Tilastollinen päättely 2, 10 op

Arto Luoma
Tilastotiede
Informaatiotieteiden yksikkö
33014 TAMPEREEN YLIOPISTO

Kevät 2014

Kirjallisuutta

- Garthwaite, Jolliffe, Jones, *Statistical Inference*, Second Edition, Prentice Hall, 2002
- Young, Smith, *Essentials of Statistical Inference*, Cambridge University Press, 2005
- Casella, Berger, *Statistical Inference*, Brooks/Cole, 2002
- Williams, *Weighing the Odds, a Course in Probability and Statistics*, Cambridge University Press, 2001
- Ross, *Introduction to Probability Models*, Academic Press, 2002
- Davison, *Statistical Models*, Cambridge University Press, 2003

Kurssin sisältö

Ensimmäinen jakso

1. bayesiläisen päättelyn perusteet (priorijakauma, uskottavuusfunktio, posteriorijakauma, Bayes-estimaattori)
2. konjugaattinen priorijakauma, epäinformatiivinen priorijakauma,
3. bayesiläiset luottamusvälit, bayesiläinen hypoteesien testaus
4. johdatus epäparametrisiin ja robusteihin menetelmiin (permutaatiotestit, järjestyslukutestit, M-estimaattorit)
5. laskentaintensiiviset menetelmät (bootstrap, linkkuveitsimenetelmä)

Toinen jakso

6. johdatus Markovin ketjuihin
7. laskentaintensiiviset menetelmät (Gibbsin poimija, Metropolis-algoritmi)
8. tilastollinen mallinvalinta (ristiinvalidointi, informaatiokriteerit)

1 Bayesiläinen tilastotiede

1.1 Perusteita

Bayesin kaava

Bayesiläinen tilastotiede perustuu Bayesin kaavan käyttöön. Oletetaan, että tapahtumat $A_1, A_2, ..., A_k$ osittavat otosavaruuden Ω, (ts. $\Omega = A_1 \cup A_2 \cup ... \cup A_k, A_i \cap A_j = \emptyset$, kun $i \neq j$), ja B on jokin otosavaruuden tapahtuma, jolle $\Pr(B) > 0$. Tällöin Bayesin kaava on

$$\Pr(A_j|B) = \frac{\Pr(A_j)\Pr(B|A_j)}{\sum_{j=1}^{k}\Pr(A_j)\Pr(B|A_j)}.$$

Bayesin kaavalla voidaan kääntää ehdollisia todennäköisyyksiä. Jos nimittäin tunnetaan osatapahtumien A_j todennäköisyydet ja tapahtuman B ehdolliset todennäköisyydet $\Pr(B|A_j), j = 1, ..., k$, kaavan avulla voidaan laskea osatapahtumien A_j ehdolliset todennäköisyydet $\Pr(A_j|B)$. Bayesiläisessä kielenkäytössä tapahtumien A_j todennäköisyyksiä $\Pr(A_j)$kutsutaan *prioritodennäköisyyksiksi* ja ehdollisia todennäköisyyksiä $\Pr(A_j|B)$ posterioritodennäköisyyksiksi. Tapahtuman A_j posterioritodennäköisyys voidaan ilmoittaa sen jälkeen, kun tiedetään, onko B tapahtunut.

Esim. Diagnostiset testit

Olkoon γ jonkin sairauden esiintyvyys ja θ osoitinmuuttuja, joka ilmoittaa, onko yksittäisellä henkilöllä kyseinen sairaus ($\theta = 1$, jos sairas, ja $\theta = 0$, jos terve). Diagnostinen testi antaa tuloksen Y, jonka jakauman kertymäfunktio on $F_1(y)$, jos henkilöllä on sairaus, ja $F_0(y)$, jos ei ole.

Yleisin testityyppi on sellainen, jossa henkilö määritetään sairastuneeksi, jos $Y > y_0$, missä y_0 on aikaisempien tutkimusten perusteella kiinnitetty raja. Nyt voidaan laskea Bayesin kaavan avulla sairauden todennäköisyys, jos raja y_0 ylittyy:

$$\begin{aligned}\Pr(\theta = 1|Y > y_0) &= \frac{\Pr(\theta = 1)\Pr(Y > y_0|\theta = 1)}{\Pr(\theta = 1)\Pr(Y > y_0|\theta = 1) + \Pr(\theta = 0)\Pr(Y > y_0|\theta = 0)} \\ &= \frac{\gamma[1 - F_1(y_0)]}{\gamma[1 - F_1(y_0)] + (1 - \gamma)[1 - F_0(y_0)]}.\end{aligned}$$

Tämä todennäköisyys on testin *positiivinen ennustearvo*, $F_0(y_0)$ on testin *spesifisyys* (tn. että terve henkilö määritetään terveeksi) ja $1 - F_1(y_0)$

testin *herkkyys* (sensitivity, tn. että sairas määritetään sairaaksi). Yleensä testi pyritään suunnittelemaan niin, että sekä herkkyys että spesifisyys ovat mahdollisimman korkeita.

Priori- ja posteriorijakauma

Yleisemmässä tapauksessa parametrilla θ voi olla useita arvoja ja niiden prioritodennäköisyyksiä voidaan kuvata *priorijakauman todennäköisyysfunktiolla* tai jatkuvassa tapauksessa *priorijakauman tiheysfunktiolla* $\pi(\theta)$. Tämä kuvaa tutkijan ennakkokäsitystä eri parametriarvojen todennäköisyyksistä. (Huomaa ero klassiseen tilastotieteeseen, jossa parametrit ovat tuntemattomia mutta kiinteitä eli epäsatunnaisia). Siis bayesiläisessä tilastotieteessä parametrejäkin käsitellään satunnaismuuttujina.

Kun on käytössä havaintoaineisto y, parametria koskevaa ennakkokäsitystä voidaan tarkentaa. Bayesin kaavan avulla voidaan laskea parametrien arvoille posterioritodennäköisyydet, jotka ovat siis ehdollisia todennäköisyyksiä, kun havaintoaineisto on annettu. Jos parametri θ voi saada arvot $1, ..., k$, ja näillä on proritodennäköisyydet $\pi_1, .., \pi_k$, posterioritodennäköisyydet ilmoittaa *posteriorijakauman todennäköisyysfunktio*

$$\Pr(\theta = j|y) = \frac{\pi_j f(y|\theta = j)}{\sum_{i=1}^{k} \pi_i f(y|\theta = i)}, \quad j = 1, ..., k. \tag{1}$$

Kun parametri θ on jatkuva-arvoinen, voidaan ilmoittaa posteriorijakauman tiheysfunktio

$$\pi(\theta|y) = \frac{\pi(\theta) f(y|\theta)}{\int \pi(\theta) f(y|\theta) d\theta}, \tag{2}$$

missä siis $\pi(\theta)$ on priorijakauman tiheysfunktio ja $f(y|\theta)$ on havaintojen yhteisjakauman tiheysfunktio eli uskottavuusfunktio.

Esim. Lääkkeen markkinaosuus

Lääketehdas aikoo markkinoida uuden lääkkeen happovaivoihin ja haluaa selvittää kyseisen lääkkeen potentiaalisen markkinaosuuden θ. Tutkimuksessa haastatellaan n:ää henkilöä ja heistä Y sanoo aikovansa ostaa lääkettä. Klassisen (ei-bayesiläisen) analyysin mukaan $\theta \in [0, 1]$, ja Y noudattaa binomijakaumaa parametrein n ja θ. Aiemman perusteella tiedämme, että $\hat{\theta} = Y/n$ on erinomainen θ:n estimaattori. Se on θ:n suurimman uskottavuuden estimaattori, harhaton, minimivarianssinen, tarkentuva ja asymptoottisesti normaalinen.

Bayesiläisen lähestymistavan mukaisesti pyritään hankkimaan etukäteen tiedossa olevaa θ:aa koskevaa informaatiota. Tässä tapauksessa voidaan käyttää hyväksi aiemmin markkinoitujen vastaavantyyppisten lääkkeiden tarjoamaa markkinainformaatiota. Jos niiden valloittamien markkinaosuuksien arvot vaihtelevat välillä $0.05 - 0.15$ ja oletamme arvot mainitulla välillä tasajakutuneeksi, niin θ:n priorijakauman tiheysfunktio on tällöin

$$\pi(\theta) = \begin{cases} 1/(0.15 - 0.05) = 10, & 0.05 \leq \theta \leq 0.15 \\ 0, & \text{muualla.} \end{cases}$$

Tällöin havainnon Y ja parametrin θ yhteisjakauman tiheysfunktio on

$$u(y, \theta) = \pi(\theta) f(y|\theta) = 10 \binom{n}{y} \theta^y (1 - \theta)^{n-y},$$

kun $y = 0, 1, ..., n$, $\theta \in [0.05, 0.15]$, ja havainnon y reunajakauman todennäköisyysfunktio

$$g(y) = 10 \binom{n}{y} \int_{0.05}^{0.15} \theta^y (1 - \theta)^{n-y} d\theta.$$

Siis θ:n posteriorijakaumaksi saadaan

$$\pi(\theta|y) = \frac{\theta^y (1 - \theta)^{n-y}}{\int_{0.05}^{0.15} \theta^y (1 - \theta)^{n-y} d\theta}.$$

Esim. Rakkularuostesieni

Yhdysvaltain luoteisosien ja Kanadan tunnetuin mäntylaji on valkomänty. Se on altis rakkularuostesairaudelle, joka aiheuttaa ruostesientä ja haavaumia kuoreen. Nämä haavaumat laajenevat ja aiheuttavat pienten oksien ja taimien kuolemia. Metsänhoitaja haluaa estimoida hehtaaria kohti sairastuneiden puiden lukumäärän keskiarvon.

Sairastuneiden puiden lukumäärää hehtaarilla voidaan pitää Poisson-jakautuneena satunnaismuutujana parametrinä λ. Koska parametri λ vaihtelee eri alueilla, pidetään sitä eksponenttijakaumaa noudattavana satunnaismuuttujana λ odotusarvona θ, joten

$$\pi(\lambda) = \begin{cases} \frac{1}{\theta} e^{-\lambda/\theta}, & \text{jos } \lambda > 0 \\ 0, & \text{muualla.} \end{cases}$$

Tutkimusta varten valitaan n:ltä erilaiselta hehtaarin alueelta n:n alkion otos. Tällöin

$$u(y_1, ..., y_n, \lambda) = \frac{1}{\theta} e^{-\lambda/\theta} \left(\frac{\lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}}{\prod_{i=1}^n y_i!} \right)$$

ja

$$\begin{aligned} g(y_1, ..., y_n) &= \frac{1}{\theta \prod_{i=1}^n y_i!} \int_0^\infty e^{-\lambda(n+1/\theta)} \lambda^{\sum_{i=1}^n y_i} d\lambda \\ &= \Gamma\left(\sum_{i=1}^n y_i + 1\right) \frac{1}{(n+1/\theta)^{\sum_{i=1}^n y_i + 1}} \cdot \frac{1}{\theta \prod_{i=1}^n y_i!}, \end{aligned}$$

missä edellä oleva integraali on laskettu ottamalla huomioon, että se on gammaintegraali parametrein $\alpha = \sum_{i=1}^n y_i + 1$ ja $\beta = n + 1/\theta$.

Posteriorijakaumaksi λ:lle saamme nyt siis

$$\pi(\lambda | y_1, ..., y_n) = \frac{(n+1/\theta)^{\sum_{i=1}^n y_i + 1}}{\Gamma\left(\sum_{i=1}^n y_i + 1\right)} \cdot \lambda^{\sum_{i=1}^n y_i} e^{-\lambda(n+1/\theta)},$$

joka on gammajakauman tiheysfunktio parametrein $\alpha = \sum_{i=1}^n y_i + 1$ ja $\beta = (n\theta + 1)/\theta$.

1.2 Bayes-estimaattorit

Bayesiläisen analyysin lopputulos on posteriorijakauma. Se sisältää sekä priori-informaation että havaintoaineiston antaman informaation. Usein kuitenkin halutaan tiivistää päättelyn tulos yhteen lukuun, joka voi olla esimerkiksi posteriorijakauman odotusarvo, mediaani tai moodi. Tutkimme seuraavaksi, millaisia teoreettisia perusteita näiden piste-estimaattoreiden käytölle on.

Bayes-estimaattori määräytyy käytettävän tappiofunktion perusteella. Oletetaan, että estimaatin $\delta(y)$ käytöstä aiheutuu tappio $L_S[\delta(y), \theta]$, kun oikea parametrin arvo on θ. Määritellään, että Bayes-estimaattori on estimaattori, joka minimoi tappiofunktion posterioriodotusarvon

$$E\{L_S[\delta(y), \theta] | y\} = \int L_S[\delta(y), \theta] \pi(\theta | y) d\theta.$$

Tappiofunktion odotusarvoa otoksen yhteisjakauman suhteen kutsutaan *riskifunktioksi*,

$$R_\delta(\theta) = E\{L_S[\delta(y), \theta]|\theta\} = \int L_S[\delta(y), \theta] f(y|\theta) dy,$$

ja riskifunktion odotusarvoa priorijakauman suhteen *Bayes-riskiksi*,

$$B_\delta^\pi = E[R_\delta(\theta)] = \int R_\delta(\theta)\pi(\theta) d\theta.$$

Koska integroimisjärjestystä voidaan vaihtaa (Fubinin lause) ja $\pi(\theta)f(y|\theta) = f(y)\pi(\theta|y)$,

$$\begin{aligned} \int R_\delta(\theta)\pi(\theta)d\theta &= \int \pi(\theta) \int L_S[\delta(y), \theta] f(y|\theta) dy d\theta \\ &= \int f(y) \int L_S[\delta(y), \theta]\pi(\theta|y) d\theta dy, \end{aligned}$$

nähdään, että Bayes-riski minimoituu, kun kullakin y:llä minimoidaan sisempää integraalia. Sisempi integraali taas minimoituu, kun käytetään Bayes-estimaattoria.

Seuraavaksi tutkimme, millaisiin Bayes-estimaattoreihin yleisimmät tappiofunktiot johtavat.

1) Yksi-nolla -tappiofunktio

$$L_S = \begin{cases} 0, & |\delta(y) - \theta| < b \\ a, & |\delta(y) - \theta| \geq b \end{cases},$$

missä $a, b > 0$ ovat jotain vakioita.

Minimoitava suure (tappiofunktion posterioriodotusarvo) on

$$\begin{aligned} \int_{-\infty}^{\infty} L_S[\delta(y), \theta]\pi(\theta|y) d\theta &= \int_{\delta(y)+b}^{\infty} \pi(\theta|y) d\theta + \int_{-\infty}^{\delta(y)-b} \pi(\theta|y) d\theta \\ &= 1 - \int_{\delta(y)-b}^{\delta(y)+b} \pi(\theta|y) d\theta, \end{aligned}$$

joka minimoituu, kun maksimoidaan

$$\int_{\delta(y)-b}^{\delta(y)+b} \pi(\theta|y) d\theta.$$

Jos $\pi(\theta|y)$ on yksihuippuinen, maksimi saavutetaan, kun $\delta(y)$ on sellaisen välin keskipiste, jonka pituus on $2b$ ja jonka päätepisteissä $\pi(\theta|y)$:lla on sama arvo. Kun $b \to 0$, Bayes-estimaattori $\delta(y)$ lähestyy posteriorijakauman moodia. Posteriorijakauman moodi on taas sama kuin suurimman uskottavuuden estimaatti, jos parametrin priorijakauma on tasajakauma.

2) Itseisarvotappiofunktio $L_S[\delta(y), \theta] = a|\delta(y) - \theta|$, missä $a > 0$ on jokin vakio.

Yleisesti on voimassa, että jos X on jokin satunnaismuuttuja, niin odotusarvo $\mathrm{E}(|X-d|)$ minimoituu, kun d on X:n jakauman mediaani. Siis Bayes-estimaattori itseisarvotappiofunktiolla on posteriorijakauman mediaani. (Ks. tarkemmat perustelut Garthwaite).

3) Neliöllinen (kvadraattinen) tappiofunktio $L_S[\delta(y), \theta] = a(\delta(y) - \theta)^2$, missä $a > 0$ on jokin vakio.

Merkitään posteriorijakauman odotusarvoa $\hat{\theta}$. Tällöin tappiofunktion posterioriodotusarvo on

$$\begin{aligned}
\mathrm{E}\left\{[\delta(y) - \theta]^2|y\right\} &= \mathrm{E}\left\{\left[[\delta(y) - \hat{\theta}] - [\theta - \hat{\theta}]\right]^2 \middle| y\right\} \\
&= \mathrm{E}\left\{[\delta(y) - \hat{\theta}]^2|y\right\} - 2[\delta(y) - \hat{\theta}]\mathrm{E}\left\{\theta - \hat{\theta}|y\right\} + \mathrm{E}\left\{[\theta - \hat{\theta}]^2|y\right\}.
\end{aligned}$$

Toinen termi on 0, sillä $\mathrm{E}\{\theta - \hat{\theta}|y\} = 0$ ja kolmas termi ei sisällä $\delta(y)$:tä. Siten odotettu tappio minimoituu, kun $\delta(y) = \hat{\theta}$. Siis Bayes-estimaattori neliöllisellä tappiofunktiolla on posteriorijakauman odotusarvo. Odotetetun tappion minimi on $\mathrm{E}\{[\theta - \hat{\theta}]^2|y\}$, joka on θ:n posteriorivarianssi.

Esim. Rakkularuostesieni, jatkoa

Aiemmassa esimerkissä estimoitiin keskimääräistä sairastuneitten mäntyjen lukumäärää λ hehtaarilla. Posteriorijakaumaksi saatiin gammajakauma $\mathrm{Gamma}(\sum_{i=1}^n y_i + 1, (n\theta + 1)/\theta)$. Bayes-estimaattori neliöllisellä tappiofunktiolla on tämän jakauman odotusarvo $(\sum_{i=1}^n y_i + 1)/(n + 1/\theta)$. Sen sijaan posteriorijakauman moodia tai mediaania ei gamma-jakauman tapauksessa ole suljetussa muodossa.

Huomattakoon, että klassinen (ei-bayesiläinen) λ:n estimaatti on $\bar{y}$.

1.3 Konjugaattinen priorijakauma

Silloin kun posteriorijakauma kuuluu johonkin yleiseen jakaumaperheeseen, se voidaan määrittää suoraviivaisesti. Posteriorijakauman tiheysfunktio voidaan nimittäin esittää muodossa

$$\pi(\theta|y) \propto \pi(\theta)L(\theta; y),$$

missä $\pi(\theta)$ on priorijakauman tiheysfunktio ja $L(\theta; y)$ uskottavuusfunktio. Merkki $\propto$ tarkoittaa sitä, että sen molemmilla puolilla olevat lausekkeet ovat suhteessa toisiinsa. Toisin sanoen

$$\pi(\theta|y) = c\ \pi(\theta)L(\theta; y),$$

missä c on θ:sta riippumaton vakio. Jos siis tulo $\pi(\theta)L(\theta; y)$ on vakiokerrointa lukuunottamatta jonkin θ:n jakauman tiheysfunktio, voidaan päätellä, että kyseinen jakauma on θ:n posteriorijakauma.

Esim. Beta-jakauma

Aiemmassa esimerkissä estimoitiin markkinaosuutta ja priorijakaumana oli tasajakauma välillä [0.05,0.15]. Sen sijaan voisimme käyttää priorijakaumana Beta(α, β)-jakaumaa, jonka tiheysfunktio on

$$\pi(\theta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{kun } 0 < \theta < 1.$$

Koska θ:n prioriodotusarvo on $\mathrm{E}(\theta) = \alpha/(\alpha+\beta)$, valitsemalla esim. $\alpha = 1$ ja $\beta = 9$, saadaan priorijakauman odotusarvoksi sama kuin aiemmassa esimerkissä. Posteriorijakauman tiheysfunktio on

$$\pi(\theta|y) \propto \pi(\theta)L(\theta; y) \propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1},$$

josta nähdään heti, että posteriorijakauma on Beta$(\alpha+y, \beta+n-y)$.

Posteriorijakauman odotusarvo (Bayes-estimaattori neliöllisellä tappiofunktiolla) on $(\alpha+y)/(\alpha+\beta+n)$ ja posteriorijakauman moodi (Bayes-estimaattori yksi-nolla-tappiofunktiolla, kun tappiottoman välin pituus lähestyy nollaa) $(\alpha+y-1)/(\alpha+\beta+n-2)$ oletuksella, että posteriorijakauma on yksihuippuinen. Posteriorijakauman mediaanin voi määrittää taulukoista tai tietokoneella.

Esim. Normaalijakauma

Olkoon $Y_1, Y_2, ..., Y_n$ satunnaisotos normaalijakaumasta $N(\theta, \sigma^2)$, missä varianssi σ^2 on tunnettu, ja olkoon tuntemattoman parametrin θ priorijakauma $N(\mu, \tau^2)$. Tällöin θ:n posteriorijakauma on

$$\begin{aligned}\pi(\theta|y) \propto \pi(\theta)L(\theta; y) &= (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right\} \\ &\quad \times \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{y_i-\theta}{\sigma}\right)^2\right\}.\end{aligned}$$

Määritettäessä posteriorijakaumaa voidaan ohittaa tekijät, jotka eivät sisällä θ:aa:

$$\begin{aligned}\pi(\theta|y) &\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu}{\tau^2}+\frac{\sum y_i}{\sigma^2}\right)\right]\right\} \\ &= \exp\left\{-\frac{(\sigma^2+n\tau^2)}{2\sigma^2\tau^2}\left[\theta - \frac{\mu\sigma^2+n\bar{y}\tau^2}{\sigma^2+n\tau^2}\right]^2 + vakio\right\}.\end{aligned}$$

Tästä nähdään, että posteriorijakauma on normaalinen odotusarvolla $(\mu\sigma^2 + n\bar{y}\tau^2)/(\sigma^2 + n\tau^2)$ ja varianssilla $[(\sigma^2 + n\tau^2)/\sigma^2\tau^2]^{-1}$. Odotusarvo voidaan kirjoittaa muodossa

$$\left(\frac{\mu}{\tau^2} + \frac{\bar{y}}{\sigma^2/n}\right) \Big/ \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right),$$

josta nähdään, että se on priorijakauman odotusarvon ja otoskeskiarvon painotettu keskiarvo painoina priorivarianssin ja otoskeskiarvon varianssin käänteisluvut. Näitä painokertoimia kutsutaan *prioritarkkuudeksi* ja *otoskeskiarvon tarkkuudeksi*, (engl. precision). Posteriorijakauman tarkkuus (varianssin käänteisluku) on puolestaan prioritarkkuuden ja otoskeskiarvon tarkkuuden summa.

