Tilastollinen päättely II, syksy 2015 – kevät 2016
Harjoitus 13 (1. ja 3. 3. 2016), ratkaisut

Tehtävät (paitsi viimeinen) liittyvät luottamusväleihin ja -joukkoihin. Monisteen luku 6.

1. (Monisteen teht. 6.1.) Olkoot $Y_1, \ldots, Y_{25} \sim N(\mu, \sigma^2)$ $\perp\!\!\!\perp$ ja $S^2 = \sum_{i=1}^{25}(Y_i - \overline{Y})^2/24$. Palauta todennäköisyyslaskennan kurssilta mieleen, miten S^2/σ^2 on jakautunut; erityisesti sen jakauma ei riipu parametreista μ ja σ^2. Etsi tämän avulla keskihajonnalle σ ylempi 95 %:n luottamusraja eli luottamusväli muotoa $(0, b)$, kun on havaittu $s = 10$.

Ratkaisu. Todennäköisyyslaskennan ja kurssimonisteen kohdan 5.4.3 perusteella tiedetään, että $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ kaikilla $\sigma^2 > 0$. Sen jakauma ei siis riipu parameterin σ^2 arvosta, eli se on saranasuure, ja

$$P\{(n-1)S^2/\sigma^2 > q_\alpha\} = 1 - \alpha,$$

missä q_α on khiin neliön jakauman vapausasteella $n-1$ α-kvantiili (eli piste, jonka vasemmalla puolella on osuus α jakauman todennäköisyysmassasta). Ratkaisemalla tämä keskihajonnan σ suhteen saadaan

$$P\left\{\sigma < \sqrt{\frac{n-1}{q_\alpha}}S\right\} = 1 - \alpha.$$

Siten keskihajonnan σ luottamusväli luottamustasolla $1 - \alpha$ on

$$A(\mathbf{y}) = \left(0, \sqrt{\frac{n-1}{q_\alpha}}s\right).$$

Tehtävän arvoilla 95% luottamusväliksi saadaan

$$\left(0, \sqrt{24/13.85} \cdot 10\right) \approx (0, 13.16).$$

2. Oletetaan, että eräässä mallissa on löydetty tunnusluku eli aineiston muunnos $t = t(\mathbf{y})$ siten, että vastaavalle satunnaismuuttujalle $T = t(\mathbf{Y})$ pätee $T - \theta \sim \text{Tas}(-1, 1)$ kaikilla θ:n arvoilla. Kyseessä on siis *saranasuure* monisteen kohdan 6.2.2 mukaisesti. Johda tämän avulla jokin 95 %:n luottamusväli θ:lle.

Ratkaisu. Tasajakauman tiheysfunktio on

$$f(x) = \begin{cases} \frac{1}{1-(-1)} = \frac{1}{2}, & x \in [-1, 1] \\ 0, & \text{muuten} \end{cases}$$

joten

$$\int_{-0.95}^{0.95} f(x)dx = 2\int_0^{0.95} \frac{1}{2}dx = 0.95.$$

Siten

$$\begin{aligned} &P\{-0.95 \leq T - \theta \leq 0.95\} = 0.95 \\ \Leftrightarrow\ &P\{T - 0.95 \leq \theta \leq T + 0.95\} = 0.95, \end{aligned}$$

joten yksi parametrin θ 95 % luottamusväli on $[t(\mathbf{y}) - 0.95, t(\mathbf{y}) + 0.95]$.

3. (Monisteen teht. 6.2.) Olkoot $Y_1 \perp\!\!\!\perp Y_2$ ja $Y_1 \sim N(\mu_1, 1)$ sekä $Y_2 \sim N(\mu_2, 1)$. Etsi luvut $a, b > 0$ siten, että

$$P\{|Y_1 - \mu_1| \leq a,\ |Y_2 - \mu_2| \leq a\} = 0.95,$$
$$P\{(Y_1 - \mu_1)^2 + (Y_2 - \mu_2)^2 \leq b^2\} = 0.95.$$

Aineisto on $(y_1, y_2) = (1, 0.5)$. Mitkä kaksi 95 %:n luottamusjoukkoa saadaan yo. yhtälöiden perusteella parametriparille (μ_1, μ_2)? Piirrä kuva. Kumpi luottamusjoukoista on mielestäsi parempi? [*Ohje.* Tarvitset jakaumien $N(0, 1)$ ja χ^2_2 taulukoita tai laskimia.]

