# Bayes-tilastotiede 2

Juha Karvanen Arto Luoma Antti Penttinen
Santtu Tikka Miika Kailas

# Sisällysluettelo

# Opintojakson kuvaus

Opintojaksolla käsitellään bayesiläistä data-analyysia ja posteriorijakauman estimoinnissa tarvittavia menetelmiä, erityisesti Markovin ketju Monte Carloa. Käytännön data-analyysiin sovelletaan R-ohjelmistoa ja erilaisia Monte Carlo -simulointiohjelmistoja. Opintojaksolla käsitellään mallinvalintaa ja mallikritiikkiä Bayes-näkökulmasta ja luodaan myös katsaus edistyneempiin Bayes-menetelmiin.

Tarvittavia ja suositeltuja esitietoja:

- R
- Bayes-tilastotiede 1
- Yleistetyt lineaariset mallit, lineaarinen sekamalli

## Osaamistavoitteet

Opintojakson suoritettuaan opiskelija osaa rakentaa hierarkkisia tilastollisia Bayes-malleja kompleksisille ongelmille, osaa käyttää mallinnukseen käytettäviä ohjelmistoja data-analyysissa, pystyy suoriutumaan vaativasta posteriorijakauman laskennasta, ymmärtää MCMC-menetelmien käyttöön liittyvät ongelmat, ja osaa arvioida Bayes-mallien sopivuutta.

## Kirjallisuutta


- Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari ja Donald B Rubin (2013). *Bayesian Data Analysis*, Third Edition, CRC Press. Saatavilla Gelmanin sivuilta ilmaisena pdf:nä
- Andrew Gelman, Jennifer Hill, Aki Vehtari (2020). *Regression and Other Stories*. https://avehtari.github.io/ROS-Examples/
- Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76(1). https://dx.doi.org/10.18637/jss.v076.i01
- Jonah Gabry, Dan Simpson, Aki Vehtari, Michael Betancourt ja Andrew Gelman (2019). *Visualization in Bayesian workflow*. Journal of the Royal Statistical Society A, 182, 389–402. doi:https://dx.doi.org/0.1111/rssa.12378

## Ohjelmistot

Kurssilla pääasiallisena työkaluna toimii probabilistinen ohjelmointikieli Stan, jota myös tuttu `brms`-paketti käyttää mallien sovittamiseen. Stanista ja muista Bayes-mallinnukseen kehitetyistä ohjelmistoista kerrotaan tarkemmin seuraavassa luvussa.

### Listaus kurssilla käytettävistä paketeista

| Paketti | Käyttötarkoitus |
|---|---|
| `rstan` | R-käyttöliittymäpaketti Stan-mallien käyttöön |
| `bayesplot` | Bayes-analyysiin liittyvien kuvien piirtäminen |
| `tidybayes` | Apuvälineitä Bayes-analyysiin |
| `loo` | Bayes-ristiinvalidointi |
| `rstanarm` | `brms`-paketin kaltainen mallinnuspaketti |
| `ggplot2` | Yleinen graafien piirto (jota `bayesplot` käyttää) |
| `dplyr` | Datan muokkaus |

Lisäksi edellä mainitut paketit tarvitsevat suuren joukon muita paketteja toimiakseen. Windows-koneen käyttäjät tarvitsevat myös Rtools-ohjelman.

Koska yllä mainitut paketit päivittyvät usein, on hyvä pitää myös R ajantasaisena. Luentomateriaalin kirjoittamishetkellä vähintään R 4.4.1 olisi hyvä olla asennettuna, mutta vanhemmallakin voi ehkä pärjätä (jos ei, niin pakettien asennusvaiheessa tulee kyllä tästä ilmoitus).

Kokeile ensin toimiiko pakettien rakentaminen:

```
library("pkgbuild") # asenna normaalisti jos puuttuu
```

```
Warning: package 'pkgbuild' was built under R version 4.4.2
```

```
check_build_tools(TRUE)
```

```
Found in Rtools 4.4 installation folder
```

```
Your system is ready to build packages!
```

RStudiota käytettäessä tämä aiheuttaa Rtools-ohjelman asennuksen, jos se puuttuu koneelta. Voit myös asentaa Rtoolsin suoraan lataamalla ohjelman täältä: https://cran.r-project.org/bin/windows/Rtools/

Asenna sitten `rstan`-paketti. Periaatteessa tämän pitäisi onnistua normaalisti komennolla

```
install.packages("rstan")
```

Kokeile sitten toimiiko `rstan` ajamalla `stan`-funktion dokumentaation esimerkki.

Lisää ohjeita tarvittaessa: https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started

# 1 Johdanto

## 1.1 Bayes-tilastotiede lyhyesti

Bayes-tilastotiede tarjoaa systemaattisen tavan ennakkotiedon ja datan yhdistämiseen. Kaiken epävarmuuden kuvaamiseen käytetään todennäköisyyksiä. Ennakkotietoa parametreista $\theta$ kuvataan priorijakaumalla $p(\theta)$. Dataa ehdolla parametrit kuvataan mallilla $p(y|\theta)$. Tavoitteena on selvittää posteriorijakauma $p(\theta|y)$. Aineisto siis päivittää priorijakauman posteriorijakaumaksi. Ehdollisen todennäköisyyden määritelmää käyttäen päädytään Bayesin kaavaan

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{1.1}$$

Bayes-päättelyssä tilastotieteilijän on siis

1. Määritettävä priorijakauma $p(\theta)$.
2. Määritettävä malli $p(y|\theta)$.
3. Laskettava posteriorijakauma Bayesin kaavan avulla.
4. Arvioitava mallin sopivuutta dataan.

Kaikkiin näihin vaiheisiin liittyy merkittäviä haasteita käytännön tilanteessa.

## 1.2 Epävarmuuden lajit

Epävarmuus voidaan jakaa satunnaiseen eli aleatoriseen epävarmuuteen ja tietämykselliseen eli episteemiseen epävarmuuteen. Aleatorinen epävarmuus on seurausta aidosta satunnaisuudesta, jota ei ole mahdollista poistaa keräämällä lisää havaintoja. Esimerkkejä ovat rahanheitto (“alea iacta est”) ja kvanttimekaniikan ilmiöt. Episteeminen epävarmuus taas liittyy siihen, että emme tiedä ilmiöstä niin paljon kuin olisi mahdollista tietää. Esimerkiksi epävarmuus väestön keskipituudesta pienenee, kun poimimme väestöstä satunnaisotoksen ja mittaamme otokseen valittujen henkilöiden pituuden.

## 1.3 Priorin määrittäminen

Ennakkotietoa parametreista kuvataan priorijakauman avulla. Vahvan ennakkotiedon perusteella muodostettua prioria kutsutaan informatiiviseksi. Subjektiivisesta priorista puhuttaessa korostetaan sitä, että eri henkilöiden ennakkotiedot ja -käsitykset saattavat poiketa toisistaan huomattavasti. Herkkyystarkasteluissa tutkitaan, kuinka erilaisten priorien käyttö vaikuttaa posteriorijakaumiin. Mitä enemmän dataa on käytettävissä, sitä vähemmän priorilla on merkitystä. Poikkeuksena on tilanne, jossa priorijakauman tiheysfunktio on nolla jollakin parametrin arvolla. Tällöin posteriorijakauman tiheysfunktiokin on aina nolla tällä parametrin arvolla.

Bayes-tilastotiedettä kohtaan esitetty kritiikki kohdentuu usein priorijakaumien käyttöön. Usein huolena on, että posteriorijakauma määräytyy liian voimakkaasti pelkästään priorijakauman perusteella. Tätä huolta on pyritty vähentämään käyttämällä epäinformatiivisia prioreita, jotka levittävät todennäköisyysmassan mahdollisimman tasaisesti parametrin määrittelyjoukon ylitse.

Modernissa Bayes-tilastotieteessä suositaan usein heikosti informatiivisia prioreita. Tällöin lähtötilanne voi olla informatiivinen priori, jota laimennetaan jakamalla todennäköisyysmassa lähtötilannetta tasaisemmin.

Toinen mahdollinen huolenaihe on se, että ennakkotiedon esittäminen jakaumamuodossa voi olla hankalaa. Paras ennakkokäsitys tutkittavasta ilmiöstä saattaa olla asiantuntijalla, jolla on vain vähäiset tiedot tilastotieteestä. On epärealistista olettaa, että asiantuntija tällöin pystyisi esittämään tietämyksensä suoraan tilastollisena jakaumana. Välittäjäksi tarvitaan tilastotieteilijä. Huolellisesti valittujen kysymysten avulla tilastotieteilijä saavuttaa ymmärryksen asiantuntijan näkemyksistä ja esittää ne priorijakaumana. Tätä kutsutaan tietämyksen koostamiseksi (*elicitation*).

## 1.4 Mallin määrittäminen

Mallin määrittäminen on tärkein tilastotieteilijän tekemä valinta sekä frekventistisessä että bayesiläisessä tilastotieteessä. Toisin kuin priorilla, mallilla on lähes aina suuri merkitys analyysin lopputulokseen. Mallin tulee olla yhdenmukainen sekä ilmiötä koskevan tietämyksen että datankeruuprosessin kanssa. Bayes-tilastotieteessä käytetään usein hierarkkisia malleja, joissa varsinaisten malliparametrien jakauma riippuu hyperparametreista.

## 1.5 Posteriorin laskeminen

Bayesin kaava 1.2 määrittelee posteriorijakauman yksikäsitteisesti, mutta kaavan käyttöön liittyy laskennallisia hankaluuksia. Tutkija määrittelee mallin $p(y|\theta)$ ja priorin $p(\theta)$, mutta datan reunajakauma tulisi laskea kaavalla:

$$p(y) = \int p(y|\theta)p(\theta)\,d\theta, \tag{1.2}$$

missä $\theta$ yleensä on moniulotteinen. Poikkeustapauksia lukuun ottamatta tämä integraali ei kuitenkaan ole lausuttavissa suljetussa muodossa. Mahdollisia ratkaisuja tähän ongelmaan ovat:

- Konjugaattiprioreiden käyttö. Tällöin priori ja malli valitaan siten, että posteriorijakauman muoto tunnetaan. Esimerkiksi jos ilmiötä mallinnetaan Poisson-jakaumalla ja intensiteettiparametrille käytetään priorina Gamma-jakaumaa, parametrin posteriorijakaumakin on Gamma-jakauma. Ratkaisu ei sovellu käytettäväksi monimutkaisten hierarkkisten mallin kanssa. Yksinkertaisissakin tapauksissa konjugaattipriorilla ei välttämättä ole mahdollista kuvata ennakkotietoja totuudenmukaisesti.
- Analyyttinen integrointi. Joissakin tapauksissa integraalin 1.2 laskeminen analyyttisesti voi olla mahdollista.
- Numeerinen integrointi. Käytännössä tämä yleensä mahdollista vain, jos parametrivektorin dimensio on yksi tai kaksi.
- Approksimointi. Oletetaan esimerkiksi, että posteriorijakauma on normaalijakauma ja määritetään tämän jakauman parametrit datasta.
- Riippumattomien havaintojen tuottaminen posteriorijakaumasta. Tässä tapauksessa luovutaan integraalin laskemista ja pyritään sen sijaan tuottamaan otos posteriorijakaumasta. Päättely perustuu tähän otokseen. Tärkeysotanta kuuluu tähän ryhmään.
- Riippuvien havaintojen tuottaminen posteriorijakaumasta. Bayes-laskennan yleisimmin käytetty työväline Markovin ketju Monte Carlo (*Markov chain Monte Carlo*, MCMC) tuottaa riippuvia havaintoja posteriorijakaumasta.

## 1.6 Mallin sopivuuden arviointi

Keskeinen osa Bayes-päättelyä on mallinnuksen tuloksena saadun posteriorijakauman kriittinen tarkastelu. Tutkija arvioi mallin sopivuutta dataan ja posteriorijakaumasta seuraavien päätelmien realistisuutta. Tarkasteluissa käytetään erityisesti posterioriennustejakaumasta generoituja havaintoja, koska niitä voidaan verrata aineistoon. Priorijakaumaan ja mallioletuksiin liittyvillä herkkyystarkasteluilla on myös tärkeä rooli. Mikäli tutkija ei ole tyytyväinen malliin, hän voi palata priorin ja mallin määrittämiseen.

## 1.7 Ohjelmistot

BUGS (Bayesian inference using Gibbs sampling) on deklaratiivinen ohjelmointikieli, jolla kuvaillaan millainen malli on mutta ei määritellä, kuinka se estimoidaan. BUGS-kehitystyö jatkui OpenBUGS-version (http://www.openbugs.net/) parissa, mutta vanhempi WinBUGS-versio (http://www.mrc-bsu.cam.ac.uk/software/bugs/) on myös edelleen käytettävissä. R-paketin `R2OpenBUGS` avulla OpenBUGS-ohjelmaa voi käyttää suoraan R:n kautta (OpenBUGS tulee asentaa ennen `R2OpenBUGS`-pakettia).

JAGS (Just Another Gibbs Sampler, http://mcmc-jags.sourceforge.net/) on vaihtoehtoinen BUGS-toteutus. Se on koodattu C++-kielellä, toisin kuin aiemmat toteutukset, jotka on koodattu vähemmän tunnetulla Component Pascal -kielellä. Tämä tekee siitä alustariippumattoman ja helpommin kehitettävän. OpenBUGS-mallit ovat yleensä käytettävissä JAGSissa suoraan tai pienten muokkausten jälkeen. JAGSia voi käyttää R-paketin `rjags` kautta.

NIMBLE (http://r-nimble.org/) on vähemmän tunnettu mutta kokeilemisen arvoinen ohjelmisto, joka käyttää ja laajentaa BUGS-syntaksia. Siinä BUGS-mallit ovat ohjelmoitavia R-olioita. Ohjelmakoodi käännetään C++-kielen välityksellä laskennan nopeuttamiseksi. BUGS:in alkuperäisten algoritmien lisäksi NIMBLen on implementoitu SMC (*Sequential Monte Carlo*). Käyttäjän on helppo implementoida uusia algoritmeja ja jakaumia. NIMBLEä voi käyttää R-paketin `nimble` kautta.

Stan (http://mc-stan.org/) on aktiivisesti kehitetty ohjelmisto, jonka syntaksi poikkeaa BUGS-syntaksista. Toisin kuin BUGS:issa, mallit määritellään imperatiivisella ohjelmointikielellä, joka määrittelee, mitä tehdään ja missä järjestyksessä. Staniin on implementoitu MCMC-menetelmistä ainoastaan HMC (Hamiltonin Morte Carlo) ja muunnelma sen adaptiivisesta versiosta NUTS (No U-Turn Sampling). HMC on yleensä (oikein säädettynä) tehokas suuriulotteisissa tapauksissa, kun posteriorijakauma on epäsäännöllisen muotoinen ja estimoitavilla parametreilla voimakkaita lineaarisia tai epälineaarisia riippuvuuksia. Algoritmi ei kuitenkaan toimi diskreeteillä posteriorijakaumilla, minkä vuoksi Stania ei voi käyttää malleissa, joissa on diskreettejä parametreja tai diskreettejä latentteja muuttujia. Joissain tapauksissa on mahdollista marginalisoida tällaiset parametrit tai muuttujat pois mallista. Saattaa olla myös toimivaa korvata diskreetit latentit muuttujat vastaavanlaisilla jatkuvilla muuttujilla. (Esim. Poisson-jakauman voisi korvata Gamma($\alpha$,1)-jakaumalla, jolla on sama odotusarvo ja varianssi.) Stan on yleensä JAGSia hitaampi sovitettaessa yksinkertaisia malleja mutta saattaa olla selvästi tehokkaampi sovitettaessa monimutkaisia malleja. Stanissa on myös implementoitu variaatiopäättely (*Variational Inference*, VI), joka on simuloinnille vaihtoehtoinen lähestymistapa approksimatiiviselle Bayes-päättelylle. Kun Stan on asennettu, sitä voidaan käyttää suoraan R-paketin `rstan` kautta.

Muissa data-analyysiin käytetyissä ohjelmointikielissä on enemmän vaihtoehtoja. Juliakieleen sisältyy paketti Turing, joka implementoi tavallisten algoritmien lisäksi kaikki edellä mainitut tehokkaat Bayes-laskennan menetelmät: SMC, VI, HMC, ja NUTS. Nämä ovat kaikki ohjelmoitu suoraan Julialla. Bayes-laskentaa (ja yleisemmin probabilistista ohjelmointia) voi tehdä Pythonin paketeilla PyMC3, Pyro ja Edvard, joista PyMC3 on tällä hetkellä pisimmälle kehitetty ja joka myös sisältää kaikki edellä mainitut menetelmät. Stania voi käyttää Pythonissa PyStan-paketin avulla.

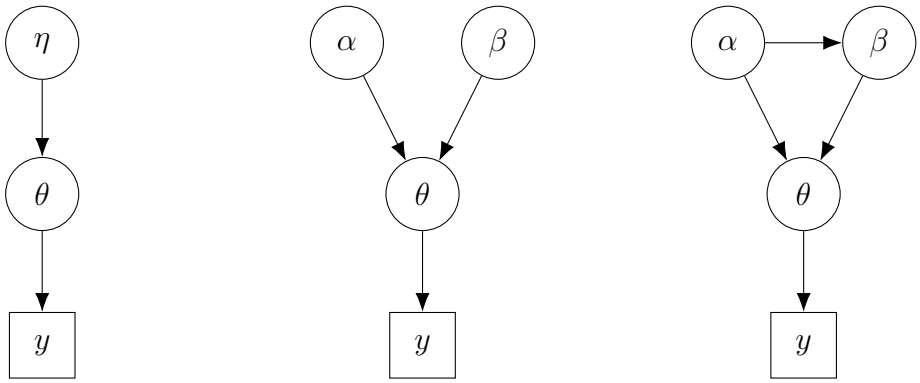
# 2 Hierarkkiset mallit

## 2.1 Aineiston rakenne

Havaintoaineistoissa on usein rakenne (hierarkia), joka vaikuttaa havaintoyksiköiden välisiin riippuvuussuhteisiin. Tällaista aineistoa kutsutaan myös klusteroituneeksi. Yksinkertainen esimerkki on koululaisia koskeva data, jossa hierarkiatasot ovat valtio, kunta, koulu, luokka ja oppilas. Tällaisen datan mallintamiseen käytetään yleisesti hierarkkisia malleja. Bayesiläisessä hierarkkisessa mallissa osa parametreista riippuu toisista parametreista, joita kutsutaan hyperparametreiksi.

## 2.2 Graafiset mallit

Hierarkkinen malli esitetään usein suunnatun silmukattoman graafin avulla (*Directed Acyclic Graph*, DAG). Yleisen käytännön mukaisesti havaittuja muuttujia ja tunnettuja vakioita merkitään graafissa neliöillä ja tuntemattomia parametreja ja latentteja muuttujia ympyröillä. Nuolet osoittavat muuttujien ja parametrien väliset riippuvuussuhteet. Kuvassa 2.1 on esitetty kolme esimerkkiä graafisesta mallista.




(a) (b) (c)

Kuva 2.1: Kolme yksinkertaista esimerkkiä graafisesta mallista. Graafissa (a) parametri $\theta$ riippuu hyperparametrista $\eta$. Graafissa (b) parametri $\theta$ riippuu hyperparametreista $\alpha$ ja $\beta$, jotka ovat toisistaan riippumattomia. Priori faktoroituu muotoon $p(\alpha)p(\beta)$. Graafissa (c) hyperparametrit $\alpha$ ja $\beta$ riippuvat toisistaan ja priori on yleistä muotoa $p(\alpha, \beta)$.

## 2.3 Vaihdannaisuus

Havainnot (satunnaismuuttujat) ovat (äärellisesti) vaihdannaisia (*finitely exchangeable*), jos niiden järjestyksen (indeksien) vaihtaminen ei muuta yhteisjakaumaa. Tällöin havainnoille $y_1, \dots, y_n$ ja mille tahansa permutaatiolle $\pi(1), \dots, \pi(n)$ pätee

$$p(y_1, \dots, y_n) = p(y_{\pi(1)}, \dots, y_{\pi(n)}).$$

Riippumattomuudesta seuraa vaihdannaisuus, mutta vaihdannaiset satunnaismuuttujat voivat olla riippuvia. Jos havainnot eivät ole vaihdannaisia, ne voivat olla ehdollisesti vaihdannaisia. Käytännössä ehdollinen vaihdannaisuus saavutetaan yleensä olettamalla, että havainnot ovat riippumattomia ehdolla parametrit. Syvällinen ja teoreettisesti merkittävä de Finettin esityslause lähtee kuitenkin liikkeelle (äärettömästä) vaihdannaisuudesta ja johtaa sen perusteella Bayes-tilastotieteen koneiston. Ääretön jono satunnaismuuttujia on (äärettömästi) vaihdannainen, jos mikä tahansa äärellinen osajono on vaihdannainen. Laajennetun version de Finettin lauseesta voi yksinkertaistettuna esittää seuraavasti

**Lause 2.1.** *Tiettyjen säännöllisyysehtojen vallitessa äärettömästi vaihdannaisen jonon osajonon $y_1, \dots, y_n$ yhteisjakauman voi esittää muodossa*

$$p(y_1, \dots, y_n) = \int \prod_{i=1}^{n} p(y_i|\theta) p(\theta)\, d\theta$$

*jollekin jakaumalle $p(\theta)$.*

Lauseen perusteella vaihdannaiset muuttujat voidaan tulkita ehdollisesti riippumattomaksi ja samoin jakautuneiksi satunnaismuuttujiksi jostakin parametrisesta jakaumasta $p(\cdot|\theta)$, jonka parametrille $\theta$ käytetään priorijakaumaa $p(\theta)$.