Konjugaattinen priorijakauma

Edellisissä esimerkeissä posteriorijakauma kuului samaan jakaumaperheeseen kuin priorijakauma. Tällaista priorijakaumaa, joka kerrottuna uskottavuusfunktiolla antaa samaan jakaumaperheeseen kuuluvan posteriorijakauman, kutsutaan *konjugaattiseksi priorijakaumaksi* θ:lle. Usein konjugaattisten priorijakaumien käyttö helpottaa oleellisesti laskentoja.

Yleensä konjugaattinen priorijakauma voidaan määrittää niin, että korvataan uskottavuusfunktion θ:sta riippuvassa osassa otoksesta riippuvat lausekkeet priorijakauman parametreilla.

Esim. Konjugaattinen priorijakauma Poisson-jakautuneelle otokselle

Olkoon $Y_1, ..., Y_n$ otos Poi(θ)-jakaumasta. Tällöin uskottavuusfunktio on

$$L(\theta; y) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \propto \theta^{\sum y_i} e^{-n\theta}.$$

Kun korvataan otoksesta riippuvat $\sum y_i$ ja n priorijakauman parametreilla α_1 ja α_2, saadaan konjugaattiseksi priorijakaumaksi

$$\pi(\theta) \propto \theta^{\alpha_1} e^{-\alpha_2 \theta},$$

joka on Gamma($\alpha_1 + 1, \alpha_2$)-jakauma. Jos parametroidaan tämä jakauma uudelleen niin, että $\alpha = \alpha_1 + 1$ ja $\beta = \alpha_2$, saadaan priorijakaumaksi Gamma(α,β).

Esim. Konjugaattinen priorijakauma tasajakautuneelle otokselle

Olkoon $Y_1, ..., Y_n$ satunnaisotos tasajakaumasta Tas$(0, \theta)$. Nyt Y_i:n tiheysfunktio on

$$f(y_i|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq y_i \leq \theta, \\ 0, & \text{muuten}, \end{cases}$$

ja uskottavuusfunktio θ:lle on

$$\begin{aligned} L(\theta; y) &= \begin{cases} \frac{1}{\theta^n} & 0 \leq y_{(1)} \leq ... \leq y_{(n)} \leq \theta \\ 0, & \text{muuten} \end{cases} \\ &= \frac{1}{\theta^n} I_{\{\theta \geq y_{(n)}\}}(y) \; I_{\{y_{(1)} \geq 0\}}(y). \end{aligned}$$

Korvaamalla otoksesta riippuvat n ja $y_{(n)}$ priorijakauman parametreilla saadaan

$$\begin{aligned} \pi(\theta) &\propto \frac{1}{\theta^\alpha} I_{\{\theta \geq \beta\}}(\theta) \\ &= \begin{cases} \frac{1}{\theta^\alpha}, & \text{kun } \theta \geq \beta, \\ 0, & \text{muuten}. \end{cases} \end{aligned}$$

Siis konjugaattinen priorijakauma on Pareto-jakauma, jonka tiheysfunktio on

$$\pi(\theta) = \begin{cases} (\alpha - 1)\frac{\beta^{\alpha-1}}{\theta^\alpha}, & \text{kun } \theta \geq \beta, \\ 0, & \text{muuten}. \end{cases}$$

Posteriorijakaumaksi saadaan

$$\begin{aligned}\pi(\theta|y) &\propto L(\theta;y)\pi(\theta)\\ &\propto \begin{cases}\frac{1}{\theta^{n+\alpha}}, & \text{kun } \theta \geq \max(\beta, y_{(n)})\\ 0, & \text{muuten.}\end{cases}\end{aligned}$$

Siis myös posteriorijakauma on Pareto-jakauma.

1.4 Epäinformatiivinen priorijakauma

Silloin kun ei ole käytettävissä ennakkoinformaatiota estimoitavista parametreista, voidaan käyttää epäinformatiivisia priorijakaumia. Niitä voidaan käyttää myös silloin, kun halutaan tutkia, kuinka paljon informatiivinen priorijakauma vaikuttaa päättelyn lopputulokseen.

Epäinformatiivisena priorijakaumana käytetään usein tasajakaumaa $\pi(\theta) \propto 1$. Tasajakauman käyttö ei ole kuitenkaan täysin ongelmatonta. Jos tasajakauma on rajoitettu jollekin välille, se ei todellisuudessa kuitenkaan ole epä-informatiivinen. Esimerkiksi priorijakauma Tas(0,1) sisältää esim. informaation, että θ on välillä [0.2,0.4] todennäköisyydellä 0.2. Tämä informaatiosisältö tulee erityisen ilmeiseksi silloin, kun tehdään parametrinmuunnos. Muunnetun parametrin jakauma ei olekaan enää tasajakauma. Esim. jos $\theta \sim \text{Tas}(0,1)$, niin $\phi = -\log(\theta) \sim \text{Exp}(1)$.

Toinen ongelma, joka liittyy tasaisen priorijakauman käyttöön on, että parametri voi saada arvoja äärettömän pitkällä välillä. Tällöin ei ole olemassa aitoa tasajakaumaa, jonka tiheysfunktio integroituisi ykköseksi. Tässä tapauksessa voidaan kuitenkin käyttää epäaitoa tasajakaumaa, jolloin posteriorijakauman tiheysfunktio on vakiokerrointa lukuun ottamatta sama kuin uskottavuusfunktio. Tällöinkin on ongelmana se, että muunnetun parametrin priorijakauma ei ole enää tasajakauma.

Jotkut parametrit, esim. asteikkoparametrit ja jakaumien varianssit, voivat saada pelkästään positiivisia reaalilukuarvoja. Tällaisille parametreille käytetään yleensä epäaitoa priorijakaumaa $\pi(\theta) \propto 1/\theta$. Tämä vastaa sitä, että $\log(\theta)$:lla on tasainen priorijakauma.

Yhden ratkaisun määrittää parametroinnista riippumaton priorijakauma on esittänyt Jeffreys. Hänen priorijakaumansa on tasajakauma sellaiselle parametrin muunnokselle $\phi(\theta)$, jonka Fisherin informaatio on vakio. Jos $\phi = \phi(\theta)$ on jokin θ:n säännöllinen, monotoninen muunnos ja $\theta = \theta(\phi)$ sen

käänteismuunnos, niin ϕ:n Fisherin informaatio on

$$I(\phi) = \mathrm{E}\left[\frac{d\ l(\theta(\phi))}{d\phi}\right]^2 = \mathrm{E}\left[\frac{d\ l(\theta)}{d\theta}\right]^2 \times \left|\frac{d\theta}{d\phi}\right|^2 = I(\theta) \times \left|\frac{d\theta}{d\phi}\right|^2,$$

josta saadaan $I(\phi)^{1/2} = I(\theta)^{1/2}|d\theta/d\phi|$. Toisaalta jos $\pi(\theta)$ on θ:n priori, sitä vastaa ϕ:n priori $\pi(\phi) = \pi(\theta)|d\theta/d\phi|$. Tästä nähdään, että priorijakauma

$$\pi(\theta) \propto I(\theta)^{1/2} \tag{3}$$

käyttäytyy johdonmukaisesti muuttujan muunnoksissa. Tätä priorijakaumaa kutsutaan *Jeffreysin prioriksi.* Kun θ on vektori, kaavassa (3) on $I(\theta)$:n tilalla informaatiomatriisin determinantti $|I(\theta)|$.

Esim. Binomijakauma

Fisherin informaatio binomijakauman parametrille θ on $I(\theta) = n/[\theta(1-\theta)]$. Siis Jeffreysin epäinformatiivinen priorijakauma $\pi(\theta) \propto [\theta(1-\theta)]^{-1/2}$, joka on Beta-jakauma parametrein $\alpha = \beta = 1/2$.

Esim. Normaalijakauman odotusarvo

Normaalijakauman odotusarvolle θ Fisherin informaatio on $I(\theta) = n/\sigma^2$. Tämä ei riipu θ:sta, joten Jeffreysin priori on vakio $\pi(\theta) \propto 1$.

Esim. Normaalijakauman varianssi

Olkoon $Y_1, ..., Y_n$ satunnaisotos normaalijakaumasta $N(\mu, \theta^2)$, missä μ on tunnettu. Tällöin logaritmoitu uskottavuusfunktio θ:lle on

$$l(\theta) = -\frac{n}{2}\log(2\pi\theta^2) - \frac{1}{2\theta^2}\sum_{i=1}^{n}(y_i - \mu)^2,$$

josta voidaan laskea Fisherin informaatio

$$\begin{aligned} I(\theta) &= -\mathrm{E}\left[\frac{d^2\ l(\theta)}{d\theta^2}\right] = -\mathrm{E}\left[\frac{n}{\theta^2} - \frac{3}{\theta^4}\sum(y_i - \mu)^2\right] \\ &= -\frac{n}{\theta^2} + \frac{3n}{\theta^2} = \frac{2n}{\theta^2}. \end{aligned}$$

Siis Jeffreysin priori on $\pi(\theta) \propto 1/\theta$.

Huom. Jos normaalijakaumassa odotusarvo μ ja hajonta σ ovat tuntemattomia, yleensä käytetään yhteispriorijakaumaa $\pi(\mu, \sigma) \propto 1/\sigma$ (tai jos parametroidaan varianssin avulla, $\pi(\mu, \sigma^2) \propto 1/\sigma^2$). Tämä on eri priori kuin se, joka saataisiin muodostamalla Jeffreysin priori parametrivektorille $\theta = (\mu, \sigma)$.

Taulukko 1: Kuolleisuudet r/m sydänleikkauksista 12 sairaalassa (Spiegelhalter et. al,*BUGS 0.5 Examples Volume 1*, Cambridge : MRC Biostatistics Unit., 1996). Taulukossa on annettu kuolemien lkm r ja leikkausten lkm m.

A	0/47	B	18/148	C	8/119	D	46/810	E	8/211	F	13/196
G	9/148	H	31/215	I	14/207	J	8/97	K	29/256	L	24/360

1.5 Bayesiläiset luottamusvälit

Posteriorijakauman informaatiota voidaan tiivistää käyttämällä piste-estimaattoreita, kuten posteriorijakauman odotuarvo tai moodi. Yleensä kuitenkin luottamusalueet tai -välit ovat hyödyllisempiä. Tason $1-2\alpha$ bayesiläinen luottamusalue on parametrin θ arvojen joukko C, johon θ kuuluu vähintään todennäköisyydellä $1-2\alpha$. Kun θ on jatkuva,

$$1-2\alpha = \Pr(\theta \in C|y) = \int_C \pi(\theta|y)d\theta.$$

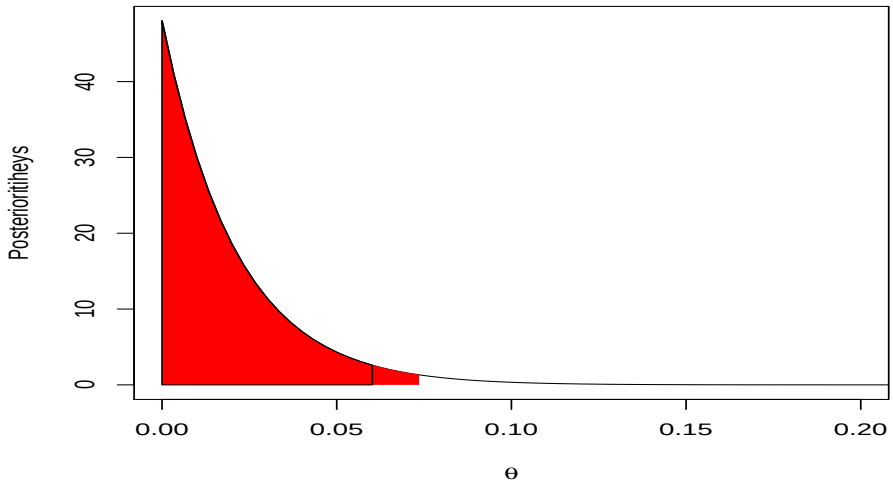
Kun θ on diskreetti, integraali korvataan summalla $\sum_{\theta\in C}\pi(\theta|y)$.

Skalaariarvoiselle parametrille θ voidaan määrittää bayesiläisiä luottamusvälejä. Tasahäntäinen (equi-tailed) bayesiläinen luottamusväli on muotoa (θ_L, θ_U), missä $\Pr(\theta < \theta_L|y) = \Pr(\theta > \theta_U|y) = \alpha$.

Käytetään myös korkeimman posterioritiheyden alueita (highest posterior density, lyh. HPD), joissa posterioritiheys on suurempi kuin missään pisteessä alueen ulkopuolella. Toisin sanoen, jos $\theta \in C$, niin $\pi(\theta|y) \geq \pi(\theta'|y)$ kaikille $\theta' \notin C$. HPD-alueella on se etu, että se on kooltaan pienin tason $1-2\alpha$ bayesiläinen luottamusalue. Toisaalta tasahäntäisellä luottamusvälillä on se etu, että tehtäessä parametrille monotoninen muunnos uuden luottamusvälin rajat saadaan tekemällä vastaava muunnos alkuperäisen luottamusvälin rajoille. HPD-välillä tällaista ominaisuutta ei ole.

Esim. Sydänleikkausaineisto

Taulukossa on aineisto vauvojen sydänleikkauksista 12 sairaalassa. Yksinkertainen malli on, että kuolemien lukumäärä r on binomijakautunut parametrein θ ja m. Esimerkiksi airaalassa A $m = 47$ ja $r = 0$, mikä antaa suurimman uskottavuuden estimaatin $\hat{\theta}_A = 0/47 = 0$. Tämä estimaatti vaikuttaa optimistiselta, kun ottaa huomioon kuolleisuuden muissa sairaaloissa. Jos käytämme priorijakaumana tasajakaumaa välillä [0,1], joka on



Kuva 1: Posterioritiheysfunktio θ_A:lle, kun priorijakauma on tasajakauma.

Beta(1,1)-jakauma, posteriorijakauma on Beta(1,48). Tällöin posteriorijakauman odotusarvo on 1/49 ja 95% HPD-posterioriväli on (0,6.05)%. Tasahäntäinen posterioriväli käyttää posteriorijakauman 2.5 ja 97.5% kvantiileja ja on (0.05,7.30)%. Kuviossa 1 on posteriorijakauman tiheysfunktio. Siinä on erotettu pystyviivoilla HPD-välin rajat; tasahäntäisen välin posterioritiheys on väritetty. Alkukohta poikkeaa niin vähän 0:sta, ettei se erotu kuvasta.

Toinen vaihtoehto estimoida θ_A olisi olettaa, että kuolleisuus on sama kaikissa sairaaloissa, ja yhdistää eri sairaaloiden tulokset. Kuitenkaan tämäkään vaihto ei tunnu täysin realistiselta. Kolmas lähestymistapa estimointiin, jota käsitellään myöhemmin, on käyttää hyväksi kaikkia sairaaloita koskevia tilastoja mutta samalla ottaa huomioon se, että sairaaloiden välillä voi esiintyä vaihtelua kuolleisuudessa.

1.6 Bayesiläinen hypoteesien testaus

Kun verrataan nollahypoteesia H_0 vaihtoehtoiseen hypoteesiin H_1, frekventistinen lähestymistapa perustuu testisuureen T käyttöön, joka saa tyypillisesti suuremman arvon silloin, kun H_1 on tosi ja H_0 epätosi. Nollahypoteesi H_0 hylätään riskitasolla α, jos havaittu testisuureen arvo t_{obs} on suurem-

pi kuin kriittinen arvo t_C, joka on määritelty niin, että T:n todennäköisyys ylittää kriittinen arvo on α nollahypoteesin ollessa tosi; $\Pr(T > t_C|H_0)=\alpha$. Usein lasketaan myös p-arvo, joka on todennäköisyys, että testisuure ylittää otoksesta lasketun arvon nollahypoteesin vallitessa, $p = \Pr(T \geq t_{obs}|H_0)$.

On huomattavaa, että frekventistisessä lähestymistavassa ei puhuta hypoteesien todennäköisyyksistä. Esim. p-arvoa ei pidä tulkita nollahypoteesin todennäköisyydeksi. Sen sijaan bayesiläisessä lähestymistavassa annetaan hypoteeseille prioritodennäköisyydet $\Pr(H_i),\ i = 0, 1$ ja lasketaan posterioritodennäköisyydet

$$\Pr(H_i|y) = \frac{\Pr(H_i)\Pr(y|H_i)}{\Pr(H_0)\Pr(y|H_0) + \Pr(H_1)\Pr(y|H_1)}, \quad i = 0, 1.$$

Frekventistisessä lähestymistavassa vaihtoehtoisen hypoteesin ei tarvitse olla täysin määritelty eikä laskettava p-arvo riipu valitusta vaihtoehtoisesta hypoteesista. Sen sijaan bayesiläisessä lähestymistavassa myös vaihtoehtoisen hypoteesin pitää olla täysin määritelty.

Joskus on hyödyllistä laskea vedonlyöntisuhde H_1:n eduksi,

$$\frac{\Pr(H_1|y)}{\Pr(H_0|y)} = \frac{\Pr(y|H_1)}{\Pr(y|H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)}. \tag{4}$$

Tämä suhde riippuu aineistosta y ainoastaan *Bayesin tekijän*

$$B_{10} = \frac{\Pr(y|H_1)}{\Pr(y|H_0)} \tag{5}$$

välityksellä. Siten *posteriorivedonlyöntisuhde=Bayesin tekijä* $\times$ *priorivedonlyöntisuhde.*

Kun molemmat hypoteesit ovat yksinkertaisia, B_{10} on uskottavuussuhde H_1:n eduksi. Yleensä kuitenkin molemmat hypoteesit sisältävät parametreja, sanotaan θ_0 ja θ_1. Tällöin

$$\Pr(y|H_i) = \int f(y|H_i, \theta_i)\pi(\theta_i|H_i)d\theta_i, \quad i = 0, 1,$$

missä $\pi(\theta_i|H_i)$ on θ_i:n priori H_i:n vallitessa. Nyt Bayesin tekijä on priorijakaumilla painotettujen uskottavuuksien suhde. Analogisesti uskottavuussuhdetestisuureen kanssa suuretta $2\log B_{10}$ käytetään ilmaisemaan evidenssiä H_1:n eduksi verrattaessa sitä H_0:n kanssa. Taulukossa 2 on esitetty karkea tulkinta Bayesin tekijälle.

Taulukko 2: Tulkinta Bayesin tekijälle B_{10}

B_{10}	$2\log B_{10}$	Evidenssi H_0:aa vastaan
1-3	0-2	tuskin mainitsemisen arvoinen
3-20	2-6	Positiivinen
20-150	6-10	Vahva (Strong)
>150	>10	Hyvin vahva

Esim. HUS-aineisto

Taulukossa 3 on esitelty eräällä Birminghamin klinikalla hoidettujen HUS-tapausten (haemolytic uraemic syndrome) lukumäärä vuodesta 1970 vuoteen 1989 (ks. tarkemmin Davison). Aineistossa näyttäisi olevan selvä lisäys vuoden 1980 paikkeilla. Testataksemme tätä oletamme, että vuosittaiset tapausten lukumäärät $y_1, ..., y_n$ ovat riippumattomia ja Poisson-jakautuneita odotusarvoilla $\mathrm{E}(Y_j) = \lambda_1$, kun $j = 1, .., \tau$, ja $\mathrm{E}(Y_j) = \lambda_2$, kun $j = \tau + 1, ..., n$. Muutospiste τ voi saada arvot $1, ..., n-1$.

Nollahypoteesina on $\lambda_1 = \lambda_2 = \lambda$ eli että muutosta ei tapahdu ja vaihtoehtoisena hypoteesi H_τ, jonka mukaan vuoden τ jälkeen tapahtuu muutos. Oletamme, että H_τ:n vallitessa parametreilla λ_1 ja λ_2 on riippumattomat priorijakaumat parametrein α ja β. Tällä tiheysfunktiolla on odotusarvo α/β ja varianssi α/β^2. Nyt $\Pr(y|H_\tau)$ on

$$\int_0^\infty \prod_{j=1}^{\tau} \frac{\lambda_1^{y_j}}{y_j!} e^{-\lambda_1} \times \frac{\beta^\alpha \lambda_1^{\alpha-1} e^{-\beta\lambda_1}}{\Gamma(\alpha)} d\lambda_1 \int_0^\infty \prod_{j=\tau+1}^{n} \frac{\lambda_1^{y_j}}{y_j!} e^{-\lambda_2} \times \frac{\beta^\alpha \lambda_2^{\alpha-1} e^{-\beta\lambda_2}}{\Gamma(\alpha)} d\lambda_2,$$

joka sievenee muotoon

$$\frac{\beta^{2\alpha}}{\Gamma(\alpha)^2 \prod_{j=1}^n y_j!} \frac{\Gamma(\alpha + s_\tau)\Gamma(\alpha + s_n - s_\tau)}{(\beta+\tau)^{\alpha+s_\tau}(\beta+n-\tau)^{\alpha+s_n-s_\tau}},$$

missä $s_\tau = y_1 + ... + y_\tau$ ja $s_n = y_1 + ... + y_n$.

Oletamme myös, että H_0:n vallitessa λ on gamma-jakautunut parametrein α ja β. Bayesin tekijäksi saadaan, kun muutospiste on vuonna τ,

$$B_{\tau 0} = \frac{\Gamma(\alpha + s_\tau)\Gamma(\alpha + s_n - s_\tau)\beta^\alpha(\beta+n)^{\alpha+s_n}}{\Gamma(\alpha)\Gamma(\alpha+s_n)(\beta+\tau)^{\alpha+s_\tau}(\beta+n-\tau)^{\alpha+s_n-s_\tau}}, \quad \tau = 1, ..., n-1.$$

Taulukko 3:

	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
y	1	5	3	2	2	1	0	0	2	1
$2\log B_{\tau 0}, \alpha = \beta = 1$	4.9	-0.5	0.6	3.9	7.5	13	24	35	41	51
$2\log B_{\tau 0}, \alpha = \beta = 0.01$	-1.3	-5.9	-4.5	-1.0	3.0	9.7	20	32	39	51
$2\log B_{\tau 0}, \alpha = \beta = 0.0001$	-10	-15	-14	-10	-6.1	0.6	11	23	30	42

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
y	1	7	11	4	7	10	16	16	9	15
$2\log B_{\tau 0}, \alpha = \beta = 1$	63	55	38	42	40	31	11	-2.9	-5.3	0
$2\log B_{\tau 0}, \alpha = \beta = 0.01$	64	57	40	47	46	38	18	1.8	1.2	0
$2\log B_{\tau 0}, \alpha = \beta = 0.0001$	55	48	31	38	37	29	8.8	-7.1	-7.7	0

Taulukosta 3 nähdään $2\log B_{\tau 0}$, kun $\tau = 1, ..., 19$ ja käytetään erilaisia priorijakaumia. Evidenssi vaihtoehtoisen hypoteesin hyväksi on erittäin vahva kaikilla priorijakaumilla silloin, kun muutosvuosi on välillä 1976-1985.

