Bayes-päättely

Kevät 2017

Ville Hyvönen
Helsingin yliopisto

Sisältö

Luku 1

Johdanto

1.1 Tilastollinen päättely

Havaitaan aineisto $\mathbf{y} = (y_1, \ldots, y_n)$. Yksittäiset komponentit voivat olla reaaliarvoisia, kuten seuraavissa esimerkeissä, tai vektoriarvoisia, kuten useimmissa oikeissa sovelluksissa Tilastollisessa päättelyssä ollaan kiinnostuneita aineiston generoineesta ilmiöstä; aineistoa voidaan pitää satunnaisvektorin $\mathbf{Y} = (Y_1, \ldots, Y_n)$ toteutuneina arvoina, ja tarkoituksena on selvittää jakauma, jota satunnaisvektori $\mathbf{Y}$ noudattaa.

Jos satunnaisvektori $\mathbf{Y}$ on diskreetti tai jatkuva, sen jakauman määrää yhteispistetodennäköisyysfunktio tai yhteistiheysfunktio $f_{\mathbf{Y}}(\mathbf{y})$. Tällä kurssilla rajoitumme **parametriseen** päättelyyn, jossa oletetaan funktion $f_{\mathbf{Y}}$ muoto tunnetuksi skalaari- tai vektoriarvoista parametria $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \Omega$ vaille. Tällöin aineiston jakaumaa merkitään joko $f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})$ tai $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$, riippuen siitä, mallinetaanko parametri satunnaismuuttujana vai vakiona. Mahdollisten parametrin arvojen joukkoa $\Omega \subset \mathbb{R}^d$ kutsutaan **parametriavaruudeksi**. Näin päättely aineiston jakaumasta palautuu parametrin $\boldsymbol{\theta}$ uskottavimman arvon, ja siihen liittyvän epävarmuuden selvittämiseen.

1.2 Frekventistinen päättely

Frekventistisessä päättelyssä puhutaan nimenomaan parametrin $\boldsymbol{\theta}$ aineiston valossa *uskottavimmasta* arvosta, sillä parametri oletetaan vakioksi, jonka meille tuntematon arvo on tarkoitus selvittää, mutta johon ei voida liittää todenäköisyysväitteitä: ei voida esimerkiksi mielekkäästi keskustella todennäköisyydestä, että parametrin todellinen arvo kuuluisi jollekin välille. Suurimman uskottavuuden estimoinnissa (maximum likelihood estimation eli MLE) tarkastellaan aineiston uskottavuusfunktiota $L(\boldsymbol{\theta}; \mathbf{y})$, joka on satunnaismuuttujien $Y_1, \ldots, Y_n$ yhteispistetodennäköisyys- tai tiheysfunktio $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ ymmärrettynä para-

metrin $\boldsymbol{\theta}$ funktiona:

$$\boldsymbol{\theta} \mapsto f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}).$$

Parametrille $\boldsymbol{\theta}$ löydetään havaitun aineiston valossa uskottavin piste-estimaatti, eli **suurimman uskottavuuden estimaatti**, maksimoimalla uskottavuusfunktiota parametrin $\boldsymbol{\theta}$ suhteen:

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ L(\boldsymbol{\theta}; \mathbf{y}) \tag{1.1}$$

Koska parametrin oletetaan vakioksi, epävarmuuttamme parametrin todellisesta arvosta kuvataan parametrin suurimman uskottavuuden *estimaattorin* $\hat{\boldsymbol{\theta}}(\mathbf{Y})$, joka saadaan korvaamalla kaavassa 1.1 havaittu aineisto $\mathbf{y}$ satunnaisvektorilla $\mathbf{Y}$, jakauman avulla, jolloin satunnaisuus liittyy satunnaisvektoriin $\mathbf{Y}$, eikä parametriin $\boldsymbol{\theta}$. Estimaattorin jakauman tarkastelu mahdollistaa esimerkiksi keskivirheiden ja luottamusvälien laskemisen ja hypoteesien testaamisen.

1.3 Bayes-päättely

Bayesiläisessä päättelyn teoriassa, jota tällä kurssilla käsitellään, parametria mallinnetaan vakion sijasta satunnaismuuttujalla $\boldsymbol{\Theta}$. Tämä perustuu **subjektiivisen todennäköisyyden** käsitteeseen. Sen mukaan todennäköisyyttä voidaan käyttää kvantifioimaan epävarmuuttamme asioiden todellisesta tilasta, esimerkiksi parametrin todellisesta arvosta. Sen lisäksi, että voidaan puhua todennäköisyydestä, että hehkulamppu palaa yli 2000 tuntia, jos sen kestoikä noudattaa eksponenttijakaumaa odotusarvolla 1000, voidaan yhtä ongelmattomasti puhua myös todennäköisyydestä, että hehkulamppujen kestoiän odotusarvo $\frac{1}{\theta}$ sijaitsee välillä (900, 1100). Sen sijaan, että käyttäisimme estimaattorin (joka siis on aineiston ymmärrettynä satunnaismuuttujana $\mathbf{Y}$ funktio) jakaumaa epävarmuutemme parametrin todellisesta arvosta ilmaisemiseen numeerisesti, voimme asettaa todennäköisyysjakauman suoraan parametrin mahdollisille arvoille.

1.3.1 Posteriorijakauman laskeminen

Bayes-päättely perustuu aineiston ja parametrin yhteisjakauman[1] $f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta}, \mathbf{y})$ määrittämiseen ja manipulointiin. Tämän yhteisjakauman reunajakaumista ja ehdollisista jakaumista käytetään Bayes-päättelyssä yleensä seuraavia nimityksiä:

[1] Jatkossa samaistamme jakauman sen tiheys- tai pistetodennäköisyysfunktioon, joka määrää jakauman yksikäsitteisesti, kanssa. Yleinen on myös tilanne, jossa osa komponenteista on diskreettejä ja osa jatkuvia. Tällöin satunnaisvektorilla $(\mathbf{Y}, \boldsymbol{\Theta})$ on ns. sekatyypin jakauma, ja funktiota $f_{\mathbf{Y},\boldsymbol{\Theta}}(\mathbf{y}, \boldsymbol{\theta})$ summataan diskreettien komponenttien, ja integroidaan jatkuvien komponenttien suhteen.

- **Aineiston jakauma / uskottavuusfunktio** (sampling distribution / likelihood function) $f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})$. Kun tarkastellaan aineiston ehdollista jakaumaa ehdolla parametri:

$$\mathbf{y} \mapsto f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}),$$

puhutaan yksinkertaisesti aineiston jakaumasta tai aineiston jakaumasta ehdolla parametri[2]. Aineiston jakauman muodon valinta on viime kädessä mielivaltainen, ja perustuu esimerkiksi ennakkotietoomme mallinnettavasta ilmiöstä. Esimerkiksi kestoikiä on luontevaa mallintaa eksponenttijakaumalla, toistuvien riippumattomien satunnaiskokeiden onnistumisten määrää binomijakaumalla, ja niin edelleen.

Kun taas on havaittu aineisto $\mathbf{Y} = \mathbf{y}$, ja halutaan selvittää parametrin todennäköisimmät arvot, tarkastellaan samaa funktiota parametrin $\boldsymbol{\theta}$ funktiona

$$\boldsymbol{\theta} \mapsto f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}).$$

Tällöin puhutaan uskottavuusfunktiosta. Koska uskottavuusfunktio on parametrin $\boldsymbol{\theta}$ funktio, se ei ole todennäköisyysjakauma. Uskottavuusfunktiosta voidaan myös tiputtaa pois parametrista $\boldsymbol{\theta}$ riippumattomat vakiot, jolloin mitä tahansa funktiota, joka on muotoa

$$\boldsymbol{\theta} \mapsto c(\mathbf{y}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}),$$

missä $c(\mathbf{y})$ on vain aineistosta $\mathbf{y}$ riippuva vakio, voidaan myös kutsua uskottavuusfunktioksi.

Varsinkin alussa tarkastelmissamme yksinkertaisissa esimerkeissä monesti oletetaan, että havainnot ovat **riippumattomia ja samoin jakautuneita** (independent and identically distributed, lyh. i.i.d.)[3] ehdolla parametri $\boldsymbol{\Theta}$. Tällöin aineiston jakauma

[2]Monesti, varsinkin frekventistisen päättelyn, mutta myös Bayes-päättelyn teoriassa tätä jakaumaa kutsutaan myös *tilastolliseksi malliksi*. Koska Bayes-päättelyssä määriteltävä malli on aineiston ja parametrin yhteisjakauma $f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta}, \mathbf{y})$, eli se sisältää aineiston ehdollisen jakauman lisäksi myös priorijakauman, emme käytä tätä nimitystä aineiston ehdollisesta jakaumasta sekaannusten välttämäseksi.

[3]Riippumattomien ja samoin jakautuneiden satunnaismuuttujien käyttämistä perustellaan monissa Bayes-päättelyn teoriaa käsittelevissä lähteissä havaintojen *vaihdettavuudella* (exchangeability), joka on tätä löyhempi oletus. Satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat vaihdettavia, jos niiden yhteisjakauma säilyy samana, vaikka niiden järjestystä vaihdetaan, eli

$$f_{Y_1,\ldots,Y_n} = f_{Y_{\sigma(1)},\ldots,Y_{\sigma(n)}}$$

kaikille indeksien $1, \ldots, n$ permutaatioille $\sigma(1), \ldots \sigma(n)$. Koko parametrisen päättelyn bayesiläisittäin perustelemiseen lähtien vaihdettavuuden käsitteestä voi tutustua esimerkiksi teoksen [1] luvusta 4 tai lyhyemmin artikkelista [2]

faktoroituu yksittäisiä havaintoja vastaavien satunnaismuuttujien jakaumien tuloksi:

$$f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i|\boldsymbol{\Theta}}(y_i|\boldsymbol{\theta}).$$

- **Priorijakauma** (priori distribution) $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. Parametrin $\boldsymbol{\Theta}$ reunajakaumaa kutsutaan Bayes-päättelyssä priorijakaumaksi, eli lyhyemmin vain **prioriksi**. Priori tarkoittaa latinaksi *ennen*, ja priorijakauman tarkoituksena onkin ilmaista numeerisessa muodossa epävarmuutemme parametrin todellisesta arvosta ennen aineiston havaitsemista.

  Parametrin $\boldsymbol{\Theta}$ priorijakaumaksi valitaan useimmiten jokin parametrinen jakauma, jolloin myös priorijakauma riippuu jostain skalaari- tai vektoriarvoisesta parametrista $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_k\}$. Näitä priorijakauman parametreja kutsutaan **hyperparametreiksi**. Ne voidaan myös merkitä eksplisiittisesti näkyviin käyttämällä priorijakaumasta merkintää $f_{\boldsymbol{\Theta}|\boldsymbol{\Phi}}(\boldsymbol{\theta}|\boldsymbol{\phi})$, mutta monesti ne jätetään pois merkintöjen selkeyttämiseksi.

- **Prioriennustejakauma** (prior predictive distribution) $f_{\mathbf{Y}}(\mathbf{y})$. Priorijennustejakaumaa kutsutaan myös evidenssiksi tai yksinkertaisesti aineiston reunajakaumaksi. Prioriennustejakauma kuvaa näkemystämme aineiston todennäköisyydestä ennen sen havaitsemista. Jos aineisto on jatkuva satunnaismuuttuja, prioriennustejakauma lasketaan integroimalla aineiston ja parametrin yhteisjakaumaa parametrin yli:

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_{\Omega} f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta}, \mathbf{y}) \,\mathrm{d}\boldsymbol{\theta} = \int_{\Omega} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta}.$$

  Jos aineisto on diskreetti, prioriennustejakauma lasketaan summaamalla yhteisjakaumaa parametrin suhteen:

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\boldsymbol{\theta}\in\Omega} f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta}, \mathbf{y}) = \sum_{\boldsymbol{\theta}\in\Omega} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}).$$

  Prioriennustejakauman kaavasta nähdään, että se on aineiston jakauman odotusarvo laskettuna parametrin $\boldsymbol{\Theta}$ priorijakauman yli:

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_{\Omega} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} = E[f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})].$$

  Sitä voidaan siis ajatella aineiston jakauman painotettuna keskiarvona, jossa eri parametrin arvojen painot määräytyvät priorijakauman mukaan, tai sekoitusmallina, jossa priorijakauma määrää sekoitussuhteen.

- **Posteriorijakauma** (posterior distribution) $f_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y})$. Parametrin $\boldsymbol{\Theta}$ ehdollista jakaumaa ehdolla aineisto kutsutaan posteriorijakaumaksi tai lyhyemmin **posterioriksi**. Posteriori tarkoittaa latinaksi *jälkeen*, ja posteriorijakauman tarkoituksena onkin kvantifioida parametrin todelliseen arvoon liittyvää epävarmuuttamme aineiston havaitsemisen jälkeen.

  Kun yhteisjakauma on määritelty priorijakauman ja uskottavuusfunktion tulona, posteriorijakauma voidaan laskea Bayesin kaavasta:

$$f_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = \frac{f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta},\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})} = \frac{f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y})}.$$

  Käytännössä ongelmana on aineiston reunajakauman $f_{\mathbf{Y}}(\mathbf{y})$ laskeminen: vähänkään monimutkaisemmissa malleissa integraalia

$$f_{\mathbf{Y}}(\mathbf{y}) = \int_{\Omega} f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta},\mathbf{y})\,\mathrm{d}\boldsymbol{\theta}$$

  ei ole mahdollista ratkaista suljetussa muodossa.

  Havaitaan kuitenkin, että aineiston reunajakauma $f_{\mathbf{Y}}(\mathbf{y})$ on vakio parametrin $\boldsymbol{\theta}$ suhteen, joten posteriorijakauma on verrannollinen[4] yhteisjakaumaan käsitettynä parametrin $\boldsymbol{\theta}$ funktiona $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta},\mathbf{y})$:

$$f_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) \propto f_{\boldsymbol{\Theta},\mathbf{Y}}(\boldsymbol{\theta},\mathbf{y}) = f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}). \tag{1.2}$$

  Normalisoimaton posteriorijakauma saadaan siis aina laskettua priorijakauman ja uskottavuusfunktion tulona. Samalla voidaan jättää pois kaikki parametrista $\boldsymbol{\theta}$ riippumattomat vakiotermit.

Huomautus 1.3. Kaikki edellämainitut kaavat pätevät sekä jatkuville, diskreeteille, että ns. sekatyypin jakaumille. Käytetyt yhteisjakauman, reunajakaumien ja ehdollisten jakaumien määritelmät, ketjusääntö ja Bayesin kaava löytyvät kattavasti esiteltyinä esimerkiksi Todennäköisyyslaskenta II-kurssin luentomonisteesta [4], josta ne voi tarvittaessa kerrata.

[4] Funktio f on verrannollinen funktioon h, jos on olemassa vakio c siten, että

$$f(x) = ch(x)$$

kaikille x. Tällöin merkitään

$$f(x) \propto h(x).$$

Huomautus 1.4. Jos palataan vielä subjektiivisen todennäköisyyden käsitteeseen, priorijakauma kvantifioi käsityksemme todennäköisimmistä parametrin arvoista ja niihin liittyvästä epävarmuudesta ennen aineiston havaitsemista. Aineiston havaitsemisen jälkeen päivitämme nämä käsityksemme Bayesin kaavan avulla, ja tuloksena on posteriorijakauma, joka kuvaa käsitystämme parametrin jakaumasta aineiston havaitsemisen jälkeen.

Havainnollistetaan vielä posteriorijakauman laskemista klassisen esimerkin kautta.

Esimerkki 1.5. Halutaan, selvittää, mikä on todennäköisyys, että nasta laskeutuu pohja alaspäin, eli piikki ylöspäin, heitettäessä. Heitetään tämän selvittämiseksi nastaa n kertaa. Määritellään satunnaismuuttuja

$$Y := \text{niiden heittokertojen määrä, joilla nasta laskeutuu pohja alaspäin.}$$

Nyt on luonnollista olettaa, että

$$Y \mid \Theta \sim \mathrm{Bin}(n, \Theta),$$

eli että Y noudattaa binomijakaumaa parametreilla n ja Θ, missä parametri Θ kuvaa yksittäisen satunnaiskokeen onnistumistodennäköisyyttä, eli todennäköisyyttä, että nasta laskeutuu pohja alaspäin. Heittojen määrä n oletetaan kiinteäksi, joten sillä ei ehdollisteta. Aineiston jakauma ehdolla parametri on siis

$$f_{Y|\Theta}(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Parametrin Θ priorijakaumaksi valitaan betajakauma $\mathrm{Beta}(\alpha, \beta)$, jolloin

$$f_\Theta(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{kun } \theta \in (0,1),$$

missä integraalia

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$$

kutsutaan Eulerin betafunktioksi, tai lyhyemmin betafunktioksi.

Priorijakauman valinta perustuu käytännöllisyyteen: betajakauma on ns. *konjugaattipriori* binomijakaumalle, jolloin posteriorijakauma kuuluu samaan jakaumaperheeseen, eli on samaa muotoa, kuin priorijakauma. Palaamme tähän vielä myöhemmin.

Posteriorijakauman selvittäminen, pitkä tapa: Yhteisjakauma voidaan laskea uskottavuusfunktion ja priorijakauman tulona:

$$\begin{aligned} f_{Y,\Theta}(y,\theta) &= f_\Theta(\theta) f_{Y|\Theta}(y|\theta) \\ &= \binom{n}{y}\frac{1}{B(\alpha,\beta)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}. \end{aligned}$$

Prioriennustejakauma, eli aineiston reunajakauma, saadaan integroimalla yhteisjakaumaa, josta tunnistamme jälleen betafunktion:

$$\begin{aligned} f_Y(y) &= \int_0^1 \binom{n}{y} \frac{1}{B(\alpha,\beta)} \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}\,\mathrm{d}\theta \\ &= \binom{n}{y} \frac{1}{B(\alpha,\beta)} \int_0^1 \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}\,\mathrm{d}\theta \\ &= \binom{n}{y} \frac{B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}. \end{aligned} \tag{1.6}$$

Tätä jakaumaa kutsutaan ns. beta-binomijakaumaksi, jota merkitään:

$$Y \sim \text{Beta-bin}(n,\alpha,\beta).$$

Nyt posteriorijakauma voidaan ratkaista Bayesin kaavasta:

$$\begin{aligned} f_{\Theta|Y}(\theta|y) &= \frac{f_{Y,\Theta}(y,\theta)}{f_Y(y)} \\ &= \frac{\binom{n}{y}\frac{1}{B(\alpha,\beta)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}}{\binom{n}{y}\frac{B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}} \\ &= \frac{1}{B(\alpha+y,\beta+n-y)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}. \end{aligned}$$

Posteriorijakaumaksi saatiin siis jälleen betajakauma, tällä kertaa parametreilla $\alpha+y$ ja $\beta+n-y$, eli

$$\Theta \,|\, Y \sim \text{Beta}(\alpha+y,\beta+n-y).$$

Posteriorijakauman selvittäminen, lyhyt tapa: Posteriorijakauma voidaan laskea myös helpommin käyttämällä Bayesin kaavasta muotoa 1.2, eli käyttämällä hyväksi sitä, että posteriorijakauma on verrannollinen priorin ja uskottavuusfunktion tuloon:

$$\begin{aligned} f_{\Theta|Y}(\theta|y) &\propto f_\Theta(\theta) f_{Y|\Theta}(y|\theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^y(1-\theta)^{n-y} \\ &= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}. \end{aligned} \tag{1.7}$$

Nyt voidaan suoraan päätellä laskematta normalisointivakion arvoa, että posteriorijakauma on betajakauma $\text{Beta}(\alpha+y,\beta+n-y)$, sillä ainoa todennäköisyysjakauma, jolla on alusta $\theta \in (0,1)$, ja joka on muotoa

$$f(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

on betajakauma Beta(a, b). Tämä johtuu siitä, että ollakseen todennäköisyysjakauman, funktion on integraalin yli sen alustan on oltava 1, jolloin ainoa mahdollinen normalisointivakio on $\frac{1}{B(a,b)}$.

Huomaa myös laskun 1.7 toisen rivin verrannollisuusmerkki: priorista ja uskottavuusfunktiosta on jätetty pois parametrista θ riippumattomat vakiot, mikä yksinkertaisti laskua entisestään ylläolevaan verrattuna.

Tämä verrannollisuustarkastelu onkin tapa, jolla posteriorijakauma kannattaa selvittää konjugaattipriorin tapauksessa.

1.3.2 Ennustaminen

Parametrien lisäksi toinen tilastollisen päättelyn keskeisistä kiinnostuksen kohteista ovat uudet havainnot tarkasteltavasta prosessista $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_m)$.

Oletetaan, että uudet havainnot ovat riippumattomia havaitusta aineistosta, jolloin ilmeisin (ja frekventistisessä päättelyssä yleensä käytetty) tapa ennustaa uusia havaintoja on sijoittaa parametrin suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\theta}}(\mathbf{y})$ uusien havaintojen jakaumaan, eli käyttää ennustamiseen jakaumaa

$$f_{\tilde{\mathbf{Y}}|\boldsymbol{\Theta}}(\tilde{\mathbf{y}}|\hat{\boldsymbol{\theta}}(\mathbf{y})).$$

Tämän lähestymistapa ei ota huomioon parametrin todelliseen arvoon liittyvää epävarmuutta ennustejakaumassa, mikä voi johtaa outoihin tuloksiin pienellä otoskoolla. Tähän palataan esimerkeissä.

