Luku 12

Väliestimointi

Estimaatteihin liittyy aina tietty epävarmuus, joka johtuu mm. otosvaihtelusta. Jos estimaattori on harhaton ja sen varianssi pieni, voidaan estimaattien odottaa osuvan lähelle parametrin arvoa. Väliestimoinnissa ilmoitetaan väli, jolle parametrin arvon arvioidaan kuuluvan, ja lisäksi ilmoitetaan väliin liittyvä luottamus tai varmuuden aste. Tässä luvussa tarkastellaan lutottamusvälejä. Merkitään otosta $X_1, \ldots, X_n$ lihavoidulla isolla kirjaimella $\boldsymbol{X}$ ja havaittua otosarvoa $x_1, \ldots, x_n$ lihavoidulla pikkukirjaimella $\boldsymbol{x}$. Aloitamme väliestimaatin määritelmällä.

Määritelmä 12.1 Olkoon $\boldsymbol{X}$ otos jostain jakaumasta F ja $\theta = \theta(F)$ on sen reaaliarvoinen parametri. On havaittu $\boldsymbol{X} = \boldsymbol{x}$. Parametrin θ *väliestimaattori* on mikä tahansa sellainen otoksen tunnuslukupari $l(\boldsymbol{X})$ ja $u(\boldsymbol{X})$, että $l(\boldsymbol{x}) \leq u(\boldsymbol{x})$ kaikilla mahdollisilla otosarvoilla $\boldsymbol{X} = \boldsymbol{x}$. Havaittu väli $[l(\boldsymbol{x}), u(\boldsymbol{x})]$ on θ:n väliestimaatti ja $[l(\boldsymbol{X}), u(\boldsymbol{X})]$ väliestimaattori.

Merkitään lyhyesti $l(\boldsymbol{X}) = L,\ l(\boldsymbol{x}) = l,\ u(\boldsymbol{X}) = U$ ja $u(\boldsymbol{x}) = u$. Olemme kiinnostuneita siitä, onko parametri θ välillä $[l, u]$ eli peittääkö väli parametrin arvon. Parametrin θ arvo on kiinteä ja väliestimaatti saadaan jonkin satunnaiskokeen tuloksena. Todennäköisyys

$$P_F[L \leq \theta \leq U]$$

on väliestimaattorin $[L, U]$ *peitetodennäköisyys*. Merkintä P_F tarkoittaa, että todennäköisyys lasketaan jakaumasta F, josta otos on tehty. Satunnaisväliä $[L, U]$ sanotaan parametrin θ luottamusväliksi luottamustasolla $(1-\alpha)100\%$, jos

$$P_F[L \leq \theta \leq U] \geq 1 - \alpha, \quad 0 < \alpha < 1,$$

kaikilla $\theta \in \Theta$. Luottamusvälin peitetodennäköisyys riippuu tavallisesti estimoitavan parametrin θ arvosta.

12.1 Keskiarvojen luottamusvälit

Olkoon $X_1, \ldots, X_n$ otos jakaumasta, jonka keskiarvo on μ ja varianssi σ^2). Tarkastelemme luottamusvälin muodostamista keskiarvolle μ.

Esimerkki 12.1 Olkoon $X_1, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$. Otoskeskiarvo $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$ on jakauman tuntemattoman keskiarvon μ harhaton estimaattori. Muodostamme $\bar{X}$:n avulla tuntemattomalle μ:lle luottamusvälin, kun varianssi $\sigma^2 = \sigma_0^2$ tunnetaan. Voimme normaalijakauman avulla määrittää sellaisen luvun $z_{\alpha/2}$, että

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha.$$

Huomaa, että $P(Z \geq z_{\alpha/2}) = P(Z \leq -z_{\alpha/2}) = \alpha/2$. Jos esimerkiksi $1 - \alpha = 0.95$, niin $z_{\alpha/2} = z_{0.025} = 1.96$. Todennäköisyys, että satunnaisväli

$$[\bar{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}]$$

sisältää tuntemattoman keskiarvon μ on $1 - \alpha$.

Kun otos on havaittu ja saatu havaintoarvot $x_1, \ldots, x_n$, voidaan laskea μ:n estimaatti $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Siitä saadaan tunnettu väli

$$[\bar{x} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}],$$

jota sanotaan μ:n $100(1 - \alpha)\%$:n luottamusväliksi. Lukua $1 - \alpha$ tai vastaavasti prosenttilukua $100(1 - \alpha)\%$ sanotaan välin luottamustasoksi tai luottamuskertoimeksi. Jos esimerkiksi $\bar{x} = 3.25$, $\sigma = 1$ ja $n = 20$, niin $3.25 \pm 1.96\frac{1}{\sqrt{20}} = [2.81, 3.69]$ on μ:n 95%:n luottamusväli. □

Vaikka ei voitaisikaan olettaa, että otos on normaalijakaumasta, voidaan silti usein saada luottamusvälin likiarvo. Keskeisen rajaväittämän nojalla $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ noudattaa likimain normaalijakaumaa $\mathrm{N}(0, 1)$, kun n on suuri. Silloin

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}) \approx 1 - \alpha.$$

ja väli

$$[\bar{x} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}],$$

on likimain $100(1-\alpha)\%$:n luottamusväli. Likiarvon täsmällisyys riippuu otoskoosta n ja jakaumasta, josta otos on peräisin.