1.7 Bayesiläinen ennustaminen

Tulevan havainnon Z arvon ennustaminen on suoraviivaista, kun parametrien priorijakauma on tiedossa. Yhteisjakauma Z:lle ja aineistolle Y voidaan kirjoittaa muodossa

$$f(y, z) = \int f(z|y, \theta) f(y|\theta) \pi(\theta) d\theta,$$

ja kun Y on saanut arvon y, Z:aan kohdistuva päättely perustuu sen *posterioriennustejakauman* tiheysfunktioon

$$f(z|y) = \int f(z|y, \theta) \pi(\theta|y) d\theta = \frac{\int f(z|y, \theta) f(y|\theta) \pi(\theta) d\theta}{\int f(y|\theta) \pi(\theta) d\theta},$$

joka on siis posteriorijakauman tiheysfunktiolla painotettu ennustejakauman $f(z|y, \theta)$ tiheysfunktio.

Posterioriennustejakaumaa voidaan simuloida generoimalla parametrit $\theta_l,\ l = 1, ..., L$ posteriorijakaumasta $\pi(\theta|y)$ ja generoimalla näitä vastaavat havainnot $Z_l,\ l = 1, ..., L$ ennustejakaumista $f(z|y, \theta_l)$.

Esim. Bernoulli-kokeet

Heitettäessä kolikkoa n ensimmäisen heiton joukossa on r kruunaa. Mikä on kruunan todennäköisyys seuraavalla heitolla?

Olkoon θ kruunan tuntematon todennäköisyys ja merkitköön $Z = 1$ tapahtumaa, että seuraava heitto antaa kruunan. Kun ehdollistetaan θ:n suhteen, Z on riippumaton edellisistä heitoista: $\Pr(Z = 1|y, \theta) = \theta$. Jos θ:n priorijakauma on $\text{Beta}(\alpha, \beta)$, niin

$$\begin{aligned}
\Pr(Z = 1|y) &= \int_0^1 \Pr(Z = 1|\theta, y)\pi(\theta|y)d\theta \\
&= \int_0^1 \theta \frac{\theta^{\alpha+r-1}(1-\theta)^{\beta+n-r-1}}{B(\alpha+r, \beta+n-r)} d\theta \\
&= \frac{\alpha+r}{\alpha+\beta+n}.
\end{aligned}$$

Kun $n, r \to \infty$, tämä lähestyy kruunien osuutta otoksessa r/n, jolloin priori-informaation vaikutus häviää.

2 Johdatus epäparametrisiin ja robusteihin menetelmiin

2.1 Johdanto

Tähän mennessä esitetyssä päättelyssä, sekä frekventistisessä että bayesiläisessä, olemme olettaneet havaintojen jakauman $f(y; \theta)$ tunnetuksi parametria θ lukuun ottamatta. Käytännössä tällainen oletus on usein epärealistinen. Usein ei ole mahdollista varmistua jakauman muodosta, esim. havaintojen vähäisen määrän vuoksi.

Tämän vuoksi on hyvä kehittää menetelmiä, jotka eivät riipu jakaumaoletuksista tai eivät ole herkkiä näille oletuksille. Nämä vaatimukset johtavat *epäparametrisiin* ja *robusteihin* menetelmiin. Seuraavassa esitetään lyhyt katsaus näihin menetelmiin.

2.2 Epäparametrinen päättely

Oletamme niin kuin aikaisemminkin, että meillä on satunnaisotos $(y_1, ..., y_n)$ jakaumasta, jonka tiheysfunktio $f(y; \theta)$ on tuntematon. Jos meillä olisi tie-

dossa jakaumaperhe $\{f(y;\theta)|\theta \in \Theta\}$, olisi selvää mitä tarkoitetaan parametria θ koskevalla päättelyllä. Kuitenkin tietyt parametrit, kuten mediaani ja muut kvantiilit, ovat hyvin määriteltyjä vaikka jakauman muotoa ei olisi oletettukaan. Käsittelemme seuraavassa tällaisiin parametreihin liittyvää päättelyä.

Kun $f(y;\theta)$ on tuntematon, yksi mahdollinen lähestymistapa on kehittää menetelmiä, jotka eivät riipu jakauman muodosta. Näitä kutsutaan *jakaumasta riippumattomiksi* (distribution free) menetelmiksi. Yleensä on mahdotonta päästä täysin eroon jakaumaan liittyvistä oletuksista, mutta oletukset pyritään pitämään mahdollisimman vähäisinä. Esimerkiksi oletetaan jakauman olevan jatkuva ja symmetrinen. *Epäparametrinen päättely* (nonparametric) viittaa tarkasti ottaen menetelmiin, joissa ei lainkaan esiinny parametreja, kuten yhteensopivuustesteihin. Kuitenkin yleisesti omaksutun kielenkäytön mukaisesti 'epäparametrinen' (tai ei-parametrinen, parametriton) tarkoittaa samaa kuin 'jakaumasta riippumaton'.

2.3 Hypoteesien testaus

Oletetaan, että $Y_1, ..., Y_n$ on satunnaisotos jakaumasta, jonka tiheysfunktio on $f(y;\theta)$, ja θ on jakauman sijaintiparametri. Tyypillisesti epäparametrisessa päättelyssä se on mediaani, mutta sen ei välttämättä tarvitse olla. Tarkoituksena on testata hypoteesia $H_0 : \theta = \theta_0$ jotain vaihtoehtoista hypoteesia vastaan.

Jos voisimme olettaa, että jakauma on normaalinen, jolloin θ on jakauman odotusarvo, tavallinen testisuure olisi t- testisuure $T = (\bar{X} - \theta_0)/(S/\sqrt{n})$. Tämä testisuure on monella tavoin optimaalinen, kun oletus normaalisuudesta pitää paikkansa. Muussa tapauksessa epäparametrinen lähestymistapa voi olla sopiva.

Permutaatiotestit

Permutaatiotesteissä on ajatuksena muodostaa testisuureen jakauma lähtemällä liikkeelle alkuperäisen otoksen permutaatioista. Tässä yhteydessä permutaatio-sanalla on yleisempi merkitys kuin kombinatoriikassa. Ensimmäinen vaihe on löytää testisuure, joka soveltuu annetuille nollahypoteesille ja vaihtoehtoiselle hypoteesille. Seuraavaksi määritellään otoksen permutaatioiden joukko niin, että kaikkien permutaatioiden todennäköisyydet tunnetaan nollahypoteesin vallitessa. Yleensä tämä joukko valitaan niin, että kaikki permutaatiot ovat yhtä todennäköisiä H_0:n vallitessa. Sitten testi-

suureen arvo lasketaan kaikille permutaatiolle ja näin saadaan testisuureelle permutaatioiden joukon yli laskettu todennäköisyysjakauma, ns. *permutaatiojakauma.* Lopuksi määritetään p-arvo vertaamalla testisuureen havaittua arvoa sen permutaatiojakaumaan. Käytännössä usein riittää tarkastella niitä permutaatioita, jotka vastaavat testisuureen permutaatiojakauman häntiä.

Esimerkki

Tarkastellaan hypoteesiparia $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$, missä θ on jakauman sijaintiparametri. Ensimmäinen askel on löytää testisuure ja t-testisuure $T = \bar{X}/(S/\sqrt{n})$ vaikuttaa järkevältä vaihtoehdolta.

Oletamme, että olemme saaneet havainnot $y_1, ..., y_n$ ja että jakauma on symmetrinen θ:n suhteen. Silloin H_0:n vallitessa on yhtä todennäköistä havaita y_i kuin $-y_i$, $i = 1, 2, ..., n$. Saamme siis 2^n yhtä todennäköistä 'permutaatiota' antamalla joko positiivisen tai negatiivisen merkin jokaiselle havainnolle. Jokaiselle permutaatiolle voidaan laskea T ja näin saadaan T:n permutaatiojakauma. Havaitun T:n arvon sijainti tässä jakaumassa määrää testin p-arvon. Esimerkiksi jos $n = 10$ ja havaittu T on kymmenenneksi suurin 2^{10} lasketun arvon joukossa, niin p-arvo on $2 \times 10/2^{10} = 0.0195$ kaksipuoleiselle testille.

Permutaatiotestin p-arvo voidaan yleensä määrittää jollain seuraavista tavoista:

1. Approksimoidaan permutaatiojakaumaa jollain parametrisella jakaumalla, yleensä normaalijakaumalla.

2. Muodostetaan kaikki permutaatiot tietokoneella

3. Poimitaan tietokoneella satunnaisotos permutaatiojakaumasta (satunnaistamistesti)

4. Jos testisuureen arvo sijoittuu permutaatiojakauman hännille, voidaan ilman tietokonettakin käydä läpi ne permutaatiot, jotka johtavat testisuureen äärimmäisiin arvoihin.

Menetelmät 2 ja 3 ovat laskentaintensiivisiä, koska ne käyttävät hyväksi tietokoneen laskentakapasiteettia.

Järjestyslukutestit

Monissa epäparametrisissa menetelmissä havainnot järjestetään jollain sopivalla tavalla ja havainnot korvataan järjestysluvuilla. Tarkastellaan seuraavaksi hypoteesin $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ testaamista merkkisellä järjestystestillä (signed rank test, Wilcoxon test). Jakauma oletetaan symmetriseksi θ:n suhteen. Testissä arvot $|Y_i - \theta_0|, i = 1, 2, ..., n$ korvataan järjestysluvuillaan ja testisuuren on niiden järjestyslukujen summa, joille $Y_i - \theta_0$ on negatiivinen.

Permutaatiot muodostetaan niin, että kukin järjestysluku $1, 2, ..., n$ saa positiivisen ja negatiivisen arvon ja testisuure lasketaan kullekin näistä 2^n vaihtoehdosta. Nollahypoteesin vallitessa kaikki permutaatiot ovat yhtä todennäköisiä, sillä silloin on yhtä todennäköistä että erotus $Y_i - \theta_0$ saa positiivisen kuin negatiivisen merkin. Perinteisesti on suurilla n:n arvoilla käytetty permutaatiojakauman normaalijakauma-approksimaatiota, mutta tietokoneiden laskentakapasiteetin kasvu mahdollistaa 'tarkkojen' jakaumien käytön. Permutaatiojakauma voidaan yksinkertaisesti muodostaa niin, että poimitaan kaikki mahdolliset osajoukot lukujen 1,2,...,n joukosta ja lasketaan näiden osajoukkojen luvut yhteen.

Esimerkki. Resistanssimittaukset

Oletetaan, että tehdään erästä 100 Ω vastuksia 10 vastuksen otos ja mitataan niiden resistanssit:

$$99.7 \quad 99.9 \quad 100.4 \quad 100.6 \quad 100.7 \quad 101.1 \quad 101.3 \quad 101.6 \quad 102.1 \quad 102.3$$

Kun nollahypoteesi on $H_0 : \theta = 100$, erotukset $Y_i - \theta_0$ $(= Y_i - 100)$ ovat

$$-0.3 \quad -0.1 \quad 0.4 \quad 0.6 \quad 0.7 \quad 1.1 \quad 1.3 \quad 1.6 \quad 2.1 \quad 2.3$$

ja itseisarvojen $|Y_i - 100|$ järjestysluvut

$$2 \quad 1 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10.$$

Testisuure W on niiden järjestyslukujen summa, joilla $Y_i - 100$ on negatiivinen, eli 2+1=3.

Jotta saataisiin selville testisuureen merkittävyys, on laskettava sellaisten permutaatioiden lukumäärä, joilla $W \leq 3$. Näitä on 5, sillä W voi saada arvot 0, 1 ja 2 yhdellä mahdollisella tavalla ja arvon 3 kahdella tavalla (joko niin, että valitaan 3 tai valitaan 1 ja 2). Siis yksisuuntaisen testin p-arvo on $5/2^{10} = 0.0049$ ja kysytyn 2-suuntaisen $2 \cdot 0.0049 = 0.0098$.

Merkkitesti

Testataan hypoteesia, että satunnaisotos $Y_1, ..., Y_n$ on jakaumasta, jonka mediaani on θ_0. Lasketaan niiden havaintojen lukumäärä, jotka ovat pienempiä kuin θ_0. Nollahypoteesin vallitessa on yhtä todennäköistä, että havainto on pienempi kuin θ_0 kuin että se on suurempi. Annetaan - -merkki niille havainnoille, jotka ovat $< \theta_0$. Nyt H_0:n vallitessa tämä - -merkkien lukumäärä S noudattaa binomijakaumaa $S \sim \text{Bin}(\frac{1}{2}, n)$, kun oletetaan, ettei mikään havainto ole tasan θ_0. Nyt testin p-arvo on todennäköisyys, että S saa vähintään yhtä äärimmäisen arvon, kuin otoksesta laskettu arvo.

Resistanssimittaukset (jatkoa)

Lasketaan merkkitestiä vastaava p-arvo resistanssimittausaineistolle. Testisuure saa arvon $S = 2$ (negatiivisten merkkien lkm). H_0:n vallitessa $S \sim \text{Bin}(10, \frac{1}{2})$ ja

$$\begin{aligned}
\Pr(S \leq 2) &= \binom{10}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{10} + \binom{10}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^9 + \binom{10}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^8 \\
&= \frac{1}{2^{10}}\left(1 + 10 + \frac{10 \cdot 9}{2}\right) = \frac{56}{2^{10}} = 0.0547,
\end{aligned}$$

josta saadaan 2-puoleisen testin p-arvo 0.109.

Testien suhteellinen tehokkuus

Edellä on esitetty muutamia tapoja samanlaisen hypoteesin testaamiseen. Luonnollinen kysymys on, miten verrata erilaisia testausmenetelmiä. Tavallisin tapa on hyödyntää asymptoottista suhteellista tehokkuutta (ARE). Jos n_1 ja n_2 ovat pienimmät mahdolliset otoskoot testeille 1 ja 2, joilla voidaan saavuttaa koko (size) α ja voimakkuus η, niin testin 1 ARE verrattuna testiin 2 on osamäärän n_2/n_1 raja-arvo, kun α pysyy kiinteänä ja $\eta \to 1$ (jolloin $n_1, n_2 \to \infty$). Jos T_1 on tavallinen t-testisuure, T_2 merkkitestisuure ja T_3 Wilcoxonin testisuure, voidaan osoittaa, että normaalijakautuneille havainnoille $\text{ARE}[T_2, T_1] = 2/\pi$ ja $\text{ARE}[T_3, T_1] = 3/\pi$. Molemmat epäparametriset testit voivat ylittää t-testin tehokkuudessa pitkähäntäisten jakaumien tapauksessa. (Ks. enemmän yksityiskohtia ARE:n laskemisesta Garthwaitesta.)

2.4 Piste- ja väliestimointi

Luottamusvälit voidaan löytää kääntämällä hypoteesitestit. Piste-estimaattien muodostamisessa puolestaan voidaan usein käyttää hyväksi sitä tapaa, jolla epäparametrinen luottamusväli on muodostettu.

Merkkitestiin perustuva luottamusväli

Tarkastellaan merkkitestiä nollahypoteesille $H_0 : \theta = \theta_0$. Testisuure S on niiden havaintojen lukumäärä (n:stä), jotka ovat pienempiä kuin θ_0. Nollahypoteesin vallitessa $S \sim \mathrm{Bin}(n, \frac{1}{2})$ (oletuksella, ettei yksikään havainto ole tasan θ_0).

Testattaessa kaksipuoleista hypoteesia vastaan kokoa α oleva kriittinen alue on muotoa $\{S \leq s-1 \text{ tai } S \geq n-s+1\}$, jota vastaa hyväksymisalue $\{s \leq S \leq n-s\}$. Voimme kääntää tämän hyväksymisalueen luottamustason $1-\alpha$ luottamusväliksi. On helppo nähdä, että joukko $\{s \leq S \leq n-s\}$ on sama kuin $\{Y_{(s)} < \theta_0 < Y_{(n-s+1)}\}$.

Nyt

$$\begin{aligned} 1-\alpha &= \Pr[s \leq S \leq n-s | \theta = \theta_0] \\ &= \Pr[Y_{(s)} < \theta_0 < Y_{(n-s+1} | \theta = \theta_0] \\ &= \Pr[Y_{(s)} < \theta < Y_{(n-s+1)}]. \end{aligned}$$

Siis $(Y_{(s)}, Y_{(n-s+1)})$ on tason $(1-\alpha)$ luottamusväli θ:lle.

Koska luottamusväli muodostettiin käytämällä otoksen kvantiileja, luonnollinen testiä vastaava piste-estimaattori on mediaani.

Merkkiseen järjestystestiin perustuva luottamusväli

Merkkiseen järjestystestiin perustuva luottamusväli voidaan muodostaa samanlaisella, vaikkakin monimutkaisemmalla, tavalla. Merkkisen järjestystestin testisuure voidaan kirjoittaa (harjoitustehtävä) muodossa

$$W = \sum_{i=1}^{n} \phi(D_i) r(|D_i|) = \sum_{i=1}^{n} \phi(D_i) + \sum_{i=1}^{n} \sum_{j=1(\neq i)}^{n} \phi(D_i)\phi(|D_j| - |D_i|), \qquad (6)$$

missä $D_i = Y_i - \theta_0$, $r(|D_i|)$ on etäisyyden $|Y_i - \theta_0|$ järjestysluku ja

$$\phi(x) = \begin{cases} 1, & x < 0 \\ 0, & \text{muuten.} \end{cases}$$

Tarkastellaan seuravaksi $\frac{1}{2}n(n+1)$ parittaista keskiarvoa $(Y_i+Y_j)/2$, missä $i \leq j$. Näitä keskiarvoja kutsutaan *Walshin keskiarvoiksi.* Osoitamme seuraavaksi, että W on yhtä suuri kuin niiden Walshin keskiarvojen lukumäärä, jotka ovat pienempiä kuin θ_0. Tarkastellaan pareja $Y_i, Y_j,\ i \leq j$ ja oletetaan yksinkertaisuuden vuoksi, että havainnot on järjestetty niin, että $Y_i \leq Y_j$. Kun ei oteta huomioon päällekkäisiä arvoja, on olemassa seuraavat mahdollisuudet:

$$
\begin{array}{ll}
(1) & i = j,\ Y_i < \theta_0 \\
(2) & Y_i < Y_j < \theta_0 \\
(3) & Y_i < \theta_0 < Y_j,\ |D_i| > |D_j| \\
(4) & i = j,\ Y_i > \theta_0 \\
(5) & \theta_0 < Y_i < Y_j \\
(6) & Y_i < \theta_0 < \ Y_j,\ |D_i| < |D_j|
\end{array}
$$

Kolmessa ensimmäisessä tapauksessa $(Y_i + Y_j)/2 < \theta_0$ ja havaintopari lisää yhtälön (6) oikeaa puolta 1:llä. Vastaavasti $(Y_i+Y_j)/2 > \theta_0$ mahdollisuuksilla (4)-(6) ja mikään niistä ei lisää yhtälön (6) oikeaa puolta. Siksi W on yhtä suuri kuin niiden Walshin keskiarvojen lukumäärä, jotka ovat pienempiä kuin θ_0.

Olkoon testin kriittinen arvo C_α valittu niin, että

$$\Pr[C_\alpha \leq W \leq \frac{1}{2}n(n+1) - C_\alpha | \theta = \theta_0] = 1 - \alpha. \tag{7}$$

Tällöin

$$\Pr[W_{(C_\alpha)} < \theta_0 < W_{(\frac{1}{2}n(n+1)-C_\alpha+1)} | \theta = \theta_0] = 1 - \alpha, \tag{8}$$

missä $W_{(k)}$ on k:nneksi pienin Walshin keskiarvo. Siis väli $(W_{(C_\alpha)}, W_{(\frac{1}{2}n(n+1)-C_\alpha+1)})$ on luottamustason $1 - \alpha$ väli θ:lle.

Testiä vastaava piste-estimaattori, ns. *Hodges-Lehman -estimaattori* on Walshin keskiarvojen mediaani.

Esim. Resistanssimittaukset, jatkoa

Resistanssimittausesimerkissä todettiin, että nollahypoteesin ollessa tosi $\Pr[W \leq 3] = 0.0049$. Symmetriasyistä $\Pr[W \geq 52] = 0.0049$, joten $\Pr[4 \leq W \leq 51] = 0.9902$, josta saadaan muunnettua 99.02 % luottamusväliksi θ:lle $(W_{(4)}, W_{(52)}) = (100.05, 101.95)$.

Muodostettaessa luottamusväliä merkkitestin avulla, valinnanvaraa valittavan luottamustason suhteen on vähemmän. Luottamusvälin $(Y_{(2)}, Y_{(9)}) = (99.9, 102.1)$ luottamustaso on $\Pr[2 \leq S \leq 8] = 0.979$ ja välin $(Y_{(3)}, Y_{(8)}) = (100.4, 101.6)$ taso on 0.891. Huomattakoon, että merkkiseen järjestystestiin perustuva noin tason 0.99 väli on lyhyempi kuin vastaava merkkitestin väli. Tämä johtuu siitä, että merkkinen järjestystesti käyttää hyväksi enemmän otoksen antamaa informaatiota.

2.5 Yhteensopivuustestit

Edellä kiinnostuksen kohteina olivat jakaumien parametrit, vaikka jakaumien muotoa ei kiinnitetty. Yhteensopivuustestien tavoitteena on sen sijaan tehdä koko jakaumaa koskevia päätelmiä.

Khin neliö -yhteensopivuustesti

Parhaiten tunnettu yhteensopivuustesti on χ^2-yhteensopivuustesti. Yksinkertaisimmassa tapauksessa oletetaan, että havainnot $Y_1, ..., Y_n$ ovat satunnaisotos jakaumasta, jonka kertymäfunktio on $F(y)$. Tarkoituksena on testata hypoteesia $H_0: \ F(y) = F_0(y)$ kaikilla y. Testissä satunnaismuuttujan Y vaihteluväli jaetaan k luokkaan ja testisuure on

$$C = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

missä O_i on havaintojen lukumäärä ja E_i havaintojen odotettu lukumäärä luokassa i hypoteesin H_0 vallitessa.