Ratkaisu. Merkitään $Z_i = Y_i - \mu_i$, jolloin $Z_i \sim N(0, 1)$. Riippumattomuuden nojalla

$$\begin{aligned}
&P\{|Y_1 - \mu_1| \leq a,\ |Y_2 - \mu_2| \leq a\} \\
=&P\{|Y_1 - \mu_1| \leq a\}P\{|Y_2 - \mu_2| \leq a\} \\
=&P\{|Z_1| \leq a\}P\{|Z_2| \leq a\} \\
=&P\{-a \leq Z_1 \leq a\}P\{-a \leq Z_2 \leq a\} \\
=&(2\Phi(a) - 1)^2,
\end{aligned}$$

missä $\Phi(\cdot)$ on standardinormaalijakauman kertymäfunktio. Ratkaistaan a

$$\begin{aligned}
(2\Phi(a) - 1)^2 &= 0.95 \\
\Leftrightarrow\ \Phi(a) &= \frac{1}{2}(\sqrt{0.95} + 1) \\
\Leftrightarrow\ a &= \Phi^{-1}(\frac{1}{2}(\sqrt{0.95} + 1)) \approx 2.24.
\end{aligned}$$

Sijoittamalla a tehtävänannon yhtälöön voidaan ratkaista 95%-luottamusjoukko

$$\begin{aligned}
A(\mathbf{y}) =&\{(\mu_1, \mu_2) : |y_1 - \mu_1| \leq a,\ |y_2 - \mu_2| \leq a\} \\
=&\{(\mu_1, \mu_2) : -a \leq y_1 - \mu_1 \leq a,\ -a \leq y_2 - \mu_2 \leq a\} \\
=&\{(\mu_1, \mu_2) : y_1 - a \leq \mu_1 \leq y_1 + a,\ y_2 - a \leq \mu_2 \leq y_2 + a\} \\
\approx&[-1.24, 3.24] \times [-1.74, 2.74].
\end{aligned}$$

Palautetaan mieleen, että jos $Z_1, Z_2 \sim N(0, 1)$, niin $X = Z_1^2 + Z_2^2 \sim \chi^2_2$. Olkoon siis $X \sim \chi^2_2$. Käyttäen lisäksi edellä määriteltyjä merkintöjä saadaan

$$\begin{aligned}
P\{(Y_1 - \mu_1)^2 + (Y_2 - \mu_2)^2 \leq b^2\} \\
= P\{(Z_1)^2 + (Z_2)^2 \leq b^2\} \\
= P\{X \leq b^2\},
\end{aligned}$$

mistä voidaan ratkaista b käyttäen χ^2_2-jakauman kvantiilifunktiota $F^{-1}_{\chi^2_2}(\cdot)$

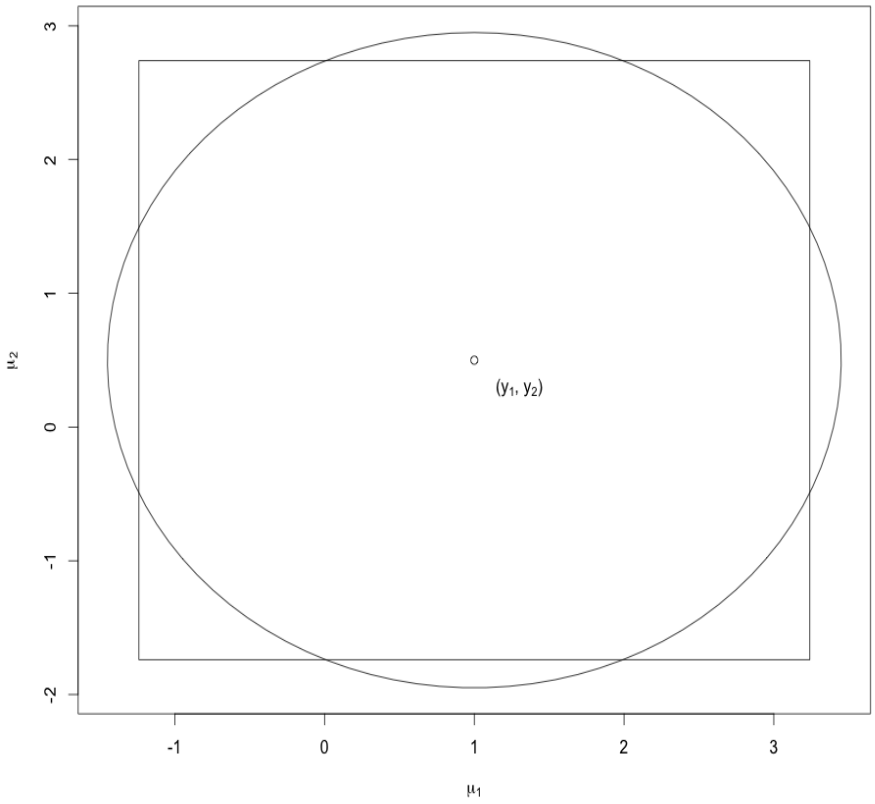
$$\begin{aligned}
P\{X \leq b^2\} &= 0.95 \\
\Leftrightarrow\ b^2 &= F^{-1}_{\chi^2_2}(0.95) \\
\Leftrightarrow\ b &= \sqrt{F^{-1}_{\chi^2_2}(0.95)} \approx 2.45.
\end{aligned}$$

