# Luku 4

# Suurimman uskottavuuden menetelmä ja momenttimenetelmä

## 4.1 Uskottavuusfunktio

Palautetaan ensin mieleen, että funktio

$$\mathbf{y} \mapsto f(\mathbf{y}; \theta)$$

eli lauseke $f(\mathbf{y}; \theta)$ ymmärrettynä argumentin $\mathbf{y}$ funktiona kiinteällä $\theta$ on satunnaisvektorin $\mathbf{Y}$ yhteistiheysfunktio tai yhteispistetodennäköisyysfunktio.

Kun aineisto $\mathbf{y}$ on havaittu, ja havaittua arvoa käytetään funktion $f(\mathbf{y}; \theta)$ ensimmäisenä argumenttina, niin tämä lauseke on enää argumentin $\theta$ funktio. Parametriavaruudella määriteltyä funktiota

$$\theta \mapsto f(\mathbf{y}; \theta)$$

kutsutaan *uskottavuusfunktioksi* (engl. *likelihood function*). Sitä merkitään

$$L(\theta) = f(\mathbf{y}; \theta).$$

Joskus tahdotaan kirjata näkyviin, että uskottavuusfunktio riippuu myös aineistosta $\mathbf{y}$, ja tällöin voidaan käyttää merkintää

$$L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta).$$

Haluttaessa voidaan sanoa tarkemmin, että kyseessä on havaintoa $\mathbf{y}$ vastaava parametrin $\theta$ uskottavuusfunktio.

Huomaa, että uskottavuusfunktion yhteydessä $\theta$ on vapaa muuttuja, eikä tarkoita parametrin todellista arvoa. Kuten aikaisemmin todettiin, tälläinen symbolien väärinkäyttö tarkoittamaan erilaisissa yhteyksissä aivan erilaisia asioita on tilastotieteen merkinnöille tyypillistä, eikä se huolellisesti käytettynä ja tulkittuna aiheuta sekaannusta.

Vaikka funktio

$$\mathbf{y} \mapsto f(\mathbf{y}; \theta)$$

on ypntf tai ytf kaikilla $\theta$, niin huomaa, että uskottavuusfunktio on funktio

$$\theta \mapsto f(\mathbf{y};\theta),$$

ja se ei ole yptnf eikä ytf.

**Esimerkki 4.1** (Uskottavuusfunktio binomikokeessa) Oletetaan pallot kulhossa -esimerkissä (jakso 2.2), että kulhossa on $N = 5$ palloa ja että nostot tehdään palauttaen ja että tulokset ovat $\mathbf{y} = (1, 0, 0, 0, 1, 0, 0)$. Tällöin valkoisten pallojen lukumäärä $n = 7$ nostossa (eli onnistumisten lukumäärä) on 2, ja uskottavuusfunktio on binomikokeen yptnf:n kaavan (2.6) mukaan

$$L(\theta) = \left(\frac{\theta}{5}\right)^2 \left(1 - \frac{\theta}{5}\right)^5, \qquad \theta = 0, 1, 2, 3, 4 \text{ tai } 5.$$

Tässä onnistumistodennäköisyys on $p = \theta/N$, joka on valkoisten pallojen suhteellinen osuus kulhossa.

Jos taas sama havainto $\mathbf{y} = (1, 0, 0, 0, 1, 0, 0)$ saadaan nasta purkissa -esimerkissä, niin tällöin onnistumistodennäköisyys on $\theta$, ja uskottavuusfunktio on

$$L(\theta) = \theta^2 (1 - \theta)^5.$$

Tässä parametriavaruudeksi ja uskottavuusfunktion määrittelyjoukoksi voidaan valita joko suljettu väli $[0, 1]$ tai avoin väli $(0, 1)$. △

Laajennamme uskottavuusfunktion määritelmää sillä tavalla, että uskottavuusfunktioksi kelpuutetaan myös mikä tahansa muotoa

$$L(\theta) = k(\mathbf{y})\, f(\mathbf{y};\theta) \tag{4.1}$$

oleva lauseke, jossa positiivinen vakio $k(\mathbf{y}) > 0$ saa riippua aineistosta $\mathbf{y}$, mutta ei saa riippua uskottavuusfunktion argumentista $\theta$. Edellä tehtiin aina valinta $k(\mathbf{y}) = 1$.

Tilastollisessa päättelyssä yleensä suositaan sellaisia menetelmiä, jotka perustuvat ainoastaan uskottavuusfunktion käyttöön ja joiden kannalta on saman tekevää, mitä verrannollisuuskerrointa $k = k(\mathbf{y}) > 0$ uskottavuusfunktion määritelmässä käytetään. Esimerkiksi uskottavuusfunktion maksimikohta pysyy samana vaikka kerrointa $k > 0$ muutetaan. Uskottavuusfunktion (tai siis minkä tahansa uskottavuusfunktion version) voidaan ajatella sisältävän kaiken aineistoon liittyvän informaation parametrin $\theta$ arvosta.

Usein uskottavuusfunktion sijasta tarkastellaan sen logaritmia.

**Määritelmä 4.1** (Logaritminen uskottavuusfunktio)**.** Uskottavuusfunktion logaritmia

$$\ell(\theta) = \log L(\theta)$$

kutsutaan logaritmiseksi uskottavuusfunktioksi tai uskottavuusfunktion logaritmiksi tai log-uskottavuusfunktioksi (engl. *log-likelihood*). Tässä log tarkoittaa luonnollista logaritmia.

Silloin, kun siirrytään uskottavuusfunktiosta $L(\theta)$ logaritmiseen uskottavuusfunktioon $\ell(\theta) = \log L(\theta)$ tehdään tavallisesti se oletus, että $L(\theta) > 0$ koko parametriavaruudessa, jolloin $\ell(\theta)$ on hyvin määritelty reaalifunktio: $\log(0)$ ei ole reaaliluku. Vaihtoehtoinen tapa selvitä tästä pulmasta on sopia, että

$\log(0) = -\infty$, joka on pienempi kuin mikään reaaliluku. Koska uskottavuusfunkio on määrätty vain positiivista verrannollisuuskerrointa $k > 0$ vaille, niin tämän seurauksena logartiminen uskottavuusfunktio on määrätty vain vakiota $\log k$ vaille; funktioon $\ell(\theta)$ voidaan lisätä mikä tahansa vakio, jos tämä yksinkertaistaa kaavoja.

Jos uskottavuusfunktio on tulomuotoa (2.2), niin logaritmin otto muuttaa sen summaksi, sillä

$$\log(\prod_{i=1}^{n} f_{Y_i}(y_i;\theta)) = \sum_{i=1}^{n} \log(f_{Y_i}(y_i;\theta)).$$

Tässä sovellettiin tuttua kaavaa

$$\log(ab) = \log(a) + \log(b), \qquad \text{kun } a > 0 \text{ ja } b > 0.$$

Tietokoneella laskettaessa logaritmointi on tärkeää, sillä uskottavuusfunktiossa esiintyvät tulon termit ovat usein erittäin pieniä lukuja, jolloin itse uskottavuusfunktion arvoksi saattaa tietokoneohjelmassa tulla tasan nolla, vaikka kyseessä olisi aidosti positiivinen luku. Logaritmin ottaminen uskottavuusfunktiosta riittää yleensä ratkaisemaan tämän ongelman.

## 4.2 Suurimman uskottavuuden estimaatti

Frekventistisessä tilastotieteessä parametria $\theta$ pidetään tuntemattomana vakiona, josta tiedetään vain, missä joukossa (eli parametriavaruudessa) sen arvot voivat olla. Parametria voidaan estimoida eli arvioida erilaisilla menetelmillä.

Tunnetuin estimointiperiaate on ns. *suurimman uskottavuuden*, eli SU-periaate (engl. *maximum likelihood, ML*), jonka mukaan parametrin parhaana estimaattina pidetään sitä parametriavaruuden arvoa $\hat{\theta}$, joka maksimoi uskottavuusfunktion. Sitä kutsutaan suurimman uskottavuuden estimaatiksi (eli SU-estimaatiksi) (engl. *maximum likelihood estimate, ML estimate, MLE*). Tämä ajatus voidaan esittää kaavalla

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta). \tag{4.2}$$

Merkintä $\arg\max L(\theta)$ tarkoittaa lausekkeen $L(\theta)$ maksimoivaa argumenttia (ts. maksimipistettä). Sen sijaan merkintä $\max L(\theta)$ tarkoittaisi lausekkeen $L(\theta)$ maksimiarvoa. Kun näitä merkintöjä käytetään, niin tällöin hiljaisesti oletetaan, että parametriavaruudessa on olemassa yksikäsitteinen maksimipiste $\hat{\theta}$, jolle

$$L(\hat{\theta}) \geq L(\theta), \quad \text{kaikille } \theta \in \Theta.$$

Koska logaritmi on aidosti kasvava funktio, on uskottavuusfunktiolla $L(\theta)$ ja logaritmisella uskottavuusfunktiolla $\ell(\theta)$ samat maksimipisteet. Tämän takia SU-estimaatti voidaan yhtä hyvin määrittää kaavalla

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta). \tag{4.3}$$

Mikäli aineistoa vastaavan satunnaisvektorin $\mathbf{Y}$ jakauma on diskreetti, niin SU-estimaatti on se parametrialueen piste, joka tekee havaitun aineiston (mallin puitteissa) mahdollisimman todennäköiseksi, eli

$$P_{\hat{\theta}}(\mathbf{Y} = \mathbf{y}) \geq P_{\theta}(\mathbf{Y} = \mathbf{y}), \qquad \text{kaikilla } \theta \in \Theta.$$

Tuntuu järkevältä suosia sellaisia parametrin arvioita, joille havainnot ovat todennäköisiä eikä sellaisia, joille ne ovat epätodennäköisiä. Koska parametriavaruudesta joudutaan yksi piste estimaatiksi valitsemaan, niin miksipä ei valittaisi sitä pistettä, joka tekee havainnot mahdollisimman todennäköisiksi.