## 2.4 Hierarkkisen mallin sovittaminen

Tarkastellaan hierarkkisen mallin sovittamista käytännön esimerkin avulla.

**Esimerkki 2.1** (PISA-tutkimus)**.** PISA-tutkimuksessa on mitattu oppilaiden lukutaitoa. Aineisto sisältää tiedon koulusta (koodattuna), mutta ei sisällä tietoa luokasta tai kunnasta. Olkoon $y_{ij}$ koulun $j$ oppilaan $i$ saama pistemäärä. Pistemäärät on standardoitu siten, että kansainvälisen vertailujakauman keskiarvo on 500 ja hajonta 100. Aineistossa on 4403 oppilasta 148 eri koulusta. Ajatellaan, että oppilaiden osaamisessa voi olla koulukohtaisia eroja eli saman koulun oppilaat on osaamistasoltaan enemmän samankaltaisia kuin eri koulujen oppilaat. Kuvataan koulun $j$ keskimääräistä tasoa parametrilla $\theta_j$. Koska normaalijakaumaoletus vaikuttaa tässä tapauksessa perustellulta, valitsemme malliksi

$$y_{ij} \sim \mathrm{N}(\theta_j, \sigma^2),$$

missä $\sigma^2$ on varianssiparametri (yhteinen kaikille kouluille). Parametreille $\theta_j$ ja $\sigma^2$ tarvitsemme priorijakaumat.

Koska tutkimuksessa mukana olevat koulut ovat satunnaisotos Suomen kaikista kouluista, on luontevaa olettaa koulukohtaisille osaamistasoille oma jakauma. Mallinnamme tätä populaatiojakaumaa normaalijakaumalla

$$\theta_j \sim \mathrm{N}(\mu, \omega^2),$$

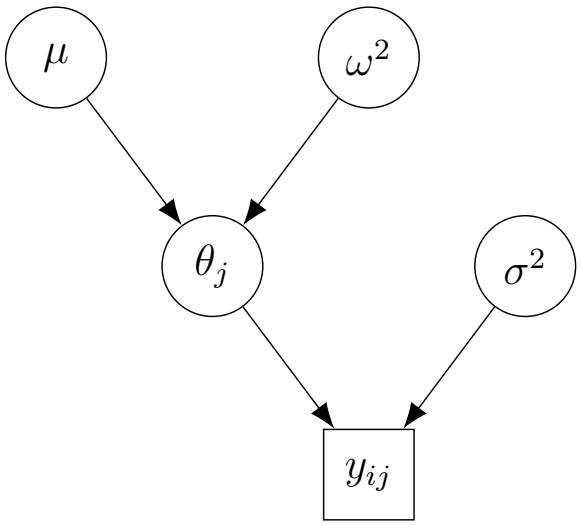
missä $\mu$ ja $\omega^2$ ovat koulutason odotusarvo- ja varianssiparametri (hyperparametreja). Odotusarvolle valitsemme priorin

$$\mu \sim \mathrm{Tas}(0, 1000).$$

Varianssiparametreille oletamme heikosti informatiiviset priorit

$$\sigma^{-2} \sim \mathrm{Gamma}(0.01, 0.01)$$
$$\omega^{-2} \sim \mathrm{Gamma}(0.01, 0.01).$$

Malliin liittyvä DAG on esitetty kuvassa 2.2.




Kuva 2.2: PISA-esimerkin graafinen malli.

**Esimerkki 2.2** (Haastattelun stressivaikutus)**.** Kokeessa osallistujilta (kiinteä lukumäärä) mitataan kyselyllä, onko henkilö kokenut ainakin yhden stressitapahtuman kuukausittain haastattelutapahtumaa edeltävänä 18 kuukauden aikana.

Tutkitaan, onko haastattelutapahtuman läheisyydellä vaikutusta stressitilaan sovittamalla yksinkertainen Poisson-jakauma.

On siis havaittu stressitapahtumien lukumäärä kuukausittain $y_1, \ldots, y_{18}$. Malli tapahtumille on

$$y_i|\lambda_i \sim \mathrm{Poisson}(\lambda_i), \quad \log \lambda_i = \beta_0 + \beta_1 i,$$

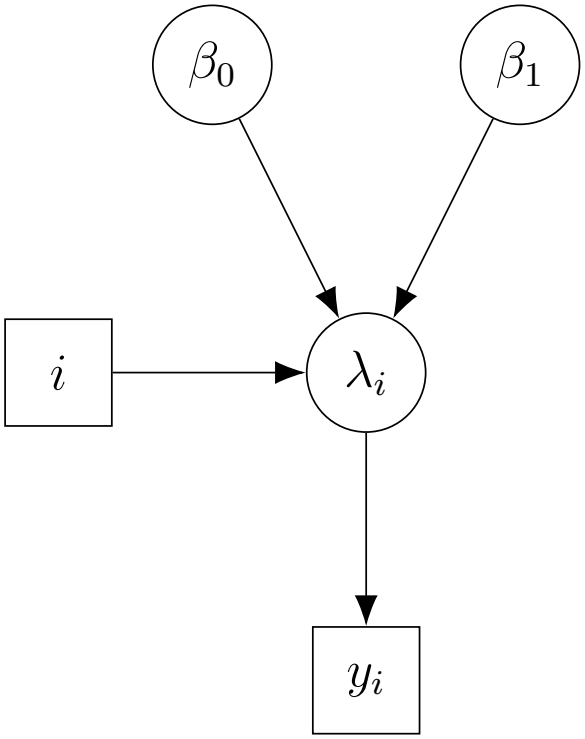
missä havainnot $y_i|\lambda_i$ oletetaan ehdollisesti riippumattomiksi. Prioriksi valitaan tasajakauma:

$$p(\beta_0, \beta_1) \propto 1.$$

Tällöin posterioriksi saadaan

$$p(\beta_0, \beta_1|y) \propto \exp\left(\sum_{i=1}^{18} [y_i(\beta_0 + \beta_1 i) - \exp(\beta_0 + \beta_1 i)]\right).$$

Malliin liittyvä DAG on esitetty kuvassa 2.3




Kuva 2.3: Stressitapahtumaesimerkin DAG.

Tässä ongelmassa kiinnostuksen kohteena on erityisesti parametri $\beta_1$ eli haastattelun läheisyyden vaikutus stressiin.

Huom. Yksilötason vaihtelua ei saada mallinnetuksi, koska kustakin henkilöstä on vain yksi havainto. Vaihtelun mallintamiseksi tarvittaisiin toistomittauksia.

# 3 Markovin ketju Monte Carlo

Bayes-tilastotieteessä Markovin ketju Monte Carloa (*Markov chain Monte Carlo*, MCMC) käytetään havaintojen simulointiin posteriorijakaumasta. Tämä tapahtuu rakentamalla Markovin ketju, jonka *tasapainojakauma* on kiinnostuksen kohteena oleva posteriorijakauma.

## 3.1 Markovin ketju

**Määritelmä 3.1** (Markovin ketju)**.** Satunnaismuuttujien jonoa $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...$ kutsutaan Markovin ketjuksi mikäli kaikilla $t > 0$

$$p(\theta^{(t+1)}|\theta^{(0)}, ..., \theta^{(t)}) = p(\theta^{(t+1)}|\theta^{(t)}).$$

Satunnaismuuttujien $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...$ mahdollisia arvoja kutsutaan usein tiloiksi ja näiden muodostamaa joukkoa tila-avaruudeksi. Markovin ketjussa tilasiirtymien todennäköisyydet (tai tiheydet, jos tila-avaruus on jatkuva) riippuvat ainoastaan nykyisestä tilasta. Sillä, kuinka nykyiseen tilaan on päädytty, ei ole merkitystä tuleviin tilasiirtymiin. Tiettyjen ehtojen vallitessa voidaan osoittaa, että kun $t \to \infty$, satunnaismuuttujan $\theta^{(t+1)}$ jakauma suppenee kohti *tasapainojakaumaa*, joka ei riipu ketjun alkuarvosta $\theta^{(0)}$.

Tehtävänä on siis luoda Markovin ketju, jonka tasapainojakauma on haluttu posteriorijakauma. Tämän toteuttamiseen on olemassa useita algoritmeja, joista tärkeimpiä käsitellään jatkossa. Lisäksi tulee valita Markovin ketjun alkuarvo ja selvittää milloin ketju on supennut tasapainojakaumaansa.

## 3.2 Globaali ja lokaali tasapainoehto

Tarkastellaan Markovin ketjua, jonka tila-avaruus $S$ on diskreetti. Tällöin siirtymää tilasta $i$ tilaan $j$ voidaan merkitä todennäköisyydellä $p_{ij}$.

**Määritelmä 3.2** (Tasapainojakauma)**.** Markovin ketjun tasapainojakauma on jakauma $\{\pi_i, i \in S\}$, joka toteuttaa globaalin tasapainoehdon

$$\sum_{i \in S} \pi_i p_{ij} = \pi_j \text{ kaikille } j \in S.$$

Globaalissa tasapainoehdossa tarkastellaan siis erilaisia tapoja saapua tilaan $j$. Kun kunkin lähtötilan $i$ todennäköisyys määrätään tasapainojakaumasta, päädytään tilaan $j$ tasapainojakauman mukaisella todennäköisyydellä $\pi_j$.

Tasapainojakauman olemassaolo voidaan osoittaa, mikäli tietyt ketjua koskevat ehdot täyttyvät. Näistä ehdoista tärkein on pelkistymättömyys (irreducibility), joka takaa sen, että siirtyminen mistä tahansa tilasta $i$ mihin tahansa tilaan $j$ on mahdollista äärellisellä määrällä siirtymiä.

Globaalin tasapainoehto ei ole kovinkaan käyttökelpoinen, kun tavoitteena on konstruoida Markovin ketju, jolla on annettu tasapainojakauma. Simulointialgoritmit perustuvatkin vahvempaan *lokaaliin tasapainoehtoon*

$$\pi_i p_{ij} = \pi_j p_{ji} \text{ kaikilla } i, j \in S.$$

Lokaalissa tasapainoehdossa tarkastellaan siis siirtymää kahden tilan välillä. Ehdon mukaan tasapainotilanteessa on yhtä todennäköistä havaita siirtymä tilasta $i$ tilaan $j$ kuin tilasta $j$ tilaan $i$. Lokaalista tasapainosta seuraa globaali tasapaino seuraavalla tavalla

$$\sum_{i \in S} \pi_i p_{ij} = \sum_{i \in S} \pi_j p_{ji} = \pi_j \sum_{i \in S} p_{ji} = \pi_j.$$

## 3.3 Gibbsin algoritmi

Gibbsin algoritmissa Markovin ketju rakennetaan siten, että mallin parametreja päivitetään yksi kerrallaan. Uusi arvo generoidaan ehdollisesta jakaumasta, jossa ehtoina ovat muut parametrit. Aluksi mallin parametrit jaetaan $k$ komponenttiin $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Yleensä komponentit ovat skalaariparametreja, mutta ne voivat olla myös vektoreita. Gibbsin algoritmin vaiheet ovat:

1. Aseta alkuarvot $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}$ ja aseta $t = 0$
2. Generoi ensin $\theta_1^{(t+1)}$ jakaumasta $p(\theta_1 | \theta_2^{(t)}, \ldots, \theta_k^{(t)})$ ja sitten $\theta_2^{(t+1)}$ jakaumasta $p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)})$ (jossa $\theta_1$ on jo siis päivitetty). Jatka näin kunnes kaikki komponentit on päivitetty ja kasvata sitten iteraatiota $t$ yhdellä.
3. Toista vaihetta 2 riittävän monta kertaa (yleensä tarvitaan tuhansia iteraatioita).

**Esimerkki 3.1** (Kaksiulotteisen normaalijakauman simulointi)**.** Oletetaan, että

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_2(\mu, \Sigma),\ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

On laskettava ehdolliset jakaumat. Tässä tapauksessa saadaan:

$$X_1 | X_2 = x_2 \sim \mathrm{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), (1 - \rho^2)\sigma_1^2\right)$$
$$X_2 | X_1 = x_1 \sim \mathrm{N}\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1 - \mu_1), (1 - \rho^2)\sigma_2^2\right),$$

missä $\rho = \dfrac{\sigma_{12}}{\sigma_1 \sigma_2}$ on korrelaatiokerroin.

Simulointi etenee seuraavasti: Olkoon alkuarvo $(x_1^{(0)}, x_2^{(0)})^T$.

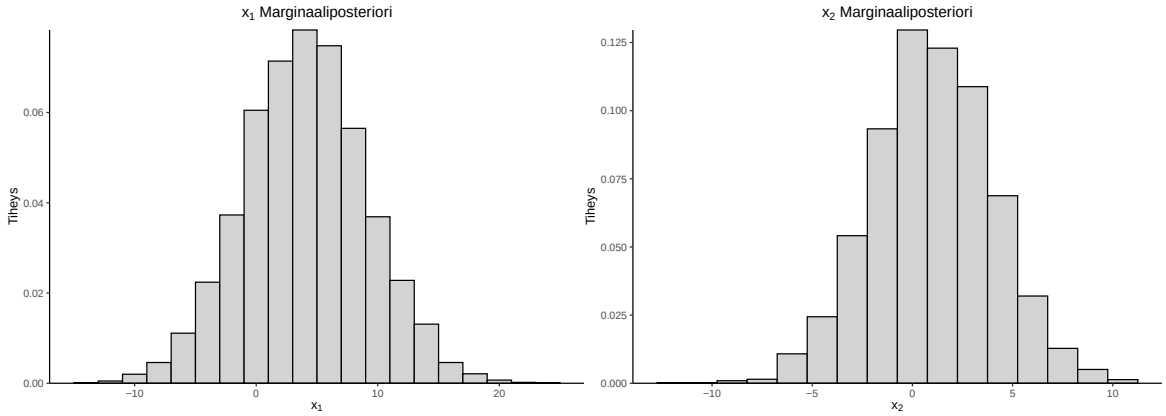
$$\begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} \to \begin{pmatrix} x_1^{(1)} \\ x_2^{(0)} \end{pmatrix} \to \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} \to \begin{pmatrix} x_1^{(2)} \\ x_2^{(1)} \end{pmatrix} \to \cdots$$

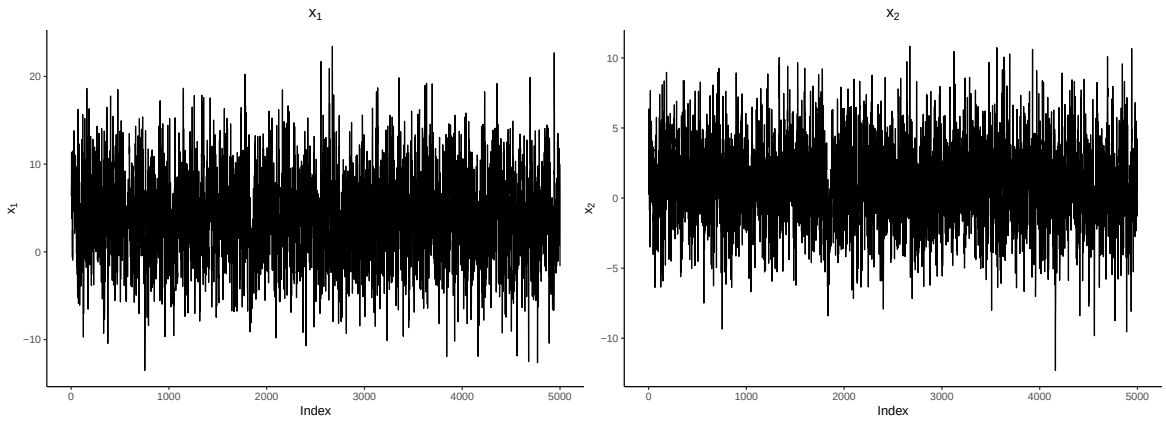
Simuloinnin voi toteuttaa esimerkiksi R-ohjelmistolla seuraavasti

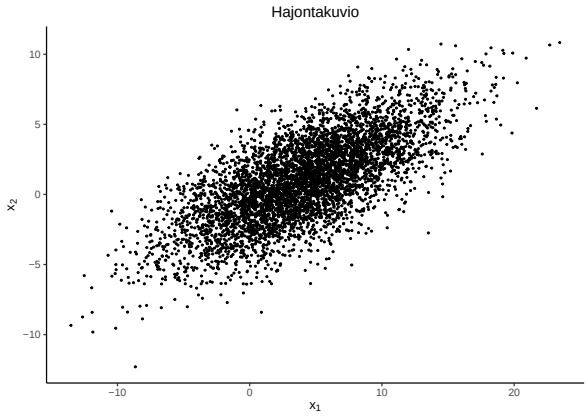
```
norm_sim <- function(mu, S, x_0, n_sim, n_burn) {
  x <- matrix(ncol = 2, nrow = n_burn + n_sim)
  x[1, ] <- x_0           # alkuarvo
  s1 <- sqrt(S[1, 1])     # x1:n keskihajonta
  s2 <- sqrt(S[2, 2])     # x2:n keskihajonta
  s12 <- S[1, 2]          # kovarianssi
  rho <- s12 / (s1 * s2) # korrelaatiokerroin
  for (i in seq(2, n_burn + n_sim)) {
    x[i, 1] <- rnorm(
      n = 1,
      mean = mu[1] + s1 / s2 * rho * (x[i - 1, 2] - mu[2]),
      sd = sqrt(1 - rho^2) * s1
    )
    x[i, 2] <- rnorm(
      n = 1,
      mean = mu[2] + s2 / s1 * rho * (x[i, 1] - mu[1]),
      sd = sqrt(1 - rho^2) * s2
    )
  }
  x[seq(n_burn + 1, n_burn + n_sim), ]
}

mu <- c(4, 1)
S <- matrix(c(25, 10.5, 10.5, 9), ncol = 2)
out <- norm_sim(mu, S, c(0, 0), 5e3, 1e3)
```

Jakaumaa voidaan kuvailla esimerkiksi sen tunnuksilla, marginaalijakaumien histogrammeilla sekä otoksen hajontakuviolla.

| | Mean | SD | 2.5 % | 97.5 % |
|---|---|---|---|---|
| $x_1$ | 4.049294 | 5.064460 | -5.868012 | 13.991879 |
| $x_2$ | 1.012624 | 3.012318 | -4.974440 | 7.037718 |


x1 Marginaaliposteriori
Tiheys
x1
x2 Marginaaliposteriori
Tiheys
x2


x1
x1
Index
x2
x2
Index




## 3.4 Metropolis-algoritmi

Metropolis-algoritmissa käytetään ehdotusjakaumaa, josta generoidaan ehdotuksia Markovin ketjun seuraavaksi tilaksi. Ehdotus voidaan hyväksyä tai hylätä. Jälkimmäisessä tapauksessa seuraava tila on sama kuin nykyinen tila. Algoritmi perustuu siihen, että posteriorijakaumien suhde voidaan laskea normeeraamattomien posteriorijakaumien avulla, vaikka Bayesin kaavan nimittäjää $p(y)$ ei tunnettaisikaan. Olkoon $\theta_a$ ja $\theta_b$ kaksi parametrin $\theta$ arvoa. Posteriorijakaumien suhteelle pätee

$$\frac{p(\theta_b|y)}{p(\theta_a|y)} = \frac{p(\theta_b)p(y|\theta_b)/p(y)}{p(\theta_a)p(y|\theta_a)/p(y)} = \frac{p(\theta_b)p(y|\theta_b)}{p(\theta_a)p(y|\theta_a)}.$$

Algoritmi voidaan esittää seuraavasti:

1. Aseta alkuarvo $\theta^{(0)}$ ja aseta $t = 0$.
2. Toista vaiheita 3–5, iteraatioilla $t = 1, 2, \ldots, T$.
3. Generoi ehdotus $\theta^*$ ehdotusjakaumasta $q(\theta^*|\theta^{(t)})$. Ehdotusjakauman tulee täyttää symmetrisyysehto $q(\theta_a|\theta_b) = q(\theta_b|\theta_a)$ kaikilla $\theta_a$, $\theta_b$.
4. Laske hyväksymistodennäköisyys normeeraamattomien posteriorijakaumien avulla
$$\alpha = \min\left\{1, \frac{p(\theta^*)p(y|\theta^*)}{p(\theta^{(t)})p(y|\theta^{(t)})}\right\}$$
5. Aseta
$$\theta^{(t+1)} = \begin{cases} \theta^* & \text{todennäköisyydellä } \alpha, \\ \theta^{(t)} & \text{muutoin.} \end{cases}$$

Metropolis-algoritmi toteuttaa lokaalin tasapainoehdon. Tämän osoittamiseksi tarkastellaan Markovin ketjun siirtymätodennäköisyyksiä

$$P(\theta^*|\theta^{(t)}) = \begin{cases} q(\theta^*|\theta^{(t)}) \min\left\{1, \frac{p(\theta^*)p(y|\theta^*)}{p(\theta^{(t)})p(y|\theta^{(t)})}\right\}, & \theta^* \neq \theta^{(t)}, \\ q(\theta^{(t)}|\theta^{(t)}) + \sum_{\theta' \neq \theta^{(t)}} q(\theta'|\theta^{(t)}) \left(1 - \min\left\{1, \frac{p(\theta')p(y|\theta')}{p(\theta^{(t)})p(y|\theta^{(t)})}\right\}\right), & \theta^* = \theta^{(t)}. \end{cases}$$

Lokaali tasapaino edellyttää ehdon $p(\theta^{(t)}|y)P(\theta^*|\theta^{(t)}) = p(\theta^*|y)P(\theta^{(t)}|\theta^*)$ voimassaoloa. Jos $\theta^* = \theta^{(t)}$, ehto toteutuu suoraan. Tilanteessa $\theta^* \neq \theta^{(t)}$ saadaan

$$\begin{aligned} p(\theta^{(t)}|y)P(\theta^*|\theta^{(t)}) &= p(\theta^{(t)}|y)q(\theta^*|\theta^{(t)}) \min\left\{1, \frac{p(\theta^*|y)}{p(\theta^{(t)}|y)}\right\} \\ &= q(\theta^{(t)}|\theta^*) \min\left\{p(\theta^*|y), p(\theta^{(t)}|y)\right\} \\ &= p(\theta^*|y)q(\theta^{(t)}|\theta^*) \min\left\{1, \frac{p(\theta^{(t)}|y)}{p(\theta^*|y)}\right\} \\ &= p(\theta^*|y)P(\theta^{(t)}|\theta^*), \end{aligned}$$

missä toisella rivillä on käytetty ehdotusjakauman symmetrisyysominaisuutta $q(\theta^*|\theta^{(t)}) = q(\theta^{(t)}|\theta^*)$.