Bayes-päättelyssä käytetään sen sijaan ennustamiseen uusien havaintojen jakaumaa ehdolla aineisto, eli ns. **posterioriennustejakaumaa** $f_{\tilde{\boldsymbol{Y}}|\mathbf{Y}}(\tilde{\mathbf{y}}|\mathbf{y})$, joka saadaan integroimalla uusien havaintojen ja parametrin yhteisjakaumaa $f_{\tilde{\boldsymbol{Y}},\boldsymbol{\Theta}|\mathbf{Y}}(\tilde{\mathbf{y}}, \boldsymbol{\theta}|\mathbf{y})$ ehdolla aineisto parametriavaruuden yli:

$$f_{\tilde{\boldsymbol{Y}}|\mathbf{Y}}(\tilde{\mathbf{y}}|\mathbf{y}) = \int_\Omega f_{\tilde{\boldsymbol{Y}},\boldsymbol{\Theta}|\mathbf{Y}}(\tilde{\mathbf{y}}, \boldsymbol{\theta}|\mathbf{y}) \,\mathrm{d}\boldsymbol{\theta}.$$

Jos $\mathbf{Y}, \tilde{\mathbf{Y}} \perp\!\!\!\perp \boldsymbol{\Theta}$, eli oletetaan, että uudet havainnot ovat riippumattomia aineistosta ehdolla parametri, posterioriennustejakauma voidaan kirjoittaa integraalina uusien havaintojen ja posteriorijakauman tulosta parametriavaruuden yli:

$$\begin{aligned} f_{\tilde{\boldsymbol{Y}}|\mathbf{Y}}(\tilde{\mathbf{y}}|\mathbf{y}) &= \int_\Omega f_{\tilde{\boldsymbol{Y}},\boldsymbol{\Theta}|\mathbf{Y}}(\tilde{\mathbf{y}}, \boldsymbol{\theta}|\mathbf{y}) \,\mathrm{d}\boldsymbol{\theta} \\ &= \int_\Omega f_{\tilde{\mathbf{Y}}|\boldsymbol{\Theta},\mathbf{Y}}(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \mathbf{y}) f_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) \\ &= \int_\Omega f_{\tilde{\mathbf{Y}}|\boldsymbol{\Theta}}(\tilde{\mathbf{y}}|\boldsymbol{\theta}) f_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}). \end{aligned} \tag{1.8}$$

Samalla tavoin kuin prioriennustejakauma, myös posterioriennustejakauma voidaan tulkita odotusarvona tai sekoitusmallina; tällä kertaa vain odotusarvo on parametrin $\boldsymbol{\Theta}$ priorijakauman sijaan sen posteriorijakauman yli, eli kyseessä on uusien havaintojen uskottavuusfunktion ehdollinen odotusarvo ehdolla havaittu aineisto $\mathbf{Y} = \mathbf{y}$:

$$f_{\tilde{\mathbf{Y}}|\mathbf{Y}}(\tilde{\mathbf{y}}|\mathbf{y}) = \int_\Omega f_{\tilde{\mathbf{Y}}|\boldsymbol{\Theta}}(\tilde{\mathbf{y}}|\boldsymbol{\theta}) f_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = E[f_{\tilde{\mathbf{Y}}|\boldsymbol{\Theta}}(\tilde{\mathbf{y}}|\boldsymbol{\theta}) \mid \mathbf{Y} = \mathbf{y}]. \tag{1.9}$$

Havainnollistetaan vielä uusien havaintojen ennustamista palaamalla nastanheittokokeeseen.

Esimerkki 1.10. Jatkoa esimerkille 1.5. Ollaan heitetty nastaa n kertaa, joista y kertaa nasta laskeutui pohja alaspäin, eli ollaan havaittu aineisto $Y = y$.

Halutaan selvittää, mikä on onnistumisten, eli heittokertojen, joilla nasta laskeutuu pohja alaspäin, määrän seuraavalla m:llä heitolla jakauma ehdolla havaittu aineisto. Merkitään tätä määrää satunnaismuuttujalla $\tilde{Y}$. Koska kyseessä on sama nasta, voidaan olettaa, että onnistumistodennäköisyys pysyy samana, eli että

$$\tilde{Y} \mid \Theta \sim \mathrm{Bin}(m, \Theta).$$

Lisäksi voidaan olettaa, että heitot ovat keskenään vaihdettavia, eli että muuttujat Y ja $\tilde{Y}$ ovat riippumattomia ehdolla parametri Θ:

$$Y \perp\!\!\!\perp \tilde{Y} \mid \Theta.$$

Haluamme siis laskea posterioriennustejakauman $f_{\tilde{Y}|Y}$. Kaavasta 1.8 näemme, että posterioriennustejakauma on muotoa

$$f_{\tilde{Y}|Y}(\tilde{y}|y) = \int_\Omega f_{\tilde{Y}|\Theta}(\tilde{y}|\theta) f_{\Theta|Y}(\theta|y)\,\mathrm{d}\theta.$$

Kyseessä siis on uusien havaintojen $\tilde{Y}$ jakauman ja parametrin Θ posteriorijakauman tulon integraalista yli välin $(0, 1)$. Mutta $\tilde{Y}$:n jakauma on binomijakauma, ja Θ:n posteriorijakauma on betajakauma, joten kyseessä on sama integraali, kuin tämän mallin prioriennustejakaumassa (kaava 1.6)! Posterioriennustejakauma siis saadaan vaihtamalla oikeat parametrit prioriennustejakauman kaavaan:

$$f_{\tilde{Y}|Y}(\tilde{y}|y) = \binom{m}{\tilde{y}} \frac{B(\alpha + y + \tilde{y}, \beta + n - y + m - \tilde{y}))}{B(\alpha + y, \beta + n - y)}.$$

Myös posterioriennustejakauma on siis beta-binomijakauma, eli

$$\tilde{Y} \mid Y \sim \text{Beta-bin}(m, \alpha + y,\ \beta + n - y).$$

Huomautus 1.11. Yllä kaikkiin jakaumiin merkittiin näkyviin satunnaismuuttujat, joiden jakaumia ne ovat, alaindekseillä. Yleensä alan kirjallisuudessa ei kuitenkaan käytetä näin tarkkoja merkintöjä. Sen sijaan käytetään lyhennysmerkintää, jossa kaikkia jakaumia merkitään samalla symbolia, Bayesian data analysis-kirjassa kirjaimella p, ja satunnaismuuttujia, joiden jakauma on kysymyksessä merkitään *muuttujilla*. Siten esimerkiksi aineiston ehdollista jakaumaa $f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})$ merkitään yksinkertaisesti $p(\mathbf{y}|\boldsymbol{\theta})$, ja Bayesin kaava kirjoitetaan muodossa:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}.$$

Tämä lyhennysmerkintä lyhentää kaavoja ja tekee niistä selkeämpiä, kunhan vain alun perin tietää, mistä merkinnät ovat lyhenteitä! Tämän takia käytimme johdannossa raskaampaa notaatiota, mutta jatkossa enimmäkseen tätä tiivimpää merkintätapaa.

Merkintöjen yksinkertaistamiseksi myös monesti merkitään satunnaismuuttujia, ja niiden realisaatioita kumpaakin samalla pienellä kirjaimella, varsinkin parametrien yhteydessä. Käytämme myös tätä tapaa jatkossa, jos sekaannuksen vaaraa ei ole.

Luku 2

Konjugaattipriorit

Konjugaattipriorilla, eli liittopriorilla, tarkoitetaan sellaista priorijakaumaa (jollekin uskottavuusfunktiolle), että posteriorijakauma kuuluu samaan parametriseen jakaumaperheeseen kuin priorijakauma. Tällöin uskottavuusfunktiosta ja priorijakaumasta puhutaan myös **konjugaattiparina**.

Esimerkiksi Poisson-jakauman konjugaattipriori on gammajakauma: riippumattomille havainnoille Poisson-jakaumasta myös posteriorijakauma on gammajakauma, kuten seuraavasta esimerkistä nähdään.

Prioriennustejakauma $p(\mathbf{y})$ on mahdollista ratkaista suljetussa muodossa ainoastaan kahdessa tapauksessa:

1. Parametriavaruus on diskreetti ja äärellinen, eli $\Omega = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p\}$, jolloin prioriennustejakauma saadaan äärellisenä summana

$$p(\mathbf{y}) = \sum_{i=1}^{p} p(\boldsymbol{\theta}_i) p(\mathbf{y}|\boldsymbol{\theta}_i).$$

2. Käytetty priorijakauma on konjugaattipriori uskottavuusfunktiolle.

Konjugaattipriorin käyttäminen siis mahdollistaa posteriorijakauman ja posterioriennustejakauman laskemisen suljetussa muodossa. Muussa tapauksessa joudumme approksimoimaan näitä jakaumia joko integroimalla numeerisesti, tai simulaatiomenetelmillä. Samalla tavoin kuin tietyn parametrisen muodon olettaminen aineiston jakaumalle, myös konjugaattipriorien käyttäminen siis yksinkertaistaa tilastotieteilijän elämää huomattavasti. Tämän takia monimutkaisempia hierarkkisia malleja, joita käsittelemme myöhemmin, rakennellaan usein juuri eksponenttiperheen jakaumista ja niiden konjugaattiprioreista.

Käsittelemme tämän luvun aluksi muutaman tärkeän esimerkin konjugaattipareista yksinkertaisimmassa tapauksessa, jossa parametri on yksiulotteinen, eli $\Omega \subset \mathbb{R}$. Sen jäl-

keen tarkastelemme mille jakaumille on olemassa konjugaattipriori, ja mitä muotoa konjugaattipriori on, kun se on olemassa. Osoittautuu, että konjugaattipriori on olemassa ainoastaan ns. eksponenttiperheen jakaumille.

Tarkastelemme myös lyhyesti erilaisia periaatteita priorijakauman valintaan. Lopuksi käsittelemme vielä yleisempää asetelmaa, jossa parametri on moniulotteinen, eli $\Omega \subset \mathbb{R}^d, d > 1$.

2.1 Yhden parametrin jakaumat

Esimerkissä 1.5 havaitsimme, että betajakauma on konjugaattipriori binomijakaumalle. Otetaan vielä esimerkki klassisesta tilastollisen päättelyn asetelmasta, jossa meillä on riippumattomia havaintoja samasta jakaumasta.

Esimerkki 2.1. Oletetaan riippumattomat havainnot Poisson-jakaumasta:

$$Y_1, \ldots, Y_n \sim \text{Poisson}(\theta), \quad Y_1, \ldots, Y_n \perp\!\!\!\perp \mid \theta,$$

ja havaitaan aineisto $\mathbf{y} = (y_1, \ldots, y_n)$. Tällöin kaikkien yksittäisten havaintojen jakauma on

$$p(y_i|\theta) = \theta^{y_i} \frac{e^{-\theta}}{y_i!},$$

ja uskottavuusfunktio voidaan havaintojen riippumattomuuden nojalla esittää tulona:

$$p(\mathbf{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta} = \theta^{n\bar{y}} e^{-n\theta},$$

missä $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ on aineiston keskiarvo.

Käytetään priorina gammajakaumaa Gamma(α, β), jolloin

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

kaikille $\theta > 0$.

Siten normalisoimattomaksi posteriorijakaumaksi saadaan

$$\begin{aligned} p(\theta|\mathbf{y}) &\propto p(\theta) p(\theta|\mathbf{y}) \\ &\propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{n\bar{y}} e^{-n\theta} \\ &= \theta^{\alpha+n\bar{y}-1} e^{(-\beta+n)\theta}. \end{aligned}$$

Tämä voidaan tunnistaa gammajakauman Gamma$(\alpha + n\bar{y}, \beta + n)$ ytimeksi. Posteriorijakauma ja priorijakauma ovat kumpikin gammajakaumia, joten gammajakauma on konjugaattipriori Poisson-jakaumalle.

Johdetaan vielä prioriennustejakauma yhdelle havainnolle $p(y_1)$. Integraali voitaisiin tunnistaa muuttujanvaihdolla gammafunktioksi, mutta ehkä helpompi tapa on huomata, että se on vakiotermiä vaille gammajakauman Gamma$(\alpha + y_1, \beta + 1)$ tiheysfunktion integraali yli sen alustan $\theta \in (0, \infty)$, ja täydentää puuttuvat vakiot, jolloin integraalin arvoksi tulee 1:

$$
\begin{aligned}
p(y_1) &= \int p(\theta) p(y_1|\theta) \, \mathrm{d}\theta \\
&= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \; \theta^{y_1} \frac{e^{-\theta}}{y_1!} \, \mathrm{d}\theta \\
&= \frac{1}{y_1!} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^{\alpha+y_1-1} e^{-(\beta+1)\theta} \, \mathrm{d}\theta \\
&= \frac{1}{y_1!} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+y_1)}{(\beta+1)^{\alpha+y_1}} \int_0^\infty \frac{(\beta+1)^{\alpha+y_1}}{\Gamma(\alpha+y_1)} \theta^{\alpha+y_1-1} e^{-(\beta+1)\theta} \, \mathrm{d}\theta \\
&= \frac{1}{y_1!} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+y_1)}{(\beta+1)^{\alpha+y_1}} \cdot 1 \\
&= \frac{\Gamma(\alpha+y_1)}{\Gamma(\alpha) y_1!} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^{y_1}.
\end{aligned}
\tag{2.2}
$$

Tätä jakaumaa kutsutaan negatiiviseksi binomijakaumaksi[1], merkitään

$$Y_1 \sim \text{Neg-bin}(\alpha, \beta).$$

Laskun 2.2 ensimmäiseltä riviltä nähdään, että negatiivinen binomijakauma on sekoitusmalli Poisson-jakaumista, joiden odotusarvot noudattavat gammajakaumaa. Negatiivista binomijakaumaa käytetäänkin ns. ylihajontaa (overdispersion), eli ilmiötä, jota voitaisiin mallintaa Poisson-jakaumalla, mutta jossa aineiston hajontaparametri on suurempi kuin keskiarvo.

[1]Tämän BDA:ssa käytetyn parametrisoinnin lisäksi toinen yleinen tapa parametrisoida negatiivinen binomijakauma on merkitä $p := \frac{\beta}{1+\beta}$, jolloin merkitään $Y \sim \text{Neg-bin}(\alpha, p)$, ja tiheysfunktio voidaan kirjoittaa muodossa

$$p(y) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha) y!} p^\alpha (1-p)^y.$$

Jos vielä lisäksi α on positiivinen kokonaisluku, jakaumaa kutsutaan monesti Pascalin jakaumaksi, ja sen tiheysfunktio voidaan kirjoittaa muodossa

$$p(y) = \binom{\alpha+y-1}{y} p^\alpha (1-p)^y.$$

Tällä jakaumalla on fysikaalinen tulkinta toistokokeen kautta: Jos merkitään satunnaismuuttujalla Y tarvittavaa epäonnistumisten määrää ennen kuin saadaan α onnistumista toistokokeessa, jonka onnistumistodennäköisyys on p, tällöin $Y \sim \text{Neg-bin}(\alpha, p)$.

Tarkastellaan vielä seuraavaa yksittäistä havaintoa samasta prosessista, eli oletetaan, että

$$\tilde{Y} \sim \text{Poisson}(\theta), \quad \tilde{Y}, \mathbf{Y} \perp\!\!\!\perp \mid \theta.$$

Tällöin posterioriennustejakauma on

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

Jälleen havaitsemme (vrt. esimerkki 1.5), että sekä posteriori- että priorijakauma ovat gammajakaumia, ja uskottavuusfunktio on kummassakin laskussa sama Poisson-jakauma eli kyseessä on sama integraali kuin prioriennustejakauman laskussa 2.2. Koska posteriorijakauma on gammajakauma $\text{Gamma}(\alpha + n\bar{y}, \beta + n)$, posterioriennustejakauma saadaan vaihtamalla nämä parametrit prioriennustejakauman lausekkeeseen, eli

$$\tilde{Y} \mid \mathbf{Y} = \mathbf{y} \sim \text{Neg-bin}(\alpha + n\bar{y}, \beta + n).$$

2.1.1 Normaalijakauma yhdellä tunnetulla parametrilla

Esimerkki 2.3. Käydään vielä läpi yksi keskeinen esimerkki jakaumista, joille on olemassa konjugaattipriori, eli normaalijakauma. Tarkastellaan ensin yksinkertaisinta tilannetta, jossa meillä on yksi havainto normaalijakaumasta, jonka varianssi oletetaan tunnetuksi, eli satunnaismuuttujaa

$$Y \sim N(\theta, \sigma_0^2),$$

missä $\sigma_0^2 \in (0, \infty)$ on tunnettu vakio. Tällöin uskottavuusfunktio on

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y-\theta)^2}{2\sigma_0^2}\right) \propto \exp\left(-\frac{\theta^2 - 2y\theta}{2\sigma_0^2}\right).$$

Valitaan odotusarvon θ priorijakaumaksi normaalijakauma $N(\mu_0, \tau_0^2)$, jolloin

$$p(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left(-\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right) \propto \exp\left(-\frac{\theta^2 - 2\mu_0\theta}{2\tau_0^2}\right).$$

Normalisoimaton posteriorijakauma voidaan tunnistaa normaalijakauman tiheysfunktion ytimeksi soveltamalla jälleen Bayesin kaavan versiota 1.2:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(\theta|y) \\ &\propto \exp\left(-\frac{\theta^2 - 2\mu_0\theta}{2\tau_0^2} - \frac{\theta^2 - 2y\theta}{2\sigma_0^2}\right) \\ &= \exp\left(-\frac{\sigma_0^2(\theta^2 - 2\mu_0\theta) + \tau_0^2(\theta^2 - 2y\theta)}{2\tau_0^2\sigma_0^2}\right) \\ &\propto \exp\left(-\frac{(\sigma_0^2 + \tau_0^2)\theta^2 - 2(\sigma_0^2\mu_0 + \tau_0^2 y)\theta}{2\tau_0^2\sigma_0^2}\right) \\ &\propto \exp\left(-\frac{\theta^2 - 2\mu_1\theta}{2\tau_1^2}\right), \end{aligned}$$

missä

$$\mu_1 = \frac{\sigma_0^2\mu_0 + \tau_0^2 y}{\sigma_0^2 + \tau_0^2},$$

ja

$$\tau_1^2 = \frac{\tau_0^2\sigma_0^2}{\sigma_0^2 + \tau_0^2}.$$

Parametrin θ posteriorijakaumaksi saadaan siis

$$\theta|y \sim N(\mu_1, \tau_1^2).$$

Posteriorijakauman parametrit voidaan kirjoittaa myös **tarkkuuden** (precision), millä nimellä kutsutaan varianssin käänteislukua $\frac{1}{\tau_1^2}$, avulla seuraavasti:

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma_0^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma_0^2}},$$

ja

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma_0^2}.$$

Posteriorijakauman tarkkuus on siis aineiston jakauman tarkkuuden (joka oletettiin tunnetuksi), ja priorijakauman tarkkuden summa, ja posteriorijakauma on lineaarikombinaatio aineiston jakauman odotusarvosta ja priorijakauman odotusarvosta painotettuna niiden tarkkuuksilla: mitä suurempi jakauman tarkkuus, sitä suurempi on vastaavan odotusarvon paino.

Posterioriennustejakauma, pitkä tapa: Lasketaan vielä posterioriennustejakauma $p(\tilde{y}|y)$ seuraavalle riippumattomalle havainnolle $\tilde{Y} \perp\!\!\!\perp Y \,|\, \theta,\ \ \tilde{Y} \sim N(\theta, \sigma_0^2)$.

