odotusarvo tunnettu

Oletetaan, että $y_i|\mu,\phi \sim \mathrm{N}(\mu,\phi)$, missä μ tunnettu ja ϕ tuntematon (ja kiinnostuksen kohde). Havainnot $y=(y_1,\ldots,y_n)$ oletetaan riippumattomiksi ehdolla ϕ.

Tällöin *uskottavuus* on

$$\begin{aligned} p(y|\phi) &\propto \phi^{-\frac{n}{2}}\, e^{-\frac{1}{2\phi}\sum_{i=1}^n (y_i-\mu)^2} \\ &\propto \phi^{-\frac{n}{2}}\, e^{-\frac{n}{2\phi}\, s_0^2}, \end{aligned}$$

missä $s_0^2 = \frac{1}{n}\sum_{i=1}^n (y_i-\mu)^2$.

Konjugaattipriori on käänteinen gammajakauma InvGamma(α,β), jonka tiheysfunktio on

$$p(\phi) \propto \phi^{-(\alpha+1)}\, e^{-\frac{\beta}{\phi}}\,,\ \phi>0\,.$$

"Käänteisyys" tarkoittaa sitä, että jos $X \sim$ Gamma(α,β), niin $1/X \sim$ InvGamma(α,β). (Johda tiheysfunktion kaava). Jakauman odotusarvo ja varianssi ovat

$$\begin{aligned} \mathsf{E}(\phi) &= \frac{\beta}{\alpha-1}\,,\ \alpha>1,\ \beta>0, \\ \mathsf{Var}(\phi) &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}\,,\ \alpha>2,\ \beta>0. \end{aligned}$$

Posteriori on nyt

$$\begin{aligned} p(\phi|y) &\propto \phi^{-(\alpha+1)}\, e^{-\frac{\beta}{\phi}}\, \phi^{-\frac{n}{2}} e^{-\frac{n}{2\phi}s_0^2} \\ &\propto \phi^{-(\alpha+\frac{n}{2}+1)}\, e^{-(\beta+\frac{n}{2}s_0^2)/\phi}, \end{aligned}$$

joka on InvGamma$(\alpha+\frac{n}{2}, \beta+\frac{n}{2}\,s_0^2)$. Posteriorijakauman tunnusluvut ovat

$$\begin{aligned} \mathsf{E}(\phi|y) &= \frac{\beta+\frac{n}{2}\,s_0^2}{\alpha+\frac{n}{2}-1}, \\ \mathsf{Var}(\phi|y) &= \frac{(\beta+\frac{n}{2}\,s_0^2)^2}{(\alpha+\frac{n}{2}-1)^2\,(\alpha+\frac{n}{2}-2)}. \end{aligned}$$

Suurilla otoksilla

$$\mathsf{E}(\phi|y) \approx s_0^2 = \frac{1}{n}\sum_{i=1}^n (y_i-\mu)^2,$$

joka on suurimman uskottavuuden estimaattori!

Luku 6

Jeffreysin epäinformatiivinen priori

Palataan ongelmaan, jossa ennakkotietoa ei ole. Seuraavia ratkaisuja on tarjolla (esimerkiksi):

- Kun $y|\theta \sim \mathrm{N}(\theta, v)$, missä v on tunnettu, käytetään prioria $\theta \sim \mathrm{N}(0, w)$, missä w on suuri.
- Kun $y|\theta \sim \mathrm{N}(\theta, v)$, missa v on tunnettu, käytetään prioria $p(\theta) \propto \kappa$.
- Kun $y|\phi \sim \mathrm{N}(\mu, \phi)$, missä μ on tunnettu, käytetään prioria $p(\phi) \propto 1/\phi$.
- Kun $y|\theta \sim \mathrm{Bin}(1, \theta)$, käytetään prioria $\theta \sim \mathrm{Beta}(1, 1)$ (tasajakauma).

Tasajakauman käyttö priorijakaumana ei ole onglematonta. Tarkastellaan seuraavanlaista tilannetta: $y|\theta \sim \mathrm{Bin}(1, \theta)$ ja meillä ei ole ennakkoinformaatiota θ:sta. On "luonnollista" valita prioriksi $p(\theta) \propto 1$ (ns. Bayes-Laplace-priori). Mutta silloin meillä ei myöskään ole informaatiota parametrista $\phi = \theta^2$. Edellinen valinta johtaa muunnoskaavan perusteella prioriin

$$p(\phi) \propto \frac{1}{\sqrt{\phi}},$$

joka ei ole tasajakauma. Johtopäätöksenä on, että tasajakauman valinnasta seuraa paradoksi. Tämä on ollut yksi bayesiläisiin kohdistunut lyömäase!

Jeffreysin menetelmä. Sir Harold Jeffreys (1891–1989) oli brittiläinen matemaatikko, tilastotieteilijä, geofyysikko ja astronomi. Hän esitti periaat-

teen epäinformatiivisen priorin konstruoimiseksi: *säännön, jolla priorijakauma määritetään, tulisi johtaa samaan tulokseen, vaikka sääntöä sovellettaisiin muunnettuun parametriin.*

Kun määritellään epäinformatiiviseksi priorijakaumaksi $p(\theta) \propto J(\theta)^{1/2}$, missä $J(\theta)$ on parametrin θ *Fisherin informaatio*

$$J(\theta) = \mathsf{E}\left[\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2 \middle| \theta\right] = -\mathsf{E}\left[\frac{d^2\log p(y|\theta)}{d\theta^2} \middle| \theta\right],$$

Jeffreysin periaate toteutuu.

Tod. Olkoon $\phi = h(\theta)$, missä $h(.)$ on aidosti monotoninen ja derivoituva funktio. Tällöin Fisherin informaatio muunnetulle parametrille ϕ on

$$J(\phi) = \mathsf{E}\left[\left(\frac{d\log p(y|\phi)}{d\phi}\right)^2 \middle| \phi\right] = \mathsf{E}\left[\left(\frac{d\log p(y|h(\theta))}{d\theta}\frac{d\theta}{d\phi}\right)^2 \middle| \theta\right] = J(\theta)\left(\frac{d\theta}{d\phi}\right)^2,$$

mistä saadaan $J(\phi)^{1/2} = J(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right|$. Tästä seuraa, että

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| \propto J(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right| = J(\phi)^{1/2}.$$

Esim. 1. Jeffreysin priori binomijakaumalle . Oletetaan, että $y|\theta \sim \text{Bin}(n,\theta)$. Tällöin

$$\begin{aligned}
p(y|\theta) &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \\
\Rightarrow \log p(y|\theta) &= y\log\theta + (n-y)\log(1-\theta) + \log\binom{n}{y} \\
\Rightarrow \frac{d}{d\theta}\log p(y|\theta) &= \frac{y}{\theta} - \frac{n-y}{1-\theta} \\
\Rightarrow -\mathsf{E}(\frac{d^2}{d\theta^2}\log p(y|\theta)\,|\theta) &= \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.
\end{aligned}$$

Jeffreysin priori on siis $p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$, joka on Beta$(\frac{1}{2},\frac{1}{2})$-jakauma. Huomaa, että epäinformatiivinen Bayes-Laplace-priori on Beta$(1,1)$ (tasajakauma).

Jeffreysin priori ei aina johda hyvään tulokseen. Moniulotteiselle parametrille sen käyttö on hankalaa.

Luku 7

Yleisiä periaatteita

7.1 Epäoleellisuus ja tyhjentävyys

Epäoleellisuus (ancillarity) ja *tyhjentävyys (sufficiency)* ovat keskeisiä tilastollisen päättelyn käsitteitä ja ne on hyvä tuntea. Seuraavassa oletetaan, että tuntematon suure on θ (sitä sen kummemmin spesifioimatta), priori θ:lle $p(\theta)$ ja uskottavuus $p(y|\theta)$.

Epäoleellisuus . Tarkastellaan seuraavassa kahta erikoistapausta:

(a) Jos aineiston y jakauma ehdolla θ (uskottavuus) $p(y|\theta)$ *ei* riipu θ:sta, niin

$$p(\theta|y) \propto p(\theta),$$

eli posteriori on sama kuin priori. Tämä merkitsee sitä, että havainnosta y ei opita, tai toisin sanoin, y on epäinformatiivinen θ:lle. Sanomme, että y on *epäoleellinen* θ:lle.

(b) Tarkastellaan tilannetta, jos aineisto y on osittain epäinformatiivinen θ:lle.Oletetaan, että $y = (x, z)$. Voidaan ajatella siten, että aineisto y havaintaan kahdessa osassa: ensin x ja sitten z.

Havaitaan ensin x, jolloin "opetus" on

$$p(\theta|x) \propto p(\theta)\, p(x|\theta).$$

Havaitaan tämän jälkeen z, jolloin opetus on

$$\begin{aligned} p(\theta|y) = p(\theta|x,z) &\propto p(\theta|x)\, p(z|x,\theta) \\ &\propto p(\theta)\, p(x|\theta)\, p(z|x,\theta). \end{aligned}$$

Jos nyt $p(x|\theta) \propto 1$, niin x on *epäoleellinen* θ:lle. Tällöin $p(\theta|x) = p(\theta)$ ja siten

$$p(\theta|y) = p(\theta|x,z) \propto p(\theta)\, p(z|x,\theta).$$

Siis tässä tapauksessa posteriori

$$p(\theta|y) \propto p(\theta)\, p(z|x,\theta)$$

riippuu z:n ehdollisesta jakaumasta ehdolla x (ja θ). Posteriori siis sisältää epäoleellisella aineiston osalla x ehdollistetun jakauman. Huomaa, että vaikka epäoleellinen otossuure ei sisällä suoraa informaatiota θ:sta, se kuitenkin vaikuttaa epäsuoralla tavalla posteriorijakaumaan!

Tyhjentävyys. Jos $p(z|x,\theta) \propto 1$, niin sanomme, että x on *tyhjentävä* θ:lle. Silloin

$$p(\theta|x,z) = p(\theta|x) \propto p(\theta)\, p(x|\theta)\,.$$

Posteriori (ja siten johtopäätökset) ei riipu lainkaan z:sta.

"Klassisessa tilastotieteessä" pyritään ehdollistamaan päättelyt epäoleellisen tunnusluvun (kiusatunnusluvun) suhteen. Toisaalta pyritään etsimään tyhjentävä tunnusluku ja käyttämään estimaattoreita ja testisuureita, jotka ovat sen funktiota. Bayes-tilastotieteessä näihin ei yleensä tarvitse kiinnittää huomiota, vaan kaikki hoituu itsetään posterioria laskettaessa.

Esim. 1. Oletetaan, että $y_i|\theta \sim N(\theta, 1)$, $i = 1, 2$. Oletetaan lisäksi, että havainnot koostuvat mittauksista

$$\begin{aligned} x &= y_1 - y_2 \\ z &= y_1 + y_2 \end{aligned}$$

(ja siis havaintoa ei ole muuttujista y_1 ja y_2).

On helppo nähdä, että

$$p(x,z|\theta) = p(x)\, p(z|\theta),$$

sillä $\{x|\theta,z\} \sim \mathrm{N}(0,2)$ ja $z|\theta \sim \mathrm{N}(2\theta,2)$. Siis z on tyhjentävä θ:lle. Koska lisäksi $x|\theta \sim N(0,2)$, jolloin $p(x|\theta)$ ei riipu θ:sta, on x epäoleellinen. Lisäksi koska $p(x|\theta) = p(x|z,\theta)$, ovat x ja z ovat riippumattomia (ehdolla θ) ja myös $p(z|x,\theta) = p(z|\theta)$.