## 3.5 Metropolis-Hastings-algoritmi

Metropolis-Hastings-algoritmissa Metropolis-algoritmia muokataan siten, että ehdotusjakauman ei tarvitse täyttää symmetrisyysehtoa. Algoritmi voidaan esittää seuraavasti:

1. Aseta alkuarvo $\theta^{(0)}$ ja aseta $t = 0$.
2. Toista vaiheita 3–5, iteraatioilla $t = 1, 2, \ldots, T$.
3. Generoi ehdotus $\theta^*$ ehdotusjakaumasta $q(\theta^*|\theta^{(t)})$. Ehdotusjakauman ei tarvitse täyttää symmetrisyysehtoa.
4. Laske hyväksymistodennäköisyys ehdotusjakauman ja normeeraamattomien posteriorijakaumien avulla
$$\alpha = \min\left\{1, \frac{p(\theta^*)p(y|\theta^*)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)})p(y|\theta^{(t)})q(\theta^*|\theta^{(t)})}\right\}$$
5. Aseta
$$\theta^{(t+1)} = \begin{cases} \theta^* & \text{todennäköisyydellä } \alpha, \\ \theta^{(t)} & \text{muutoin.} \end{cases}$$

Metropolis-Hastings-algoritmia voidaan käyttää Gibbsin algoritmin osana. Parametri, jonka ehdollista jakaumaa ei kyetä laskemaan, voidaan päivittää Metropolis-Hastings-algoritmilla. Toisaalta Gibbsin algoritmi voidaan nähdä Metropolis-Hastings-algoritmin varianttina, jossa ehdotukset tuotetaan komponenteittain ehdollisista jakaumista ja hyväksytään todennäköisyydellä 1.

## 3.6 Ketjun suppenemisen tarkastelu

MCMC-menetelmiä käytettäessä tulee varmistaa, että ketju on supennut tasapainojakaumaan. Ketjun alkupään havainnot eivät ole peräisin tasapainojakaumasta eikä niitä tästä syystä tule käyttää posteriorijakaumaa estimoitaessa. Ketjun alkua kutsutaan nimellä lämmitys tai sisäänajo (englanniksi *burn-in* tai *warm-up*). Monet ohjelmistot poistavat oletusarvoisesti ketjun ensimmäisen puoliskon, ellei käyttäjä toisin määrää.

Ensimmäinen tapa arvioida ketjun suppenemista on silmämääräinen tarkastelu. Ketjun tulisi näyttää satunnaiselta vaihtelulta vakiona pysyvän keskiarvon ympärillä. Silmämääräisen tarkastelun avulla on usein mahdollista todeta, että ketju ei ole vielä supennut, mutta suppenemisen varmistaminen pelkästään tällä tavalla ei ole täysin luotettavaa. On yleistä, että jokin mallin parametri on supennut mutta jokin toinen parametri ei ole. Kaikkien parametrien systemaattinen tarkastelu kuvaajien avulla voi viedä paljon aikaa, koska Bayes-malleissa usein satoja tai tuhansia parametreja.

Edistyneemmät suppenemistarkastelut vaativat usean eri ketjun käyttämistä. Ketjujen alkuarvojen tulisi selvästi poiketa toisistaan, jotta alkuarvojen vaikutusta suppenemiseen voidaan tarkastella. Samaa alkuarvoa käytettäessä osa mahdollisista ongelmista voi jäädä huomaamatta. Ketjujen käyttäytymistä voi edelleen tarkastella silmämääräisesti esimerkiksi piirtämällä (saman parametrin) eri ketjut samaan kuvaan. Suppenemisen tarkastelemiseen on myös ehdotettu erilaisia tunnuslukuja, joista seuraavaksi esitellään usean ohjelmiston käyttämä tunnusluku *mahdollisen skaalan pienenemisen kerroin* $\hat{R}$ (*potential scale reduction factor*).

Tunnusluku perustuu ketjujen sisäisen ja ketjujen välisen vaihtelun vertaamiseen ja siitä on esitetty useita eri versioita kirjallisuudessa. Seuraava määritelmä on niin sanottu jaettujen ketjujen $\hat{R}$ (*split* $\hat{R}$). Idea tässä tunnusluvun variantissa on, että sen tulisi huomioida samanaikaisesti ketjujen mahdollinen epästationaarisuus ja huono sekoittuminen. Olkoon käytössä $m_0$ ketjua, joista jokaisesta on simuloitu $n_0$ havaintoa. Jätetään nyt ensimmäinen puolikas havainnoista pois lämmittelyjaksona (eli $n_0/2$ havaintoa). Jaetaan nyt jokainen ketju kahteen osaan siten, että ensimmäiseen osaan valitaan ensimmäiset $n_0/4$ jäljelle jääneistä havainnoista, ja toiseen osaan viimeiset $n_0/4$ havaintoa. Ajatellaan nyt jaettuja ketjuja erillisinä siten, että käsiteltävä ketjujen määrä on $m = 2m_0$ joista jokaisesta on $n = n_0/4$ havaintoa.

Skalaariparametrista $\theta$ on käytössä simuloidut arvot $\theta^{(ij)}$, $i = 1, \dots, n$, $j = 1, \dots, m$. Ketjujen välistä vaihtelua kuvaa neliösumma

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\theta}^{(\cdot j)} - \overline{\theta}^{(\cdot\cdot)})^2, \text{ missä } \overline{\theta}^{(\cdot j)} = \frac{1}{n} \sum_{i=1}^{n} \theta^{(ij)} \text{ ja } \overline{\theta}^{(\cdot\cdot)} = \frac{1}{m} \sum_{j=1}^{m} \overline{\theta}^{(\cdot j)}. \tag{3.1}$$

Ketjujen sisäistä vaihtelua kuvaa neliösumma

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \text{ missä } s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\theta^{(ij)} - \overline{\theta}^{(\cdot j)})^2.$$

Sisäinen varianssi $W$ aliestimoi posteriorivarianssia $\mathrm{Var}(\theta|y)$, koska äärellinen ketju ei välttämättä ole käynyt läpi koko parametriavaruutta. Sen sijaan estimaattori

$$\widehat{\mathrm{Var}}^{+}(\theta|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

yliestimoi posteriorivarianssia. Sekä $W$ että $V$ ovat posteriorivarianssin tarkentuvia estimaattoreita. Tunnusluku $\hat{R}$ perustuu näiden kahden estimaattorin suhteeseen

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}(\theta|y)}{W}},$$

joka suppenee kohti arvoa 1, kun $n \to \infty$. Käytännössä ketjun katsotaan supenneen, jos $\hat{R} < 1.05$ kaikille parametreille.

## 3.7 Tehokas otoskoko

Koska Markovin ketjut tuottavat riippuvia havaintoja, otos sisältää tyypillisesti vähemmän informaatiota kuin samankokoinen otos riippumattomia havaintoja. Tehokkaan otoskoon tarkoituksena on kertoa, kuinka suurta riippumatonta otosta käytettävissä oleva ketju vastaa. Esimerkiksi Stan estimoi tehokkaan otoskoon automaattisesti. Tehokkaan otoskoon voi estimoida esimerkiksi myös `coda`-paketin funktiolla `effectiveSize`.

Mikäli ketju on tasapainotilassa, riippuvuudella ei ole vaikutusta odotusarvoon. Ketjujen välinen vaihtelu $B$ (kaava 3.1) sen sijaan on suurempaa kuin riippumattomalla otoksella. Mitä suurempi ketjujen autokorrelaatio on, sitä enemmän $B$ poikkeaa riippumattoman otoksen vaihtelusta.

Tehokas otoskoko määritellään ketjun autokorrelaation avulla eri viiveitä (*lags*) tarkastelemalla. Viiveen $k \geq 0$ autokorrelaatio $\rho_k$ ketjulle, jonka yhteisjakauman on $p(\theta)$ odotusarvolla $\mu$ ja varianssilla $\sigma^2$, määritellään seuraavasti

$$\rho_k = \frac{1}{\sigma^2}\int_S (\theta^{(t)} - \mu)(\theta^{(t+k)} - \mu)p(\theta)\, d\theta. \tag{3.2}$$

Autokorrelaatio mittaa ketjujen $\theta^{(t)}$ ja $\theta^{(t+k)}$ välistä korrelaatiota, eli siis alkuperäisen ketjun $\theta^{(t)}$ ja siitä $k$ indeksiä siirtämällä saadun ketjun välistä korrelaatiota.

Otoskokoa $n$ vastaava tehokas otoskoko määritellään autokorrelaatioiden avulla seuraavasti:

$$n_{\mathrm{eff}} = \frac{n}{\sum_{k=-\infty}^{\infty}\rho_k} = \frac{n}{1 + 2\sum_{k=1}^{\infty}\rho_k}.$$

Otoksesta laskettuja autokorrelaatioita käyttämällä tehokkaalle otoskoolle voidaan johtaa estimaattori

$$\hat{n}_{\mathrm{eff}} = \frac{mn}{1 + 2\sum_{k=1}^{K}\hat{\rho}_k},$$

missä $\hat{\rho}_k$ on estimaatti viiveen $k$ autokorrelaatiolle joka on laskettu $m$ ketjusta. Estimoinnissa käytetty maksimiviive $K$ on pienin pariton luku, jolle $\hat{\rho}_{K+1} + \hat{\rho}_{K+2} < 0$.

# 4 Mallikritiikki ja mallin valinta

Mallikritiikissä tarkastellaan mallin sopivuutta dataan. Mallikritiikkiin liittyvät menetelmät ja periaatteet ovat aktiivinen tutkimusaihe Bayes-tilastotieteessä.

Mallin sopivuutta tulisi arvioida uudella datalla, jota ei ole käytetty mallin sovittamisessa. Tämän toteuttamiseksi data voidaan jakaa opetusaineistoon $y_f$ ja testiaineistoon $y_c$. Jos tällainen jako toteutetaan usealla eri tavalla, puhutaan ristiinvalidoinnista. Erilaiset informaatiokriteerit pyrkivät myös mittaamaan mallin sopivuutta uudessa datassa. Informaatiokriteereiden arvo riippuu uskottavuudesta ja mallin parametrien lukumäärästä.

## 4.1 Posterioriennustejakauman tarkastelu

Bayes-tilastotieteessä tulokset esitetään parametrien posteriorijakaumana. Parametrien posteriorijakaumaa $p(\theta|y_f)$ ei voi suoraan verrata todellisiin arvoihin $y_c$, vaan vertailu tulee tehdä posterioriennustejakauman $p(\tilde{y}_c|y_f)$ avulla. Posterioriennustejakauma saadaan integroimalla posteriorijakauman ylitse

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

Käytännössä tämä toteutetaan siten, että jokaisella MCMC-ketjun arvolla $\theta^{(i)}$ generoidaan mallista uusi havaintoaineisto $\tilde{y}_c^{(i)}$. Näitä posterioriennustejakauman tuottamia toistoaineistoja verrataan testiaineistoon $y_c$.

Käytännössä ei yleensä jaeta aineistoa vaan käytetään samaa aineistoa sekä estimointiin että mallintarkistukseen ($y_c = y_f$). Tällöin diagnostiikat ovat todennäköisesti konservatiivisia, eli ne eivät osoita poikkeamaa mallin ja havaintoaineiston välillä kovinkaan herkästi.

## 4.2 Bayes-p-arvo

Bayes-p-arvo (eli posterioriennuste-p-arvo) kertoo, kuinka hyvin testidata ja posterioriennustejakauma vastaavat toisiaan jonkin kiinnostuksen kohteena olevan ominaisuuden suhteen. Olkoon $T(y_c)$ jokin datasta laskettu tunnusluku, jota voi nimittää myös testisuureeksi. Esimerkiksi aineiston minimi tai maksimi voi olla tällainen tunnusluku. Bayes-p-arvo lasketaan todennäköisyytenä, että

$$\begin{aligned} p_{\text{Bayes}} =& P(T(\tilde{y}_c) \geq T(y_c)|y_f) \\ =& \int I(T(\tilde{y}_c) \geq T(y_c))p(\tilde{y}_c|y_f)\,d\tilde{y}_c, \end{aligned}$$

missä $I$ on indikaattorifunktio. Bayes-p-arvon laskemiseksi tarvitaan siis posterioriennustejakaumasta useita simuloituja otoksia, joiden koko on sama kuin testiaineiston $y_c$. Tällöin p-arvo saadaan sellaisten otosten suhteellisena osuutena, joilla niistä laskettu testitunnusluku ylittää testiaineistosta lasketun tunnusluvun. Jos $p_{\text{Bayes}}$ tai $1 - p_{\text{Bayes}}$ on pieni, malli sopii huonosti dataan sen ominaisuuden osalta, jota tunnusluku $T(y_c)$ kuvaa.

Testisuure voidaan yleistää niin, että se riippuu myös mallin tuntemattomista parametereista. Olkoon $T(y, \theta)$ tällainen testisuure. Tällöin Bayes-p-arvo määritellään

$$\begin{aligned} p_{\text{Bayes}} &= P(T(\tilde{y}_c, \theta) \geq T(y_c, \theta)|y_f) \\ &= \int\int I(T(\tilde{y}_c, \theta) \geq T(y_c, \theta)) p(\tilde{y}_c|\theta) p(\theta|y_f)\, d\tilde{y}_c d\theta. \end{aligned}$$

## 4.3 Jäännökset

Frekventistisessä tilastotieteessä mallikritiikki perustuu usein mallin jäännösten tarkastelemiseen. Jäännöksiä voi tarkastella myös Bayes-tilastotieteessä.

Standardoitu Pearsonin jäännös $r_i$ havainnolle $y_i$ lasketaan parametrien funktiona

$$r_i(\theta) = \frac{y_i - \mathrm{E}(y_i|\theta)}{\sqrt{\mathrm{Var}(y_i|\theta)}}.$$

Standardointi toimii jatkuvien muuttujien tapauksessa. Jäännöksiä $r_i(\theta)$ voi verrata $\mathrm{N}(0, 1)$-jakaumaan, mutta formaali inferenssi on hankalaa, koska jäännökset eivät ole toisistaan riippumattomia.

Jos vastemuuttuja on diskreetti, mallin sopivuutta voi tarkastella vertailemalla havaittuja ja ennustettuja suureita sopivasti määritellyissä luokissa. Niin sanottujen devianssijäännösten tarkastelu voi myös olla hyödyllistä joissakin tapauksissa.

## 4.4 Devianssi

Devianssi on $-2$ kertaa log-uskottavuus määriteltynä parametrien funktiona

$$D(\theta) = -2 \log p(y|\theta).$$

Stanissa devianssille voidaan laskea posteriorijakauma samaan tapaan kuin malliparametreille. JAGSissa devianssin posteriorin määrittämiseksi tulee log-uskottavuus määritellä eksplisiittisesti mallissa. NIMBLEssä voi tehdä samoin tai määritellä log-uskottavuus erillisenä funktiona. Devianssia ei yleensä yritetä tulkita sellaisenaan, vaan sitä käytetään informaatiokriteereiden osana.

## 4.5 Informaatiokriteerit

Informaatiokriteereitä käytetään valittaessa parasta mallia tarjolla olevista ehdokkaista. Uutta aineistoa ei eksplisiittisesti ole käytössä, mutta malleja rangaistaan käytetystä tehokkaasta parametrien lukumäärästä. Tehokas parametrien lukumäärä ottaa huomioon parametrien välisen riippuvuuden. Jos malli parametrisoidaan toisella tavalla niin tehokas parametrien määrä voi muuttua.

**Devianssi-informaatiokriteeri** (*Deviance Information Criterion*, DIC) perustuu odotettuun devianssiin

$$\overline{D} = \mathrm{E}_{\theta|y}(D(\theta)) = \int -2 \log p(y|\theta) p(\theta|y)\, d\theta$$

ja posterioriodotusarvon devianssiin $D(\overline{\theta})$, missä $\overline{\theta} = \mathrm{E}(\theta|y)$. Parametrien tehokas määrä saadaan näiden erotuksena

$$p_D = \overline{D} - D(\overline{\theta}).$$

Jos osa parametreista on kategorisia, posterioriodotusarvon $\overline{\theta}$ määrittäminen on ongelmallista. DIC määritellään odotetun devianssin ja tehokkaan parametrien määrän summana

$$\begin{aligned} \mathrm{DIC} &= \overline{D} + p_D = 2\overline{D} - D(\overline{\theta}) \\ &= \int -4 \log p(y|\theta) p(\theta|y)\, d\theta - D(\overline{\theta}). \end{aligned}$$

Malli, jonka DIC-arvo on pienin, on kriteerin mukaan paras. Kriteeri voidaan esittää myös muodossa

$$\begin{aligned} \mathrm{DIC} &= \overline{D} + p_D = D(\overline{\theta}) + 2p_D \\ &= -2 \log p(y|\overline{\theta}) + 2p_D, \end{aligned}$$

joka muistuttaa frekventistisessä tilastotieteessä käytettyä Akaiken informaatiokriteeriä (*Akaike Information Criterion*, AIC), kun $\overline{\theta}$ tulkitaan suurimman uskottavuuden estimaatiksi ja $p_D$ parametrien lukumääräksi. JAGS-mallista voidaan generoida DIC:n otoksia funktiolla `dic.samples`.

**Watanabe-Akaike-informaatiokriteeri** (*Watanabe-Akaike Information Criterion* tai *Widely Available Information Criterion*, WAIC) käyttää odotetun devianssin ja posterioriodotusarvon sijaan havaintokohtaisia lausekkeita ja noudattaa täten Bayes-filosofiaa paremmin kuin DIC. Eräs WAIC-versio määritellään seuraavasti

$$\mathrm{WAIC} = -2 \sum_{i=1}^{n} \left[ 2 \log \left( \int p(y_i|\theta) p(\theta|y)\, d\theta \right) - \int \log(p(y_i|\theta)) p(\theta|y)\, d\theta \right].$$

Kun käytössä on otos posteriorista $\theta^{(1)}, \ldots, \theta^{(T)}$, WAIC voidaan estimoida kaavasta

$$\widehat{\mathrm{WAIC}} = -2 \sum_{i=1}^{n} \left[ 2 \log \left( \frac{1}{T} \sum_{t=1}^{T} p(y_i|\theta^{(t)}) \right) - \frac{1}{T} \sum_{t=1}^{T} \log(p(y_i|\theta^{(t)})) \right].$$

NIMBLE sisältää implementaation WAICin laskemiseksi, mutta se käyttää tehokkaan parametrien lukumäärän estimoinnissa edellisestä poiketen otosvariansseja, jolloin WAIC:in estimaattori on

$$\widehat{\text{WAIC}}_2 = -2\sum_{i=1}^{n}\left[\log\left(\frac{1}{T}\sum_{t=1}^{T}p(y_i|\theta^{(t)})\right) - \frac{1}{T-1}\sum_{t=1}^{T}\left(\log(p(y_i|\theta^{(t)}) - \overline{\log(p(y_i|\theta^{(t)}))}\right)^2\right],$$

missä $\overline{\log(p(y_i|\theta^{(t)}))} = \frac{1}{T}\sum_{t=1}^{T}\log(p(y_i|\theta^{(t)}))$.

## 4.6 Bayesin tekijä

Bayesin tekijää (*Bayes Factor*) voi käyttää kahden mallin vertailemiseen. Olkoot $M_1$ ja $M_2$ kaksi mallia, joiden prioritodennäköisyydet ovat $p(M_1)$ ja $p(M_2)$. Bayesin tekijäksi kutsutaan suhdetta

$$\text{B} = \frac{p(y|M_1)}{p(y|M_2)},$$

joka kertoo, kuinka paljon data muuttaa ennakkokäsityksiä. Mallien posterioritodennäköisyyksien suhde on

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(M_1)p(y|M_1)}{p(M_2)p(y|M_2)} = \frac{p(M_1)}{p(M_2)}\text{B}.$$

Bayesin tekijä ei ota huomioon parametrien määrää malleissa. Usein mallien keskiarvoistaminen tai laajentaminen on parempi vaihtoehto kuin mallin valinta Bayesin tekijän avulla.