Jos σ^2 on tuntematon ja otoskoko kohtuullisen suuri ($n \geq 30$), noudattaa $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ likimain normaalijakaumaa, vaikka otos ei ole peräisin normaalijakaumasta. Koska otosvarianssi S^2 on σ^2:n tarkentuva estimaattori, eli $S^2 \xrightarrow{P} \sigma^2$ otoskoon n kasvaessa, seuraa tulos keskeisestä rajaväittämästä ja Slutskyn lauseesta (Lause 9.22). Jos perusjakauma, josta otos tehdään, on esimerkiksi voimakkaasti vino, saattaa vielä otoskoko 30 olla liian pieni, jotta likiarvo olisi hyvä.

Esimerkki 12.2 Jos otos on normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, niin

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

noudattaa t-jakaumaa vapausastein $n-1$. Silloin t-jakaumasta vapausastein $n-1$ voidaan määrittää luku $t_{\alpha/2;n-1}$ siten, että

$$P(-t_{\alpha/2;n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2;n-1}) = 1 - \alpha.$$

Otoksesta laskettujen estimaattien $\bar{x}$ ja s^2 perusteella saadaan μ:n $100(1-\alpha)\%$:n luottamusväli

$$[\bar{x} - t_{\alpha/2;n-1}\frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}]. \tag{12.1.1}$$

Jos ei voida otaksua otoksen olevan normaalijakaumasta, on väli (12.1.1) vain likimain μ:n $100(1-\alpha)\%$:n luottamusväli. Likiarvo ei ole herkkä poikkeamiselle normaalisuusoletuksesta. □

Joissain sovelluksissa tarvitaan vain esimerkiksi μ:n alarajan (ylärajan) arvio. Olkoon otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$. Silloin

$$P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha) = 1 - \alpha$$

tai vastaavasti

$$P[\bar{X} - z_\alpha\frac{\sigma}{\sqrt{n}} \leq \mu] = 1 - \alpha.$$

Kun $\bar{X}$:n arvo on havaittu, saadaa μ:n yksipuolinen $100(1-\alpha)\%$:n luottamusväli $(\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty)$. Tämä yksipuolinen väli antaa μ:lle alarajan.

Usein luottamusväli voidaan muodostaa laskennallisesti helpoimmin suurimman uskottavuuden estimaattorin jakauman likiarvon avulla. Tuloksen (11.2.2) nojalla suurimman uskottavuuden estimaattori noudattaa asymptoottisesti normaalijakaumaa. Välin

$$\left[\hat{\theta} \pm \frac{c}{\sqrt{I(\hat{\theta})}}\right]$$

peitetodennäköisyys on

$$P_\theta\left(\hat{\theta} - \frac{c}{\sqrt{I(\hat{\theta})}} \leq \theta \leq \hat{\theta} + \frac{c}{\sqrt{I(\hat{\theta})}}\right) = P_\theta\left(-c \leq (\hat{\theta} - \theta)\sqrt{I(\hat{\theta})} \leq c\right).$$

Tuloksen (11.2.4) mukaan

$$P_\theta\left(-c \leq (\hat{\theta} - \theta)\sqrt{I(\hat{\theta})} \leq c\right) \approx P(-c \leq Z \leq c),$$

missä $Z \sim \mathrm{N}(0,1)$. Jos peitetodennäköisyydeksi asetetaan 0.95, niin väli

$$\left[\hat{\theta} - \frac{1.96}{\sqrt{I(\hat{\theta})}},\ \hat{\theta} + \frac{1.96}{\sqrt{I(\hat{\theta})}}\right] \tag{12.1.2}$$

on θ:n luottamusväli, jonka luottamustaso on likimain 95 %:n, koska $P(-1.96 \leq Z \leq 1.96) = 0.95$.

Esimerkki 12.3 Oletetaan, että jakaumasta $\mathrm{Bin}(100, \theta)$ on saatu havainto $x = 17$. Lasketaan θ:n likimääräinen 95 %:n luottamusväli. Nyt θ:n suurimman uskottavuuden estimaatti on $\hat{\theta} = x/n = 0.17$ ja

$$l(\theta) - l(\hat{\theta}) = 17 \log \theta + 83 \log(1-\theta) + 45.581, \qquad 0 < \theta < 1.$$

Laskemalla voidaan todeta, että $l(\theta) - l(\hat{\theta}) \geq \log 0.147$, kun $0.105 \leq \theta \leq 0.251$. Tämä on θ:n 14.7 %:n uskottavuusväli ja likimain 95 %:n luottamusväli.

