Bayes-päättely, 3. harjoitukset (8.–9.2.2017)

1. Oletetaan riippumaton otos Bernoulli-jakaumasta:

$$Y_1, \ldots, Y_n \perp\!\!\!\perp | \, \theta, \quad Y_i \sim B(\theta)$$

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

(a) Osoita, että Bernoulli-jakauma kuuluu eksponenttiperheeseen. Mikä on sen luonnollinen parametri $\phi(\theta)$?

(b) Esitä satunnaisvektorin $\mathbf{Y} = (Y_1, \ldots, Y_n)$ jakauma $p(\mathbf{y}|\theta)$ eksponenttiperheen yleisessä muodossa (luentojen huomautuksen 2.15 muodossa), ja osoita tämän avulla, että konjugaattiprioriit sen jakaumalle ovat betajakaumia, eli muotoa

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1},$$

missä $\alpha, \beta > 0$.

(c) Johda eksponenttiperheen posteriorijakauman yleisen muodon (luentojen lause 2.17) avulla posteriorijakauma parametrille θ konjugaattipriorilla $\theta \sim \text{Beta}(\alpha, \beta)$ (Käytä yleistä muotoa, vaikka lasku on varmasti helpompi tehdä suoraan. Tämä esimerkki osoittaa, että vaikka yleinen muoto on olemassa, niin laskut voi monesti olla helpompi tehdä suoraan.).

(d) Oletetaan m uutta riippumatonta havaintoa $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_m)$ samasta Bernoulli-jakaumasta. Johda eksponettiperheen posteriorijennustejakauman yleisen muodon (luentojen lause 2.17) avulla posterioriennustejakauma $p(\tilde{\mathbf{y}}|\mathbf{y})$ uusille havainnoille ehdolla havaittu aineisto.

(e) Vertaa posteriorijakaumaasi luentojen esimerkin 1.5 tulokseen, ja posterioriennustejakaumaasi luentojen esimerkin 1.9 tulokseen. Miten selittäisi tulosta?

2. Oletetaan riippumattomat havainnot normaalijakaumasta tunnetulla varianssilla $\sigma_0^2 \in (0, \infty)$:

$$Y_1, \ldots, Y_n \perp\!\!\!\perp | \, \theta, \quad Y_i \sim N(\theta, \sigma_0^2)$$

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

(a) Osoita, että normaalijakauma tunnetulla varianssilla ja tunnetulla odotusarvolla θ kuuluu eksponenttiperheeseen.

(b) Esitä satunnaisvektorin $\mathbf{Y} = (Y_1, \ldots, Y_n)$ jakauma $p(\mathbf{y}|\theta)$ eksponenttiperheen yleisessä muodossa (luentojen huomautuksen 2.15 muodossa), ja osoita tämän avulla, että konjugaattiprioriit sen jakaumalle ovat normaalijakaumia, eli muotoa

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

missä $\mu_0 \in \mathbb{R}$, $\tau_0 \in (0, \infty)$.

(c) Johda eksponenttiperheen posteriorijakauman yleisen muodon (luentojen lause 2.17) avulla posteriorijakauma parametrille θ konjugaattipriorilla $\theta \sim N(\mu_0, \tau_0)$.

(d) Oletetaan m uutta riippumatonta havaintoa $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_m)$ samasta normaalijakaumasta. Johda eksponettiperheen posteriorijennustejakauman yleisen muodon (luentojen lause 2.17) avulla posterioriennustejakauma $p(\tilde{\mathbf{y}}|\mathbf{y})$ uusille havainnoille ehdolla havaittu aineisto.

3. Lopultakin oikeaa aineistoa! (Kirjan tehtävä 2.21) Tiedosto `pew_research_center_june_elect_wknd_data.dta` sisältää vuoden 2008 Yhdysvaltojen presidentinvaalien alla tehdyn mielipidekyselyn tulokset, ja `2008ElectionResult.csv` sisältää näiden vaalien tulokset osavaltioittain.

Lataa aineistot, ja laske osavaltiokohtaiset keskiarvot 'hyvin liberaalien' vastaajien osuuksille (mielipidekyselyn `ideo`-muuttujan arvo `'very liberal'`). Piirrä hajontakuva, jossa kullekin osavaltiolle x-akselilla on hyvin liberaalien osuus vastanneista, ja y-akselilla on Obamaa äänestäneiden osuus. Piirrä osavaltiot käyttäen niiden lyhenteitä `state.abb`-vektorista.

Piirrä myös hajontakuva, jossa x-akselilla on osavaltion vastaajamäärä, ja y-akselilla osavaltion hyvin liberaalien osuus.

Vihjeitä:

- `dta`-tiedoston saat ladattua `foreign`-paketin `read.dta`-funktiolla.
- Mielipidekyselyn ja tulosten osavaltiot ovat aakkosjärjestyksessä. Voit käyttää tätä hyväksi tulosten yhdistämisestä
- Poista `state.abb`-vektorista Alaska (`'AK'`) ja Havaiji (`'HI'`), niin sekin on samassa järjestyksessä kuin tulokset Washington D.C.:tä / District of Columbiaa lukuunottamatta, jota ei ole lyhenteissä, koska se ei ole oikea osavaltio. Voit poistaa sen tuloksista, tai sitten lisätä esim. lyhenteellä `'DC'` oikeaan kohtaan lyhenteisiin.
- **Bonus**: Voit myös piirtää osavaltiot niiden alueiden mukaisilla väreillä, jotka löytyvät `state.region`-vektorista.