## 4.7 Mallien keskiarvoistaminen

Yksi ratkaisu mallin valintaan on mallien keskiarvoistaminen. Tällöin sovitetaan kaikki mallit ja ennusteina käytetään painotettua keskiarvoa mallien antamista ennusteista. Olkoot $p(M_1), \ldots, p(M_K)$ mallien $M_1, \ldots, M_K$ prioritodennäköisyydet. Mallien posterioritodennäköisyyksiä

$$p(M_r|y) = \frac{p(M_r)p(y|M_r)}{\sum_{k=1}^{K}p(M_k)p(y|M_k)}$$

käytetään painoina laskettaessa posterioriennustejakaumaa

$$p(\tilde{y}|y) = \sum_{k=1}^{K}p(\tilde{y}|y, M_k)p(M_k|y).$$

## 4.8 Valintapriorit

Tarkastellaan mallinnustilannetta, jossa potentiaalisia kovariaatteja on paljon, mutta on syytä uskoa, että vain pieni määrä niistä on tärkeitä. Toisin sanoen kovariaattien uskotaan kuuluvan kahteen eri ryhmään, mutta etukäteen ei pystytä kertomaan mihin ryhmään kukin

kovariaatti kuuluu. Tällaisen ennakkotiedon kuvaamiseen voi käyttää valintaprioreita. Liitetään kuhunkin kovariaattiin $j = 1, \ldots, J$ indikaattorimuuttuja $I_j$, joka saa arvon 1 mikäli kyseessä on tärkeä selittäjä ja arvon 0 muutoin. Priorit määrätään siten, että todennäköisyys $p(I_j = 1)$, $j = 1, \ldots, J$ on pieni. Kovariaattien regressiokertoimille $\beta_j$, $j = 1, \ldots, J$ määritetään ehdolliset priorit

$p(\beta_j | I_j = 0)$ todennäköisyysmassa keskittyy origon ympäristöön (tai origoon)
$p(\beta_j | I_j = 1)$ todennäköisyysmassa leviää laajalle alueelle.

Priorijakaumasta tulee tällöin seosjakauma

$$p(\beta_j) = p(I_j = 1)p(\beta_j | I_j = 1) + p(I_j = 0)p(\beta_j | I_j = 0).$$

Malli voi olla esimerkiksi yleistetty lineaarinen malli. Posterioritodennäköisyys $p(I_j = 1|y)$ kertoo, onko kovariaatti $j$ tärkeä selittäjä vai ei. On mahdollista sovittaa toinen malli, johon otetaan mukaan vain kovariaatit, jotka ovat tärkeitä valintaposteriorin perusteella.

# 5 Regressiomalleja

## 5.1 Sensuroidut havainnot

Sensurointi on mahdollista ottaa huomioon posteriorin määrittämisessä. Tarkastellaan sensuroinnin käsittelyä esimerkkien avulla:

**Esimerkki 5.1** (Eksponenttijakautunut elinaika)**.** Oletetaan, että havaintoihin $y_1, \dots, y_n$ liittyvät seuraavat oletukset.

1. Havainnot noudattavat jakaumaa $\text{Exp}(\theta)$ jollakin $\theta > 0$.
2. Havainnot ovat ehdollisesti riippumattomia ehdolla $\theta$, ts. ne ovat vaihdannaisia.
3. Havainnoinnissa on tehty sensurointi oikealta tunnettuun arvoon $a > 0$: Jos $y_i \leq a$, havainto tehdään ja jos $y_i > a$, niin havainto on sensuroitu (niistä pidetään lukua).

Näistä oletuksista ensimmäinen koskee populaatiojakaumaa, kun taas oletukset 2–3 koskevat havainnointia. Tällöin saadaan

$$p(y|\theta) = \prod_{y_i \leq a} p(y_i|\theta) \times \prod_{y_i > a} S(a|\theta),$$

missä $S(a|\theta) = P(y_i > a|\theta)$ on elossaolofunktio. Siten

$$p(y_1, \dots, y_n|\theta) = \prod_{i=1}^{m} \left[\theta e^{-\theta y_i}\right] \times \left[e^{-\theta a}\right]^{n-m} = \theta^m e^{-\theta\left[\sum_{i=1}^{m} y_i + (n-m)a\right]},$$

missä $m$ on sensuroimattomien havaintojen lukumäärä (ja $n - m$ on sensuroitujen havaintojen lukumäärä).

Oletetaan, että $\theta$ noudattaa gammajakaumaa $\text{Gamma}(\alpha, \beta)$. Silloin

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

Priorin myötä syntyy kaksi uutta parametria, gammajakauman muotoparametri $\alpha$ ja käänteinen skaalaparametri $\beta$. Näille tulee asettaa priori: jos $\alpha$ ja $\beta$ oletetaan a priori riippumattomiksi, niin silloin on määriteltävä $p(\alpha)$ ja $p(\beta)$, muulloin yhteisjakauma $p(\alpha, \beta)$.

Oletetaan nyt, että $\alpha$ ja $\beta$ ovat a priori riippumattomia ja tunnettuja vakioita: $\alpha = A$ ja $\beta = B$. (Nämä ovat degeneroituneita jakaumia). Vaihdannaiset havainnot ovat $y_1, \ldots, y_m$ ja $n - m$ havaintoa on sensuroitu arvoon $\alpha > 0$. Tällöin

$$\begin{aligned} p(\theta|y) &\propto \theta^m e^{-\theta[\sum_{i=1}^m y_i + (n-m)a]} \theta^{A-1} e^{-B\theta} \\ &= \theta^{A+m-1} e^{-\theta[B + \sum_{i=1^m} y_i + (n-m)a]}, \end{aligned}$$

kun $\theta > 0$. Tämä jakauma on

$$\text{Gamma}(A + m, B + \sum_{i=1}^{m} y_i + (n - m)a).$$

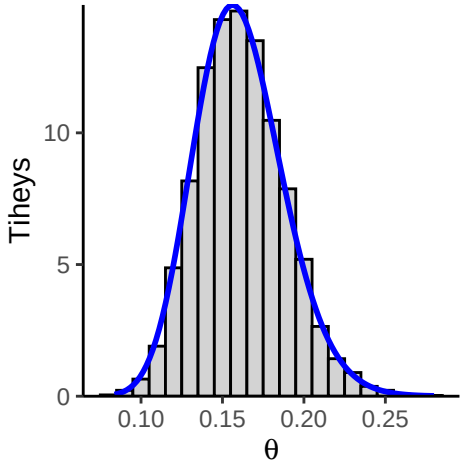
Tarkastellaan seuraavaksi aineistoa `device.dat`, joka sisältää 50 eri laitteen käyttöajat vuosissa. Mikäli laite on ollut käytössä yli 8 vuotta, ei sen todellista käyttöaikaa ole havaittu. Käyttöaikojen oletetaan noudattavan jakaumaa Exp($\theta$). Sovitetaan nyt edellä kuvailtu malli Stanilla.

```
scode <- "
data {
  int<lower=0> n; // havaintojen lukumäärä
  int<lower=0> m; // sensuroimattomien havaintojen lukumäärä
  real a;         // sensuroinnin raja
  real y[m];      // havainnot
}

parameters {
  real<lower=0> theta;
}

model {
  theta ~ gamma(1, 1);
  y ~ exponential(theta);
  target += (n-m) * exponential_lccdf(a | theta);
  // eksponenttijakauman komplementaarisen kumulatiivisen jakaumafunktion
  // eli eksponenttijakauman elossaolofunktion logaritmi
}
"

# Poimitaan sensuroimattomat havainnot erilleen
device_obs <- device[device$time < 8, ]

# Kaikkien ja sensuroimattomien havaintojen lukumäärät
n <- nrow(device)
m <- nrow(device_obs)

# Sensuroinnin raja datassa
```

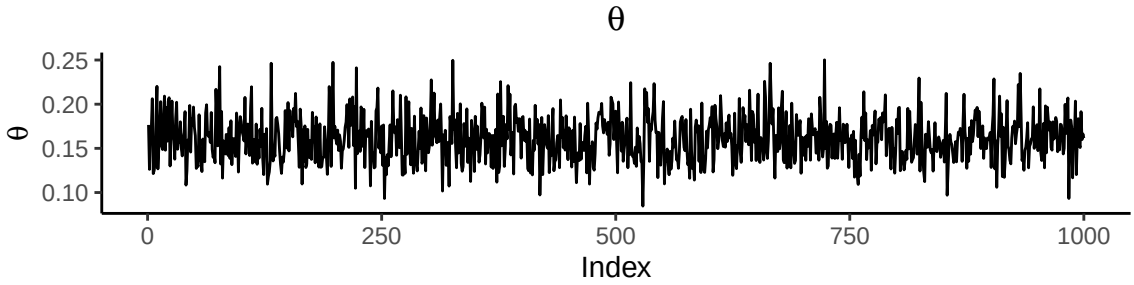
```
a <- 8.0

fit <- stan(
  model_code = scode,
  data = list(n = n, m = m, a = a, y = device_obs$time),
  iter = 2000,
  verbose = FALSE
)
```

Parametrin $\theta$ posteriorijakauman odotusarvo on 0.161 ja keskihajonta 0.027, ja 95 %:n posteriorivälikSi saadaan (0.114, 0.218).

Posteriorijakaumaotoksen histogrammi ja teoreettisen Gamma-posteriorin kuvaaja (sinisellä):




Parametrin $\theta$ otospolku lämmittelyjakson jälkeen yhdessä Stanin ketjuista on




**Esimerkki 5.2** (Kemoterapia-aineisto)**.** Tarkastellaan aineistoa, jossa kultakin potilaalta on mitattu elinaika hoidon aloittamisesta alkaen. Elinajat ovat osittain sensuroituja, ja koeasetelmassa on kaksi ryhmää, koeryhmä ja kontrolliryhmä. Kovariaattina on potilaan ikä. Kiinnostuksen kohteena on hoidon mahdollinen vaikutus elinaikaan. (`chemo.dat`)

On siis havaittu elinajat $t_1, \dots, t_n$, sensurointi $\delta_1, \dots, \delta_n$ sekä kovariaatit $x_{ij}, \dots, x_{nj}$, missä $j = 1, \dots, p$.

Sensurointimuuttuja $\delta_i$ määritellään sellaiseksi, että $\delta_i = 1$ jos havainto $i$ on tehty sensuroimattomana, muuten $\delta_i = 0$. Malli elinajoille on

$$\log t_i = \mu + \sum_{j=1}^{p} \beta_j x_{ij} + \sigma \epsilon_i,$$

Oletetaan, että virhetermit $\epsilon_i$ ovat riippumattomia ja noudattavat tyypin 1 Gumbel-jakaumaa, jonka tiheysfunktio on

$$p(\epsilon) = \exp(\epsilon - \exp(\epsilon)).$$

Gumbel-jakauman karakterisointeja:

1. log-Weibull -jakauma.
2. eräs GEV-perheen jakauma (generalized extreme value distribution).
3. kaksinkertainen eksponenttinen jakauma.

Mallissa on siis $p + 2$ parametria: $\mu, \sigma, \beta_1, \dots, \beta_p$. Merkitään:

- $y_i = \log t_i$
- $\beta = (\beta_1, \dots, \beta_p)$
- $z_i = (y_i - \mu - \sum_{j=1}^{p} \beta_j x_{ij})/\sigma$
- $p_i(y_i|\beta, \mu, \sigma) = \frac{1}{\sigma} \exp(z_i - \exp(z_i))$
- $S_i(y_i|\beta, \mu, \sigma) = \exp(-\exp(z_i))$ (elossaolofunktio)

Uskottavuudeksi saadaan

$$p(y|\beta, \mu, \sigma) = \prod_{i=1}^{n} \left[ p_i(y_i|\beta, \mu, \sigma)^{\delta_i} \, S_i(y_i)^{1-\delta_i} \right].$$

Oletetaan, että parametrit ovat a priori riippumattomat. Valitaan

- $p(\beta_j) \propto 1$, $j = 1, \dots, p$
- $p(\mu) \propto 1$
- $p(\sigma) \propto \frac{1}{\sigma}$

| Vuosikymmen | Onnettomuuksien lukumäärä |
|---|---|
| 1851–1860 | 4, 5, 4, 1, 0, 4, 3, 4, 0, 6 |
| 1861–1870 | 3, 3, 4, 0, 2, 6, 3, 3, 5, 4 |
| 1871–1880 | 5, 3, 1, 4, 4, 1, 5, 5, 3, 4 |
| 1881–1890 | 2, 5, 2, 2, 3, 4, 2, 1, 3, 2 |
| 1890–1900 | 1, 1, 1, 1, 1, 3, 0, 0, 1, 0 |
| 1901–1910 | 1, 1, 0, 0, 3, 1, 0, 3, 2, 2 |
| 1911–1920 | 0, 1, 1, 1, 0, 1, 0, 1, 0, 0 |
| 1921–1930 | 0, 2, 1, 0, 0, 0, 1, 1, 0, 2 |
| 1931–1940 | 2, 3, 1, 1, 2, 1, 1, 1, 1, 2 |
| 1941–1950 | 4, 2, 0, 0, 0, 1, 4, 0, 0, 0 |
| 1951–1960 | 1, 0, 0, 0, 0, 0, 1, 0, 0, 1 |
| 1961–1962 | 0, 0 |

Normeeraamaton posteriori on

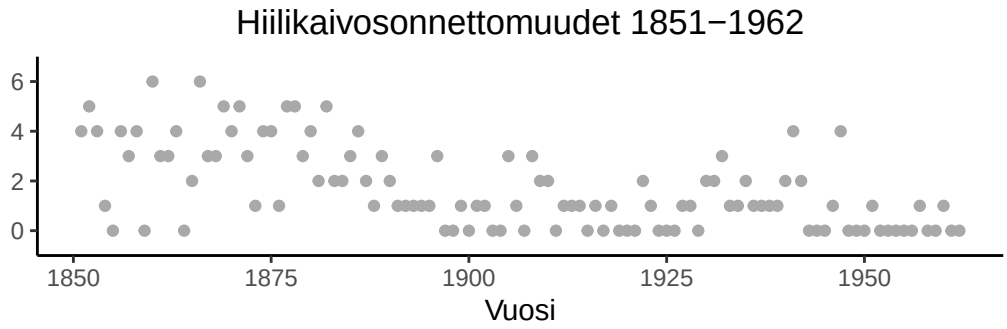
$$p(\beta, \mu, \sigma|y) \propto \frac{1}{\sigma} p(y|\beta, \mu, \sigma).$$

Posteriorijakauman avulla voidaan tutkia esimerkiksi käsittelyn vaikutusta, eri kovariaattien vaikutusta tai käsittelyryhmään kuuluvan henkilön selviytymistä. Mallin sovittaminen esimerkiksi JAGSissa vaatii erityistoimenpiteitä, sillä Gumbel-jakaumaa ei suoraan ole saatavilla. Tällöin voidaan nojata esimerkiksi ns. “nolla-yksi -temppuun”, joka esitellään luvussa @ref(sec:newdist). Stanissa Gumbel-jakauma on suoraan saatavilla, mutta sen parametrisointi poikkeaa edellä esitetystä. NIMBLEssä on mahdollista määritellä uusi jakauma.

## 5.2 Muutospisteongelma

Stokastisten prosessien ominaisuudet eivät välttämättä pysy vakioina ajan suhteen. Esimerkiksi prosessin odotusarvo, varianssi tai jonkin muu ominaisuus voi muuttua yllättäen. Muutospisteongelmalla tarkoitetaan yleisesti tilannetta, jossa jonkin prosessin parametrien osajoukon tiedetään muuttuvan vähintään kerran tietyn tarkastelujakson aikana. Tarkoituksena on tällöin selvittää ajanhetket, joissa muutokset tapahtuvat.

**Esimerkki 5.3** (Hiilikaivosonnettomuudet)**.** Seuraava aineisto kuvaa hiilikaivosonnettomuuksien määrää vuosina 1851–1962 ($n = 112$, `coal.dat`).




Havaitaan, että onnettomuuksien intensiteetti vuosina 1851–1882 on luokkaa 4 per vuosi, kun taas siitä eteenpäin tämä intensiteetti on hieman yli 1 per vuosi. Yksinkertainen malli on Poisson-malli, jossa intensiteetti muuttuu yhden kerran tarkastelujakson aikana. Muutoskohta pidetään tuntemattomana.

**Uskottavuus.** Olkoon $\lambda$ intensiteetti muutospisteeseen $k$ saakka ja $\mu$ siitä alkaen. Oletetaan, että on olemassa täsmälleen yksi muutospiste. Silloin $k \in \{1, \dots, n-1\}$, missä $n$ on havaintojen lukumäärä. Oletetaan edelleen, että

$$\begin{aligned} y_i|\lambda, \mu, k &\sim \text{Poisson}(\lambda), \quad i = 1, \dots, k, \quad \text{vaihdannaisia, ja} \\ y_i|\lambda, \mu, k &\sim \text{Poisson}(\mu), \quad i = k+1, \dots, n, \quad \text{vaihdannaisia.} \end{aligned}$$

Lisäksi $y_i$ ja $y_j$, $i \leq k, j > k$ ovat ehdollisesti riippumattomia ehdolla $\lambda, \mu, k$.

**Priorit.** Asetetaan prioreiksi

- $\lambda \sim \text{Gamma}(A, B)$
- $\mu \sim \text{Gamma}(C, D)$
- $k \sim \text{Tas}(\{1, \dots, n-1\})$

Parametrit $\lambda, \mu$ ja $k$ oletetaan lisäksi riippumattomiksi a priori. Valitaan heikosti informatiiviset priorit asettamalla $A = C = 0.01$ ja $B = D = 0.01$.

**Posteriori.** Edellä olevat valinnat johtavat normeeraamattomaan posterioriin

$$\begin{aligned} p(\lambda, \mu, k|y) &\propto p(k)p(\lambda)p(\mu)p(y|\lambda, \mu, k) \\ &\propto \lambda^{A+\sum_{i=1}^{k} y_i - 1} \mu^{C+\sum_{j=k+1}^{n} y_j - 1} e^{-[(B+k)\lambda + (D+(n-k))\mu]}. \end{aligned}$$

**Posteriorin simulointi MCMC-menetelmällä.** Suoraviivaisin algoritmi on komponenteittain tapahtuva päivittäminen, jossa yhdellä kierroksella päivitetään

- $\lambda$ Gibbsin algoritmilla
- $\mu$ edelleen Gibbsin algoritmilla
- $k$ Metropolisin algoritmilla.

Kyseessä on ns. “Metropolis within Gibbs” -lähestymistapa. Gibbsin algoritmia käytetään päivityksissä silloin, kun parametrien ehdolliset jakaumat ovat helposti laskettavissa. Muulloin käytetään Metropolisin päivitystä.

Toimitaan siis seuraavasti:

$$(\lambda, \mu, k) \to (\lambda^*, \mu, k) \to (\lambda^*, \mu^*, k) \to (\lambda^*, \mu^*, k^*).$$

**1. $\lambda$ :n päivitys**

Lasketaan ehdollinen jakauma $p(\lambda|\mu, k, y)$

$$p(\lambda|\mu, k, y) \propto \lambda^{A+\sum_{i=1}^{k} y_i - 1} e^{-(B+k)\lambda},$$

joka on Gamma$(A + \sum_{i=1}^{k} y_i, B + k)$.

**2. $\mu$ :n päivitys**

Lasketaan ehdollinen jakauma $p(\mu|\lambda, k, y)$

$$p(\mu|\lambda, k, y) \propto \lambda^{C+\sum_{j=k+1}^{n} y_j - 1} e^{-[D+(n-k)]\lambda},$$

joka on Gamma$(C + \sum_{j=k+1}^{n} y_j, D + (n - k))$.