Informaatiofunktio on

$$I(\theta) = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}, \qquad 0 < \theta < 1.$$

Sijoittamalla informaatiofunktioon $\theta = \hat{\theta}$ saadaan

$$I(\hat{\theta}) = \frac{x}{\hat{\theta}^2} + \frac{n-x}{(1-\hat{\theta})^2} = \frac{n}{\hat{\theta}} + \frac{n}{1-\hat{\theta}} = \frac{n}{\hat{\theta}(1-\hat{\theta})}.$$

Nyt (12.1.2):n mukaan

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.17 \pm 0.0736$$

on θ:n likimääräinen 95 %:n luottamusväli. Väli ei kuitenkaan ole uskottavuusväli, sillä välin alarajan 0.096 suhteellinen uskottavuus $R(0.096) = 0.072$ on paljon pienempi kuin ylärajan 0.244 suhteellinen uskottavuus $R(0.244) = 0.200$. □

Esimerkki 12.4 Oletetaan, että $X \sim \mathrm{Bin}(3, \theta)$. Silloin

$$\begin{aligned} l(\theta) &= x \log \theta + (3-x) \log(1-\theta) \\ &= 3[\hat{\theta} \log \theta + (1-\hat{\theta}) \log(1-\theta)] \quad \text{ja} \\ l(\hat{\theta}) &= 3[\hat{\theta} \log \hat{\theta} + (1-\hat{\theta}) \log(1-\hat{\theta})], \end{aligned}$$

missä $\hat{\theta} = x/3$ ja $0 \leq \theta \leq 1$. Estimaattori $\hat{\theta}$ voi saada arvot 0, $\frac{1}{3}$, $\frac{2}{3}$ ja 1. Parametrin 10 %:n uskottavuusväli on

$$\mathrm{uv}(x; 10\,\%) = \{\, \theta \mid l(\theta) - l(\hat{\theta}) \geq \log 0.1 \,\}.$$

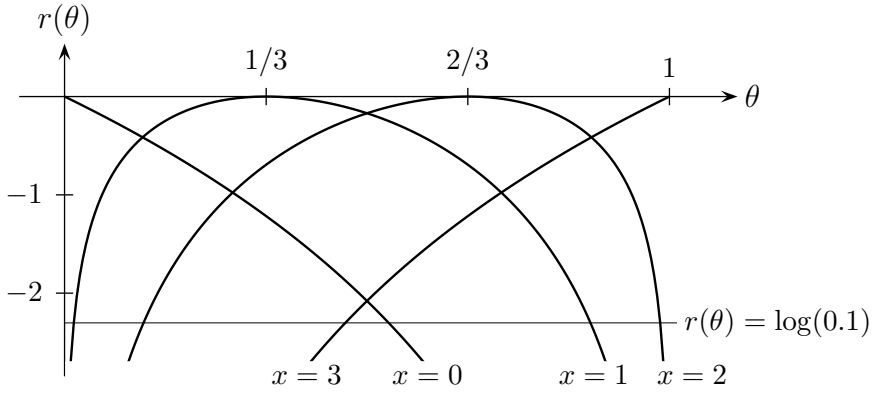
Jos esimerkiksi $x = 1$, niin

$$\begin{aligned} \mathrm{uv}(x; 10\,\%) &= \{\, \theta \mid \log\theta + 2\log(1-\theta) \geq -0.39 \,\} \\ &= [0.015, 0.869]. \end{aligned}$$

Eri x:n arvoilla saadaan seuraavat θ:n 10 %:n uskottavuusvälit

$$\begin{aligned} \mathrm{uv}(0) &= [0, 0.536], \\ \mathrm{uv}(1) &= [0.015, 0.869], \\ \mathrm{uv}(2) &= [0.131, 0.985], \\ \mathrm{uv}(3) &= [0.464, 1]. \end{aligned}$$

Todennäköisyys, että väli peittää parametrin todellisen arvon, riippuu nyt parametrin arvosta.



Kuvio 12.1. Logaritmoitu normitettu uskottavuusfunktio $r(\theta) = l(\theta) - l(\hat{\theta})$, kun $x = 0, 1, 2$ ja 3.