Posteriori on nyt

$$p(\theta|x,z) \propto p(\theta)\, p(z|\theta) \propto p(\theta|z),$$

mistä nähdään, että posteriorijakauma y:n suhteen on sama kuin posteriorijakauma z:n suhteen.

Esim. 2. Oletetaan, että $x_i|\theta \sim \text{Bin}(1,\theta)$ ja havainnot $x = (x_1,\ldots,x_n)$ ovat riippumattomia ehdolla θ. Merkitään $s = \sum_{i=1}^n x_i$ ("onnistumisten" määrä).

Tällöin uskottavuus on

$$\begin{aligned} p(x|\theta) &= \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^s(1-\theta)^{n-s}. \end{aligned}$$

Tässä $p(x|s,\theta) = \binom{n}{s}^{-1} \propto 1$, joka ei riipu θ:sta. Siis s on tyhjentävä θ:lle.

Tulkinta: Kun $s = \sum_{i=1}^n x_i$ tunnetaan, niin havaintojen $x_1,\ldots,x_n$ tunteminen ei opeta lisää θ:sta. Voidaan siis toimia posteriorilla $p(\theta|x)$ tai $p(\theta|s)$.

7.2 Uskottavuusperiaate ja epäinformatiivinen pysäyttäminen

Uskottavuusperiaatteen mukaan kahden todennäköisyysmallin, joilla on vakiolla kertomista lukuun ottamatta sama uskottavuusfunktio, pitäisi johtaa samaan päättelyyn θ:n suhteen. Bayes-päättely toteuttaa tämän periaatteen, koska posteriorijakauma riippuu aineistosta ainoastaan uskottavuusfunktion välityksellä.

(Huomaa kuitenkin, että Jeffreysin periaatteella muodostettu priorijakauma riippuu otantajakaumasta. Tällöin priorit voivat poiketa vaikka uskottavuusfunktiot olisivat verrannolliset eli vakiokerrointa lukuun ottamatta samat. Siksi Jeffreysin periaate on ristiriidassa uskottavuusperiaatteen kanssa!)

Esim. 3. Jatketaan edellistä esimerkkiä. Oletetaan, että n *ei ole kiinnitetty.* Jatketaan toistoja, kunnes saadaan r onnistumista. Nyt n on satunnainen. Uskottavuus on

$$\begin{aligned} p(n|\theta) &= \left(\begin{array}{c} n-1 \\ r-1 \end{array}\right)\theta^{r-1}(1-\theta)^{(n-1)-(r-1)}\cdot\theta \\ &= \left(\begin{array}{c} n-1 \\ r-1 \end{array}\right)\theta^{r}(1-\theta)^{n-r} \\ &\propto \theta^r(1-\theta)^{n-r}\,,\ n = r, r+1,\ldots\,. \end{aligned}$$

Kun n on kiinnitetty (ja r satunnainen), niin

$$\begin{aligned} p(r|\theta, n) &= \left(\begin{array}{c} n \\ r \end{array} \right) \theta^r (1-\theta)^{n-r} \\ &\propto \theta^r (1-\theta)^{n-r}. \end{aligned}$$

Koska kummassakin tapauksessa uskottavuus on sama, niin myös *posteriori ja siten Bayes-tilastotieteen johtopäätökset* ovat samat, jos priori on sama. Klassisessa tilastotieteessä su-estimaattori on kummassakin tapauksessa $\hat{\theta} = r/n$, mutta jakaumateoria on eri (miksi?).

Tärkeä uskottavuusperiaatteen sovellus liittyy otannan pysäyttämissääntöihin. Pysäyttämissääntöä sanotaan *epäinformatiiviseksi,* jos peräkkäisotannassa todennäköisyys keskeyttää otanta ei riipu parametrista θ tai riippuu siitä ainoastaan havaintojen välityksellä. Tällöin pysäyttämissääntö ei vaikuta uskottavuusfunktioon eikä uskottavuusperiaatteen mukaan myöskään parametriin θ liittyvään tilastolliseen päättelyyn. Bayes-päättelyssä pysäyttämissääntö on epäinformatiivinen, jos pysäyttämissääntö ei vaikuta uskottavuusfunktioon eikä priorijakaumaan.

Oletetaan, että havainnot $y = (y_1, ..., y_n)$ ovat riippumattomia ehdolla θ ja noudattavat jakaumaa $p(y_i|\theta)$. Oletetaan lisäksi, että pysäyttämissääntö i:nnen havainnon kohdalla riippuu edellisistä havainnoista $y_1, ...y_{i-1}$ Merkitään pysäyttämishetki käyttäen indikaattorimuuttujaa $\delta_i,\ i = 1, ..., n+1$, ja määritellään, että

$$\delta_i = \begin{cases} 1, \text{ jos pysäytetään otanta,} \\ 0, \text{ jos poimitaan havainto } y_i. \end{cases}$$

Tällöin todennäköisyysmalli, joka sisältää sekä havainnot että pysäyttämissäännön, voidaan kirjoittaa muodossa

$$p(y, \delta|\theta) = \left[\prod_{i=1}^{n} \mathsf{P}(\delta_i = 0|\theta, y_1, ..., y_{i-1}) p(y_i|\theta) \right] \times \mathsf{P}(\delta_{n+1} = 1|\theta, y_1, ..., y_n).$$

Jos nyt pysäyttämistodennäköisyys i:nnen havainnon kohdalla määräytyy edellisten havaintojen $y_1, ..., y_{i-1}$ perusteella kaikilla $i = 1, ..., n+1$, niin

$$\mathsf{P}(\delta_i = 0|\theta, y_1, ..., y_{i-1}) = \mathsf{P}(\delta_i = 0|y_1, ..., y_{i-1}),\ i = 1, 2, ...$$

Tällöin uskottavuusfunktio on

$$p(y, \delta|\theta) \propto \prod_{i=1}^{n} p(y_i|\theta) = p(y|\theta),$$

joten pysäyttämissääntö ei vaikuta uskottavuuspäättelyyn.

Esim 3. (jatkoa) . Kun jatketaan Bernoulli-kokeita, kunnes "onnistumisia" on r, pysäyttäminen määräytyy edellisistä havainnoista. Tällöin uskottavuuspäättelyssä ei tarvitse ottaa huomioon pysäyttämissääntöä. Jos otoskoko n on ennalta kiinnitetty, se on parametrista θ riippumaton eikä pysäytyssääntöä tarvitse ottaa huomioon.

Luku 8

Hypoteesien testaus

Olkoon parametriavaruus Θ ja aineisto y. *Hypoteesilla* H tarkoitetaan parametriavaruuden osajoukkoa $\Theta_H \subset \Theta$ (kuten "klassisessa" tilastotieteessä). Olkoon testattava hypoteesi $H_0 : \theta \in \Theta_0$ ja vastahypoteesi $H_1 : \theta \in \Theta_1$, missä $\Theta_0 \cup \Theta_1 = \Theta$ ja $\Theta_0 \cap \Theta_1 = \emptyset$. Testaustilanteessa on valittava H_0:n ja H_1:n väliltä.

"Klassisessa" testiteoriassa testi määritellään *hylkäämisalueen*

$$R = \{y \,|\, y \text{ johtaa } H_0\text{:n hylkäämiseen } \},$$

avulla. Hylkäämisalue on havaintoavaruuden osa. Testiä kuvaavavat todennäköisyydet

$$\begin{aligned} \mathsf{P}(y \in R|\theta), &\quad \theta \in \Theta_0, \\ 1 - \mathsf{P}(y \in R|\theta), &\quad \theta \in \Theta_1, \end{aligned}$$

joita kutsutaan I lajin ja II lajin virheiksi.

Bayes-tilastotieteessä tilanne on vieläkin yksinkertaisempi. Hypoteeseille määritellään *prioritodennäköisyydet*

$$\begin{aligned} \pi_0 &= \mathsf{P}(\theta \in \Theta_0), \\ \pi_1 &= \mathsf{P}(\theta \in \Theta_1) \end{aligned}$$

sekä *posterioritodennäköisyydet*

$$\begin{aligned} p_0 &= \mathsf{P}(\theta \in \Theta_0|y), \\ p_1 &= \mathsf{P}(\theta \in \Theta_1|y). \end{aligned}$$

Priorivedonlyöntisuhde π_0/π_1 kuvaa, kuinka paljon uskomme hypoteesiin H_0 suhteessa hypoteesiin H_1 a priori, ja posteriorivedonlyöntisuhde p_0/p_1 kuvaa tilannetta havainnon y jälkeen. *Bayes-tekijäksi* sanotaan suhdetta

$$B = \frac{p_0/p_1}{\pi_0/\pi_1}.$$

Siten B mittaa, kuinka aineisto muuttaa priorikäsitystä. Todennäköisyys $p_0 = \mathsf{P}(\theta \in \Theta_0|y)$ voidaan esittää B:n avulla seuraavasti:

$$p_0 = \frac{1}{[1 + \frac{\pi_1}{\pi_0} B^{-1}]}.$$

Bayes-tekijä soveltuu hyvin *pistehypoteesien* tarkasteluun: Oletetaan , että $\Theta_0 = \{\theta_0\}$ ja $\Theta_1 = \{\theta_1\}$. Tällöin

$$\begin{aligned} p_0 &= \mathsf{P}(\theta = \theta_0|y) \propto \pi_0\, p(y|\theta_0) \\ p_1 &= \mathsf{P}(\theta = \theta_1|y) \propto \pi_1\, p(y|\theta_1). \end{aligned}$$

Edelleen

$$\frac{p_0}{p_1} = \frac{\pi_0}{\pi_1} \times \frac{p(y|\theta_0)}{p(y|\theta_1)},$$

jolloin $B = p(y|\theta_0)/p(y|\theta_1)$ on *uskottavuusosamäärä*.

Jos sen sijaan Θ_0 ja Θ_1 ovat *joukkohypoteeseja*, tarvitaan integrointia parametriavaruuden yli:

$$B = \frac{p(H_0|y)/\pi_0}{p(H_1|y)/\pi_1} = \frac{\int_{\Theta_0} p(\theta|y)\mathrm{d}\theta/\pi_0}{\int_{\Theta_1} p(\theta|y)\mathrm{d}\theta/\pi_1} = \frac{\int_{\Theta_0} p(y|\theta)p(\theta)/\pi_0\,\mathrm{d}\theta}{\int_{\Theta_1} p(y|\theta)p(\theta)/\pi_1\,\mathrm{d}\theta},$$

missä $\pi_i = \int_{\Theta_i} p(\theta)d\theta$, $i = 0, 1$.

Tapaus, missä H_0 on pistehypoteesi ja H_1 on yhdistetty hypoteesi, ei ole monien mielestä (esim. Gelman et al. 2013) mielekäs (bayesiläisessä asiayhteydessä). Kuitenkin muodollisesti voidaan määritellä hypoteesit $H_0:\ \theta = \theta_0$ ja $H_1:\ \theta \neq \theta_0$, jolloin Bayesin tekijä on

$$B = \frac{p(y|\theta_0)}{\int_{\Theta\setminus\{\theta_0\}} p(y|\theta)p(\theta|H_1)\mathrm{d}\theta},$$

missä $p(\theta|H_1)$ on hypoteesia H_1 vastaava priorijakauma. On huomattava, että lopputulos riippuu oleellisesti siitä, miten informatiivista priorijakaumaa käytetään.

Esim. 1. (Cavendishin aineisto). Oletetaan, että havainnot y_i, $i = 1, ..., n$, ovat riippumattomia ehdolla θ ja $y_i \sim \mathrm{N}(\theta, 0.1)$. Testataan hypoteesia $H_0 : \theta = 5.4$ vs. $\theta = 5.5$.