Jatkuvan yhteisjakauman tapauksessa SU-menetelmän motivointi on samantapainen, mutta monimutkaisempi. Oletamme, että havaintosatunnaisvektorin yhteisjakauma on jatkuva ja että satunnaismuuttujat $Y_i$ ovat riippumattomia, kuten kaavassa (2.2). SU-estimaatti on se parametriarvo, joka maksimoi yhteistiheysfunktion arvon laskettuna aineistolle $\mathbf{y}$. Koska yhteisjakauma on jatkuva, niin yhteispistetodennäköisyys $P_\theta(\mathbf{Y} = \mathbf{y}) = 0$, joten tätä tarkastelemalla emme saa aikaan järkevää kriteeriä. Sen sijaan tarkastelemme todennäköisyyttä, että kukin satunnaismuuttuja $Y_i$ saa arvonsa sellaiselta lyhyeltä väliltä $[a_i, b_i]$, joka sisältää havainnon $y_i$.

$$\begin{aligned} P_\theta(a_1 \le & Y_1 \le b_1, a_2 \le Y_2 \le b_2, \ldots, a_n \le Y_n \le b_n) \\ &= \int_{a_1}^{b_1} f_{Y_1}(t_1;\theta)\,\mathrm{d}t_1 \int_{a_2}^{b_2} f_{Y_2}(t_2;\theta)\,\mathrm{d}t_2 \ \ldots \int_{a_n}^{b_n} f_{Y_n}(t_n;\theta)\,\mathrm{d}t_n \\ &\approx (b_1 - a_1) f_{Y_1}(y_1;\theta)\ (b_2 - a_2) f_{Y_2}(y_2;\theta)\ \ldots\ (b_n - a_n) f_{Y_n}(y_n;\theta) \\ &= f(\mathbf{y};\theta) \prod_{i=1}^{n} (b_i - a_i) \end{aligned}$$

Tässä ensin vedottiin satunnaismuuttujien $Y_i$ riippumattomuteen, ja sen jälkeen kussakin integraalissa tehtiin seuraava approksimaatio. Lyhen välin $[a_i, b_i]$ yli laskettu integraali funktiosta $f_{Y_i}(t_i;\theta)$ on (integraalilaskennan väliarvolauseeseen nojaten) osapuilleen sama kuin sen suorakaiteen pinta-ala, jonka kanta on kyseisen välin pituus ja korkeus $f_{Y_i}(y_i;\theta)$. (Tiheysfunktiot $y_i \mapsto f_{Y_i}(y_i;\theta)$ oletetaan jatkuviksi.) Todennäköisyydeksi saatiin osapuilleen välien pituuksien tulo kertaa yhteistiheysfunktion arvo $f(\mathbf{y};\theta)$. Tämän tarkastelun jälkeen näemme, että SU-menetelmän motivaatio on myös jatkuvan yhteisjakauman tapauksessa se, että yritämme valita sellaisen parametriavaruuden pisteen, joka tekee havainnot mahdollisimman todennäköiseksi.

**Esimerkki 4.2** (Jatkoa esimerkille 4.1, pallot kulhossa) Valkoisten pallojen lukumäärä $\theta$ on yksi luvuista $0, 1, 2, 3, 4$ tai $5$, ja uskottavuusfunktio on

$$\begin{aligned} L(\theta) &= \left(\frac{\theta}{5}\right)^2 \left(1 - \frac{\theta}{5}\right)^5 \\ &= \begin{cases} 0, & \text{jos } \theta = 0, \\ 1024/5^7, & \text{jos } \theta = 1, \\ 972/5^7, & \text{jos } \theta = 2, \\ 288/5^7, & \text{jos } \theta = 3, \\ 16/5^7, & \text{jos } \theta = 4, \\ 0, & \text{jos } \theta = 5. \end{cases} \end{aligned}$$

Havaintojen takia voimme sulkea pois arvot $\theta = 0$ ja $\theta = 5$, koska kulhosta ei voitaisi nostaa valkoisia (mustia) palloja, jos niitä ei siellä alunperin lainkaan olisi. Voimme myös sanoa, että arvo $\theta = 3$ on uskottavampi kuin arvo $\theta = 4$,

koska $L(3) > L(4)$. Todennäköisyys poimia valkoinen pallo kaksi kertaa seitsemässä nostossa on suurempi, mikäli $\theta = 3$ kuin siinä tapauksessa, että $\theta = 4$. Kaikista uskottavin arvo eli SU-estimaatti on $\hat{\theta} = 1$. △

*Varoitus.* SU-estimaatti $\hat{\theta}$ on edellisessä esimerkissä se arvo, joka tekee havainnot (mallin puitteissa) mahdollisimman todennäköisiksi. Sen sijaan olisi vakava väärinkäsitys väittää, että $\hat{\theta}$ eli uskottavin parametrin olisi parametrin todennäköisin arvo. Frekventistisen tilastotieteen puitteissa tällainen lausuma on mieltä vailla, koska parametrin arvoa koskevia todennäköisyyksiä ei frekventistisessä mallissa ole määriteltynä. Juuri tästä syystä Fisher otti käyttöön termin *uskottavuus*.

Pallot kulhossa -esimerkissä parametriavaruus on diskreetti. Koska se ei esimerkissä koostu kovin monesta pisteestä, pystymme laskemaan uskottavuusfunktion arvon jokaisessa parametriavaruuden pisteessä. Tämän jälkeen valitsemme sen pisteen, jossa suurin arvo saavutetaan.

Jos parametriavaruus on jatkuva, niin tälläinen menettely ei tule kyseeseen, vaan maksimoinnissa käytetään hyväksi derivaattaa. Tarkastelmme ensin yhden parametrin $\theta$ tapausta. Yksinkertaisissa tilanteissa SU-estimaatti saadaan ratkaistua algebrallisesti etsimällä logaritmisen uskottavuusfunktion derivaatan nollakohdat, eli ratkaisemalla ns. uskottavuusyhtälö

$$\ell'(\theta) = 0.$$

Tämä perustuu siihen, että mikäli (jatkuvasti derivoituva) yhden muuttujan funktio saavuttaa maksimin jossakin määrittelyjoukkonsa sisäpisteessä, niin kyseisessä pisteessä funktion derivaatta saa arvon nolla. Tämän jälkeen pitää funktion kriittisistä pisteistä eli derivaatan nollakohdista valita ne, jotka ovat maksimipisteitä. Tämä onnistuu joko tarkastelemalla derivaatan merkkikaaviota tai $\ell$:n toista derivaattaa (jos $\ell'(\theta_0) = 0$ ja $\ell''(\theta_0) < 0$, niin $\theta_0$ on maksimipiste). Lisäksi pitää kiinnittä huomiota (log-)uskottavuusfunktion käyttäytmiseen, kun lähestytään parametriavaruuden reunapisteitä. Tällä tavalla löydetään kaikki paikalliset maksimipisteet, ja lopulta niistä valitaan globaali maksimi, eli se piste, jossa $\ell$ saavuttaa suurimman arvonsa koko parametriavaruudessa. Näemme tästä menettelystä esimerkkejä seuraavissa jaksossa.

Jos parametreja on useita, niin kaikki $\ell$:n ensimmäisen kertaluvun osittaisderivaatat häviävät maksimipisteessä, joten tällöin uskottavuusyhtälö on yhtälöryhmä. Esim. kahden parametrin $\theta = (\mu, \phi)$ tapauksessa pitäisi etsiä ne pisteet, joissa molemmat yhtälöt

$$\frac{\partial}{\partial \mu}\ell(\mu, \phi) = 0, \qquad \frac{\partial}{\partial \phi}\ell(\mu, \phi) = 0$$

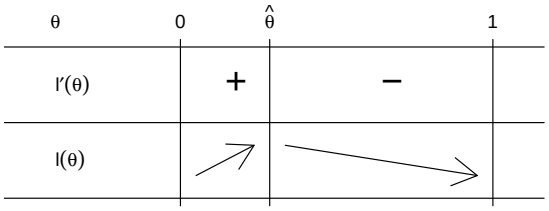
toteutuvat. Kriittisen pisteen laadun (minimi, maksimi, satulapiste) voi tarkistaa toisen kertaluvun osittaisderivaattojen avulla.

Monimutkaisemmissa tapauksissa maksimipisteitä ei enää pystytä määrittämään algebrallisesti, vaan ne haetaan tietokoneen avulla soveltamalla jotakin numeerista maksimointimenetelmää.

## 4.3 SU-estimaatti binomikokeessa

Binomikokeessa uskottavuusfunktion tai sen logaritmin arvo voidaan laskea kaavan (2.6) heti, kun tiedetään onnistumisten lukumäärä, toistojen lukumäärä

**Kuva 4.1** Logaritmisen uskottavuusfunktion kulkukaavio binomikokeessa, kun $n = 7$ toistossa havaitaan $x = 2$ onnistumista.




sekä parametriavaruus. Tätä varten ei tarvitse tietää, missä järjestyksessa onnistumiset ja epäonnistumiset sattuivat aineistossa $(y_1, \ldots, y_n)$. Johdamme seuraavaksi SU-estimaatin kaavan, kun $n$ toistossa onnistutaan $x$ kertaa, ja onnistumistodennäköisyys yhdessä toistossa on $\theta$. Oletamme, että parametriavaruus $\Theta$ on joko avoin väli $(0, 1)$ tai suljettu väli $[0, 1]$. Tällainen tilanne oli nasta purkissa -esimerkissä (mutta pallot pallot kulhossa -esimerkissä parametriavaruus oli diskreetti).

Käsittelemme ensin sen tapauksen, jossa onnistumisten lukumäärä $x$ on välillä $1 \leq x \leq n - 1$. Logaritminen uskottavuusfunktio on

$$\ell(\theta) = \log\left(\theta^x \, (1-\theta)^{n-x}\right) = x \, \log\theta + (n-x) \, \log(1-\theta),$$

joka on hyvin määritelty, kun $0 < \theta < 1$.

Ratkaisemme seuraavaksi logaritmisen uskottavuusfunktion derivaatan nollakohdat. Kun $0 < \theta < 1$, niin

$$\ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x - n\theta}{\theta\,(1-\theta)}$$

Derivaatan ainoa nollakohta on $\hat{\theta} = x/n$, ja kyseessä on maksimipiste, sillä derivaatan merkki vaihtuu siinä positiivisesta negatiiviseksi. (Nimittäjä $\theta\,(1-\theta)$ on positiivinen.) Tämän takia kyseessä on maksimipiste. Derivaatan merkistä nähdään myös, että $\ell(\theta)$ kasvaa välillä $(0, \hat{\theta})$ ja vähenee välillä $(\hat{\theta}, 1)$, joten $\hat{\theta}$ on globaali maksimi. Kuvassa 4.1 esitetään funktion $\ell$ kulkukaavio siinä tilanteessa, kun $n = 7$ toistossa havaitaan $x = 2$ onnistumista.

Tapauksessa $x = n$ uskottavuusfunktio on

$$L(\theta) = \theta^n,$$

ja tämä on selvästi aidosti kasvava funktio välillä $(0, 1)$. Jos parametriavaruus on $[0, 1]$, niin SU-estimaatti on $\hat{\theta} = 1 = x/n$. Huomaa, että SU-estimaatti ei tässä tapauksessa löydy derivaatan nollakohdasta, vaan parametriavaruuden reunalta. Jos parametriavaruus kuitenkin on avoin väli $(0, 1)$, niin tällöin joudumme

toteamaan, että SU-estimaattia ei ole olemassa, koska uskottavuusfunktio ei saavuta missään parametriavaruuden pisteessä maksimiarvoaan.