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)\,\mathrm{d}\theta \\
&= \frac{1}{\sqrt{2\pi\sigma_0^2}} \int \exp\left(-\frac{(\tilde{y}-\theta)^2}{2\sigma_0^2}\right) \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right) \mathrm{d}\theta \\
&= \frac{1}{(2\pi)\sqrt{\sigma_0^2\tau_1^2}} \int \exp\left(-\frac{(\sigma_0^2+\tau_1^2)\theta^2 - 2(\tau_1^2\tilde{y}+\sigma_0^2\mu_1)\theta + \tau_1^2\tilde{y}^2 + \sigma_1^2\mu_1^2}{2\sigma_0^2\tau_1^2}\right) \mathrm{d}\theta \\
&= \frac{1}{(2\pi)\sqrt{\sigma_0^2\tau_1^2}} \int \exp\left(-\frac{\theta^2 - 2\frac{\tau_1^2\tilde{y}+\sigma_2\mu_1}{\sigma_0^2+\tau_1^2}\theta + \frac{\tau_1^2\tilde{y}^2+\sigma_1^2\mu_1^2}{\sigma_0^2+\tau_1^2}}{2\frac{\sigma_0^2\tau_1^2}{\sigma_0^2+\tau_1^2}}\right) \mathrm{d}\theta \\
&= \frac{\sqrt{2\pi\frac{\sigma_0^2\tau_1^2}{\sigma_0^2+\tau_1^2}}}{(2\pi)\sqrt{\sigma_0^2\tau_1^2}} \exp\left(-\frac{\frac{\tau_1^2\tilde{y}^2+\sigma_0^2\mu_1^2}{\sigma_0^2+\tau_1^2} - \left(\frac{\tau_1^2\tilde{y}+\sigma_0^2\mu_1}{\sigma_0^2+\tau_1^2}\right)^2}{2\frac{\sigma_0^2\tau_1^2}{\sigma_0^2+\tau_1^2}}\right) \int \frac{1}{\sqrt{2\pi\frac{\sigma_0^2\tau_1^2}{\sigma_0^2+\tau_1^2}}} \exp\left(-\frac{\left(\theta - \frac{\tau_1^2\tilde{y}+\sigma_0^2\mu_1}{\sigma_0^2+\tau_1^2}\right)^2}{2\frac{\sigma_0^2\tau_1^2}{\sigma_0^2+\tau_1^2}}\right) \mathrm{d}\theta \\
&= \frac{1}{\sqrt{2\pi(\sigma_0^2+\tau_1^2)}} \exp\left(-\frac{\tau_1^2\tilde{y}^2 + \sigma_0^2\mu_1^2 - \frac{(\tau_1^2\tilde{y}+\sigma_0^2\mu_1)^2}{\sigma_0^2+\tau_1^2}}{2\sigma_0^2\tau_1^2}\right) \\
&= \frac{1}{\sqrt{2\pi(\sigma_0^2+\tau_1^2)}} \exp\left(-\frac{\frac{\sigma_0^2+\tau_1^2}{\sigma_0^2}\tilde{y}^2 + \frac{\sigma_0^2+\tau_1^2}{\tau_1^2}\mu_1^2 - \frac{\tau_1^2}{\sigma_0^2}\tilde{y}^2 - 2\tilde{y}\mu_1 - \frac{\sigma_0^2}{\tau_1^2}\mu_1^2}{2(\sigma_0^2+\tau_1^2)}\right) \\
&= \frac{1}{\sqrt{2\pi(\sigma_0^2+\tau_1^2)}} \exp\left(-\frac{(\tilde{y}-\mu_1)^2}{2(\sigma_0^2+\tau_1^2)}\right).
\end{aligned}
$$

Posterioriennustejakaumaksi saatiin siis normaalijakauma $N(\mu_1, \sigma_0^2+\tau_1^2)$. Laskussa käytettiin jälleen hyväksi sitä, että jakauman, tässä tapauksessa normaalijakauman integraali yli sen alustan (normaalijakauman tapauksessa koko reaaliakselin) on 1.

Posterioriennustejakauma, lyhyempi tapa: Ylläoleva integrointi on työläs ja virhealtis. Helpommin tämän posterioriennustejakauman voi selvittää käyttämällä hyväksi normaalijakauman ominaisuuksia. Kaksiulotteisen normaalijakauman tiheysfunktio on muotoa

$$p(x,y) \propto \exp\left(ax^2 + by^2 + cx + dy + exy\right),$$

ja havaitsemme, että ylläolevan laskun integrandi, eli uuden havainnon ja parametrin yhteisjakauma ehdolla havaittu aineisto

$$
\begin{aligned}
p(\tilde{y}, \theta|y) &= p(\tilde{y}|\theta)p(\theta|y) \\
&\propto \exp\left(-\frac{(\tilde{y}-\theta)^2}{2\sigma_0^2}\right) \exp\left(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right) \\
&\propto \exp\left(-\frac{(\sigma_0^2+\tau_1^2)\theta^2 - 2\tau_1^2\tilde{y}\theta + 2\sigma_0^2\mu_1\theta + \tau_1^2\tilde{y}^2}{2\sigma_0^2\tau_1^2}\right)
\end{aligned}
$$

on myös tätä muotoa. Muistamme todennäköisyyslaskennasta ([4], lause 10.2), että multinormaalijakauman reunajakaumat ovat myös normaalijakaumia, joten posterioriennustejakauma, joka saadaan integroimalla tätä ehdollista yhteisjakaumaa

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

on myös jokin normaalijakauma.

Koska $\tilde{Y}$ ja Y ovat riippumattomia ehdolla parametri,

$$E[\tilde{Y} \,|\, \theta, Y] = E[\tilde{Y} \,|\, \theta] = \theta$$

ja

$$\mathrm{Var}[\tilde{Y} \,|\, \theta, Y] = \mathrm{Var}[\tilde{Y} \,|\, \theta] = \sigma_0^2.$$

Siten tarkasteltavan normaalijakauman odotusarvo saadaan soveltamalla todennäköisyyslaskennasta tuttua ns. iteroidun odotusarvon lakia ([4], lause 8.3):

$$E[\tilde{Y} \,|\, Y] = E[E(\tilde{Y} \,|\, \theta, Y) \,|\, Y] = E[\theta \,|\, Y] = \mu_1,$$

ja varianssi vastaavasta kaavasta satunnaismuuttujan varianssille ([4], lause 8.4):

$$\begin{aligned}\mathrm{Var}[\tilde{Y} \,|\, Y] &= E[\mathrm{Var}(\tilde{Y} \,|\, \theta, Y) \,|\, Y] + \mathrm{Var}[E(\tilde{Y} \,|\, \theta, Y) \,|\, Y] \\ &= E[\sigma^2 \,|\, Y] + \mathrm{Var}[\theta \,|\, Y] \\ &= \sigma^2 + \tau_1^2.\end{aligned}$$

Huomautus 2.4. Harjoituksissa osoitetaan että kun n:lle riippumattomasti samaa normaalijakaumaa noudattavalle satunnaismuuttujalle

$$Y_1, \ldots, Y_n \perp\!\!\!\perp \,|\, \theta, \quad Y_i \sim N(\mu, \sigma_0^2) \quad \text{kaikille } i = 1, \ldots, n$$

missä varianssi $\sigma_0^2 \in (0, \infty)$ on tunnettu vakio, parametrin θ posteriorijakaumaksi samalla priorijakaumalla

$$\theta \sim N(\mu_0, \tau_0^2)$$

saadaan

$$\theta \,|\, \mathbf{y} \sim N(\mu_n, \tau_n^2),$$

missä

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}},$$

ja

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}.$$

Posteriorijakauman tarkkuus on siis priorijakauman tarkkuus plus aineiston koko kertaa aineiston tarkkuus, ja posteriorijakauman odotusarvo on priorijakauman odotusarvon ja aineiston keskiarvon priorijakauman tarkkuudella, ja n kertaa aineiston tarkkuudella painotettu summa. Havaitaan siis, että kun otoskoko kasvaa, priorijakauman kontribuutio posteriorijakauman odotusarvoon ja tarkkuuteen vähenee, ja aineiston kontribuutio kasvaa, lineaarisesti.

Käsitellään vielä vastakkainen esimerkki, eli normaalijakauma, jonka odotusarvo on tunnettu, mutta varianssi tuntematon.

Esimerkki 2.5. Oletetaan satunnaismuuttujat $Y_1, \ldots, Y_n \perp\!\!\!\perp | \, \sigma^2$, $Y_i \sim N(\theta_0, \sigma^2)$ kaikille $i = 1, \ldots, n$, missä $\theta_0 \in \mathbb{R}$ on tunnettu vakio.

Tämän mallin uskottavuusfunktio on

$$\begin{aligned} p(\mathbf{y}|\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \theta_0)^2}{2\sigma^2}\right) \\ &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \theta_0)^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-n/2} \exp\left(-\frac{nv}{2\sigma^2}\right), \end{aligned}$$

missä

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2.$$

Valitaan varianssin σ^2 priorijakaumaksi ns. skaalattu ja invertoitu khiin neliön jakauma (scaled inverse-χ^2 distribution) vapausasteella ν_0, ja skaalaparametrilla σ_0^2, eli $\sigma^2 \sim$ Inv-$\chi^2(\nu_0, \sigma_0^2)$, jolloin

$$\begin{aligned} p(\sigma^2) &= \frac{(\nu_0/2)^{-\nu_0/2}}{\Gamma(\nu_0/2)}(\sigma_0^2)^{\nu_0/2}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right), \end{aligned}$$

kun $\sigma^2 > 0$.

Siten normalisoimattomaksi posteriorijakaumaksi saadaan

$$\begin{aligned} p(\sigma^2|\mathbf{y}) &\propto p(\sigma^2)p(\mathbf{y}|\sigma^2) \\ &\propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)(\sigma^2)^{-n/2} \exp\left(-\frac{nv}{2\sigma^2}\right) \\ &= (\sigma^2)^{-((\nu_0+n)/2+1)} \exp\left(-\frac{(\nu_0\sigma_0^2+nv)}{2\sigma^2}\right) \\ &= (\sigma^2)^{-(\nu_n/2+1)} \exp\left(-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right), \end{aligned}$$

missä

$$\nu_n = \nu_0 + n,$$

ja

$$\sigma_n^2 = \frac{\nu_0\sigma_0^2 + nv}{\nu_0 + n}.$$

Varianssin σ^2 posteriorijakauma on siis jälleen skaalattu invertoitu χ^2-jakauma, eli

$$\sigma^2|\mathbf{y} \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2),$$

missä vapausaste on priorijakauman vapausasteen ν_0 ja otoskoon n summa, ja skaalaparametri on vapausasteilla painotettu keskiarvo priorin skaalaparametrista ν_0 ja aineiston keskineliövirheestä v (vertaa Huomautus 2.4).

2.2 Eksponenttiperhe

Kerrataan aluksi lyhyesti, mitä tarkoitetaan tyhjentävällä tunnusluvulla. Aihetta käsitellään tarkemmin esimerkiksi Tilastollisen päättelyn kurssin luentomonisteen [5] luvussa 4, tai Essentials of statistical inference-kirjan [6] luvussa 6.

2.2.1 Tyhjentävä tunnusluku

Mitä tahansa havaitun aineiston $\mathbf{y}$ skalaari- tai vektoriarvoista funktiota $\mathbf{t}(\mathbf{y})$ kutsutaan **tunnusluvuksi**. Tunnuslukuja ovat esimerkiksi:

- Aineiston summa $t(\mathbf{y}) = \sum_{i=1}^n y_i$,
- Otoskeskiarvon ja otoskeskihajonnan muodostama vektori

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

- Suurin havainto $y_{(1)} = \max\{y_1, \ldots, y_n\}$,
- Aineisto itse $\mathbf{y} = (y_1, \ldots, y_n)$.

Jos korvataan havainnot $\mathbf{y}$ tunnusluvun määritelmässä sitä vastaavalla satunnaismuuttujalla $\mathbf{Y}$, satunnaismuuttujaa $\mathbf{T} = \mathbf{t}(\mathbf{Y})$ kutsutaan myös tunnusluvuksi.

Tyhjentävän tunnusluvun määritelmän mukaan (löyhästi) tunnusluku $\mathbf{T}$ on parametrin $\boldsymbol{\theta}$ **tyhjentävä tunnusluku**, jos satunnaismuuttujan $\mathbf{Y}$ jakauma ehdolla $\mathbf{T}$ ei riipu parametrista $\boldsymbol{\theta}$.

Yleensä tunnusluvun tyhjentävyyden tarkasteluun käytetään seuraavaa, varsinaista määritelmää näppärämpää, tulosta:

Lause 2.6. *(Neymanin) Faktorointikriteeri. Tunnusluku* $\mathbf{T} = \mathbf{t}(\mathbf{Y})$ *on parametrin* $\boldsymbol{\theta}$ *tyhjentävä tunnusluku, jos ja vain on olemassa funktiot* h *ja* g*, siten että sen tiheysfunktio (vastaavasti ptnf) voidaan esittää muodossa*

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y})g(\mathbf{t}(\mathbf{y}), \boldsymbol{\theta}).$$

Huomautus 2.7. Koska uskottavuusfunktiosta voidaan pudottaa pois kaikki parametrista $\boldsymbol{\theta}$ riippumattomat vakiot, tässä tapauksessa siis $h(\mathbf{y})$, faktorointikriteeri tarkoittaa yksinkertaisesti sitä, että tunnusluku $\mathbf{t}$ on tyhjentävä, jos ja vain jos uskottavuusfunktio riippuu aineistosta ainoastaan sen kautta.

Esimerkki 2.8. Tarkastellaan riippumattomia havaintoja Poisson-jakaumasta:

$$Y_1, \ldots, Y_n \sim \text{Poisson}(\theta), \quad Y_1, \ldots, Y_n \perp\!\!\!\perp | \, \theta.$$

Satunnaisvektorin $\mathbf{Y} = (Y_1, \ldots, Y_n)$ ptnf voidaan esittää muodossa

$$p(\mathbf{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \frac{1}{\prod_{i=1}^{n} y_i!} e^{-n\theta} \theta^{\sum_{i=1}^{n} y_i} = h(\mathbf{y})g(t(\mathbf{y}), \theta),$$

missä

$$h(\mathbf{y}) = \frac{1}{\prod_{i=1}^{n} y_i!},$$
$$g(t, \theta) = e^{-n\theta}\theta^t,$$

ja

$$t(\mathbf{y}) = \sum_{i=1}^{n} y_i.$$

Havaintojen summa on siis tyhjentävä tunnusluku riippumattomille samaa Poisson-jakaumaa noudattaville satunnaismuuttujille. Seuraavaksi tarkastellaan, minkä jakaumien i.i.d.-malleille on mahdollista löytää tällainen tyhjentävä tunnusluku, jonka dimensio ei riipu havaintojen määrästä n.

2.2.2 Eksponenttiperheen jakaumat

Suurin osa käytetyimmistä parametrisista jakaumista, esimerkiksi binomijakauma, normaalijakauma, Poisson-jakauma, eksponenttijakauma ja gammajakauma kuuluvat eksponenttiperheeseen. Tällä tarkoitetaan, että niiden tiheysfunktiot (tai vastaavasti ptnf:t) ovat tiettyä, seuraavaksi määriteltävää, muotoa.

Määritelmä 2.9. Jos satunnaisvektorin $\mathbf{Y}$ jakauma riippuu d-ulotteisesta parametrista $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ tiheysfunktion (tai vastaavasti ptnf:n) $p(\mathbf{y}|\boldsymbol{\theta})$, joka on muotoa

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y})g(\boldsymbol{\theta}) \exp\left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) u_j(\mathbf{y}) \right\},$$

kautta, sanotaan, että sen jakauma kuuluu d-ulotteiseen **eksponenttiperheeseen**.

Vektoria $\boldsymbol{\phi}(\boldsymbol{\theta}) = (\phi_1(\boldsymbol{\theta}), \ldots, \phi_d(\boldsymbol{\theta}))$ kutsutaan jakauman **luonnolliseksi parametriksi**, ja jakauma on mahdollista parametroida uudelleen sen kautta.

Huomautus 2.10. Jos jakauma kuuluu eksponenttiperheeseen, ja sen alusta ei riipu parametrista $\boldsymbol{\theta}$, sanotaan että jakauma kuuluu **säännölliseen** eksponenttiperheeseen.

Esimerkki epäsäännöllisestä eksponenttiperheen jakaumasta on välin $(0, \theta)$ tasajakauma $\mathrm{Tas}(0, \theta)$: tämän jakauman alusta, eli joukko, jossa tiheysfunktio on aidosti positiivinen, riippuu tarkasteltavasta parametrista θ.

Huomautus 2.11. Suoraan eksponenttijakauman määritelmästä nähdään, että faktorointikriteerin nojalla d-ulotteiseen eksponenttiperheeseen kuuluvalla jakaumalla on aina tyhjentävä tunnusluku

$$\mathbf{u}(\mathbf{y}) = (u_1(\mathbf{y}), \ldots, u_d(\mathbf{y})),$$

jonka dimensio on siis sama kuin parametrin $\boldsymbol{\theta}$.

Huomautus 2.12. Jos satunnaismuuttujat $\mathbf{Y} = (Y_1, \ldots, Y_n)$ ovat riippumattomasti samoin jakautuneita ehdolla parametri $\boldsymbol{\theta}$, ja niiden tiheysfunktio (vastaavasti ptnf) kuuluu eksponenttiperheeseen, eli on muotoa

$$p(y_i|\boldsymbol{\theta}) = h(y_i)g(\boldsymbol{\theta}) \exp\left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) u_j(y_i) \right\},$$

niin aineiston, eli satunnaisvektorin $\mathbf{Y}$, jakauma on muotoa

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) &= \left(\prod_{i=1}^{n} h(y_i) \right) g(\boldsymbol{\theta})^n \exp\left\{ \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) u_j(y_i) \right) \right\} \\ &= \left(\prod_{i=1}^{n} h(y_i) \right) g(\boldsymbol{\theta})^n \exp\left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) \left(\sum_{i=1}^{n} u_j(y_i) \right) \right\}, \end{aligned}$$

eli se kuuluu myös eksponenttiperheeseen samalla luonnollisella parametrilla $\boldsymbol{\phi}(\boldsymbol{\theta})$. Siten aineiston koosta n riippumatta aineiston jakaumalla on tässä tapauksessa d-ulotteinen tyhjentävä tunnusluku

$$\mathbf{t}(\mathbf{y}) = \sum_{i=1}^{n} \mathbf{u}(y_i) = \left(\sum_{i=1}^{n} u_1(y_i), \ldots, \sum_{i=1}^{n} u_d(y_i) \right).$$

Osoittautuu, että muutamaa poikkeusta lukuunnottamatta kaikki jakaumat, joille on mahdollista löytää tällainen aineiston koosta riippumaton tyhjentävä tunnusluku, kuuluvat eksponenttiperheeseen.

Esimerkki 2.13. Osoitetaan, että Poisson-jakauma kuuluu eksponenttiperheeseen, ja havainnollistetaan vielä edellistä tulosta, eli näytetään, että siten myös i.i.d. Poisson-mallin jakauma kuuluu eksponenttiperheeseen samalla luonnollisella parametrilla.

(a) Tarkastellaan ensin yhtä havaintoa Poisson-jakaumasta $Y \sim \text{Poisson}(\theta)$. Sen ptnf voidaan kirjoittaa muodossa

$$p(y|\theta) = \frac{1}{y!} e^{-\theta} \theta^y = \frac{1}{y!} e^{-\theta} \exp\{y \log \theta\} = h(y) g(\theta) \exp\{\phi(\theta) u(y)\},$$

missä

$$h(y) = \frac{1}{y!}, \quad g(\theta) = e^{-\theta}, \quad \phi(\theta) = \log \theta, \text{ ja } u(y) = y.$$

Poisson-jakauma siis kuuluu yksiulotteiseen eksponenttiperheeseen luonnollisella parametrilla $\phi(\theta) = \log \theta$.

(b) Tarkastellaan riippumattomia havaintoja Poisson-jakaumasta:

$$Y_1, \ldots, Y_n \sim \text{Poisson}(\theta), \quad Y_1, \ldots, Y_n \perp\!\!\!\perp \theta.$$

Satunnaisvektorin $\mathbf{Y} = (Y_1, \ldots, Y_n)$ ptnf voidaan esittää muodossa

$$\begin{aligned} p(\mathbf{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) &= \frac{1}{\prod_{i=1}^{n} y_i!} e^{-n\theta} \theta^{\sum_{i=1}^{n} y_i} \\ &= \prod_{i=1}^{n} \frac{1}{y_i!} \left(e^{-\theta}\right)^n \exp\left\{ \log \theta \sum_{i=1}^{n} y_i \right\} \\ &= \left(\prod_{i=1}^{n} h(y_i) \right) g(\theta)^n \exp\left\{ \phi(\theta) \sum_{i=1}^{n} u(y_i) \right\}. \end{aligned}$$

Riippumaton otos Poisson-jakaumasta siis kuuluu eksponenttiperheeseen samalla luonnollisella parametrilla $\log \theta$, ja sillä on yksiulotteinen tyhjentävä tunnusluku

$$t(\mathbf{y}) = \sum_{i=1}^{n} u(y_i) = \sum_{i=1}^{n} y_i.$$

joka on summa yksittäisten havaintojen tunnusluvuista $u(y_i) = y_i$.

2.2.3 Konjugaattianalyysi

Määrittelimme luvun alussa konjugaattipriorin jollekin aineiston jakaumalle $p(\mathbf{y}|\boldsymbol{\theta})$ siten, että tällä priorilla $p(\boldsymbol{\theta})$ posteriorijakauma $p(\boldsymbol{\theta}|\mathbf{y})$ kuuluu samaan jakaumaperheeseen kuin aineiston jakauma. Tämä määritelmä on kuitenkin vielä epämääräinen, sillä se täyttyy, jos jakaumaperhe määritellään riittävän laajasti, esimerkiksi kaikkien parametristen jakaumien joukko.

Meidän täytyy siis määritellä tarkemmin, mitä tarkoitamme konjugaattipriorilla, jotta niiden käyttäminen mahdollistaisi prioriennustejakauman

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

ja siten posteriorijakauman $p(\boldsymbol{\theta}|\mathbf{y})$, ratkaisemisen suljetussa muodossa. Lisäksi konjugaattipriorin tulee olla muodoltaan riittävän monimutkainen, jotta se mahdollistaa parametrin arvoa koskevien uskomustemme muotoilemisen.

Osoittautuu, että edellä määritelty tyhjentävän tunnusluvun käsite auttaa löytämään tällaisen konjugaattipriorien perheen. Jos nimittäin parametrilla $\boldsymbol{\theta}$ on tyhjentävä tunnusluku $\mathbf{t}(\mathbf{y})$, faktorointikriteerin nojalla

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})g(\boldsymbol{\theta}, \mathbf{t}(\mathbf{y})),$$

eli posteriorijakauma riippuu aineistosta ainoastaan tyhjentävän tunnusluvun $\mathbf{t}$ kautta!

