Tilastollinen päättely II, syksy 2015 – kevät 2016
Harjoitus 5 (1. ja 3. 12. 2015)

1. Paluu toistokokeeseen. Luennolla ja monisteessa on tarkasteltu kahta erilaista mallia toistokokeelle: Bernoulli-jakaumaan perustuvaa ja binomijakaumaan perustuvaa (monisteen esim. 2.1.2 ja 2.1.5).

Kolmas vaihtoehto: Suoritetaan toistoja, kunnes ennalta päätetty määrä k onnistumisia on sattunut. Olkoon tarvittavien toistojen lukumäärä n, jolloin selvästi $n \geq k$. Johda n:ää vastaavan satunnaismuuttujan N pistetodennäköisyysfunktio, kun yhden toiston onnistumistodennäköisyys on $0 < \theta < 1$. Mikä on havaintoa n vastaava uskottavuusfunktio parametrille θ? Vertaa sitä binomijakaumamallin uskottavuusfunktioon. Eroavatko uskottavuuteen perustuvat päätelmät θ:sta toisistaan näissä malleissa?

Pohdittavaksi: Keksitkö mitään sovellustilannetta, jossa toistokoetta olisi luontevaa lähestyä kolmannen mallin esittämällä tavalla? Voidaanko aina edellyttää toistokokeen tekijältä (esim. puhelinhaastattelija, katugallupin tekijä, lantinheittäjä tai lääkäri, joka haluaa selvittää, kuinka suuri osuus tietyn sydänleikkauksen läpikäyneistä potilaista kuolee ensimmäisen vuorokauden aikana), että hän osaa selvästi kertoa, mihin toistojen tekeminen päättyi: ennalta valitun toistojen määrän tultua täyteen, ennalta valitun "onnistumisten" lukumäärän tultua täyteen vai johonkin muuhun syyhyn (kaveri soitti ja pyysi tulemaan bileisiin, joten tutkimus keskeytyi)?

Ohje. Tapahtuma $\{N = n\}$ merkitsee, että $n-1$ ensimmäisessä toistossa on sattunut $k-1$ onnistumista ja $n-k$ epäonnistumista ja että n:s toisto on ollut onnistuminen. Toistot oletetaan riippumattomiksi. N:n jakaumalla on vakiintunut nimikin; mikähän se on?

Ratkaisu 1

Kuvataan i:nnen kokeen tulosta sm:lla Y_i siten, että $Y_i = 1$, jos koe onnistuu ja $Y_i = 0$, jos se epäonnistuu. Tällöin Y_i:t ovat riippumattomia ja $Y_i \sim B(\theta)$. Tapahtuma $\{N = n\}$ on yhdiste kaikista tapahtumista $\{Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}, Y_n = y_n\}$, joille

$$\sum_{i=1}^{n-1} y_i = k - 1 \text{ ja } y_n = 1.$$

Yhden tällaisen tapahtuman todennäköisyys on Y_i:den riippumattomuuden nojalla

$$\prod_{i=1}^{n} P\{Y_i = y_i\} = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} = \theta^k(1-\theta)^{n-k}.$$

Lisäksi nämä tapahtumat ovat erillisiä, joten niiden yhdisteen todennäköisyys saadaan niiden todennäköisyyden summana, ja koska $k-1$ onnistumista ($Y_n = 1$ aina) voidaan sijoittaa $n-1$:een toistoon $\binom{n-1}{k-1}$ erilaisella tavalla,

$$f_N(n;\theta) = \binom{n-1}{k-1}\theta^k(1-\theta)^{(n-k)}$$

Tätä jakaumaa kutsutaan negatiiviseksi binomijakaumaksi. Sen uskottavuusfunktiosta voidaan jättää parametrista θ riippumaton vakio $\binom{n-1}{k-1}$ pois, jolloin sen uskottavuusfunktioksi saadaan

$$L(\theta;n) = \theta^k(1-\theta)^{n-k}.$$

Päädytään siis samaan uskottavuusfunktioon, kuin jos onnistumisten lukumäärää K n:ssä toistossa mallinnettaisiin binomijakaumalla, eli $K \sim Bin(n,\theta)$. Uskottavuuteen perustuvat päätelmät ovat siis samat kummassakin mallissa.

Tehtävät 2–4 liittyvät harhattomuuteen. Monisteen jakso 3.2.