Tapauksessa $x = 0$ nädään vastaavasti, että SU-estimaatti on $\hat{\theta} = 0 = x/n$, mikäli parametriavaruus on $[0, 1]$. Jos parametriavaruus kuitenkin on $(0, 1)$, niin SU-estimaattia ei ole olemassa.

Mikäli binomikokeessa tahdotaan käyttää SU-estimointia, niin tästä syystä on kätevää valita parametriavaruudeksi suljettu väli $[0, 1]$. Tällöin SU-estimaatti saadaan kaikissa tapauksissa kaavalla

$$\hat{\theta} = \frac{x}{n} \tag{4.4}$$

eli SU-estimaatti on onnistumisten $x$ suhteellinen osuus $n$ toistossa.

Olemme jo edellä jaksossa 3.4 nähneet, että vastaava estimaattori on harhaton ja että sen (otantajakauman) varianssi saadaan kaavalla

$$\frac{1}{n}\,\theta\,(1-\theta).$$

Tämän ansiosta SU-estimaatin $\hat{\theta}$ keskivirhe voidaan laskea kaavalla

$$\sqrt{\frac{1}{n}\,\hat{\theta}\,(1-\hat{\theta})}, \tag{4.5}$$

jossa tuntematon parametrinarvo $\theta$ on korvattu sen estimaatilla $\hat{\theta}$.

Kuvassa 4.2 esitetään binomikokeen uskottavuusfunktio ja logaritminen uskottavuusfunktio kahdella erilaisella otoskoolla. Näissä kuvissa tilanne on valittu siten, että $\hat{\theta} = x/n = 0.2$ on molemmilla otoskoilla. Huomaa, että pienellä otoskoolla uskottavuusfunktio on selvästi laakeampi kuin suurella otoskoolla. Suurella otoskoolla uskottavat parametrinarvot ovat melko kapealla välillä SU-estimaatin ympärillä, joten intuitio sanoo, että suurella otoskoolla parametrin arvosta voi tehdä tarkempia päätelmiä kuin pienellä. Tämän asian näkee myös laskemalla estimaattien keskivirheet kaavalla (4.5), jolloin otoskoolla $n = 5$ saadaan keskivirhe

$$\sqrt{\frac{1}{5} \times 0.2 \times 0.8} = 0.18$$

ja otoskoolla $n = 80$ keskivirhe

$$\sqrt{\frac{1}{80} \times 0.2 \times 0.8} = 0.045.$$

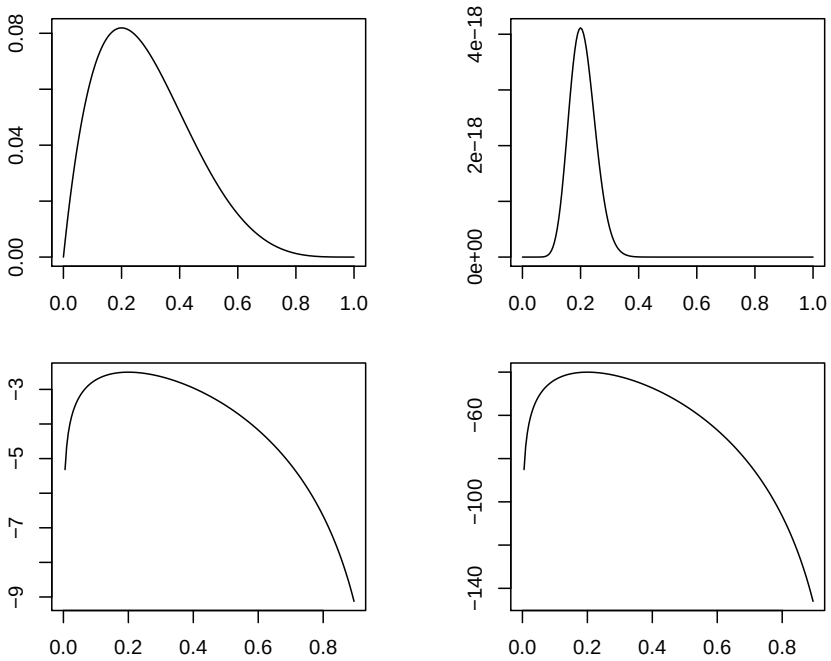
## 4.4 Normaalijakauman parametrien estimointi

Tarkastelemme tilannetta, jossa mallinnamme aineiston $\mathbf{y} = (y_1, \ldots, y_n)$ siten, että vastaavat satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$. Ts. oletamme, että satunnaismuuttujat $Y_i$ ovat riippumattomia, ja kukin niistä noudattaa normaalijakaumaa $N(\mu, \sigma^2)$. Tässä $\mu \in \mathbb{R}$ ja $\sigma^2 > 0$ voivat molemmat olla tuntemattomia parametreja, tai sitten toinen niistä voi olla tunnettu vakio ja toinen tuntematon parametri.

Kunkin yksittäisen satunnaismuuttujan $Y_i$ tiheysfunktio on

$$g(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\,\exp\left(-\frac{1}{2\,\sigma^2}(y-\mu)^2\right).$$

**Kuva 4.2** Uskottavuusfunktio ja logaritminen uskottavuusfunktio binomikokeessa kahdella eri otoskoolla, kun parametriavaruus on jatkuva. Vasemmalla $n = 5$ ja oikealla $n = 80$; ylempänä on uskottavuusfunktio ja alempana sen logaritmi. Molemmissa tapauksissa onnistumisten suhteellinen osuus $k/n = 0.2$. Suuremmalla otoskoolla uskottavuusfunktio ja sen logaritmi ovat selvästi terävämpihuippuisia funktioita kuin pienellä; logaritmisten uskottavuusfunktioden kohdalla $y$-akselien skaalat ovat tyystin erilaiset.



Tässä exp tarkoittaa eksponenttifunktiota, eli

$$\exp(x) = \mathrm{e}^x, \qquad \text{kun } x \in \mathbb{R}.$$

Parametrien $\mu$ ja $\sigma^2$ merkitys on se, että kullakin $i$

$$EY_i = \mu, \qquad \operatorname{var} Y_i = \sigma^2.$$

Parametri $\mu$ on paitsi normaalijakauman $N(\mu, \sigma^2)$ odotusarvo, myös sen moodi ja mediaani. Normaalijakauman tiheysfunktio on symmetrinen odotusarvon suhteen. Varianssiparametri kuvaa sitä, miten tiukasti jakauma on keskittynyt keskikohtansa ympärille: mitä pienempi varianssi, sitä keskittyneempi jakauma.

Havaintosatunnaisvektorin $\mathbf{Y}$ yhteistiheysfunktio on

$$\begin{aligned} f(\mathbf{y}; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left(-\frac{1}{2\,\sigma^2}(y_i - \mu)^2\right). \\ &= (2\,\pi)^{-n/2}\,(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \end{aligned} \tag{4.6}$$

Johdossa sovellettiin tuttua kaavaa

$$\mathrm{e}^a\,\mathrm{e}^b = \mathrm{e}^{a+b}, \quad \text{eli} \quad \exp(a)\,\exp(b) = \exp(a+b),$$

joka pätee kaikille reaaliluvuille $a$ ja $b$.

Jätetään uskottavuusfunktiosta $2\pi$:n potenssit pois, jolloin kaavasta (4.6) saadaan havaintoa $\mathbf{y}$ vastaavalle logaritmiselle uskottavuusfunktiolle lauseke

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{4.7}$$

Ylläolevassa kaavassa voidaan neliöiden summa hajottaa kahteen osaan (harjoitustehtävä)

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n\,(\bar{y} - \mu)^2, \tag{4.8}$$

jossa $\bar{y}$ on lukujen $y_i$ otoskeskiarvo,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.9}$$

Tämä huomio helpottaa SU-estimaattien löytämistä.

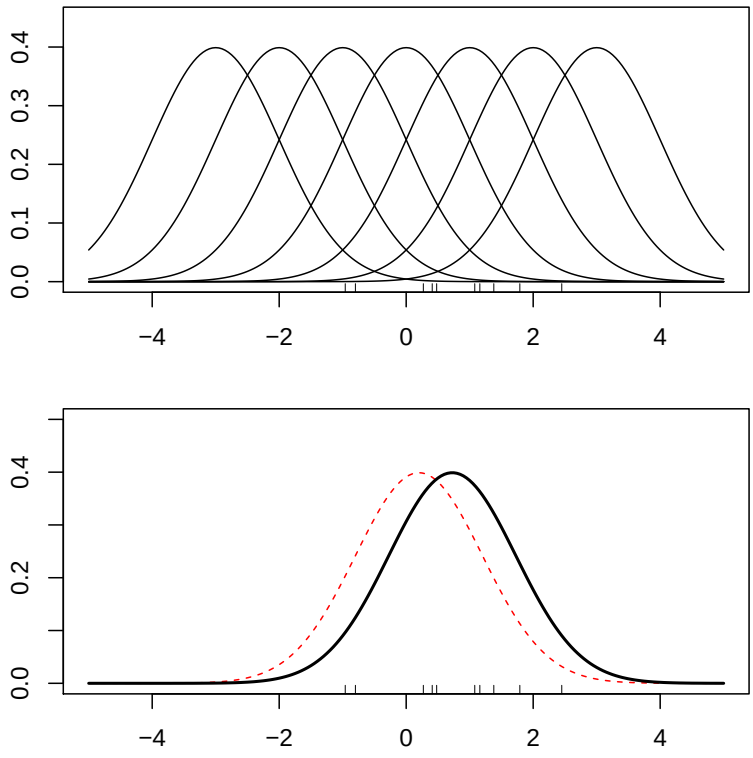
### 4.4.1 Varianssi tunnettu

Jos normaalijakaumaperheessä varianssi $\sigma^2$ on tunnettu luku, niin mallissa on jäljellä vain yksi tuntematon parametri $\mu$. Kuvassa 4.3 näytetään muutama $N(\mu, \sigma^2)$-jakaumaperheen tiheysfunktio. Yksi tällainen tiheysfunktio pitää nyt valita kuvaamaan $x$-akselille lyhyillä viivoilla merkittyä aineistoa.