Normaalijakaumaan perustuvan likiarvon avulla johdettu tavanomainen 95 %:n luottamusväli on muotoa

$$l_V(\hat{\theta}) = \hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{3}}.$$

Koska $\hat{\theta}$:n mahdolliset arvot ovat 0, $\frac{1}{3}$, $\frac{2}{3}$ ja 1, niin mahdolliset luottamusvälit ovat

$$0, \quad [0.061, 0.605], \quad [0.395, 0.939] \quad \text{ja} \quad 1.$$

Havaintoarvoilla $X = 0$ ja $X = 3$ väli degeneroituu yhdeksi pisteeksi. Kun $0 < \theta < 0.061$ tai $0.939 < \theta < 1$, on luottamusvälin peitetodennäköisyys 0. □

12.1.1 Napasuureet

Olkoon $X_1, \ldots, X_n$ otos tasajakaumasta $\mathrm{Tas}(0, \theta)$ ja olkoon $Y = X_{(n)}$ havaintojen maksimi. Muodostetaan tuntemattomalle parametrille θ väliestimaattori. Tarkastellaan kahta vaihtoehtoa:

$$\begin{aligned} &[Y, aY], && a > 1; \\ &[Y, Y + b], && b > 0, \end{aligned}$$

missä a ja b ovat annettuja vakioita. Ensimmäisen välin peitetodennäköisyys on

$$P(\theta \in [Y, aY]) = P(Y \leq \theta \leq aY) = P\left(\frac{1}{a} \leq \frac{Y}{\theta} \leq 1\right).$$

Koska Y:n tiheysfunktio on $f_Y(y) = ny^{n-1}/\theta^n$, $0 \leq y \leq \theta$, niin satunnaismuuttujan $T = Y/\theta$ tiheysfunktio on $f_T(t) = nt^{n-1}$, $0 \leq t \leq 1$. Siksi peitetodennäköisyys on

$$P\left(\frac{1}{a} \leq T \leq 1\right) = \int_{1/a}^{1} nt^{n-1}\,\mathrm{d}t = 1 - \left(\frac{1}{a}\right)^n.$$

Peitetodennäköisyys ei riipu θ:sta ja siksi välin $[Y, aY]$ luottamustaso on $1 - (1/a)^n$ kaikilla $\theta > 0$.

Toisen välin peitetodennäköisyys on

$$\begin{aligned} P(\theta \in [Y, Y + b]) &= P(Y \leq \theta \leq Y + b) \\ &= P\left(1 - \frac{b}{\theta} \leq T \leq 1\right) \\ &= \int_{1-b/\theta}^{1} nt^{n-1}\,\mathrm{d}t = 1 - \left(1 - \frac{b}{\theta}\right)^n. \end{aligned}$$

Tässä tapauksessa peitetodennäköisyys riippuu parametrista θ.

Sanomme satunnaismuuttujaa $T = t(\hat{\theta}; \theta)$ *napasuureeksi* (*pivotal quantity* tai *pivot*), jos T:n jakauma ei riipu parametrista θ. Tässä $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ on θ:n estimaattori. Napasuureen avulla voidaan siis konstruoida luottamusvälejä, joiden peitetodennäköisyys ei riipu estimoitavasta parametrista.

Esimerkki 12.5 Tarkastellaan nyt normaalijakauman $\mathrm{N}(\mu, \sigma^2)$ parametrin σ^2 väliestimointia. Koska

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \mathrm{Khi2}(n-1)$$

on napasuure, niin

$$\begin{aligned} 1 - \alpha &= P(a \leq V \leq b), \qquad a < b \\ &= P\left(\frac{a}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{(n-1)S^2}\right) = P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) \end{aligned}$$

on σ^2:n luottamusvälin $[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}]$ luottamustaso. □

12.2 Kahden keskiarvon erotuksen luottamusvälit

Olkoot $X_1, \ldots, X_n$ ja $Y_1, \ldots, Y_m$ kaksi riippumatonta otosta, joista ensimmäinen on normaalijakaumasta $\mathrm{N}(\mu_X, \sigma_X^2)$ ja toinen normaalijakaumasta $\mathrm{N}(\mu_Y, \sigma_Y^2)$. Oletetaan, että varianssit σ_X^2 ja σ_Y^2 tunnetaan. Koska otokset ovat riippumattomat, niin myös otoskeskiarvot $\bar{X}$ ja $\bar{Y}$ ovat riippumattomat ja niiden jakaumat ovat $\bar{X} \sim \mathrm{N}(\mu_X, \sigma_X^2/n)$ ja $\bar{Y} \sim \mathrm{N}(\mu_Y, \sigma_Y^2/m)$. Otoskeskiarvojen erotuksen $W = \bar{X} - \bar{Y}$ jakauma on $\mathrm{N}(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$ ja

$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Kun havainnot on tehty, saadaan havaitut otoskeskiarvot $\bar{x}$ ja $\bar{y}$ ja keskiarvojen erotuksen $100(1-\alpha)\%$:n luottmusväli on

$$[\bar{x} - \bar{y} - z_{\alpha/2}\sigma_W, \bar{x} - \bar{y} + z_{\alpha/2}\sigma_W],$$

missä σ_W on W:n hajonta.

Jos variansseja σ_X^2 ja σ_Y^2 ei tunneta, mutta otoskoot n ja m ovat suuret, niin varianssit σ_X^2 ja σ_Y^2 voidaan korvata varianssien harhattomilla estimaateilla s_x^2 ja s_y^2. Silloin saadaan likimääräinen $100(1-\alpha)\%$:n luottmusväli

$$\bar{x} - \bar{y} - z_{\alpha/2}\sigma_W, \bar{x} - \bar{y} + z_{\alpha/2}s_W,$$

missä $s_W = \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$ on W:n hajonnan estimaatti.