**3. $k$ :n päivitys**

Tässä ei ole mahdollista päätyä yksinkertaiseen ehdolliseen jakaumaan $p(k|\lambda, \mu, y)$. Päädytään suosiolla Metropolis-Hastings -menetelmään. Ensin on päätettävä ehdotusjakauma, jonka arvojoukko on $\{1, \ldots, n-1\}$. Pyritään symmetriseen ehdotusjakaumaan (vaikka se ei olekaan tarpeen) yhdistämällä aika-akselin ensimmäinen ja viimeinen arvo:

$$q(k, k^*) = \begin{cases} \frac{1}{3} & \text{kun } k = k^*, \\ \frac{1}{3} & \text{kun } |k - k^*| = 1, \\ 0 & \text{muutoin,} \end{cases}$$

mutta kuitenkin siten, että

$$q((n-1), 1) = \frac{1}{3}$$
$$q(1, (n-1)) = \frac{1}{3}.$$

Hyväksymistodennäköisyys on $\alpha = \min(1, r)$, missä

$$r = \lambda^{\sum_{i=1}^{k^*} y_i - \sum_{j=1}^{k} y_j} \mu^{\sum_{i=k^*+1}^{n} y_i - \sum_{j=k+1}^{n} y_j} e^{-(k^*-k)\lambda + (k^*-k)\mu}.$$

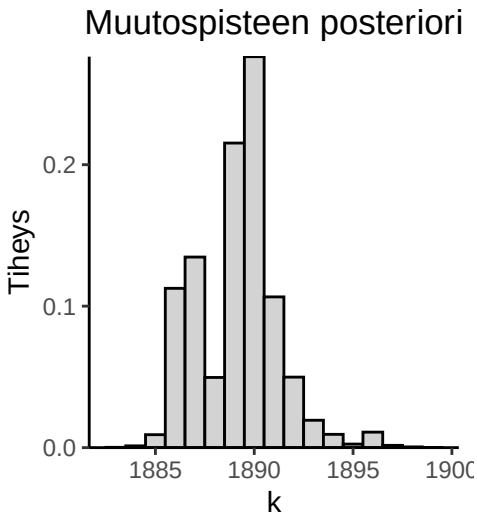
**Implementointi**

```r
disaster_gibbs <- function(y, n_sim, n_burn, theta_0, A, B, C, D) {
  n <- length(y)
  # Tässä theta = (lambda, mu, k)
  theta <- matrix(ncol = 3, nrow = n_sim + n_burn)
  theta[1, ] <- theta_0
  k <- theta[1, 3]
  for (i in 2:(n_sim + n_burn)) {
    lambda <- rgamma(1, A + sum(y[1:k]), rate = B + k)
    mu <- rgamma(1, C + sum(y[(k + 1):n]), rate = D + (n - k))
    k_star <- sample((k - 1):(k + 1), size = 1)
    if (k_star == 0) {
      k_star <- n - 1
    }
    else if (k_star == n) {
      k_star <- 1
    }
    r <- lambda^(sum(y[1:k_star]) - sum(y[1:k])) *
      mu^(sum(y[(k_star + 1):n]) - sum(y[(k + 1):n])) *
      exp(-(k_star - k) * lambda + (k_star - k) * mu)
    k <- ifelse(rbinom(1, size = 1, prob = min(1, r)), k_star, k)
    theta[i,] <- c(lambda,mu,k)
  }
  theta[seq(n_burn + 1, n_burn + n_sim), ]
}
dis_res <- disaster_gibbs(hiili, 2e4, 1e4, c(4, 1, 40), 0.01, 0.01, 0.01, 0.01)
dis_res[, 3] <- dis_res[, 3] + 1850 # muunnetaan erotus vuodesta 1850 vuosiluvuksi
```

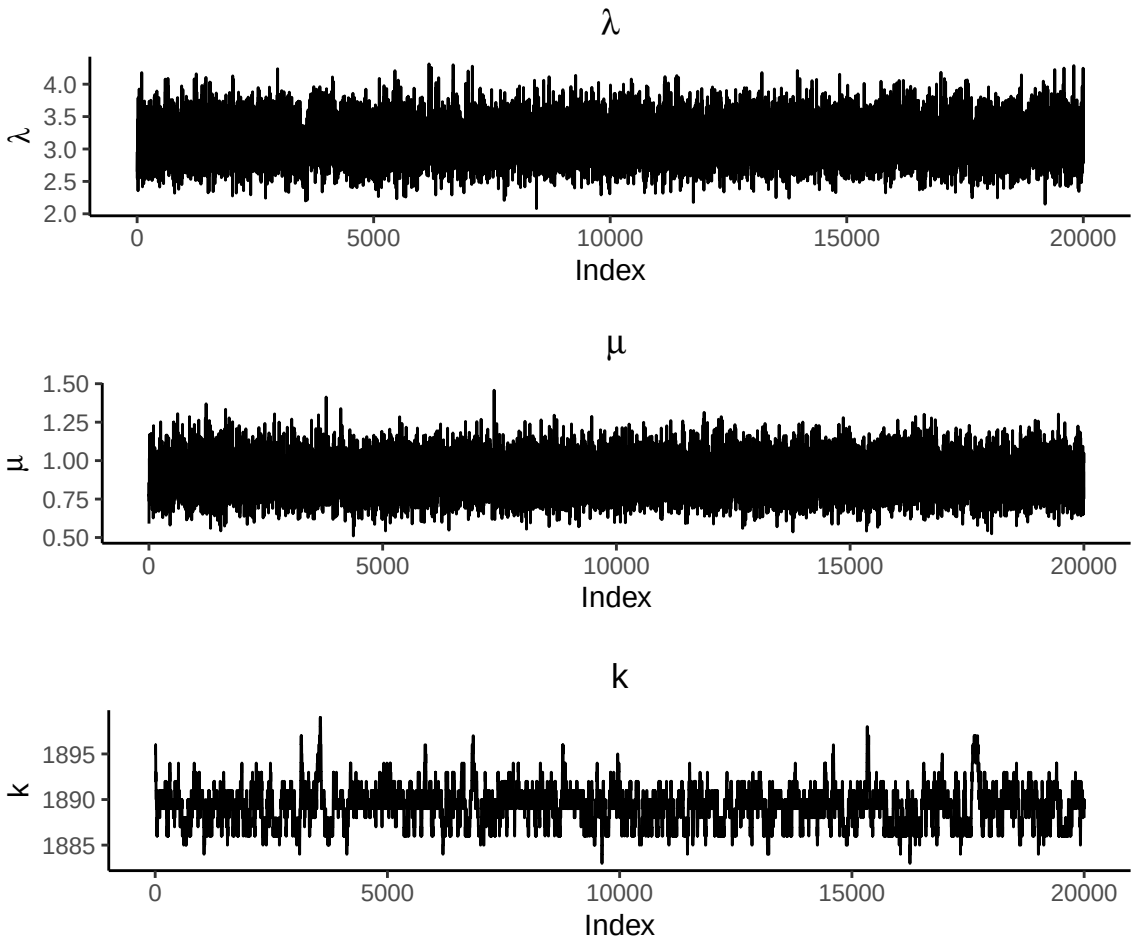
Simuloinnin tuloksena saadaan seuraavat tunnusluvut parametreille $\lambda, \mu$ ja $k$

| | Mean | SD | 2.5 % | 97.5 % |
|---|---|---|---|---|
| $\lambda$ | 3.1415781 | 0.2935152 | 2.5898433 | 3.739479 |
| $\mu$ | 0.8944647 | 0.1140691 | 0.6851683 | 1.129907 |
| $k$ | 1889.1730000 | 2.0584419 | 1886.0000000 | 1894.000000 |

Muutospisteen $k$ posterioritodennäköisyysjakauma on




Parametrien otospolut ovat



### Onnettomuusintensiteetin estimointi Stanilla

Edellistä mallia ei voi suoraan estimoida Stanilla, koska muutospisteparametri $k$ saa diskreettejä arvoja. Myös suora parametrin $k$ tulkitseminen reaaliarvoiseksi on ongelmallista, koska posterioritiheys riippuu $k$:n arvosta epäjatkuvasti. Sen sijaan voidaan määritellä lähes vastaava malli, missä onnettomuuksien määrää kuvaavan Poisson-jakautuneen muuttujan parametri riippuu jatkuvasti vuodesta ja parametrista $k$. Olkoon $y_i$ vuonna $t_i$ tapahtuneiden hiilikaivosonnettomuuksien lukumäärä. Määritellään malli

$$y_i \sim \text{Poisson}\left(\lambda + (\mu - \lambda) f\left(\frac{t_i - k}{\sigma}\right)\right),$$

missä $f(x)$ on sopiva jatkuva funktio, jolla $f(x) \approx 0$ kun $x \ll 0$ ja $f(x) \approx 1$ kun $x \gg 0$; valitaan tässä

$$f(x) = \frac{1}{1 + e^{-x}}$$

eli käänteinen logit-muunnos. Parametreilla $\lambda$, $\mu$ ja $k$ on sama tulkinta kuin edellä. Niiden lisäksi mallissa on mukana onnettomuusintensiteetin muutoksen nopeutta kuvaava parametri $\sigma$, jonka pienet arvot tarkoittavat jyrkkää muutosta. Valitsemalla heikosti informatiivisia prioreja parametreille päädytään malliin

$$\begin{aligned} y_i|\lambda, \mu, k, \sigma &\sim \text{Poisson}\left(\lambda + (\mu - \lambda) f\left(\frac{t_i - k}{\sigma}\right)\right) \\ \lambda &\sim \text{Puoli}-\text{N}(0, 10^2) \\ \mu &\sim \text{Puoli}-\text{N}(0, 10^2) \\ k &\sim \text{Tas}(1851, 1962) \\ \sigma &\sim \text{Puoli}-\text{N}(0, 2^2). \end{aligned}$$

Jyrkkyysparametrille $\sigma$ on annettu informatiivisempi, pieniä arvoja suosiva priori. Stanilla malli toteutetaan seuraavasti.

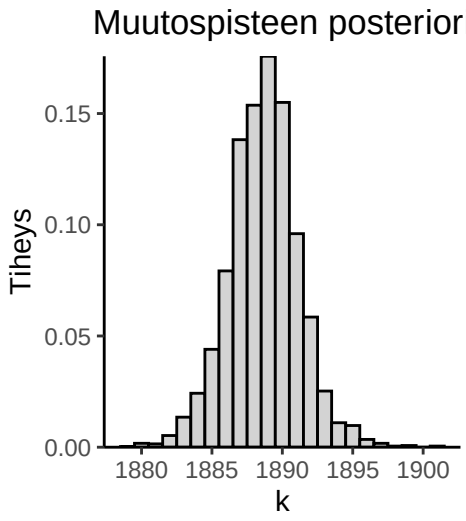
```
scode <- "
data {
  int<lower=0> n;    // vuosien määrä
  int<lower=0> y[n]; // onnettomuuksien määrä
  vector[n] year;    // vuosi
}

parameters {
  real<lower=min(year), upper=max(year)> k;
  real<lower=0> lambda;
  real<lower=0> mu;
  real<lower=0> sigma;
}

model {
  sigma ~ normal(0, 2);
```

```
  lambda ~ normal(0, 10);
  mu ~ normal(0, 10);
  y ~ poisson(lambda + (mu - lambda) * inv_logit((year - k) / sigma));
}
"

n <- length(hiili)
fit <- stan(
  model_code = scode,
  data = list(n = n, y = hiili, year = 1851:1962),
  iter = 3000,
  warmup = 2000,
  control = list(adapt_delta = 0.95),
  verbose = FALSE
)
```

Stan-mallin estimointi tuottaa seuraavat tunnusluvut

| | Mean | SD | 2.5 % | 97.5 % |
|---|---|---|---|---|
| $k$ | 1888.6797466 | 2.4663607 | 1883.6767283 | 1893.575823 |
| $\lambda$ | 3.2791552 | 0.3219494 | 2.6903658 | 3.968823 |
| $\mu$ | 0.8776066 | 0.1163778 | 0.6651125 | 1.124654 |
| $\sigma$ | 1.9713758 | 1.0991079 | 0.2475208 | 4.342358 |

Huomataan, että tulokset pitkälti vastaavat aikaisemmin saatuja. Muutospisteen $k$ posterioritodennäköisyysjakauma tässä mallissa on

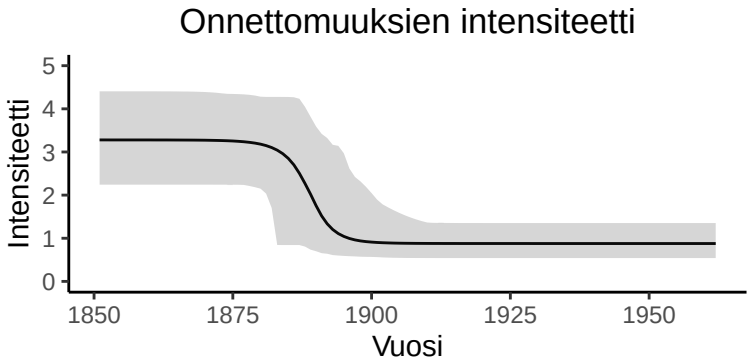



**Keskiarvoistaminen**

Analyysissa voidaan mennä vielä pidemmälle: lasketaan intensiteettifunktio ajan funktiona ja sille rajat (envelopes). Rajat kuvaavat väliä, jossa käyrä on suurella posterioritodennäköisyydellä. Tämä on alkeellinen esimerkki epäparametrisesta mallista, jossa yksinkertaista funktiota sovitetaan aineistoon suuri määrä ja niistä lasketaan keskiarvo sekä rajat.

Määritellään funktio, joka laskee intensiteetille minimit, maksimit sekä keskiarvon kullekin vuodelle

```
profile <- function(x) {
  res <- t(sapply(1:ncol(x), function(i) {
    c(
      mean = mean(x[, i]),
      min = min(x[, i]),
      max = max(x[, i]),
      year = 1850 + i
    )
  }))
  as.data.frame(res)
}
```




Intensiteetin mallinnuksessa voitaisiin mennä pidemmälle ja käyttää joustavampia parametrisiä funktioperheitä. Olisi esimerkiksi suoraviivaista käyttää useampaa kuin yhtä muutospistettä, tai lineaarista mallia log-intensiteetille. Vieläkin joustavampia funktioperheitä saataisiin sovitettua epäparametrisilla malleilla. Voitaisiin esimerkiksi määritellä, että onnettomuuksien log-intensiteetti eri vuosina on gaussinen vektori, jolla on epätriviaali kovarianssimatriisi eli log-intensiteetti mallinnettaisiin gaussisena prosessina.

## 5.3 Nollapaisutettu Poisson-jakauma

Tarkastellaan tilannetta, jossa muuttuja saa ei-negatiivisia kokonaislukuarvoja $\{0, 1, 2, ...\}$. Perusteltu malli "harvinaisille riippumattomille tapahtumille" on Poisson($\mu$) -jakauma.

Usein Poisson-malli ei ole sellaisenaan sopiva syystä, että arvo 0 esiintyy useammin kuin jakauma edellyttää. Esimerkkinä olkoon astmakohtaukset annetulla aikavälillä: On niitä,

jotka voivat saada allergiakohtauksen ja kohtausten määrä on Poisson-jakautunut, kun taas osa henkilöistä ei voi saada kohtausta lainkaan (ovat täysin terveitä). Kuitenkaan aineiston keruussa tätä erottelua ei voida tehdä. Komplikaationa on, että aineistossa on kahdenlaisia nollia: sellaisia, jotka eivät ole saaneet kohtausta (mutta voisivat saada) ja sellaisia, jotka eivät edes voi saada kohtausta (rakenteellisia nollia).

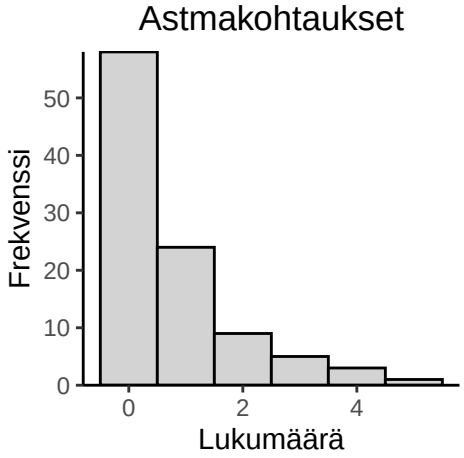
Tällaiseen tilanteeseen on ehdotettu nollapaisutettua Poisson-jakaumaa (*Zero-inflated Poisson Distribution*, ZIP):

$$p(k|\mu,\omega) = \begin{cases} (1-\omega) + \omega e^{-\mu} & \text{kun } k = 0, \\ \omega \frac{\mu^k e^{-\mu}}{k!} & \text{kun } k \geq 1. \end{cases}$$

Tässä parametrisoinnissa $\omega$ on todennäköisyys sille, että henkilö kuuluu siihen populaation osaan, joka *voi* sairastua. Tämä jakauma on seos jakaumista Poisson($\mu$) ja Poisson(0).

**Esimerkki 5.4** (Astmakohtaukset)**.** On annettu aineisto (astmakohtausten lukumäärä viimeisen kuukauden aikana) sadalta tutkimukseen osallistuneelta. Voidaan olettaa, että havainnot ovat vaihdannaisia (`asthma.dat`).

Havaitaan, että aineiston keskiarvo on 0.74 ja varianssi 1.2448, joten aineistossa on ylihajontaa.




Sovitetaan nollapaisutettu Poisson-jakaumamalli tähän aineistoon Stania käyttäen.

Oletetaan, että otosyksiköille $i$, joille on mahdollista, että $y_i > 0$:

$$\begin{aligned} y_i|\mu &\sim \text{Poisson}(\mu) \\ \mu &\sim \text{Gamma}(0.5, 0.01) \end{aligned}$$

Näiden otosyksiköiden osuus populaatiossa on $\omega$, jolle asetetaan

$$\omega \sim \text{Tas}(0, 1).$$

Uskottavuusfunktio voidaan siis kirjoittaa muodossa

$$p(y_i|\mu,\omega) = \begin{cases} (1-\omega) + \omega e^{-\mu} & \text{kun } y_i = 0, \\ \omega \frac{\mu^{y_i} e^{-\mu}}{y_i!} & \text{kun } y_i \geq 1, \end{cases}$$

joka voidaan implementoida suoraan Stanissa.

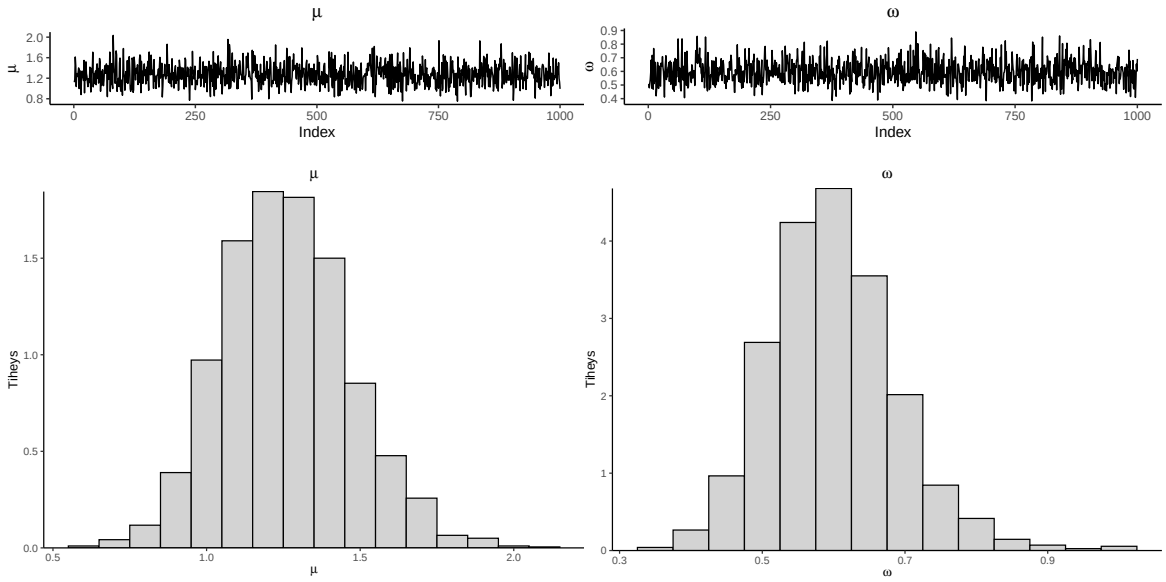
```
scode <- "
data {
  int<lower=0> n; // koehenkilöiden lukumäärä
  int y[n];       // astmakohtausten määrä
}

parameters {
  real<lower=0, upper=1> omega;
  real<lower=0> mu;
}

model {
  mu ~ gamma(0.5, 0.01);
  for (i in 1:n) {
    if (y[i] > 0) target += log(omega) + poisson_lpmf(y[i] | mu);
    else target += log((1-omega) + omega*exp(-mu));
  }
}
"

n <- length(astma)
fit <- stan(
  model_code = scode,
  data = list(n = n, y = astma),
  iter = 2000,
  verbose = FALSE
)
```

Simuloinnin tuloksena saadaan:

| | Mean | SD | 2.5 % | 97.5 % |
|---|---|---|---|---|
| $\mu$ | 1.2596638 | 0.2075539 | 0.8815041 | 1.6943848 |
| $\omega$ | 0.5990272 | 0.0895373 | 0.4415202 | 0.7945041 |

Parametrien otospolut yhdessä ketjussa ovat

Marginaaliposteriorit ovat




Voidaan sanoa, että sairastavien ryhmässä intensiteetin 95 %:n posterioriväli on (0.88, 1.69) ja populaatiosta 60 % kuuluu tähän ryhmään. Jälkimmäinen arvio on melko epätarkka: sen 95 %:n posterioriväli on (0.44, 0.79).