Sijoittamalla b tehtävänannon yhtälöön saadaan 95%-luottamusjoukoksi

$$\begin{aligned}
B(\mathbf{y}) =&\{(\mu_1, \mu_2) : (y_1 - \mu_1)^2 + (y_2 - \mu_2)^2 \leq b^2\} \\
=&\overline{B}((y_1, y_2), b) \\
\approx&\overline{B}((1, 0.5), 2.45),
\end{aligned}$$

missä $\overline{B}((y_1, y_2), b)$ on $\mathbb{R}^2$:n (y_1, y_2) -keskinen suljettu kuula, jonka säde on $b \approx 2.45$.

Luottamusjoukkoja $A(\mathbf{y})$ ja $B(\mathbf{y})$ voidaan vertailla niiden pinta-alojen avulla. Luottamusjoukkoa $A(\mathbf{y})$ vastaava pinta-ala on $4a^2 \approx 20.0$ ja vastaavasti $B(\mathbf{y})$:n pinta-ala on $\pi b^2 \approx 18.8$. Luottamusjoukko $B(\mathbf{y})$ antaa siten tarkemman arvion parametreista μ_1, μ_2. Toisaalta neliön muotoinen luottamusjoukko on selkeämpi esittää.



Kuva 1: Luottamusjoukot parametreille μ_1, μ_2.

4. Vanha koetehtävä. Oletetaan, että $Y_1, \ldots, Y_n$ ovat riippumattomia ja noudattavat kukin jatkuvaa jakaumaa, jonka tiheysfunktio on

$$f(y; \lambda) = 2\lambda y \exp(-\lambda y^2), \qquad y > 0,$$

ja jossa λ on positiivinen parametri.

a) Muodosta tilastollinen malli $f_{\mathbf{Y}}$ ja johda parametrin λ suurimman uskottavuuden estimaattori $\hat{\lambda}$ sekä Fisherin informaatio $i(\lambda)$.

b) Mitä normaalijakaumaa $\hat{\lambda}$ approksimatiivisesti noudattaa, kun n on suuri?

c) Muodosta Waldin testiin eli yo. normaaliapproksimaatioon perustuva approksimatiivinen 95 %:n luottamusväli parametrille λ, kun aineisto on $\mathbf{y} = (y_1, \ldots, y_n)$. Mikä on $\hat{\lambda}$:n keskivirhe?