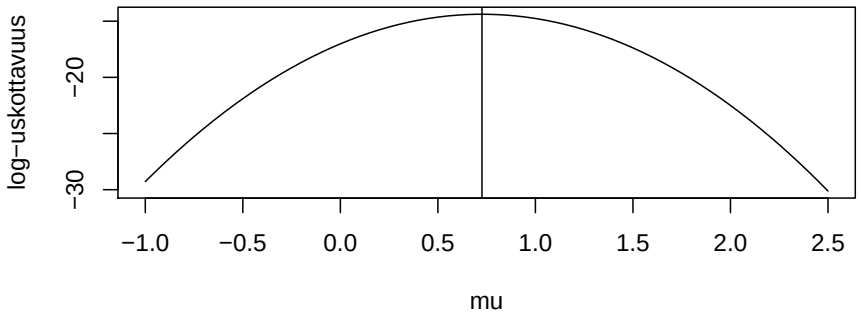
Logaritminen uskottavuusfunktio on kaavojen (4.7) ja (4.8) mukaan

$$\begin{aligned} \ell(\mu) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\,\sigma^2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 + n\,(\bar{y} - \mu)^2\right) \\ &= \text{vakio} - \frac{n}{2\,\sigma^2}(\bar{y} - \mu)^2 \end{aligned}$$

**Kuva 4.3** Parametrin $\mu$ estimointi normaalijakaumaperheelle $N(\mu, \sigma^2)$, kun $\sigma^2$ on tunnettu luku (tässä $\sigma^2 = 1$). Ylemmässä kuvassa esitetään normaalijakaumaperheen $N(\mu, 1)$ tiheysfunktioita muutamilla eri parametrin $\mu$ arvoilla sekä eräästä normaalijakaumasta $N(\mu, 1)$ simuloitu aineisto (lyhet viivat $x$-akselin yläpuolella). Alemmassa kuvassa on paljastettu todellinen simuloinnissa käytetty tiheysfunktio (katkoviiva) sekä SU-estimaattia vastaava estimoitu tiheysfunktio (yhtenäinen viiva). Todellisessa tilastollisen päättelyn tilanteessa katkoviivalla merkittyä todellista tiheysfunktiota ei tunnettaisi.



**Kuva 4.4** Parametrin $\mu$ logaritminen uskottavuusfunktio. SU-estimaatti on merkitty pystyviivalla.




Tässä vakioksi merkitty termi ei riipu $\mu$:sta. Koska kerroin $n/(2\,\sigma^2)$ on positiivinen, niin logaritminen uskottavuusfunktio maksimoituu täsmälleen silloin, kun lauseke $(\bar{y} - \mu)^2$ minimoituu, eli silloin, kun $\mu = \bar{y}$. Logaritminen uskottavuusfunktio on esitetty kuvassa 4.4 kuvan 4.3 aineistolle.

Tässä tapauksessa SU-estimaatti on *otoskeskiarvo* (engl. *sample mean; average*), eli

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{4.10}$$

Vastaava estimaattori $\frac{1}{n} \sum_{i=1}^{n} Y_i$ on harhaton, ja sen varianssi on

$$\mathrm{var}_\mu \left( \frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \frac{1}{n} \sigma^2,$$

joka on tässä mallissa tunnettu vakio. Tämän luvun neliöjuuri on SU-estimaatin keskivirhe.

### 4.4.2 Molemmat parametrit tuntemattomia

Nyt molemmat parametrit $\mu$ ja $\sigma^2$ ovat tuntemattomia, joten satunnaisvektorin $\mathbf{Y}$ jakauman kiinnittämiseksi pitäisi tuntea parametrivektorin $\boldsymbol{\theta} = (\mu, \sigma^2)$ arvo. Kuvassa 4.5 näytetään muutama $N(\mu, \sigma^2)$-jakaumaperheen tiheysfunktio. Yksi tällainen tiheysfunktio pitää jälleen valita kuvaamaan $x$-akselille lyhyillä viivoilla merkittyä aineistoa.

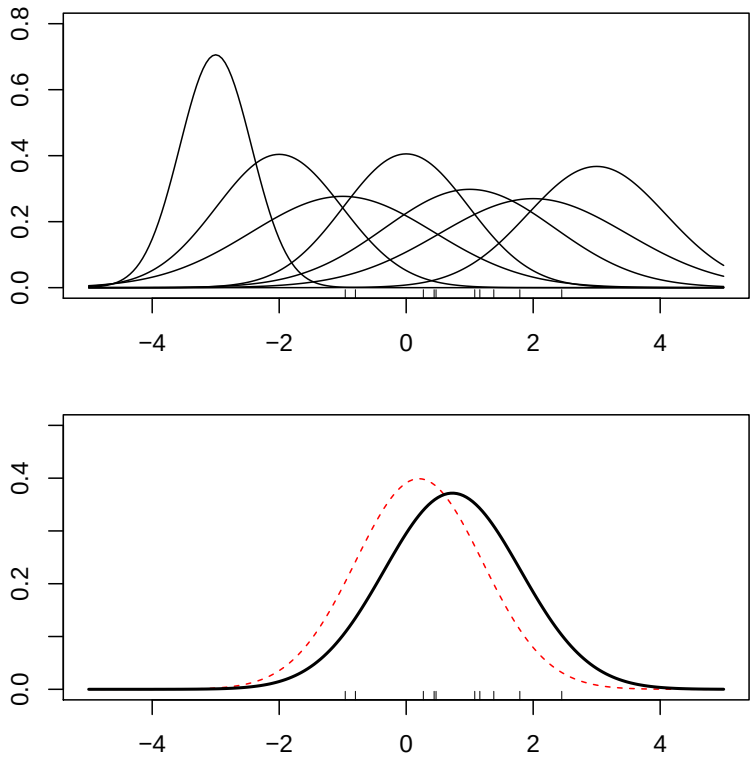
Logaritminen uskottavuusfunktio on kaavojen (4.7) ja (4.8) mukaan

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{n}{2\,\sigma^2} (\bar{y} - \mu)^2$$

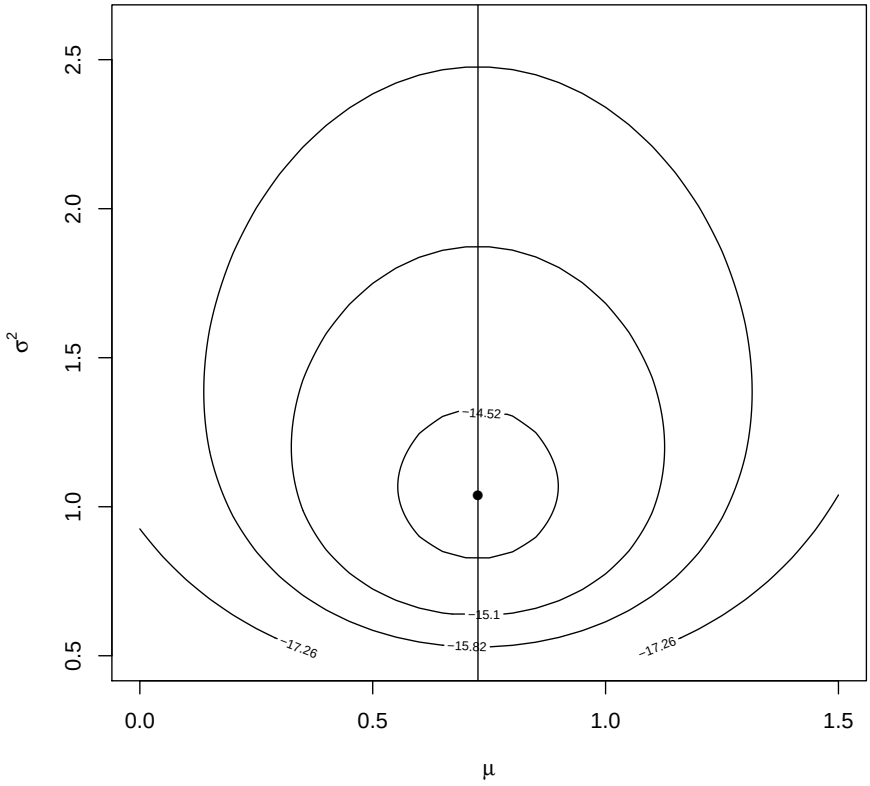
Logaritminen uskottavuusfunktio on esitetty kuvassa 4.6 kuvan 4.3 aineistolle.

Logaritminen uskottavuusfunktio riippuu $\mu$:n arvosta vain sen viimeisen termin kautta. Oli varianssiparametrin $\sigma^2 > 0$ arvo mikä tahansa, niin funktion

**Kuva 4.5** Parametrin $(\mu, \sigma^2)$ estimointi normaalijakaumaperheelle $N(\mu, \sigma^2)$, kun sekä $\mu$ että $\sigma^2$ ovat tuntemattomia. Ylemmässä kuvassa esitetään normaalijakaumaperheen $N(\mu, \sigma^2)$ tiheysfunktioita muutamilla eri parametrivektorin $(\mu, \sigma^2)$ arvoilla sekä eräästä normaalijakaumasta simuloitu aineisto (lyhet viivat $x$-akselin yläpuolella). Alemmassa kuvassa on paljastettu todellinen simuloinnissa käytetty tiheysfunktio (katkoviiva) sekä SU-estimaattia vastaava estimoitu tiheysfunktio (yhtenäinen viiva). Todellisessa tilastollisen päättelyn tilanteessa katkoviivalla merkittyä todellista tiheysfunktiota ei tunnettaisi.



**Kuva 4.6** Parametrivektorin $(\mu, \sigma^2)$ logaritminen uskottavuusfunktio $\ell(\mu, \sigma^2)$ esitettynä tasa-arvokäyriensä avulla. SU-piste on merkitty pallolla. Millä tahansa varianssiparametrin arvolla funktion $\mu \mapsto \ell(\mu, \sigma^2)$ maksimi löytyy pisteestä $\mu = \bar{y}$, joka on osoitettu suoralla.