## 5.4 Moniulotteiset vasteet

Oletetaan, jokaiselta yksilöltä on saatu on $m$ mittausta. Olkoot indeksi $i = 1, \ldots, n$, joka liittyy yksilöön, ja $j = 1, \ldots, m$, joka liittyy havaintoon. Tällöin $y_{ij}$ on yksilöltä $i$ saatu mittaus $j$. Mikäli mittaukset saavat arvoja reaaliakselilta, voi normaalijakauma olla varteenotettava vaihtoehto havaintojen malliksi. Koska samalta yksilöltä tehdyt mittaukset ovat mahdollisesti korreloituneita, niin oletetaan että ne noudattavat moniulotteista normaalijakaumaa, missä odotusarvovektori $\mu$ ja kovarianssimatriisi $\Sigma$ ovat tuntemattomia:

$$y_i = (y_{i1}, \ldots, y_{im})^T | \mu, \Sigma \sim \mathrm{N}_m(\mu, \Sigma), \quad i = 1, \ldots, n.$$

Jos on lisäksi havaittu joukko kovariaatteja, kuten ikä mittaushetkellä, voidaan odotusarvon $\mu$ komponentteja kuvata monimuuttujaisella lineaarisella regressiomallilla:

$$\mu_j = \beta_0 + \sum_{k=1}^{p} \beta_k x_{kj}, \quad j = 1, \ldots, m,$$

missä $x_{kj}$ on $j$:s kovariaatin $k$ arvo. Tyypillisesti regressiokertoimille $\beta$ oletetetaan epäinformatiiviset normaalijakaumapriorit ja kovarianssimatriisille oletetaan käänteinen Wishart-jakauma, jolloin $\Sigma^{-1} \sim \mathrm{Wishart}(R, \rho)$ (Wishart-jakauma on moniulotteisen normaalijakauman käänteisen kovarianssimatriisin konjugaattipriori). Tässä $R$ on niin sanottu "skaalamatriisi" ja $\rho$ on jakauman vapausasteiden lukumäärä. Wishart-jakauma voidaan mieltää khiin-neliö -jakauman moniulotteiseksi yleistykseksi, joka ilmaantuu klassisessa tilastotieteessä neliösummien ja -tulojen matriisin jakaumana moniulotteisesta normaalijakaumasta

tehdyn otannan yhteydessä. Odotusarvovektoriin $\mu$ voidaan liittää muunkinlaisia parametrisointeja, mikäli mittaukset $y_{ij}$ eivät esimerkiksi ole mittauksia samasta vasteesta eri ajanhetkillä vaan mittauksia eri vasteista samalla ajanhetkellä. Mikäli mitatut vasteet muuttuvat ajassa, on myös mahdollista tarkastella ajassa muuttuvia kovariaatteja.

Vähiten informatiivinen aito Wishart-priori saadaan asettamalla $\rho = r$, missä $r$ on jakauman dimensio (tyypillisesti $r = m$). Prioriodotusarvo on $\rho R^{-1}$, jolloin hyvä valinta matriisiksi $R$ on $\rho\Sigma_0$, missä $\Sigma_0$ on jokin ennakkoarvaus kovarianssimatriisiksi.

Monimuuttujainen lineaarinen regressio voidaan helposti yleistää epälineaarisiin regressionmalleihin, aivan kuten yksiulotteisessa tapauksessa. Virhetermeille voidaan määrittää moniulotteinen $t$-jakauma poikkeavia havaintoyksiköitä varten. Priorijaukaumien valinta ei kuitenkaan ole aivan yhtä suoraviivaista. Kovarianssimatriisin tulee olla positiivisesti definiitti, ja Wishart-jakauma on ainoa tyypillisesti käytetty jakauma, joka toteuttaa tämän rajoitteen luonnollisella tavalla. (Tosin esim. Stanissa on implementoitu Lewandowski–Kurowicka–Joe-jakauma (LKJ) korrelaatiomatriiseille, jonka pohjalta voidaan rakentaa prioreja kovarianssimatriiseille.) Jos halutaan käyttää priorijakaumaa, jossa tämä kovarianssimatriisia koskeva rajoite ei suoraan toteudu, on se otettava mallissa huomioon muilla keinoin.

**Esimerkki 5.5** (Leukaluut)**.** Elston ja Grizzle (1962) esittävät aineiston, jossa 20 pojalta on mitattu leukaluun korkeudet 8, 8.5, 9 ja 9.5 vuosien iässä. Tarkoituksena on mallintaa keskimääräistä kasvukäyrää.

Yksinkertainen lineaarinen malli voidaan sovittaa Stanilla seuraavasti:

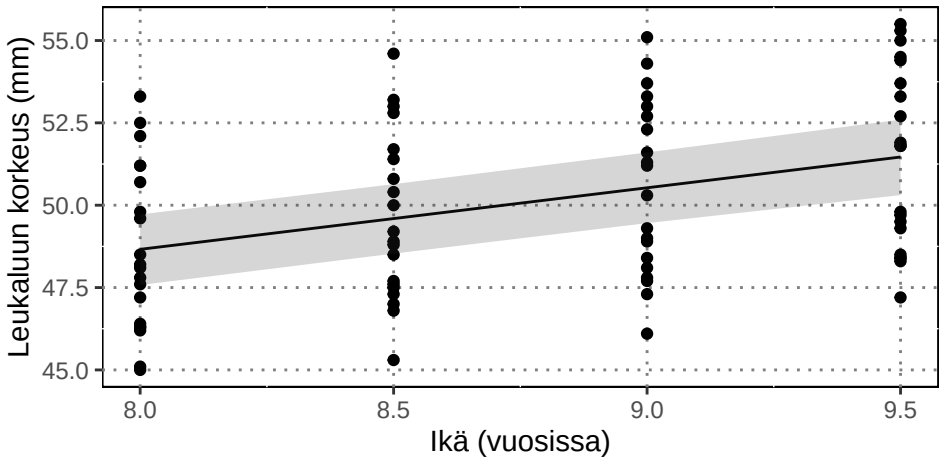
```
scode <- "
data {
  int<lower=0> n; // mittausten lukumäärä
  int<lower=0> m; // kovariaattien määrä
  vector[m] y[n]; // mittaukset
  vector[m] x;    // kovariaatit
}

parameters {
  real beta[2];
  cov_matrix[m] Sigma;
}

transformed parameters {
  vector[m] mu;
  mu = beta[1] + beta[2]*x;
}

model {
  Sigma ~ inv_wishart(m, diag_matrix(rep_vector(1, m)));
  y ~ multi_normal(mu, Sigma);
}
```

```
"

fit <- stan(
  model_code = scode,
  data = list(n = nrow(jaws), m = ncol(jaws), y = jaws, x = c(8, 8.5, 9, 9.5)),
  iter = 2000,
  verbose = FALSE
)
```

Simuloinnin tuloksena saadaan:

| | Mean | SD | 2.5 % | 97.5 % |
|---|---|---|---|---|
| $\beta_0$ | 33.714609 | 1.9957438 | 29.780970 | 37.642819 |
| $\beta_1$ | 1.868243 | 0.2265965 | 1.419098 | 2.308369 |
| $\mu_1$ | 48.660556 | 0.5519952 | 47.576506 | 49.720066 |
| $\mu_2$ | 49.594678 | 0.5413184 | 48.530595 | 50.639149 |
| $\mu_3$ | 50.528799 | 0.5540988 | 49.456173 | 51.601688 |
| $\mu_4$ | 51.462921 | 0.5888109 | 50.306480 | 52.597680 |

Kuviossa ovat alkuperäiset havainnot, keskiarvosovite sekä 95 % posterioriväli.




## 5.5 Regularisointi

Tarkastellaan lineaarista regressiomallia

$$y = X\beta + \epsilon,$$

missä

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{2n} & \cdots & x_{nm} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Tällöin PNS-ratkaisun estimointiyhtälöt ovat (Gauss-Markovin lauseen nojalla)

$$X^\top X\beta = X^\top y.$$

Tämän yhtälön ratkaisu $\hat{\beta} = (X^\top X)^{-1}X^\top y$ edellyttää, että matriisi $X^\top X$ on täysiasteinen, jolloin käänteismatriisi $(X^\top X)^{-1}$ on olemassa.

On tilanteita, joissa matriisi $X^\top X$ on "lähes" singulaarinen, esimerkiksi silloin, kun matriisin $X$ sarakkeissa on (yksi tai useampi) vahvasti korreloiva muuttujapari. Tämä aiheuttaa ongelmia estimoinnissa.

Yksi ratkaisu ongelmaan on ridge-regressio (harjanneregressio), jossa matriisi $X^\top X$ regularisoidaan lisäämällä sen diagonaalille positiivisia alkioita. Ridge-estimaattori on

$$\hat{\beta}_R = (X^\top X + \lambda I)^{-1}X^\top y.$$

Tämä estimaattori on harhainen, mutta estimaattorin komponenteilla voi olla pienemmät keskivirheet kuin PNS-estimaattorin komponenteilla.

Ridge-regressiolle voidaan löytää bayesiläinen formulointi, joka auttaa tulkinnassa.

**Ridge-regression Bayes-formulointi.** Oletetaan, että varianssi $1/\tau$ on tunnettu. Tällöin

$$p(y|\beta) \propto \exp\left\{-\frac{1}{2}\tau(y - X\beta)^\top(y - X\beta)\right\}.$$

Asetetaan prioriksi:

$$p(\beta) = \left(\frac{\kappa}{2\pi}\right)^{-M/2} \exp\left\{-\frac{1}{2}\kappa\beta^\top\beta\right\},$$

eli siis $\beta \sim \mathrm{N}_m(0, I/\kappa)$, missä $\kappa$ on tunnettu vakio. Tällöin

$$\log p(\beta|y) \propto -\frac{1}{2}\left[\tau(y - X\beta)^\top(y - X\beta) + \kappa\beta^\top\beta\right].$$

Derivoimalla edellinen lauseke vektorin $\beta$ komponenttien suhteen saadaan

$$\frac{\partial \log p}{\partial \beta} = \tau X^\top y - \tau X^\top X\beta - \kappa\beta.$$

Asettamalla derivaatta nollaksi saamme yhtälön

$$(X^\top X + \frac{\kappa}{\tau}I)\beta = X^\top y,$$

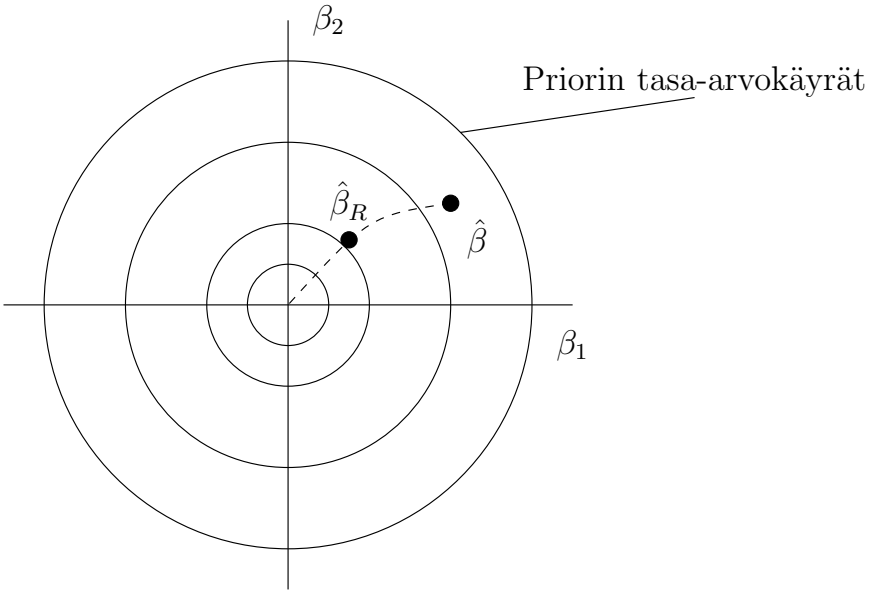
jonka ratkaisu on

$$\overline{\beta} = (X^\top X + \lambda I)^{-1} X^\top y, \text{ missä } \lambda = \frac{\kappa}{\tau}.$$

Posteriori on $m$-ulotteinen normaalijakauma ja sen moodi (ja myös odotusarvo) on $\overline{\beta}$.

Tulos: Harjanne-estimaattori on posterioriodotusarvo silloin, kun priori on $\beta \sim \mathrm{N}(0, I/\kappa), \kappa = \lambda\tau$.

Priorilla on "kutistava" vaikutus:




Kuva 5.1: Havainnollistus ridge-regressiosta.

## 5.6 Puuttuva tieto

Lähes kaikki käytännön aineistot sisältävät puuttuvaa tietoa. Havaintoyksiköltä voi esimerkiksi puuttua joidenkin muuttujien arvoja, tai havaintoyksikkö voi puuttua kokonaan. Puuttuvaa tietoa syntyy myös sensuroinnin seurauksena.

Oletetaan, että data on $y$. Se jaetaan kahteen osaan, havaittuun dataan $y^{\text{obs}}$ ja puuttuvaan tietoon $y^{\text{mis}}$, eli $y = (y^{\text{obs}}, y^{\text{mis}})$. Klassisessa tilastotieteessä käytetään pääasiassa neljää tapaa puuttuvan tiedon käsittelyssä:

1. Puuttuva tieto jätetään pois (naiivi menetelmä).
2. Puuttuva tieto imputoidaan yksinkertaisesti, esim. keskiarvolla.
3. Käytetään moni-imputointia (Rubin).
4. Mallinnetaan puuttuvuusmekanismi ja sovelletaan EM-algoritmia (*Expectation Maximization*) tai jotain muuta menetelmää.

Näistä vaihtoehto (1) johtaa lähes aina virheellisiin päätelmiin, mutta voi toimia joissain harvoissa erikoistapauksissa, vaihtoehto (2) vähintäänkin aliarvioi todellista vaihtelua, (3) on monipuolinen vaihtoehto, mutta voi olla usein hankala toteuttaa käytännössä ja (4) sopii suurimman uskottavuuden lähestymistavan yhteyteen ja toimii oikein.

Periaatetasolla puuttuvan tiedon käsittely Bayes-lähestymistavassa on varsin yksinkertaista. Puuttuva tieto $y^{\text{mis}}$ voidaan ajatella tuntemattomaksi parametriksi, aivan kuten mallin parametrit $\theta$ tai muut havaitsemattomat suureet. Siksi on luonnollista pyrkiä konstruoimaan posteriori

$$p(\theta, y^{\text{mis}}|y^{\text{obs}}).$$

Silloin parametrien $\theta$ marginaaliposteriori on

$$p(\theta|y^{\text{obs}}) = \int p(\theta, y^{\text{mis}}|y^{\text{obs}})\, dy^{\text{mis}}.$$

MCMC-menetelmillä posteriori saadaan tietysti käsittelemällä vain $\theta$:sta tehtyjä simulointeja. Posteriori $p(\theta, y^{\text{mis}}|y^{\text{obs}})$ on kuitenkin simuloitava!

Sivutuotteena puuttuvasta tiedosta saadaan informaatiota marginaaliposteriorijakauman $p(y^{\text{mis}}|y^{\text{obs}})$ kautta, sillä siihen liittyvä epävarmuus tiedetään.

Oletetaan, että täydellinen aineisto on

$$y = (y_1, \dots, y_n),$$

ja mallina on $p(y|\theta)$. Otetaan käyttöön indikaattorifunktio

$$I = (I_1, \dots, I_n),$$

missä

$$I_i = \begin{cases} 1 & \text{jos } y_i \text{ havaitaan,} \\ 0 & \text{muuten.} \end{cases}$$

Esimerkiksi jos $y_1, y_2, y_5, y_6$ ovat havaitut (aikasarjan) arvot ja havainnot $y_3, y_4$ ovat puuttuvia, niin silloin:

Täydellinen aineisto: $y_1, \dots, y_6$

Havaittu aineisto: $y_1, y_2, y_5, y_6$ ja $I = (1, 1, 0, 0, 1, 1)$.

Oletetaan edelleen, että

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

missä oikean puolen ensimmäinen termi on täydellisen aineiston uskottavuus. Toinen termi on puuttuvuuden malli ja $\phi$ sisältää siihen liittyvät parametrit. Oletetaan siis, että puuttuminen ei riipu parametreista $\theta$.

Lisäksi oletetaan, että jokainen $I_i, i = 1, \dots, n$ havaitaan. Tällöin havaittujen suureiden uskottavuus on

$$p(y^{\text{obs}}, I|\theta, \phi) = \int p(y^{\text{obs}}, y^{\text{mis}}|\theta)p(I|y^{\text{obs}}, y^{\text{mis}}, \phi)\, dy^{\text{mis}}.$$

Puuttuvien havaintojen luonne määrä sen, minkälainen uskottavuus lopulta on.

Vaikka puuttuvan tiedon käsittely Bayes-tilastotieteessä on ajatuksellisesti yksinkertaista, ei se ole välttämättä sitä käytännössä. Puuttuvuudesta täytyy olla tietoa, jotta puuttuvuusmekanismi voidaan mallintaa.

Rubin (1987) määrittelee kolme puuttumisen muotoa:

- Täysin satunnainen puuttuminen (*Missing Completely At Random*, MCAR, ), jos

$$p(I|y^{\text{obs}}, y^{\text{mis}}, \phi) = p(I|\phi).$$

- Satunnainen puuttuminen (*Missing At Random*, MAR), jos

$$p(I|y^{\text{obs}}, y^{\text{mis}}, \phi) = p(I|y^{\text{obs}}, \phi).$$

- Ei-satunnainen (informatiivinen) puuttuminen (*Missing Not At Random*, MNAR), jos $p(I|y^{\text{obs}}, y^{\text{mis}}, \phi)$ riippuu sekä havaitusta aineistosta $y^{\text{obs}}$ että puuttuvasta aineistosta $y^{\text{mis}}$.

Satunnaisen puuttumisen (MAR) tapauksessa uskottavuus on

$$\begin{aligned} p(y^{\text{obs}}, I|\theta, \phi) &= \int p(y^{\text{obs}}, y^{\text{mis}}|\theta) p(I|y^{\text{obs}}, y^{\text{mis}}, \phi)\, dy^{\text{mis}} \\ &= p(I|y^{\text{obs}}, \phi) \int p(y^{\text{obs}}, y^{\text{mis}}|\theta)\, dy^{\text{mis}} \\ &= p(I|y^{\text{obs}}, \phi) p(y^{\text{obs}}|\theta). \end{aligned}$$

Jos oletetaan, että $p(\theta, \phi) = p(\theta)p(\phi)$, niin posteriori on silloin

$$\begin{aligned} p(\theta, \phi|y^{\text{obs}}, I) &= \frac{p(\theta)p(\phi)p(y^{\text{obs}}, I|\theta, \phi)}{\int p(\theta)p(\phi)p(y^{\text{obs}}, I|\theta, \phi)\, d\theta d\phi} \\ &= \frac{p(\theta)p(y^{\text{obs}}|\theta)}{\int p(\theta)p(y^{\text{obs}}|\theta)\, d\theta} \frac{p(\phi)p(I|y^{\text{obs}}, \phi)}{\int p(\phi)p(I|y^{\text{obs}}, \phi)\, d\phi} \\ &= p(\theta|y^{\text{obs}})p(\phi|I, y^{\text{obs}}). \end{aligned}$$

Siten parametrien $\theta$ marginaaliposteriori on

$$p(\theta|y^{\text{obs}}, I) = p(\theta|y^{\text{obs}}) \int p(\phi|I, y^{\text{obs}})\, d\phi = p(\theta|y^{\text{obs}}).$$

**Seuraus:** parametreja $\theta$ koskeva Bayes-inferenssi perustuu posterioriin

$$p(\theta|y^{\text{obs}}) \propto p(\theta)p(y^{\text{obs}}|\theta).$$

Sanotaan, että puuttuvuusmekanismi on *epäoleellinen* (*ignorable*). Käytännössä tämä merkitsee sitä, että puuttumista ei tässä tapauksessa tarvitse ottaa huomioon.

## Posteriorin laskenta ja moni-imputointi

Posteriori on (yleisessä muodossa)

$$p(\theta, \phi, y^{\text{mis}}|y^{\text{obs}}).$$

Sen simulointi MCMC-menetelmällä johtaa simuloituihin arvoihin $(\theta^{(t)}, \phi^{(t)}, {y^{\text{mis}}}^{(t)})$, $t = 1, \ldots, T$, josta saadaan approksimaatio marginaaliposteriorille $p(\theta|y^{\text{obs}})$ tarkastelemalla vain simuloitua jonoa $\{\theta^{(t)}\}$. Puuttuvien arvojen $y^{\text{mis}}$ laskemista simuloinneista $\{{y^{\text{mis}}}^{(t)}\}$ sanotaan *moni-imputoinniksi*.

On syytä huomata *imputoinnin* ja *moni-imputoinnin* ero: Moni-imputoinnissa pyritään tuottamaan useita arvoja puuttuville havainnoille $y^{\text{mis}}$. Tämä toteutuu Bayes-lähestymistavassa luonnollisella tavalla.

**Esimerkki 5.6** (Sensurointi puuttuvana tietona)**.** Oletetaan, että $y_i|\theta \sim \text{Exp}(\theta)$, $i = 1, \ldots, n$, ehdollisesti riippumattomia ehdolla $\theta$. Havainto $y_i$ puuttuu, jos $y_i > y_0$. Tällöin

$$I_i = \begin{cases} 1 & \text{jos } y_i \leq y_0 \\ 0 & \text{jos } y_i > y_0. \end{cases}$$

Täydellinen uskottavuus on

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

kun taas havaittu uskottavuus on

$$p(y^{\text{obs}}, I|\theta, y_0) = \prod_{i=1}^{n} \left[ p(y_i|\theta)^{I_i} P(y_i > y_0|\theta, y_0)^{1-I_i} \right].$$

Tästä voidaan päätellä, että $p(I|y^{\text{obs}}, y^{\text{mis}}, y_0)$ riippuu puuttuvista havainnoista, joten otanta ei ole epäoleellinen.