Jos lisäksi tyhjentävän tunnusluvun dimensio d ei riipu otoskoosta n, niin voimme korvata sen uskottavuusfunktion lausekkeessa hyperparametrilla $\boldsymbol{\tau} \in \mathbb{R}^d$, ja näin saamme priorijakauman, joka on samaa muotoa kuin uskottavuusfunktio. Määritellään tämä ns. luonnollisten konjugaattipriorien perhe (jatkossa puhuessamme konjugaattiprioreista tarkoitamme juuri tämän luonnollisen konjugaattiperheen jäseniä) vielä tarkemmin.

Merkitään vielä jatkossa satunnaisvektorin $\mathbf{Y}$ dimensio, eli aineiston otoskoko, tyhjentävän tunnusluvun ensimmäiseksi komponentiksi:

$$\mathbf{t}_n(\mathbf{y}) = (n, \mathbf{t}(\mathbf{y})).$$

Huomaa, että otoskokoa n ei kuitenkaan ajatella satunnaiseksi, vaan se on etukäteen kiinnitetty vakio.

Määritelmä 2.14. Jos kaikille jakaumaperheen $\mathcal{F}$ jakaumille $p(\mathbf{y}|\boldsymbol{\theta})$ on aina olemassa tyhjentävä tunnusluku $\mathbf{t}_n(\mathbf{Y})$, jonka dimensio $d+1$ on riippumaton satunnaisvektorin $\mathbf{Y}$ dimensiosta n, eli nämä jakaumat voidaan esittää muodossa

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y})g(\boldsymbol{\theta}, \mathbf{t}_n(\mathbf{y})),$$

tälle jakaumaperheelle voidaan määritellä **luonnollinen konjugaattiperhe**

$$\mathcal{P} = \left\{ \frac{g(\boldsymbol{\theta}, \boldsymbol{\tau})}{c(\boldsymbol{\tau})} : \boldsymbol{\tau} \in \mathcal{T} \right\},$$

missä

$$\mathcal{T} = \left\{ \boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_d) : 0 < c(\boldsymbol{\tau}) = \int g(\boldsymbol{\theta}, \boldsymbol{\tau}) \, \mathrm{d}\boldsymbol{\theta} < \infty \right\}$$

Luonnolliset konjugaattipriorit saadaan siis korvaamalla tyhjentävä tunnusluku $\mathbf{t}_n$ uskottavuusfunktion lausekkeessa hyperparametrilla $\boldsymbol{\tau}$, ja normalisoimalla tulos todennäköisyysjakauman tiheysfunktioksi parametrin $\boldsymbol{\theta}$ suhteen. Ehto $\boldsymbol{\tau} \in \mathcal{T}$ tarkoittaa yksinkertaisesti sitä, että tämä normalisointi on mahdollista tehdä.

Huomautus 2.15. Jakaumaperheen, johon kuuluvat satunnaisvektorit $\mathbf{Y} = (Y_1, \ldots, Y_n)$, joiden komponentit noudattavat riippumattomasti samaa eksponenttiperheen jakaumaa, eli jonka jäsenet voidaan kirjoittaa muodossa

$$p(\mathbf{y}|\boldsymbol{\theta}) = \left(\prod_{i=1}^{n} h(y_i) \right) g(\boldsymbol{\theta})^n \exp \left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) \mathbf{t}_j(\mathbf{y}) \right\},$$

luonnolliset konjugaattipriorit ovat määritelmän 2.14 nojalla muotoa

$$p(\boldsymbol{\theta}|\boldsymbol{\tau}) = \frac{g(\boldsymbol{\theta})^{\tau_0} \exp \left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) \tau_j \right\}}{c(\boldsymbol{\tau})},$$

missä $\boldsymbol{\tau}$:lle pätee, että

$$0 < c(\boldsymbol{\tau}) = \int g(\boldsymbol{\theta})^{\tau_0} \exp \left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) \tau_j \right\} \mathrm{d}\boldsymbol{\theta} < \infty.$$

Esimerkki 2.16. Jatkoa esimerkille 2.13. Tarkastellaan jälleen riippumattomia samaa Poisson-jakaumaa noudattavia satunnaismuuttujia

$$Y_1, \ldots, Y_n \sim \text{Poisson}(\theta), \quad Y_1, \ldots, Y_n \perp\!\!\!\perp \mid \theta.$$

Jakauman $p(\mathbf{y}|\theta)$ luonnollinen konjugaattipriori on edellisen huomion nojalla muotoa

$$p(\theta) \propto g(\theta)^{\tau_0} \exp \{ \phi(\theta) \tau_1 \} = e^{-\theta \tau_0} \theta^{\tau_1}.$$

Käytetään hieman kätevämpää parametrisointia $\alpha = \tau_1 + 1$ ja $\beta = \tau_0$, jolloin

$$p(\theta) \propto \theta^{\alpha - 1} e^{-\beta \theta}$$

tunnistetaan gammajakauman tiheysfunktion ytimeksi, eli luonnolliset priorijakaumat ovat muotoa Gamma(α, β), missä $\alpha, \beta > 0$.

Edellä olemme laskeneet posteriorijakaumat ja posterioriennustejakaumat binomijakaumalle, Poisson-jakaumalle ja normaalijakaumalle tunnetulla varianssilla. Laskuista voi havaita, että ne noudattavat varsin samaa kaavaa. Osoittautuukin, että konjugaattiprioria käytettäessä säännöllisten eksponenttiperheen jakaumien posteriorijakauma ja posterioriennustejakauma voidaan ratkaista seuraavassa yleisessä muodossa.

Lause 2.17. *Olkoot* $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *riippumattomia samoin jakautuneita satunnaismuuttujia, joiden jakauma* $p(\mathbf{y}|\boldsymbol{\theta})$ *kuuluu säännölliseen eksponenttiperheeseen (eli on huomautuksen 2.12 muotoa) tyhjentävällä tunnusluvulla* $\mathbf{t}(\mathbf{y})$*, ja* $p(\boldsymbol{\theta}|\boldsymbol{\tau}) \in \mathcal{P}$ *tämän jakauman luonnollinen konjugaattipriori.*

(i) Parametrin $\boldsymbol{\theta}$ *posteriorijakauma on*

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\tau}) = p(\boldsymbol{\theta}|\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y})).$$

(ii) Aineiston prioriennustejakauma on

$$p(\mathbf{y}|\boldsymbol{\tau}) = \left(\prod_{i=1}^{n} h(y_i)\right) \frac{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}))}{c(\boldsymbol{\tau})}.$$

(iii) Uusien riippumattomien havaintojen samasta jakaumasta $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_m)$ *posterioriennustejakauma on*

$$p(\tilde{\mathbf{y}}|\mathbf{y}, \boldsymbol{\tau}) = \left(\prod_{i=1}^{m} h(\tilde{y}_i)\right) \frac{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}) + \mathbf{t}_m(\tilde{\mathbf{y}}))}{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}))}.$$

Todistus. (i) Bayesin kaavan nojalla

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\tau}) &\propto p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\tau}) \\ &\propto g(\boldsymbol{\theta})^n \exp\left\{\sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) t_j(\mathbf{y})\right\} g(\boldsymbol{\theta})^{\tau_0} \exp\left\{\sum_{j=1}^{d} \phi_j(\boldsymbol{\theta}) \tau_j\right\} \\ &\propto g(\boldsymbol{\theta})^{n+\tau_0} \exp\left\{\sum_{j=1}^{d} \phi_j(\boldsymbol{\theta})(\tau_j + t_j(\mathbf{y}))\right\} \\ &\propto p(\boldsymbol{\theta}|\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y})). \end{aligned}$$

(ii) Prioriennustejakauma saadaan integroimalla yhteisjakaumaa $p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\tau})$ parametrin $\boldsymbol{\theta}$ yli:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\tau}) &= \int p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\tau}) \, \mathrm{d}\boldsymbol{\theta} \\ &= \left(\prod_{i=1}^{n} h(y_i) \right) \frac{1}{c(\boldsymbol{\tau})} \int g(\boldsymbol{\theta})^{n+\tau_0} \exp\left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta})(t_j(\mathbf{y}) + \tau_j) \right\} \mathrm{d}\boldsymbol{\theta} \\ &= \left(\prod_{i=1}^{n} h(y_i) \right) \frac{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}))}{c(\boldsymbol{\tau})}, \end{aligned}$$

(iii) ja posterioriennustejakauma samalla periaatteella:

$$\begin{aligned} &p(\tilde{\mathbf{y}}|\mathbf{y}, \boldsymbol{\tau}) = \int p(\tilde{\mathbf{y}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\tau}) \, \mathrm{d}\boldsymbol{\theta} \\ &= \left(\prod_{i=1}^{m} h(\tilde{y}_i) \right) \frac{1}{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}))} \int g(\boldsymbol{\theta})^{\tau_0+n+m} \exp\left\{ \sum_{j=1}^{d} \phi_j(\boldsymbol{\theta})(\tau_j + t_j(\mathbf{y}) + t_j(\tilde{\mathbf{y}})) \right\} \mathrm{d}\boldsymbol{\theta} \\ &= \left(\prod_{i=1}^{m} h(\tilde{y}_i) \right) \frac{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}) + \mathbf{t}_m(\tilde{\mathbf{y}}))}{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}))}. \end{aligned}$$

□

Esimerkki 2.18. Jatkoa esimerkeille 2.13 ja 2.16. Käytetään luonnollista konjugaattipriooria Gamma(α, β) riippumattomille samaa Poisson-jakaumaa Poisson(θ) noudattaville satunnaismuuttujille. Merkitään priorijakauman parametreja $\boldsymbol{\tau} = (\beta, \alpha)$.

(a) Nyt lauseen 2.17 ensimmäisen kohdan nojalla parametrin θ posteriorijakauma on Gamma$(\alpha + t(\mathbf{y}), \beta + n)$, missä $t(\mathbf{y}) = \sum_{i=1}^{n} y_i$ on havaintojen summa.

(b) Käytetään lauseen 2.17 toista kohtaa, jonka nojalla prioriennustejakauma $p(\mathbf{y})$ saadaan ratkaistua gammajakauman normalisointivakion käänteisluvun

$$c(\boldsymbol{\tau}) = \frac{\Gamma(\alpha)}{\beta^{\alpha}}$$

avulla:

$$\begin{aligned} p(\mathbf{y}) &= \left(\prod_{i=1}^{n} h(y_i) \right) \frac{c(\boldsymbol{\tau} + \mathbf{t}_n(\mathbf{y}))}{c(\boldsymbol{\tau})} \\ &= \frac{1}{(\prod_{i=1}^{n} y_i)!} \frac{\Gamma(\alpha + t(\mathbf{y}))}{(\beta + n)^{\alpha + t(\mathbf{y})}} \frac{\beta^{\alpha}}{\Gamma(\alpha)}. \end{aligned}$$

(c) Seuraavan riippumattoman havainnon samasta Poisson-jakaumasta $\tilde{Y}_1 \sim \text{Poisson}(\theta)$ posterioriennustejakauma saadaan samalla tavalla sijoittamalla sijoittamalla oikeat termit gammajakauman normalisointivakion käänteislukuun:

$$\begin{aligned} p(\tilde{y}_1|\mathbf{y}, \boldsymbol{\tau}) &= h(\tilde{y_1}) \frac{c(\boldsymbol{\tau} + t_n(\mathbf{y}) + t_1(\tilde{y}_1))}{c(\boldsymbol{\tau} + t_n(\mathbf{y}))} \\ &= \frac{1}{\tilde{y}_1!} \frac{\Gamma(\alpha + t(\mathbf{y}) + \tilde{y}_1)}{(\beta + n + 1)^{\alpha + t(\mathbf{y}) + \tilde{y}_1}} \frac{(\beta + n)^{\alpha + t(\mathbf{y})}}{\Gamma(\alpha + t(\mathbf{y}))} \\ &= \frac{\Gamma(\alpha_n + \tilde{y}_1)}{\Gamma(\alpha_n)\tilde{y}_1!} \left(\frac{\beta_n}{\beta_n + 1}\right)^{\alpha_n} \left(\frac{1}{\beta_n + 1}\right)^{\tilde{y}_1}. \end{aligned}$$

Posterioriennustejakauma yhdelle uudelle havainnolle on siten negatiivinen binomijakauma (vrt. esimerkki 2.1) Neg-bin(α_n, β_n), missä

$$\alpha_n = \alpha + t(\mathbf{y}),$$

ja

$$\beta_n = \beta + n.$$

2.3 Priorin valinta

Jos meillä ei ole selkeää käsitystä parametrin todellisesta arvosta, tai haluamme että nämä ennakkokäsityksemme vaikuttaisivat mahdollisimman vähän päättelyn tuloksiin, valitsemme priorijakaumaksi jakauman, joka jakaa todennäköisyysmassan mahdollisimman tasaisesti parametriavaruuteen. Tällöin puhutaan **epäinformatiivisesta** priorista. Epäinformatiivisen priorin tarkoituksena on antaa päättelyssä aineistolle mahdollisimman suuri, ja priorijakaumalle mahdollisimman pieni rooli, ja tehdä päättelystä tässä mielessä mahdollisimman objektiivista.

Epäinformatiivisen priorin käsite on informaali, sillä se riippuu jakauman parametrisoinnista: jos priori on tasainen parametrin suhteen, se ei yleensä säily tasaisena tämän parametrin muunnokselle. Jos parametriavaruus ei ole rajoitettu, niin myöskään tällöin todennäköisyysmassan jakaminen täysin tasaisesti ei onnistu, sillä tällöin jakaumaa $p(\boldsymbol{\theta}) \propto 1$ ei voida normalisoida todennäköisyysjakaumaksi. Tarkastellaan näitä ongelmia esimerkin avulla.

Esimerkki 2.19. Jatkoa esimerkille 1.5. Heitetään nastaa n kertaa, ja merkitään kertojen määrää, jolloin nasta laskeutuu kanta alaspäin, satunnaismuuttujalla Y, jolloin

$$Y \,|\, \theta \sim \text{Bin}(n, \theta),$$

ja käytetään konjugaattiprioria $\theta \sim \text{Beta}(\alpha, \beta)$.

Miten meidän tulisi valita priorijakauman parametrit α ja β, jos haluamme, että priorijakauma on epäinformatiivinen, eli jos haluamme, että ennakkokäsityksemme parametrin θ mahdollisista arvoista vaikuttaa posteriorijakaumaamme mahdollisimman vähän?

Tarkastellaan mahdollisia vastauksia:

(i) Ilmeisin vaihtoehto on käyttää tasajakaumaa $U(0, 1)$, eli jakaumaa $\text{Beta}(1, 1)$. Tämä priori vastaa ennakkokäsitystä, että kaikki parametrin arvot ovat yhtä todennäköisiä, ja jakaa todennäköisyysmassan tasaisesti kaikille mahdollisille parametrin θ arvoille, eli välille $(0, 1)$.

Betajakauman odotusarvo on $\frac{\alpha}{\alpha+\beta}$, joten parametrin θ posteriorijakauman $\text{Beta}(\alpha + y, \beta + n - y)$ odotusarvo on

$$E[\theta \mid Y] = \frac{\alpha + y}{\beta + n - y}.$$

Tästä muotoilusta nähdään, että hyperparametrit voidaan tulkitaa **pseudo-havaintoina**: oikeisiin havaintoihin lisätään priorijakaumasta α onnistumista, ja β epäonnistumista, jolloin priorijakauman implikoiman pseudo-otoksen koko on $\alpha + \beta$. Siten käytettäessä priorina tasajakaumaa $\text{Beta}(1, 1)$ posteriorijakauman odotusarvoksi saadaan

$$E[\theta \mid Y] = \frac{1 + y}{1 + n - y}.$$

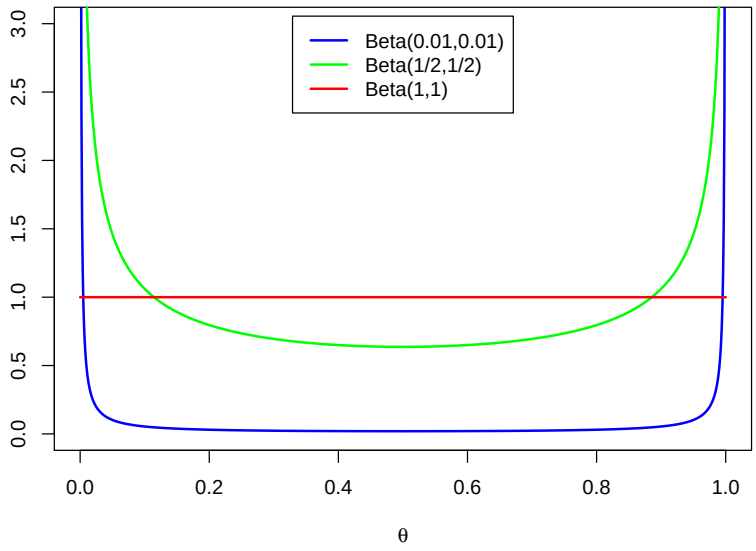
Tämä havainnollistaa priorijakauman tulkintaa tasoituksena (**smoothing**) tai **regularisaationa** (regularisation), joita käytetään myös ei-Bayesiläisessa mallintamisessa, ja jotka voidaan monesti tulkita priorijakauman käyttämisenä parametrille: yhden pseudohavainnon lisääminen kumpaankin luokkaan, eli nastan laskeutumiseen kanta ylös- tai alaspäin voidaan siis tulkita tasajakauman $\text{Beta}(1, 1)$ käyttämisenä priorina. Tasajakaumaa priorina binomimallille kutsutaan monesti *Laplacen prioriksi*.

Mutta entä jos mallin parametrina halutaankin käyttää logaritmista vedonlyöntisuhdetta (**log-odds**)? Tällöin parametri on

$$\phi = \text{logit}(\theta) = \log\left(\frac{\theta}{1 - \theta}\right),$$

missä θ on onnistumistodennäköisyys. Tämän parametrisoinnin kätevä puoli on se, että uusi parametri ϕ voi saada arvoja koko reaaliakselilta. Nyt kuitenkaan priori, joka on

Kuva 2.1: Heikko priori, Jeffreys'n priori ja tasajakaumapriori binomimallille.



tasajakautunut parametrin θ suhteen ei noudata tasajakaumaa ϕ:n suhteen. Vastaavasti ϕ:lle ei voida edes määrittää tasajakaumaa

$$p(\phi) \propto 1,$$

joka olisi oikea todennäköisyysjakauma: vakion integraali yli koko reaaliakselin on aina ääretön, joten normalisoimatonta jakaumaa $p(\phi) \propto 1$ ei voida normalisoida todennäköisyysjakaumaksi.

Tämä ei kuitenkaan välttämättä ole ongelma, sillä osoittautuu, että priorin ei välttämättä tarvitse olla kunnollinen todennäköisyysjakauma, vaan riittää että posteriorijakaumasta tulee todennäköisyysjakauma.

Yksiulotteisessa tapauksessa intuitiivinen tasajakauman käyttäminen epäinformatiivisena priorina osoittautuu kuitenkin ongelmalliseksi korkeampiulotteisilla parametreilla: korkeammissa ulottuvuuksissa tasajakauma sijoittaa suurimman osan jakauman todennäköisyysmassasta parametriavaruuden reunoille, jolloin siitä tuleekin informatiivinen siten, että se suosii äärimmäisiä arvoja.

2.3.1 Jeffreys'n priori

Vastauksena intuitiivisen tasajakauman ongelmiin fyysikko/tilastotieteilijä Harold Jeffreys (1891-1989) kehitti periaatteen, jolla saadaan johdettua uskottavuusfunktiolle priorijakauma, joka on invariantti mallin parametrisoinnin suhteen, eli ei riipu parametrisoinnista, jota uskottavuusfunktiolle käytetään.

Merkitään Jeffreys'n priorin parametrille θ (oletetaan tässä, että parametri on yksiulotteinen, sillä Jeffreys'n priori toimii parhaiten nimenomaan yksiulotteisessa tapauksessa) tiheysfunktiota $J(\theta) = p(\theta)$. Se määritellään

$$J(\theta) \propto \sqrt{i(\theta)},$$

missä

$$i(\theta) = E\left[-\frac{\partial^2 \log f_{\mathbf{Y}|\Theta}(\mathbf{Y}|\theta)}{\partial\theta^2}\,\middle|\,\Theta = \theta\right]$$

on parametrin θ Fisherin informaatio, joka siis on aineiston logaritmisen uskottavuusfunktion vastaluvun odotusarvo[2].

Jos siis käytetään mallin parametrina alkuperäisen parametrin muunnosta $\phi = g(\theta)$, saadaan sama tulos riippumatta siitä, johdetaanko Jeffreys'n priori suoraan uudesta parametrille:

$$J_\Phi(\phi) \propto \sqrt{i_\Phi(\phi)},$$

vai sijoitetaanko muunnos suoraan alkuperäisen mallin Jeffreys'n prioriin:

$$J_\Theta(g(\theta)) \propto \sqrt{i_\Theta(g(\theta))},$$

eli tällöin

$$J_\Phi(\phi) = J_\Theta(g(\theta)).$$

Yllä merkittiin selkeyden vuoksi alaindeksillä parametria, jonka suhteen Fisherin informaatio ja Jeffreys'n priori lasketaan.