2. (Vrt. monisteen teht. 3.6.) Olkoot $x_1, x_2, \ldots$ annettuja nollasta eroavia reaalilukuja (selittävän muuttujan arvoja). Tarkastellaan regressiomallia $Y_1, \ldots, Y_n \perp\!\!\!\perp$, $Y_i \sim N(\beta x_i, \sigma^2)$, jossa β ja σ^2 ovat tuntemattomia parametreja.

a) Varmista, että parametrin β su-estimaattori on

$$\hat{\beta} = \sum_{i=1}^{n} x_i Y_i \Big/ \sum_{i=1}^{n} x_i^2.$$

Totea, että se on harhaton.

b) Totea, että myös β:n estimaattori

$$T = \sum_{i=1}^{n} Y_i \Big/ \sum_{i=1}^{n} x_i$$

on harhaton.

Ratkaisu 2

a) Lähdetään selvittämään suurimman uskottavuuden estimaattia. Koska satunnaismuuttujat noudattavat normaalijakaumaa saadaan yhteistiheysfunktioksi

$$f_{\mathbf{Y}}(\mathbf{y}; \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{\sum_{i=1}^{n}(y_i - \beta x_i)^2}{2\sigma^2} \right\}$$

Joten uskottavuusfunktio on

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{\sum_{i=1}^{n}(y_i - \beta x_i)^2}{2\sigma^2} \right\}$$

Log-uskottavuusfunktioksi saadaan

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

Parametrin β suurimman uskottavuuden estimaatti löydetään minimoimalla neliösumma $S(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i)^2$. Etsitään sen derivaatan nollakohdat:

$$S'(\beta) = -2\sum_{i=1}^{n}(x_i y_i) + 2\sum_{i=1}^{n}(x_i^2)\beta = 0 \Leftrightarrow 2\sum_{i=1}^{n}(x_i^2)\beta = 2\sum_{i=1}^{n}(x_i y_i) \Leftrightarrow \beta = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sum_{i=1}^{n}(x_i^2)}$$

Joka on minimikohta, sillä

$$S''(\beta) = 2\sum_{i=1}^{n}(x_i^2) > 0$$

Joten $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sum_{i=1}^{n}(x_i^2)}$. Saatu estimaatti $\hat{\beta}$ on harhaton sillä

$$E(\hat{\beta}) = E\left(\frac{\sum_{i=1}^{n}(x_i Y_i)}{\sum_{i=1}^{n}(x_i^2)} \right) = \frac{\sum_{i=1}^{n}(x_i E(Y_i))}{\sum_{i=1}^{n}(x_i^2)} = \frac{\sum_{i=1}^{n}(x_i^2)\beta}{\sum_{i=1}^{n}(x_i^2)} = \beta$$

b) Myös estimaatti $T = T = \sum_{i=1}^n Y_i \Big/ \sum_{i=1}^n x_i$ on harhaton, sillä

$$E(T) = E\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right) = \frac{\sum_{i=1}^n E(Y_i)}{\sum_{i=1}^n x_i} = \frac{\beta \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i} = \beta$$

3. (Monisteen teht. 3.4.) Havainnoista $Y_1, \ldots, Y_n$ oletetaan samoin kuin harjoituksen 4 tehtävässä 4: ne ovat riippumattomia ja noudattavat jakaumaa, jolla on odotusarvo μ ja varianssi σ^2. Keksi jokin harhaton estimaattori odotusarvon neliölle μ^2.

Ratkaisu 3

Tiedetään, että

$$E(\overline{Y}^2) = \mathrm{var}\overline{Y} + \mu^2 = \frac{\sigma^2}{n} + \mu^2.$$

Hyödynnetään viikon 4 tehtävän 4 tulosta poistamaan ensimmäinen termi ja valitaan satunnaismuuttuja

$$Z := \overline{Y}^2 - \frac{S^2}{n},$$

jossa $S^2 = \frac{1}{n-1}\sum_{i=1}^n y_i$ on parametrin σ^2 harhaton estimaattori.
Tälle pätee odotusarvon lineaarisuuden nojalla

$$E(Z) = E(\overline{Y}^2 - S^2/n) = \frac{\sigma^2}{n} + \mu^2 - \frac{1}{n}\sigma^2 = \mu^2.$$

Siis $Z = \overline{Y}^2 - S^2/n$ on eräs parametrin μ^2 harhaton estimaattori.