**Esimerkki 5.7** (Sensurointi puuttuvana tietona. (jatk.))**.** Oletetaan, että havainnointi lopetetaan, kun $m_0 < n$ havaintoa on tehty, missä $m_0$ ei riipu havainnoista $y_i$. Oletetaan katkaisumäärä $m_0$ tunnetuksi. Katkaisu on siis havaintosarjassa riippumatta havaintojen $y_i$ arvoista. Tällöin

$$y^{\text{obs}} = (y_1, \ldots, y_{m_0})$$
$$y^{\text{mis}} = (y_{m_0+1}, \ldots, y_n).$$

Edelleen $I_i = 1$, kun $i = 1, \ldots, m_0$ ja $I_i = 0$, kun $i = m_0 + 1, \ldots, n$. Havaittuun tietoon perustuva uskottavuus on

$$p(y^{\text{obs}}, I|\theta, m_0) = \prod_{i=1}^{n} p(y_i|\theta)^{I_i}.$$

Tällöin $p(I|y^{\text{obs}}, y^{\text{mis}}, m_0) = p(I|m_0)$. Kyseessä on täydellisesti satunnainen puuttuminen (MCAR).

**Esimerkki 5.8** (Puuttuva suurin havainto)**.** Oletetaan, että vaihdannaisista havainnoista $y_1, \dots, y_n$, $y_{\max} = \max\{y_1, \dots, y_n\}$ puuttuu, ts. $n-1$ havaintoarvoa on saatu ja tiedetään, että suurin havainto puuttuu. Minkälaisesta puuttumisesta on kyse? Pohdi, voidaanko tämä mallintaa!

## 5.7 Aikasarjamallinnus

**Esimerkki 5.9** (AR-malli)**.** Oletetaan, että $y_0, y_1, \dots, y_n$ on aikasarja, jolle

$$y_0|\sigma_0^2 \sim \mathrm{N}(0, \sigma_0^2)$$
$$y_i|y_{i-1}, \dots, y_1, y_0, \alpha, \sigma^2 \sim \mathrm{N}(\alpha y_{i-1}, \sigma^2),\ i = 1, \dots, n.$$

Tutumpi esitys on

$$y_i = \alpha y_{i-1} + \epsilon_i,\ i = 1, \dots, n,$$

missä $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ ovat riippumattomia. Tämä on AR(1)-prosessi. Uskottavuusfunktio on muotoa

$$p(y|\alpha, \sigma^2) = p(y_0) \prod_{i=1}^{n} p(y_i|y_{i-1}, \alpha, \sigma^2),$$

kun oletetaan, että $\sigma_0^2$ on tunnettu. Havainnot eivät ole vaihdannaisia.

Oletetaan nyt, että havainto $y_i$ puuttuu riippumatta sen arvosta. Silloin

$$y^{\mathrm{obs}} = (y_0, \dots, y_{i-1}, y_{i+1}, \dots, y_n) = y_{-i}$$
$$y^{\mathrm{mis}} = y_i.$$

Tuntematon parametri on $\theta = (\alpha, \sigma^2)$. Sille asetetaan priori $p(\theta)$. Posteriori on

$$p(\theta, y^{\mathrm{mis}}|y^{\mathrm{obs}}) = p(\theta, y_i|y_{-i}).$$

Simulointi on mahdollista (ja yksinkertaista) tehdä komponenteittain: Oletetaan, että on simuloitu $(\theta^{(k)}, y_i^{(k)})$: $y_i^{(k+1)}$ simuloidaan jakaumasta $p(y_i|\theta^{(k)}, y_{-i})$ ja tämän jälkeen $\theta^{(k+1)}$ jakaumasta $p(\theta|y_i^{(k+1)}, y_{-i})$.

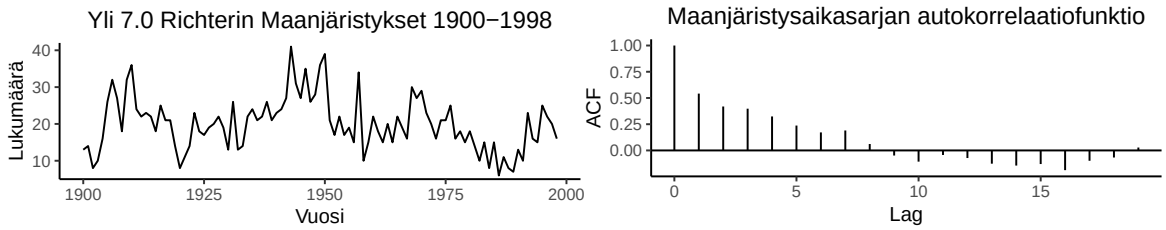
Huomaa, että

$$p(y_i|y_{-i}, \theta) = \frac{p(y_i|y_{i-1}, \theta)p(y_{i+1}|y_i, \theta)}{\int p(y_i|y_{i-1}, \theta)p(y_{i+1}|y_i, \theta)\, dy_i}$$

ja

$$p(\theta|y_{-i}, y_i^{(k+1)})$$

on posteriori parametrille $\theta$ ehdolla data, jossa puuttuva arvo $y_i$ on korvattu edellisen päivityksen antamalla arvolla. Ensimmäinen askel on helppo tehdä lasketusta jakaumasta, toisen voi tehdä Metropolisin algoritmilla. NIMBLEssä riittää korvata puuttuva havainto `NA`:lla.




**Esimerkki 5.10** (Maanjäristykset)**.** Seuraava aineisto esittää vuosihavaintona 7.0 Richterin ylittävien maanjäristysten lukumäärää vuosilta 1900-1998 (99 havaintoa, `earthq.dat`).

Aineiston kuvaaja ja autokorrelaatiofunktio ovat:

Aikasarjaa voidaan mallintaa gaussisena AR(1)-prosessina

$$y_i - m = \alpha(y_{i-1} - m) + \epsilon_i,\ i = 1, \ldots, n,$$

missä $m$ on aikasarjan odotusarvo. Suoritetaan Bayes-mallinnus tälle aikasarjalle seuraavin oletuksin:

Odotusarvo $m$ estimoidaan aineistosta

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

ja

$$\begin{aligned} y_1 &\sim \mathrm{N}(\hat{m}, 1/0.03) \\ y_i|\alpha, \tau, y_{i-1} &\sim \mathrm{N}(\hat{m} + \alpha(y_{i-1} - \hat{m}), 1/\tau),\ i = 2, \ldots, n \\ \alpha &\sim \mathrm{N}(0.5, 1/5) \\ \tau &\sim \mathrm{Gamma}(0.3, 1), \end{aligned}$$

missä $\tau = 1/\sigma^2$. Mallin Stan-koodi on

```
scode <- "
data {
  int<lower=0> n; // vuosien lukumäärä
  real y[n];      // havainnot
  real m_hat;     // prosessin keskiarvo
}

parameters {
  real alpha;
  real<lower=0> tau;
}

transformed parameters {
  real<lower=0> sigma;
  sigma = 1 / sqrt(tau);
```

```
}

model {
  alpha ~ normal(0.5, 1 / sqrt(5));
  tau ~ gamma(0.3, 1);

  y[1] ~ normal(m_hat, 1 / sqrt(0.03));
  for (i in 2:n) {
    y[i] ~ normal(m_hat + alpha * (y[i - 1] - m_hat), sigma);
  }
}
"

fit <- stan(
  model_code = scode,
  data = list(n = length(earthq), m_hat = mean(earthq), y = earthq),
  iter = 2000,
  verbose = FALSE
)
```

Posteriorijakauman kuvaus marginaaliposteriorien tunnusten avulla on seuraava:

| | Mean | SD | 2.5 % | 97.5 % |
|---|---|---|---|---|
| $\alpha$ | 0.54103 | 0.0844993 | 0.3732556 | 0.7071583 |
| $\sigma^2$ | 37.82581 | 5.6222920 | 28.4887300 | 50.6538331 |

Vertailun vuoksi:

$$\begin{aligned} \text{Yule-Walker}: &\quad \hat{\alpha} = 0.547 \quad \hat{\sigma}^2 = 37.66 \\ \text{SU-menetelmä}: &\quad \hat{\alpha} = 0.543 \quad \hat{\sigma}^2 = 36.7 \end{aligned}$$

**Esimerkki 5.11** (Maanjäristykset (jatk.))**.** Poistetaan havainto no. 93 (arvo 23). Malli on sama kuin edellä. Lasketaan aikasarjan tasoparametri $m$ vain havaituista arvoista.

Stanissa puuttuva havainto koodataan parametriksi. Tässä esimerkissä voidaan jakaa data kahteen osioon: havaintoihin $y_1, \ldots, y_{92}$ ennen puuttuvaa havaintoa ja havaintoihin $y_{94}, \ldots, y_{99}$ puuttuvan havainnon jälkeen.

```
scode2 <- "
data {
  int<lower=0> n1; // vuosien lukumäärä ennen puuttuvaa havaintoa
  int<lower=0> n2; // vuosien lukumäärä puuttuvan havainnon jälkeen
  real y1[n1];     // havainnot ennen puuttuvaa havaintoa
  real y2[n2];     // havainnot puuttuvan havainnon jälkeen
  real m_hat;      // prosessin keskiarvo
```

```
}

parameters {
  real alpha;
  real<lower=0> tau;
  real y_mis; // puuttuva havainto
}

transformed parameters {
  real<lower=0> sigma;
  sigma = 1 / sqrt(tau);
}

model {
  alpha ~ normal(0.5, 1 / sqrt(5));
  tau ~ gamma(0.3, 1);

  y1[1] ~ normal(m_hat, 1 / sqrt(0.03));
  for (i in 2:n1) {
    y1[i] ~ normal(m_hat + alpha * (y1[i - 1] - m_hat), sigma);
  }
  y_mis ~ normal(m_hat + alpha * (y1[n1] - m_hat), sigma);
  y2[1] ~ normal(m_hat + alpha * (y_mis - m_hat), sigma);
  for (i in 2:n2) {
    y2[i] ~ normal(m_hat + alpha * (y2[i - 1] - m_hat), sigma);
  }
}
"

earthq1 <- earthq[1:92]
earthq2 <- earthq[94:99]
m_hat <- mean(c(earthq1, earthq2))
fit2 <- stan(
  model_code = scode2,
  data = list(
    n1 = length(earthq1), n2= length(earthq2), m_hat = m_hat,
    y1 = earthq1, y2 = earthq2
  ),
  iter = 2000,
  verbose = FALSE
)
```

Tulokseksi saadaan

| | Mean | SD | 2.5 % | 97.5 % |
|---|---|---|---|---|
| $\alpha$ | 0.558408 | 0.0835179 | 0.3930069 | 0.7217718 |
| $\sigma^2$ | 36.861388 | 5.5024450 | 27.4053442 | 49.3123260 |
| $y_{93}$ | 14.145270 | 5.3173522 | 3.7618671 | 24.8534227 |

Saamme nyt marginaaliposteriorin myös puuttuvalle havainnolle $y_{93}$. Tämän jakauman odotusarvo on 14.1452696 ja keskihajonta 5.3173522 ja 95 %:n posterioriväli (3.76186708645141, 24.8534226891283).

# 6 Päätöksentekoteoriaa

## 6.1 Päätöksenteko Bayes-ongelmana

Päätöksenteossa tavoitteena on valita paras päätös $d$ mahdollisten päätösten joukosta $\mathcal{D}$. Päätöksen seurauksia kuvaa hyötyfunktio $U(d, \theta)$, joka kertoo päätöksestä $d$ seuraavan utiliteetin eli hyödyn, kun maailman tila on $\theta$. Maailman tilaan liittyy yleensä epävarmuutta, jota bayesilaisittain kuvataan posteriorijakaumalla $p(\theta|y)$. Optimaalinen päätös maksimoi odotetun hyödyn

$$U(d|y) = \int U(d, \theta)p(\theta|y)\, d\theta.$$

Hyötyä mitataan tavallisimmin rahassa, mutta hyöty ei välttämättä ole lineaarinen funktio rahan suhteen.

Usein päätöksentekotilanne on kaksi- tai useampivaiheinen siten, että ensimmäinen päätös koskee lisätietojen hankkimista. Merkitään tutkimusasetelmaa (design) symbolilla $\eta$. Yksinkertaisimmillaan $\eta$ pelkistyy otoskooksi, joka tutkijan tulee valita. Asetelmaan $\eta$ liittyvä maksimoitu odotettu hyöty saadaan kaavalla

$$\overline{U}(\eta) = \int_y \left[ \max_{d\in\mathcal{D}} \int_\Theta U(d, \theta)p(\theta|y)\, d\theta \right] p(y|\eta)\, dy.$$

Asetelma $\eta_0$ tarkoittaa, että uutta tietoa ei kerätä lainkaan. Päätöksenteko perustuu tällöin ainoastaan prioritietoon.

## 6.2 Informaation arvo

Lisätietojen avulla käsitys maailman tilasta tarkentuu ja päätöksen odotettu hyöty kasvaa. Oletetaan, että hyötyä mitataan suoraan rahassa. Tällöin odotetun hyödyn muutosta kutsutaan informaation (rahalliseksi) arvoksi ja se saadaan erotuksena

$$\overline{U}(\eta) - \overline{U}(\eta_0) = \int_y \left[ \max_{d\in\mathcal{D}} \int_\Theta U(d, \theta)p(\theta|y)\, d\theta \right] p(y|\eta)\, dy - \max_{d\in\mathcal{D}} \int_\Theta U(d, \theta)p(\theta)\, d\theta. \qquad (6.1)$$

Lisätietojen hankkimiseen liittyviä kustannuksia tulee verrata informaation arvoon. Informaation arvo 6.1 on mahdollista yleistää tilanteeseen, jossa hyötyä ei mitata suoraan rahassa.

Taulukko 6.1: Hyötyfunktion arvot erilaisille päätöksille ja säätiloille.

| Päätös | Poutaa | Sadetta |
|---|---|---|
| A) Kotiin jääminen | 3 | 3 |
| B) Ulos ilman sateenvarjoa | 6 | 1 |
| C) Ulos sateenvarjon kanssa | 2 | 4 |

## 6.3 Esimerkkejä

Päätöksentekoa ja informaation arvon käsitettä selvennetään kahdella esimerkillä.

**Esimerkki 6.1** (Sateenvarjo-ongelma)**.** Herra Bayes harkitsee kolmea vaihtoehtoa, jotka ovat A) kotiin jääminen, B) lähteminen ulos ilman sateenvarjoa ja C) lähteminen ulos sateenvarjon kanssa. Edellisen päivän sääennusteen mukaan sateen todennäköisyys on 0.5. Päätöksiin ja säätilaan liittyvät hyötyfunktion arvot on esitetty taulukossa 6.1. Oletetaan, että viimeisin sääennuste antaa täsmällisen tiedon päivän säästä (sataa tai ei sada todennäköisyydellä 1). Kuinka paljon herra Bayesin kannattaa enintään maksaa tästä tiedosta?

Lasketaan ensin optimaalinen päätös, kun käytettävissä on tieto edellisen päivän sääennusteesta. Odotetuiksi hyödyiksi saadaan päätökselle A: $0.5 \cdot 3 + 0.5 \cdot 3 = 3$, päätökselle B: $0.5 \cdot 6 + 0.5 \cdot 1 = 3.5$ ja päätökselle C: $0.5 \cdot 2 + 0.5 \cdot 4 = 3$. Päätös B, ulos lähteminen ilman sateenvarjoa, on siis optimaalinen päätös, koska odotettu hyöty on suurin.

Tarkastellaan seuraavaksi päätöksentekotilannetta, jossa päivän säätila tunnetaan. Taulukosta 6.1 nähdään, että poutasäällä optimaalinen päätös on lähteä ulos ilman sateenvarjoa ja sadesäällä optimaalinen päätös on lähteä ulos sateenvarjon kanssa. Odotettu hyöty on siis $0.5 \cdot 6 + 0.5 \cdot 4 = 5$. Huomaa, että laskussa tarvitaan sateen prioritodennäköisyyttä, vaikka päätös vaihtoehtojen A, B ja C välillä tehdäänkin ilman epävarmuutta säätilasta.

Täsmällisestä sääennusteesta kannattaa siis maksaa enintään $5 - 3.5 = 1.5$. Tämä erotus on informaation arvo.

**Esimerkki 6.2** (Kalankasvattajan ongelma)**.** Kalankasvatusaltaassa on 10 000 kalaa. Tuntematon osuus kaloista kantaa tautia, joka ei tartu toisiin kaloihin mutta johtaa hoitamattomana kuolemaan. Sairaat kalat on mahdollista parantaa lääkitsemällä koko allas, mikä maksaa 15 000 euroa. Kasvatusjakson lopussa kasvattaja myy elossa olevat kalat 5 euron kappalehintaan. Käytettävissä oleva ennakkotieto sairaiden kalojen osuudesta $\theta$ voidaan esittää priorijakaumana $p(\theta) = \text{Beta}(\theta|3, 3)$.

Sairaiden kalojen osuudesta on mahdollista kerätä lisätietoja diagnosoimalla satunnaisesti valittuja kaloja. Diagnostinen testi maksaa 10 euroa kalaa kohden. Kuinka monta kalaa kasvattajan kannattaa diagnosoida ennen hoitopäätöksen tekemistä?

Aluksi määrittelemme kalojen rahallisen arvon

$$v(\theta, d) = \begin{cases} (1-\theta)10000 \times 5, & d = \text{ei lääkitystä} \\ 10000 \times 5 - 15000, & d = \text{lääkitys}, \end{cases}$$

jossa luvut ovat euroja. Oletamme hyödyn olevan suoraan verrannollinen rahalliseen arvoon, jolloin

$$U(\theta, d, n_\eta, y) = U(\theta, d, n_\eta) = v(\theta, d) - 10n_\eta. \tag{6.2}$$

Koska ennakkotietojen perusteella $\mathrm{E}(\theta) = 3/(3+3) = 0.5$, optimaalinen päätös pelkän prioritiedon perusteella on kasvatusaltaan hoitaminen, joka johtaa maksimoituun odotettuun hyötyyn $\overline{U}(\eta_0) = 35000$. Jotta lisätiedon hankkiminen kaloja diagnosoimalla olisi järkevää, maksimoidun odotetun hyödyn tulisi ylittää 35 000 euroa.

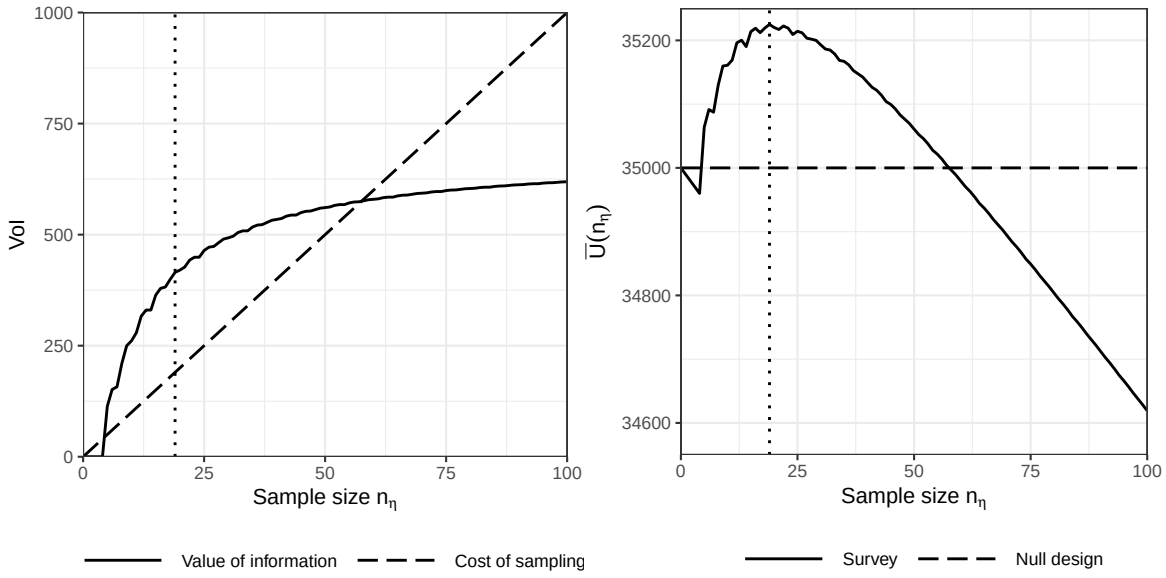
Oletetaan yksinkertaisuuden vuoksi, että otanta toteutetaan palauttaen. Olkoon $n_\eta$ diagnosoitujen kalojen määrä ja $y$ tautia kantavien kalojen määrä otoksessa. Päädytään malliin $y|\theta, n_\eta \sim \mathrm{Bin}(y|n_\eta, \theta)$. Otoskokoon $n_\eta$ liittyvä odotettu hyöty saadaan laskemalle jokaiselle mahdolliselle sairaiden kalojen määrälle $y$ maksimoitu odotettu hyöty ja painottamalla näitä kunkin mahdollisuuden prioritodennäköisyydellä.Kaavana tämä voidaan esittää muodossa

$$\overline{U}(n_\eta) = \sum_{y=0}^{n_\eta} \text{Beta-Bin}(y|n_\eta, 3, 3) \max_{d\in\mathcal{D}} \int U(\theta, d, n_\eta)\text{Beta}(\theta|y+3, n_\eta - y + 3)d\theta.$$

missä $U(\theta, d, n_\eta)$ on määritelty kaavassa @ref(eq:farmersutility). Maksimoitua odotettua hyötyä $\overline{U}(n_\eta)$ laskettaessa hyötyfunktio $U(\theta, d, n_\eta)$ integroidaan mahdollisten otosten ja posteriorijakauman $p(\theta|y)$ ylitse sekä maksimoidaan päätöksen $d$ suhteen. Vaihtoehtoinen tapa ratkaista ongelma on laskea otokseen $n_\eta$ liittyvä informaation arvo

$$\overline{v}(n_\eta) - \overline{v}(n_0) = \sum_{y=0}^{n_\eta} \text{Beta-Bin}(y|n_\eta, 3, 3) \max_{d\in\mathcal{D}} \int v(\theta, d)\text{Beta}(\theta|y+3, n_\eta - y + 3)d\theta - 35000,$$

ja verrata sitä otantakustannuksiin. Molemmat tavat johtavat samaan lopputulokseen. Kuvassa 6.1 esitetään informaation arvo, otantakustannukset ja maksimoitu odotettu hyöty, kun otoskoko vaihtelee nollasta sataan. Hyöty on positiivinen, kun otoskoko on välillä $[5, 57]$. Suurin hyöty saavutetaan, kun $n_\eta = 19$.