Tarkastellaan seuraavaksi kahden normaalijakauman keskiarvojen erotuksen luottamusvälin määrittämistä, kun variansseja ei tunneta ja otoskoot ovat pienet. Olkoon $X_1, \ldots, X_n$ otosnormaalijakaumasta $\mathrm{N}(\mu_X, \sigma_X^2)$ ja $Y_1, \ldots, Y_m$ normaalijakaumasta $\mathrm{N}(\mu_Y, \sigma_Y^2)$ ja otokset ovat toisistaan riippumattomat. Käsitellään ensin tilannetta, jossa voidaan olettaa $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Silloin satunnaismuuttuja

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}$$

noudattaa normaalijakaumaa $\mathrm{N}(0, 1)$.

Koska otokset ovat toisistaan riippumattomat, niin

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}$$

on kahden riippumattoman Khi2-jakaumaa noudattavan satunnaismuuttujan summa ja $U \sim \mathrm{Khi2}(n + m - 2)$. Määritemän mukaan

$$T = \frac{Z}{U/(n+m-2)}$$

noudattaa t-jakaumaa vapausastein $n+m-2$. Kun tähän sijoitetaan edellä esitetyt Z:n ja U:n lausekkeet, saadaan

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

missä

$$S_P = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}.$$

Nyt

$$P(t_{\alpha/2;n+m-2} \leq T \leq t_{\alpha/2;n+m-2}) = 1 - \alpha$$

ja siksi

$$P\big(|\bar{X} - \bar{Y} - (\mu_X - \mu_Y)| \leq t_{\alpha/2;n+m-2}S_P\sqrt{\frac{1}{n} + \frac{1}{m}}\big) = 1 - \alpha.$$

Jos $\bar{x}, \bar{y}$ ja s_P ovat satunnaismuuttujien $\bar{X}, \bar{Y}$ ja S_P javaitut arvot, niin saadaan $(\mu_X - \mu_Y)$:n $100(1-\alpha)\%$:n luottamusväli

$$[\bar{x} - \bar{y} - t_{\alpha/2;n+m-2}s_P\sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{x} - \bar{y} + t_{\alpha/2;n+m-2}s_P\sqrt{\frac{1}{n} + \frac{1}{m}}].$$

Jos tunnetaan varianssien suhde σ_X^2/σ_Y^2, voidaan keskiarvojen erotukselle $(\mu_X - \mu_Y)$ johtaa luottamusväli t-jakauman avulla vastaavasti kuin tilanteessa $\sigma_X^2 = \sigma_Y^2$. Jos kuitenkaan varianssien suhdetta ei tunneta, tarkastellaan suuretta

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}}.$$

Jos n ja m ovat tarpeeksi suuria, niin W noudattaa likimain normaalijakaumaa ja

$$P(z_{\alpha/2} \leq W \leq z_{\alpha/2}) \approx 1 - \alpha.$$

Jos n ja m eivät ole kovin suuria, käyteetään t-jakaumaan perustuvaa Welchin likiarvoa. Lasketaan

$$r = \frac{(s_x^2/n + s_y^2/m)^2}{\frac{(s_x^2/n)^2}{n-1} + \frac{(s_y^2/m)^2}{m-1}}$$

ja r pyöristetään alspäin lähimpään kokonaislukuun $(= \lfloor r \rfloor)$. Silloin $(\mu_X - \mu_Y)$:n likimäärin $100(1-\alpha)\%$:n luottamusväli on

$$\bar{x} - \bar{y} \pm t_{\alpha/2;\lfloor r \rfloor}\sqrt{s_x^2/n + s_y^2/m}.$$

Joissakin sovelluksissa mittaukset X ja Y ovat toisistaan riippuvat. Mitataan esimerkiksi n:n henkilön paino ennen ja jälkeen laihdutuskuurin ja saadaan mittaukset $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, joka on otos kaksiulotteisesta jakaumasta. Silloin eri mittausparit (X_i, Y_i) ja (X_j, Y_j), $i \neq j$, ovat

toisistaan riippumattomat, mutta mittaukset X_i ja Y_i ovat riippuvia. Muodostetaan erotukset $D_i = X_i - Y_i,\ i = 1, 2, \ldots, n$. Usein voidaan olettaa, että $D_1, D_2, \ldots, D_n$ on otos normaalijakaumasta $\mathrm{N}(\mu_X - \mu_Y, \sigma_D^2)$, missä σ_D^2 on erotusten varianssi. Erityisesti, jos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ on otos kaksiulotteisesta normaalijakaumasta $N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, niin $D_1, D_2, \ldots, D_n$ on otos normaalijakaumasta $\mathrm{N}(\mu_X - \mu_Y, \sigma_D^2)$. Silloin $(\mu_X - \mu_Y)$:n luottamusväli voidaan muodostaa suureen

$$T = \frac{\bar{D} - (\mu_X - \mu_Y)}{S_D/\sqrt{n}}$$

avulla, missä $\bar{D}$ on erotusten otoskeskiarvo ja S_D^2 erotusten otosvarianssi. Erotuksen $\mu_X - \mu_Y$) $100(1-\alpha\%$:n luottamusväli on

$$\bar{d} \pm t_{\alpha/2;n-1} \frac{s_d}{\sqrt{n}}$$

missä $\bar{d}$ on havaintojen keskiarvo ja s_d niiden hajonta.