Ratkaisu. a) Yhteistiheysfunktio on

$$f_{\mathbf{Y}}(\mathbf{y}; \lambda) = \prod_{i=1}^{n} 2\lambda y_i \exp(-\lambda y_i^2) = (2\lambda)^n \left(\prod_{i=1}^{n} y_i \right) \exp(-\lambda \sum_{i=1}^{n} y_i^2),$$

ja kun valitaan $c(\mathbf{y}) = (2^n \prod_{i=1}^n y_i)^{-1}$, uskottavuusfunktioksi saadaan

$$L(\lambda; \mathbf{y}) = \lambda^n \exp(-\lambda \sum_{i=1}^n y_i^2),$$

ja edelleen logaritminen uskottavuusfunktio on

$$l(\lambda; \mathbf{y}) = n \log \lambda - \lambda \sum_{i=1}^n y_i^2.$$

Logaritmisen uskottavuusfunktion derivaatta on

$$l'(\lambda; \mathbf{y}) = \frac{n}{\lambda} - \sum_{i=1}^n y_i^2.$$

Ratkaistaan SU-estimaatti laskemalla tämän nollakohta:

$$\begin{aligned} l'(\lambda; \mathbf{y}) &= \frac{n}{\lambda} - \sum_{i=1}^n y_i^2 = 0 \\ \Leftrightarrow \lambda = \hat{\lambda} &= \frac{n}{\sum_{i=1}^n y_i^2}. \end{aligned}$$

Log-uskottavuusfunktion toinen derivaatta on

$$l''(\lambda; \mathbf{y}) = -\frac{n}{\lambda^2} < 0 \quad \forall \lambda > 0,$$

joten $\hat{\lambda}$ on uskottavuusfunktion globaali maksimikohta ja siten parametrin λ SU-estimaatti. Fisherin informaatioksi saadaan

$$i(\lambda) = E[-l''(\lambda; \mathbf{Y})] = \frac{n}{\lambda^2}.$$

b) SU-estimaattorin asymptoottisen normaalisuuden (monisteen kohta 3.6.5) nojalla

$$\hat{\lambda} \underset{as}{\sim} N\left(\lambda, \frac{1}{i(\lambda)}\right) = N\left(\lambda, \frac{\lambda^2}{n}\right).$$

c) Sijoitetaan ensin SU-estimaatti mallin Fisherin informaatioon

$$i(\hat{\lambda}) = \frac{n}{\hat{\lambda}^2} = n\left(\frac{\sum_{i=1}^n y_i^2}{n}\right)^2 = \frac{\left(\sum_{i=1}^n y_i^2\right)^2}{n},$$

ja lasketaan tämän avulla SU-estimaatin keskivirhe:

$$\text{s.e.}(\hat{\lambda}) = \frac{1}{i(\hat{\lambda})^{1/2}} = \frac{\sqrt{n}}{\sum_{i=1}^n y_i^2}.$$

Siten Waldin 95%:n approksimatiiviseksi luottamusväliksi saadaan (monisteen kohta 6.4.1)

$$\left(\hat{\lambda} - z_{0.025}\,\text{s.e.}(\hat{\lambda}), \hat{\lambda} + z_{0.025}\,\text{s.e.}(\hat{\lambda})\right) = \left(\frac{n - 1.96\sqrt{n}}{\sum_{i=1}^n y_i^2}, \frac{n + 1.96\sqrt{n}}{\sum_{i=1}^n y_i^2}\right).$$

5. Osoitteessa http://www.fsd.uta.fi/menetelmaopetus/paattely/paattely.html#luottamusvali kuvataan luottamusvälin ja luottamustason käsitteitä seuraavasti:

"Tilastollisen päättelyn kaksi keskeistä käsitettä ovat luottamusväli ja luottamustaso. **Luottamusväli** (*confidence interval*) kertoo millä välillä todellinen perusjoukon tunnusluvun arvo on tietyllä todennäköisyydellä. Käyttäen edelleen Nato-kyselyä esimerkkinä, voidaan kuvitella, että otoksessa 45 % kaikista vastaajista ilmoitti kannattavansa Suomen Nato-jäsenyyttä. Koska tähän lukuun

vaikuttavat monet satunnaiset tekijät, emme voi suoraan päätellä, että myös perusjoukossa (kaikki täysi-ikäiset suomalaiset) vastaava osuus on täysin sama. On kuitenkin todennäköistä, että perusjoukon mielipidettä kuvaava arvo on lähellä otoksesta saatua arvoa. Voimme esimerkiksi päätellä, että 95 %:n todennäköisyydellä Nato-jäsenyyttä kannattavien ihmisten osuus perusjoukossa on välillä 40–50 %. Tätä väliä kutsutaan luottamusväliksi.