$\mu \mapsto \ell(\mu, \sigma^2)$ maksimoi arvo $\hat{\mu} = \bar{y}$. Tämän ansiosta maksimointi saadaan palautettua yhdestä muuttujasta riippuvan funktion $u$ maksimointitehtäväksi, jossa

$$\begin{aligned} u(\sigma^2) &= \max_{\mu} \ell(\mu, \sigma^2) = \ell(\bar{y}, \sigma^2) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{aligned}$$

Tämän funktion maksimi puolestaan löytyy pisteestä

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Näiden tarkastelujen jälkeen ollaan saatu selville, että parametrin $(\mu, \sigma^2)$ SU-estimaatti on

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{4.11}$$

Estimaattori

$$\hat{\mu}(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

on harhaton, mutta varianssiparametrin SU-estimaattori

$$\hat{\sigma}^2(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

on harhainen, sillä sen odotusarvo on (harjoitustehtävä)

$$E_{(\mu,\sigma^2)}[\hat{\sigma}^2(\mathbf{Y})] = \frac{n-1}{n}\sigma^2.$$

Koska harhan saa helposti korjattua, niin varianssin estimaattina käytetään tavallisesti SU-estimaatin sijasta *otosvarianssia* (engl. *sample variance*)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{4.12}$$

Sitä vastaava estimaattori

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{4.13}$$

on harhaton (varianssiparametrille $\sigma^2$), sillä

$$E_{(\mu,\sigma^2)}[S^2] = E_{(\mu,\sigma^2)}[\frac{n}{n-1}\hat{\sigma}^2(\mathbf{Y})] = \frac{n}{n-1}\,\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Estimaattorien $(\bar{Y}, S^2)$ yhteisotantajakauma tunnetaan. Esim. aineopintojen todennäköisyyslaskennan kurssilla todistetaan, että kun (mallin oletusten mukaan) $Y_1, \ldots, Y_n$ ovat riippumattomia normaalijakaumaa $N(\mu, \sigma^2)$ noudattavia

satunnaismuuttujia, niin tällöin

$$\bar{Y} \text{ ja } S^2 \text{ ovat riippumattomia,} \tag{4.14}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \frac{1}{n}\,\sigma^2), \tag{4.15}$$

$$\frac{n-1}{\sigma^2}S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2_{n-1}. \tag{4.16}$$

Tässä $\chi^2_{n-1}$ tarkoittaa khiin neliön jakaumaa vapausasteluvulla $n-1$, joka on eräs kuuluisa positiivisella reaaliakselilla määritelty jatkuva jakauma. Sovellamme näitä tietoja myöhemmin.

Keskiarvoa $\bar{Y}$ koskeva jakaumatulos (4.15) on helppo johtaa. Normaalijakauman yhteenlaskuominaisuuden mukaan riippumattomien satunnaismuuttujien $Y_1$ ja $Y_2$ summalla on normaalijakauma, jonka parametrit saadaan laskemalla yhteen $Y_1$:n ja $Y_2$:n jakaumien parametrit, eli

$$Y_1 + Y_2 \sim N(\mu + \mu, \sigma^2 + \sigma^2).$$

(Varoitus: tämä on nimenomaan normaalijakaumaa, riippumattomia satunnaismuuttujia ja yhteenlaskua koskeva ominaisuus. Vastaavat kaavat eivät automaattisesti pidä paikkaansa muille jakaumille, riippuville satunnaismuuttujille, tai muille laskutoimituksille.) Tätä päättelyä voidaan jatkaa, jolloin summan jakaumaksi saadaan

$$Y_1 + \cdots + Y_n \sim N(n\mu, n\sigma^2).$$

Kun nyt muistetaan, että tässä ensimmäinen parametri on odotusarvo ja toinen varianssi, niin nähdään helposti, että luvulla $1/n$ skaalatun summan jakauma on

$$\bar{Y} \sim N(\mu, \frac{1}{n}\,\sigma^2).$$

Sen sijaan satunnaismuuttujan $S^2$ jakauman johtaminen on paljon monimutkaisempaa, ja se väite, että $\bar{Y}$ ja $S^2$ ovat riippumattomia voi ensinäkemältä herättää hämmennystä, sillä satunnaismuuttuja $S^2$ määritellään satunnaismuuttujan $\bar{Y}$ avulla.

Usein normaalijakaumamallissa ollaan tosiasiassa kiinnostuneita lähinnä populaation odotusarvosta $\mu$, ja populaation varianssi $\sigma^2$ on ns. *haittaparametri* (engl. *nuisance parameter*), joka tarvitaan mallin spesifioimiseksi, mutta jonka arvosta ei olla kiinnostuneita. Tässä tapauksessa parametrin $\mu$ estimaatti on otoskeskiarvo $\bar{y}$. Vastaavan estimaattorin $\bar{Y}$ (otantajakauman) varianssi on $\sigma^2/n$. Kun tähän kaavaan sijoitetaan tuntemattoman populaatiovarianssin $\sigma^2$ tilalle sen otosestimaatti $s^2$, päädytään siihen, että keskiarvon keskivirhe lasketaan kaavalla

$$\frac{1}{\sqrt{n}}\,s,$$

jossa *otoskeskihajonta* (engl. *sample standard deviation*) $s$ on otosvarianssin (4.12) neliöjuuri, eli

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{4.17}$$

Otoskeskihajonta estimoi populaation keskihajontaa. Sen sijaan *keskiarvon keskivirhe* (engl. *standard error of the mean*)

$$\frac{1}{\sqrt{n}}\, s = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4.18}$$

estimoi satunnaismuuttujan $\bar{Y}$ keskihajontaa $\sigma/\sqrt{n}$.

Jos odotusarvo on tuntematon, niin myös muulloin kuin normaalijakautuneen populaation tapauksessa populaation varianssia usein estimoidaan otosvarianssilla $s^2$ (4.12), jota vastaava estimaattori $S^2$ (4.13) on populaation varianssin harhaton estimaattori aina, kun käsitellään satunnaisotosta populaatiosta, jonka varianssi on $\sigma^2$. Populaation keskihajontaa $\sigma = \sqrt{\sigma^2}$ on myös tapana estimoida otoskeskihajonnalla, vaikka vastaava estimaattori $S = \sqrt{S^2}$ ei ole harhaton.

## 4.5 Momenttimenetelmä

Momenttimenetelmä (engl. *method of moments*) on SU-menetelmää varhaisempi menetelmä estimaattorin määrittelemiseksi. Tarkastelemme tätä menetelmää siinä tapauksessa, jossa käsitellään satunnaisotosta $Y_1, \ldots, Y_n$ jakaumasta, jonka ptnf/tf on $g(y; \boldsymbol{\theta})$. Otamme käyttöön vielä satunnaismuuttujan $Y$ jolla myöskin on on ptnf/tf $g(y; \boldsymbol{\theta})$.

Jakauman $k$:s momentti ($k = 1, 2, \ldots$) määritellään kaavalla

$$\mu_k(\boldsymbol{\theta}) = EY^k = \begin{cases} \sum_y y^k\, g(y; \boldsymbol{\theta}) & \text{jos jakauma on diskreetti,} \\ \int y^k\, g(y; \boldsymbol{\theta})\, \mathrm{d}y & \text{jos jakauma on jatkuva.} \end{cases} \tag{4.19}$$

Tutuille jakaumille momenttien kaavat tunnetaan. Momenttia $\mu_k(\theta)$ voidaan estimoida $k$:nnella otosmomentilla

$$m_k(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} Y_i^k, \tag{4.20}$$

joka on populaatiomomentin $\mu_k(\boldsymbol{\theta})$ harhaton estimaattori.

Momenttimenetelmässä estimaatti (tai estimaattori) muodostetaan ratkaisemalla yhtälöryhmästä

$$\begin{cases} \mu_1(\boldsymbol{\theta}) &= m_1 \\ \mu_2(\boldsymbol{\theta}) &= m_2 \\ &\vdots \\ \mu_r(\boldsymbol{\theta}) &= m_r \end{cases} \tag{4.21}$$

tuntematon suure $\boldsymbol{\theta}$, jossa otosmomentit $m_1, \ldots, m_r$ lasketaan aineistosta. Ehtoja asetetaan niin monta, että yhtälöryhmällä on yksikäsitteinen ratkaisu parametriavaruudessa. Tavallisesti yhtälöitä asetetaan niin monta, kuin parametrivektorissa on komponentteja.

Tällä tavalla saadaan aikaan näppäriä kaavoja estimaateille joissakin sellaisissa tilanteissa, joissa SU-estimaatit jouduttaisiin määrittämään numeerisesti.

**Esimerkki 4.3** (Eksponenttijakauman parametrin estimointi momenttimenetelmällä) Eksponenttijakaumaa noudattava satunnaismuuttuja $Y$ voi saada kaikkia positiivisia reaaliarvoja, ja sillä on tiheysfunktio

$$g(y;\lambda) = \lambda \, \mathrm{e}^{-\lambda \, y}, \qquad y > 0. \tag{4.22}$$

jossa jakauman parametria on merkitty kirjaimella $\lambda > 0$. Jakaumasta käytetään lyhennettä $\mathrm{Exp}(\lambda)$. Jos $Y \sim \mathrm{Exp}(\lambda)$, niin sen odotusarvo on tunnetusti

$$EY = \frac{1}{\lambda}.$$

Tarkastelemme satunnaisotosta $Y_1, \ldots, Y_n$ eksponenttijakaumasta $\mathrm{Exp}(\lambda)$, ja havaitut arvot ovat $y_1, \ldots, y_n$. Koska parametreja on vain yksi, momenttimenetelmässä tarvitaan vain yksi yhtälö

$$EY = \frac{1}{\lambda} = m_1 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}.$$

Momenttimenetelmän mukainen parametrin $\lambda$ estimaatti on

$$\hat{\lambda} = \frac{1}{\bar{y}}.$$

Vaihtoehtoisesti eksponenttijakauma $\mathrm{Exp}(\lambda)$ voitaisiin parametroida sen odotusarvolla $\theta = 1/\lambda$. Momenttimenetelmä antaa tälle parametrille estimaatin

$$\hat{\theta} = \bar{y}.$$

Eksponenttijakauman parametrille voi helposti johtaa myös SU-estimaatin kummalla tahansa parametroinnilla. Tässä esimerkissä momenttimenetelmä antaa samat estimaatit kuin SU-menetelmä, mutta yleisesti ottaen nämä menetelmät voivat tuottaa erilaiset estimaatit. △

# Luku 5

# Luottamusvälit ja luottamusjoukot

## 5.1 Johdanto

On epärealistista ajatella, että piste-estimaatilla löydettäisiin juuri oikea parametrinarvo. Siksi on tarpeen arvioida piste-estimaatin tarkkuutta. Edellisessä luvussa tätä tarkoitusta varten laskettiin keskivirheitä. Tässä luvussa parametriavaruudesta rajataan joukko (mielellään mahdollisimman pieni joukko), joka sisältää todellisen parametrinarvon suurella todennäköisyydellä (toistetussa otannassa). Tällöin puhutaan luottamusjoukosta.

Jos estimoitava parametri on yksiulotteinen ja jos luottamusjoukko on väli, niin silloin sitä kutsutaan luottamusväliksi.

Useissa tilastollisissa malleissa joudutaan tyytymään likimääräisiin luottamusväleihin (tai -joukkoihin).

Luottamusvälien sijasta on joskus mielekästä tarkastella aivan muunlaisia välejä, esim. enuustevälejä.