Esimerkki 2.20. Jatkoa esimerkille 2.19. Lasketaan binomimallille Jeffreys'n priori. Binomijakaumaa $\mathrm{Bin}(n, \theta)$ noudattavan satunnaismuuttujan Y logaritminen uskottavuusfunktio on

$$\log p(y|\theta) \propto y \log\theta + (n - y)\log(1 - \theta).$$

[2]Yllä merkittiin $\mathbf{Y}$ tiheysfunktion argumenttina isolla kirjaimella sen korostamiseksi, että Fisherin informaatiota laskettaessa se käsitetään satunnaismuuttujana, jonka suhteen odotusarvo lasketaan kiinteällä parametrin arvolla θ. Kyseessä on siis tilastollisen päättelyn kurssilta tuttu käsite, jota merkittiin siellä $E[-l''(\theta; \mathbf{Y})]$.

Tämän ensimmäinen derivaatta on

$$\frac{\partial \log p(y|\theta)}{\partial \theta} \propto \frac{y}{\theta} - \frac{n-y}{1-\theta},$$

toinen derivaatta on

$$\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2} \propto -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2},$$

ja siten käyttämällä hyväksi tietoa, että binomijakauman odotusarvo on

$$E[Y|\theta] = n\theta,$$

parametrin θ Fisherin informaatioksi saadaan

$$i(\theta) = E\left[-\frac{\partial^2 \log p(Y|\theta)}{\partial \theta^2} \,\middle|\, \Theta = \theta\right] \propto \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Näin Jeffreys'n priori on

$$J(\theta) \propto \sqrt{i(\theta)} \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

Tämä voidaan tunnistaa betajakauman tiheysfunktion ytimeksi, eli

$$J(\theta) \stackrel{d}{=} \text{Beta}(1/2, 1/2).$$

Jeffreys'n priorilla lasketun posteriorijakauman $\text{Beta}(y+1/2, n-y+1/2)$ odotusarvo on

$$E[\theta \,|\, y] = \frac{y+1/2}{n+1},$$

eli se vastaa puolen pseudohavainnon lisäämistä sekä onnistumisiin että epäonnistumisiin.

2.3.2 Epäoleelliset priorit

Vaikka binomijakauman tapauksessa Jeffreys'n priori sattui olemaan konjugaattipriori, näin ei ole yleisessä tapauksessa. Tarkastellaan esimerkkinä eksponenttijakaumaa.

Esimerkki 2.21. Oletetaan satunnaisotos $Y_1, \ldots, Y_n$ eksponenttijakaumasta $\text{Exp}(\theta)$.

(a) Johdetaan mallin Jeffreys'n priori. Fisherin informaatio on additiivinen riippumattomille satunnaismuuttujlle, joten i.i.d.-mallin Fisherin informaatio $i_n(\theta)$ on n kertaa sen yksittäisen komponentin informaatio $i(\theta)$:

$$i_n(\theta) = n \cdot i(\theta) \propto i(\theta).$$

Siten Jeffreys'n priorin selvittämiseksi riittää laskea yhden komponentin Y_1 Fisherin informaatio. Satunnaismuuttujan $Y_1 \sim \text{Exp}(\theta)$ log-uskottavuusfunktio on

$$\log p(y_1|\theta) = \log \theta - y_1\theta.$$

Tämän ensimmäinen derivaatta on

$$\frac{\partial \log p(y_1|\theta)}{\partial \theta} = \frac{1}{\theta} - y_1,$$

toinen derivaatta on

$$\frac{\partial^2 \log p(y_1|\theta)}{\partial \theta^2} = -\frac{1}{\theta^2}$$

ja siten mallin Jeffreys'n priori on

$$J(\theta) \propto \frac{1}{\theta}, \quad \text{kun } \theta > 0.$$

Koska tämän integraali yli positiivisen reaaliakselin on ääretön, kyseessä ei ole todennäköisyysjakauman tiheysfunktio.

(b) Havaitaan aineisto $\mathbf{y} = (y_1, \ldots, y_n)$. Käytettäessä Jeffreyn prioria $p(\theta) \propto \frac{1}{\theta}$ parametrin θ normalisoimattomaksi posteriorijakaumaksi saadaan

$$p(\theta|\mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) \propto \theta^{n-1}e^{-\sum_{i=1}^n y_i\theta},$$

eli posteriorijakauma on Gamma$(n, \sum_{i=1}^n y_i)$. Vaikka priorijakauma ei ole oikea todennäköisyysjakauma, niin posteriorijakauma on, ja se voidaan ajatella, että se johdettiin konjugaattipriorilla Gamma$(0, 0)$.

Yllä tarkasteltu priori

$$p(\theta) \propto \frac{1}{\theta}\,\mathbb{1}_{(0,\infty)}$$

on esimerkki **epäoleellisesta priorista** (improper prior). Tällä tarkoitetaan funktiota, joka ei ole minkään todennäköisyysjakauman tiheysfunktio[3], mutta jota voidaan käyttää priorijakauman tapaan sijoittamalla se Bayesin kaavaan

$$f(\boldsymbol{\theta}, \mathbf{y}) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}$$

[3]Täsmällisemmin epäoleellinen priori on jokin mitta, joka ei ole todennäköisyysmitta, jolloin Bayesin kaavaan sijoitetaan tämän mitan tiheysfunktio (tämän kurssin tapauksissa Lebesguen mitan suhteen).

priorijakauman paikalle. Jos integraali

$$p(\mathbf{y}) = \int p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

on äärellinen, eli näin saatu funktio $f(\boldsymbol{\theta}, \mathbf{y})$ on todennäköisyysjakauman tiheysfunktio parametrin $\boldsymbol{\theta}$ suhteen, voidaan merkitä

$$p(\boldsymbol{\theta}|\mathbf{y}) := f(\boldsymbol{\theta}, \mathbf{y})$$

ja kutsua jakaumaa $p(\boldsymbol{\theta}|\mathbf{y})$ posteriorijakaumaksi, vaikka se ei olekaan tarkkaan ottaen ehdollinen todennäköisyysjakauma.

Käytännössä epäoleellista Jeffreys'n prioria käyttämällä saadut posteriorijakaumat kannattaa tulkita raja-arvoina, jotka saadaan, kun konjugaattipriorin hyperparametri viedään sen määrittelyjoukon rajalle. Esimerkiksi yllä laskettu posteriorijakauma Gamma$(n, \sum_{i=1}^{n} y_i)$ voidaan ajatella posteriorijakaumana konjugaattipriorille Gamma(α, β), kun

$$\alpha \to 0, \quad \beta \to 0.$$

Esimerkki 2.22. Jatkoa esimerkeille 2.19 ja 2.20. Binomijakaumalla Bin(n, θ) epäoleellinen priori on ns. Haldanen priori

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1},$$

eli Beta$(0, 0)$. Tällä priorilla posteriorijakauma

$$p(\theta|y) \stackrel{d}{=} \text{Beta}(y, n-y)$$

riippuu ainoastaan aineistosta, eli kyseessä on itse asiassa uskottavuusfunktio normalisoituna todennäköisyysjakaumaksi parametrin θ suhteen.

Tasajakaumapriori vastasi yhden, ja Jeffreys'n priori puolen, pseudohavainnon lisäämistä sekä onnistumisiin että epäonnistumisiin. Vastaavasti nähdään, että Haldanen priorilla posteriorijakauman odotusarvo on

$$E[\theta \,|\, y] = \frac{y}{n},$$

eli aineistoon ei lisätä yhtään pseudohavaintoa, ja siten posteriorijakauman odotusarvo on sama kuin parametrin θ suurimman uskottavuuden estimaatti $\hat{\theta}$.

Jos ei havaita yhtään onnistumista, eli $y = 0$, tai yhtään epäonnistumista, eli $y = n$, posteriorijakauma ei kuitenkaan ole oikea todennäköisyysjakauma.

Sen sijaan heikosti informatiivista prioria, esim Beta$(0.01, 0.01)$ käytettäessä posteriorijakauma Beta$(y + 0.01, n - y + 0.01)$ on aina oikea todennäköisyysjakauma.

2.3.3 Referenssipriorit

Jeffreys'n priorit toimivat hyvin yksiulotteiselle parametrille, mutta ovat ongelmallisempia moniulotteisessa tapauksessa. **Referenssipriorien** (reference prior) tarkoituksena on formalisoida epäinformatiivisen priorin käsite informaatioteoreettisten työkalujen avulla. Ajatuksena on valita priori, joka maksimoi priori- ja posteriorijakauman välisen etäisyyden, Kullback-Leibler-divergenssin, odotusarvon asymptoottisesti, eli kun otoskoko lähestyy ääretöntä. Tämä voidaan tulkita intuitiivisesti siten, että valitaan priori, jolle posteriorijakauman ja priorijakauman ero on mahdollisimman suuri, kun meillä on käytössämme ääretön aineisto, eli täydellinen informaatio; referenssipriori on siis sellainen priori, joka sisältää mahdollisimman pienen osan priorin ja aineiston yhteisestä kokonaisinformaatiosta koskien parametrin todellista arvoa.

Myös referenssipriorit ovat invariantteja mallin parametrisoinnin suhteen, ja osoittautuukin, että yksiulotteiselle parametrille referenssipriorien teoria antaa samat priorit kuin Jeffreys'n periaate.

2.3.4 Informatiiviset priorit

Jos taas haluamme käyttää käyttää hyväksi etukäteistietoamme mallinnettavasta ilmiöstä, eli haluamme, että priorijakaumallamme on selkeä vaikutus posteriorijakaumaan, puhutaan **informatiivisesta priorista** (informative prior). Kuten epäinformatiivinen priori, myöskään tämä ei ole formaalisti määritelty käsite.

Käytännön esimerkki informatiivisen priorin käytöstä voisi olla vedonlyönnin kerrointen laskenta, jossa voidaan yhdistää asiantuntijoiden arvioita priorijakauman muodossa havaitun aineiston pohjalta tehtyyn tilastolliseen malliin. Vedonlyöntiyhtiöllä ei ole varaa antaa pelkästään 'aineiston puhua puolestaan', sillä jos kertoimet eivät vastaa todellisia todennäköisyyksiä, se voi hävitä paljon rahaa.

Ongelmana informatiivisten priorien käytössä on se, että ihmiset, asiantuntijat mukaan lukien, ovat käytännössä erittäin huonoja todennäköisyyksien arvioimiseen; yleensä asiantuntijoiden näkemykset ennustustensa tarkkuudesta ovat huomattavan ylioptimistisia. Myöskään monimutkaisempien mallien parametreille ei välttämättä ole mitään intuitiivista tulkintaa, joten niitä koskevien käsitysten muotoileminen voi olla vähintäänkin vaikeaa.

Monesti epäinformatiivisen tai informatiivisen priorin valinta ei olekaan niinkään filosofinen kuin käytännön kysymys. Voidaan ajatella kaksi ääritapausta: ollaanko tekemässä tieteellistä koetta, joka halutaan pitää mahdollisimman puhtaana ennakko-oletusten vaikutuksista, vai lyömässä vetoa, jolloin halutaan hyödyntää kaikki kvalitatiivinen ja kvantitatiivinen informaatio? Yleensä tilanne ei tietenkään ole näin selkeä, mutta priorin valintaa voidaa tarkastella laajemmin esimerkiksi päätösteoreettisessa viitekehyksessä.

2.4 Usean parametrin jakaumat

Olemme oikeastaan jo käsitelleet posteriorijakauman laskemisen moniulotteisessa tapauksessa, sillä olemme koko ajan olettaneet, että parametri $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ on vektori, ja käsitelleet yksiulotteista parametria erityistapauksena tästä. Esimerkiksi laskuharjoituksissa käsitellyssä multinomijakaumamallissa $Y \sim \text{Multin}(n, \boldsymbol{\theta})$ monesti ajatellaan, että olemme kiinnostuneita kaikkien vaihtoehtojen todennäköisyyksistä, eli kaikista parametrivektorin $\boldsymbol{\theta}$ komponenteista. Vaikka parametri on vektoriarvoinen, voimme siis ajatella, että koko parametrin posteriorijakauma $p(\boldsymbol{\theta}|\mathbf{y})$ on haluttu lopputulos, eli prosessi on täysin analoginen yksiulotteisen parametrin tapauksen kanssa.

Monesti emme kuitenkaan ole kiinnostuneet koko parametrivektorin $\boldsymbol{\theta}$ posteriorijakaumasta

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}),$$

vaan ainoastaan osan parametreista reunaposteriorijakaumista. Klassinen esimerkki on tilanne, jota mallinnetaan mittausten arvoja jostain suureesta normaalijakaumalla, ja olemme kiinnostuneet suureen todellisesta arvosta, eli jakauman odotusarvosta, emme niinkään mittauvirheestä, eli varianssista, joka oletetaan myös tuntemattomaksi. Tässä tapauksessa mallin parametri on kaksiulotteinen $\boldsymbol{\theta} = (\mu, \sigma^2)$. Parametria, jonka arvosta emme sinällään ole kiinnostuneita, mutta jota kuitenkin tarvitaan mallintamiseen, eli varianssia tässä tilanteessa, kutsutaan ns. *haittaparametriksi* (nuisance parameter).

Tarkastellaan siis jatkossa tilannetta, jossa moniulotteinen parametri voidaan jakaa kahteen (mahdollisesti myös moniulotteiseen) osaan $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, jossa $\boldsymbol{\theta}_2$ edustaa parametreja, joiden arvoista olemme kiinnostuneita, ja $\boldsymbol{\theta}_2$ muita mallin parametreja, eli ns. haittaparametreja.

2.4.1 Reunaposteriorijakauman laskeminen

Parametrin $\boldsymbol{\theta}_1$ jakaumaa ehdolla aineisto kutsutaan sen **reunaposteriorijakaumaksi**, ja tässä tilanteessa päättelyn tavoitteena on selvittää nimenomaan tämä parametrin kiinnostavan osan reunaposteriorijakauma. Periaatteessa se onnistuu yksinkertaisesti integroimalla posteriorijakaumaa haittaparametrin yli:

$$p(\boldsymbol{\theta}_1|\mathbf{y}) = \int p(\boldsymbol{\theta}|\mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}_2.$$

Tämä integraali voidaan myös faktoroida muotoon

$$p(\boldsymbol{\theta}_1|\mathbf{y}) = \int p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})p(\boldsymbol{\theta}_2|\mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}_2.$$

Tästä muodosta nähdään, että reunaposteriorijakauma voidaan tulkita sekoitusmallina kiinnostavan parametrin ehdollisista jakaumasta ehdolla haittaparametri, missä sekoituspainot tulevat haittaparametrin posteriorijakaumasta $p(\boldsymbol{\theta}_2|\mathbf{y})$.

Käytännössä monimutkaisemmissa malleissa usein edes posteriorijakaumalle ei ole suljettua muotoa, joten myös reunaposteriorijakaumaa approksimoidaan numeerisesti tai simuloimalla. Käydään kuitenkin seuraavaksi läpi esimerkki normaalijakaumamallista, jolle nämä reunaposteriorijakaumat pystytään ratkaisemaan.

Esimerkki 2.23. Oletetaan n riippumatonta havaintoa samasta normaalijakaumasta:

$$Y_1, \ldots, Y_n \perp\!\!\!\perp | \, (\mu, \sigma^2), \quad Y_i \sim N(\mu, \sigma^2)$$

kaikille $i = 1, \ldots, n$.

Käytetään epäinformatiivista priorijakaumaa

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

jolloin normalisoimattomaksi posteriorijakaumaksi saadaan tilastollisen päätelyn kurssilta tutun hajotelman

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

avulla

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

missä otoskeskiarvo

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

ja otoskeskihajonta

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

muodostavat kaksiulotteisen tyhjentävän tunnusluvun mallin parametrille (μ, σ^2).

Oletetaan, että olemme kiinnostuneet odotusarvon reunaposteriorijakaumasta $p(\mu|\mathbf{y})$, joka saadaan integroimalla posteriorijakaumaa varianssin yli. Normaalijakauman tapauksessa tämä integraali voidaan ratkaista esimerkiksi täydentämällä se käännetyn gammajakauman (inverse gamma distribution) tiheysfunktion

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\beta}{x}}$$

integraaliksi yli sen alustan:

$$\begin{aligned}
p(\mu|\mathbf{y}) &= \int p(\mu, \sigma^2|\mathbf{y})\,\mathrm{d}\sigma^2 \\
&\propto \int_0^\infty \sigma^{-n-2} \exp\left\{-\frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{2\sigma^2}\right\}\,\mathrm{d}\sigma^2 \\
&= \int_0^\infty (\sigma^2)^{-\left(\frac{n}{2}+1\right)} \exp\left\{-\frac{c}{\sigma^2}\right\}\,\mathrm{d}\sigma^2 \\
&\propto \frac{1}{c^{\frac{n}{2}}} \int_0^\infty \frac{c^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} (\sigma^2)^{-\left(\frac{n}{2}+1\right)} \exp\left\{-\frac{c}{\sigma^2}\right\}\,\mathrm{d}\sigma^2 \\
&= \left((n-1)s^2 + n(\bar{y}-\mu)^2\right)^{-\frac{n}{2}} \\
&= \left(1 + \frac{1}{(n-1)}\left(\frac{\mu-\bar{y}}{s/\sqrt{n}}\right)^2\right)^{-\frac{(n-1)+1}{2}}.
\end{aligned}$$

Tämä voidaan tunnistaa (ei-standardin) t:n jakauman vapausasteella $n-1$ tiheysfunktion ytimeksi, eli

$$\mu \,|\, \mathbf{Y} \sim t_{n-1}(\bar{y}, s^2/n).$$

Siten havaitaan, että sopivasti skaalattu ja siirretty keskiarvoparametri noudattaa standardia t:n jakaumaa vapausasteella $n-1$:

$$\left.\frac{\mu-\bar{y}}{s/\sqrt{n}}\right| \mathbf{Y} \sim t_{n-1}$$

Sen jakauma ei siis riipu aineistosta, eikä parametrista, eli kysessä on saranasuure.

Luku 3

Hierarkkiset mallit

Monesti havaintoyksiköitä ei voida pitää riippumattomia, vaan ne ovat jakautuneet ryhmiin, eli niillä on jokin luonnollinen hierarkia. Esimerkiksi samaa lääkettä voidaan testata useammalla kokeella, joissa kussakin on useita koehenkilöitä, jolloin yhteen kokeeseen osallistuneet henkilöt (tai oikeastaan heidän tuloksensa) muodostavat yhden ryhmän. Tällaista usean samanlaisen kokeen tulosten yhdistämistä kutsutaan meta-analyysiksi. Vastaavasti pitkän matematiikan ylioppilaskokeen tuloksia voidaan mallintaa hierarkkisesti, jolloin yhden lukion tulokset muodostuvat yhden ryhmän. Ei ole järkevää olettaa, että oppilaiden tulokset koko maassa olisivat riippumattomia, mutta yhden lukion sisällä tuloksia voidaan mallintaa keskenään riippumattomina (ellei oppilaista ole muita tietoja, joita voidaan käyttää taustamuuttujina).

Hierarkiatasoja voi olla myös useampia: voidaan ajatella, että saman kunnan lukoiden tulokset riippuvat toisistaan, ja edelleen, että saman maakunnan kuntien tulokset riippuvat toisistaan. Hierarkkisten mallien käsittely Bayesiläisessä viitekehyksessa on periaatteessa yksinkertaista: malliin vain lisätään uusia hierarkiatasoja ja parametreja. Käytännössä priorien valinta ja laskenta monimutkaistuvat.

3.1 Hierarkkisen mallin rakenne

Yksinkertaisessa hierarkkisessa mallissa meillä on J ryhmää, joiden sisällä havainnot oletetaan keskenään riippumattomiksi:

$$Y_{1j}, \ldots, Y_{n_j j} \perp\!\!\!\perp | \, \theta_j \quad \text{kaikille } j = 1, \ldots J.$$

Merkitään ryhmän j havaintoja $Y_{\cdot j} = (Y_{1j}, \ldots, Y_{n_j j})$, jolloin niiden jakauma faktoroituu muotoon

$$p(y_{\cdot j}|\theta_j) = \prod_{i=1}^{n_j} p(y_{ij}|\theta_j).$$

Ylioppilaskokeen tapauksessa $Y_{\cdot j}$ on vektori, joka sisältää j:nnen koulun oppilaiden, joita on n_j kappaletta, tulokset.

Oletetaan myös, että eri ryhmien tulokset ovat keskenään riippumattomia ehdolla parametri:

$$Y_{\cdot 1}, \ldots, Y_{\cdot J} \perp\!\!\!\perp \mid \boldsymbol{\theta}, \tag{3.1}$$

jolloin aineiston jakauma faktoroituu muotoon

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{j=1}^{J} p(y_{\cdot j}|\theta_j) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij}|\theta_j).$$

Ryhmien parametreja

$$\theta_1, \ldots, \theta_J \perp\!\!\!\perp \mid \boldsymbol{\phi} \tag{3.2}$$

voidaan edelleen mallintaa riippumattomana otoksena samasta **populaatiojakaumasta**, eli

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

Hierarkkisessa mallissa kaikki satunnaisvaihtelu ei siis ole yksittäisten havaintoyksiköiden Y_{ij} välillä ryhmien sisällä, vaan oletetaan, että myös ryhmien parametrien θ_j arvoihin liittyy satunnaisvaihtelua siten, että ne noudattavat samaa jakaumaa.

Jos ylioppilaskokeen tapauksessa parametrit ovat lukiokohtaisia keskiarvoja, niin tällöin niihin liittyvää satunnaisvaihtelua mallinnetaan olettamalla ne satunnaisotoksena yhteisestä jakaumasta.

