**Tilastollinen päättely I, kevät 2017**
**Harjoitus 5 – palautuspäivämäärä 25. 4.**

**Tehtävät 1–4** koskevat bayesläistä päättelyä (monisteen jaksot 10.1–10.6). Tehtävät 1–3 ovat perustehtäviä ja tehtävä 4 on "täydentävää ainesta".

**1.** Vuonna 1975 uutisoitiin tutkimuksesta, jonka mukaan 50 % kanadalaisista miehistä käytti värillisiä (eli muita kuin valkoisia) alushousuja kun taas amerikkalaisista miehistä sellaisia käytti vain 20 %. Bermudalaisen hotellin asiakaskunta koostui yksinomaan amerikkalaisista ja kanadalaisista siten, että miesasiakkaista 80 % oli amerikkalaisia ja 20 % kanadalaisia. Todennäköisyyslaskentaa opiskellut siivooja huomasi miesasiakkaan huoneessa punaiset alushousut. Millä todennäköisyydellä hän päätteli asiakkaan olevan kanadalainen? Muotoile sopivat tapahtumat ja sovella niihin Bayesin kaavaa.

**2.** Kulhossa on 4 palloa, joista $\theta$ on valkoisia ja loput mustia. Henkalla ei ole mitään tietoa siitä, miten pallojen värit ovat määräytyneet, joten hänen ennakkokäsityksensä mukaan kaikki vaihtoehdot $\theta$:n arvolle (0,1,2,3,4) ovat yhtä todennäköisiä. Hän nostaa korista umpimähkään ja palauttaen kaksi palloa: kumpikin niistä on valkoinen. (Tätä satunnaiskoetta kuvaava malli on esitetty monisteen jaksossa 10.2.)

Esitä arvot luettelemalla ja halutessasi myös kaavalla i) Henkan priorijakauma, ii) uskottavuusfunktio, iii) Henkan posteriorijakauma. Mikä on Henkan mielestä havaintojen teon jälkeen todennäköisin $\theta$:n arvo?

**3.** Jatkoa edelliseen tehtävään. Ville seuraa sivusta Henkan koetta. Samalla hän tietää, että pallot päätyivät kulhoon seuraavasti: Oli 4 samanlaista valkoista palloa. Kunkin pallon kohdalla heitettiin harhatonta lanttia, ja mikäli saatiin kruunu, pallo värjättiin mustaksi. Sitten pallot pantiin kulhoon. Mitkä ovat Villen priori- ja posteriorijakauma (luettele arvot)? Mikä on hänen mielestään havaintojen teon jälkeen todennäköisin $\theta$:n arvo?

**4.** Tilastollinen malli aineistolle $\mathbf{y} = (y_1, \ldots, y_n)$ on satunnaisotos normaalijakaumasta $N(\theta, \sigma^2)$ eli

$$f(\mathbf{y}|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\},$$

jossa $\theta$ on reaalinen parametri ja $\sigma^2 > 0$ on jokin tunnettu luku. Priorijakaumaksi $p(\theta)$ valitaan normaalijakauma $N(0, \sigma_0^2)$, jossa $\sigma_0^2 > 0$ on tunnettu luku. Osoita, että posteriorijakauma $p(\theta|\mathbf{y})$ on eräs normaalijakauma $N(\mu_1, \sigma_1^2)$ ja lausu $\mu_1$ sekä $\sigma_1^2$ lukujen $\sigma_0^2$, $\sigma^2$, $n$ ja $\bar{y}$ avulla.

Tämä lasku osoittaa, että normaalijakauma on normaalimallin uskottavuuden *liittopriori* (vrt. monisteen jakso 10.5). Tulos on voimassa myös yleisemmälle priorille $N(\mu_0, \sigma_0^2)$; oletus $\mu_0 = 0$ tekee vain laskun hieman yksinkertaisemmaksi.

*Apu.* Tässä lienee kätevää edetä verrannollisuustarkastelun $p(\theta|\mathbf{y}) \propto p(\theta)f(\mathbf{y}|\theta)$ kautta (vrt. monisteen s. 123).

KÄÄNNÄ!

**Tehtävää 5** varten perehdy monisteen jaksoihin 9.1 ja 9.2, jotka koskevat kahden muuttujan välisen lineaarisen riippuvuuden kuvaamista ja pienimmän neliösumman menetelmää. Luennoilla aihetta käsitellään viimeisellä luentoviikolla.

**5. PNS-suoran sovitus (ks. monisteen jaksot 9.1–9.3).** Taulukossa alla on 12 naisen iät ($x$) ja levossa mitatut systoliset verenpaineet ($y$):

| $x$ | 56 | 42 | 72 | 36 | 63 | 47 | 55 | 49 | 38 | 42 | 68 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 147 | 125 | 160 | 118 | 149 | 128 | 150 | 145 | 115 | 140 | 152 | 155 |

Sijoita pisteparit $(x_i, y_i)$, $i = 1, \ldots, 12$, $xy$-koordinaatistoon ja totea, että niiden välinen riippuvuus näyttää suunnilleen lineaariselta. Aineistosta voidaan laskea seuraavat tunnusluvut:

$$\bar{x} = 52.33 \qquad \bar{y} = 140.33$$
$$\textstyle\sum_i (x_i - \bar{x})^2 = 1550.67 \qquad \sum_i (y_i - \bar{y})^2 = 2500.67$$
$$\textstyle\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 1764.67.$$

Laske PNS-suoran kertoimet ja ilmoita sen yhtälö. Sijoita PNS-suora piirtämääsi hajontakuvioon.

**Tehtävää 6** varten perehdy alustavasti testiteorian yleiseen asetelmaan ja $p$-arvon käsitteeseen sekä tulkintaan (jakso 6.1 ja erityisesti 6.4 sekä 6.9). Luvun 6 aiheista *emme* tällä kurssilla käsittele lainkaan testien voiman käsitettä (jakso 6.3 ja osittain 6.5), testien ja luottamusjoukkojen duaalisuutta (jakso 6.6) ja binomijakauman testausta (jakso 6.8).

**6.** Eräässä väestössä on otantatutkimuksen avulla tutkittu tuloeroja yhtäältä miesten ja naisten välillä ja toisaalta mustien ja valkoisten välillä. Tutkimuksen tulokset raportoidaan tällaiseen hyvin tyypilliseen tapaan: ”Miesten ja naisten keskituloissa havaittiin merkitsevä ero ($p = 0.009$), miesten ansaitessa keskimäärin enemmän. Mustien ja valkoisten keskituloissa sen sijaan ei havaittu eroa ($p = 0.11$).”

a) Ystäväsi ei ole opiskellut lainkaan tilastotiedettä. Miten selität tutkimuksen tulokset ja siinä esiintyvät lukuarvot hänelle?

b) Ystäväsi kysyy: ”Onko nyt siis osoitettu, että ko. väestössä miehet ansaitsevat keskimäärin enemmän kuin naiset ja että mustilla ja valkoisilla on olennaisesti samat keskitulot?” Miten vastaat? Perustele.