## 5.2 Luottamusjoukon määritelmä

Tarkastelemme frekventististä tilastollista mallia eli jakaumaperhettä

$$\{f(\mathbf{y};\theta),\ \theta \in \Theta\},$$

sekä satunnaisvektoria $\mathbf{Y}$, joka noudattaa jakaumaa $f(\mathbf{y};\theta)$ jollakin parametrinarvolla $\theta \in \Theta$.

**Määritelmä 5.1** (Luottamusjoukko)**.** Olkoon $0 < \alpha < 1$ jokin luku. Aineistosta riippuva $\Theta$:n osajoukko $A(\mathbf{y})$ on parametrin $\tau = k(\theta)$ *luottamusjoukko* (engl. *confidence set*) *luottamustasolla* $1-\alpha$ (engl. *confidence level; confidence coefficient*), mikäli vastaava satunnaisvektorista $\mathbf{Y}$ laskettu joukko toteuttaa ehdon

$$P_\theta\left(\tau \in A(\mathbf{Y})\right) \geq 1-\alpha, \qquad \text{kaikilla } \theta \in \Theta. \tag{5.1}$$

Luottamusväli on luottamusjoukko, joka on lukusuoran väli, joten se voidaan määritellä seuraavasti.

**Määritelmä 5.2** (Luottamusväli)**.** Aineistosta laskettua väliä $[L, U]$ sanotaan skalaariparametrin $\tau = k(\theta)$ luottamusväliksi (engl. *confidence interval, CI*) luottamustasolla $1 - \alpha$, jos vastaaville satunnaisille välin päätepisteille $L(\mathbf{Y})$ ja $U(\mathbf{Y})$ pätee

$$P_\theta(L(\mathbf{Y}) \leq \tau \leq U(\mathbf{Y})) \geq 1 - \alpha \tag{5.2}$$

Huomautuksia

- Tässä (kuten tilastollisissa testeissä) $\alpha$ on virhetodennäköisyys. Se on tavallisesti pieni luku, ja tyypillisin valinta on $\alpha = 0.05$, jolloin luottamustaso on $1 - \alpha = 0.95$, eli 95%. Tällöin usein sanotaan lyhyesti, että $A(\mathbf{y})$ on parametrin $\tau$ 95%:n luottamusjoukko. Toinen tavanomainen valinta on $\alpha = 0.01$, mikä vastaa luottamustasoa 99%.

- Satunnaisuus viittaa aineistoa vastaavan satunnaisvektorin $\mathbf{Y}$ jakaumaan (tai toistettuun otantaan).

- Frekventistisessä päättelyssä parametri $\theta$ ei ole satunnainen, vaan kiinteä. Havaintoaineistosta laskettu luottamusjoukko $A(\mathbf{y})$ joko sisältää tai ei sisällä todellista parametrinarvoa $\tau = k(\theta)$, eikä tähän sisälly enää mitään satunnaisuuta. Tämän takia tarvitaan taas uusi termi: luottamusjoukko, luottamusväli. (Ei voida puhua esim. todennäköisyysvälistä.)

- Tahtoisimme luottamusjoukon olevan jollakin tavalla pieni. Koko parametriavaruus $A(\mathbf{y}) = \Theta$ olisi minkä tahansa tason $1 - \alpha$ luottamusjoukko mallin parametrille $\theta$, mutta tämä triviaali luottamusjoukko ei kiinnosta ketään.

- Kaikkein mieluiten konstruoisimme luottamusjoukon sillä tavalla, että kaavassa (5.1) peittotodennäköisyys (engl. *coverage probability*)

  $$P_\theta\left(\tau \in A(\mathbf{Y})\right)$$

  olisi tasan $1-\alpha$ koko parametriavaruudessa. Tietyissä yksinkertaisissa malleissa tämä on mahdollista. Toisinaan tätä vaatimusta on kuitenkin mahdotonta toteuttaa, ja sen takia määritelmässä sallitaan myös epäyhtälö.

## 5.3 Saranasuure

Jos havaintojen jakauma on jatkuva ja jos parametriavaruus on jatkuva, niin eräissä tärkeissä malleissa on mahdollista löytää luottamusjoukko, jolla on tarkalleen haluttu peittotodennäköisyys $1 - \alpha$. Konstruktioon tarvitaan ns. saranasuure.

**Määritelmä 5.3** (Saranasuure)**.** Parametrin $\tau = k(\theta)$ ja satunnaisvektorin $\mathbf{Y}$ funktiota, jonka jakauma ei riipu parametrinarvosta, kutsutaan *saranasuureeksi* (tai napamuuttujaksi) (engl. *pivotal quantity, pivot*) parametrille $\tau$.

**Esimerkki 5.1** Jos $Y_1, \ldots, Y_n$ on satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, ja varianssiparametri $\sigma^2$ on tunnettu luku, niin tällöin

$$\bar{Y} \sim N(\mu, \frac{1}{n}\sigma^2),$$

josta nähdään, että

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

joten $Z$ on saranasuure. Huomaa, että se ei ole tunnusluku, koska sen arvoa ei pystytä laskemaan, jos tunnetaan $Y$:n arvo, mutta ei parametrinarvoa $\theta = \mu$ (tässä $\sigma^2$ on tunnettu luku). △

Jos normaalijakauman varianssi on tuntematon, niin osoittautuu että analogisesti muodostetulla saranasuureella on ns. $t$-jakauma tietyllä vapausasteparametrilla $\nu$. Nämä $t$-jakaumat ovat sellainen jakaumaperhe, jossa jokaista positiivista reaalilukua $\nu > 0$ kohti on olemassa vastaava jakauma $t_\nu$.

## 5.4 Ala- ja yläkvantiilit

Luottamusvälin konstruoimiseen tarvitsemme saranasuureen jakauman ns. kriittisiä arvoja, jotka lasketan sen kvantiilifunktion avulla. Kvantiilifunktion arvoja kutsutaan myös (jakauman) kvantiileiksi tai fraktiileiksi. Määrittelemme kvantiilifunktion vain jatkuvassa tapauksessa.

Olkoon satunnaismuuttujalla $X$ jatkuva jakauma. Oletamme lisäksi, että sen *kertymäfunktio* (engl. *(cumulative) distribution function*

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(v)\,\mathrm{d}v$$

on aidosti kasvava jollakin välillä $(a, b)$, joka sisältää tämän jakauman koko todennäköisyysmassan, ts. $P(X \in (a, b)) = 1$. Tässä yhteydessä salimme välin $(a, b)$ päätepisteille myös arvot $a = -\infty$ tai $b = \infty$. Esimerkiksi

- standardinormaalijakaumalle $N(0, 1)$ tai $t$-jakaumalle $t_\nu$ tällainen väli on $(-\infty, \infty)$;
- khiin neliön jakaumalle $\chi^2_\nu$ tällainen väli on $(0, \infty)$.

Edeltävillä oletuksilla millä tahansa $0 < u < 1$ on olemassa yksikäsitteinen piste $x \in (a, b)$ siten, että

$$F_X(x) = u \tag{5.3}$$

Tämän yhtälön ratkaisua $x = q(u) \in (a, b)$ kutsutaan satunnaismuuttujan $X$ (tai sen jakauman) *u-kvantiiliksi* $q$ (engl. *u quantile*) eli sen *kvantiilifunktion* (engl. *quantile function*) arvoksi pisteessä $0 < u < 1$. Huomaa, että

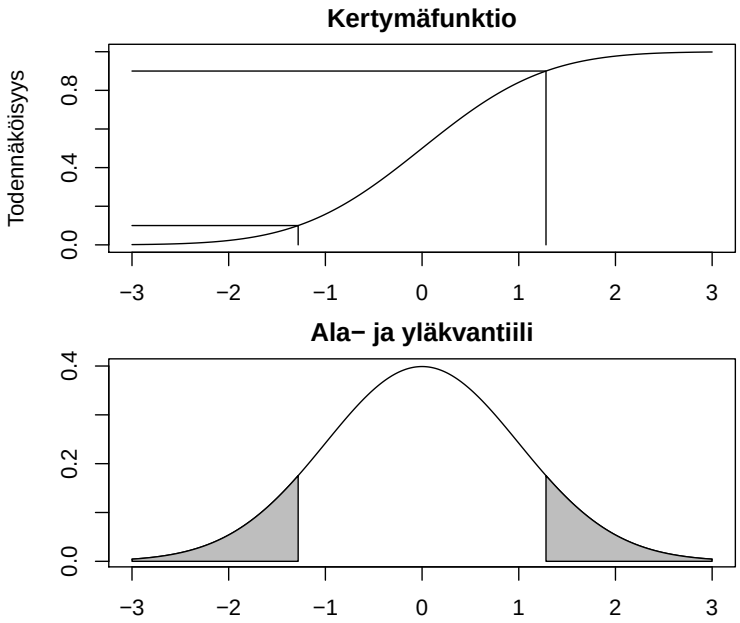
$$q(u) = x \quad \text{eli} \quad F_X(x) = u$$

täsmälleen silloin kuin

$$P(X \leq x) = P(X < x) = u \quad \text{ja} \quad P(X > x) = P(X \geq x) = 1 - u.$$

Ylläolevia todennäköisyyksiä kutsutaan usein *häntätodennäköiyyksiksi* (engl. *tail probability*) tai häntäalueen todennäköisyyksiksi (engl. *tail-area probability*). Jatkuvien jakaumien kohdalla voidaan puhua häntäalueiden pinta-aloista, ks. esim. kuvaa 5.4.

**Kuva 5.1** Standardinormaalijakauman $N(0,1)$ kertymäfunktio sekä sen ala- ja yläkvantiilit, kun $u = 0.1$. Kummankin varjostetun häntäalueen pinta-ala on $u$.




**Määritelmä 5.4** (Ala- ja yläkvantiilit)**.** Sellaista pistettä, josta oikealle jää satunnaismuuttujan todennäköisyysmassasta osuus $0 < u < 1$ kutsutaan ko. jakauman $u$-yläkvantiiliksi (engl. *upper u quantile*).

Sellaista pistettä, josta vasemmalle jää satunnaismuuttujan todennäköisyysmassasta osuus $0 < u < 1$ kutsutaan ko. jakauman $u$-kvantiiliksi tai $u$-alakvantiiliksi.