Hypoteesin H_0 vallitessa testisuureen C asymptoottinen jakauma on $\chi^2_{(k-1)}$. Testisuure on suurella otoskoolla likimain $-2\ln\lambda = 2\sum_{i=1}^{k} O_i \ln(O_i/E_i)$, missä λ on uskottavuussuhdetestin testisuure (harjoitustehtävät). Toisaalta C voidaan johtaa suoraan pistemäärätestistä. Waldin testi johtaa samankaltaiseen testisuureeseen $C' = \sum_{i=1}^{k}(O_i - E_i)^2/O_i$.

Testi voidaan helposti laajentaa tilanteeseen, jossa jakauma $F_0(y)$ ei ole täysin määritelty vaan siinä on joitain tuntemattomia parametrejä. Kun nämä parametrit estimoidaan suurimman uskottavuuden menetelmällä, χ^2-jakauman vapausasteita vähennetään estimoitavien parametrien lukumäärällä.

Kolmogorov-Smirnov-testi

Kolmogorov-Smirnov-testiä voidaan käyttää sekä hypoteesin $H_0: \ F(y) = F_0(y)$ testaamiseen että luottamusvälin muodostamiseen kertymäfunktiolle $F(y)$, kun kyseessä on jatkuva jakauma.

Olkoot otoksen järjestyssuureet $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(n)}$. Tällöin *empiirinen kertymäfunktio* otokselle on $F_n(y) = k/n$, kun $X_{(k)} \leq y < X_{(k+1)}, \ k = 0, 1, ..., n$, missä määritellään, että $Y_{(0)} = -\infty$ ja $Y_{(n+1)} = \infty$. Testisuure on muotoa

$$D_n = \sup_y |F_n(y) - F_0(y)|.$$

Voidaan osoittaa, että testi on jakaumasta riippumaton. Testisuureen jakauma nollahypoteesin vallitessa voidaan laskea ja suurella otoskoolla n testisuureen kertymäfunktiolle on olemassa yksinkertainen approksimaatio.

Olkoon $D_{n;\alpha}$ testisuureen kriittinen arvo riskitasolla α. Siis

$$\Pr(D_n \geq D_{n;\alpha} | H_0) = \alpha.$$

Tästä voidaan helposti johtaa, että tason $1 - \alpha$ luottamusvyö kertymäfunktiolle $F(y)$ on

$$(F_n(y) - D_{n;\alpha}, \ F_n(y) + D_{n;\alpha}).$$

Siis kertymäfunktio $F(y)$ on tällä välillä samanaikaisesti *kaikilla* arvoilla y todennäköisyydellä $1 - \alpha$.

Kolmogorov-Smirnovin testin etu verrattuna χ^2-testiin on, että siinä ei tarvitse tehdä vaihteluvälin jakoa luokkiin ja luottamusvyön muodostaminen kertymäfunktiolle on helppoa. Testi voidaan myös sovittaa myös tilanteeseen, jossa on yksisuuntainen vaihtoehto $H_1: \ F(y) \geq F_0(y)$ kaikilla y. Lisäksi koska testisuureen eksakti jakauma on mahdollista määrittää, testiä voidaan soveltaa myös pienellä otoskoolla. Toisaalta χ^2-testin etuna on, että se soveltuu myös diskreeteille jakaumille ja testin yleistäminen tapaukseen, jossa nollahypoteesin jakauma ei ole täysin määritelty, on helppoa.

2.6 Semi-parametriset menetelmät

Tietyissä tilastollisissa menetelmissä osa mallista parametroidaan mutta osa jätetään parametrittomaksi. Näitä menetelmiä kutsutaan semi-parametrisiksi. Tunnettu esimerkki on suhteellisen riskin malli (proportional hazards model, Cox regression model). Määritellään *vaarafunktio (hazard function)* $h(y;\theta) = f(y;\theta)/(1 - F(y;\theta)$, missä $f(y;\theta)$ on elinajan tiheysfunktio ja $F(y;\theta)$ kertymäfunktio. Suhteellisen riskin mallissa oletetaan, että vaarafunktio on muotoa

$$h(y;\theta) = \exp(\theta_1 z_1 + ... + \theta_p z_p) h_0(y),$$

missä $z_1, ..., z_p$ ovat selittäviä muuttujia ja $h_0(y)$ on tuntematon perusvaarafunktio. Yleensä ollaan kiinnostuneita parametreista $\theta_1, ..., \theta_p$, ts. siitä miten eri tekijät kasvattavat riskiä eikä niinkään riskin suuruudesta itsessään. Mallin epäparametrinen osa $h_0(y)$ on mallissa ns. haittafunktio (nuisance function), joka eliminoidaan.

2.7 Robusti päättely

Tilastollista menetelmää sanotaan *robustiksi*, jos sen avulla tehtävät päätelmät eivät ole herkkiä poikkeamille käytettävän mallin oletuksista. Edellä käsitellyt epäparametriset ja jakaumasta riippumattomat menetelmät ovat usein luonteeltaan robusteja, mutta robustin menetelmän ei tarvitse olla jakaumasta riippumaton. Robustissa päättelyssä lähtökohtana on parametrinen malli $f(y;\theta)$, mutta siinä otetaan huomioon havaintoaineiston mahdollinen poikkeaminen oletuksesta.

Useimmiten poikkeamalla oletuksista tarkoitetaan sitä, että havaintojen todellinen jakauma $f(y;\theta)$ poikkeaa oletetusta jakaumasta $f_0(y;\theta)$, tai sitä, että aineistossa on poikkeavia havaintoja. Poikkeavien havaintojen tapauksessa voidaan ajatella, että todellinen jakauma $f(y;\theta)$ on sekoitus (mixture) jakaumasta $f_0(y;\theta)$ ja poikkeavia havaintoja generoivasta jakaumasta. Jakaumaoletuksen lisäksi voidaan tarkastella robustisuutta esim. havaintojen riippumattomuusoletuksen suhteen tai priorijakauman valinnan suhteen bayesiläisessä analyysissä.

Robustisuutta voidaan tarkastella eri ominaisuuksien suhteen. Esimerkiksi hypoteesien testauksen yhteydessä robustisuus tarkoittaa sitä, että ensimmäisen lajin virheen todennäköisyys pysyy suhteellisen muuttumattomana ja että testin voimakkuusfunktio ei ratkaisevasti heikkene, kun havaintojen jakauma poikkeaa oletetusta.

Piste-estimointi

Olkoon $Y_1, ..., Y_n$ satunnaisotos jakaumasta $F(y;\theta)$ ja $\hat{\theta}_n(Y_1, ..., Y_n)$ estimaattori parametrille θ. Laadullinen määritelmä robustisuudelle on, että pienet muutokset jakaumassa $F(y;\theta)$ aiheuttavat ainoastaan pienen muutoksen estimaattorin $\hat{\theta}$ jakaumaan.

Oletamme seuraavassa, että havaintojen todellinen jakauma $F(y;\theta)$ on ideaalisen jakauman $F_0(y;\theta)$ ja saastuttavan jakauman $F_1(y;\theta)$ sekajakauma:

$$F(y;\theta) = (1-\epsilon)F_0(y;\theta) + \epsilon F_1(y;\theta).$$

Luku ϵ voidaan ymmärtää todellisen jakauman $F(y;\theta)$ ja ideaalisen jakauman $F_0(y;\theta)$ etäisyytenä.

Usein on olemassa luvulle ϵ kriittinen arvo ϵ^*, jonka alapuolella estimaattori $\hat{\theta}_n$, joka perustuu otokseen jakaumasta $F(y;\theta)$, saa samanlaisia arvoja kuin estimaattori, joka perustuu otokseen jakaumasta $F_0(y;\theta)$. Kun tämä raja ylittyy, estimaattori voi saada ratkaisevasti erilaisia arvoja sen mukaan, onko otos jakaumasta $F(y;\theta)$ vai jakaumasta $F_0(y;\theta)$. Tätä kriittistä arvoa ϵ^* kutsutaan estimaattorin murtumapisteeksi (breakdown point).

Voidaan osoittaa (ks. viitteitä Garthwaiten et al. kirjasta), että kun jakauma $F_0(y;\theta)$ on normaalinen ja estimoidaan odotusarvoa, otoskeskiarvon murtumapiste on $\epsilon^* = 0$, otosmediaanin murtumapiste $\epsilon^* = 0.5$ ja typistetyn keskiarvon murtumapiste $\epsilon^* = \alpha$, missä α on keskiarvosta pois jätettävien havaintojen osuus.

Yksittäisen havainnon vaikutusta estimaattoriin voidaan tutkia ns. *vaikutuskäyrän (influence curve)*, tai *vaikutusfunktion*, avulla. Olkoon $T(F)$ sellainen kertymäfunktioille F määritelty funktionaali, että $T(F(y;\theta)) = \theta$. Tarkastellaan häirittyä jakaumaa F^*, jossa osuus ϵ, $0 < \epsilon < 1$, todennäköisyysmassasta siirretään pisteeseen y, ts. $F^* = (1-\epsilon)F + \epsilon\Delta_y$, missä Δ_y on kertymäfunktio satunnaismuuttujalle, joka saa arvon y todennäköisyydellä 1. Vaikutuskäyrä määritellään

$$IF(y;T,F) = \lim_{\epsilon\to 0}\left[\frac{T(F^*) - T(F)}{\epsilon}\right].$$

Otosvastine vaikutuskäyrälle on ns. *herkkyyskäyrä (sensitivity curve)*, joka määritellään

$$SC_{n-1}(y) = n\left[\hat{\theta}_n(y_1, ..., y_{n-1}, y) - \hat{\theta}_{n-1}(y_1, ..., y_{n-1})\right].$$

Herkkyyskäyrä osoittaa, kuinka paljon estimaatti muuttuu, kun otokseen lisätään havaintoarvo y.

Esimerkki. Keskiarvon ja mediaanin herkkyyskäyrät.

Tarkastellan jälleen jakauman odotusarvon estimointia. Jos $\hat{\theta}_n = \bar{y}_n$, niin

$$\begin{aligned} SC_{n-1}(y) &= n\left[\left(\sum_{i=1}^{n-1} y_i + y\right)/n - \left(\sum_{i=1}^{n-1} y_i\right)/(n-1)\right] \\ &= \left((n-1)y - \sum_{i=1}^{n-1} y_i\right)/(n-1) \\ &= y - \bar{y}_{n-1}, \end{aligned}$$

josta nähdään, että estimaatti voi muuttua rajatta, kun y vaihtelee.

Oletetaan seuraavaksi, että estimaatti on otosmediaani ja $y_{(i)}$ on i:nnes järjestyssuure otokselle $y_1, ...y_{n-1}$. Jos n on pariton ja $n = 2k+1$, niin $\hat{\theta}_{n-1} = \frac{1}{2}[y_{(k)} + y_{(k+1)}]$ ja

$$\hat{\theta}_n = \begin{cases} y_{(k)}, & \text{kun } y < y_{(k)} \\ y, & \text{kun } y_{(k)} \leq y \leq y_{(k+1)} \\ y_{(k+1)}, & \text{kun } y > y_{(k+1)}. \end{cases}$$

Täten

$$SC_{n-1}(y) = \begin{cases} -\frac{n}{2}[y_{(k+1)} - y_{(k)}], & \text{kun } y < y_{(k)} \\ \frac{n}{2}[2y - y_{(k)} - y_{(k+1)}, & \text{kun } y_{(k)} \leq y \leq y_{(k+1)} \\ \frac{n}{2}[y_{(k+1)} - y_{(k)}], & \text{kun } y > y_{(k+1)}. \end{cases}$$

Siis riippumatta siitä, miten suuri tai pieni y on, mediaanin muutos on aina rajojen $\pm\frac{1}{2}[y_{(k+1)} - y_{(k)}]$ sisäpuolella.

M-estimaatit

Edellä havaitsimme, että otosmediaani on hyvin robusti poikkeavia havaintoja vastaan. Toisaalta se ei ole läheskään yhtä tehokas kuin otoskeskiarvo normaalijakauman tapauksessa. Yleisesti ottaen sellainen estimaattori olisi hyvä, joka olisi tehokas ja jolla olisi korkea murtumapiste. On esitetty eri-laisia periaatteita tehokkaiden robustien estimaattorien konstruoimiseksi. Seuraavaksi tutustumme M-estimaattoreihin, jotka ovat suurimman uskottavuu-den estimaattorin tyyppisiä (kirjain M tulee nimityksestä 'maximum likelihood type').

M-estimaatti saadaan minimoimalla summa $\sum_{i=1}^{n} \rho(y_i; \hat{\theta}_n)$, missä $\rho(y;\theta)$ on jokin y:n funktio parametrin arvolla θ. Suurimman uskottavauuden estimoinnin tapauksessa $\rho(y;\theta) = -\log(f(y;\theta))$, missä $f(y;\theta)$ on yksittäisen havainnon tiheysfunktio. Jos $\rho(y;\theta)$ on derivoituva θ:n suhteen, estimaatti saadaan ratkaisemalla yhtälö $\sum_{i=1}^{n} \psi(y_i; \hat{\theta}_n) = 0$, missä

$$\psi(y;\theta) = -\frac{\partial}{\partial\theta}\rho(y;\theta).$$

(Vakiintuneen käytännön mukaan lausekkeessa on – -merkki, vaikka se ei ole ratkaisun kannalta välttämätön.)

Siinä erikoistapauksessa, että θ on sijaintiparametri, funktio ρ voidaan esittää muodossa $\rho(y-\theta)$, ja M-estimaatti ratkaistaan yhtälöstä

$$\sum_{i=1} \psi(y_i - \hat{\theta}_n) = 0. \tag{9}$$

M-estimaattien ja vaikutusfunktioiden välillä on yhteys. Voidaan osoittaa (ks. viite Garthwaite et al.), että vaikutusfunktio on vakiokerrointa lukuun ottamatta yhtä suuri kuin ψ. Täten ψ voidaan valita niin, että sillä on toivotunlaisen vaikutusfunktion ominaisuudet.

Esimerkki (jatkoa).

Jos haluamme estimoida jakauman odotusarvon, valitsemalla $\rho(y;\theta) = (y-\theta)^2$ saamme pienimmän neliösumman ratkaisun $\hat{\theta}_n = \bar{y}$. Vaikutusfunktio on tässä tapauksessa $\psi(y;\theta) = -\partial\rho(y;\theta)/\partial\theta = 2(y-\theta)$.

Jos valitsemme $\rho(y;\theta) = |y-\theta|$, niin $\hat{\theta}_n$ on otosmediaani ja vaikutusfunktio on

$$\psi(y;\theta) = \begin{cases} -1, & y < \theta \\ 0, & y = \theta \\ 1, & y > \theta. \end{cases}$$

Tämä on rajoitettu, toisin kuin otoskeskiarvon vaikutusfunktio.

On ehdotettu useita muita M-estimaattoreita, jotka parantavat otoskeskiarvon robustisuutta mutta eivät menetä paljon tehokkuutta normaalisen aineiston tapauksessa. Tällainen on Huberin M-estimaattori, jolla

$$\rho(y;\theta) = \begin{cases} (y-\theta)^2/2, & |y-\theta| \le c \\ c|y-\theta| - c^2/2, & |y-\theta| > c, \end{cases}$$

missä c on jokin sopivasti valittu vakio. Tästä saadaan vaikutusfunktio

$$\psi(y;\theta) = \begin{cases} -c, & y < \theta - c \\ y - \theta, & \theta - c \leq y \leq \theta + c \\ c, & y > \theta + c. \end{cases}$$

Yleensä M-estimaatin laskenta joudutaan tekemään iteratiivisesti. Yksi vaihtoehto on lähteä liikkeelle yhtälöstä

$$\sum_{i=1}^{n} w_i(y_i - \hat{\theta}_n) = 0,$$

missä

$$w_i = \frac{\psi(y_i;\hat{\theta}_n)}{y_i - \hat{\theta}_n}. \tag{10}$$

Ratkaisu on havaintojen painotettu keskiarvo

$$\hat{\theta}_n = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}. \tag{11}$$

Mutta koska painot w_i riippuvat estimaatista $\hat{\theta}_n$, asetetaan estimaatille aluksi alkuarvo ja toistetaan yhtälöiden (10) ja (11) laskentaa, kunnes saavutetaan suppeneminen.

Ongelmana edellä esitetyllä M-estimaatin määritelmällä on se, että se ei ota huomioon muuttujan asteikkoa. Tämän korjaamiseksi menetelmää voidaan muuttaa niin, että yhtälön (9) sijasta ratkaistaan

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \hat{\theta}_n}{S_n}\right) = 0,$$

missä S_n on asteikkoparametrin robusti estimaatti. Muut yhtälöt voidaan muuntaa vastaavalla tavalla.

L- ja R-estimaattorit

On ehdotettu myös monia muita tapoja muodostaa robusteja estimaattoreita. L-estimaattorit ovat järjestyssuureiden lineaarikombinaatioita. Esimerkiksi otosmediaani ja typistetty keskiarvo kuuluvat tähän luokkaan. R-estimaattorit on johdettu järjestystesteistä, esim. aiemmin mainittu Hodges-Lehman-estimaattori kuuluu tähän luokkaan. Muita estimaattoreita ovat A-, D- ,P- ,S- ja W-estimaattorit.

Muunnellut suurimman uskottavuuden estimaattorit

Tiku, Tan ja Balakrishnan (*Robust inference*, Marcel Dekker, New York, 1986) ovat kuvailleet toisenlaisen lähestymistavan robustiin estimointiin. Ideana on poistaa suurimmat ja pienimmät havainnot samoin kuin typistetyssä keskiarvossa. Kuitenkin näiden havaintojen heittäminen kokonaan pois olisi informaation hukkaamista. Sen sijaan näitä havaintoja voidaan käsitellä sensuroituina. Merkitään jälleen järjestyssuureita $y_{(1)}, ..., y_{(n)}$. Jos sensuroimme r_1 ensimmästä ja r_2 viimeistä havaintoa, uskottavuusfunktio voidaan esittää muodossa

$$\frac{n!}{r_1!r_2!}\left(F\left[\frac{y_{(r_1+1)}-\theta}{\sigma}\right]\right)^{r_1}\left(1-F\left[\frac{y_{(n-r_2)}-\theta}{\sigma}\right]\right)^{r_2}$$
$$\times\ \sigma^{-(n-r_1-r_2)}\prod_{i=r_1+1}^{n-r_2} f\left(\frac{y_{(i)}-\theta}{\sigma}\right),$$

missä $f((y-\theta)/\sigma)/\sigma$ on yksittäisen havainnon tiheysfunktio. Maksimoimalla tämä uskottavuusfunktio saadaan estimoitua sijaintiparametri θ ja asteikkoparametri σ. Kuitenkin monilla jakaumilla estimaattoreiden ratkaiseminen on vaikeaa, ja Tiku et al. (1986) ovat kehittäneet likimääräisen menetelmän, jota sanotaan muunnelluksi suurimman uskottavuuden estimoinniksi (modified maximum likelihood, MML). Approksimaatio on hyvä, vaikka otoskoko olisi vain 10.

Hypoteesien testaus ja luottamusvälit

Useimmat robustit estimoinnin ideat voidaan siirtää hypoteesien testaamiseen. Esimerkiksi testattessa jakauman sijaintia tavallinen t-testisuure $T = \bar{X}/(S/\sqrt{n})$ voidaan muuntaa testisuureeksi $\hat{\mu}/(\hat{\sigma}/\sqrt{n})$, missä $\hat{\mu}$ ja $\hat{\sigma}$ ovat MML-estimaattorit. Tiku et al (1986) ovat osoittaneet, että jos havaintojen jakauma on normaalinen, tämän muunnellun testisuureen jakauma nollahypoteesin vallitessa on asymptoottisesti t-jakauma $n-2r$ vapausasteella, missä r on kummaltakin hännältä sensuroitujen havaintojen lukumäärä. Vastaavasti tason $1-\alpha$ luottamusvälin rajoiksi voidaan asettaa $\hat{\mu} \pm t_{(n-2r);\alpha/2}\hat{\sigma}/\sqrt{n}$.

3 Laskentaintensiiviset menetelmät

3.1 Permutaatio- ja satunnaistamistestit

Permutaatio- ja satunnaistamistestejä voidaan toteuttaa hyödyntämällä tietokoneiden laskentakapasiteettia. Permutaatiotesteissä muodostetaan permutaatioiden joukko alkuperäisestä otoksesta niin, että tunnetaan eri permutaatioiden todennäköisyydet nollahypoteesin vallitessa. Yleensä permutaatiot ovat yhtä todennäköisiä. Testisuureen permutaatiojakauma nollahypoteesin vallitessa saadaan laskemalla testisuure kustakin permutaatiosta. Joskus permutaatioiden lukumäärä on niin suuri, ettei ole mahdollista edes tietokoneella käydä läpi kaikkia permutaatioita. Tällöin voidaan tehdä *satunnaistamistesti* poimimalla satunnaisotos permutaatioiden joukosta ja laskea testisuure ainoastaan tämän otoksen permutaatioille.

Termillä 'satunnaistamistesti' viitataan toisinaan myös koeasetelman satunnaistamiseen. Tällöin eri käsittelyt arvotaan satunnaisesti koeyksiköille. Testissä muodostettavat permutaatiot vastaavat niitä tuloksia, jotka olisivat voineet syntyä koeasetelmaa satunnaistettaessa ja nollahypoteesin ollessa tosi.

Esim. 2-suuntainen varianssianalyysi

Tarkastellaan kahden faktorin, A ja B, koetta, jossa A:lla on kolme ja B:llä 4 tasoa. Mahdollisia faktorien tasojen yhdistelmiä on tällöin 12, ja nämä on jaettu satunnaisesti 24 koeyksikölle niin, että kullakin yhdistelmällä on kaksi yksikköä. Havainnot on annettu taulukossa 4.

Taulukko 4: Aineisto 2 faktorin kokeesta

	B_1	B_2	B_3	B_4	Yhteensä
A_1	5 8	11 9	5 0	12 16	66
A_2	0 4	5 10	0 6	16 15	56
A_3	0 8	12 8	5 10	22 26	91
Yhteensä	25	55	26	107	213

Olettakaamme, että haluamme testata hypoteesia, ettei faktorilla A ole vaikutusta vastemuuttujaan. Tällöin koeyksikköjen permutointi ei vaikuta niiden vasteeseen, jos vain faktorin B taso pysyy permutoinnissa samana

kullekin koeyksikölle. Toisin sanoen tarkastelemme sellaisia koeyksiköiden permutointeja, jotka olisivat voineet toteutua koeasetelmassa ja jotka eivät muuta faktorin B tasoja. Siis ne kuusi yksikköä, jotka saivat käsittelyn B_1, permutoidaan keskenään, samoin ne kuusi, jotka saivat käsittelyn B_2, keskenään, jne.