4. Jatkoa edelliseen tehtävään. Johdantoa hierarkkisiin malleihin: tehdään alkeellinen versio hierarkkisesta mallista, jossa priorijakauman parametrit estimoidaan aineistosta (ns. empirical Bayes). Oletetaan, että hyvin liberaalien osuudet osavaltioittain (Obaman äänimäärät edellisessä tehtävässä olivat hämäystä: ne eivät liity varsinaiseen mallintamiseen mitenkään) Y_j noudattavat binomijakaumaa kukin omalla parametrillaan θ_j, ja nämä parametrit noudattavat kaikki yhteistä beta-jakaumaa:

$$Y_j|\theta_j \sim \text{Binom}(n_j, \theta_j), \quad \theta_j \sim \text{Beta}(\alpha, \beta)$$

kaikille $j = 1, \ldots, 49$.

(a) Estimoi beta-priorin parametrit koko aineistosta, eli hyvin liberaalien osuuksista kaikista vastaajista. Voit käyttää momenttimenetelmää, tai jos et jaksa, niin voit vain estimoida priorijakauman parametrit α ja β komennolla:

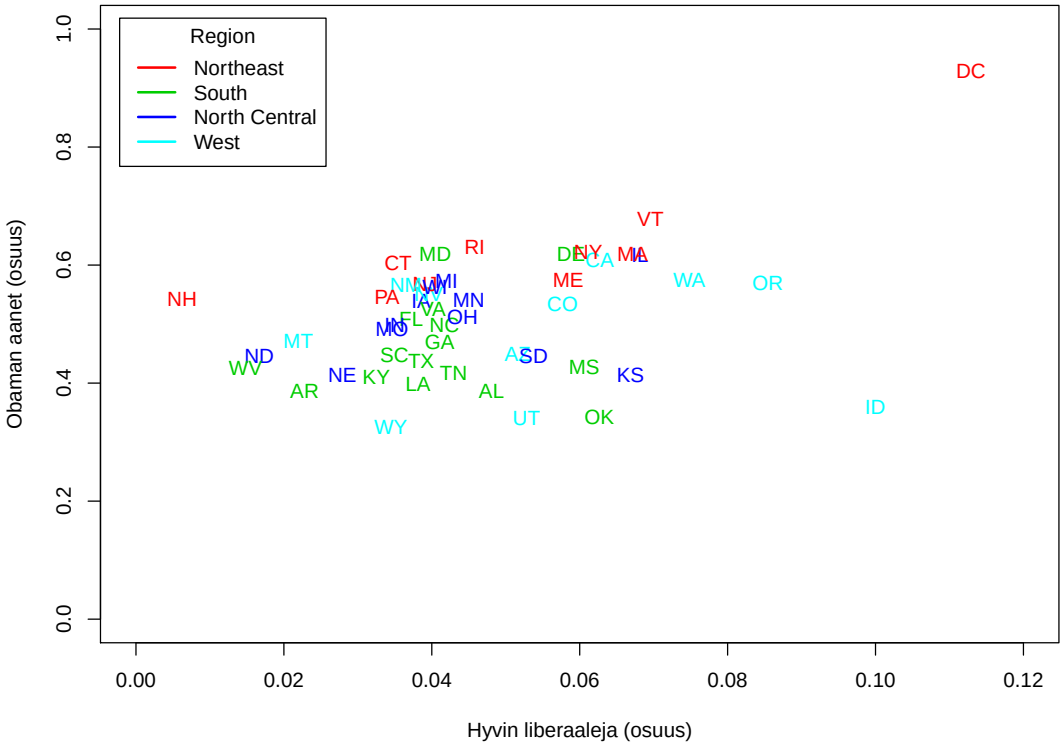
```
install.packages('VGAM')
library(VGAM)

# negative log likelihood of data given alpha; beta
ll <- function(alpha, beta) {
  -sum(dbetabinom.ab(y, n, alpha, beta, log = TRUE))
}

mm <- mle(ll, start = list(alpha = 1, beta = 10), method = "L-BFGS-B")
alpha <- coef(mm)[1]
beta <- coef(mm)[2]
```

Ylläolevassa koodissa oletetaan, että `y` on vektori, joka sisältää osavaltiokohtaiset hyvin liberaalien määrät, ja `n` vektori, joka sisältää osavaltiokohtaiset vastanneiden osuudet.

Kuva 1: Kuvien on siis tarkoitus olla jotain tämän näköistä, tässä tehtävän 3 ensimmäinen kuva. Värit ja niihin liittyvät alueet ovat extraa.



(b) Laske posteriorijakauma hyvin liberaalien osuuksille θ_j kullekin osavaltiolle, ja näiden avulla posteriorijakauman keskiarvot näille osuuksille. Piirrä hajontakuva, jossa x-akselilla on osavaltion vastaajamäärä, ja y-akselilla osavaltion hyvin liberaalien osuuden posteriorijakauman keskiarvo osavaltiolle. Mitä muutoksia huomasit edellisen tehtävän vastaavaan kuvaan, ja miten selittäisit niitä (vrt. kirjan kuvat 2.8 ja 2.9 sivulla 50)?