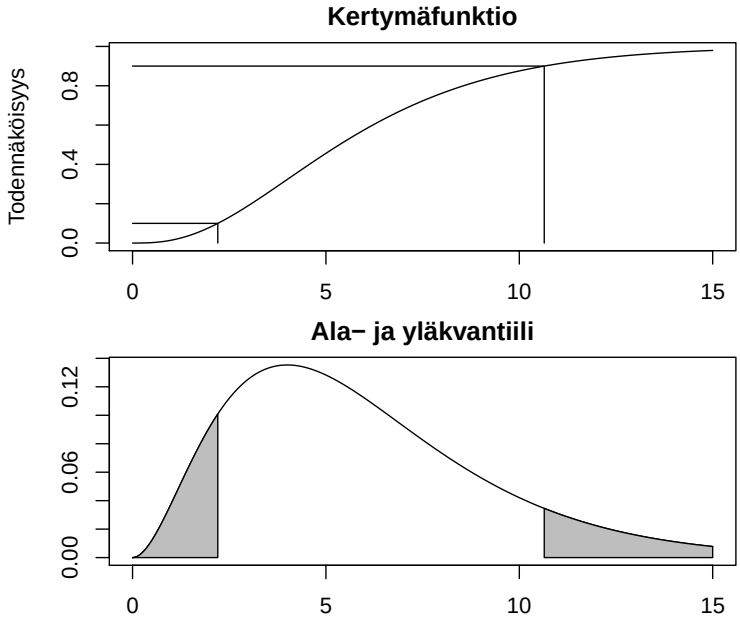
Huomautuksia:

- Termit alakvantiili ja yläkvantiili eivät ole kovin yleisessä käytössä; yleensä käytetään pidempiä ilmaisuja.
- Kvantiilifunktion $q$ avulla lausuttuna $u$-kvantiili eli $u$-alakvantiili on $q(u)$ ja $u$-yläkvantiili on $q(1-u)$.
- Kvantiileja kutsutaan myös fraktiileiksi, ja usein luku $u$ ilmaistaan prosenteissa. Tällöin alakvantiilille käytetään myös nimeä persentiili tai prosenttipiste.

Kuvassa 5.4 havainnollistetaan ala- ja yläkvantiileja sekä vastaavia häntäalueita standardinormaalijakaumalle $N(0,1)$, ja kuvassa 5.4 taas tietylle khiin neliön jakaumalle.

Vanhemmissa tilastotieteen oppikirjoissa on liitteenä laajat taulukot esim. standardinormaalijakauman, t-jakauman ja khiin neliön jakauman kvantiilifunktioista (tai kriittisistä pisteistä). Tällaiset taulukot ovat nykyaikana tarpeettomia. Tilastollisilla ohjelmistoilla saadaan nykyään (tietokoneella tai jopa älypuhelimella) vaivattomasti selville päättelyssä tarvittavat ala- ja yläkvantiilit. Niitä löytyy myös monilta verkkosivuilta, kuten esim.

**Kuva 5.2** Khiin neliön $\chi^2_\nu$ kertymäfunktio sekä sen ala- ja yläkvantiilit, kun $u = 0.1$ ja vapausasteluku $\nu = 6$. Kummankin varjostetun häntäalueen pinta-ala on $u$.




`http://www.statsoft.com/textbook/distribution-tables/`

Esimerkiksi R-ohjelmistolla standardinormaalijakauman alakvantiilit pisteissä 0.1, 0.05, 0.025, 0.01 ja 0.005 saadaan laskettua komennoilla

```
u <- c(0.1, 0.05, 0.025, 0.01, 0.005)
qnorm(u)

## [1] -1.282 -1.645 -1.960 -2.326 -2.576
```

ja yläkvantiilit samoissa pisteissä komennolla

```
qnorm(u, lower = FALSE)

## [1] 1.282 1.645 1.960 2.326 2.576
```

Vastaavasti $t$-jakauman ala- ja yläkvantiilit saadaan laskettua (annetulla $\nu$:n arvolla) komennoilla

```
nu <- 6
qt(u, df = nu)

## [1] -1.440 -1.943 -2.447 -3.143 -3.707

qt(u, df = nu, lower = FALSE)

## [1] 1.440 1.943 2.447 3.143 3.707
```

ja khiin neliön jakauman ala- ja yläkvantiilit komennoilla

```
qchisq(u, df = nu)

## [1] 2.2041 1.6354 1.2373 0.8721 0.6757

qchisq(u, df = nu, lower = FALSE)

## [1] 10.64 12.59 14.45 16.81 18.55
```

Jos jakauma on symmetrinen (ts. sen tiheysfunktio on parillinen funktio), niin tällöin $u$-alakvantiili on $u$-yläkvantiilin vastaluku, sillä symmetriselle jakaumalle

$$q(1-u) = -q(u) \quad \text{kaikille } 0 < u < 1,$$

vrt. kuva 5.4. Tämän takia symmetrisille jakaumille ei tarvita kuin toista jakauman häntää vastaavat kvantiilit. Näille käytetään usein lyhyitä merkintöjä. Tässä monisteessa

$$z_u \quad \text{on } N(0,1)\text{-jakauman } u\text{-yläkvantiili} \tag{5.4}$$
$$t_\nu(u) \quad \text{on } t_\nu\text{-jakauman } u\text{-yläkvantiili.} \tag{5.5}$$

*Varoitus:* Merkinnät ovat eri lähteissä erilaisia. Useissa kirjoissa $z_\alpha$ tarkoittaa $N(0,1)$-jakauman $u$-kvantiilia eikä $u$-yläkvantiilia. Joissakin lähteissä $z_\alpha$ tarkoittaa $N(0,1)$-jakauman $\alpha/2$-yläkvantiilia. Vapausasteluvun merkintä $t$-jakauman yhteydessä on kirjavaa.

## 5.5 Luottamusjoukon muodostaminen saranasuureen avulla

Olkoon nyt $h(\tau, \mathbf{Y})$ saranasuure parametrille $\tau = k(\theta)$. Määritelmän mukaan tämä tarkoittaa sitä, että saranasuureen jakauma on sama riippumatta siitä, mikä on parametrinarvo $\theta \in \Theta$. Oletamme, että tämä jakauma on jatkuva, ja merkitsemme sen kvantiilifunktiota kirjaimella $q$.

Mikäli $0 < \alpha < 1$ on annettu, ja valitsemme luvut $\alpha_1 > 0$ ja $\alpha_2 > 0$ siten, että

$$\alpha = \alpha_1 + \alpha_2$$

niin tällöin

$$P_\theta\left[q(\alpha_1) \le h(\tau, \mathbf{Y}) \le q(1-\alpha_2)\right] = 1 - \alpha, \qquad \text{kaikilla } \theta$$

sillä alempaan jakauman häntään jää saranasuureen jakauman todennäköisyysmassasta osuus $\alpha_1$ ja ylempään häntään osuus $\alpha_2$. Tästä näemme, että

$$A(\mathbf{y}) = \{\tau : q(\alpha_1) \le h(\tau, \mathbf{y}) \le q(1-\alpha_2)\} \tag{5.6}$$

on parametrin $\tau$ luottamusjoukko (luottamus-)tasolla $1-\alpha$. Rajankäynnillä ($\alpha_1 \to 0$ tai $\alpha_2 \to 0$) saadaan vielä seuraavat luottamusjoukot

$$A(\mathbf{y}) = \{\tau : h(\tau, \mathbf{y}) \le q(1-\alpha)\}$$
$$A(\mathbf{y}) = \{\tau : q(\alpha) \le h(\tau, \mathbf{y})\}$$

Se miten virhetodennäköisyys $\alpha$ jaetaan alemmalle ja ylemmälle saranasuureen jakauman hännälle riippuu siitä, minkälainen joukko parametrille saadaan ratkaisemalla ko. epäyhtälöt: epäyhtälöpari (5.6) tai nämä yksittäiset epäyhtälöt. Yleisin valinta on

$$\alpha_1 = \alpha_2 = \alpha/2,$$

ja tällöin voidaan puhua tasahantäisestä (engl. *equal tail*) luottamusvälistä.

Jotta luottamujoukko ei olisi tarpeettoman suuri, niin saranasuureen pitäisi olla järkevä. Se ei saisi (jossain mielessä) hukata aineistoon sisältyvää informaatiota parametrin todellisesta arvosta. Normaalijakaumamallin tapauksessa tulemme käyttämään tällaisia järkeviä saranasuureita.

## 5.6 Luottamusvälejä normaalijakaumamallissa

Tarkastelemme satunnaisotosta $Y_1, \ldots, Y_n$ normaalijakaumasta $N(\mu, \sigma^2)$. Ts. satunnaismuuttujat $Y_i$ ovat riippumattomia, ja niillä on kaikilla normaalijakauma $N(\mu, \sigma^2)$. Muodostamme saranasuureen avulla luottamusvälin parametrille $\mu$ kahdessa tilanteessa.

1) Kun varianssiparametri on tunnettu, jolloin mallin parametri on $\mu$.

2) Kun sekä $\mu$ että $\sigma^2$ ovat tuntemattomia, jolloin mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Lopuksi muodostamme vielä luottamusvälin varianssiparametrille $\sigma^2$.

### 5.6.1 Odotusarvon luottamusväli, kun varianssi on tunnettu

Tämä on se tapaus, jossa luottamusvälin muodostaminen on helpointa ymmärtää. Valitettavasti tätä tapausta ei käytännössä tarvita juuri koskaan, sillä hyvin harvoin normaalijakauman varianssi on tunnettu mutta sen odotusarvo on tuntematon.

Käytämme saranasuuretta (vrt. esimerkki 5.1)

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \tag{5.7}$$

joka noudattaa standardinormaalijakaumaa $N(0, 1)$. Jos $0 < \alpha < 1$ on annettu, ja luvut $\alpha_1 > 0$ ja $\alpha_2 > 0$ on valittu niin, että $\alpha_1 + \alpha_2 = \alpha$, niin todennäköisyydellä $1 - \alpha$ pätee epäyhtälöpari

$$q(\alpha_1) \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq q(1 - \alpha_2), \tag{5.8}$$

missä $q$ on $N(0, 1)$-jakauman kvantiilifunktio.

Merkitään väliaikaisesti

$$q_1 = q(\alpha_1), \qquad \text{ja} \quad q_2 = q(1 - \alpha_2),$$

ja ratkaistaan kaksoisepäyhtälö (5.8) $\mu$:n suhteen:

$$\begin{aligned} & q_1 \leq \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq q_2 \\ \Leftrightarrow \quad & q_1 \frac{\sigma}{\sqrt{n}} \leq \bar{Y}-\mu \leq q_2 \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow \quad & -q_2 \frac{\sigma}{\sqrt{n}} \leq \mu-\bar{Y} \leq -q_1 \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow \quad & \bar{Y}-q_2 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y}-q_1 \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Ratkaisu on väli, joten tulokseksi saadaan luottamusväli parametrille $\mu$.