Sellaisia permutaatioita tasolla B_1, joissa kullekin A:n tasolle tulee 2 yksikköä, on 6!/(2!2!2!)=90. Täten permutaatioita, joissa B:n taso ei muutu, on kaikestaan $90^4 = 6561 \times 10^4$, eli niitä on liikaa, jotta ne kannattaisi kaikki käydä läpi. Sen sijaan teemme satunnaistamistestin, jossa poimimme satunnaisesti 5000 permutaatiota.

Tarvitsemme testisuureen, joka on herkkä faktorin A tasojen muutoksille. Voimme käyttää esimerkiksi tavanomaista 2-suuntaisen varianssianalyysin F-testisuuretta, jota käytetään rivifaktorin omavaikutuksen testaamiseen, $F = MS_A/MS_E$, missä $MS_A = SS_A/(I-1)$ ja $MS_E = SS_E/[(IJ(K-1)]$; I on rivien, J sarakkeiden ja K toistojen lukumäärä ja $N = IJK$. Neliösummat voidaan laskea kaavoilla

$$\begin{aligned}
SS_A &= \sum_i y_{i..}^2/(JK) - y_{...}^2/N, \\
SS_B &= \sum_j y_{.j.}^2/(IK) - y_{...}^2/N, \\
SS_{AB} &= \sum_i \sum_j y_{ij.}^2/K - y_{...}^2/N - SS_A - SS_B \\
SS_E &= \sum_i \sum_j \sum_k y_{ijk}^2 - \sum_i \sum_j y_{ij.}^2/K,
\end{aligned}$$

missä pisteet alaindekseissä tarkoittavat summausta kyseisen indeksin yli.

Aineistosta voimme laskea, että

$$MS_A = \{\frac{1}{8}(66^2 + 56^2 + 91^2) - \frac{1}{24}213^2\}/2 = 40.63$$

ja

$$MS_E = \{5^2 + 8^2 + 0^2 + ... + 26^2 - (13^2 + 4^2 + ... + 48^2)/2\}/12 = 10.54,$$

joten F=40.63/10.54=3.854.

Jos rivifaktorilla A ei ole omavaikutusta ja havainnot ovat normaalijakautuneita, testisuure noudattaa jakaumaa $F \sim F(I-1, IJ(K-1)$. Tätä vastaava p-arvo on $\Pr(F(2,12) > 3.854) = 0.051$.

Permutaatiojakauman avulla määritetyt merkittävyystasot on päteviä, vaikka normaalisuutta ei voitaisikaan olettaa. Simuloitaessa permutaatiojakaumaa 273 permutaatiosta laskettu F ylitti 3.854, jota vastaa p-arvo 273/5000= 0.0546. Huomaa, että F-testisuureen tilalla voitaisiin käyttää pelkästään MS_A:ta, sillä permutaatiojakaumaa käytettäessä testisuuretta ei ole tarpeen suhteuttaa solujen sisäiseen varianssiin.

3.2 Linkkuveitsimenetelmä

Linkkuveitsi- (jackknife) ja bootstrap-menetelmät ovat otoksen uudelleen käyttöön perustuvia menetelmiä (resampling methods) samoin kuin permutaatiotestit. Permutaatiotesteissä muodostetaan alkuperäisen otoksen permutaatiota ja bootstrap-menetelmässä ns. bootstrap-otoksia. Linkkuveitsimenetelmässä estimoitavalle parametrille θ muodostetaan ns. valearvot alkuperäisen otoksen perusteella. Bootstrap-otoksia ja valearvoja käytetään sitten parametria koskevien päätelmien tekoon alkuperäisen otoksen sijasta. Sekä linkkuveitsi- että bootstrap-menetelmää voidaan käyttää estimaattorien harhan poistamiseen, estimaattorien varianssien arvioimiseen ja luottamusvälien muodostamiseen.

Olkoon $\hat{\theta}$ estimaattori parametrille θ, joka perustuu alkuperäiseen otokseen $Y_1, Y_2, ..., Y_n$. Linkkuveitsimenetelmässä muodostetaan estimaatit $\hat{\theta}_{-i}$, $i = 1, .., n$ poistamalla otoksesta vuorotellen i:s havainto Y_i ja laskemalla estimaattori θ:lle loppujen $n-1$ havainnon perusteella. Määritellään *valearvot (pseudo values)*

$$\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}, \quad i = 1, 2, ..., n \tag{12}$$

ja niiden avulla *linkkuveitsiestimaattori*

$$\tilde{\theta}_J = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i. \tag{13}$$

Selvästikin $\tilde{\theta}_J$ on harhaton, jos $\hat{\theta}$ on harhaton. Osoitamme seuraavaksi, että linkkuveitsiestimaattori vähentää alkuperäisen estimaattorin harhaa.

Oletamme, että $\hat{\theta}$:n odotusarvo voidaan esittää muodossa

$$\mathrm{E}[\hat{\theta}] = \theta + \sum_{i=1}^{\infty} c_i/n^i, \tag{14}$$

missä c_i:t voivat olla θ:n mutta eivät n:n funktioita. Tällöin $\mathrm{E}[\hat{\theta}_{-i}] = \theta + \sum_{i=1}^{\infty} c_i/(n-1)^i$, joten

$$\begin{aligned}
\mathrm{E}[\tilde{\theta}_i] &= n\left[\theta + c_1/n + \sum_{i=2}^{\infty} c_i/n^i\right] - (n-1)\left[\theta + c_1/(n-1) + \sum_{i=2}^{\infty} c_i/(n-i)^i\right] \\
&= \theta + \sum_{i=2}^{\infty}(c_i/n^{i-1} - c_i/(n-1)^{i-1}).
\end{aligned}$$

Siis $\mathrm{E}[\tilde{\theta}_i] = \theta + O(n^{-2})$, josta seuraa, että $\mathrm{E}[\tilde{\theta}_J] = \theta + O(n^{-2})$.

Siis jos $\hat{\theta}$:n harha on muotoa

$$\mathrm{E}[\hat{\theta}] - \theta = c_1/n + O(1/n^2),$$

linkkuveitsi poistaa suuruusluokkaa $1/n$ olevan termin. Jos $\mathrm{E}(\hat{\theta}) = \theta + c_1/n$, linkkuveitsi poistaa harhan kokonaan.

Esim. Varianssin estimointi

Halutaan estimoida $\theta = \mathrm{var}(Y)$ ja käytetään estimaattoria

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

joka on suurimman uskottavuuden estimaattori havaintojen ollessa normaalisia. Koska otosvarianssi on harhaton θ:lle, $\mathrm{E}[\sum(Y_i - \bar{Y})^2/(n-1)] = \theta$, saamme, että $\mathrm{E}[\hat{\theta}] = \theta - \theta/n$. Koska harha on muotoa θ/n, linkkuveitsiestimaattori on harhaton. Tämä voidaan todeta helposti suoraankin, sillä

$$\begin{aligned}
\mathrm{E}[\tilde{\theta}_i] &= n\mathrm{E}(\hat{\theta}) - (n-1)\mathrm{E}(\hat{\theta}_{-i}) \\
&= n[\theta - \theta/n] - (n-1)[\theta - \theta/(n-1)] = \theta.
\end{aligned}$$

Siis $\tilde{\theta}_i$ on harhaton, joten $\tilde{\theta}_J = \sum \tilde{\theta}_i/n$ on harhaton.

Linkkuveitsimenetelmää voidaan soveltaa myös väliestimoinnissa. Tällöin ideana on käsitellä valearvoja $\tilde{\theta}_1, \tilde{\theta}_2, ..., \tilde{\theta}_n$ aivan kuin ne olisivat riippumattomia havaintoja jakaumasta, jonka odotusarvo on θ. Tällöin voidaan olettaa, että testisuure

$$\frac{\tilde{\theta}_J - \theta}{\sqrt{\hat{\sigma}^2/n}}, \tag{15}$$

missä $\hat{\sigma}^2 = \sum_i(\tilde{\theta}_i - \tilde{\theta}_J)^2/(n-1)$ on pseudoarvojen estimoitu varianssi, noudattaa likimain t-jakaumaa vapausastein $n-1$. Tästä saadaan θ:lle likimääräinen $100(1-\alpha)\%$ luottamusväli $\tilde{\theta}_J \pm t_{(n-1);\alpha/2}\sqrt{\hat{\sigma}^2/n}$.

Yksinkertainen esimerkki, jossa valearvot ovat riippumattomia, saadaan, kun estimoidaan jakauman odotusarvoa tavallisella keskiarvolla. Tällöin valearvot ovat suoraan tavallisia havaintoja. Yleensä valearvot eivät ole täysin riippumattomia. On kuitenkin osoitettu (ks. viitteitä Garthwaitesta) että monissa tapauksissa t-jakauman avulla muodostettu luottamusväli on melko tarkka, vaikka poikkeama valearvojen riippumattomuudesta olisi huomattava. Yleissääntönä voidaan sanoa, että linkkuveitsimenetelmää voidaan käyttää luottamusväliestimoinnissa ja estimaattorin varianssin arvioinnissa, jos estimaattori perustuu otosmomentteihin. Jos se perustuu otoskvantiileihin, menetelmä ei toimi.

Esim. Lineaarinen regressio

Taulukossa 5 on esitetty 10 havaintoparin (x_i, y_i) aineisto, jonka avulla voidaan havainnollistaa linkkuveitsimenetelmää. Estimoitaessa lineaarinen regressiomalli $y = \alpha + \beta x + \varepsilon$ tavallisella pienimmän neliösumman menetelmällä saadaan estimaateiksi $\hat{\alpha} = 39.156$ ja $\hat{\beta} = 2.434$. Kullakin taulukon rivillä i on esitetty estimaatit $\hat{\alpha}_{-i}$ ja $\hat{\beta}_{-i}$, jotka on saatu jättämällä aineistosta pois havaintopari (x_i, y_i). Näiden avulla voidaan laskea valearvot $\tilde{\beta}_1 = 10 \cdot 2.434 - 9 \cdot 2.564 = 1.264, \tilde{\beta}_2 = 10 \cdot 2.434 - 9 \cdot 2.190 = 4.630, ..., \tilde{\beta}_{10} = 10 \cdot 2.434 - 9 \cdot 2.544 = 1.444$. Näistä saadaan linkkuveitsiestimaatti β:lle $\tilde{\beta}_J = (1.264 + 4.630 + ... + 1.444)/10 = 2.419$. Vastaavasti voidaan laskea $\tilde{\alpha}_J = 39.7$.

Määritämme seuraavaksi luottamusvälin β:lle. Nyt valearvojen varianssi on $\hat{\sigma}^2 = \{(1.264 - 2.419)^2 + (4.630 - 2.419)^2 + ... + (1.444 - 2.419)^2\}/9 = 1.296$. Koska $t_{9;0.025} = 2.262$, likimääräinen 95 % luottamusväli β:lle on $2.419 \pm 2.262\sqrt{1.296/10} = 2.419 \pm 0.814$. (Tavanomainen luottamusväli, joka perustuu oletukselle virheiden normaalijakautuneisuudesta, on 2.434 ± 0.766.)

3.3 Bootstrap-menetelmä

Bootstrap-menetelmä perustuu siihen, että approksimoidaan havaintojen jakaumaa empiirisellä jakaumalla ja simuloidaan tätä empiiristä jakaumaa. Olkoon $Y_1, ..., Y_n$ satunnaisotos jakaumasta, jonka kertymäfunktio on $F(y)$ ja olkoon $\hat{\theta}$ tähän otokseen perustuva estimaattori jakauman parametrille θ. Otoksen empiirinen kertymäfunktio antaa todennäköisyyden $1/n$ kullekin havaitulle arvolle $y_1, ..., y_n$;

$$\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} I_{\{y|y_i \leq y\}}(y). \tag{16}$$

Taulukko 5: Regressioanalyysiaineisto linkkuveitsimenetelmää ja ristiinvalidointia varten (poimittu Garthwaitesta)

i	x_i	y_i	$\hat{\alpha}_{-i}$	$\hat{\beta}_{-i}$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	22.0	96	34.7	2.564	91.1	24.3
2	23.4	88	47.8	2.190	99.0	121.1
3	24.9	105	34.9	2.550	98.4	43.2
4	28.5	111	38.3	2.453	108.2	7.8
5	29.8	107	40.1	2.419	112.2	27.3
6	31.6	113	39.2	2.443	116.4	11.7
7	34.2	132	41.6	2.318	120.9	123.5
8	36.4	122	36.1	2.556	129.2	51.2
9	37.7	135	42.1	2.320	129.6	29.3
10	39.0	131	36.2	2.544	135.4	19.7

Tästä jakaumasta poimittua n havainnon otosta, merk. $Y_1^*, ..., Y_n^*$ sanotaan *bootstrap-otokseksi*. Bootstrap-otos saadaan poimimalla n havainnon satunnaisotos palauttaen joukosta $\{y_1, ..., y_n\}$. Bootstrap-otoksesta lasketaan estimaattori $\hat{\theta}^*$ parametrille θ samalla tavoin kuin $\hat{\theta}$ lasketaan alkuperäisestä otoksesta. Bootstrap-estimaattorin $\hat{\theta}^*$ jakaumaa, joka syntyy toistettaessa bootstrap-otantaa, kutsutaan *bootstrap-jakaumaksi*, koska se määräytyy empiirisen kertymäfunktion $\hat{F}(y)$ perusteella.

Oletetaan seuraavassa yksinkertaisuuden vuoksi, että $\hat{\theta}$ on empiirisen jakauman $\hat{F}$ parametri, joka vastaa alkuperäisen jakauman F parametria θ. Esimerkiksi, jos θ on alkuperäisen jakauman odotusarvo, $\hat{\theta}$ on empiirisen jakauman $\hat{F}$ odotusarvo, eli havaintojen $y_1, ..., y_n$ keskiarvo. Vastaavasti jos θ on alkuperäisen jakauman varianssi, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Bootstrap-jakaumaa voidaan simuloida generoimalla suuri määrä (B kpl) bootstrap-otoksia ja laskemalla näistä estimaatit $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$. Estimaattori bootstrap-jakauman odotusarvolle on

$$\bar{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

ja varianssille

$$\hat{\sigma}^2(\hat{\theta}^*) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\theta}^*)^2.$$

Yleisesti ottaen bootstrap-menetelmä perustuu siihen, että $F(y) \approx \hat{F}(y)$

ja että $\hat{\theta}$:n suhde θ:aan on samanlainen kuin $\hat{\theta}^*$:n suhde $\hat{\theta}$:aan. Tällöin parametria θ koskeva päättely voidaan perustaa bootstrap-jakauman käyttöön. Erityisesti jos voimme olettaa, että erotuksen $\hat{\theta}-\theta$ jakauma on likimain sama kuin erotuksen $\hat{\theta}^* - \hat{\theta}$ jakauma, saamme seuraavat tulokset:

Estimaattorit $\hat{\theta}$:n varianssille ja harhalle. Varianssi $\hat{\theta}$:lle on

$$\begin{aligned} \mathrm{Var}(\hat{\theta}) &= \mathrm{Var}_F[\hat{\theta} - \theta] \approx \mathrm{Var}_{\hat{F}}[\hat{\theta}^* - \hat{\theta}] \\ &= \mathrm{Var}_{\hat{F}}[\hat{\theta}^*] \approx \hat{\sigma}^2(\hat{\theta}^*) \end{aligned}$$

Harha $\hat{\theta}$:lle on

$$\begin{aligned} \mathrm{E}(\hat{\theta} - \theta) &\approx \mathrm{E}_{\hat{F}}[\hat{\theta}^* - \hat{\theta}] \\ &= \mathrm{E}_{\hat{F}}(\hat{\theta}^*) - \hat{\theta} \approx \bar{\theta}^* - \hat{\theta}, \end{aligned}$$

josta saadaan θ:lle harhakorjattu estimaattori

$$\hat{\theta} - (\bar{\theta}^* - \hat{\theta}) = 2\hat{\theta} - \bar{\theta}^*. \tag{17}$$

Luottamusväli. Merkitään θ^*_α:lla tason α kvantiilia bootstrap-jakaumalle, ts. $\Pr_{\hat{F}}[\hat{\theta}^* \leq \theta^*_\alpha] = \alpha$. Saamme

$$\begin{aligned} \alpha &= \Pr_{\hat{F}}(\hat{\theta}^* \leq \theta^*_\alpha) \\ &= \Pr_{\hat{F}}(\hat{\theta}^* - \hat{\theta} \leq \theta^*_\alpha - \hat{\theta}) \\ &\approx \Pr(\hat{\theta} - \theta \leq \theta^*_\alpha - \hat{\theta}), \end{aligned}$$

joten

$$1 - \alpha \approx \Pr(\theta < \hat{\theta} + [\hat{\theta} - \theta^*_\alpha]).$$

Vastaavasti

$$\alpha \approx \Pr(\theta < \hat{\theta} - [\theta^*_{1-\alpha} - \hat{\theta}]).$$

Tästä saadaan tason $100 \cdot (1-2\alpha)\%$ luottamusväli $(\hat{\theta} - [\theta^*_{1-\alpha} - \hat{\theta}],\ \hat{\theta} + [\hat{\theta} - \theta^*_\alpha])$. Kutsumme sitä *perus-bootstrap-väliksi.*

Muita bootstrap-luottamusvälejä. 1) Jos voidaan olettaa, että $\hat{\theta}$ on likimain normaalijakautunut, voidaan käyttää väliä $\hat{\theta} \pm z_\alpha\ \hat{\sigma}(\hat{\theta}^*)$, missä z_α on

standardoidun normaalijakauman kvantiili $\Phi^{-1}(\alpha)$. Jos lisäksi tehdään harhakorjaus, saadaan väli

$$\hat{\theta} - (\bar{\theta}^* - \hat{\theta}) \pm z_\alpha \; \hat{\sigma}(\hat{\theta}^*)$$

2) Jos voidaan olettaa $\hat{T} = (\hat{\theta} - \theta)/\hat{\sigma}(\hat{\theta})$ on likimain samoin jakautunut kuin $\hat{T}^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*(\hat{\theta}^*)$, missä $\hat{\sigma}(\hat{\theta})$ on estimaattorin $\hat{\theta}$ estimoitu keskivirhe ja $\hat{\sigma}^*(\hat{\theta}^*)$ on bootstrap-otoksen $y_1^*, ..., y_n^*$ perusteella laskettava estimaattori $\hat{\theta}^*$:n hajonnalle, saadaan ns. *studentoitu* luottamusväli eli *bootstrap t -väli.* On havaittu, että studentoitu väli on yleensä parempi kuin aiemmin mainitut välit.

Merkitään $\hat{T}^*$:n jakauman kvantiileja T_α^* ja $T_{1-\alpha}^*$. Tällöin

$$\alpha \approx \Pr\left[(\hat{\theta} - \theta)/\hat{\sigma}(\hat{\theta}) < T_\alpha^*\right] = \Pr\left[\theta > \hat{\theta} - \hat{\sigma}(\hat{\theta})T_\alpha^*\right],$$

joten $[\hat{\theta} - \hat{\sigma}(\hat{\theta})T_{1-\alpha}^*,\ \hat{\theta} - \hat{\sigma}(\hat{\theta})T_\alpha^*]$ voidaan ottaa tason $1-2\alpha$ luottamusväliksi θ:lle.

Menetelmä vaatii estimaattoreita $\hat{\sigma}(\hat{\theta})$ ja $\hat{\sigma}^*(\hat{\theta}^*)$. Ensimmäisen voidaan laskea aiemmin kuvatulla tavalla ($\hat{\sigma}(\hat{\theta}) = \hat{\sigma}(\hat{\theta}^*)$) ja jälkimmäinen voidaan laskea kullekin bootstrap-otokselle erikseen linkkuveitsimetelmällä tai deltamenetelmällä tai bootstrappaamalla bootstrap-otosta (2. kertaluvun bootstrap), joka on erittäin laskentaintensiivinen menetelmä.

3) Edellisten lisäksi käytetään *perusprosenttipisteväliä* (θ_α^*, $\theta_{1-\alpha}^*$), mutta sen vaatimat oletukset eivät yleensä ole voimassa. Menetelmä ei yleensä tuota kovinkaan hyviä luottamusvälejä. On myös kehitetty paljon muita bootstrap-menetelmiä; ks. Davison, Hinkley (1997): *Bootstrap Methods and their Applications,* Cambridge University Press.

4 Johdatus Markovin ketjuihin

4.1 Markovin prosessit

Tähän mennessä olemme käsitelleet havaintoja, jotka ovat toisistaan riippumattomia. Ehkä yksinkertaisimpia satunnaisprosesseja, joissa peräkkäisillä havainnoilla on riippuvuusrakenne, ovat *Markovin ketjut.*

Tarkastelemme seuraavaksi yleisesti satunnaisprosessia $\{X_t\}$, jonka havainnot kuuluvat joukkoon $\mathcal{S} = \{1, 2, ..., S\}$, missä S voi olla ääretön. Arvoja

$1, ..., S$ kutsutaan prosessin tiloiksi ja joukkoa $\mathcal{S}$ on prosessin *tila-avaruudeksi.* Stokastisen prosessin indeksiä t kutsutaan ajaksi, vaikka kaikissa sovelluksissa se ei tarkoita aikaa.

Jos prosessi havaitaan ajanhetkinä $0 < t_1 < t_2 < ... < t_k$, havaintojen yhteisjakauman todennäköisyysfunktio voidaan esittää muodossa

$$\begin{aligned} & \Pr(X_0 = s_0, ..., X_{t_k} = s_k) \\ = \quad & \Pr(X_0 = s_0) \prod_{j=1}^{k} \Pr(X_{t_j} = s_j | X_{t_{j-1}} = s_{j-1}, ..., X_0 = s_0). \end{aligned}$$

Prosessi yksinkertaistuu huomattavasti, jos sillä on ns. *Markovin ominaisuus*

$$\Pr(X_{t_j} = s_j | X_{t_{j-1}} = s_{j-1}, ..., X_0 = s_0) = \Pr(X_{t_j} = s_j | X_{t_{j-1}} = s_{j-1}).$$

Tällöin prosessin 'tuleva' havainto X_{t_j} on riippumaton 'menneistä' havainnoista $X_{t_{j-2}}, ..., X_0$, kun 'nykyinen' havainto $X_{t_{j-1}}$ on annettu. Toisin sanoen kaikki informaatio X_t:n tulevasta kehityksestä sisältyy prosessin nykyiseen tilaan. Satunnaisprosessia, jolla on Markovin ominaisuus, sanotaan *Markovin prosessiksi.* Havaintojen yhteisjakauman todennäköisyysfunktio voidaan nyt esittää

$$\Pr(X_0 = s_0, ..., X_{t_k} = s_k) = \Pr(X_0 = s_0) \prod_{j=1}^{k} \Pr(X_{t_j} = s_j | X_{t_{j-1}} = s_{j-1}). \tag{18}$$

Markovin prosessia $\{X_t\}$ sanotaan *aikahomogeeniseksi*, jos ehdolliset todennäköisyydet $\Pr(X_{t+h} = s | X_t = r), \quad r, s \in \mathcal{S}$, eivät muutu ajan t kasvaessa ja aikaeron $h > 0$ pysyessa vakiona. Aikahomogeenisuudesta seuraa, että

$$\Pr(X_t = s | X_u = r) = \Pr(X_{t-u} = s | X_0 = r).$$

4.2 Markovin ketjut

Markovin prosessia, jolla on diskreetti (äärellinen tai numeroituva) tila-avaruus ja jonka havainnot ovat tasaisin aikavälein, kutsutaan *Markovin ketjuksi.* Tarkastelemme seuraavassa aikahomogeenisia Markovin ketjuja, jotka on havaittu aikapisteissä $t = 0, 1, .., k$ ja joilla on tila-avaruus $\mathcal{S} = \{1, 2, ..., S\}$. Ehdollisia todennäköisyyksiä siirtyä tilasta toiseen

$$p_{rs} = \Pr(X_1 = s | X_0 = r), \quad r, s \in \mathcal{S}$$

kutsutaan Markovin ketjun *siirtymätodennäköisyyksiksi*. Kun tilojen määrä S on äärellinen, voidaan muodostaa $S \times S$ *siirtymämatriisi* P, jonka elementtejä ovat siirtymätodennäköisyydet p_{rs}. Siirtymämatriisia sanotaan *stokastiseksi*, koska sen rivisummat ovat ykkösiä.