Sinänsä tässä mallissa ei vielä ole mitään Bayesiläistä; monitasoisisia malleja voidaan rakentaa myös klassisen teorian pohjalta. Jos ja kun haluamme kuitenkin rakentaa täysin Bayesiläisen mallin, oletamme myös populaatiojakauman parametrit satunnaismuuttujiksi, ja niille priorijakauman

$$p(\boldsymbol{\phi}).$$

3.1.1 Osittaiset konjugaattimallit

Yllä kuvatussa yksinkertaisessa hierarkkisessa mallissa päättelyn kohteena on parametrin $(\boldsymbol{\theta}, \boldsymbol{\phi})$ posteriorijakauma

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) \propto p(\boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\mathbf{y}|\boldsymbol{\theta}). \tag{3.3}$$

Yllä olevassa kaavassa aineiston jakauma yksinkertaistui muotoon

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\mathbf{y}|\boldsymbol{\theta}),$$

sillä oletetaan, että ylempien hierarkiatasojen parametrit vaikuttavat aineiston $\mathbf{Y}$ jakaumaan ainoastaan ryhmäkohtaisen parametrin $\boldsymbol{\theta}$ kautta.

Vähänkään monimutkaisemmissa malleissa tätä posteriorijakaumaa ei pystytä ratkaisemaan suljetussa muodossa, vaan sitä joudutaan approksimoimaan simuloimalla. Kuitenkin, jos populaatiojakauma $p(\boldsymbol{\theta}|\boldsymbol{\phi})$ on konjugaattijakauma aineiston jakaumalle $p(\mathbf{y}|\boldsymbol{\theta})$, osa ehdollisista posteriorijakaumista ja reunaposteriorijakaumista voidaan ratkaista analyyttisesti, mikä helpottaa simulointia:

- Koska ryhmien parametrit θ_j oletettiin riippumattomaksi satunnaisotokseksi populaatiojakaumasta, parametrin $\boldsymbol{\theta}$ ehdollinen posteriorijakauma faktoroituu J:n riippumattomaan komponenttiin:

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{y}) \propto p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{j=1}^{J} p(\theta_j|\boldsymbol{\phi})p(y_{\cdot j}|\theta_j), \tag{3.4}$$

  ja nämä komponentit osataan laskea konjugaattioletuksen nojalla.

- Periaatteessa populaatiojakauman parametrin $\boldsymbol{\phi}$ reunaposterijakauma $p(\boldsymbol{\phi}|\mathbf{y})$ voidaan laskea integroimalla posteriorijakaumaa parametrin $\boldsymbol{\theta}$ yli:

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

  Joillekin konjugaattimalleille, kuten normaalijakaumalle tunnetulla varianssilla ja normaalipriorilla odotusarvolle, tämä voidaan ratkaista helposti sjoittamalla yhteisposteriorijakauma 3.3 ja edellä laskettu ehdollinen posteriorijakauma 3.4 ehdollisen todennäköisyyden määritelmään:

$$p(\boldsymbol{\phi}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y})}{p(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{y})}. \tag{3.5}$$

Näiden jakaumien avulla on mahdollista simuloida otos posteriorijakaumasta:

- Simuloi ensin otos parametrin $\boldsymbol{\phi}^{\text{sim}}$ reunaposteriorijakaumasta $p(\boldsymbol{\phi}|\mathbf{y})$.
- Simuloi sitten otos $\boldsymbol{\theta}^{\text{sim}}$ ehdollisesta posteriorijakaumasta $p(\boldsymbol{\theta}|\boldsymbol{\phi}^{\text{sim}}, \mathbf{y})$ käyttäen edellä generoituja $\boldsymbol{\phi}^{\text{sim}}$:n arvoja.

Haluttaessa voidaan vielä simuloida havaintoja posterioriennustejakaumasta uusille havainnoille $\tilde{\mathbf{Y}}$. Voidaan generoida kahdenlaisia uusia havaintoja:

- Havaintoja $\tilde{y}_{1j}, \ldots, \tilde{y}_{m_j j}$ jo olemassa olevista ryhmistä $j = 1, \ldots J$. Tällöin generoidaan arvoja jakaumasta $p(\tilde{\mathbf{y}}|\boldsymbol{\theta}^{\text{sim}})$ edellä simuloiduille arvoille $\boldsymbol{\theta}^{\text{sim}}$.

Taulukko 3.1: Havaittujen harjoitusvaikutusten keskiarvot ja näiden keskivirheet kouluittain.

Koulu	Keskiarvo $\bar{y}_{.j}$	Keskivirhe σ_j
1	28	15
2	8	10
3	-3	16
4	7	11
5	-1	9
6	1	11
7	18	10
8	12	18

- Simuloidaan ensin uusia ryhmiä $\tilde{\theta}_{J+1} \ldots, \tilde{\theta}_K$ samasta superpopulaatiosta kuin alkuperäiset ryhmät, ja sen jälkeen uusia havaintoja $\tilde{y}_{1k}, \ldots, \tilde{y}_{m_k k}$ näistä ryhmistä kaikille $k = 1, \ldots, K$.

3.2 Esimerkki: 8 koulun vertailu

Tutustutaan vielä konkreettisesti yksinkertaiseen hierarkkiseen malliin esimerkin (poimittu BDA:n [3] luvusta 5.5) kautta.

Testataan harjoitusohjelmien vaikutusta standardoidun SAT-V (Scholastic aptitude test - verbal) koululaisten kielellistä lahjakkuutta mittaavan testin tuloksiin. Testin tarkoitus olisi olla siinä mielessä robusti, että sen tuloksiin ei voisi juurikaan vaikuttaa lyhytaikaisilla 'preppausohjelmilla'. Harjoitusohjelmien tehon tutkimiseksi tarkastellaan kahdeksaa koulua, joissa kussakin on toteutettu oma harjoitusohjelmansa testiä varten. Harjoitusohjelmien keskimääräiset vaikutukset keskivirheineen löytyvät taulukosta 3.1 (testin pisteet ovat yleensä välillä 200-800, ja keskiarvo n. 500 ja keskihajonta n. 100). Nämä on estimoitu vertaamalla oppilaiden koetuloksia heidän ennen harjoitusohjelmaa tekemänsä preliminäärikokeen tuloksiin; huomioon on myös otettu kovariaatteja.

Voidaan kuitenkin ajatella, että oppilaiden havaitut harjoitusvaikukset jokaiselle koululle j noudattavat normaalijakaumaa

$$Y_{ij} \sim N(\theta_j, \sigma^2) \quad \text{kaikille } i = 1, \ldots, n_j,$$

missä θ_j on tämän koulun harjoitusvaikuksen keskiarvo, ja σ^2 kaikille kouluille yhteiseksi ja tunnetuksi oletettu keskihajonta. Kaikissa kouluissa otoskoko suurehko (yli 30), joten

normaalijakaumaoletusta ja oletusta tunnetusta ja yhteisestä varianssista voidaan pitää hyväksyttävinä approksimaatioina. Siten voidaan tarkastella pelkästään kouluittaisten otoskeskiarvojen $\bar{Y}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$ jakaumia

$$\bar{Y}_{.j} \,|\, \theta_j \sim N(\theta_j, \frac{\sigma^2}{n_j}) \quad \text{kaikille } j = 1, \ldots J.$$

Merkitään vielä tunnettuja koulujen harjoitusvaikutusten θ_j keskivirheitä

$$\sigma_j = \frac{\sigma}{n_j}.$$

Siten voimme tarkastella havaintoina riippumattomia (ei samaa) normaalijakaumaa noudattavia otoskeskiarvoja:

$$\bar{Y}_{.j} \,|\, \theta_j \sim N(\theta_j, \sigma_j), \quad j = 1, \ldots J.$$

Malli

Oletetaan lisäksi, että koulukohtaisten harjoitusvaikutusten keskiarvot $\theta_1, \ldots, \theta_J$ ovat riippumaton otos samasta normaalijakaumasta. Tämän populaatiojakauman keskiarvolle μ ja keskihajonnalle τ oletetaan priorijakauma, joka määritellään myöhemmin. Siten tämä hierarkkinen malli voidaan kirjoittaa (priorijakauman määrittelyä vaille) seuraavasti kaikille $j = 1, \ldots, J$:

$$\begin{aligned} \bar{Y}_{.j} \,|\, \theta_j &\sim N(\theta_j, \sigma_j), \\ \theta_j \,|\, \mu, \tau &\sim N(\mu, \tau^2), \\ p(\mu,&\tau). \end{aligned}$$

Tällaisessä hierarkkisen mallin merkinnässä oletetaan implisiittisesti (ellei toisin mainita) havaintojen riippumattomuus ehdolla parametrit (3.1):

$$\bar{Y}_{.1} \ldots, \bar{Y}_{.J} \perp\!\!\!\perp | \, \theta_1, \ldots, \theta_J,$$

ja parametrien riippumattomuus ehdolla posteriorijakauman parametrit (3.2):

$$\theta_1, \ldots, \theta_J \perp\!\!\!\perp | \, \mu, \tau.$$

Samoin oletetaan, että havainnot riippuvat populaatiojakauman parametreista vain ryhmäkohtaisten parametrien kautta, eli että

$$\bar{Y}_{.1} \ldots, \bar{Y}_{.J} \perp\!\!\!\perp (\mu, \tau) | \, \theta_1, \ldots, \theta_J.$$

Priorijakauma

Hajotetaan priorijakauma muotoon

$$p(\mu, \tau) = p(\tau)p(\mu|\tau).$$

Odotusarvo μ on sijaintiparametri, joten sen ehdolliseksi (epäoleelliseksi) priorijakaumaksi voidaan valita tasajakauma koko reaaliakselilla:

$$p(\mu|\tau) \propto 1.$$

Siten

$$p(\mu, \tau) \propto p(\tau),$$

eli meidän tulee vielä määrittää priorijakauma keskihajonnalle τ. Johdetaan ensin malli

$$\begin{aligned} \bar{Y}_{.j} &\sim N(\theta_j, \sigma_j), \\ \theta_j &\sim N(\mu, \tau^2), \\ p(\mu, \tau) &\propto p(\tau) \end{aligned}$$

ehdolla τ:n priori, ja tarkastellaan sen jälkeen erilaisia mahdollisuuksia valita se. Huomaa, että ehdollistukset parametreilla on jätetty pois mallin merkinnöistä: kirjoitetaan esimerkiksi $\theta_j \sim N(\mu, \tau^2)$ sen sijaan, että merkittäisiin eksplisiittisesti $\theta_j \mid \mu, \tau \sim N(\mu, \tau^2)$. Näin tehdään monesti, jolloin oletetaan implisiittisesti, että ehdollistetaan aina alemman tason parametreilla.

Posteriorijakaumat

Mallin normalisoimaton yhteisposteriorijakauma voidaan kirjoittaa (kaava 3.3):

$$p(\boldsymbol{\theta}, \mu, \tau|\mathbf{y}) \propto p(\tau) \prod_{j=1}^{J} p(\theta_j|\mu, \tau)p(\bar{y}_{.j}|\theta_j).$$

Siten parametrin $\boldsymbol{\theta}$ ehdolliseksi posteriorijakaumaksi ehdolla μ ja τ saadaan [1] (kaava 3.4):

$$p(\boldsymbol{\theta}|\mu, \tau) \propto \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2)N(\bar{y}_{.j}|\theta_j, \sigma_j^2).$$

[1] Tässä käytetään kätevää lyhennysmerkintää $p(x|\mu, \tau) = N(x|\mu, \tau^2)$, joka tarkoittaa, että x:n (tai oikeastaan satunnaismuuttujan X, jonka realisaatio x on) tiheysfunktio on normaalijakauman odotusarvolla μ ja varianssilla τ^2 tiheysfunktio

Tämän takia populaatiojakauma kannatti valita konjugaattijakaumaksi aineiston jakaumalle: jokainen parametrin $\boldsymbol{\theta}$ ehdollisen posteriorijakauman J:stä komponentista on normaalijakaumamalli tunnetulla varianssilla, ja tämä malli ratkaistiin esimerkissä 2.3! Parametrien θ_j ehdolliset posteriorijakaumat voidaan siten ratkaista erikseen kaikille $j = 1, \ldots, J$:

$$\theta_j \,|\, \mu, \tau, \mathbf{y} \sim N(\hat{\theta}_j, V_j),$$

missä

$$\hat{\theta}_j = \frac{\frac{1}{\tau^2}\mu + \frac{1}{\sigma_j^2}\bar{y}_{.j}}{\frac{1}{\tau^2} + \frac{1}{\sigma_j^2}},$$

ja

$$\frac{1}{V_j} = \frac{1}{\tau^2} + \frac{1}{\sigma_j^2}.$$

Mallin osittaista konjugaattiutta voidaan hyödyntää myös populaatiojakauman parametrien (μ, τ) reunaposteriorijakauman laskemiseen: aineiston reunajakauma voidaan esittää yksittäisten koulujen reunajakaumien tulona:

$$p(\mathbf{y}|\mu, \tau) = \prod_{j=1}^{J} \int N(\bar{y}_{.j}|\theta_j, \sigma_j) N(\theta_j|\mu, \tau)\, \mathrm{d}\theta_j,$$

jonka komponentit ovat normaalijakauman tunnetulla varianssilla prioriennustejakaumia. Nämä saadaan jälleen esimerkistä 2.3 (jossa tosin laskettiin posterioriennustejakauma, mutta prioriennustejakauma saadaan samalla laskulla):

$$\bar{Y}_{.j} \,|\, \mu, \tau \sim N(\mu, \tau^2 + \sigma_j^2).$$

Siten parametrien μ ja τ reunaposteriorijakaumaksi saadaan

$$p(\mu, \tau|\mathbf{y}) \propto p(\mu, \tau) p(\mathbf{y}|\mu, \tau) = p(\tau) \prod_{j=1}^{J} N(\bar{y}_{.j}|\mu, \tau^2 + \sigma_j^2). \tag{3.6}$$

Olemme siis ratkaisseet kaavan 3.5 reunaposteriorijakauman. Periaattessa voisimme jo aloittaa simuloinnin simuloimalla parametrien μ ja τ arvoja tästä kaksiulotteisesta jakaumasta. Voimme kuitenkin ratkaista mallia vielä tästäkin eteenpäin, jolloin meidän tarvitsee enää simuloida yksiulotteisesta parametrin τ reunajakaumasta $p(\tau|\mathbf{y})$.

Parametrin μ posteriorijakauma ehdolla τ on siis:

$$p(\mu|\tau, \mathbf{y}) \propto p(\mu, \tau|\mathbf{y}) \propto \prod_{j=1}^{J} N(\bar{y}_{.j}|\mu, \tau^2 + \sigma_j^2) \tag{3.7}$$

Mutta jälleen havaitaan, että tämä on yleistys esimerkin 2.3 posteriorijakauman laskusta, jossa kerrotaan kaksi normaalijakauman ydintä, ja tuloksena saadaan normaalijakauma, jonka odotusarvo on alkuperäisten varianssien käänteisluvuilla, eli tarkkuusparametreilla, painotettu keskiarvo, ja tarkkuusparametri alkuperäisten tarkkuusparametrien summa. Kaavan 3.7 laskussa kerrottavia normaalijakauman ytimiä on J kappaletta, joten

$$\mu|\tau, \mathbf{y} \sim N(\hat{\mu}, V_\mu), \tag{3.8}$$

missä

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2} \bar{y}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}},$$

ja

$$\frac{1}{V_\mu} = \sum_{j=1}^{J} \frac{1}{\tau^2+\sigma_j^2}.$$

Parametrin τ reunaposteriorijakauma saadaan nyt sijoittamalla 3.6 ja 3.8 kaavaan 3.5 (sijoitetaan $\boldsymbol{\phi} = \tau$ ja $\boldsymbol{\theta} = \mu$; samat kaavat toimivat siis myös useampitasoisissa hierarkkisissa malleissa):

$$p(\tau|\mathbf{y}) = \frac{p(\tau, \mu|\mathbf{y})}{p(\mu|\tau, \mathbf{y})} = p(\tau) \frac{\prod_{j=1}^{J} N(\bar{y}_{.j}|\mu, \tau^2+\sigma_j^2)}{N(\mu|\hat{\mu}, V_\mu)}.$$

Sijoitetaan vielä $\mu = \hat{\mu}$, jolloin

$$p(\mu = \hat{\mu}|\tau, y) \propto V_\mu^{-1/2},$$

ja siten parametrin τ normalisoimattomaksi reunaposteriorijakaumaksi saadaan

$$\begin{aligned} p(\tau|\mathbf{y}) &\propto p(\tau) \frac{\prod_{j=1}^{J} N(\bar{y}_{.j}|\hat{\mu}, \tau^2+\sigma_j^2)}{N(\hat{\mu}|\hat{\mu}, V_\mu)} \\ &\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\tau^2+\sigma_j^2)^{-1/2} \exp\left\{ -\frac{(\bar{y}_{.j}-\hat{\mu})^2}{2(\tau^2+\sigma_j^2)} \right\}. \end{aligned} \tag{3.9}$$

Tämä ei ole minkään tunnetun jakauman tiheysfunktio, mutta se on yksiulotteisen jakauman normalisoimaton tiheysfunktio, joten sen simuloiminen onnistuu helposti. Ensin meidän täytyy kuitenkin vielä valita priorijakauma $p(\tau)$.

Priorijakauman valinta

Huomaa, että tähän mennessä on ratkaistu yhteisposteriorijakauman

$$p(\boldsymbol{\theta}, \mu, \tau|\mathbf{y}) = p(\tau|\mathbf{y})p(\mu|\tau, \mathbf{y})p(\boldsymbol{\theta}|\mu, \tau, \mathbf{y})$$

komponenteista $p(\boldsymbol{\theta}|\mu, \tau, \mathbf{y})$ ja $p(\mu|\tau, \mathbf{y})$, jotka siis ovat kunnollisia todennäköisyysjakaumia. Siten yhteisposteriorijakauman olemassaolon varmistamiseksi riittää valita priorijakauma parametrille τ siten, että sen reunaposteriorijakauma $p(\tau|\mathbf{y})$ on olemassa, eli että kaavasta 3.9 saatavan normalisoimattoman reunaposteriorijakauman integraali yli positiivisen reaaliakselin on äärellinen:

$$\int_0^\infty p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\tau^2 + \sigma_j^2)^{-1/2} \exp\left\{ -\frac{(\bar{y}_{.j} - \hat{\mu})^2}{2(\tau^2 + \sigma_j^2)} \right\} \mathrm{d}\tau < \infty. \tag{3.10}$$

Ja vastaavasti, jos tämä integraali ei ole äärellinen, myöskään yhteisposteriorijakaumaa ei ole olemassa.

Käytetään τ:n normalisoimattomasta reunaposteriorijakaumasta, eli lausekkeen 3.10 integrandista, merkintää $f(\tau, \mathbf{y})$. Integraalin suppenemisen selvittämiseksi jaetaan se kahteen osaan:

$$\int_0^\infty f(\tau, \mathbf{y})\, \mathrm{d}\tau = \int_0^1 f(\tau, \mathbf{y})\, \mathrm{d}\tau + \int_1^\infty f(\tau, \mathbf{y})\, \mathrm{d}\tau,$$

ja tarkastellaan sen käyttäytymistä erikseen kummallakin rajalla. Kun $\tau \to 0$, integrandi lähestyy $C(\mathbf{y})p(\tau)$:ta, missä $C(\mathbf{y})$ on parametrista τ riippumaton vakio, eli

$$f(\tau, \mathbf{y}) \sim p(\tau), \quad \text{kun } \tau \to 0.$$

Koska $\int_0^1 \frac{1}{x}\, \mathrm{d}x$ hajaantuu, normaalijakauman epäoleelliselle referenssipriorille $p(\tau) \propto 1/\tau$ posteriorijakauman integraali hajaantuu, ja siten tämä priori ei johda kunnolliseen posteriorijakaumaan.

Sen sijaan jos parametrin τ priorijakaumana käytetään tasajakaumaa positiivisella reaaliakselilla

$$p(\tau) \propto 1,$$

integraali suppenee (koska $\int_0^1 \mathrm{d}x$ suppenee) alarajalla. Tarkastetaan vielä suppeneminen toisella rajalla, eli että tasapriorille myös

$$\int_1^\infty f(\tau, \mathbf{y})\, \mathrm{d}\tau < \infty.$$

Eksponenttitermi on aina pienempää kuin yksi, joten:

$$\begin{aligned} f(\tau, \mathbf{y}) &< p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\tau^2 + \sigma_j^2)^{-1/2} \\ &= p(\tau) \left(\sum_{j=1}^{J} \frac{1}{\tau^2 + \sigma_j^2} \right)^{-1/2} \prod_{j=1}^{J} (\tau^2 + \sigma_j^2)^{-1/2} \\ &= p(\tau) \left(\sum_{j=1}^{J} \frac{\prod_{k \neq j} (\tau^2 + \sigma_k^2)}{\prod_{j=1}^{J} (\tau^2 + \sigma_j^2)} \right)^{-1/2} \prod_{j=1}^{J} (\tau^2 + \sigma_j^2)^{-1/2} \\ &= p(\tau) \left(\sum_{j=1}^{J} \prod_{k \neq j} (\tau^2 + \sigma_k^2) \right)^{-1/2}. \end{aligned}$$

Tarkasteltavassa integraalissa $\tau > 1$, joten edelleen

$$f(\tau, \mathbf{y}) < p(\tau)(J\tau^{2(J-1)})^{-1/2} = p(\tau) J^{-1/2} \tau^{1-J}.$$

Koska

$$\int_1^\infty \frac{1}{x^p} \, \mathrm{d}x$$

suppenee, kun $p > 1$, tasapriorilla $p(\tau) \propto 1$ integraali suppenee tälläkin rajalla, ja täten posteriorijakauma on olemassa, kun $J > 2$.