Olkoon $\pi_0 = 0.8$ ja $\pi_1 = 0.2$ (jolloin otetaan vahvasti kantaa H_0:n puolesta). Tällöin priorivedonlyöntisuhde on $\pi_0/\pi_1 = 4.0$ ja Bayesin tekijä

$$B = \frac{p(y|\theta_0)}{p(y|\theta_1)} = e^{-\frac{1}{2 \cdot 0.1}[\sum_i (y_i - 5.4)^2 - \sum (y_i - 5.5)^2]} \approx 0.449.$$

Posteriorivedonlyöntisuhde on $p_0/p_1 = (\pi_0/\pi_1) \times B \approx 1.797$ eli havainto muuttaa käsitystämme H_1:n suuntaan.

Esim. 2. Hemofilian periytyminen. Miehen perimä sisältää kromosomiparin XY ja naisen perimä parin XX. Merkitään **X**, jos kromosomi X sisältää sairauden aiheuttajan. Seuraavat mahdollisuudet esiintyvät:

$$\text{Mies on } \begin{cases} \text{terve, jos hänellä on XY,} \\ \text{sairas, jos hänellä on } \mathbf{X}\text{Y.} \end{cases}$$

$$\text{Nainen on } \begin{cases} \text{terve, jos hänellä on XX,} \\ \text{kantaja, jos hänellä on } \mathbf{X}\text{X,} \\ \text{sairas, jos hänellä on } \mathbf{XX}. \end{cases}$$

Ongelma: Tiedetään, että nainen ei ole sairas (vaan terve tai kantaja).Lisäksi tiedetään, että naisen äiti ei ole sairas, veli on sairas ja naisen kaksi poikaa ovat terveitä. On pääteltävä, onko nainen taudinkantaja (vai terve).

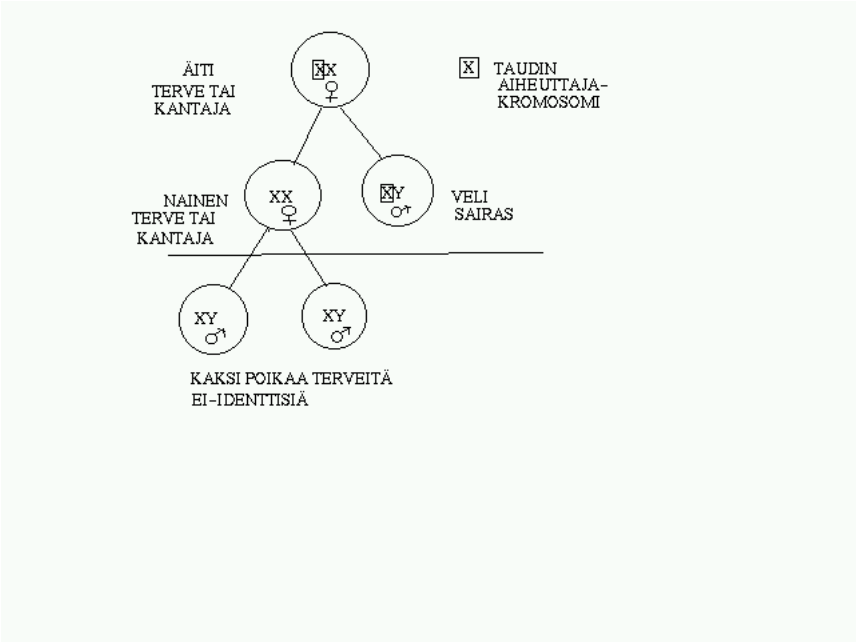
Tiedetään välittömästi (ks. kaavio), että naisen äiti on kantaja. Merkitään $\theta = 1$ jos nainen on kantaja, $\theta = 0$ jos terve (ja ei-kantaja). Olkoon *priori* $\mathsf{P}(\theta = 1) = \frac{1}{2}$.

Aineisto: Merkitään $y_i = 1$ jos poika i on sairas, $y_i = 0$ jos terve, $i = 1, 2$. Havaitaan $y = (0, 0)$.

Uskottavuus:

$$\begin{aligned} \mathsf{P}(y_1 = 0, y_2 = 0|\theta = 1) &= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \\ \mathsf{P}(y_1 = 0, y_2 = 0|\theta = 0) &= 1 \end{aligned}$$

Kaavio perimästä:



Posteriori saadaan laskemalla:

$$\begin{aligned} & \mathsf{P}(\theta = 1 | y = (0,0)) \\ = & \frac{\mathsf{P}(y = (0,0)|\theta = 1)\mathsf{P}(\theta = 1)}{\mathsf{P}(y = (0,0)|\theta = 1)\mathsf{P}(\theta = 1) + \mathsf{P}(y = (0,0)|\theta = 0)\mathsf{P}(\theta = 0)} \,. \\ = & \frac{1/4 \cdot 1/2}{1/4 \cdot 1/2 + 1 \cdot 1/2} = \frac{1}{5} = 0.20 \,. \end{aligned}$$

Nähdään, että terveiden poikien määrä laskee posterioritodennäköisyyttä.

Esim. 2 (jatkoa). Muotoillaan esimerkin kysymys bayesiläisenä hypoteesintestauksena, missä $H_0 : \theta = 0$, $H_1 : \ \theta = 1$.

Tällöin vedonlyöntisuhteet ovat:

$$\begin{aligned} \frac{\pi_0}{\pi_1} &= \frac{\mathsf{P}(H_0)}{\mathsf{P}(H_1)} = 1 \\ \frac{p_0}{p_1} &= \frac{\mathsf{P}(H_0|y)}{\mathsf{P}(H_1|y)} = \frac{0.80}{0.20} = 4.0 \end{aligned}$$

ja Bayes-tekijä on

$$B = \frac{p_0/p_1}{\pi_0/\pi_1} = 4.0\,.$$

On siis positiivista näyttöä siitä, että nainen ei ole kantaja.

Huomaa, että voidaan laskea hypoteesien posterioritodennäköisyydet: $p_0 = \mathsf{P}(H_0|y) = 0.80$ ja $p_1 = \mathsf{P}(H_1|y) = 0.20$. Nämä ovat todellakin todennäköisyyksiä. On muistettava, että klassinen P-arvo on ihan jotain muuta.

Tässä tilanteessa hypoteesintestaus on järkevää, koska on vain kaksi toisistaan erillistä mallia (pistehypoteesia), joita verrataan. Mallien välissä ei ole muita vaihtoehtoja.

Luku 9

Johdatus moniparametrisiin malleihin

9.1 Kiusaparametrien eliminointi

Olkoon tuntemattoman suureen (usein parametrin) θ ulottuvuus (dimensio) suurempi kuin 1, merk. $\dim(\theta) = p > 1$.

Kaikki parametrit eivät välttämättä kiinnosta, mutta niitä tarvitaan, jotta mallin rakenne olisi järkevä. Tällaisia "ylimääräisiä" parametreja sanotaan *kiusaparametreiksi* (nuisance). Merkitään $\theta = (\theta_1\ ,\ \theta_2)$, missä θ_1 koostuu kiinnostavista parametreista ja θ_2 kiusaparametreista.

Esim. 1. Normaalijakauman tapauksessa, $y|\mu, \sigma^2 \sim \mathrm{N}(\mu, \sigma^2)$, kiinnostus kohdistuu usein sijaintiparametriin μ hajontaparametrin σ^2 ollessa kiusaparametri. Tällöin $\theta = (\theta_1, \theta_2)$, missä $\theta_1 = \mu$, $\theta_2 = \sigma^2$.

Oletetaan, että aineisto on y ja parametrivektori $\theta = (\theta_1, \theta_2)$. Jos θ_2 on kiusaparametri, meitä ei niinkään kiinnosta yhteisposteriori $p(\theta_1, \theta_2|y)$ vaan *reunaposteriori* $p(\theta_1|y)$.

Reunaposteriori voidaan laskea kahdella tavalla:

1. Yhteisposteriorista, joka on

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)\, p(\theta_1, \theta_2)\,.$$

Tällöin reunaposteriori saadaan integroimalla kiusaparametri ulos, ts.

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)\, \mathrm{d}\theta_2.$$

2. Yhteisposteriori voidaan jakaa tekijöihin seuraavasti:

$$p(\theta_1, \theta_2|y) = p(\theta_1|\theta_2, y)\, p(\theta_2|y)\,,$$

jolloin

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)\, p(\theta_2|y)\, \mathrm{d}\theta_2\,.$$

Simuloitaessa posteriorijakaumaa voidaan simuloida parametrivektoreita $\theta = (\theta_1, \theta_2)$ posteriorijakaumasta ja yksinkertaisesti "tiputtaa" ei-kiinnostava θ_2 pois. Jos käytetään kohdassa 2 esitettyä tekijöihin jakoa, voidaan ensin generoida θ_2 sen reunaposteriorista $p(\theta_2|y)$ ja sitten θ_1 ehdollisesta posteriorista $p(\theta_1|\theta_2, y)$.

9.2 Normaalinen otos

Oletetaan seuraavaksi, että aineisto y sisältää n havaintoa normaalijakaumasta $N(\mu, \sigma^2)$, missä μ ja σ^2 ovat tuntemattomia. Asetetaan priorijakauma

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1},$$

joka on tasajakauma parametrivektorille $(\mu, \log(\sigma^2))$. Tällöin posteriorijakauma on

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\ &= (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right]\right) \\ &= (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]\right), \end{aligned}$$

missä

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

on otosvarianssi. Tyhjentävät otossuureet ovat $\bar{y}$ ja s^2.

Posteriorijakauman faktorointi.

Posteriorijakauma voidaan esittää tulona $p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y)$. Aiemmin johdettiin, että kun μ:n priorijakauma on tasainen ja σ^2 tunnettu, μ:n posteriorijakauma on

$$\mu|\sigma^2, y \sim N(\bar{y}, \sigma^2/n). \tag{9.1}$$

Reunaposteriori σ^2:lle

Reunaposteriorijakauma $p(\sigma^2|y)$ saadaan yhteisjakaumasta integroimalla μ:n yli:

$$\begin{aligned}
p(\sigma^2|y) &\propto \int_{-\infty}^{\infty} (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu \\
&\propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \int_{-\infty}^{\infty} \left(-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right) d\mu \\
&\propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{2\pi\sigma^2/n} \\
&\propto (\sigma^2)^{-(\frac{n-1}{2}+1)} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right),
\end{aligned}$$

joka on käänteinen gammajakauma InvGamma$((n-1)/2, (n-1)s^2/2)$.

Jakaumaa kutsutaan myös *skaalatuksi käänteiseksi χ^2-jakaumaksi*. Jos $x \sim \text{Gamma}(\nu/2, 1/2)$, niin $x \sim \chi^2(\nu)$ (χ^2-jakauma vapausastein ν). Lisäksi $1/x$ noudattaa käänteistä gammajakaumaa InvGamma$(\nu/2, 1/2)$ eli käänteistä χ^2-jakaumaa Inv-$\chi^2(\nu)$. Skaalattu versio edellisestä on $\nu s^2/x$, joka noudattaa *skaalattua* käänteistä χ^2-jakaumaa Inv-$\chi^2(\nu, s^2)$.

Siis reunaposteriori σ^2:lle on

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2).$$

Huomaa, että tämä tulos on yhtenevä vastaavan otantateorian tuloksen $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ kanssa, joskin otantateorian tapauksessa μ ja σ^2 ajatellaan kiinteinä ja s^2 satunnaisena.