Esim. Sään ennustaminen

Oletetaan, että sateen todennäköisyys huomenna riippuu siitä, sataako tänään vai ei, mutta ei edellisten päivien sateisuuksista, kun tiedetään tämän päivän tilanne. Oletetaan myös, että sateen todennäköisyys huomenna on 0.4, kun tänään sataa, ja 0.3, kun tänään ei sada. Jos sanomme, että prosessi on tilassa 1, kun sataa, ja tilassa 2, kun ei sada, voimme mallintaa sitä kaksitilaisella Markovin ketjulla, jonka siirtymämatriisi on

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.3 & 0.7 \end{pmatrix}.$$

Esim. DNA-ketju

Perimän DNA on kaksoiskierre, jonka kummassakin kierteessä esiintyy emäksiä adeniinini (A), sytosiini (C), guaniini (G) ja tymiini (T). DNA:n rakenne määräytyy yhden ketjun perusteella, sillä A liittyy T:hen toisessa ketjussa ja C liittyy G:hen.

Yksinkertaisin tilastollinen malli olisi olettaa peräkkäiset emäkset riippumattomiksi. Tämä ei kuitenkaan vastaa todellisuutta. Taulukossa 4.2 on tarkasteltu siirtymien suhteellisia osuuksia emäksestä toiseen eräässä geenissä. Noin 80 % tapauksista emästä C seuraa A tai T, kun taas G on harvinainen. Sen sijaan G esiintyy paljon useammin A, G ja T:n jälkeen. Täten emästen jono ei vaikuta puhtaasti satunnaiselta.

DNA-ketjua voitaisiin yrittää mallintaa Markovin prosessilla, jonka tila-avaruus on $\mathcal{T} = \{A, C, G, T\}$. Siirtymätodennäköisyyksien estimaatteina voitaisiin käyttää taulukossa esiintyviä suhdelukuja. Tässä tapauksessa ei ole kysymys ajassa peräkkäisistä havainnoista, kuten Markovin ketjuissa yleensä.

Taulukko 6: Siirtymien suhteelliset osuudet emäksestä toiseen preproglucagon-geenissä. Tarkasteltavassa geenin osassa on 1572 emästä. (Avery, Henderson: Fitting Markov chain models to discrete state series such as DNA sequences. *Applied Statistics*, 1999, **48**, 53-61)

	Toisen emäksen suht. osuus				
Ensimmäinen emäs	A	C	G	T	Yhteensä
A	0.359	0.143	0.167	0.331	1.0
C	0.384	0.156	0.023	0.437	1.0
G	0.305	0.199	0.150	0.345	1.0
T	0.284	0.182	0.177	0.357	1.0
Keskimäärin	0.328	0.167	0.144	0.360	1.0

Chapmanin-Kolmogorovin yhtälöt

Merkitään n askeleen siirtymistodennäköisyyttä $p_{rs}^{(n)}$, siis

$$p_{rs}^{(n)} = \Pr(X_n = s | X_0 = r),$$

ja n askelen siirtymätodennäköisyyksien matriisia $P^{(n)}$. On helppo nähdä, että siirtymätodennäköisyyksille ovat voimassa ns. Chapmanin-Kolmogorovin yhtälöt

$$p_{rs}^{(n+m)} = \sum_{k=1}^{S} p_{rk}^{(n)} p_{ks}^{(m)}, \text{ missä } n, m \geq 1,$$

jotka voidaan esittää matriisimuodossa $P^{(n+m)} = P^{(n)} P^{(m)}$. On helppo induktiivisesti päätellä, että $P^{(n)} = P^n$.

Esim. Säätilan ennustaminen (jatkoa)

Jos halutaan ennustaa, sataako 3 päivän kuluttua, voimme laskea 3 askeleen siirtymämatriisin

$$P^3 = \begin{pmatrix} 0.4 & 0.6 \\ 0.3 & 0.7 \end{pmatrix}^3 = \begin{pmatrix} 0.334 & 0.666 \\ 0.333 & 0.667 \end{pmatrix}.$$

Siis jos tänään sataa, kolmen päivän päästä sataa todennäköisyydellä 0.334. Huomaa, että ennustetodennäköisyyteen ei juurikaan vaikuta se, sataako tänään vai ei.

4.3 Markovin ketjujen luokittelu

Markovin ketjujen tiloja voidaan luokitella monella tavoin. Tila s on *palautuva* (recurrent), jos

$$\Pr(X_t = s \text{ jollakin t>0 } |X_0 = s) = 1,$$

mikä tarkoittaa sitä, että ketju palaa varmasti tilaan s lähdettyään siitä. Tila s on *väistyvä* (transient), jos se ei ole palautuva eli todennäköisyys palata tilaan on aidosti pienempi kuin yksi.

Määritellään, että $T_{rs} = \min\{t > 0 : X_t = s|X_0 = r\}$ on *siirtymäaika* (first-passage time) tilasta r tilaan s ja $\mathrm{E}(T_{ss})$ *keskimääräinen paluuaika* (mean recurrence time) tilaan s. Jos tila s on väistyvä, voidaan osoittaa, että $\mathrm{E}(T_{ss}) = \infty$. Jos $\mathrm{E}(T_{ss}) < \infty$ palautuvalle tilalle s, sitä sanotaan *positiivisesti palautuvaksi*, ja jos $\mathrm{E}(T_{ss}) = \infty$, sitä sanotaan *nollapalautuvaksi*. Voidaan osoittaa, että Markovin ketjulla, jonka tilojen määrä on äärellinen ($S < \infty$), on ainakin yksi palautuva tila ja että kaikki palautuvat tilat ovat positiivisesti palautuvia.

Tilan s *jakso* (period) on $d = \mathrm{syt}\{n : p_{ss}^{(n)} > 0\}$, suurin yhteinen tekijä niille ajoille, joilla paluu tilaan s on mahdollista. Tila s on *jaksoton* (aperiodic), jos $d = 1$, ja *jaksollinen* (periodic) muuten.

Markovin ketjun tilojen jakaminen toisensa poissulkeviin luokkiin

Sanotaan, että tilasta r *on pääsy* tilaan s (s is accessible from r), merk. $r \to s$, jos $p_{rs}^{(n)} > 0$ jollakin $n \geq 0$. Jos r:stä on pääsy s:ään ja s:stä on pääsy r:ään, sanotaan, että tilat r ja s *kommunikoivat*, merk. $r \leftrightarrow s$. Jokainen tila kommunikoi itsensä kanssa, sillä $p_{rr}^{(0)} = \Pr(X_0 = r|X_0 = r) = 1$. Kommunikointi on ekvivalenssirelaatio, sillä sillä on seuraavat kolme ominaisuutta:

1. $r \leftrightarrow r$

2. Jos $r \leftrightarrow s$, niin $s \leftrightarrow r$.

3. Jos tila $r \leftrightarrow k$ ja $k \leftrightarrow s$, niin $r \leftrightarrow s$.

Ekvivalenssirelaatio jakaa tila-avaruuden toisensa poissulkeviin luokkiin. Eri luokkien tilat eivät kommunikoi keskenään. Markovin ketjua sanotaan *pelkistymättömäksi* (irreducible), jos siinä on vain yksi luokka.

Esimerkki pelkistymättömästä Markovin ketjusta

Markovin ketju, jolla on siirtymätodennäköisyyksien matriisi

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

on pelkistymätön.

Esimerkki pelkistyvästä Markovin ketjusta

Markovin ketjulla, jonka siirtymätodennäköisyyksien matriisi on

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

on luokat $\{1, 2\}$, $\{3\}$ ja $\{4\}$.

Voidaan osoittaa, että edellä esitetyt tilojen ominaisuudet, kuten palautuvuus, nollapalautuvuus ja jaksollisuus, ovat luokkaominaisuuksia. Tämä tarkoittaa sitä, että jos Markovin ketjun tilalla on jokin ominaisuus, myös sen kanssa samaan luokkaan kuuluvilla tiloilla on sama ominaisuus. Esimerkiksi jos jokin tila on palautuva, niin sen kanssa kommunikoituva tila on myös palautuva. Voidaan myös osoittaa, että kommunikoivilla tiloilla on sama jakso. Luokka voidaan sanoa palautuvaksi, väistyväksi jne. sen tilojen ominaisuuksien perusteella.

Luokkaa, josta ei ole poispääsyä, sanotaan *suljetuksi*. On helppo päätellä, että palautuva luokka on suljettu. Kääntäen voidaan osoittaa, että suljettu luokka, jossa on äärellinen määrä tiloja, on positiivisesti palautuva. Yhden tilan muodostamaa suljettua luokkaa sanotaan *absorptiotilaksi*. Edellisessä esimerkissä luokka $\{1, 2\}$ on palautuva, $\{3\}$ on väistyvä ja $\{4\}$ on absorptiotila.

Markovin ketjua kutsutaan jaksottomaksi, positiivisesti palautuvaksi ja niin edelleen, jos sen kaikilla tiloilla on sama ominaisuus. Jaksotonta, pelkistymätöntä ja positiivisesti palautuvaa ketjua sanotaan *ergodiseksi*.

4.4 Stationaarinen jakauma

Voidaan osoittaa, että ergodisella Markovin ketjulla on olemassa raja-arvot $\lim_{n\to\infty} P_{rs}^{(n)}, \quad s \in \mathcal{S}$ ja ne eivät riipu r:stä. Lisäksi raja-arvot

$$\pi_s = \lim_{n\to\infty} p_{rs}^{(n)}, \quad s \in \mathcal{S}$$

ovat yksikäsitteinen ei-negatiivinen ratkaisu yhtälöille

$$\pi_s = \sum_{r=1}^{S} \pi_r p_{rs}, \quad s \in \mathcal{S}; \quad \sum_{s=1}^{S} \pi_s = 1.$$

Jakaumaa $\{\pi_s, s \in \mathcal{S}\}$ sanotaan *stationaariseksi*, sillä jos

$$P(X_0 = s) = \pi_s, \quad s \in \mathcal{S},$$

niin

$$P(X_t = s) = \pi_s \quad \text{kaikilla } s \in \mathcal{S} \text{ ja } t \geq 0,$$

eli eri tilojen jakauma pysyy samana koko prosessin aikana.

Koska prosessi palaa tilaan s keskimäärin joka $\mathrm{E}(T_{ss})$:nnellä askelella, voidaan päätellä, että $\pi_s = 1/\mathrm{E}(T_{ss})$. Tarkempi perustelu tälle väitteelle edellyttää perehtymistä uusiutumisprosessien (renewal process) teoriaan, ks. esim. Ross: Stochastic Processes.

Esim. Säätilan rajatodennäköisyydet (jatkoa)

Säätilan stationaariset todennäköisyydet voidaan ratkaista yhtälöistä

$$\begin{cases} \pi_1 &= 0.4\pi_1 + 0.3\pi_2 \\ \pi_2 &= 0.6\pi_1 + 0.7\pi_2 \\ \pi_1 + \pi_2 &= 1 \end{cases}$$

Jompi kumpi kahdesta ensimmäisestä yhtälöistä on tarpeeton, sillä yhtälöitä on kolme ja tuntemattomia vain kaksi. Ratkaisuksi saadaan $\pi_1 = 1/3$ ja $\pi_2 = 2/3$. Siis keskimäärin joka kolmantena päivänä sataa tämän mallin mukaan.

Koska keskimääräiset paluuajat ovat $\mathrm{E}(T_{11}) = 1/\pi_1 = 3$ ja $\mathrm{E}(T_{22}) = 1/\pi_2 = 3/2$, sateisen päivän jälkeen saa odottaa keskimäärin 3 päivää, että tulee seuraava sateinen päivä, ja kuivan päivän jälkeen 1,5 päivää, että tulee seuraava kuiva päivä.

Tietyn kuvion esiintyminen rahanheitossa

Tarkastellaan riippumattomia rahan heittoja, missä kruunan (K) todennäköisyys on p ja klaavan (L) todennäköisyys $q = 1 - p$. Mikä on odotettu heittojen määrä, että esiintyy kuvio KLKL?

Heitot voidaan mallintaa Markovin ketjulla, kun määrittelemme, että tila hetkellä t, $t \geq 4$, on neljä viimeistä heittotulosta. Tällöin tilan KLKL rajatodennäköisyys π_{KLKL} on käänteisluku keskimääräiselle ajalle kulkea tilasta KLKL tilaan KLKL. Koska heitot ovat riippumattomia,

$$\Pr(X_t = KLKL) = p^2q^2, \quad t \geq 4,$$

joten

$$\pi_{KLKL} = \lim_{t\to\infty} \Pr(X_t = KLKL) = p^2q^2.$$

Siis $1/(p^2q^2)$ on keskimääräinen aika päästä tilasta KLKL tilaan KLKL.

Koska tilan KLKL saavuttamiseksi täytyy ensin saada peräkkäiset heitot KL, odotettu aika kuviosta KL kuvioon KLKL on sama kuin kuviosta KLKL kuvioon KLKL, joten

$$\mathrm{E}(\text{aika kuvioon KLKL}) = \mathrm{E}(\text{aika kuvioon KL}) + \frac{1}{p^2q^2}.$$

Tarkastelemme seuraavaksi Markovin ketjua, jonka tila määräytyy kahden viimeisen heiton perusteella. Keskimääräinen aika kulkea tilasta KL tilaan KL on $1/\pi_{KL} = 1/(pq)$. Koska tämä aika on sama kuin keskimääräinen aika heittosarjan alusta tilaan KL,

$$\mathrm{E}(\text{aika kuvioon KLKL}) = \frac{1}{pq} + \frac{1}{p^2q^2}.$$

Ergodinen lause

Vahvan suurten lukujen lain mukaan

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g(X_t) = \mathrm{E}(g(X_t))$$

todennäköisyydellä 1, kun $\{X_t,\ t = 1, 2, ...\}$ on jono riippumattomia, samoin jakautuneita satunnaismuuttujia, ja odotusarvo $\mathrm{E}(g(X_t))$ on olemassa. Tämän lain yleistystä Markovin ketjuille kutsutaan ergodiseksi lauseeksi ja se on voimassa ergodisille Markovin ketjuille.

Ergodinen lause. Olkoon $\{X_t, t = 1, 2, ..\}$ ergodinen Markovin ketju, jonka stationaarinen jakauma on $\{\pi_s,\ s \in \mathcal{S}\}$. Tällöin jos g on sellainen funktio, että odotusarvo

$$\mathrm{E}_\pi(g(X_t)) = \sum_{s\in\mathcal{S}} \pi_s\ g(s)$$

on määritelty, niin

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g(X_t) = \mathrm{E}_\pi(g(X_t))$$

todennäköisyydellä 1.

Todistus. Olkoon $N_s(T)$ niiden kertojen määrä, joina Markovin ketju on tilassa s ajanhetkinä $1, 2, .., T$. Tällöin $N_s(T)/T \to \pi_s$, kun $T \to \infty$ todennäköisyydellä 1, josta seuraa, että

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g(X_t) &= \lim_{T\to\infty} \frac{1}{T} \sum_{s\in\mathcal{S}} g(s) N_s(T) \\
&= \sum_{s\in\mathcal{S}} g(s) \lim_{T\to\infty} \frac{N_s(T)}{T} \\
&= \sum_{s\in\mathcal{S}} g(s)\ \pi_s
\end{aligned}$$

todennäköisyydellä 1. (Aputulos $\Pr\{N_s(T)/T \to \pi_s\} = 1$ on intuitiivinen, mutta sen voi perustella käyttämällä uusiutumisprosessien vahvaa lakia, jota ei tässä esitetä; ks. Ross).

Aikakäännettävät Markovin ketjut

Ergodista Markovin ketjua, jolla on stationaariset todennäköisyydet $\pi_s = \lim_{n\to\infty} p_{rs}^{(n)}$, $\quad s \in \mathcal{S}$, sanotaan *aikakäännettäväksi* (time reversible), jos se toteuttaa ns. *yksityiskohtaisen tasapainoehdon* (detailed balance condition)

$$\pi_r p_{rs} = \pi_s p_{sr} \quad \text{kaikilla } r, s \in \mathcal{S}. \tag{19}$$

Tämän kanssa on yhtäpitävää, että

$$\Pr(X_t = r, X_{t+1} = s) = \Pr(X_t = s, X_{t+1} = r) \tag{20}$$

silloin, kun Markovin ketju on saavuttanut *vakaan tilan* (steady state) eli eri tilojen todennäköisyydet ovat yhtäsuuria vastaavien stationaaristen todennäköisyyksien kanssa.

Toisaalta, jos tarkastellaan mitä tahansa aikahomogeenista, pelkistymätöntä Markovin ketjua siirtymätodennäköisyyksillä p_{rs}; $r, s \in \mathcal{S}$ ja löydetään ei-negatiiviset luvut π_s, jotka toteuttavat ehdot

$$\pi_r p_{rs} = \pi_s p_{sr} \quad \text{kaikilla } r, s \in \mathcal{S} \text{ ja } \sum_{r \in \mathcal{S}} \pi_r = 1,$$

niin summaamalla indeksin r yli saadaan

$$\sum_{r \in \mathcal{S}} \pi_r p_{rs} = \pi_s \sum_{r \in \mathcal{S}} p_{sr} = \pi_s,$$

josta nähdään, että $\{\pi_s,\ s \in \mathcal{S}\}$ on Markovin ketjun stationaarinen jakauma. Tämä jakauma on yksikäsitteinen, koska ketju on pelkistymätön. Lisäksi tiedetään, että Markovin ketju on positiivisesti palautuva, koska väistyvällä tai nollapalautuvalla ketjulla ei ole stationaarista jakaumaa.

4.5 Uskottavuuspäättely Markovin ketjuille

Oletetaan, että aikahomogeenisesta äärellistilaisesta Markovin ketjusta on havainnot $s_0, s_1, ..., s_k$ ajanhetkiltä $0, 1, 2, ..., k$. Uskottavuusfunktio on

$$\begin{aligned}
\Pr(X_0 = s_0, ..., X_k = s_k) &= \Pr(X_0 = s_0) \prod_{t=0}^{k-1} \Pr(X_{t+1} = s_{t+1} | X_t = s_t) \\
&= \Pr(X_0 = s_0) \prod_{t=0}^{k-1} p_{s_t s_{t+1}} \\
&= \Pr(X_0 = s_0) \prod_{r=1}^{S} \prod_{s=1}^{S} p_{rs}^{n_{rs}}, \qquad (21)
\end{aligned}$$

missä n_{rs} on havaittu siirtymien lukumäärä tilasta r tilaan s. Jos jätetään ensimmäinen termi lausekkeessa (21) ottamatta huomioon, logaritmoitu uskottavuusfunktio on

$$l(p) = \sum_{r=1}^{S} \sum_{s=1}^{S} n_{rs} \log p_{rs}. \qquad (22)$$

Huomaa, että estimoinnissa tarvittava informaatio sisältyy $S \times S$-matriisiin, jonka elementteinä ovat siirtymien lukumäärät n_{rs}. Tämän matriisin r. rivi $(n_{r1}, ..., n_{rS})$ on multinomijakautunut parametrein $n_{r\cdot}$ ja $(p_{r1}, ..., p_{rS})$, missä $n_{r\cdot} = n_{r1} + ... + n_{rS}$. Mallissa on $S(S-1)$ vapaata parametria, koska rivisummat ovat ykkösiä, $\sum_{s=1}^{S} p_{rs} = 1$.

Log-uskottavuusfunktio (22) maksimoituu, kun kullakin r:n arvolla maksimoidaan summa $\sum_{s=1}^{S} n_{rs} \log p_{rs}$ rajoitteella $\sum_{s=1}^{S} p_{rs} = 1$. Tämä johtaa suurimman uskottavuuden estimaattoreihin $\hat{p}_{rs} = n_{rs}/n_{r\cdot}$. Kun $p_{rs} > 0$ kaikilla s, r, tavallinen uskottavuusestimointiin liittyvä asymptotiikka toimii ja estimaattorit ovat likimain normaalijakautuneita variansseilla ja kovariansseilla

$$\operatorname{cov}(\hat{p}_{rs}, \hat{p}_{tu}) = \begin{cases} \hat{p}_{rs}(1 - \hat{p}_{rs})/n_{r\cdot}, & r = t, s = u \\ -\hat{p}_{rs}\hat{p}_{ru}/n_{r\cdot}, & r = t, s \neq u, \\ 0, & \text{muuten.} \end{cases}$$

(Nämä varianssit ja kovarianssit perustuvat multinomijakauman ominaisuuksiin).

Tähän mennessä olemme jättäneet huomiotta ensimmäisen termin lausekkeessa (21). Tällä ei ole kovin paljon merkitystä, jos havaintoja on paljon. Jos havaintoja on vähän, $\Pr(X_0 = s_0)$ voidaan korvata stationaarisella todennäköisyydellä π_{s_0}, joka on siirtymämatriisin P elementtien funktio. Tällöin uskottavuusfunktion maksimointi on hankalampaa ja täytyy yleensä tehdä numeerisesti.