12.3 Suhteellisten osuuksien luottamusvälit

Olkoon $X_1, X_2, \ldots, X_n$ otos Bernoullin jakaumasta $\mathrm{Ber}(p)$. Silloin onnistumisten lukumäärä $Y = X_1 + \cdots X_n$ noudattaa Binomijakaumaa $\mathrm{Bin}(n, p)$. Otoskeskiarvo Y/n on parametrin p harhaton estimaattori. Suure

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{Y/n - p}{\sqrt{p(1-p)/n}}$$

noudattaa keskeisen rajaväittämän nojalla likimain normaalijakaumaa $\mathrm{N}(0, 1)$, kun n on tarpeeksi suuri. Voimme siis olettaa, että

$$P\big(-z_{\alpha/2} \le \frac{Y/n - p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}\big) \approx 1 - \alpha, \tag{12.3.1}$$

josta saadaan

$$P\big[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\big] \approx 1 - \alpha.$$

Koska epäyhtälön päätepisteissä esiintyy estimoitava tuntematon parametri p, ei tästä tuloksesta saada suoraan luottamusväliä. Tarvitaan toinen likiarvo, korvataan päätepisteissä p harhattomalla estimaatorilla Y/n. Suurella havaintojen lukumäärällä n pitää edelleen paikkansa, että

$$P\big[\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}}\big] \approx 1 - \alpha.$$

Jos havaitaan $Y = y$, niin suurilla n:n arvoilla p:n likimääräinen $100(1-\alpha)\%$:n luottamusväli on

$$\frac{y}{n} \pm z_{\alpha/2}\sqrt{\frac{(y/n)(1-y/n)}{n}}.$$

Muodostamme nyt luottamusvälin kahden onnistumistodennäköisyyden p_1 ja p_2 erotukselle $p_1 - p_2$. Oletetaan, että onnistumisten lukumäärät Y_1 ja Y_2 kahdessa riippumattomassa kokeessa noudattavat binomijakaumaa siten, että $Y_i \sim \text{Bin}(n_i, p_i),\ i = 1, 2$. Satunnaismuuttujat Y_1 ja Y_2 ovat siis riippumattomat. Koska Y_i/n_i on p_i:n, $i = 1, 2$ harhaton estimaattori ja Y_1 ja Y_2 ovat riippumattomat, niin $Y_1/n_1 - Y_2/n_2$ on $p_1 - p_2$:n harhaton estimaattori ja

$$\text{Var}(Y_1/n_1 - Y_2/n_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Voidaan osoittaa, että suure

$$\frac{(Y_1/n_1 - Y_2/n_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

noudattaa likimain normaalijakaumaa $\text{N}(0, 1)$, kun n on suuri. Jos nimittäjässä p_1 ja p_2 korvataan estimaattoreillaan, on

$$P(-z_{\alpha/2} \leq \frac{(Y_1/n_1 - Y_2/n_2) - (p_1 - p_2)}{\sqrt{Y_1/n_1(1-Y_1/n_1)/n_1 + Y_2/n_2(1-Y_2/n_2)/n_2}} \leq z_{\alpha/2}) \approx 1 - \alpha.$$

suurilla n:n arvoilla. Tästä saadaan erotuksen $p_1 - p_2$ likimain $100(1-\alpha)\%$:n luottamusväli

$$\frac{y_1}{n_1} - \frac{y_2}{n_2} \pm z_{\alpha/2}\sqrt{\frac{y_1(1-y_1)}{n_1} + \frac{y_2(1-y_2)}{n_2}}.$$

12.4 Otoskoko

Jos haluamme, että odotusarvon μ luottamusväli, $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$, ei ole pidempi kuin annettu väli $\bar{x} \pm \varepsilon$, niin asetetaan

$$\varepsilon = \frac{z_{\alpha/2}\sigma}{\sqrt{n}},\ \text{josta seuraa} \quad n = \frac{z_{\alpha/2}^2\sigma^2}{\varepsilon^2}.$$

Suuretta $\varepsilon = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$ kutsutaan usein estimaatin maksimivirheeksi. Suhteellisen osuuden p likimäärin $100(1-\alpha)\%$:n luottamusväli on

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

missä $\hat{p} = y/n$. Haluamme määrittää otoskoon niin, että estimaatin $\hat{p} = y/n$ maksimivirhe on $\varepsilon = z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. Koska $\hat{p}$ on tuntematon ennen koetta, siitä ei ole hyötyä otoskoon määrittämisessä. Jos tiedetään, että p:n arvo on noin p^* tai korkeintaan p^*, niin silloin tarvittava otoskoko

$$n = \frac{z^2_{\alpha/2} p^*(1-p^*)}{\varepsilon^2}.$$

Jos meillä ei ole riittävän luotettavaa ennakkoarviota p:n arvosta, voidaan käyttää varovaista (riittävän suurta) otoskoon arviota

$$n = \frac{z^2_{\alpha/2}}{4\varepsilon^2},$$

sillä $p(1-p) \leq 1/4$ kaikilla $p \in [0,1]$.