Reunaposteriorijakauma μ:lle

Vastaavalla tavalla voimme johtaa μ:n reunaposteriorijakauman. Merkitsemällä $A = (n-1)s^2 + n(\mu - \bar{y})^2$ ja täydentämällä integrandi käänteisen gamma-jakauman tiheysfunktioksi saamme

$$\begin{aligned}
p(\mu|y) &\propto \int_0^\infty (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}A\right) d\sigma^2 \\
&\propto \frac{\Gamma(n/2)}{(A/2)^{n/2}} \int_0^\infty \frac{(A/2)^{n/2}}{\Gamma(n/2)} (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}A\right) d\sigma^2 \\
&\propto [(n-1)s^2 + n(\mu-\bar{y})^2]^{-n/2} \\
&\propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}.
\end{aligned}$$

Tämä on Studentin t-jakauma vapausastein $n-1$, sijaintiparametrilla $\bar{y}$ ja asteikkoparametrilla $s/\sqrt{n}$, merk. $t_{n-1}(\bar{y}, s^2/n)$.

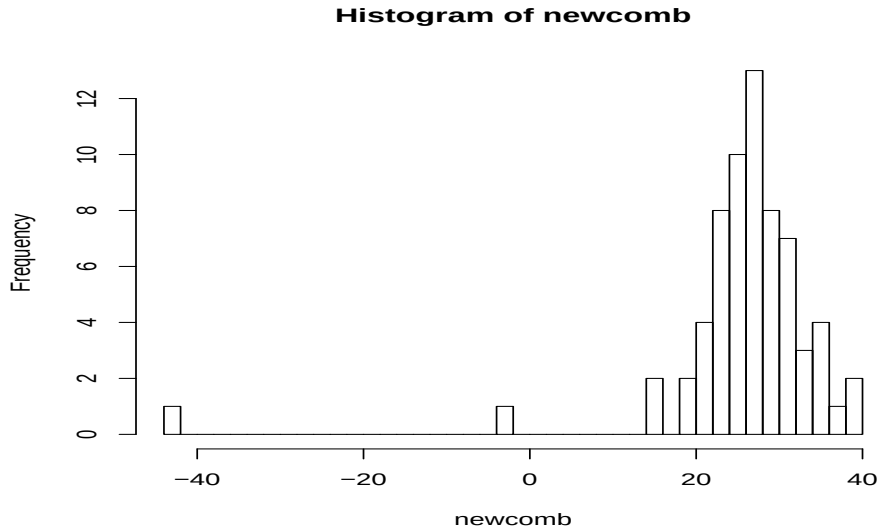
Toisin sanoen, kun parametrivektorilla $(\mu, \log(\sigma^2))$ on tasainen priorijakauma,

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} \mid y \sim t_{n-1}.$$

Tämä tulos on jälleen sopusoinnussa otantateorian kanssa. Otantateoriassa kuitenkin pidetään otossuureita $\bar{y}$ ja s^2 satunnaisina ja parametreja μ ja σ^2 kiinteinä.

Posterioriennustejakauma tulevalle havainnolle

Tulevan havainnon posterioriennustejakauma voidaan esittää sekajakaumana $p(\tilde{y}|\bar{y}) = \int\int p(\tilde{y}|\mu, \sigma^2, y) p(\mu, \sigma^2|y) d\mu d\sigma^2$. Jakaumasta voidaan simuloida generoimalla ensin μ ja σ^2 niiden posteriorijakaumasta ja sen jälkeen $\tilde{y}$ jakaumasta $N(\mu, \sigma^2)$. Voidaan myös osoittaa, että $\tilde{y}|y \sim t_{n-1}(\bar{y}, (1 + \frac{1}{n})s^2)$.



Esim. 1. Valon nopeuden estimointi. Simon Newcomb teki 1882 kokeen, jossa hän mittasi valon nopeutta. Hän mittasi ajan, joka kuluu, kun valo kulkee 7442 metriä. Kuviossa havainnot on ilmoitettu poikkeamina 24800 nanosekunnista. Kuviossa voidaan havaita kaksi hyvin poikkeavaa havaintoa, joten normaalijakauma ei ole hyvä malli ainakaan aineistolle sellaisenaan. Harjoituksen vuoksi kuitenkin oletamme, että havainnot ovat $N(\mu, \sigma^2)$-jakautuneita. Keskiarvo 66 havainnolle on $\bar{y} = 26.2$ ja otoshajonta $s = 10.8$. Kun oletamme tavanomaisen epäinformatiivisen priorijakauman $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$, saamme 95% posteriorivälikSi $[\bar{y} \pm t_{0.975;65} s/\sqrt{66}] = [23.6, 28.8]$. Tämä väli ei sisällä oikeana pidettävää arvoa 33.0, mikä ei ole yllättävää, kun ottaa huomioon mallin huonouden.

Posteriorijakaumaa voitaisiin simuloida poimimalla ensin σ^2 jakaumasta Inv-$\chi^2(65, s^2)$ (eli $\sigma^2 \sim 65s^2/\chi^2_{65}$) ja sen jälkeen μ jakaumasta $N(26.2, \sigma^2/66)$.

9.3 Kahden normaalipopulaation odotusarvojen vertailu

Oletetaan, että $x_1, \ldots, x_m$, missä $x_i|\lambda, \phi \sim \mathrm{N}(\lambda, \phi)$, on riippumaton otos ehdolla $\lambda,\ \phi$. Vastaavasti oletetaan, että $y_1, \ldots, y_n$, missä $y_i|\mu, \psi \sim \mathrm{N}(\mu, \psi)$, on toinen, riippumaton otos ehdolla $\mu,\ \psi$. Molemmat otokset oletetaan toisistaan riippumattomiksi.

Koska kiinnostuksen kohteena on odotusarvojen λ ja μ vertailu, päättely voidaan tehdä erotuksen $\delta = \lambda - \mu$ perusteella. Tilanne voidaan jakaa neljään tapaukseen (kuten "klassisessa" tilastotieteessä – nyt tulee peruskurssiasiaa!):

1. parivertailu (käytetään myös nimityksiä "riippuvat otokset", "verrannolliset parit"),
2. riippumattomat otokset, varianssit tunnetut,
3. varianssit yhtäsuuret mutta tuntemattomat,
4. varianssit tuntemattomat.

Näistä kohta 4 on jakaumateoreettisesti vaikea mutta voidaan käsitellä myöhemmin esiteltävällä simulointimenetelmällä helposti.

1. **Parivertailu.** Tällöin $m = n$ ja (x_i, y_i) muodostavat "verrannolliset parit" ollen usein mittauksia samoista yksilöistä.

   Tilanne palautuu yhden otoksen tapaukseen. Merkitään $z_i = x_i - y_i$ ja $\delta = \lambda - \mu$. Tällöin

   $$z_i|\delta, \omega \ \sim \ \mathrm{N}(\delta, \omega)$$

   missä $\omega = \phi + \psi$ (koska havainnot x_i ja y_i ovat ehdollisesti riippumattomia).

2. **Riippumattomat otokset, varianssit tunnetut.** Yksinkertaisin tilanne on se, jossa λ ja μ ovat riippumattomia a priori ja $p(\lambda) \propto 1$, $p(\mu) \propto 1$. Tällöin

   $$\begin{aligned} \lambda|x &\sim \mathrm{N}(\bar{x}, \frac{\phi}{m}), \\ \mu|y &\sim \mathrm{N}(\bar{y}, \frac{\psi}{n}), \\ \delta|x, y &\sim \mathrm{N}(\bar{x} - \bar{y}, \frac{\phi}{m} + \frac{\psi}{n}). \end{aligned}$$

3. **Varianssit yhtäsuuret mutta tuntemattomat.** Oletetaan, että $\psi = \phi$ ja että priori on epäinformatiivinen

   $$p(\lambda, \mu, \phi) \propto \frac{1}{\phi}\,.$$

Tällöin posteriori on

$$p(\lambda, \mu, \phi|x, y) = p(\lambda, \mu, \phi)\, p(x|\lambda, \phi)\, p(y|\mu, \phi)\,.$$

Suorien laskujen jälkee tästä saadaan

$$p(\delta, \phi|x, y) = p(\phi|s^2)\, p(\delta|\bar{x} - \bar{y}, \phi)\,,$$

missä $p(\delta|\bar{x} - \bar{y}, \phi) = \mathrm{N}(\delta|\bar{x} - \bar{y}, \phi\,(\frac{1}{m} + \frac{1}{n}))$.

Posteriori on

$$\left.\frac{\delta - (\bar{x} - \bar{y})}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}\right| \; x, y \sim \mathrm{t}_{n+m-2},$$

missä

$$s^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m + n - 2}\,.$$

4. **Varianssin erisuuret ja tuntemattomat.** Oletetaan epäinformatiivinen priori

$$p(\lambda, \mu, \phi, \psi) \propto \frac{1}{\phi\psi}.$$

Tällöin

$$\begin{aligned} \phi|x &\sim \text{Inv-}\chi^2(m-1, s_x^2), \\ \psi|y &\sim \text{Inv-}\chi^2(n-1, s_y^2), \\ \delta|\phi, \psi, x, y &\sim N(\bar{x} - \bar{y}, \frac{\phi}{m} + \frac{\psi}{n}). \end{aligned}$$

Nyt posteriorijakaumaa voidaan simuloida seuraavasti: simuloidaan ensin ϕ ja ψ reunaposteriorijakaumistaan ja sitten δ ehdollisesta posteriorista $\{\delta|\phi, \psi, x, y\}$.

Esim 2. Käenmunat. Simulointi yksinkertaistaa tämänkin tilanteen käytännön data-analyysissa. Esimerkkiaineistona on kahden lintulajin (rautiainen, ruokokerttunen) pesästä poimittujen käenmunien läpimitat [mm]:

```
rautiainen     22.0 23.9 20.9 23.8 25.0 24.0 21.7 23.8 22.8 23.1
ruokokerttunen 23.2 22.0 22.2 21.2 21.6 21.9 22.0 22.9 22.8
```

Aineiston tunnuskuluja (laskettuna R-ympäristössä):

```
> x <- c(22.0,23.9,20.9,23.8,25.0,24.0,21.7,23.8,22.8,23.1)
> y <- c(23.2,22.0,22.2,21.2,21.6,21.9,22.0,22.9,22.8)
>
> round(c(x.mean=mean(x),x.var=var(x),y.mean=mean(y),y.var=var(y)),1)
x.mean  x.var y.mean  y.var
23.1    1.6   22.2    0.4
```

Ilmeisesti tässä tapauksessa ei ole perusteltua olettaa, että populaatioiden varianssit olisivat yhtäsuuret. Siis kyseessä on edellä esitelty tapaus 4. Laskemme seuraavaksi simuloimalla posterioritodennäköisyyden, että rautiaisen munat ovat keskimäärin suurempia kuin ruokokerttusen:

```
> n.sim <- 10000; m <- length(x); n <- length(y)
> phi <- (m-1)*var(x)/rchisq(n.sim,df=m-1)
> psi <- (n-1)*var(y)/rchisq(n.sim,df=n-1)
> delta <- rnorm(n.sim,(mean(x)-mean(y)),phi/m+psi/n)
> mean(delta>0)
[1] 0.9937
```

9.4 Multinomimalli

Multinomijakauma saadaan yleistyksenä binomijakaumasta, kun riippumattomien satunnaiskokeiden mahdollisia lopputuloksia on enemmän kuin kaksi. Kun toistetaan n riippumatonta satunnaiskoetta, joissa on k vaihtoehtoa ja $\theta = (\theta_1, ..., \theta_k)$ ilmoittaa eri vaihtoehtojen todennäköisyydet ja havaintovektori $y = (y_1, ..., y_k)$ ilmoittaa eri vaihtoehtojen toteutumisten lukumäärät, otantajakauma on

$$p(y|\theta) = \begin{pmatrix} n \\ y_1 \quad \cdots \quad y_k \end{pmatrix} \theta_1^{y_1} \cdots \theta_k^{y_k},$$

missä $\sum_{j=1}^{k} \theta_j = 1$ ja $\sum_{j=1}^{k} y_j = n$. Tätä kutsutaan *multinomijakaumaksi*.