Jos yhden pitkän Markovin ketjun sijasta on käytettävissä useita lyhyitä Markovin ketjuja, joilla on samat parametrit, voidaan pyrkiä käyttämään paremmin hyväksi ensimmäisten havaintojen antama informaatio ja pyrkiä estimoimaan ensimmäisen havainnon todennäköisyysjakauma siirtymätodennäköisyyksien lisäksi. Oletamme kuitenkin jatkossa, että meillä on käytössä vain yksi pitkä ketju.

Jos peräkkäiset havainnot ovat riippumattomia, siirtymätodennäköisyydet eivät riipu nykyisestä tilasta, $p_{rs} = p_s$. Tässä yksinkertaisessa *riippumattomuusmallissa* on vapaita parametreja vain $S - 1$ ja niiden suurimman uskottavuuden estimaatit ovat $\hat{p}_s = n_{\cdot s}/n_{\cdot\cdot}$, missä $n_{\cdot s} = n_{1s} + ... + n_{Ss}$ ja $n_{\cdot\cdot} = \sum_s n_{\cdot s}$. Tavallisia Markovin ketjuja, joita on käsitelty tähän mennessä, voidaan sanoa ensimmäisen kertaluvun Markovin ketjuiksi, koska siirtymätodennäköisyydet riippuvat ainoastaan tämänhetkisestä tilasta. Jos siirtymätodennäköisyydet riippuvat n peräkkäisen tilan arvosta, voidaan puhua n. kertaluvun Markovin ketjusta. Riippumattomuusmallia voidaan sanoa nollannen kertaluvun Markovin ketjuksi.

Uskottavuussuhdetestisuure riippumattomuushypoteesin testaamiseksi on

$$W = 2 \sum_{r,s} n_{rs} \log \left(\frac{\hat{p}_{rs}}{\hat{p}_s} \right) = 2 \sum_{r,s} n_{rs} \log \left(\frac{n_{rs} n_{\cdot\cdot}}{n_{r\cdot} n_{\cdot s}} \right).$$

Taulukko 7: Havaitut frekvenssit 16 mahdolliselle peräkkäisten emästen muodostamalle parille DNA-ketjussa. Aineisto sama kuin taulukossa 6.

	Toisen emäksen frekvenssit				
Ensimmäinen emäs	A	C	G	T	Yhteensä
A	185	74	86	171	516
C	101	41	6	115	263
G	69	45	34	78	226
T	161	103	100	202	566
Yhteensä	516	263	226	566	1571

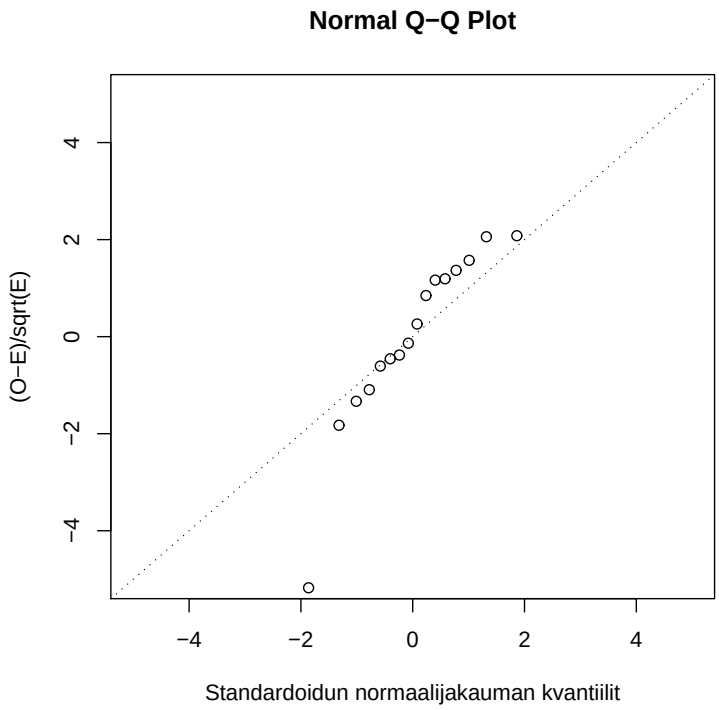
Nollahypoteesin vallitessa testisuure W noudattaa χ^2-jakaumaa vapausastein $S(S-1)-(S-1) = (S-1)^2$. Testisuure on tavanomainen uskottavuustestisuure testattaessa rivi- ja sarakemuuttujien riippumattomuutta ristiintaulukossa. W on likimain yhtä suuri kuin Pearsonin testisuure $P = \sum(O_{rs} - E_{rs})^2/E_{rs}$, missä O_{rs} on havaittu lukumäärä n_{rs} ja E_{rs} odotettu lukumäärä riippumattomuuden vallitessa, $n_{r\cdot}n_{\cdot s}/n_{\cdot\cdot}$.

Esim. DNA-aineisto (jatkoa)

Testataan peräkkäisten emästen riippumattomuutta DNA-ketjussa. Taulukosta 7 saadaan riippumattomuushypoteesia vastaavat estimaatit $\hat{p}_A = 516/1571 = 0.328$, $\hat{p}_C = 263/1571 = 0.167$ jne. Tavallista Markovin ketjua vastaavat estimaatit ovat $\hat{p}_{AA} = 185/516 = 0.359$, $\hat{p}_{AC} = 74/516 = 0.143$, $\hat{p}_{CG} = 74/516 = 0.023$ ja niin edelleen. Nollahypoteesin (riippumattomuuden) vallitessa W on noudattaa χ^2_9, jakaumaa, mutta havaittu testisuure $w = 64.45$ on tähän verrattuna hyvin suuri. Pearsonin testisuureen P arvo on 50.3.

Jos nollahypoteesi on oikea, suureiden $(O_{rs} - E_{rs})/E_{rs}^{1/2}$ pitäisi noudattaa likimain standardoitua normaalijakaumaa. Kuviossa 2 on näiden suureiden kvantiili-kvantiilikuvio. Yksi havaittu frekvenssi näyttää poikkeavan huomattavasti odotetusta frekvenssistä ($(O_{rs} - E_{rs})/E_{rs}^{1/2}$ alle -5). Se vastaa solua CG.

Testaamista voitaisiin jatkaa niin, että tutkittaisiin ensimmäisen kertaluvun Markovin ketjun riittävyyttä testaamalla sitä toisen kertaluvun Markovin ketjua vastaan. Osoittautuu, että tällä aineistolla tavallinen, 1. kertaluvun Markovin ketju on riittävä, ks. tarkemmin Davison.



Kuva 2: QQ-kuvio riippumattomuusmallin sopivuuden tutkimiseen.

5 MCMC-menetelmät

Esittelemme seuraavaksi Markovin ketjuihin pohjautuvia laskentaintensiivisiä menetelmiä, joita voidaan käyttää moniulotteisten jakaumien simuloinnissa. Erityisen hyödyllisiä MCMC-menetelmät (Monte Carlo Markov Chain) ovat bayesiläisen tilastotieteen laskennoissa. Periaatteena menetelmissä on generoida aikakäännettävä Markovin ketju, jolla on haluttu stationaarinen jakauma.

5.1 Gibbsin poimija

Gibbsin poimijaa (Gibbs sampler) voidaan käyttää simuloimaan kaksi- tai useampiulotteista jakaumaa, jonka todennäköisyysfunktio (tai tiheysfunktio) on $p(\mathbf{x})$. Gibbsin poimija voidaan toteuttaa, kun voidaan generoida satunnaislukuja täydellisisistä ehdollisista jakaumista (full conditional di-

stributions), joilla on todennäköisyysfunktiot $p_i(x_i|\mathbf{x}_{-i})$, $i = 1, ..., d$, missä $\mathbf{x}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$.

Algoritmi toimii niin, että valitaan ensin alkuarvot $\mathbf{x}^0 = (x_1^0, ..., x_d^0)$. Kun on saatu generoitua satunnaisvektorit $\mathbf{x}^1, ..., \mathbf{x}^t$, voidaan $\mathbf{x}^{t+1}$ generoida komponenteittain seuraavasti:

- Generoidaan x_1^{t+1} ehdollisesta jakaumasta $p_1(x_1|x_2^t, ..., x_d^t)$
- Generoidaan x_2^{t+1} ehdollisesta jakaumasta $p_2(x_2|x_1^{t+1}, x_3^t, ..., x_d^t)$
- Generoidaan x_3^{t+1} ehdollisesta jakaumasta $p_3(x_3|x_1^{t+1}, x_2^{t+1}, x_4^t, ..., x_d^t)$

  ...
- Generoidaan x_d^{t+1} ehdollisesta jakaumasta $p_d(x_d|x_1^{t+1}, x_2^{t+1}, ..., x_{d-1}^{t+1})$

Tuloksena saadaan yleensä ergodinen Markovin ketju, jonka stationaarisen jakauman todennäköisyysfunktio on $p(\mathbf{x})$.

Todistus. Algoritmi tuottaa Markovin ketjun, sillä selvästi satunnaisvektorin $\mathbf{x}^{t+1}$ jakauma riippuu vain $\mathbf{x}^t$:stä. Ketju on aikahomogeeninen, sillä siirtymätodennäköisyydet määräytyvät koko ajan ehdollisen jakaumien $p_j(x_j|\mathbf{x}_{-j})$ perusteella. Ketju on pelkistymätön, kun Gibbsin poimija pääsee liikkumaan jokaiseen tila-avaruuden pisteeseen. Näin on silloin, kun arvoalue, jossa todennäköisyysfunktio $p(\mathbf{x})$ saa positiivisia arvoja, on riittävän yhtenäinen.

Osoitetaan seuraavaksi, että $p(\mathbf{x})$ toteuttaa yksityiskohtainen tasapainoehdon

$$p(\mathbf{x}) \Pr(\mathbf{X}^{t+1} = \mathbf{x}^*|X^t = \mathbf{x}) = p(\mathbf{x}^*) \Pr(\mathbf{X}^{t+1} = \mathbf{x}|\mathbf{X}^t = \mathbf{x}^*),$$

missä $\mathbf{x} = (x_1, ..., x_j, ..., x_d)$ ja $\mathbf{x}^* = (x_1, ..., x_j^*, ..., x_d)$. Tässä tarkastellaan ketjua poikkeuksellisesti niin, että yksi aika-askel vastaa yhden satunnaisvektorin komponentin generointia. Saamme

$$\begin{aligned} p(\mathbf{x}) \Pr(\mathbf{X}^{t+1} = \mathbf{x}^*|\mathbf{X}^t = \mathbf{x}) &= p(\mathbf{x}) p_j(x_j^*|\mathbf{x}_{-j}) = p(\mathbf{x}) \frac{p(\mathbf{x}^*)}{p(\mathbf{x}_{-j})} \\ &= p(\mathbf{x}^*) \frac{p(\mathbf{x})}{p(\mathbf{x}_{-j})} = p(\mathbf{x}^*) p_j(x_j|\mathbf{x}_{-j}) \\ &= p(\mathbf{x}^*) \Pr(\mathbf{X}^{t+1} = \mathbf{x}|\mathbf{X}^t = \mathbf{x}^*), \end{aligned}$$

joten $p(\mathbf{x})$ on stationaarinen jakauma. Pelkistymättömyydestä seuraa, että se on yksikäsitteinen. Ketju on positiivisesti palautuva, koska väistyvillä ja nollapalautuvilla ketjuilla ei ole stationaarista jakaumaa. Se on myös jaksoton, koska on mahdollista, että uusi generoitu arvo on sama kuin vanha. Näistä ominaisuuksista seuraa, että ketju on ergodinen.

Huom. Gibbsin poimijan toimivuus osoitettiin diskreetille jakaumalle $p(\mathbf{x})$. Yleensä Gibbsin poimijaa ja muita MCMC-menetelmiä sovelletaan kuitenkin jatkuville jakaumille. Tällöin periaatteessa tarvittaisiin yleisiin tila-avaruuksiin soveltuvaa Markovin ketjujen teoriaa. Koska kuitenkin jatkuvia jakaumia voidaan approksimoida mielivaltaisen tarkasti diskreettien jakaumien avulla, tässä esitettyä teoriaa voidaan pitää riittävänä.

Jotta saamme mielekkään esimerkin, jossa voimme soveltaa Gibbsin poimijaa, esittelemme seuraavaksi bayesiläisen hierarkisen mallin.

5.2 Bayesiläinen hierarkinen malli

Hierarkiset mallit ovat hyödyllisiä silloin, kun aineistossa on eri vaihtelun tasoja. Sairauden esiintyvyys voi vaihdella alueittain ja alueiden sisällä voi olla vaihtelua esimerkiksi köyhyyden ja saastuneisuuden vuoksi. Kun alueellisia ja paikallisia esiintyvyyksiä pidetään satunnaismuuttujina, voidaan kuvitella hierarkinen tilanne, jossa sairastuneiden ihmisten lukumäärä riippuu satunnaisesta paikallisesta esiintyvyydestä, joka puolestaan riippuu satunnaisesta alueellisesta esiintyvyydestä.

Yksinkertaisimmassa tilanteessa voidaan tutkimusaineisto jakaa ryhmiin ja olettaa, että j:nnen ryhmän havainnot Y_{ij} ovat satunnaisotos jakaumasta $f(y_{ij}|\theta_j)$, missä θ_j on j. ryhmää kuvaava parametrivektori. Ryhmien välistä vaihtelua voidaan kuvata parametrivektorien θ_j:n jakaumalla; oletetaan että θ_j:t ovat riippumattomia havaintoja jakaumasta $g(\theta_j|\phi)$, jolloin niiden yhteisjakauman tiheysfunktio on

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

Jos ϕ:n priorijakauma on $\pi(\phi)$, parametrivektorin θ reunajakaumaksi saadaan

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

Parametrit θ_j eivät ole riippumattomia, kun niiden jakaumaa ei tarkastella kiinteällä ϕ:n arvolla. Sen sijaan ne ovat *vaihdettavia* (exchangeable). Vaihdettavuus on bayesiläisessä hierarkisessa mallinnuksessa tärkeä käsite. Yleisesti määritellään, että satunnaismuuttujat $U_1, U_2, ..., U_n$ ovat vaihdettavia, jos niiden tiheysfunktiolla (tai todennäköisyysfunktiolla) on ominaisuus

$$f(u_1, ..., u_n) = f(u_{\xi(1)}, ..., u_{\xi(n)})$$

mille tahansa joukon $\{1, ..., n\}$ permutaatiolle ξ.

Esim. Sydänleikkausaineisto (jatkoa)

Taulukossa 1 on sydänleikkausaineisto, jossa on esitetty vauvojen kuolleisuus 12 sairaalassa tehdyissä leikkauksissa. Näyttäisi siltä, että sairaaloiden välillä esiintyy vaihtelua kuolleisuudessa, mitä emme voineet ottaa huomioon soveltaessamme binomimallia kullekin sairaalalle erikseen tai yhdistämällä eri sairaaloiden tulokset. Mallinnamme ilmiötä seuraavassa käyttäen hierarkista bayesiläistä mallia. Oletamme, että kuolleisuudet θ_i on vaihdettavia satunnaismuuttujia, jolloin prioriodotuksen mukaan sairaalat ovat samanlaisessa asemassa. (Meillä ei ole käytössä selittäviä muuttujia, jotka voisivat selittää eroja sairaaloiden välillä).

Tarkemmin ottaen mallimme on seuraava: Kuolleiden lukumäärä y_j sairaalassa j noudattaa binomijakaumaa $\text{Bin}(\theta_j, m_j)$, $j = A, ..., L$, ja parametrit $\theta_A, ..., \theta_L$ ovat otos jakaumasta $g(\theta_j|\phi)$ ja ϕ:n priorijakauma on $\pi(\phi)$. Yksinkertainen tapa mallintaa θ:n jakaumaa on käyttää logit-linkkiä, jolloin $\beta_j = \log\{\theta_j/(1-\theta_j)\} \sim N(\mu, \sigma^2)$. Siis tässä $\phi = (\mu, \sigma^2)$. Edelleen oletamme, että $\mu \sim N(\mu_0, c^2)$ ja $\sigma^2 \sim \text{InvGamma}(a, b)$[1], missä a,b ja c ovat kiinteitä hyperparametreja, joiden avulla voidaan sisällyttää priori-informaatiota. Asetamme $a = b = 10^{-3}$, jolloin $1/\sigma^2$:n prioriodotusarvo on 1 ja -varianssi 10^3, ja $\mu_0 = 0$, $c = 10^3$. Meillä on siis aidot priorijakaumat, kuten on yleensä syytä olla hierarkisissa malleissa, mutta niiden priori-informaatio on heikko.

Parametrien uskottavuusfunktio on

$$L(\beta; y) = \prod_j \binom{m_j}{y_j} \theta_j^{y_j} (1-\theta_j)^{m_j - y_j} \propto \prod_j \frac{e^{y_j \beta_j}}{(1+e^{\beta_j})^{m_j}},$$

ja posteriorijakauman tiheysfunktio $\pi(\beta, \mu, \sigma^2|y) \propto \pi(\beta, \mu, \sigma^2) L(\beta; y)$, missä

$$\pi(\beta, \mu, \sigma^2) \quad \propto \quad \pi(\mu, \sigma^2)\ g(\beta|\mu, \sigma^2)$$

[1]InvGamma tarkoittaa käänteistä gammajakaumaa; $1/\sigma^2 \sim \text{Gamma}(a, b)$

$$\propto \quad e^{-\frac{1}{2c^2}(\mu-\mu_0)^2} \times \left(\frac{1}{\sigma^2}\right)^{a+1} e^{-b\frac{1}{\sigma^2}} \times \prod_j \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(\beta_j-\mu)^2}.$$

Gibbsin poimijaa varten (esitellään myöhemmin)tarvitaan parametrien ehdolliset posteriorijakaumat, jotka saadaan poimittua yhteisposteriorijakaumasta. Parametrille μ ehdollinen posteriorijakauma on $\mu|\sigma^2, \beta, y \sim N(\hat{\mu}, \sigma_\mu^2)$, missä

$$\hat{\mu} = \frac{\frac{1}{\sigma^2}\sum \beta_j + \frac{1}{c^2}\mu_0}{\frac{J}{\sigma^2} + \frac{1}{c^2}}, \quad \sigma_\mu^2 = \frac{1}{\frac{J}{\sigma^2} + \frac{1}{c^2}},$$

ja σ^2:n ehdollinen posteriorijakauma on käänteinen gammajakauma,

$$\sigma^2|\mu, \beta, y \sim \text{InvGamma}\left\{a + J/2, b + \frac{1}{2}\sum_j (\beta_j - \mu)^2\right\}.$$

Parametrien β_j ehdolliset posteriorit ovat vähän hankalampi tapaus, sillä ne eivät ole mitään yleisiä jakaumia:

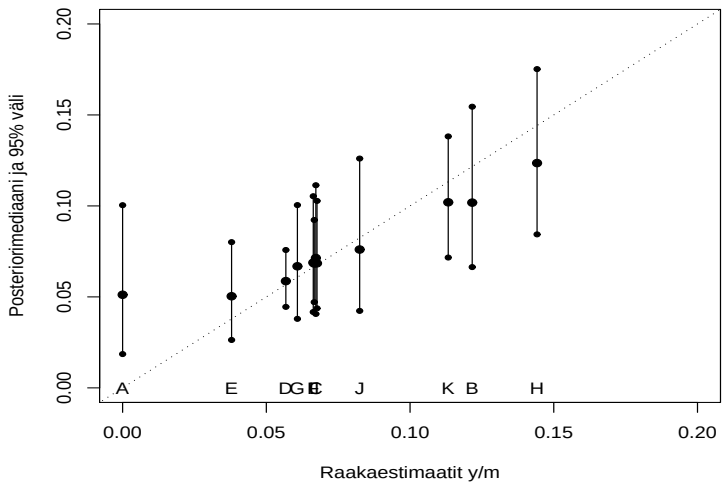
$$\pi(\beta_j|\beta_{-j}, \mu, \sigma^2, y) \propto \frac{e^{y_j\beta_j}}{(1+e^{\beta_j})^{m_j}} e^{-\frac{1}{2\sigma^2}(\beta_j-\mu)^2}. \tag{23}$$

Näistä jakaumista voidaan kuitenkin generoida satunnaislukuja esim. adaptiivista hylkäyspoimintaa käyttäen (ks. lähemmin Davison). Adaptiivinen hylkäyspoiminta toimii log-konkaaveille tiheysfunktioille. Tiheysfunktion (23) logaritmi on konkaavien (concave, kovera) funktioiden summa, joten se on konkaavi, jolloin logaritmoimaton tiheysfunktio on log-konkaavi.

Kun estimointi tehtiin Gibbsin poimijalla, se suppeni varsin nopeasti stationaariseen jakaumaan. Kuviossa 3 on 999 simulointikierrokseen perustuvat 95% posteriorivälit parametreille θ_j. Vaaka-akselilla on raaka-estimaatit y_j/m_j ja pystyakselilla posteriorivälit ja -mediaanit. Nähdään, että hierarkisessa mallissa yksittäisten sairaaloiden kuolleisuusestimaatit siirtyvät kohti koko aineiston keskiarvoa kohti.

5.3 Metropolis- ja Metropolis-Hastings-algoritmit

Metropolis-algoritmin käyttö ei edellytä kykyä generoida satunnaisvektorin komponenttien ehdollisista jakaumista. Riittää, että tunnetaan tavoitteena olevan jakauman todennäköisyysfunktio (tiheysfunktio) vakiokerrointa lukuunottamatta. Usein bayesiläisessä analyysissä on helppo määrittää



Kuva 3: Kuolleisuuslukujen posteriorimediaanit ja 95% -välit vs. raakaestimaatit y_j/m_j, kun on sovellettu hierarkista mallia 12 sairaalan sydänleikkausaineistoon

posteriorijakauman tiheysfunktio vakiokerrointa lukuunottamatta, jonka jälkeen on suoraviivaista soveltaa Metropolis-algoritmia posteriorijakauman simulointiin.

Oletetaan, että halutaan simuloida jakaumaa, jonka todennäköisyysfunktio on $p(x)$. Tässä x voi olla skalaari tai vektori. Metropolis-algoritmia varten tarvitaan hyppäysjakauma eli ehdotusjakauma (jumping distribution, proposal distribution) $J(y|x)$, jonka avulla voidaan generoida ehdotus y, kun simuloitavan ketjun nykyinen arvo on x. Metropolis-algoritmissa oletetaan hyppäysjakauman symmetrisyys $J(y|x) = J(x|y)$.

Metropolis-algoritmi toimii niin, että valitaan alkuarvo x_0. Kun on generoitu vektorit $x_0, ..., x_t$, generoidaan y hyppäysjakaumasta $J(y|x_t)$. Uusi arvo y hyväksytään todennäköisyydellä

$$\min\left(1, \frac{p(y)}{p(x_t)}\right).$$

Jos uusi arvo hyväksytään, asetetaan $x_{t+1} = y$, ja jos se hylätään, säilytetään vanha arvo, jolloin $x_{t+1} = x_t$.