12.5 Mediaanin jakaumasta vapaa luottamusväli

Olkoon $X_1, X_2, \ldots, X_n$ otos jatkuvasta jakaumasta, jota ei tarkemmin tunneta. Muodostetaan nyt jakauman mediaanille luottamusväli. Huomattakoon, että jakauman odotusarvo ei ole välttämättä olemassa. Kun luottamusväli muodostetaan jakaumaa koskevien varsin niukkojen oletusten varassa, menetelmää sanotaan jakaumasta vapaaksi. Luottamusvälin muodostamisessa käytetään järjestyssuureita.

Olkoon $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ järjestetty otos, jossa siis $X_{(1)} > X_{(2)} > \cdots > X_{(n)}$. Mediaanin m luottamusväliksi voidaan ajatella esimerkiksi havaintojen vaihteluväliä $(X_{(1)}, X_{(n)})$, missä $X_{(1)}$ on pienin ja $X_{(n)}$ suurin havaintoarvo. Välin luottamustaso on silloin todennäköisyys $P(X_{(1)} < m < X_{(n)})$, missä m on määritelmänsä mukaan jakauman 50%:n fraktiili eli $\pi_{0.5}$ eli $P(X < m) = 0.5$. Otoksen $X_1, X_2, \ldots, X_n$ avulla voidaan määritellä riippumattomat satunnaismuuttujat $I_1, I_2, \ldots, I_n$ siten, että $I_j = 1$, kun $X_j < m$ ja muutoin $I_j = 0$. Onnistumisten lukumäärä $L = I_1 + I_2 + \cdots + I_n$ on siis mediaania pienempien havaintojen lukumäärä ja $L \sim \text{Bin}(n, 1/2)$.

Olkoon esimerkiksi $n = 5$. Jos kaikki havainnot ovat mediaania pienempiä ($L = 5$) tai kaikki havainnot ovat mesiaania suurempia ($L = 0$), niin mediaani ei ole välillä $(X_{(1)}, X_{(5)})$. Muutoin mediaani on välillä $(X_{(1)}, X_{(5)})$. Näin siis

$$\begin{aligned} P(X_{(1)} < m < X_{(5)}) &= 1 - P(L = 0) - P(L = 5) \\ &= 1 - (1/2)^5 - (1/2)^5 = 15/16 \end{aligned}$$

ja $(x_{(1)}, x_{(5)})$ on 94%:n luottamusväli ($15/16 \approx 0.94$).

Yleisesti välin $(X_{(1)}, X_{(n)})$ luottamustaso on

$$\begin{aligned} P(X_{(1)} < m < X_{(n)})) &= 1 - P(L = 0) - P(L = n) \\ &= \sum_{k=1}^{n-1} \binom{n}{k} (1/2)^k (1/2)^{n-k} \\ &= 1 - (1/2)^n - (1/2)^n = 1 - (1/2)^{n-1}. \end{aligned}$$

Kasvattamalla otoskokoa saadaan todennäköisyys $P(X_{(1)} < m < X_{(n)})$ mielivaltaisen lähelle ykköstä. On kuitenkin huomattava, että myös välin $(x_{(1)}, x_{(n)})$ pituus kasvaa n:n kasvaessa. Lyhempiä välejä (tarkempia estimaatteja) saadaan käyttämällä jotain muuta järjestyssuureisiin perustuvaa väliä $(X_{(i)}, X_{(j)})$, missä $i < j$. Esimerkiksi voitaisiin kokeilla väliä $(X_{(2)}, X_{(n-1)})$ tai $(X_{(3)}, X_{(n-2)})$. Vastaavalla päättelyllä kuin edellä saadaan välin $(X_{(i)}, X_{(j)})$ luottamustaso

$$P(X_{(i)} < m < X_{(j)})) = \sum_{k=i}^{j-1} \binom{n}{k} (1/2)^k (1/2)^{n-k} = 1 - \alpha.$$

Edellä esitettyä menetelmää voidaan käyttää minkä tahansa jatkuvan jakauman prosenttipisteen π_p luottamusvälin määrittämiseen. Medianin tapauksessa käytetty onnistumistodennäköisyys vain korvataan onnistumistodennäköisyydellä $P(X < \pi_p) = p$.