4. (Vrt. monisteen teht. 3.10.) Mallissa $Y_1, \ldots, Y_n \sim \mathrm{Tas}(0, \theta)$ $\perp\!\!\!\perp$ on su-estimaattoriksi saatu $\hat{\theta} = \max(Y_1, \ldots, Y_n)$ (ks. kohta 2.2.8).

a) Muodosta $\hat{\theta}$:n kertymäfunktio F lähtien havainnosta

$$P\{\hat{\theta} \leq t\} = P\{Y_1 \leq t\} \cdots P\{Y_n \leq t\}$$

ja derivoi siitä tiheysfunktio $f = F'$.

b) Laske $\hat{\theta}$:n odotusarvo ja totea, että $\hat{\theta}$ on harhainen mutta asymptoottisesti harhaton.

Ratkaisu 4

a) $\hat{\theta} = \max\{Y_1, \ldots, Y_n\} \leq y$, jos ja vain jos $Y_i \leq y$ kaikilla $i = 1, \ldots, n$. Siten riippumattomuuden nojalla

$$F(y) = P(\hat{\theta} \leq y) = P(Y_1 \leq y, \ldots, Y_n \leq y) = P(Y_1 \leq y) \cdots P(Y_n \leq y).$$

Koska kaikilla $i = 1, \ldots, n$

$$P(Y_i \leq y) = \begin{cases} 0, & y \in (-\infty, 0], \\ \theta^{-1} y, & y \in (0, \theta), \\ 1, & y \in [\theta, \infty), \end{cases}$$

niin

$$F(y) = \begin{cases} 0, & y \in (-\infty, 0], \\ \theta^{-n} y^n, & y \in (0, \theta), \\ 1, & y \in [\theta, \infty). \end{cases}$$

Derivoimalla saadaan siis

$$f(y) = F'(y) = \begin{cases} 0, & y \in (-\infty, 0], \\ \theta^{-n} n y^{n-1}, & y \in (0, \theta), \\ 0, & y \in [\theta, \infty). \end{cases}$$

Koska tiheysfunktiota f käytetään ainoastaan integraaleissa, ei haittaa että F' ei ole määritelty pisteissä 0 ja θ. Itse asiassa voidaan asettaa $f(0) := f(\theta) := 0$, jolloin $f(y) = \theta^{-n} n y^{n-1} I_{(0,\theta)}(y),\ y \in \mathbb{R}$.

b) Odotusarvon määritelmän perusteella saadaan

$$E(\hat{\theta}) \quad = \quad \int_{-\infty}^{\infty} y f(y) dy = \frac{n}{\theta^n} \int_0^{\theta} y^n dy = \frac{n}{n+1} \cdot \frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1}\theta, \quad \theta > 0.$$

Estimaattori $\hat{\theta}$ on siis harhainen mutta asymptoottisesti harhaton, koska

$$\lim_{n\to\infty} E(\hat{\theta}) = \lim_{n\to\infty} \frac{n}{n+1}\theta = \theta.$$

Tehtävää 5 varten tutustu alustavasti monisteen jakson 3.4 alkuun.

5. (Monisteen teht. 3.1.) Johda funktion $g(\theta)$ estimaattorin T keskineliövirheelle monisteen kohdassa 3.4.1 mainittu hajotelma

$$E_\theta[(T - g(\theta))^2] = Var_\theta(T) + b(\theta)^2,$$

jossa $b(\theta)$ on estimaattorin harha.

Ratkaisu 5

Laskemalla neliö auki, sijoittamalla estimaattorin harhan kaavasta saatava

$$g(\theta) = E_\theta(T) - b(\theta),$$

saadaan:

$$\begin{aligned} E_\theta[(T - g(\theta))^2] &= E_\theta(T^2) - 2g(\theta)E_\theta(T) + g(\theta)^2 \\ &= E_\theta(T^2) - 2(E_\theta(T) - b(\theta))E_\theta(T) + (E_\theta(T) - b(\theta))^2 \\ &= E_\theta(T^2) - 2(E_\theta(T))^2 + 2b(\theta)E_\theta(T) + (E_\theta(T))^2 - 2b(\theta)E_\theta(T) + b(\theta)^2 \\ &= E_\theta(T^2) - (E_\theta(T))^2 + b(\theta)^2 \end{aligned}$$

ja lopuksi soveltamalla varianssin kaavaa

$$var(T) = E(T^2) - (E(T))^2,$$

saadaan:

$$E_\theta[(T - g(\theta))^2] = E_\theta(T^2) - (E_\theta(T))^2 + b(\theta)^2 = var_\theta(T) + b(\theta)^2$$