Uskottavuusfunktiossa voidaan tiputtaa alussa oleva multinomikerroin pois:

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}.$$

Konjugaattinen priorijakauma on Beta-jakauman moniulotteinen yleistys Dirichlet:n jakauma Dirichlet$(\alpha_1, ..., \alpha_k)$, jonka tiheysfunktio on

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1},$$

missä $\theta_j \in [0,1]$ ja $\sum_{j=1}^{k} \theta_j = 1$. Tasainen priorijakauma saadaan asettamalla $\alpha_j = 1$ kaikilla j. Jos asetetaan $\alpha_j = 0, \quad j = 1, .., k$, saadaan epäaito priorijakauma, joka on tasainen $\log(\theta_j)$:lle. Dirichlet'n jakauma palautuu beeta-jakaumaan Beta(α_1, α_2), kun $k = 2$.

Konjugaattipriorin tapauksessa *posteriori* on

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)\, p(\theta) \\ &\propto \prod_{i=1}^{k} \theta_i^{y_i} \cdot \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1}, \end{aligned}$$

joka on Dirichlet$(\alpha_1 + y_1, \ldots, \alpha_k + y_k)$.

Esim. 3. Puolueiden kannatus. Radiouutisten syyskuussa 2001 tekemässä haastattelututkimuksessa kysyttiin puolueiden kannatusta. Haastateltavana oli 1962 henkilöä. Tulos (laskettuna annetuista kannatusprosenteista) oli

SDP	Kesk	Kok	Vihr	Vas	muu	yht
471	453	396	243	177	222	1962
24.0	23.1	20.2	12.4	9.0	11.2	100%

Valitaan prioriksi Dirichlet$(1, 1, 1, 1, 1, 1)$. Reunaposteriorien kuvaus:

```
        Mean       SD    2.5%     50%  97.5%
SDP  0.24000 0.009718 0.22170 0.23980 0.2592
Kesk 0.23070 0.009305 0.21270 0.23060 0.2491
Kok  0.20180 0.008961 0.18460 0.20160 0.2198
Vihr 0.12400 0.007439 0.10990 0.12380 0.1390
Vas  0.09036 0.006496 0.07802 0.09035 0.1032
Muu  0.11330 0.007145 0.09971 0.11310 0.1275
```

Edellä oleva posteriori on laskettu käyttäen pakettiin MCMCpack kuuluvaa funktiota rdirichlet. Funktio mcmc muuntaa simulointituloksen luokkaan mcmc kuuluvaksi olioksi, jolloin geneerinen funktio summary antaa haluttuja yhteenvetotuloksia.

```
library(MCMCpack)
puol <- c("SDP","Kesk","Kok", "Vihr","Vas","Muu" )
#eri puolueiden kannattajien lkm otoksessa noin
y <- c(471,453,396,243,177,222)

#Posteriorijakauma
#Dirichlet parametreilla y+1
theta.post <- rdirichlet(10000,y+1)
colnames(theta.post) <- puol
b <- summary(mcmc(theta.post))
signif(cbind(b$statistics[,1:2],b$quantiles[,c(1,3,5)]),4)
```

Mikä on posterioritodennäköisyys, että SDP:n kannatus on suurempi kuin a) Keskustan b) Kokoomuksen? Käytetään edelleen samaa Dirichlet'n prioria, jossa $\alpha_i = 1$ kaikille i. Vastaus saadaan seuraavalla koodilla:

```
#Millä todennäköisyydellä SDP:n kannatus on suurempi kuin Keskustan?
mean(theta.post[,"SDP"]>theta.post[,"Kesk"])
#[1] 0.7162

#Millä todennäköisyydellä SDP:n kannatus on suurempi kuin Kokoomuksen?
mean(theta.post[,"SDP"]>theta.post[,"Kok"])
#[1] 0.9946
```

Luku 10

Usein käytettyjä perusmalleja

Bayes-tilastotieteen laskentaa varten on olemassa erityisohjelmistoja, kuten BUGS, JAGS ja Stan. Näitä voidaan käyttää R:stä käsin sopivilla R-paketeilla. Pythonille on olemassa Bayes-laskentaa varten aitoja Python-paketteja, kuten PyMC3, Edward ja Pyro. R-paketissa MCMCpack on implementoitu muutamia yleisimpiä tilastollisia malleja, ja seuraavissa esimerkeissä käytetään tämän paketin funktioita.

10.1 Regressiomalli

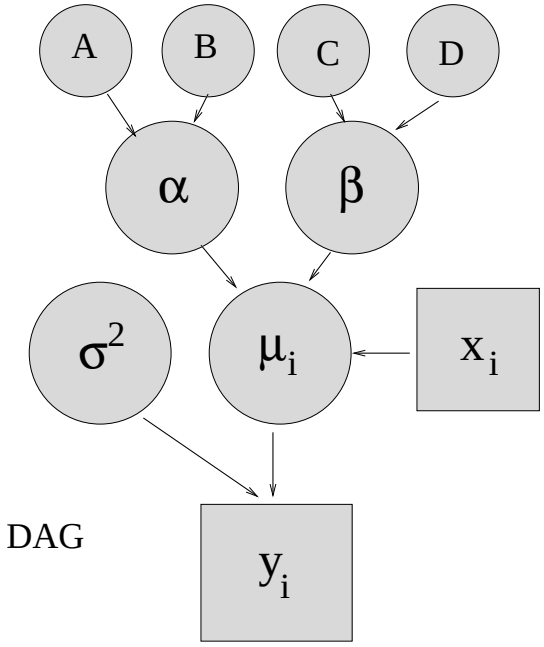
Tarkastellaan yhden selittäjän lineaarista regressiomallia

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

missä virhetermit $\epsilon_i,\ i = 1, ..., n$ ovat $N(0, \sigma^2)$-jakautuneita ja riippumattomia ehdolla σ^2. Voimme täydentää mallin Bayes-malliksi priorioletuksilla seuraavasti:

- $y_i|\mu_i, \sigma^2 \sim \mathrm{N}(\mu_i, \sigma^2)$,
- $\mu_i = \alpha + \beta x_i$,
- $\alpha \sim \mathrm{N}(A, B)$,
- $\beta \sim \mathrm{N}(C, D)$,
- $\sigma^2 \sim \mathrm{InvGamma}(E, F)$.

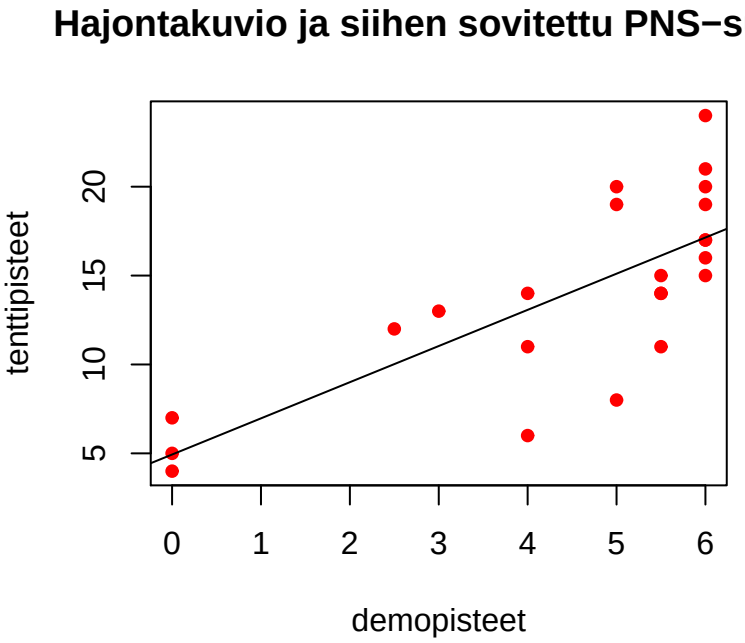
Mallia voidaan havainnollistaa suunnatulla syklittömällä verkolla (DAG, directed acyclic graph).



Esim. 1. Harjoitusten vaikutus tentissä osaamiseen Bayes-tilastotieteen kurssilla kl. 2007.

Havaintoaineisto on:

demo	tentti	demo	tentti	demo	tentti
6.0	15	0.0	7	5.5	11
5.0	20	6.0	17	5.5	14
6.0	16	3.0	13	2.5	12
5.5	15	5.0	19	4.0	6
5.5	14	4.0	14	0.0	5
6.0	21	6.0	19	6.0	20
6.0	24	4.0	11	0.0	4
6.0	17	5.0	8		



Sovitetaan regressiomalli

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

missä y_i on tenttipistemäärä ja x_i demopistemäärä. Priorijakaumilla $\alpha \sim \mathrm{N}(5, 1)$, $\beta \sim \mathrm{N}(2, 1)$ ja $\sigma^2 \sim \mathrm{InvGamma}(0.001, 0.01)$ saamme seuraavat tulokset:

```
             Mean    SD 2.5%   50% 97.5%
(Intercept)  5.00 0.869 3.30  5.01  6.72
x            2.02 0.225 1.57  2.02  2.47
sigma2      13.90 4.650 7.56 13.00 25.30
```

Tulokset saadaan seuraavalla R-koodilla:

```
a2 <- MCMCregress(y~x,b0=c(5,2),B0=diag(2),c0=2*0.001,d0=2*0.01)
a2.s <- summary(a2)
signif(cbind(a2.s$statistics[,c(1,2)],a2.s$quantiles[,c(1,3,5)]),3)
```

Oleellinen tulos (mallin perusteella) on se, että pisteen arvoinen demo-osallistuminen näkyy tenttisuorituksessa kahtena pisteenä. Jonkinlainen kynnyspistemäärä on 5 ja siihen yltää ne, jotka eivät ole demopisteitä saaneet. Mallista voi nyt laskea kiinnostavien hypoteesien todennäköisyyksiä!