Epäoleellisia prioreja käytettäessä on siis tärkeää aina tarkastaa posteriorijakauman olemassaolo, joko analyyttisesti, kuten tässä tehtiin, tai tarkastelemalla simuloidun posteriorijakauman käyttäymistä.

Simulointi

Nyt posteriorijakaumasta $p(\boldsymbol{\theta}, \mu, \tau | \mathbf{y})$ simuloiminen on yksinkertaista:

1. Generoi τ^{sim} yksiulotteisesta jakaumasta $p(\tau | \mathbf{y})$.
2. Generoi μ^{sim} normaalijakaumasta $p(\mu | \tau^{\text{sim}}, \mathbf{y}) = N(\mu | \hat{\mu}, V_\mu)$.
3. Kaikille $j = 1, \ldots J$: generoi θ_j^{sim} normaalijakaumasta $p(\theta_j | \mu^{\text{sim}}, \tau^{\text{sim}}, \mathbf{y}) = N(\theta_j | \hat{\theta}_j, V_j)$.

Parametrin τ posteriorijakaumasta voi simuloida esimerkiksi luomalla tiheän gridin välille, jossa oletetaan suurimman osan jakauman todennäköisyysmassasta olevan, esimerkiksi välille $(0, 50)$ (yksiulotteisessa tapauksessa tämä väli on helppo löytää kokeilemalla ja vertaamalla tuloksia silmämääräisesti), sen jälkeen laskemalla normalisoimattoman

tiheysfunktion arvot gridin pisteissä, ja normalisoimalla tulos siten, että lasketut arvot summautuvat ykköseksi. Lopuksi voit vielä generoida arvoja gridin välin pituisesta nollakeskeisestä tasajakaumasta, ja lisätä ne generoimiisi arvoihin, jotta saat ne pois gridin pisteistä. Seuraavassa vielä resepti yhden pisteen τ^{sim} (samalla tavalla voidaan tietenkin simuloida myös n_{sim} kappaleen otos) generoimiseen τ:n posteriorijakaumasta väliltä $(0, 50)$ ja gridivälillä 0.1:

1. Laske normalisoimattoman tiheysfunktion $f(\tau, \mathbf{y})$ arvot pisteissä

$$\tau_1 = 0.01, \tau_2 = 0.02, \ldots, \tau_{n-1} = 49.99, \tau_n = 50.$$

2. Normalisoi tulos:

$$p(\tau_j|\mathbf{y}) = \frac{f(\tau_j, \mathbf{y})}{\sum_{i=1}^{n} f(\tau_i, \mathbf{y})} \quad \text{kaikille } j = 1, \ldots, n.$$

3. Valitse satunnaisesti τ^{sim} gridin pisteistä $\tau_1 \ldots, \tau_n$ siten, että niiden valintatodennäköisyydet ovat $p(\tau_1|\mathbf{y}), \ldots, p(\tau_n|\mathbf{y})$.

4. Generoi $X \sim U(-0.005, 0.005)$, ja lisää se generoimaasi arvoon:

$$\tau^{\text{sim}} = \tau^{\text{sim}} + X.$$

Parametrien μ ja θ ehdollisista posteriorijakaumista simuloiminen onnistuu tämän jälkeen helposti normaalijakauman satunnaisgeneraattorilla, esimerkiksi `R`:n `rnorm`-funktiolla.

Luku 4

Päättely ja mallinvalinta

4.1 Mallinvalinta

Tilastolliset mallit sovitetaan havaittuun aineistoon $\mathbf{y} = (y_1, \ldots, y_n)$, mutta todellisuudessa olemme kiinnostuneita mallin sopivuudesta tarkasteltavan ilmiön kuvaamiseen, eli siitä, miten hyvin malli ennustaa uusia havaintoja samasta ilmiöstä.

Käytännössä meillä ei ole käytettävissämme ääretöntä otosta uusista havainnoista, emmekä tiedä todellista jakaumaa, josta aineisto ja uudet havainnot ovat peräisin (sitähän juuri pyrimme selvittämään!), joten emme voi tietää tätä tarkasti. Jos meillä kuitenkin on käytössämme **testiaineisto** $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$, voimme approksimoida mallin todellista ennustustarkkuutta sen avulla.

Jos esimerkiksi tavoitteenamme on sään ennustaminen, voimme sovittaa kymmenen edellisen vuoden säätietojen pohjalta erilaisia tilastollisia malleja, joiden tarkoituksena on ennustaa, sataako seuraavana päivänä. Voimme testata näiden mallien sopivuutta sään ennustamiseen tarkkailemalla vuoden ajan, miten hyvin malli ennustaa seuraavan päivän sään. Tässä siis testiaineisto $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$ on seuraavan vuoden säätilat.

Usein kuitenkaan käytössä ei ole erillistä testiaineistoa, joten mallien sopivuutta täytyy mitata samalla aineistolla, jota käytetään niiden estimointiin. Tähän käytetään yleensä joko ristiinvalidointia tai informaatiokriteerejä. Ennen näihin tutustumista tarkastellaan kuitenkin lyhyesti tärkeää kysymystä siitä, kuinka mallien ennustustarkkuutta tulisi mitata.

4.1.1 Kustannusfunktio

Käytännössä tilastollisen mallintamisen tarkoituksena on usein toimia pohjana päätöksenteolle, joten on tärkeää selvittää, mikä on mallin käyttötarkoitus, sillä se määrää vääriin ennustuksiin liittyvät kustannukset ja vastaavasti oikeisiin ennusteisiin liittyvät hyödyt.

Jos sääennustetta on esimerkiksi tarkoitus käyttää maanviljelyssä, voi olla että on haitallisempaa ennustaa virheellisesti poutaa sadepäivälle kuin päinvastoin, sillä sade pilaa heinänteon. Tämä voidaan formalisoida *kustannusfunktion* (loss function) $L(\mathbf{y}, d(\mathbf{y}))$ avulla (tai vastaavasti sen vastaluvun, eli *hyötyfunktion* [utility function] avulla), joka liittää jokaiseen mahdolliseen aineistoon $\mathbf{y}$ ja mallin perusteella sen pohjalta tehtyyn päätökseen $d(\mathbf{y})$ liittyvän kustannuksen (tai hyödyn): esimerkiksi virheellisen poudan ennustamiseen voi liittyä keskimäärin 1000 euron tappio, kun taas virheelliseen sateen ennustamisen vain 100 euron tappio.

Tällaista laajempaa viitekehystä, jossa tilastollista mallintamista voidaan tarkastella, kutsutaan *päätösteoriaksi* (decision theory). Usein kuitenkaan mallin käyttötarkoitus ei ole mallintamisvaiheessa selvä, vaan tavoitteena on yksinkertaisesti ennustaa tutkittavaa ilmiötä mahdollisimman tarkasti. Tai vaikka käyttötarkoitus tiedettäisiinkin, usein on hankalaa kvantifioida oikein ja väärin menneiden ennusteiden hyötyjä ja kustannuksia.

Jatkossa tarkastellaankin yleisiä mittareita mallin ennustustarkkuudelle, eli käytetään kustannusfunktiota, joissa päätös $d(\mathbf{y})$ on joko mallin antama piste-estimaatti (piste-ennustaminen) tai todennäköisyysjakauma (probabilistinen ennustaminen).

Oletetaan yksinkertaisuuden vuoksi jatkossa, että havaittu aineisto $\mathbf{y} = (y_1, \ldots, y_n)$ ja uudet havainnot $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$ ovat riippumattomien (ehdolla parametri $\boldsymbol{\theta}$) satunnaismuuttujien

$$Y_1, \ldots, Y_n,\ \tilde{Y}_1, \ldots, \tilde{Y}_m \perp\!\!\!\perp \boldsymbol{\theta} \tag{4.1}$$

realisaatioita. Tällöin aineiston (ja vastaavasti uusien havaintojen) jakauma faktoroituu muotoon

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}).$$

4.1.2 Piste-ennustaminen

Piste-ennustamisessa kustannusfunktio $L(\tilde{\mathbf{y}}, \hat{\mathbf{y}})$ vertaa mallin tuottamia piste-estimaatteja $\hat{y}_1, \ldots, \hat{y}_m$ testiaineiston todellisiin arvoihin $\tilde{y}_1, \ldots, \tilde{y}_m$. Yksinkertainen ja teoreettisesti perusteltu tapa on tutkia **keskineliövirhettä** (mean squared error)

$$L_{\text{MSE}}(\tilde{\mathbf{y}}, \hat{\mathbf{y}}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - \tilde{y}_i)^2. \tag{4.2}$$

Sateenennustamisesimerkissämme (jos $y_i = 1$ tarkoittaa sadetta, ja $y_i = 0$ poutaa, i:nnelle päivälle) keskineliövirhe on yksinkertaisesti mallin oikein ennustamien päivien osuus testivuoden kaikista päivistä. Jos taas ennustettava muuttuja on jatkuva, esimerkiksi sademäärä, keskineliövirhe on keskimääräinen ennustetun ($\hat{y}_i$) ja toteutuneen ($\tilde{y}_i$) sademäärän erotuksen neliö.

Mitä pienempi keskineliövirhe, sitä paremmin mallin voidaan ajatella mallintavan tutkittavaa ilmiötä. Teoriassa olemme siis kiinnostuneet keskineliövirheen odotusarvosta (ehdolla aineiston todellinen jakauma), mutta koska emme tiedä aineiston todellista jakaumaa (jos tietäisimme, niin mallinnusongelmamme olisi ratkaistu!), voimme vain approksimoida keskineliövirheen odotusarvoa laskemalla kaavasta 4.2 neliövirheen keskiarvon testiaineistolle.

Käytännössä täytyy siis muistaa, että testiaineisto on äärellisen kokoinen, joten toteunut keskineliövirhe, kuten kaikki muutkin ennustuskyvyn mittarit, riippuu aina juuri tästä nimenomaisesta testiaineistosta: jos esimerkiksi testivuosi sattuu olemaan erityisen sateinen, niin tämä testiaineisto suosii malleja, jotka ennustavat sadetta useammin.

4.1.3 Probabilistinen ennustaminen

Piste-estimaattien käyttäminen ennustustarkkuuden mittaamiseen ei kuitenkaan ota huomioon niihin liittyvää epävarmuutta: jos ensimmäinen malli ennustaa, että huomenna sataa 10% todennäköisyydellä, ja toisen mallin mukaan huomenna sataa 40% todennäköisyyllä, on loogista pitää toisen mallin ennustetta parempana jos seuraavana päivänä sataa, ja vastaavasti ensimmäisen mallin ennustetta parempana, jos seuraavana päivänä on poutaa. Siten olisi järkevää ajatella, että jos vertailtavat mallit antavat piste-estimaattien lisäksi myös todennäköisyysjakauman uusille havainnoille, myös mittarimme mallien ennusteille ottavat huomioon tämän koko jakauman.

Käytämmekin jatkossa mallien ennustustarkkuuden mittaamiseen **logaritmista uskottavuusfunktiota**[1] (log likelihood), eli todennäköisyyden tai tiheysfunktion arvon $\hat{p}(\tilde{\mathbf{y}})$, jonka malli antaa testiaineistolle $\hat{\mathbf{y}}$, logaritmia $\log \hat{p}(\tilde{\mathbf{y}})$. Riippumattomuusoletuksen 4.1 nojalla tämä voidaan esittää summana testiaineiston pisteiden todennäköisyyksien (tai vastaavasti tf:n arvojen) logaritmeista:

$$\log \hat{p}(\tilde{\mathbf{y}}) = \log \prod_{i=1}^{m} \hat{p}(\tilde{y}_i) = \sum_{i=1}^{m} \log \hat{p}(\tilde{y}_i).$$

Sateenennustamisesimerkissämme tämä siis on summa mallin ennustamien sateen todennäköisyyksien $\hat{p}(\tilde{y}_i = 1)$ logaritmeista sadepäiville, ja mallin ennustamien poudan todennäköisyyksien $\hat{p}(\tilde{y}_i = 0)$ logaritmeista poutapäiville. Mitä suuremman todennäköisyyden malli antaa toteutuneelle testiaineistolle, sitä suurempi log-uskottavuusfunktion arvo.

Olemme tähän mennessä merkinneet mallin ennustetta testiaineiston $\tilde{\mathbf{y}}$ todennäköisyydelle pelkästään $\hat{p}(\tilde{\mathbf{y}})$. Tämä johtuu siitä, että periaatteessa voimme vertailla mitä

[1] Jos logaritmisen uskottavuusfunktion haluaa ajatella kustannusfunktiona, sen merkin voi muuttaa negatiiviseksi. Osoittautuu, että normaalijakautuneille virheille, esimerkiksi lineaarissa regressiossa, logaritmisen uskottavuusfunktion vastaluku on suoraan verrannollinen keskineliövirheeseen (mieti miksi!).

tahansa ennusteita keskenään. Vaikka sääennusteet perustuvat yleensä (tietoon ilmäkehässä vallitsevista fysiikan laeista perustuviin) tilastollisiin malleihin, lopulliset ennusteet voivat olla yhdistelmä simulaatiomallin tuloksista ja asiantuntijan arviosta. Vastaavasti voimme myös kysyä ”sammakkomieheltä” tai vastaavalta paikalliselta ennustajalta arviot sateen todennäköisyyksistä $\hat{p}(\tilde{y}_1 = 1), \ldots, \hat{p}(\tilde{y}_{365} = 1)$ seuraavan vuoden päiville, ja verrata näitä muiden mallien ennusteisiin.

Rajoitutaan jatkossa kuitenkin havaitun aineiston $\mathbf{y} = (y_1, \ldots, y_n)$ pohjalta estimoitujen tilastollisten mallien vertailuun, ja tarkastellaan kolmea eri tapaa määritellä ennusteet uusien havaintojen $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$ todennäköisyyksille:

1. Sijoitetaan suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\theta}}_{\text{MLE}}(\mathbf{y})$ uusien havaintojen jakauman $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ parametriksi, jolloin tarkasteltava logaritminen uskottavuusfunktio on muotoa
$$\log \hat{p}(\tilde{\mathbf{y}}) = \log p(\tilde{\mathbf{y}}|\hat{\boldsymbol{\theta}}_{\text{MLE}}).$$

2. Sijoitetaan bayesiläinen piste-estimaatti, eli posteriorijakauman $p(\boldsymbol{\theta}|\mathbf{y})$ odotusarvo $\hat{\boldsymbol{\theta}}_{\text{Bayes}}(\mathbf{y}) = E[\boldsymbol{\theta}|\mathbf{y}]$, uusien havaintojen jakauman parametriksi, jolloin tarkasteltava logaritminen uskottavuusfunktio on muotoa
$$\log \hat{p}(\tilde{\mathbf{y}}) = \log p(\tilde{\mathbf{y}}|\hat{\boldsymbol{\theta}}_{\text{Bayes}}).$$

3. Otetaan odotusarvo uusien havaintojen jakaumasta parametrin $\boldsymbol{\theta}$ posteriorijakauman yli, eli käytetään uusien havaintojen posterioriennustejakaumaa:
$$\log \hat{p}(\tilde{\mathbf{y}}) = \log p(\tilde{\mathbf{y}}|\mathbf{y}) = \log \int p(\tilde{\mathbf{y}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}.$$

Näistä kolmesta tavasta ennustaa uusia havaintoja ensimmäinen on täysin frekventistinen: suurimman uskottavuuden estimaattia ei yleensä käytetä Bayes-päättelyssä[2]. Vaikka toisessa tavassa käytetään posteriorijakauman odotusarvoa, joka on bayesiläinen piste-estimaatti, sekin on edelleen piste-estimaatti. Tämäkään tapa ei siis ota huomioon posteriorijakauman kuvaamaa epävarmuutta parametrin $\boldsymbol{\theta}$ todellisesta arvosta, vaan tiivistää koko posteriorijakauman sen odotusarvoon.

[2]Bayes-päättelyn vastine suurimman uskottavuuden estimaatille, eli uskottavuusfunktion maksimikohdalle, on posteriorijakauman $p(\boldsymbol{\theta}|\mathbf{y})$ maksimikohta, eli ns. maximum a posteriori (MAP) - estimaatti
$$\hat{\boldsymbol{\theta}}_{\text{MAP}}(\mathbf{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{\theta}|\mathbf{y}).$$
Yleensä Bayes-päättelyssä posteriorijakauman piste-estimaattina käytetään posteriorijakauman odotusarvoa sen maksimikohdan sijaan, sillä ajatellaan, että odotusarvo edustaa todennäköisyysmassan painopisteenä paremmin koko jakaumaa.

Sen sijaan kolmas tapa, jossa tarkastellaan posterioriennustejakaumaa, eli odotusarvoa uusien havaintojen jakaumasta parametrin $\boldsymbol{\theta}$ posteriorijakauman yli (kts. kaavan 1.9 muoto posterioriennustejakaumasta ehdollisena odotusarvona), on täysin bayesiläinen, sillä se ottaa huomioon myös todelliseen parametrin arvoon liittyvän epävarmuuden ennusteissa uusien havaintojen todennäköisyyksille.

Logaritmisen posterioriennustejakauman uudelle havainnolle $\tilde{y}_i$ voi kätevästi approksimoida posteriorijakaumasta simuloitujen parametrin arvojen $\boldsymbol{\theta}_1^{\text{sim}}, \ldots, \boldsymbol{\theta}_S^{\text{sim}}$ avulla:

$$\log p(\tilde{y}_i|\mathbf{y}) \approx \log\left(\frac{1}{S}\sum_{s=1}^{S} p(\tilde{y}_i|\boldsymbol{\theta}_s^{\text{sim}})\right), \tag{4.3}$$

ja riippumattomuusoletuksen 4.1 nojalla koko testiaineiston logaritmista posterioriennustejakaumaa voidaan approksimoida yksittäisten havaintojen arvojen summana

$$\log p(\tilde{\mathbf{y}}|\mathbf{y}) \approx \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(\tilde{y}_i|\boldsymbol{\theta}_s^{\text{sim}})\right).$$

Yleensä kuitenkin haluamme soveltaa tilastollista malliamme heti, eikä meillä ole aikaa kerätä uusia havaintoja testiaineistoksi, jolla voisimme testata malliemme sopivuutta tarkasteltavan ilmiön kuvaamiseen. Sateenennustamisesimerkissämme tuskin haluaisimme odottaa vuotta, jotta tietäisimme, mikä sääennusteista on luotettavin.

Yksi tapa ratkaista tämä ongelma on jakaa käytettävissämme oleva aineisto harjoitusaineistoon, jota käytämme mallin sovittamiseen, ja testiaineistoon, jota käytämme mallin sopivuuden testaamiseen. Sateenennustamistapauksessa voisimme sovittaa erilaiset mallimme 9 ensimmäisen vuoden aineiston pohjalta, ja käyttää viimeisen vuoden aineistoa niiden ennustuskyvyn testaamiseen.

Tämän ratkaisun (ja testiaineiston käytön yleensä) ongelma on se, että testiaineisto ei välttämättä ole edustava otos kaikista mahdollisista havainnoista tarkasteltavasta ilmiöstä. Siten pienellä testiaineistolla tämän nimenomaisen testiaineiston erityispiirteet voivat vaikuttaa liikaa tuloksiin. Todellisuudessa olemme kiinnostuneet logaritmisen uskottavuusfunktion odotusarvosta

$$E[\log \hat{p}(\tilde{\mathbf{Y}})], \tag{4.4}$$

missä odotusarvo otetaan uusien havaintojen $\tilde{\mathbf{Y}}$ (tuntemattoman) jakauman yli, ja testiaineiston perusteella laskettu log-uskottavuusfunktion arvo $\log \hat{p}(\tilde{\mathbf{y}})$ on vain approksimaatio tästä.

Toisaalta mitä suurempi osa aineistosta lohkaistaan testiaineistoksi, sitä pienempi osuus jää harjoitusaineistoksi, jota käytetään mallin sovittamiseen. Tässä mielessä aineiston jakaminen erilliseen testiaineistoon ja harjoitusaineistoon ”hukkaa informaatiota”.

4.1.4 Ristiinvalidointi

Jos käytössä ei ole erillista testiaineistoa, yleensä kannattaakin jakaa aineisto useampaan kertaan testi- ja harjoitusaineistoon, jolloin kaikkia aineiston pisteitä käytetään vuorollaan testiaineistona. Tätä kutsutaan **ristiinvalidoinniksi** (cross-validation).

Esimerkiksi 10-kertaisessa ristiinvalidoinnissa (10-fold cross-validation) aineisto jaetaan kymmeneen yhtä suureen osaan. Sen jälkeen malli sovitetaan 10 kertaa siten, että jokaisella kerralla kukin osa aineistosta jätetään testiaineistoksi, ja malli estimoidaan käyttäen yhdeksää muuta osaa harjoitusaineistona. Sen jälkeen näiden kymmenen kerran logaritmisen uskottavuusfunktion[3] arvot lasketaan yhteen, jolloin saadaan arvio logaritmisen uskottavuusfunktion odotusarvosta 4.4.