(a) Informaation arvo (b) Maksimoitu odotettu hyöty

Kuva 6.1: Informaation arvo (vasen kuva) ja maksimoitu odotettu hyöty (oikea kuva) erikokoisille otoksille. Pystyviiva kuvaa optimaalista otoskokoa, jolle informaation arvon ja otantakustannusten erotus on suurin. Informaation arvo on negatiivinen, kun otoskoko on neljä tai pienempi, koska kalojen lääkitseminen on tällöin aina optimaalinen päätös diagnosoinnin tuloksesta riippumatta.

# 7 Posteriorin approksimointi

MCMC-simuloinnin lisäksi posteriorijakauman estimointiin on olemassa muitakin menetelmiä. Tiettyihin tilanteisiin on olemassa menetelmiä, jotka saattavat olla huomattavasti MCMC-simulointia nopeampia. Toisaalta approksimointi on ainoa vaihtoehto silloin, kun malli niin monimutkainen ettei sitä laskennallisten rajoitteiden takia ole käytännössä mahdollista sovittaa MCMC-menetelmin.

## 7.1 Numeerinen integrointi

Kun mallin parametrien lukumäärä on pieni, käytännössä yksi tai kaksi, integraalin $\int p(\theta)p(y|\theta)\,d\theta$ voi laskea numeerisella integroinnilla. Tähän soveltuu esimerkiksi R-ohjelman `integrate`-funktio. Parametrien lukumäärän kasvaessa numeerisen integroinnin laskennallinen vaativuus kasvaa nopeasti. Moniulotteisessa tapauksessa voidaan käyttää esimerkiksi R pakettia `cubature`, joka sisältää useita adaptiivisia menetelmiä numeeriseen integrointiin.

**Esimerkki 7.1** (Genetiikkaesimerkki)**.** Koe-eläimet ($n = 197$) on jaettu neljään luokkaan geneettisin perustein:

| Kategoria | AA | AB | BA | BB |
|---|---|---|---|---|
| Frekvenssi | 125 | 18 | 20 | 34 |

Olkoon $y_i$ luokan $i$ frekvenssi. Oletetaan, että luokittelu tapahtuu riippumattomasti, jolloin yleisin malli havaintovektorille $y = (y_1, y_2, y_3, y_4)$ on multinomiaalimalli:

$$y_1, y_2, y_3, y_4|\pi_1, \pi_2, \pi_3, \pi_4 \sim \text{Mult}(n; \pi_1, \pi_2, \pi_3, \pi_4),$$

missä $\sum_{i=1}^{4} y_i = \text{n}$, $0 < \pi_i < 1$, $i = 1, 2, 3, 4$ ja $\sum_{i=1}^{4} \pi_i = 1$.

Geneettisen teorian perusteella asetetaan malli todennäköisyyksille $\pi_i$

$$\pi = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

Mallissa on siten yksiulotteinen parametri $\theta$. Huomataan, että summaehto täyttyy aina ja lisäksi voidaan todeta, että $0 < \theta < 1$.

Valitaan prioriksi tasajakauma $p(\theta) \propto 1$. Posterioriksi saadaan (ilman normeerausta)

$$p(\theta|y) \propto \left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1}{4}(1-\theta)\right)^{y_2}\left(\frac{1}{4}(1-\theta)\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}.$$

Lasketaan

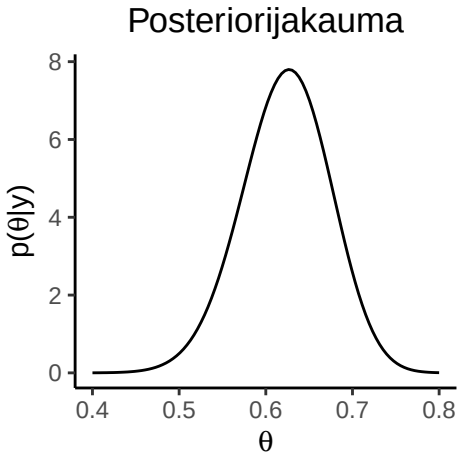
$$p(y) = \int_0^1 \left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1}{4}(1-\theta)\right)^{y_2}\left(\frac{1}{4}(1-\theta)\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4} d\theta$$

numeerisella integroinnilla.

```
y <- c(125, 18, 20, 34)

# Normeeraamaton posteriori
p_un <- function(theta, y) {
  (0.5 + 0.25 * theta)^y[1] * (0.25 * (1 - theta))^y[2] *
    (0.25 * (1 - theta))^y[3] * (0.25 * theta)^y[4]
}
I <- integrate(p_un, lower = 0, upper = 1, y = y)$value
I
```

```
[1] 5.84345e-91
```

Parametrin $\theta$ posteriorijakauman kuvaaja on




Numeerisella integroinnilla saadaan posteriorijakaumalle posteriorikeskiarvo $\approx 0.623$ ja posteriorikeskihajonta $\approx 0.051$.

## 7.2 Laplace-approksimaatio

Tiettyjen säännöllisyysehtojen vallitessa voidaan osoittaa, että posteriorijakauma lähestyy asymptoottisesti normaalijakaumaa, kun havaintojen määrä kasvaa. Tätä ominaisuutta hyödynnetään Laplace-approksimaatiossa. Jos havaintojen määrä on pieni, approksimaatio ei välttämättä toimi luotettavasti.

Laplace-approksimaatio perustuu posteriorimaksimin $\hat{\theta}$ ympärille kehitettyyn Taylorin sarjaan. Merkitään $\theta = (\theta_1, \dots, \theta_m)^\top$. Taylorin kehitelmässä käytetään normeeraamattoman posteriorijakauman logaritmia

$$h(\theta) = \log[p(y|\theta)p(\theta)].$$

Lasketaan ensimmäisen ja toisen kertaluvun osittaisderivaatat

$$\begin{aligned} h'(\theta) =& \frac{\partial h(\theta)}{\partial \theta} = \left( \frac{\partial h(\theta)}{\partial \theta_1}, \dots, \frac{\partial h(\theta)}{\partial \theta_m} \right)^T \\ h''(\theta) =& \frac{\partial^2 h(\theta)}{\partial \theta^2} = \left( \frac{\partial^2 h(\theta)}{\partial \theta_r \partial \theta_s} \right), \quad r, s = 1, \dots, m \end{aligned}$$

missä $h''(\theta)$ on siis $m \times m$-matriisi. Posteriorimaksimille $\hat{\theta}$ pätee $h'(\hat{\theta}) = 0$. Taylorin kehitelmästä saadaan

$$h(\theta) \approx h(\hat{\theta}) + (\theta - \hat{\theta})^\top h'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^\top h''(\hat{\theta})(\theta - \hat{\theta}).$$

Kun $\theta - \hat{\theta}$ on pieni, korkeamman kertaluvun termit voidaan jättää ottamatta huomioon. Tällöin posteriorijakaumalle pätee

$$p(\theta|y) \propto \exp(h(\theta)) \propto \exp\left( \frac{1}{2}(\theta - \hat{\theta})^\top h''(\hat{\theta})(\theta - \hat{\theta}) \right).$$

Kirjoittamalla $V = (-h''(\hat{\theta}))^{-1}$ saadaan muoto

$$p(\theta|y) \propto \exp\left( -\frac{1}{2}(\theta - \hat{\theta})^\top V^{-1}(\theta - \hat{\theta}) \right),$$

joka vastaa normeeraamatonta normaalijakaumaa. Laplace-approksimaatioksi saadaan tämän perusteella

$$\theta|y \sim N_m(\hat{\theta}, V).$$

On siis laskettava funktion $h(\theta)$ maksimi (joko derivoimalla tai numeerisella optimoinnilla) sekä sen toiset derivaatat maksimikohdassa. On huomattava, että tämä on approksimaatio. Jos normeeraustekijää tarvitaan, niin se voidaan myös laskea

$$\begin{aligned} p(y) &\approx \int \exp(h(\hat{\theta})) \exp\left( -\frac{1}{2}(\theta - \hat{\theta})^T V^{-1}(\theta - \hat{\theta}) \right) d\theta \\ &= p(\hat{\theta})p(y|\hat{\theta})(2\pi)^{\frac{m}{2}} \det(V)^{\frac{1}{2}} \times \int (2\pi)^{-\frac{m}{2}} \det(V)^{-\frac{1}{2}} \exp\left( -\frac{1}{2}(\theta - \hat{\theta})^T V^{-1}(\theta - \hat{\theta}) \right) d\theta \\ &= p(\hat{\theta})p(y|\hat{\theta})(2\pi)^{\frac{m}{2}} \det(-h''(\hat{\theta}))^{-\frac{1}{2}}. \end{aligned}$$

Approksimaatio voidaan kytkeä Newtonin algoritmiin:

- Asetetaan alkuarvo $\widehat{\theta}^{(0)}$.
- Päivitys: $\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} - h''(\widehat{\theta}^{(k)})^{-1} h'(\widehat{\theta}^{(k)})$.

Jos algoritmi konvergoi kun $k \to \infty$, niin silloin

- $\widehat{\theta}^{(k)} \to \widehat{\theta}$
- $h'(\widehat{\theta}^{(k)}) \to 0$
- $-h''(\widehat{\theta}^{(k)}) \to V^{-1}$

**Esimerkki 7.2** (Genetiikkaesimerkki (jatk.))**.** Palataan genetiikkaesimerkkiin, jossa

$$\begin{aligned} p(\theta|y) &\propto \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1} \left(\frac{1}{4}(1-\theta)\right)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} \\ &\propto (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4}. \end{aligned}$$

Nyt

$$\begin{aligned} h(\theta) &= y_1 \log(2+\theta) + (y_2+y_3)\log(1-\theta) + y_4 \log(\theta) \\ h'(\theta) &= \frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta} \\ h''(\theta) &= -\frac{y_1}{2+\theta} - \frac{y_2+y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2}. \end{aligned}$$

Newtonin algoritmilla saadaan $\widehat{\theta} = 0.6268215$ ja $h''(\widehat{\theta}) = -377.5169$. Posteriori on approksimatiivisesti normaalijakauma, jonka odotusarvo on 0.627 ja varianssi 0.0026.

Posteriorista voidaan laskea esimerkiksi approksimatiivinen 90 %:n Bayes-väli parametrille $\theta$: $0.627 \pm 1.644 \times 0.0516$ ja saadaan tulokseksi $(0.590, 0.711)$. Tulos vastaa varsin hyvin numeerisella integroinnilla saatua (jota voidaan pitää tarkkana). Tässä tapauksessa Laplace-approksimaatio toimii siis hyvin.

## 7.3 Variaatiopäättely

Normaalijakauman sijaan posteriorijakaumaa voi approksimoida myös jollain muulla jakaumalla. Bayes-tilastotieteessä voidaan käyttää variaatiolaskentaa (*variational inference*) parhaan approksimaation löytämiseen. Approksimaation ominaisuuksia ei vielä tunneta perinpohjaisesti.

Menetelmässä posteriorijakaumaa $p(\theta|y)$ approksimoidaan parametrisella jakaumalla $q_\eta(\theta)$. Useimmiten käytetään tulomuotoisia jakaumia

$$q_\eta(\theta) = \prod_{i=1}^{m} q_\eta^{(i)}(\theta_i)$$

jollekin parametrivektorinositukselle $\{\theta_1, \ldots, \theta_m\}$. Approksimaatioon liittyvät parametrit $\eta$ määritetään siten, että approksimaation $q_\eta(\theta)$ etäisyys posteriorista $p(\theta|y)$ minimoituu. Jakaumien välistä etäisyyttä mitataan Kullback–Leibler-divergenssilla

$$\int q_\eta(\theta) \log\left(\frac{q_\eta(\theta)}{p(\theta|y)}\right) d\theta.$$

Tulomuotoisen approksimaation tapauksessa variaatiolaskennan avulla voidaan osoittaa, että etäisyys minimoituu, kun

$$\log q_\eta^{(i)}(\theta_i) \propto \int q_\eta^{(-i)}(\theta_{-i}) \log(p(y,\theta))\, d\theta_{-i}, \tag{7.1}$$

missä alaindeksi $-i$ tarkoittaa ositteen $i$ poisjättämistä. Ratkaisu 7.1 tulee normalisoida jakaumaksi. Käytännössä laskenta vaatii usein iterointia, koska ratkaisu ositteelle $\theta_i$ riippuu muiden ositteiden ratkaisuista. Yleisemmin Kullback–Leibler-divergenssi voidaan Bayesin kaavan mukaan kirjoittaa muodossa

$$\int q_\eta(\theta) \log\left(\frac{q_\eta(\theta)}{p(\theta|y)}\right) d\theta = -\int q_\eta(\theta) \log p(y|\theta)\, d\theta + \int q_\eta(\theta) \log\left(\frac{q_\eta(\theta)}{p(\theta)}\right) d\theta + \log p(y),$$

mistä nähdään, että optimaalinen variationaalinen approksimaatio $q_\eta(\theta)$ tasapainottaa uskottavuusfunktion maksimoimista ja pientä etäisyyttä priorijakaumasta. Variaatioapproksimaation parametrit $\eta$ voidaan optimoida esim. stokastista gradienttimenetelmää käyttäen.

## 7.4 INLA

Integroitu sisäkkäinen Laplace-approksimaatio (*Integrated Nested Laplace Approximation*, INLA) on Laplace-approksimaation yleistys. Menetelmää voi käyttää latenteille gaussisille malleille, joita ovat muun muassa yleistetyt lineaariset mallit sekä monet spatiaaliset mallit. INLA on käyttökelpoinen menetelmä silloin, kun hyperparametrien määrä $\theta$ on pieni, tyypillisesti alle 6. Latenttien muuttujien määrä sen sijaan saa olla suuri.

Latentti gaussinen malli koostuu kahdesta osasta: latentit muuttujat $x$ ovat normaalijakautuneita mallin $p(x|\theta)$ mukaisesti ja havainnot $y$ noudattavat ehdollista jakaumaa $p(y|x,\theta)$. Keskeinen idea on kirjoittaa hyperparametrien posteriori muodossa

$$p(\theta|y) \propto \frac{p(\theta)p(x|\theta)p(y|x,\theta)}{p(x|\theta,y)} \tag{7.2}$$

ja käyttää Laplace-approksimaatiota nimittäjälle $p(x|\theta,y)$. Approksimaatio on varsin tarkka, koska jakauma $p(x|\theta)$ on normaalijakauma. Kaava 7.2 tuottaa approksimaation normeeraamattomalle posteriorille $p(\theta)p(y|\theta)$. Lisää approksimaatioita tarvitaan, kun halutaan määrittää marginaaliposteriorijakaumat parametrivektorin $\theta$ komponenteille. Nämä approksimaatiot ovat teknisesti hankalia ja toimivat ainoastaan silloin, kun parametrivektorin $\theta$ dimensio on pieni.

Lisätietoja menetelmästä on saatavilla INLAn kotisivulta (https://www.r-inla.org/home).

## 7.5 Hylkäysotanta

Hylkäysotannassa (*rejection sampling*) pyritään tuottamaan havaintoja kiinnostuksen kohteena olevasta posteriorijakaumasta $p(\theta|y)$ käyttämällä hyväksi ehdotusjakaumaa $g(\theta)$. Hylkäysotanta edellyttää, että seuraavat ominaisuudet ovat voimassa ehdotusjakaumalle:

- Havaintoja voidaan tuottaa jakaumasta $g(\theta)$.
- *Tärkeyssuhteelle* on voimassa $p(\theta|y)/g(\theta) \leq M$ kaikilla $\theta$, missä $M$ on jokin positiivinen vakio.

Hylkäysotanta etenee seuraavasti (yhden realisaation tuottamiseksi):

1. Generoi ehdotus $\theta^*$ jakaumasta $g(\theta)$.
2. Hyväksy $\theta^*$ realisaatioksi jakaumasta $p(\theta|y)$ todennäköisyydellä $p(\theta|y)/(Mg(\theta))$. Jos ehdotus hylätään, palaa kohtaan 1.

Tärkeyssuhteelle voimassaoleva ehto $p(\theta|y)/g(\theta) \leq M$ takaa sen, että ehdotuksen hyväksymistodennäköisyyden lauseke on aina $\leq 1$ kaikilla $\theta$:n arvoilla. Menetelmän toimivuutta voi arvioida hyväksyttyjen realisaatioiden osuuden avulla. Hylkäysotanta toimii sitä tehokkaammin, mitä enemmän $g(\theta)$ muistuttaa posterioria $p(\theta|y)$.

## 7.6 Tärkeysotanta

Tärkeysotannassa (*importance sampling*) jostakin sopivasta ehdotusjakaumasta generoituja havaintoja painotetaan siten, että ne edustavat kiinnostuksen kohteena olevaa posteriorijakaumaa. Ehdotusjakauman $g(\theta)$ tulisi muistuttaa posteriorijakaumaa ja havaintojen generoimisen tulisi olla helppoa.

Olkoon kiinnostuksen kohteena odotusarvo $\mathrm{E}(\xi(\theta)|y)$, missä $\xi$ on jokin tunnettu funktio. Tärkeysotanta perustuu tulokseen

$$\begin{aligned}\mathrm{E}(\xi(\theta)|y) &= \int \xi(\theta)p(\theta|y)\,d\theta = \frac{\int \xi(\theta)p(\theta)p(y|\theta)\,d\theta}{\int p(\theta)p(y|\theta)\,d\theta} \\ &= \frac{\int \xi(\theta)\frac{p(\theta)p(y|\theta)}{g(\theta)}g(\theta)\,d\theta}{\int \frac{p(\theta)p(y|\theta)}{g(\theta)}g(\theta)\,d\theta} = \frac{\mathrm{E}_g(w(\theta)\xi(\theta))}{\mathrm{E}_g(w(\theta))},\end{aligned}$$

missä painot määräytyvät kaavasta $w(\theta) = p(\theta)p(y|\theta)/g(\theta)$.

Generoidaan nyt riippumattomia havaintoja $\theta^{(1)}, \ldots, \theta^{(n)}$ jakaumasta $g(\theta)$. Keskiarvon $\overline{\xi} = \mathrm{E}(\xi(\theta)|y)$ tärkeysotantaan perustuva estimaattori on painotettu keskiarvo

$$\overline{\xi}_{IS} = \frac{\sum_{j=1}^{n} w(\theta^{(j)})\xi(\theta^{(j)})}{\sum_{j=1}^{n} w(\theta^{(j)})},$$

ja estimoitu keskivirhe saadaan painotettujen havaintojen varianssin avulla

$$\mathrm{SE}(\bar{\xi}_{IS}) = \frac{\sqrt{\sum_{j=1}^{n} [w(\theta^{(j)})]^2 \, (\xi(\theta^{(j)}) - \bar{\xi}_{IS})^2}}{\sum_{j=1}^{n} w(\theta^{(j)})}.$$

Jos ehdotusjakauma on valittu huonosti, suurin osa painoista voi olla lähellä nollaa samalla, kun muutamat painot ovat todella suuria. Tällöin tärkeysotanta toimii huonosti. Ehdotusjakauman voi valita esimerkiksi Laplace-approksimaation pohjalta. Joissain tapauksissa varianssin suurentaminen saattaa olla tarpeen liian suurien painojen välttämiseksi. Ehdotusjakaumana voi käyttää myös informatiivista prioria. Tällöin painot yksinkertaistuvat uskottavuudeksi

$$w(\theta) = \frac{p(\theta)p(y|\theta)}{g(\theta)} = \frac{p(\theta)p(y|\theta)}{p(\theta)} = p(y|\theta).$$

**Esimerkki 7.3** (Genetiikkaesimerkki (jatk.))**.** Lasketaan posterioriodotusarvo tärkeysotannalla käyttäen simulointijakaumana Laplacen approksimaatiota $\mathrm{N}(0.6268215, 0.0514673^2)$:

```
importance_sampling <- function(y, mu, sigma, n_sim) {
  theta <- rnorm(n_sim, mu, sigma)
  dens <- dnorm(theta, mu, sigma)
  p_un <- (0.5 + 0.25 * theta)^y[1] * (0.25 * (1 - theta))^y[2] *
    (0.25 * (1 - theta))^y[3] * (0.25 * theta)^y[4]
  w <- p_un / dens
  theta_is <- sum(w * theta) / sum(w)
  theta_is
}

mu <- 0.6268215
sigma <- 0.0514673
y <- c(125, 18, 20, 34)
importance_sampling(y, mu, sigma, 1e4)
```

```
[1] 0.6229489
```

Mikä on hyvin lähellä numeerisella integroinnilla saatua arvoa.