Vertailun vuoksi "klassisen" regressioanalyysin tulos:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9349     1.8516   2.665   0.0145 *
dat$demo      2.0341     0.3796   5.359 2.58e-05 ***
---
Residual standard error: 3.61 on 21 degrees of freedom
Multiple R-Squared: 0.5776,     Adjusted R-squared: 0.5575
F-statistic: 28.71 on 1 and 21 DF,  p-value: 2.583e-05
```

Voimme myös muodostaa posterioriennustejakauman opiskelijalle, jolla 2 demopistettä:

```
ypred <- a2[,"(Intercept)"]+a2[,"x"]*2+rnorm(10000,0,sqrt(a2[,"sigma2"]))
ypred.s <- summary(ypred)
signif(c(ypred.s$statistics[1:2],ypred.s$quantiles[c(1,3,5)]),3)

Mean    SD  2.5%   50% 97.5%
9.09  3.81  1.66  9.06 16.80
```

10.2 Binäärinen regressio

Binäärisessä regressiossa vastemuuttuja voi saada kaksi arvoa, jotka yleensä koodataan 1 "onnistuminen" ja 0 "epäonnistuminen". Onnistumisen todennäköisyyttä θ_i tietyllä selittäjän arvolla x_i voidaan mallintaa esimerkiksi logit-funktion avulla:

$$\text{logit}(\theta_i) = \log \frac{\theta_i}{1-\theta_i} = \alpha + \beta\, x_i.$$

Kun havainnot ovat riippumattomia (ehdolla mallin parametrit) ja havaintoja tietyllä selittäjän arvolla x_i on n_i, onnistumisten lukumäärä y_i noudattaa binomijakaumaa

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i)\,,\ i = 1, \ldots, k.$$

Ehdollisesta riippumattomuudesta seuraa, että uskottavuus on

$$p(y|\alpha, \beta, x, n) \propto \prod_{i=1}^{k} p(y_i|\alpha, \beta, x_i, n_i),$$

missä

$$p(y_i|\alpha,\beta,n_i,x_i) \;\propto\; \theta_i^{y_i}(1-\theta)^{n_i-y_i}$$
$$= \left[\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right]^{y_i}\left[\frac{1}{1+\exp(\alpha+\beta x_i)}\right]^{n_i-y_i}.$$

Jos priorijakauma on $p(\alpha,\beta)$, posteriori on

$$p(\alpha,\beta|y) \propto p(\alpha,\beta)p(y|\alpha,\beta,x,n).$$

Esim. 2. Biologinen eläinkoeaineisto. Tarkastellaan seuraavaa eläinkoeaineistoa:

annostus (log g/ml)	eläinten määrä n_i	kuolleiden määrä y_i
-0.863	5	0
-0.296	5	1
-0.053	5	3
0.727	5	5

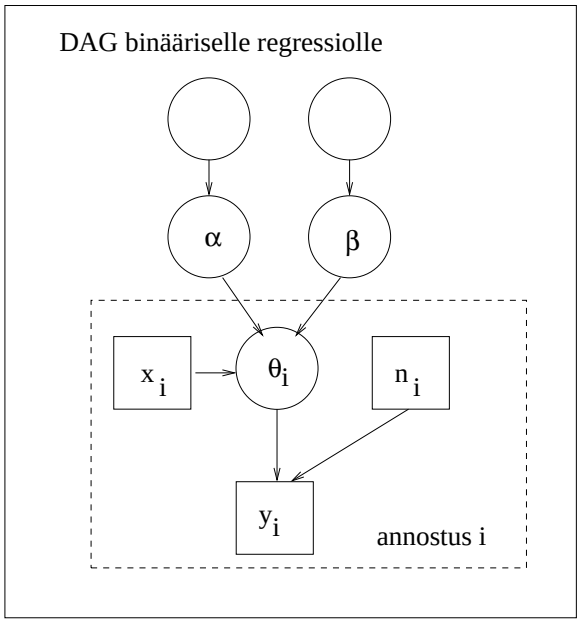
Aineisto koostuu havainnoista (x_i, y_i, n_i), $i = 1,\ldots,4$. Vasteena on kuolleiden määrä y_i ja selittäjänä x_i koe-eläimelle annettu annos jotain lääkettä tai kemiallista yhdistettä. Koe-eläimiä kussakin käsittelyryhmässä on n_i. Muuttujat x_i ja n_i ovat tutkijan kontrolloitavissa. Asetetaan logit-malli

$$\text{logit}(\theta_i) = \log\frac{\theta_i}{1-\theta_i} = \alpha + \beta\, x_i,$$

missä θ_i on kuoleman todennäköisyys annostuksella x_i.

Tutkimusongelmana on, miten annostus x_i vaikuttaa vasteeseen y_i. Tämä taas riippuu todennäköisyydestä θ_i. Siten annostuksen vaikutus kohdistuu juuri parametriin θ_i.

Mallia voidaan kuvata suunnatulla syklittömällä graafilla.



Priori. Oletetaan että α ja β ovat riippumattomia a priori, ja asetetaan epäinformatiiviset priorit $\alpha \sim N(0, 10^6)$ ja $\beta \sim N(0, 10^6)$.

Posteriori. Posteriori voidaan määrittää simuloimalla. Käyttämällä R-paketin MCMCpack funktiota MCMClogit saamme seuraavat tulokset:

```
               Mean     SD      2.5%       50%  97.5%
(Intercept)   1.400 1.100 -5.3e-01   1.30000  3.700
annos        12.000 6.000  3.6e+00 11.00000 26.000
theta1        0.007 0.022  1.9e-09   0.00027  0.063
theta2        0.150 0.120  4.4e-03   0.12000  0.470
theta3        0.650 0.180  2.8e-01   0.66000  0.930
theta4        0.990 0.026  9.3e-01   1.00000  1.000
LD           -0.110 0.092 -2.8e-01  -0.11000  0.094
```

Vertailun vuoksi, su-menetelmä antaa $\alpha = 0.8734$ s.e. 1.040, $\beta = 7.913$ s.e. 5.062, jäännösdevianssi 0.047306, $df = 2$.

LD50. Tällaisten kokeiden yhteydessä yleensä lasketaan LD50, joka on se annostus, joka aiheuttaa 50%:n kuolleisuuden ja ratkaistaan yhtälöstä

$$1 - \theta = \frac{1}{1 + \exp(\alpha + \beta x)} = 0.5$$

eli $x_{LD50} = -\frac{\alpha}{\beta}$. Bayes-estimaatti (posterioriodotusarvo) on $x_{LD50} = -0.1096$, sd $= 0.0935$. (Vastaava su-estimaatti on -0.1140.)

```
library(boot)
Y=c(0,1,3,5)
n=c(5,5,5,5)
x=c(-0.863,-0.296,-0.053,0.727)

#Aineisto "pitkässä" muodossa:
y <- c(0,0,0,0,0,1,0,0,0,0,1,1,1,0,0,1,1,1,1,1)
annos <- rep(x,each=5)

#Bayes-tulokset
b <- MCMClogit(y~annos,b0=0,B=1e-6)
theta <- inv.logit(b[,"(Intercept)"]+b[,"annos"]%*%t(x))
colnames(theta)<-paste("theta",1:4,sep="")
LD <- -b[,"(Intercept)"]/b[,"annos"]
b.s <- summary(mcmc(cbind(b,theta,LD)))
signif(cbind(b.s$statistics[,c(1,2)],b.s$quantiles[,c(1,3,5)]),2)