Yleisemmin k-kertaisessa ristiinvalidoinnissa aineisto jaetaan k:n yhtä suureen osaan $\mathbf{y}_1, \ldots, \mathbf{y}_k$, ja malli sovitetaan k kertaa. Merkitään

$$\hat{p}_{-i}(\mathbf{y}_i)$$

aineiston osien $\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, \ldots, \mathbf{y}_k$ perusteella sovitetun mallin ennustetta aineiston osan $\mathbf{y}_i$ todennäköisyydelle. Lopuksi lasketaan näiden yhteen näin saadut k logaritmisen uskottavuusfunktion (tai vastaavasti keskineliövirheen tai muun ennustustarkkuuden mittarin) arvot:

$$\log \hat{p}(\mathbf{y})_{CV_{(k)}} = \sum_{i=1}^{k} \log \hat{p}_{-i}(\mathbf{y}_i).$$

Tärkeä erikoistapaus k-kertaisesta ristiinvalidoinnista on $k = n$. Tätä kutsutaan nimellä leave-one-out cross-validation (LOOCV). Siinä malli sovitetaan aineistoon n kertaa, siten että yksi piste vuorollaan jätetään testiaineistoksi, ja loppuja $n-1$ pistettä käytetään harjoitusaineistona:

$$\log \hat{p}(\mathbf{y})_{LOOCV} = \log \hat{p}(\mathbf{y})_{CV_{(n)}} = \sum_{i=1}^{n} \log \hat{p}_{-i}(y_i).$$

Tämä on siinä mielessä optimaalinen jako, että jokaisella kerralla käytetään aineiston sovittamiseen maksimaalinen määrä aineistosta. Toisaalta se on laskennallisesti raskain ristiinvalidoinnin muoto, sillä malli täytyy sovittaa n kertaa (ellei käytössä ole mitään maagista oikotietä; esimerkiksi lineaarisen regression tapauksessa LOOCV:n voi suorittaa sovittamalla mallin ainoastaan kerran). Vaikka se estimoi mallin yleistysvirhettä mahdollisimman tarkasti, toisaalta myös sen varianssi on suuri. Käytännössä monesti käytetäänkin 5- tai 10-kertaista ristiinvalidointia.

[3] Tässä käytetään logaritmista uskottavuusfunktiota mallin ennustuskyvyn mittarina, mutta ristiinvalidointi toimii täsmälleen samalla tavalla myös muille kustannusfunktioille, esimerkiksi keskineliövirheelle.

Jos mallin k-kertainen ristiinvalidointi ei vie kovin paljon aikaa/laskentatehoa, se kannattaa suorittaa useaan kertaan jakamalla aina aineisto satunnaisesti k:n osaan, ja laskemalla keskiarvo tuloksista. Näin vähennetään aineiston satunnaisesta jakamisesta k:n osaan johtuvaa varianssia.

Samalla tapaa kuin testiaineistoa käytettäessä, myös ristiinvalidoinnissa logaritmisen uskottavuusfunktion paikalle voidaan sijoittaa mikä tahansa edellä käsitellystä kolmesta vaihtoehdosta.

4.1.5 Informaatiokriteerit

Informaatiokriteerit ovat ristiinvalidointia karkeampi approksimaatio mallin yleistettävyydelle: jos ristiinvalidointia ei haluta tai ei ole mahdollista tehdä joko laskennan raskauden takia, tai siksi että aineistoa ei voi jakaa n:n erilliseen osaan, mallin yleistettävyyttä voi approksimoida myös mallin sopivuuden havaittuun aineistoon, eli arvon, jonka mallin ennustejakauma $\hat{p}(\mathbf{y})$ antaa havaitulle aineistolle, avulla. Koska malli on sovitettu juuri havaittuun aineistoon, sen sopivuus havaittuun aineistoon on odotusarvomielessä parempi kuin uuteen samaa jakaumaa noudattavaan aineistoon. Informaatiokriteerit ottavat huomioon tämän ylioppimisen havaittuun aineistoon ns. rankaisutermillä, joka riippuu mallin monimutkaisuudesta: mitä monimutkaisempi malli, sitä enemmän se (odotusarvomielessä) sovittuu havaitun aineiston satunnaisvaihtelusta johtuviin erityispiirteisiin, ja vastaavasti sitä suurempi on rankaisutermi, jonka tarkoitus on tasapainottaa tämän ylioppimisen vaikutusta arviossa mallin yleistettävyydestä.

Kaikki tässä käsiteltävät informaatiokriteerit ovat muotoa

$$-2\log\hat{p}(\mathbf{y}) + 2p, \tag{4.5}$$

missä $\hat{p}(\mathbf{y})$ on mallin antama todennäköisyys havaitulle aineistolle $\mathbf{y}$, ja p on mallin **efektiivisten parametrien** määrä, joka kuvaa mallin monimutkaisuutta. Yksinkertaisimmillaan se on mallin vapaiden parametrien määrä.

Mallin logaritminen ennustejakauma on kerrottu miinus kahdella, joten mitä suuremman todennäköisyyden malli antaa aineistolle, sitä pienempi on informaatiokriteerin arvo (jos rankaisutermiä $2p$ ei oteta huomioon): hyvä malli siis on sellainen, jolle tarkasteltavan informaatiokriteerin arvo on mahdollisimman pieni. Muunnoksen $-2\log\hat{p}(\mathbf{y})$ käyttö liittyy asymptoottisiin tarkasteluihin, joilla informaatiokriteerien käyttö mallin yleistettävyyden mittarina perustellaan.

Seuraavaksi käsiteltävät kolme informaatiokriteeriä (AIC, DIC ja WAIC) saadaan sijoittamalla yleiseen kaavaan 4.5 kappaleen 4.1.3 kolme eri vaihtoehtoa valita ennustejakauma $\hat{p}(\mathbf{y})$, ja vastaavasti kolme erilaista tapaa mitata efektiivisten parametrien määrää p. Voidaan osoittaa, että kukin informaatiokriteeri on asymptoottisesti (otoskoon n

kasvaessa kohti ääretöntä) ekvivalentti sitä vastaavan tavan tehdä ristiinvalidointia kanssa.

AIC

Akaiken informaatiokriteerissä (Akaike information criterion, AIC) ennustejakauma $\hat{p}(\mathbf{y})$ on uskottavuusfunktio $p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})$, jonka parametriksi $\boldsymbol{\theta}$ on sijoitettu sen suurimman uskottavuuden estimaatti, ja efektiivisten parametrien määrä p on mallin vapaiden parametrien määrä k:

$$\mathrm{AIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}) + 2k.$$

Akaiken informaatiokriteeri käyttää suurimman uskottavuuden estimaattia, joten se ei ole bayesiläinen informaatiokriteeri. Se ei myöskään sovi hierarkkisille malleille, joissa voi olla paljon parametreja, jotka kuitenkin riippuvat toisistaan populaatiojakauman kautta, jolloin parametrien määrä yliarvioi mallin todellisen monimutkaisuuden.

DIC

Devianssi-informaatiokriteerissä (Deviance information criterion, DIC) ennustejakauma $\hat{p}(\mathbf{y})$ on uskottavuusfunktio $p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathrm{Bayes}})$, jonka parametriksi $\boldsymbol{\theta}$ on sijoitettu posteriorijakauman odotusarvo

$$\hat{\boldsymbol{\theta}}_{\mathrm{Bayes}} = E[\boldsymbol{\theta}|\mathbf{y}].$$

DIC on siten muotoa

$$\mathrm{DIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathrm{Bayes}}) + 2p_{\mathrm{DIC}},$$

missä efektiivinen parametrien määrä[4] lasketaan kaavasta:

$$p_{\mathrm{DIC}} = 2 \left(\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathrm{Bayes}}) - E_{\mathrm{post}} \left[\log p(\mathbf{y}|\boldsymbol{\theta}) \right] \right).$$

Tässä odotusarvo $E_{\mathrm{post}} \left[\log p(\mathbf{y}|\boldsymbol{\theta}) \right]$ lasketaan parametrin $\boldsymbol{\theta}$ posteriorijakauman $p(\boldsymbol{\theta}|\mathbf{y})$ yli. Efektiivinen parametrien määrä siis kuvaa mallin monimutkaisuutta, ja voi poiketa mallin vapaiden parametrien määrästä. Osoittautuu, että yksinkertaisille malleille efektiivinen parametrien määrä p_{DIC} on sama kuin mallin vapaiden parametrien määrä k.

DIC:n etuna AIC:iin verrattuna on se, että se soveltuu myös hierarkkisille malleille, ja toisaalta se käyttää posteriorijakauman odotusarvoa, joka on bayesiläinen pisteestimaatti.

[4] DIC:stä on olemassa myös versio, jossa efektiivinen parametrien määrä lasketaan kaavasta

$$p_{\mathrm{DIC2}} = 2 \, \mathrm{Var}_{\mathrm{post}} \left[\log p(\mathbf{y}|\boldsymbol{\theta}) \right], \tag{4.6}$$

missä varianssi $\mathrm{Var}_{\mathrm{post}} \left[\log p(\mathbf{y}|\boldsymbol{\theta}) \right]$ lasketaan parametrin $\boldsymbol{\theta}$ posteriorijakauman yli. Tämä versio antaa hyvin samankaltaisia tuloksia, mutta sen etuna on, että se on aina positiivinen. Toisaalta ensimmäinen versio on numeerisesti vakaampi.

WAIC

DIC kuitenkin tiivistää edelleen posteriorijakaumaan sisältyvän epävarmuuden parametrin todellisesta arvosta piste-estimaattiin. **Watanabe-Akaike-informaatiokriteeri** (Watanabe-Akaike information criterion) sen sijaan on täysin bayesiläinen informaatiokriteeri: siinä käytetään ennustejakaumana $\hat{p}$ uusien havaintojen (pisteittäistä) posteriorienn ustejakaumaa[5]

$$p(\tilde{y}_i = y_i|\mathbf{y}).$$

Efektiivinen parametrien määrä[6] taas lasketaan kaavasta

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\Big(\log p(\tilde{y}_i = y_i|\mathbf{y}) - E_{\text{post}}[\log p(y_i|\boldsymbol{\theta})]\Big). \tag{4.7}$$

Tämä muistuttaa hyvin paljon kaavaa 4.6 DIC:n efektiivisten parametrien määrälle, mutta ensinnäkin siinä käytetään logaritmista posterioriennustejakaumaa $\log p(\tilde{y}_i = y_i|\mathbf{y})$ log-uskottavuusfunktion, johon on sijoitettu posteriorijakauman keskiarvo, sijaan, ja toiseksi se lasketaan pisteittäin, kun taas kaavan 4.6 erotus lasketaan koko aineistolle. Tämä on myös WAIC:n heikkous verrattuna kahteen muuhun informaatiokriteeriin: kuten ristiinvalidointi, myös se vaatii, että aineisto on jaettavissa n:ään havaintoon.

Logaritmisen posterioriennustejakauman arvoa yksittäiselle havainnolle y_i voi kätevästi approksimoida (vrt kaava 4.3) posteriorijakaumasta simuloidun otoksen $\boldsymbol{\theta}_1^{\text{sim}}, \ldots, \boldsymbol{\theta}_S^{\text{sim}}$ avulla:

$$\log p(\tilde{y}_i = y_i|\mathbf{y}) \approx \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}_s^{\text{sim}})\right).$$

Myös kaavan 4.7 summan toista termiä pystyy approksimoimaan vastaavalla tavalla:

$$E_{\text{post}}[\log p(y_i|\boldsymbol{\theta})] \approx \frac{1}{S}\sum_{s=1}^{S} \log p(y_i|\boldsymbol{\theta}_s^{\text{sim}}).$$

[5]Tässä kohtaa käytetty lyhennysmerkintä on hieman hämäävä. Kyseessä on siis uuden havainnon $\tilde{Y}_i$ (satunnaismuuttujana) jakauma ehdolla havaittu aineisto $\mathbf{Y} = \mathbf{y}$:

$$f_{\tilde{Y}_i|\mathbf{Y}(\cdot|\mathbf{y})},$$

johon on sijoitettu argumentin paikalle havaittu piste y_i, eli yksinkertaisesti lasketaan arvo, jonka posterioriennustejakauma antaa havaitulle aineiston pisteelle.

[6]Myös WAIC:lle efektiivisten parametrien määrälle on olemassa toinen muoto, jossa käytetään varianssia:

$$p_{\text{WAIC2}} = \sum_{i=1}^{n} \text{Var}_{\text{post}}\left[\log p(y_i|\theta)\right].$$

Ensimmäinen termi siis on logaritminen posteriorienn ustejakauma, eli logaritmi uskottavuusfunktion odotusarvosta posteriorijakauman yli, ja toinen termi taas odotusarvo logaritmisesta uskottavuusfunktiosta posteriorijakauman yli.

4.1.6 Bayes-faktori

Malleja voi vertailla myös täysin bayesiläisessä viitekehyksessä, jolloin niille annetaan prioritodennäköisyydet, lasketaan aineiston reunajakauma ehdolla malli, eli reunauskottavuusfunktio (marginal likelihood), ja sen jälkeen vertaillaan mallien posterioritodennäköisyyksiä.

Kahden vaihtoehtoisen mallin H_0 ja H_1 tapauksessa tarkastellaan niiden posterioritodennäköisyyksien suhdetta

$$\frac{p(H_1|\mathbf{y})}{p(H_0|\mathbf{y})} = \frac{p(H_1)p(\mathbf{y}|H_1)}{p(H_0)p(\mathbf{y}|H_0)} = \frac{p(H_1)}{p(H_0)}\ \mathrm{BF}(H_1, H_0).$$

Bayes-faktoriksi kutsutaan aineiston reunauskottavuusfunktioiden suhdetta

$$\mathrm{BF}(H_1, H_0) = \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)}.$$

Bayes-faktori siis antaa kertoimen, jolla aineiston $\mathbf{y}$ havaitseminen muuttaa käsitystämme mallien H_1 ja H_0 todennäköisyyksien suhteesta. Jos mallien prioritodennäköisyydet ovat yhtä suuret, eli $p(H_1) = 1/2 = p(H_0)$, mallien posterioritodennäköisyyksien suhde on niiden Bayes-faktori.

Bayes-faktori voi saada arvoja väliltä $(0, \infty)$; jos se on suurempaa kuin yksi, aineisto tukee mallia H_1, ja jos se on pienempää kuin yksi, aineisto tukee mallia H_0. Mitä suurempi tai pienempi Bayes-faktori on, sitä suurempi aineiston tuki kyseiselle mallille.

Reunauskottavuusfunktiot saadaan integroimalla mallia sen parametrin yli,

$$p(\mathbf{y}|H_i) = \int p(\boldsymbol{\theta}_i|H_i)p(\mathbf{y}|\boldsymbol{\theta}_i, H_i)\,\mathrm{d}\boldsymbol{\theta}_i$$

eli ne ovat aineiston tästä mallista laskettuja prioriennustejakaumia.

4.2 Bayesiläiset uskottavuusvälit

Bayesiläisellä 95% (tai vastaavasti minkä tahansa muun osuuden) uskottavuusvälillä (credible interval[7]) tarkoitetaan väliä, jolla sijaitsee 95% posteriorijakauman todennäköisyysmassasta. Siten voidaan yksinkertaisesti sanoa, että parametrin todellinen arvo sijaitsee

[7]Luontevin suomennus olisi ehkä uskottavuusväli, mutta valitettavasti sillekin on frekventistisessä tilastotieteessä vakiintunut merkitys. Kutsutaan siten näitä välejä hieman epätyydyttävästi bayesiläisiksi uskottavuusväleiksi.

95% prosentin Bayesiläisellä uskottavuusvälillä 95% prosentin todennäköisyydellä! Bayesiläiset uskottavuusvälit ovatkin käsitteellisesti ja tulkinnallisesti huomattavasti frekventistisiä luottamusvälejä kätevämpiä. Vaikea osuus on tehty jo posteriorijakaumaa laskettaessa tai simuloitaessa; tämän jälkeen kaikki päätelmät parametrien arvoista ja niihin liittyvästä epävarmuudesta voidaan tehdä tästä lasketusta tai simuloidusta posteriorijakaumasta.

Jatkuvasta jakaumasta voidaan valita väli, joka sisältää 95% sen todennäköisyysmassasta äärettömän monella eri tavalla. Yleensä tehdään kuten frekventististen luottamusvälien tapauksessa, eli valitaan tämä väli jakauman keskeltä. Siten esimerkiksi 95% bayesiläinen uskottavuusväli valitaan yleensä siten, että sen alaraja on posteriorijakauman 0.025-kvantiili, ja yläraja sen 0.975-kvantiili. Ellei toisin mainita, bayesiläisellä uskottavuusvälillä tarkoitetaan tätä (todennäköisyysmassan mielessä) keskitettyä väliä, jota kutsutaan joskus myös nimellä equal-tailed credible interval.

Vaihtoehtoinen tapa valita bayesiläinen uskottavuusväli on valita kapein mahdollinen väli, joka sisältää 95% jakauman todennäköisyysmassasta, eli ns. highest posterior density region (HPD). Symmetristen yksihuippuisten jakaumien, kuten normaalijakauman, tapauksessa nämä kaksi tapaa antavat saman tuloksen, mutta jos jakauma on monihuippuinen, ei HPD välttämättä ole väli, vaan se voi koostua useammasta erillisestä välistä välistä.

Luku 5

Lineaariset mallit

Lineaarinen malli ja sen yleistykset ovat keskeisimpiä tilastotieteen soveltajan työkaluja. Käsitteellisesti lineaarinen malli ei tuo juurikaan uutta tähän mennessä opittuun, vaan sen perustana olevat normaalijakaumalaskut (tai oikestaan niiden yksiulotteiset versiot) on jo käyty läpi.

Erona aiempaan on, että nyt jokaiseen havaintoon y_i liittyvät selittävien muuttujien arvot $\mathbf{x}_i = (x_{11}, \ldots, x_{1k})$. Olemme kiinnostuneet Y_i:n arvoista ehdolla selittävien muuttujien arvot, joten selitettävien muuttujien arvoja voidaan pitää vakioina.

Kerätään selitettävän muuttujan arvot vektoriin

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

ja selittävien muuttujien arvot matriisiin

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nk} \end{bmatrix},$$

jolloin jatkossa voidaan käyttää käteviä matriisimerkintöjä mallille. Monesti malliin halutaan mukaan vakiotermi, jolloin selittävien muuttujien matriisin ensimmäiksi sarakkeeksi annetaan ykkösvektori: $(x_{11}, \ldots, x_{n1}) = \mathbf{1}_n$. Mallin parametria $\boldsymbol{\beta}$, eli regressiokertoimia merkitään myös vektorilla

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

5.1 Klassinen lineaarinen malli

Klassisessa lineaarisessa mallissa (ordinary least squares regression) oletetaan, että selitettävän muuttujan arvot noudattavat normaalijakaumaa ehdolla selitävien muuttujien arvo, ja että tämän normaalijakauman odotusarvo on lineaarikombinaatio parametrivektorin $\boldsymbol{\beta}$ muuttujien arvoista:

$$E[Y_i \,|\, \boldsymbol{\beta}, \mathbf{x}_i] = \mathbf{x_i}^T \boldsymbol{\beta} = x_{i1}\beta_1 + \cdots + x_{ik}\beta_k,$$

ja että näiden normaalijakaumien varianssi on vakio. Toisin sanoen oletetaan, että aineiston jakauma on muotoa:

$$Y_1, \ldots, Y_n \perp\!\!\!\perp | \, \boldsymbol{\beta}, \sigma^2, \quad Y_i \,|\, \boldsymbol{\beta}, \sigma \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad \text{kaikille } i = 1, \ldots, n,$$

eli

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}),$$

yllä määritellyillä matriisimerkinnöillä.

Mallin parametri on siis $(\boldsymbol{\beta}, \sigma^2)$. Yleisesti käytetty epäinformatiivinen priorijakauma on epäoleellinen

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Parametrin $\boldsymbol{\beta}$ reunaposteriorijakauma $p(\boldsymbol{\beta}, |\sigma^2, \mathbf{y})$ ehdolla varianssi σ^2 on

$$\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, V_{\boldsymbol{\beta}} \sigma^2),$$

missä

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

ja

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1}.$$

Varianssin σ^2 reunaposteriorijakauma on

$$\sigma^2 | \mathbf{y} \sim \text{Inv-}\chi^2_{n-k}(s^2),$$

missä

$$s^2 = \frac{1}{n-k} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Osoittautuu, että posteriorijakauma on olemassa, jos aineiston otoskoko on suurempi kuin selittäjien määrä, eli $n > k$, ja matriisin $\mathbf{X}$ sarakkeet ovat lineaarisesti riippumattomia,eli se on täyttä astetta, eli sen aste on k. Epäoleellisella priorilla $p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$ saadaan lineaariselle regressiolle siis oleellisesti samat tulokset kuin klassisessa teoriassa.

Kirjallisuutta

[1] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley Series in Probability & Statistics. Wiley, 1994.

[2] José M Bernardo. The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122, 1996.

[3] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[4] Petri Koistinen. Todennakoisyyslaskenta. `http://wiki.helsinki.fi/pages/viewpage.action?pageId=196948970`, 2013.

[5] Pekka Nieminen and Saikkonen Pentti. Tilastollinen paattely. `http://wiki.helsinki.fi/pages/viewpage.action?pageId=164335164`, 2013.

[6] G.A. Young and R.L. Smith. *Essentials of Statistical Inference*. Cambridge Series in Statistica. Cambridge University Press, 2005.