Luottamustaso (*confidence level*) kertoo, millä todennäköisyydellä perusjoukkoa kuvaava tunnusluku on jollain tietyllä luottamusvälillä. Esimerkiksi 95 %:n todennäköisyydellä 40–50 % suomalaisista haluaa Suomen liittyvän Natoon. Luottamustaso on tällöin 95 %:n todennäköisyys.

Luottamustaso ja luottamusväli ovat siis täysin toisiinsa sitoutuneita käsitteitä. Tieto luottamusvälistä ei ole mielekäs, jos ei ole tietoa luottamustasosta ja päinvastoin. Olennaista on, että luottamustason kasvaessa laajenee myös luottamusväli. Toisin sanoen tämä tarkoittaa siis sitä, että mitä suuremmalla varmuudella haluamme tietää, millä välillä jokin perusjoukon tunnusluku sijaitsee, sitä suurempi on luottamusväli. Jos esimerkiksi haluaisimme tietää, millä välillä suomalaisten Nato-jäsenyyden kannatus on 99 %:n luottamustasolla, luottamusväli olisi suurempi kuin 95 prosentin luottamustasolla (esimerkiksi 30–60 %). Jos olisimme valmiita tyytymään esimerkiksi 90 %:n luottamustasoon, väli voisi olla 43–47 %."

Mikä tässä kuvauksessa on hyvää ja mikä antaa kritiikin aihetta?

Kommentteja

Nea:

- Tekstissä puhutaan todennäköisyyksistä: voisi tarkentaa.
- Luottamusväli määritellään jollain luottamustasolla, eli tekstin kuvaus luottamusvälin ja luottamustason yhteydestä on outo. Siinä on kuitenkin hyvä pointti se, että luottamusväli ei kerro mitään, jos ei tiedä millä luottamustasolla se on määritelty.

Henkka:

- Hyvää: Luottamusvälin leveys on ymmärretty oikein päin: suurempi luottamustaso leventää luottamusväliä.
- Kritiikkiä: Parametriin on liitetty jakauma, joka on sitten mennyt sekaisin luottamusvälin kanssa.

Pekka:

- Ongelmia: Koko ajan puhutaan parametriin (tai siis "perusjoukkoon") liittyvästä tn:stä, mikä ei frekventistisen käsityksen mukaan ole tavallisesti mielekästä. Lause "luottamustaso on tällöin ... todennäköisyys"on tavallaan aika hassu: herää kysymys, miksi lanseerata uusi käsite "luottamus", jos kerran se on sama asia kuin "todennäköisyys"? Toistetusta aineistonkeruusta ei puhuta yhtään mitään, vaikka se on ainoa tapa ymmärtää/tulkita luottamusväliin liittyvä luottamustaso (OK, joku voisi nähdä häivähdyksen toistetun aineistonkeruun ajatusta sivulauseessa "koska tähän lukuun vaikuttavat monet satunnaiset tekijät").
- Hyvää: Puhe luottamusvälistä ei ole hyödyllistä ellei tiedetä, millä luottamustasolla se on muodostettu. Luottamusväli kasvaa kun luottamustaso kasvaa.

6. Keskusteltavaksi:

a) Mainitse pari kolme mielestäsi tärkeintä tai mielenkiintoisinta tilastollisen päättelyn piiriin kuuluvaa käsitettä tai menetelmää, jotka olet tällä kurssilla oppinut. Perustele!

b) Miten käsityksesi tilastollisesta päättelystä on muuttunut tällä kurssilla verrattuna *Johdatus tilastolliseen päättelyyn* -kurssiin (tai vastaavaan)?

Kommentteja

Nea: Su-estimointi on mielestäni tärkeä aihe, sillä sitä tarvitsee muillakin tilastotieteen kursseilla, joten siitä (ja muistakin syistä) se on tärkeä aihe hallita hyvin. Kun suoritin kurssin itse koin kiinnostavimmaksi ehkä hypoteesien testauksen, sillä aihetta ei juurikaa oltu käsitelty aiemmilla kursseilla, mutta siitä puhutaan paljon.

Henkka: Kurssin tärkeimmäksi ja mielenkiintoisimmaksi sisällöksi koen SU-estimaattorin asymptotiikan, informaation käsitteen ja luentomonisteen lopussa käsiteltävän testiteorian. Näiden ymmärtämisestä voi olla paljonkin iloa sekä teoriasta että sovelluksista kiinnostuneille.