Metropolis-algoritmi tuottaa Markovin ketjun, sillä uuden arvon x_{t+1}:n jakauma riippuu ainostaan nykyisestä arvosta x_t. Ketju on myös aikahomogeeninen, sillä siirtymätodennäköisyydet määräytyvät hyppäysjakauman $J(y|x)$ avulla, joka pysyy muuttumattomana simuloinnin aikana. Ketju on lisäksi pelkistyvä, jos hyppäysjakauma on sellainen, että se mahdollistaa siirtymisen jokaiseen tila-avaruuden pisteeseen. Näin yleensä on asianlaita. Jakauma $p(x)$ toteuttaa yksityiskohtaisen tasapainoehdon, sillä jos x ja x^* ovat kaksi eri tila-avaruuden pistettä ja oletetaan, että $p(x^*) \leq p(x)$, niin

$$\begin{aligned} p(x)\Pr(X_{t+1} = x^*|X_t = x) &= p(x)J(x^*|x)\frac{p(x^*)}{p(x)} = p(x^*)J(x|x^*) \\ &= p(x^*)\Pr(X_{t+1} = x|X_t = x^*). \end{aligned}$$

Siis $p(x)$ on syntyvän ketjun stationaarinen jakauma ja ketju on positiivisesti palautuva. Koska ketju on selvästi jaksoton, se on myös ergodinen.

Hyppäysjakauman valinnasta

Sanotaan, että Markovin ketju sekoittuu (is mixing) huonosti, jos peräkkäiset havainnot ovat voimakkaasti korreloituneita. Tällöin ergodinen lause toimii huonosti ja esimerkiksi simuloitujen havaintojen keskiarvo lähestyy hyvin hitaasti stationaarisen jakauman odotusarvoa. Tähän ongelmaan voi olla kaksi syytä. Jos hyppäysjakauman hajonta on hyvin pieni jollekin simuloitavan jakauman komponentille, ketju liikkuu hitaasti kyseisen komponentin osalta. Jos hajonta on taas liian suuri, ketju polkee paikallaan, koska vain hyvin harvat uudet ehdotukset hyväksytään. On tutkittu, että jos hyppäysjakauma on d-ulotteinen normaalijakauma ja tavoiteltavan stationaarisen jakauman kovarianssimatriisi on Σ, optimaalinen hyppäysjakauman kovarianssimatriisi on $c^2\Sigma$, missä $c \approx 2.4/\sqrt{d}$ (ks. Gelman: Bayesian Data Analysis).

Esim. Weibull-jakauma

Tarkastellaan kaksiparametrista Weibullin jakaumaa, jonka tiheysfunktio on

$$f(x;\beta,\delta) = \frac{\delta}{\beta^\delta}x^{\delta-1}\exp\left\{-\left(\frac{x}{\beta}\right)^\delta\right\}, \quad x,\beta,\delta > 0.$$

Kun meillä on satunnaisotos $y_1, ...y_n$ tästä jakaumasta, uskottavuusfunktio on

$$L(y) = \frac{\delta^n}{\beta^{n\delta}}\left(\prod_i y_i\right)^{\delta-1}\exp\left\{-\sum_i\left(\frac{y_i}{\beta}\right)^\delta\right\}.$$

Kun valitaan priorijakaumaksi $\pi(\beta, \delta) \propto 1/(\beta\delta)$, posteriorijakaumaksi saadaan

$$\pi(\beta, \delta|y) \propto \frac{\delta^{n-1}}{\beta^{n\delta+1}} \left(\prod_i y_i\right)^{\delta-1} \exp\left\{-\sum_i \left(\frac{y_i}{\beta}\right)^{\delta}\right\}.$$

Nyt olisi mahdollista muodostaa parametrien ehdolliset posteriorijakaumat ja simuloida posteriorijakaumaa Gibbsin poiminnalla. Ehdollisista jakaumista voitaisiin generoida satunnaislukuja käyttäen adaptiivista hylkäysalgoritmia. Yksinkertaisempaa on kuitenkin soveltaa Metropolis-algoritmia.

Kuviossa 4 on Metropolis-algoritmilla simuloitu 10000:n mittainen ketju lähtien alkuarvoista $\delta = \beta = 1$, kun havaintoaineisto on keinotekoinen ja saatu generoimalla 100 havaintoa Weibull(0.3,10)-jakaumasta. Hyppäysjakaumana on käytetty kaksiulotteista normaalijakaumaa, jonka odotusarvovektori on vanha havaintovektori ja kovarianssimatriisi diag$(0.01, 10)$. Kuvion perusteella suppeneminen stationaariseen jakaumaan tapahtuu melko nopeasti, mutta β:n ketju näyttää sekoittuvan huonosti.

Seuraavaksi on simuloitu 10000 uutta havaintoa käyttäen optimaalista kovarianssimatriisia $2.4^2\Sigma/2$, missä Σ on tavoitejakauman kovarianssimatriisi, joka on estimoitu käyttämällä alkuperäisen simulointiketjun loppupään havaintoja. Alkuarvoina on käytetty ensimmäisen simulointiketjun viimeistä havaintovektoria. Kuvan 5 perusteella sekoittuminen on nyt parempaa. Kuvassa 6 on piirretty kumulatiivisten 2.5%, 50 % ja 97.5 % kvantiilien kuvaajat.

Metropolis-Hastings-algoritmi

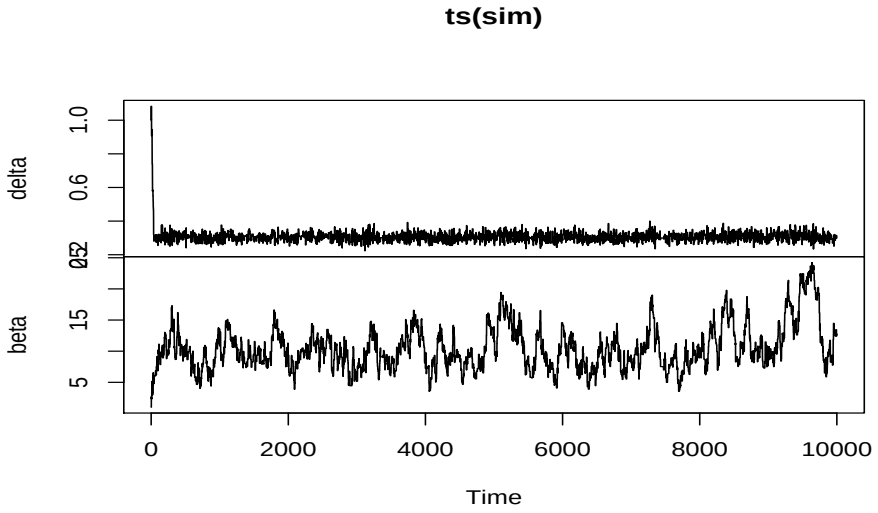
Metropolis-Hastings-algoritmi on muuten samanlainen kuin Metropolis-algoritmi, mutta siinä ei oleteta, että hyppäysjakauma $J(y|x)$ on vanhan arvon x suhteen symmetrinen. Ehdotuksen hyväksymistodennäköisyys on tällöin

$$\min\left(1, \frac{p(y)/J(y|x_t)}{p(x_t)/J(x_t|y)}\right).$$

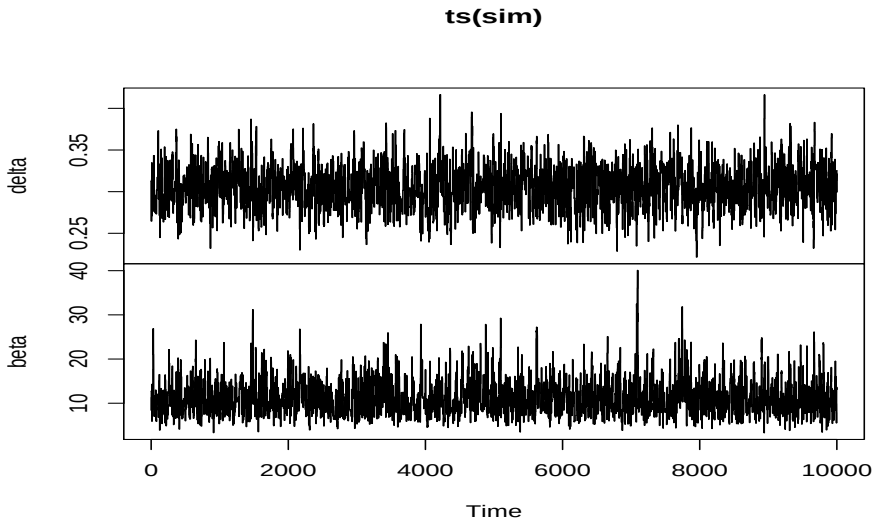
Voidaan osoittaa, että syntyy Markovin ketju, jonka stationaarinen jakauma on $p(x)$ (harj. tehtävä).

5.4 Tulosten analysointi

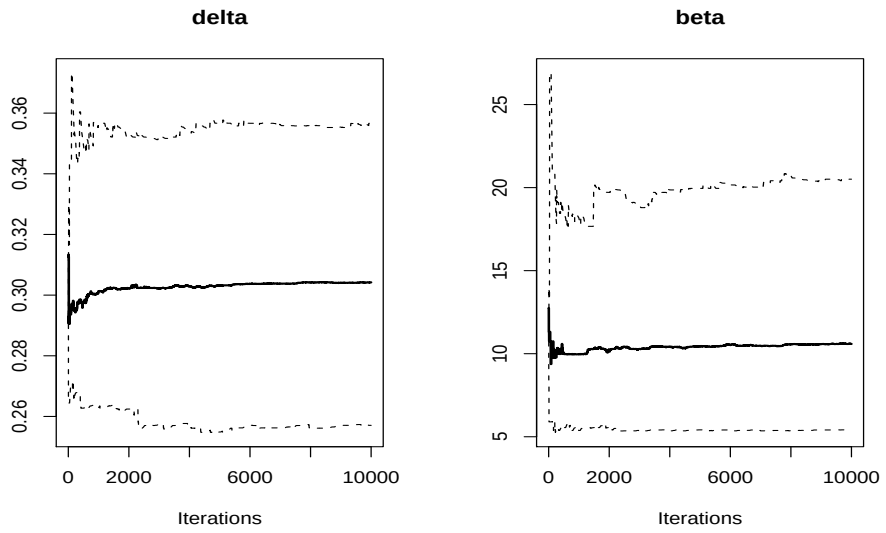
Markovin ketjun simulointia on jatkettava, kunnes on saavutettu stationaarinen jakauma, ja tämän jälkeen vielä niin pitkälle, että on saatu luotettavat estimaatit stationaarisen jakauman yhteenvetosuureille. On useita



Kuva 4: Weibull-jakauman parametrien estimointi Metropolis-algoritmilla. Parametri β näyttää sekoittuvan huonosti.



Kuva 5: Weibull-jakauman parametrien estimointi Metropolis-algoritmilla, kun hyppäysjakaumaa on optimoitu



Kuva 6: Kumulatiivinen kvantiilikuvio kvantiiliestimaattien suppenemisen tarkasteluun; piirretty 2.5%, 50% ja 97.5 % kvantiilit.

tapoja tutkia, onko ketju supennut stationaariseen jakaumaan. Ainoa varma tapa todeta suppeneminen on tutkia teoreettisesti suppenemisnopeutta. Käytännössä tämä on mahdollista vain muutamissa erikoistapauksissa.

Käytännössä suppeneminen voidaan todeta käyttäen suppenemista mittaavia testisuureita (convergence diagnostics) ja tutkimalla aikasarjakuvioita. Suppenemistestisuureet voivat perustua simuloidun ketjun alku- ja loppupään vertaamiseen (Geweken suppenemisdiagnostiikka) tai usean rinnakkaisen simuloidun ketjun vertaamiseen (Gelmanin ja Rubinin suppenemisdiagnostiikka). Tulosten analysoinnissa hyödyllisiä funktioita on R:n coda-paketissa.

On myös hyödyllistä tutkia erityyppisiä aikasarjakuvioita. Karkean kuvan suppenemistilanteesta saa tutkimalla simuloitavan prosessin eri komponenttien muodostamia aikasarjoja. Niistä voi helposti todeta, mitkä komponentit ovat ongelmallisia ja suppenevat hitaasti. Kaikkien komponenttien yhteistä suppenemista voidaan tutkia tarkastelemalla logaritmoidun uskottavuusfunktion tai posteriorijakauman tiheysfunktion saaman arvon muodostamaa aikasarjaa. Toinen hyödyllinen kuvio on tutkia mediaanin ja muiden kvantiilien kumulatiivisia aikasarjoja. Niiden avulla voidaan nähdä, miten tarkko-

ja ovat esimerkiksi simulointien perusteella muodostettavat posteriorivälit. Tutkimalla autokorrelaatiokuviota voidaan puolestaan nähdä, miten autokorreloituneita peräkkäiset simuloidut havainnot ovat. Jos autokorrelaatio on suurta, ketju sekoittuu hitaasti ja tarvitaan pitkä simuloitu ketju, jotta yhteenvetosuureet, kuten keskiarvot, antaisivat luotettavan kuvan stationaarisesta jakaumasta.

Monihuippuiset kohdejakaumat voivat olla ongelmallisia, sillä voi olla epävarmaa, että simuloitu ketju on käynyt kaikissa tärkeissä huipuissa. Tämän vuoksi voi olla hyödyllistä simuloida useita Markovin ketjuja käyttäen toisistaan huomattavasti poikkeavia alkuarvoja.

Yleensä simuloiduista ketjuista erotetaan alkuosa, ns. *burn in* -jakso, ja käytetään loppuosaa stationaarista jakaumaa koskevien päätelmien tekoon. Jäljelle jääneiden havaintojen kvantiilien perusteella voidaan muodostaa esimerkiksi posteriorivälit estimoitaville parametreille. Yleisesti jos halutaan estimoida $\mu = \mathrm{E}(m(U))$, missä U on stationaarisen jakauman havaintovektori tai sen komponentti ja $m(.)$ on sellainen funktio, että $\mathrm{E}(|m(U)|) < \infty$, tämä voidaan tehdä käyttämällä ergodista keskiarvoa

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} m(u_t),$$

missä $u_1, u_2, ..., u_T$ ovat peräkkäisiä MCMC-menetelmällä simuloituja arvoja.

6 Tilastollinen mallinvalinta

6.1 Ristiinvalidointi

Validointi ja ristiinvalidointi (cross-validation) ovat laskentaintensiivisiä tekniikoita, joita voidaan käyttää tilastollisessa mallinvalinnassa. Tavoitteena on löytää malli, jolla on mahdollisimman hyvä ennustetarkkuus uusille, otokseen kuulumattomille havainnoille. Ennustettava suure Y voi olla esimerkiksi selitettävä muuttuja regressioanalyysissä, ennustettava luokka erotteluanalyysissä tai tuleva havainto aikasarja-analyysissä. Yleensä kunkin havainnon ennuste perustuu perustuu havaintoa vastaavaan ennustavien muuttujien vektoriin $\mathbf{x}$.

Validoinnissa otos jaetaan kahteen osaan, joita käytetään opetus- ja testijoukkoina. Opetusjoukkoa käytetään mallin parametrien estimointiin, jonka jälkeen mallin avulla ennustetaan testijoukon havainnot. Ennusteen hyvyys

voidaan mitata laskemalla yhteen ennustevirheistä aiheutuvat tappiot, kun käytetään jotain sopivaa tappiofunktiota. Validointi tehdään kaikille vertailtaville malleille ja valitaan se malli, jolla paras ennustetarkkuus.

Ristiinvalidoinnissa pyritään otoksen antama informaatio tehokkaammin. Siinä otos jaetaan k suunnilleen yhtä suureen osaan. Tämän jälkeen kunkin osan havainnot ennustetaan käyttäen muiden $k-1$ osan havaintojen avulla estimoitua mallia ja lasketaan yhteen eri havaintojen ennustevirheistä aiheutuvat tappiot. Siis kukin havainto tulee ennustetuksi kertaalleen. Usein k valitaan väliltä 3-20, mutta k voi olla myös havaintojen lukumäärä n, jolloin kukin havainto ennustetaan käyttäen muiden havaintojen avulla estimoitua mallia. Tämä viimeksi mainittu vaihtoehto vaatii eniten laskentakapasiteettia.

Esim. Lineaarinen regressiomalli

Tarkastellaan Taulukossa 5 esitettyä aineistoa. Siinä on annettu estimaatit $\hat{\alpha}_{-i}$ ja $\hat{\beta}_{-i}$, jotka on saatu estimoimalla regressiomalli

$$\mathrm{E}(Y) = \alpha + \beta x,$$

kun i. havainto on pudotettu pois. Taulukossa on annettu myös ennusteet $\hat{y}_i$, jotka on saatu käyttämällä estimaatteihin $\hat{\alpha}_{-i}$ ja $\hat{\beta}_{-i}$ perustuvaa malli. Viimeisenä rivillä on annettu neliölliseen tappiofunktioon perustuva tappiot $(y_i - \hat{y}_i)^2$. Mallin kokonaistappio on

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 24.3 + 112.1 + ... + 19.7 = 495.0.$$

Jos toistetaan vastaava menettely neliölliselle mallille $\mathrm{E}(Y) = \alpha + \beta_1 x + \beta_2 x^2$, kokonaistappioksi saadaan 569.2, joka on suurempi kuin yksinkertaisemman mallin tapauksessa. Siis ristiinvalidointimenettely johtaa tässä tapauksessa vähäparametrisemman regressiomallin valintaan.

6.2 Informaatiokriteerit

Tilastollisessa mallinvalinnassa käytetään rutiininomaisesti informaatiokriteereitä, joista tunnetuimpia ovat AIC ja BIC. Näitä kriteereitä käytetään niin, että pyritään valitsemaan malli, jolla on pienin kriteerin arvo. Informaatiokriteereiden lausekkeessa esiintyy yleensä uskottavuusfunktion maksimi, joka mittaa mallin sopivuutta havaintojen kanssa, sekä mallin parametrien lukumäärän funktio, joka sakottaa liian monista parametreista. Yleensä

uskottavuusfunktio voi saada sitä suuremman arvon mitä enemmän mallissa on parametreja mutta samalla mallin ennustekyky heikkenee.

Akaiken informaatiokriteeri (AIC) perustuu ajatukseen valita malli, jonka Kullback-Leibler-etäisyys oletetusta oikeasta mallista on mahdollisimman pieni. Kriteeri määritellään

$$\mathrm{AIC} = -2\log(L(\hat{\theta}; y) + 2k,$$

missä $\hat{\theta}$ on suurimman uskottavuuden estimaatti parametrivektorille θ ja k on mallin parametrien lukumäärä. Akaiken informaatiokriteeri on harhainen ja erityisesti pienellä otoskoolla on syytä käyttää siitä korjattua versiota

$$\mathrm{AICC} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1},$$

missä n on havaintojen lukumäärä.

Voidaan osoittaa, että Akaiken informaatiokriteerin käyttö on yhtäpitävää ristiinvalidoinnin kanssa suurella otoskoolla, jos tappioksi määritellään

$$-\sum_i \log f_{-i}(y_i|\mathbf{x}_i),$$

missä $f_{-i}(.|\mathbf{x}_i)$ on i. havainnon ennustejakauma, kun muut havainnot paitsi y_i on annettu.

Bayesiläinen informaatokriteeri (BIC), jota kutsutaan myös Schwarzin-Rissasen kriteeriksi (SIC), määritellään

$$\mathrm{BIC} = -2\log(L(\hat{\theta}; y) + k\log(n),$$

missä n on havaintojen lukumäärä. Lausekkeesta nähdään, että BIC rankaisee parametreista enemmän kuin AIC, joten sillä on taipumus valita vähäparametrisempi malli. Kriteeri perustuu ajatukseen valita malli, jolla on suurin posterioritodennäköisyys, kun eri mallien prioritodennäköisyydet ovat yhtä suuret. BIC on tarkentuva, mikä tarkoittaa sitä, että todennäköisyys valita oikea malli lähestyy ykköstä, kun havaintojen lukumäärä lähestyy ääretöntä. Tämä ominaisuus on merkityksellinen, jos voidaan olettaa, että yksi vertailtavista malleista on "oikea".

Bayesiläisen simuloinnin yhteydessä voidaan estimoida devianssin $D(y, \theta) = -2\log(L(\theta; y))$posterioriodotusarvo

$$D_{avg}(y) = E(D(y, \theta)|y)$$

laskemalla keskiarvo $\hat{D}_{avg} = \frac{1}{L}\sum_{l=1}^{L} D(y, \theta_l)$, missä θ_l:t on saatu simuloimalla posteriorijakaumaa. Jälleen voidaan valita malli, jolla $\hat{D}_{avg}(y)$ on pienin.

Mallinen vertailussa voidaan käyttää myös *devianssi-informaatiokriteeriä* (DIC), joka määritellään

$$\mathrm{DIC} = 2\hat{D}_{avg}(y) - D(y, \hat{\theta}),$$

missä $\hat{\theta}$ on jokin piste-estimaatti, esim. posteriorijakauman odotusarvo. Kriteeri perustuu asymptoottisiin tuloksiin ja sitä on ehdotettu käytettäväksi, kun halutaan valita malli, joka ennustaa mahdollisimman hyvin otoksen ulkopuolisia havaintoja.

DIC on yksinkertaisten mallien tapauksessa likimain sama kuin AIC(C), mutta esimerkiksi hierarkisten mallien tapauksessa se antaa eri tuloksen, sillä se perustuu parametrien tehokkaaseen lukumäärään. Riippuu tilanteesta, kumpi kriteeri soveltuu paremmin (ks. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/DIC-slides.pdf). On hyvä myös pitää mielessä, että edellä esitellyt informaatiokriteerit AIC(C), BIC ja DIC perustuvat multinormaalijakauma-approksimaatioon (suurimman uskottavuuden estimaattorin jakauma tai posteriorijakauma), joka ei välttämättä kaikissa tilanteissa päde, ei ainakaan pienellä otoskoolla.

Mallin valinnassa käytetään laajalti myös ns. *lyhimmän kuvauksen periaatetta* (MDL, Minimum Description Length), jonka on kehittänyt suomalainen tutkija Jorma Rissanen. Siinä on ajatuksena valita malli, jonka avulla havaintoaineisto voidaan kuvata mahdollisimman lyhyesti. Kuvauspituuteen lasketaan itse mallin kuvaus ja mallin avulla koodattu havaintoaineisto. (Ks. tarkemmin http://www.mdl-research.org).