12.6 Yhden selittäjän lineaarinen regressiomalli

12.6.1 Ehdollinen normaalimalli

Oletetaan, että satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat riippumattomat ja

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2), \qquad 1 \le i \le n. \tag{12.6.1}$$

Havaittu aineisto muodostuu n:stä arvoparista $(x_1, y_1), \ldots, (x_n, y_n)$. Ennustemuuttujan x arvot $x_1, \ldots, x_n$ ajatellaan tunnetuiksi vakioiksi. Regressiofunktio on siis muotoa

$$E(Y \mid x) = \alpha + \beta x$$

ja kaikilla satunnaismuuttujilla Y_i on sama varianssi σ^2. Malli (12.6.1) voidaan myös lausua muodossa

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad 1 \le i \le n,$$

missä $\varepsilon_1, \ldots, \varepsilon_n$ ovat riippumattomat ja $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $1 \le i \le n$.

Havaintojen $Y_1, \ldots, Y_n$ yhteisjakauman tiheysfunktio on

$$\begin{aligned} f(y_1, \ldots, y_n \mid \alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} f_1(y_i \mid \alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right]. \end{aligned}$$

Tästä nähdään, että uskottavuusfunktio on

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

ja logaritmoitu uskottavuusfunktio on

$$l(\alpha, \beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2. \tag{12.6.2}$$

Funktiosta (12.6.2) voidaan ratkaista parametrien α, β ja σ^2 suurimman uskottavuuden estimaattorit

$$\begin{aligned} \hat{\beta} &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i, \\ \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{x} \quad \text{ja} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2, \end{aligned}$$

missä $\bar{x}$ on koevakioiden $x_1, \ldots, x_n$ ja $\bar{Y}$ on havaintojen $Y_1, \ldots, Y_n$ keskiarvo sekä $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Estimaattorit $\hat{\alpha}$ ja $\hat{\beta}$ ovat harhattomia, mutta $\hat{\sigma}^2$ on σ^2:n harhainen estimaattori. Sen sijaan estimaattori

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

on σ^2 harhaton estimaattori. Jotta voidaan esittää näihin estimaattoreihin perustuvat estimointi ja testausmenettelyt, täytyy tuntea estimaattoreiden otantajakaumat. Otantajakaumia koskevat tulokset on esitetty seuraavassa lauseessa.

Lause 12.1 *Normaalimallissa (12.6.1) estimaattoreiden $\hat{\alpha}, \hat{\beta}$ ja $\hat{\sigma}^2$ otantajakaumat ovat*

$$\hat{\alpha} \sim \mathrm{N}(\alpha, \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n} x_i^2), \quad \hat{\beta} \sim \mathrm{N}(\beta, \frac{\sigma^2}{S_{xx}}), \quad \textit{missä} \quad \mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2\bar{x}}{S_{xx}}.$$

Lisäksi, $(\hat{\alpha}, \hat{\beta})$ *ja* S^2 *ovat riippumattomat ja*

$$\frac{(n-2)S^2}{\sigma^2} \sim \text{Khi2}(n-2).$$

Näiden jakaumatulosten avulla voidaan johtaa parametreille luottamusvälit tässä luvussa esitettyjen periaatteiden mukaisesti.

12.6.2 Yksinkertainen lineaarinen regressio

Lineaarisessa regressiossa oletetaan, että vastemuuttuja riippuu lineaarisesti selittäjistä. Malli on muotoa

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad 1 \leq i \leq n,$$

missä Y_i on havaittava satunnaismuuttuja ja ε_i on virhetermi, α ja β ovat tuntemattomia vakioita ja $x_1, \ldots, x_n$ ovat tuntemattomia koevakioita. Oletamme, että $E(\varepsilon_i) = 0$, joten

$$E(Y_i) = \alpha + \beta x_i. \tag{12.6.3}$$

Itse asiassa (12.6.3) on ehdollinen odotusarvo

$$E(Y_i \mid x_i) = \alpha + \beta x_i.$$

Oletetaan, että regressiofunktio on lineaarinen, mutta ei tehdä normaalisuusoletusta kuten edellisessä alaluvussa. Pienimmän neliösumman keinolla saadaan parametrien estimaateiksi

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{r s_x s_y}{s_x^2} = \frac{s_y}{s_x} \cdot r$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

missä s_{xy} on otoskovarianssi ja s_x^2 sekä s_y^2 ovat otosvariansseja. Voidaan osoittaa, että $\hat{\alpha}$ ja $\hat{\beta}$ ovat minimivarianssisia kaikkien α:n ja β:n lineaaristen harhattomien estimaattorien joukossa. Voidaan osoittaa, että tässäkin tapauksessa estimaattorit $\hat{\alpha}$ ja $\hat{\beta}$ noudattavat likimain normaalijakaumaa suurilla n:n arvoilla. Siksi normaalijakauman avulla voidaan johtaa likimääräiset luottamusvälit.