#Suurimman uskottavuuden estimaatit
b2 <- glm(cbind(Y,n-Y)~x,family=binomial)
summary(b2)
```

Luku 11

Hierarkkinen malli

Moniparametrisissa tilastollisissa malleissa parametrien välillä saattaa olla ongelman rakenteesta aiheutuvia riippuvuussuhteita. Näitä riippuvuussuhteita saattaa olla järkevää mallintaa hierarkkisilla malleilla. Esimerkiksi tutkittaessa sydänhoidon tehokkuutta, voidaan potilaan henkiinjäämistodennäköisyyksiä θ_j eri sairaaloissa pitää otoksena sairaaloiden populaatiossa. Ellei ole käytössä sairaaloita erottavaa informaatiota, parametreja θ_j voidaan pitää vaihdannaisina ja niiden voidaan ajatella olevan satunnaisotos ns. *populaatiojakaumasta*, jolla on omat parametrinsa, ns. *hyperparametrit*. Tämän populaatiojakauman ominaisuuksia voidaan estimoida käyttäen havaintoja y_{ij}, missä indeksit viittaavat j:nnen ryhmän i:nteen yksilöön.

11.1 Parametroidun priorijakauman konstruointi

Esittelemme aluksi esimerkin, jossa populaatiojakaumalle ei tehdä täydellistä todennäköisyysmallia. Sen sijaan käytämme aiemmin tehtyjä tutkimuksia parametrin θ priorijakauman parametrisointiin, mitä kutsutaan *empiiriseksi bayesiläisyydeksi*.

Esim. 1. Kasvaimen riskin estimointi ryhmällä rottia. Tutkittaessa lääkkeiden mahdollista soveltuvuutta sairaanhoidossa niitä kokeillaan säännönmukaisesti jyrsijöille. Oletetaan, että tarkoituksena on estimoida todennäköisyys θ tietyntyyppiselle kasvaimelle (endometrial stromal polyp) tyyppiä F344 olevalle laboratorion naarasrottapopulaatiolle, kun lääkettä ei anneta (kontrolliryhmä). Kokeessa 4 rotalle 14:sta kehittyi kyseisen tyyppinen kasvain. On luonnollista olettaa binomijakauma kasvaintapausten lukumäärälle, kun θ on annettu. Laskujen helpottamiseksi käytämme konjugaattista priorijakaumaa $\theta \sim \text{Beta}(\alpha, \beta)$.

Aikaisemmat kokeet:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Nykyinen koe:
4/14

Taulukko 11.1: Kasvaimen esiintyminen aikaisemmilla kontrolliryhmillä ja nykyisellä rottaryhmällä (Tarone, 1982). Taulukon arvot ovat y_j/n_j: (kasvaimellisten rottien lkm)/ (rottien kokonaislkm)

Kasvaimen todennäköisyys θ vaihtelee populaatiosta toiseen, koska rottapopulaatioissa ja laboratorioiden koeolosuhteissa esiintyy vaihtelua. Mallinnamme tätä vaihtelua siis Beta-jakaumalla. Taulukon 11.1 historiallisen aineiston perusteella voimme yrittää estimoida tämän jakauman parametrit. Aineiston perusteella esiintyvyyden y_j/n_j keskiarvo on 0.136 ja hajonta 0.103. Kun merkitsemme nämä yhtä suuriksi beta-jakauman vastaavien parametrien kanssa voimme ratkaista beta-jakauman parametrit α ja β ja saamme estimaatin $(\alpha, \beta) = (1.4, 8.6)$. Tällöin nykyiselle populaatiolle esiintyvyyden posteriorijakauma on $\theta_{71} \sim \text{Beta}(5.4, 18.6)$, jonka odotusarvo on 0.223 ja hajonta 0.083. Posterioriodotusarvo 0.223 on huomattavasti alhaisempi kuin raakaestimaati 4/14=0.286.

Aikaisemmista kokeista estimoitua priorijakaumaa voitaisiin käyttää myös estimoimaan $\theta_1, ...\theta_{70}$ eli kasvaimen todennäköisyydet aikaisemmille tutkimusryhmille. Tällöin ilmenevät seuraavata periaatteelliset ja käytännölliset ongelmat:

- Kun 70 ensimmäistä koetta käytetään priorijakauman estimointiin ja sitten kullekin ryhmälle estimoidaan θ_j, aineisto tulee käytetyksi kahteen kertaan. Tämä voi aiheuttaa tarkkuuden yliarviointia.
- Piste-estimaatin valinta parametreille (α, β) ei ole yksikäsitteistä. Joka tapauksessa piste-estimaatin käyttö jättää ottamatta huomioon osan posteriorijepävarmuudesta.
- Voimme myös kysyä, onko tarpeen lainkaan estimoida (α, β) etukäteen. Niitä voidaan pitää osana priori-informaatiota, jota ei tarvitse bayesiläisen päättelyn logiikan mukaan tarvitse tuntea ennen aineiston hankintaa.

Näistä ongelmakohdista huolimatta on selvää, että parametrit θ_j kannattaa estimoida yhdessä. Rakennamme seuraavaksi hierarkkisen mallin, jonka avulla samanaikainen estimointi on mahdollista ja edellä esitetyt ongelmakohdat tulevat ratkaistuksi.

11.2 Vaihdannaisuus ja hierarkkisen mallin rakentaminen

Tarkastelemme seuraavaksi kokeita $j = 1, ..., J$, joissa kokeella j on aineisto (vektori) y_j, parametri θ_j ja uskottavuusfunktio $p(y_j|\theta_j)$. Jotkut parametrit voivat olla yhteisiä eri kokeille, esim. σ^2 on yhteinen, jos $\theta_j = (\mu_j, \sigma^2)$ normaalijakaumamallissa.

Ellei ole käytettävissä informaatiota, joka erottaisi kokeet toisistaan, niitä kannattaa pitää vaihdannaisina. Yksinkertaisin tapa mallintaa vaihdannaisuutta on olettaa, että kokeita vastaavat parametrit ovat riippumaton otos populaatiojakaumasta, jolla on tuntematon parametrivektori ϕ. Tällöin

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi).$$

Parametrien θ_j reunajakauma on silloin

$$p(\theta) = \int \left[\prod_{j=1}^{J} p(\theta_j|\phi) \right] p(\phi) d\phi.$$

Vaihdannaisuutta voidaan mallintaa myös silloin, kun on käytössä koetilanteita/ryhmiä erottavaa informaatiota. Jos ryhmän j parametria θ_j selittää muuttujan x arvo x_j, parametrien θ_j riippumattomuus saavutetaan ehdollistamalla x_j:den suhteen:

$$p(\theta_1, ..., \theta_J|x_1, ..., x_J) = \int [\prod_{j=1}^{n} p(\theta_j|\phi, x_j)] p(\phi|x) d\phi,$$

missä $x = (x_1, ..., x_J)$. Tällöin vektorit (θ_j, x_j) ovat vaihdannaisia.

Hierarkisen mallin täysi bayesiläinen käsittely

Populaatiojakauman parametri ϕ on tuntematon ja merkitsemme sen priorijakaumaa $p(\phi)$. Tällöin vektorin (ϕ, θ) yhteispriorijakauma on

$$p(\phi, \theta) = p(\phi)p(\theta|\phi)$$

ja yhteisposteriorijakauma

$$\begin{aligned} p(\phi, \theta|y) &\propto p(\phi, \theta)p(y|\phi, \theta) \\ &= p(\phi, \theta)p(y|\theta), \end{aligned}$$

missä jälkimmäinen yhtäsuuruus on seurausta siitä, että aineisto y riippuu ainoastaan θ:sta; y riippuu hyperparametrista ϕ ainoastaan θ:n välityksellä.

Tähän mennessä olemme olettaneet ϕ:n tunnetuksi; nyt oletamme sen tuntemattomaksi ja asetamme sille oman jakauman, jota kutsumme *hyperpriorijakaumaksi*. Jos ϕ:stä ei ole paljon tietoa, sille voi asettaa epäinformatiivisen priorijakauman. Tällöin on kuitenkin varmistettava, että posteriorijakaumasta tulee aito. Erityisesti hierarkkisten mallien tapauksessa on vaarana, että posteriorijakauma on epäaito, jos hyperparametrin priorijakauma on epäaito.

Posterioriennustejakaumat

Hierarkisten mallien tapauksessa esiintyy kahdenlaisia posterioriennustejakaumia: 1) tulevan havainnon $\tilde{y}$ jakauma, joka liittyy olemassaolevaan parametriin θ_j, ja 2)tulevan havainnon $\tilde{y}$ jakauma, joka liittyy tulevaan parametriin θ_j, joka on poimittu samasta superpopulaatiosta. Merkitsemme tulevia parametrin arvoja $\tilde{\theta}$. Esimerkin rottakokeessa vaihtoehto 1) tarkoittaa uutta rottaa olemassaolevassa kokeessa ja 2) uutta rottaa uudessa kokeessa.

11.3 Laskenta hierarkkisten mallien yhteydessä

Laskennassa voidaan käyttää moniparametristen mallien yhteydessä esitettyjä yleisiä periaatteita, kun käsitellään hyperparametria ϕ haittaparametrina. Laskenta voi kuitenkin olla käytännössä vaikeampaa, sillä parametrien lukumäärä voi olla suuri hierarkkisilla malleilla. Laskentoja helpottaa, jos populaatiojakauma $p(\theta|\phi)$ on konjugaattinen uskottavuudelle $p(y|\theta)$. Ei-konjugaattiset hierarkkiset mallit voivat vaatia käyttämään edistyneempiä laskentamenetelmiä, jotka esitellään myöhemmin.

Esim. 2. Rottien kasvaimet (jatkoa). (Ks. esimerkkikoodi hierExample.r). Oletamme jälleen, että rottakokeiden $j = 1, ..., J$, $J = 71$, tulokset noudattavat binomijakaumia

$$y_j \sim \text{Bin}(n_j, \theta_j),$$

missä rottien lukumäärät n_j ovat tunnettuja. Parametrien θ_j ajatellaan olevan riippumattomia havaintoja beta-jakaumasta

$$\theta_j \sim \text{Beta}(\alpha, \beta).$$

Tällöin kaikkien parametrien yhteisposteriorijakauma on

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta) \\ &\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}. \quad (11.1) \end{aligned}$$

Kun α ja β on annettu, parametrivektorin θ posteriorijakauma on

$$p(\theta|\alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}. \quad (11.2)$$

Tämän jälkeen (α, β):n reunaposteriorijakauma saadaan jakamalla yhteisposteriorijakauma (11.1) ehdollisella posteriorijakaumalla (11.2):

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}.$$

Yleisesti käytetty uudelleenparametrointi beta-jakauman parametreille on $\text{logit}(\frac{\alpha}{\alpha+\beta}) = \log(\frac{\alpha}{\beta})$ ja $\log(\alpha+\beta)$. Osoittautuu, että aidon posteriorijakauman tuottaa esimerkiksi epäinformatiivinen priorijakauma

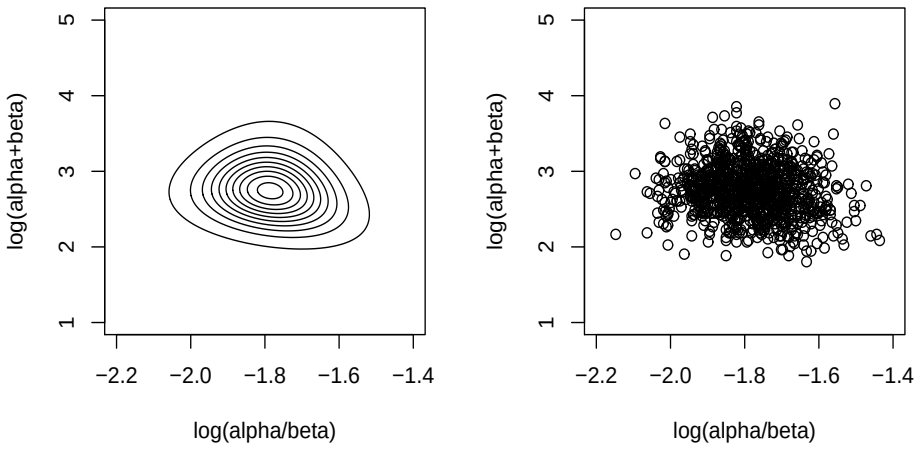
$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2},$$

jota vastaa muunnetulla asteikolla

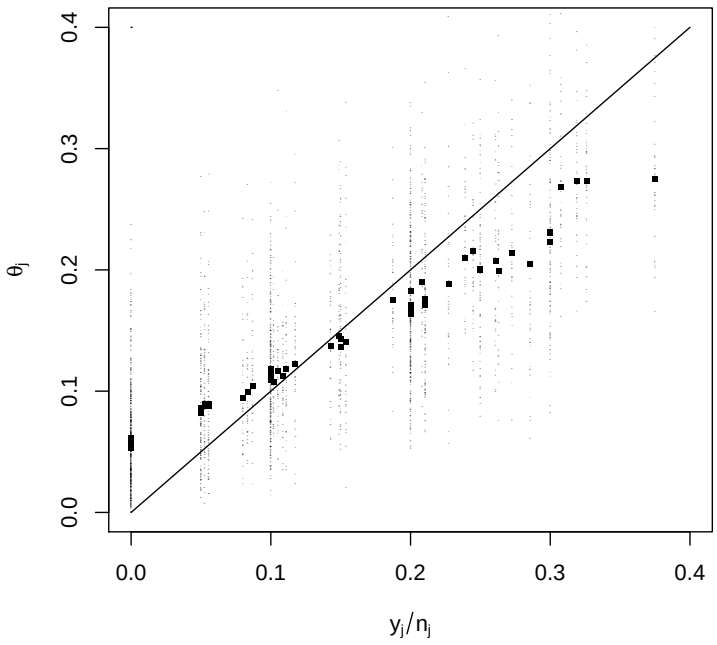
$$p\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2}.$$

Nyt voimme tehdä posteriorisimulointeja diskretoimalla muunnettujen parametrien kaksiulotteinen jakauma. Ylivuotojen välttämiseksi voi olla syytä laskea taulukkoon logaritmoidun posteriorijakauman arvot, vähentää näistä arvoista moodi ja eksponoida tämän jälkeen.

Kun on simuloitu esim. 1000 havaintoa, ne voidaan muuntaa alkuperäisiksi hyperparametreiksi (α, β). Tämän jälkeen voidaan simuloida populaatioparametrit θ_j ehdollisista jakaumista $\theta_j|\alpha, \beta, y \sim \text{Beta}(\alpha+y_j, \beta+n_j-y_j)$. Kuviossa 11.2 on nähtävillä simulaatioiden tulokset. Populaatioiden posteriorimediaanit ovat siirtyneet raakaestimaateista y_j/n_j kohti populaatiojakauman odotusarvoa 0.14.



Kuvio 11.1: a) Korkeuskäyräkuvio hyperparametrien reunaposteriorijakaumalle esimerkin rottakokeessa. Käyrien korkeudet suhteessa posteriorijakauman moodiin ovat 0.05, 0.15...,0.95. b) Pisteparvi 1000 simuloidulle havainnolle.



Kuvio 11.2: Rottien kasvainten esiintymistodennäköisyyksien θ_j simuloidut posteriorijakaumat havaittujen esiintyvyyksien y_j/n_j eri arvoilla. Kuvioissa on merkitty pienillä pisteillä populaatioparametrien θ_j simuloituja arvoja ja isoilla pisteillä simuloitujen jakaumien mediaaneja. Kuvion 45° suora vastaa raakaestimaatteja $\hat{\theta} = y_j/n_j$.

Luku 12

Mallikritiikki

Tähän mennessä on käsitelty mallin konstruointia ja tulkintaa sekä posteriorin laskemista. Sen sijaan emme ole puuttuneet *mallikritiikkiin*, joka on tärkeä osa Bayes-tilastotiedettä. Mallikritiikkiin sisältyy *yhteensopivuuden tutkiminen* ja *herkkyysanalyysi*. Yhteensopivuuden tutkiminen tarkoittaa *mallin oletusten* sopivuuden tarkastelua aineiston kannalta. Herkkyysanalyysi voidaan karakterisoida seuraavasti: Useat todennäköisyysmallit "sopivat" aineistoon, mutta *kuinka paljon posteriori-päättely muuttuu, kun siirrytään mallista toiseen?*

Mallikritiikki "klassisessa" tilastotieteessä

"Klassisessa" tilastotieteessä mallikritiikki yleensä tarkoittaa

- mallin yhteensopivuuden testausta,
- jäännösanalyysia,
- mallien vertailua (mallinvalintakriteerein).

Yhteensopivuuden ("goodness of fit") testaus johtaa usein asymptoottiiseen χ^2 -testiin (lineaarisissa malleissa F-testiin). Tyypillisesti yhteensopivuutta mitataan neliösummalla kuten

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

tai uskottavuusosamäärään perustuvalla mitalla

$$-2\left[\log p(y|\hat{\theta}) - \log p(y|\hat{\theta}^{\text{sat})}\right],$$

missä $\hat{\theta}^{\text{sat}}$ vastaa kyllästettyä (vähärajoitteista) mallia.

Jäännösten tutkiminen tarkoittaa esimerkin avulla seuraavaa: Oletetaan lineaarinen malli

$$y_i = \alpha + \sum_j \beta_j \, x_{ij} + \epsilon_i,$$

missä $\epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)$ riippumattomia. Parametrien estimoinnin jälkeen estimoidaan jäännökset (jokaiselle havaintoyksikölle).

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \sum_j \hat{\beta}_j \, x_{ij}.$$

Jäännösten estimaateista tutkitaan mm. samavarianssisuus, korreloituneisuus ja normaalisuus.

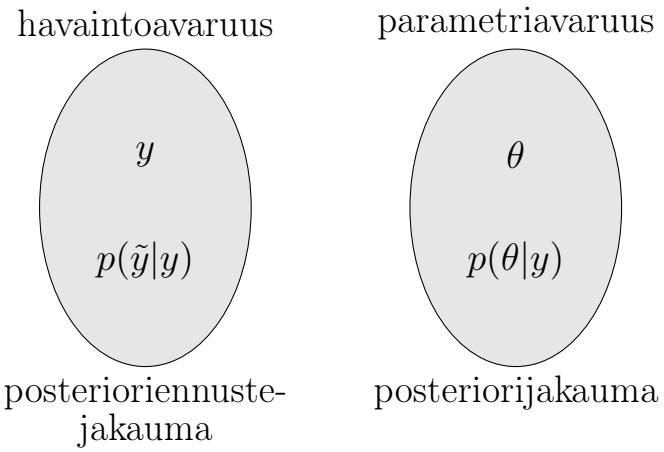
Vertailtaessa malleja käytetään tyypillisesti sellaisia informaatiokriteereitä kuin AIC (Akaiken informaatiokriteeri) ja BIC (Bayesiläinen informaatiokriteeri).

Mallikritiikki Bayes-tilastotieteessä

Klassisessa tilastotieteessä kaikki satunnaisuus on havaintoavaruudessa $\mathcal{Y}$. Jäännösanalyysi on suoraviivaista. Bayes-tilastotieteessä satunnaisuus on parametriavaruudessa Θ. Voisi kuvitella, tästä aiheutuu ongelma mallikritiikille, sillä havainnot ovat avaruudessa $\mathcal{Y}$. Huomaa kuitenkin, että posterioriennustejakauma

$$p(y^{\text{uusi}}|\text{data})$$

on määritelty havaintoavaruudessa $\mathcal{Y}$! Mallikritiikki paljolti perustuu juuri tähän jakaumaan.



Tarjolla on useita menetelmiä. Tarkastellaan seuraavia lähestymistapoja:

1. Sisältöön liittyvä mallin validointi
2. Ulkopuolisen aineiston käyttö mallivalidoinnissa
3. Toistoaineistojen käyttö
4. Prioriennustejakauman käyttö
5. Ristiinvalidointi
6. Jäännöstarkastelut
7. Poikkeamaindeksi (DIC)
8. Bayes-tekijä

Silti mallikritiikissä on edelleen vaikeuksia, varsinkin kompleksisten mallien yhteydessä: Kritiikin pitäisi yltää mallin joka osaan. Usein "hyvät" menetelmät ovat laskennallisesti vaativia.

12.1 Sisältöön liittyvä mallin validointi

Sisältöön liittyvää mallin validointia tehdään tarkastelemalla mallin antamien tuloksia (myös ei-havaittujen muuttujien osalta) substanssin eli asiasisällön kannalta. Ovatko analyysin tulokset järkeviä maalaisjärjellä ajateltuina? Entä ovatko ne yhteensopivia sovellustieteen teorioiden kanssa?

12.2 Validointiaineiston käyttö

Jos aineisto on suuri, niin silloin se on mahdollista jakaa

a) *opetusaineistoon* (learning data), jota käytetään mallin rakentamisessa sekä posteriorin laskemisessa, ja

b) *validointiaineistoon* (validation data), jota käytetään mallin hyvyyden ja ennustuskyvyn tutkimiseen.

Oletetaan, että mallin rakentaminen ja posteriorin laskeminen perustuu aineistoon y. Tämän lisäksi on olemassa validointiaineisto y^{uusi}, jota ei ole käytetty mallin rakentamisessa (eli posteriorin konstruoinnissa).

Työkaluna on *posteriorienustejakauma*

$$p(y^{\text{uusi}}|y) = \int p(y^{\text{uusi}}|\theta)p(\theta|y)\mathrm{d}\theta,$$

joka perustuu havaittuun aineistoon y. Ideana on laskea $p(y^{\text{uusi}}|y)$ ja verrataan uuden aineiston havaittuja arvoja tähän jakaumaan.

12.3 Toistettujen aineistojen käyttö

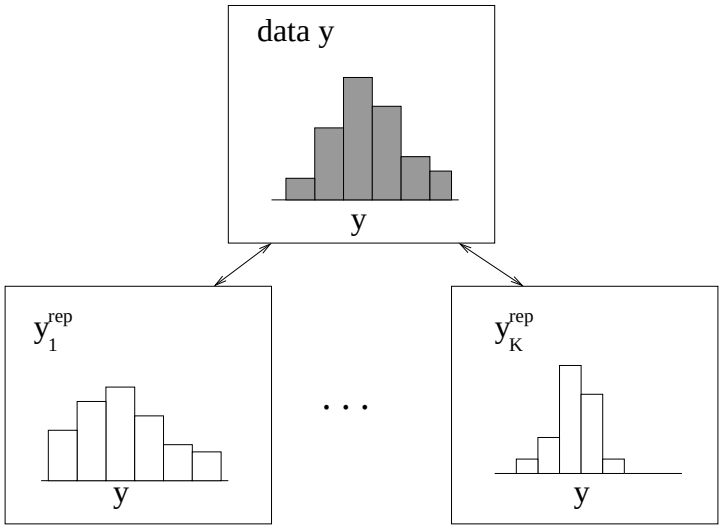
Oletetaan, että validointiaineistoa ei ole käytössä. Aineisto on y ja posteriorijakauma $p(\theta|y)$.

Olkoon y^{rep} "toistettu aineisto" eli replikaatti: se on aineisto, joka saadaan ennustamalla aineistoa mallilla. Esimerkiksi, jos mallissa on selittävä muuttuja x, niin aineiston y^{rep} konstruoinnissa mallin avulla on käytetty niitä x:n arvoja, joita on havaitussa aineistossa y.

Mallikritiikissä työskennellään y^{rep}:n jakaumalla, joka saadaan tämän hetken tiedon perusteella. Se on posterioriennustejakauma

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)\, p(\theta|y)\, \mathrm{d}\theta.$$

On mahdollista simuloida MCMC-menetelmällä replikaatteja $y_1^{\text{rep}}, \ldots, y_K^{\text{rep}}$ jakaumasta $p(y^{\text{rep}}|y)$. Jos y on yksiulotteinen datavektori, niin siitä voidaan laskea esimerkiksi histogrammi (tai joku muu tunnus) ja verrata toistoja aineistosta y laskettuun vastaavaan tunnukseen.



Esim. 1. Newcombin valonnopeusmittaukset (jatkoa). Kappaleen 9.2 esimerkin 1 kuviossa on esitetty Simon Newcombin 66 valonnnopeusmittausta. Kuvion perusteella kaksi pienintä havaintoa näyttävät poikkeavilta. Tutkimme mallin sopivuutta generoimalla 20 toistoaineistoa y^{rep}, jotka on saatu poimimalle ensin μ ja σ^2 niiden posteriorijakaumsta ja sen jälkeen 66 havaintoa normaalijakaumasta $N(\mu, \sigma^2)$. Niiden histogrammit ovat alla olevassa kuviossa. Kuten voidaan havaita missään niistä ei ole sellaisia poikkeavia havaintoja kuin alkuperäisessä aineistossa.

Testisuureet

Mittaamme eroavuutta mallin ja aineiston välillä *testisuureiden* (test quantitities) avulla. Määrittelemme että testisuure $T(y, \theta)$ on parametrien ja aineiston funktio, joka mittaa aineiston ja mallin eroavaisuutta jonkin ominaisuuden suhteen. Bayes-analyysissä testisuureita käytetään mallin tarkistuksessa samalla tavoin kuin testitunnuslukuja käytetään klassisessa analyysissä hypoteesien testaukseen. *Testitunnusluku* (test statistic) on testisuure, joka riippuu pelkästään aineistosta.

Aineiston yhteensopivuutta posterioriennustejakauman välillä voidaan mitata testisuureen jakauman häntäalueen todennäköisyydellä eli p-arvolla. *Klassinen p-arvo* testitunnusluvulle $T(y)$ on

$$p_C = \Pr(T(y^{rep}) \geq T(y)|\theta),$$

jossa todennäköisyys lasketaan y^{rep}:n jakauman suhteen pitämällä θ kiinnitettynä. Jotta p-arvo voidaan laskea, θ:n paikalle on pantava nollahypoteesia vastaava arvo tai jokin piste-estimaatti, kuten suurimman uskottavuuden estimaatti.

Bayesiläinen p-arvo (posterioriennustejakauman p-arvo) määritellään todennäköisyydeksi, että toistettu aineisto on poikkeavampi kuin havaittu aineisto testisuureella mitattuna:

$$p_B = \Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y).$$

Tässä todennäköisyys lasketaan θ:n posteriorijakauman ja y^{rep}:n posterioriennustejakauman suhteen (eli yhteisjakauman $p(\theta, y^{rep}|y)$ suhteen). P-arvot, jotka ovat lähellä nollaa tai ykköstä kertovat poikkeamasta mallin ja aineiston välillä.

Käytännössä p_B voidaan laskea simuloimalla seuraavasti: Generoidaan L havaintoa θ^l posteriorijakaumasta ja näitä vastaavat toistoaineistot $y^{rep\ l}$ posterioriennustejakaumasta. Tällöin p_B on likimain niiden tapausten suhde L simuloinnista, kun $T(y^{rep\ l}, \theta^l) \geq T(y, \theta^l),\ l = 1, ...L$.

Esim. 2. Newcombin valonnopeusmittaukset (jatkoa). (Ks. esimerkkikoodi testExample.r.) Tutkiessamme normaalimallin soveltuvuutta Newcombin aineistoon käytämme *testitunnuslukua* $T(y) = \min(y)$ ja *testisuuretta* $T(y, \theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$. Kuviossa vasemmalla on piirretty toistoaineistojen minimiarvon jakauma. Kaikissa tapauksissa minimiarvo on suurempi kuin alkuperäisen aineiston pienin havainto, joka on merkitty pystyviivalla. Malli ei siis näytä yhteensopivalta alkuperäisen havaintoaineiston kanssa. Mallia voisi korjata käyttämällä epäsymmetristä epäpuhdasta (contaminated) normaalijakaumaa tai jotain symmetristä paksuhäntäistä jakaumaa. Testitunnuslukuun liittyvä p-arvo on 1.

Oikeanpuoleisessa kuviossa on parien $(T(y, \theta), T(y^{rep}, \theta))$ pisteparvi, joka perustuu 1000 posteriorisimulaatioon (θ, y^{rep}). Niiden pisteiden osuus, jotka ylittävät 45^o suoran, vastaa testiin liittyvää p-arvoa, joka on tässä tapauksessa noin 0.5. Siis varianssi testisuureena ei "havaitse" poikkeamaa mallin ja havaintoaineiston välillä.