Tässä tapauksessa on tavanomaista jakaa virhetodennäköisyys tasan alemman ja ylemmän saranasuureen jakauman hännän kesken, jolloin valitaan

$$\alpha_1 = \alpha_2 = \frac{\alpha}{2}.$$

Tällöin $N(0,1)$-jakauman symmetrisyyden ja sopimuksen (5.4) mukaan

$$q_1 = q(\alpha/2) = -z_{\alpha/2} \quad \text{ja} \quad q_2 = q(1-\alpha/2) = z_{\alpha/2},$$

joten

$$P_\mu\left(\bar{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{5.9}$$

Olemme johtaneet parametrin $\mu$ luottamustason $1-\alpha$ luottamusvälin, kun normaalijakaumaa noudattavan populaation varianssi $\sigma^2$ on tunnettu luku, nimittäin

$$[\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] \tag{5.10}$$

Sitä kutsutaan toisinaan $z$-luottamusväliksi, jotta se erotettaisiin myöhemmin käsiteltävästä ns. $t$-luottamusvälistä. Nimi $z$ tulee viitejakaumana käytettävästä $N(0,1)$-jakaumasta, jota noudattavaa satunnaismuuttujaa usein merkitään kirjaimella $Z$. Luottamusväli (5.10) on symmetrinen piste-estimaatin $\bar{y}$ suhteen, ja se voidaan ilmoittaa myös kaavalla

$$\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Aikaisemmin opitun mukaisesti $\sigma/\sqrt{n}$ on $\mu$:n piste-estimaatin (eli otoskeskiarvon $\bar{y}$, joka on SU-estimaatti) keskivirhe. Huomaa, että luottamusvälin leveys riippuu luotamustasosta ja otoskoosta. Otoskoon nelinkertaistaminen puolittaa tämän luottamusvälin leveyden.

Luottamusväli (5.10) on kaksisuuntainen (eli kaksitahoinen) (engl. *two-sided*). On myös mahdollista johtaa yksisuuntaiset (engl. *one-sided*) luottamusvälit. Todennäköisyydellä $1-\alpha$ pätee epäyhtälö

$$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq q(1-\alpha) = z_\alpha,$$

ja kun tämä ratkaistaan $\mu$:n suhteen, nähdään että

$$P_\mu\left(\mu \geq \bar{Y} - z_\alpha\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{5.11}$$

Toisaalta todennäköisyydellä $1-\alpha$ pätee myöskin epäyhtälö

$$-z_\alpha = q(\alpha) \leq \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}},$$

ja kun tämä ratkaistaan, nähdään että

$$P_\mu\left(\mu \leq \bar{Y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{5.12}$$

Ts. seuraavat aineistosta $\mathbf{y}$ lasketut yksisuuntaiset välit ovat luottamustason $1-\alpha$ luottamusvälejä

$$[\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty) \tag{5.13}$$

$$(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}] \tag{5.14}$$

### 5.6.2 Aineistosta lasketun luottamusvälin tulkinta

Lasketaan nyt 95% luottamusväli (5.10) (eli kaksisuuntainen $z$-luottamusväli) populaation odotusarvolle $\mu$ käyttämällä kuvan 4.3 aineistoa olettaen, että tiedämme että $\sigma^2 = 1$. (Simuloinnissa käytettiin tätä varianssia.) Käyttämällä tietoja

$$\bar{y} = 0.726, \quad n = 10, \quad z_{0.025} = 1.96$$

saadaan laskettua parametrille $\mu$

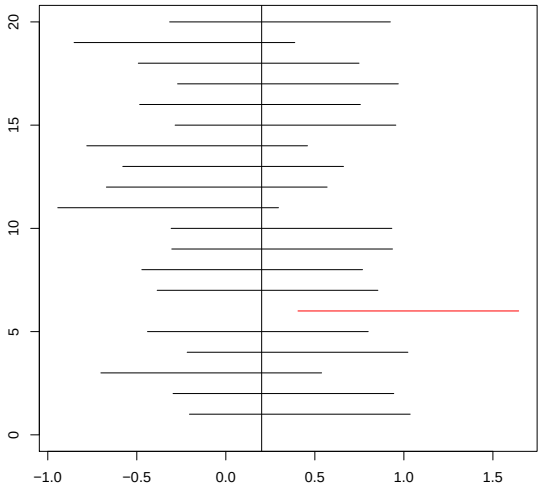
- piste-estimaatti 0.73 (eli SU-estimaatti $\bar{y}$)
- estimaatin keskivirhe 0.32 (eli $\sigma/\sqrt{n}$)
- 95%:n luottamusväli $[0.10, 1.35]$ (eli $\bar{y} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$).

Simuloinnissa käytetty todellinen parametrinarvo $\mu = 0.2012$ kuuluu laskettuun luottamusväliin.

R:n peruspaketeissa ei ole toteutettuna $z$-luottamusväliä. Ohjelmiston kehittäjät ovat luultavasti arvioineet, ettei sitä todellisuudessa koskaan tarvita. Tarvittavat laskut saadaan tehtyä esim. seuraavasti.

```
y <- c(1.38, -0.96, 1.08, 0.41, 0.48, 2.45, -0.8, 0.27, 1.79,
    1.16)
n <- length(y)
z <- qnorm(0.05/2, lower = FALSE)
sigma <- 1
sigma/sqrt(n)

## [1] 0.3162

mean(y) - z * sigma/sqrt(n)

## [1] 0.1062

mean(y) + z * sigma/sqrt(n)

## [1] 1.346
```

**Kuva 5.3** 20 kappaletta kaavalla (5.10) laskettua $z$-luottamusväliä jakaumasta $N(\mu, 1)$ simuloiduille, kokoa $n = 10$ oleville aineistoille. Todellinen parametrinarvo on merkitty pystyviivalla. Kun normaalijakauman varianssi tunnetaan, niin luottamusvälin leveys pysyy vakiona.




Ennen aineiston keräämistä (ts. simulointia) tiedämme, että aineistosta laskettava 95%:n luottamusväli tulee sisältämään todellisen populaation keskiarvon todennäköisyydellä 95%. Sitten aineisto kerättiin (tässä: simuloitiin), ja luottamusväliksi saatiin $[0.10, 1.35]$.

*Kysymys: Voimmeko sanoa, että $\mu \in [0.10, 1.35]$ todennäköisyydellä 0.95?*

Pysähdy pohtimaan tätä kysymystä, ja muodosta asiasta oma mielipiteesi ennen kuin luet alla olevan vastauksen!

*Vastaus:*

- Aineistosta laskettu luottamusväli joko sisältää todellisen parametrinarvon tai tai ei sisällä sitä. Emme voi pelkästään aineistoa tarkastelemalla sanoa mitään sen enempää, vaan tätä varten pitäisi tuntea todellinen parametrinarvo.

- Frekventistisessä tilastotieteessä parametri on tuntematon, mutta kiinteä (siis ei-satunnainen). Tämän lähestymistavan puitteissa väite $\mu \in [0.10, 1.35]$ on joko tosi tai epätosi (nyt se on tosi). Tällaisen väitteen todennäköisyys ei taatusti ole 0.95.

Tämä tulkinnallinen vaikeus ei liity kaavaan (5.10), vaan luottamusvälin käsitteeseen. Luottamusvälin määritelmässä todennäköisyys viittaa siihen, että aineistoa pidetään satunnaisvektorina, jolla on jakauma $f(\mathbf{y}; \theta)$. Tällöin luottamusvälin päätepisteet eli tunnusluvut $L(\mathbf{Y})$ ja ovat $U(\mathbf{Y})$ ovat satunnaismuuttujia, ja todennäköisyydellä $1-\alpha$ todellinen parametrinarvo sisältyy satunnaiselle välille $[L(\mathbf{Y}), U(\mathbf{Y})]$.

Tätä tulkintaa voidaan havainnollistaa ajattelemalla toistettua aineistonkeruuta, jota on havainnollistettu kuvassa 5.3. Jos laskemme luottamusvälin (5.10) suurelle määrälle $r$ normaalijakaumasta $N(\mu, \sigma^2)$ simuloituja kokoa $n$ olevia otoksia (jossa $\sigma^2$ on tunnettu)

$$\mathbf{y}_1, \ldots, \mathbf{y}_r,$$

niin saamme $r$ kappaletta luottamusvälejä

$$[L(\mathbf{y}_1), U(\mathbf{y}_1)], \ldots, [L(\mathbf{y}_r), U(\mathbf{y}_r)].$$

Näistä osapuilleen $r(1-\alpha)$ kappaletta sisältää todellisen parametrinarvon ja $r\alpha$ kappaletta ei sisällä sitä.

*Kysymys:* Hyvä on, *en saa sanoa*, että $\mu \in [0.10, 1.35]$ todennäköisyydellä 0.95. Miten sitten *saan* tulkita aineistosta lasketun luottamusvälin?

*Vastaus:* Aineiston perusteella paras arvauksemme parametrinarvolle on pisteestimaatti 0.73. 95%:n luottamusvälillä $[0.10, 1.35]$ olevat arvot ovat kaikki kohtuullisessa sopusoinnussa havaintojen kanssa. Sekä luottamusvälin leveys että estimaatin keskivirhe kuvastavat tietomme epävarmuutta parametrinarvosta tämän aineiston valossa. Väli on laskettu sellaisella menetelmällä, joka toistetussa aineistonkeruussa mallin oletukset toteuttavasta populaatiosta sisältäisi todellisen parametrinarvon noin 95% toistoista. Ennen aineistonkeruuta todennäköisyys oli 95%, että siitä laskettava 95%:n luottamusväli tulee sisältämään oikean parametrinarvon (olettaen tietenkin, että populaatio toteuttaa mallioletukset).

### 5.6.3 Odotusarvon luottamusväli, kun varianssi on tuntematon

Tässä tilanteessa sekä $\mu$ että $\sigma^2$ ovat tuntemattomia, jolloin mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$. Haluamme muodostaa luottamusvälin odotusarvoparametrille

$$\mu = k(\mu, \sigma^2).$$

(Tässä funktio $k$ vain palauttaa ensimmäisen argumenttinsa arvon.)

Kun varianssi oli tunnettu, käytimme saranasuuretta

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}.$$

Kun varianssi on tuntematon, $Z$ ei ole saranasuure, koska se riippuu paitsi aineistosta ja kiinnostusparametrista $\mu$, myös haittaparametrista $\sigma^2$. Ajatuksena on kuitenkin matkia mahdollisimman tarkoin aikaisempaa konstruktiota. Koska populaation keskihajonta $\sigma$ on tuntematon, sen tilalle sijoitetaan otoskeskihajontaa (4.17) vastaava satunnaismuuttuja $S$. Tässä mallissa satunnaismuuttuja

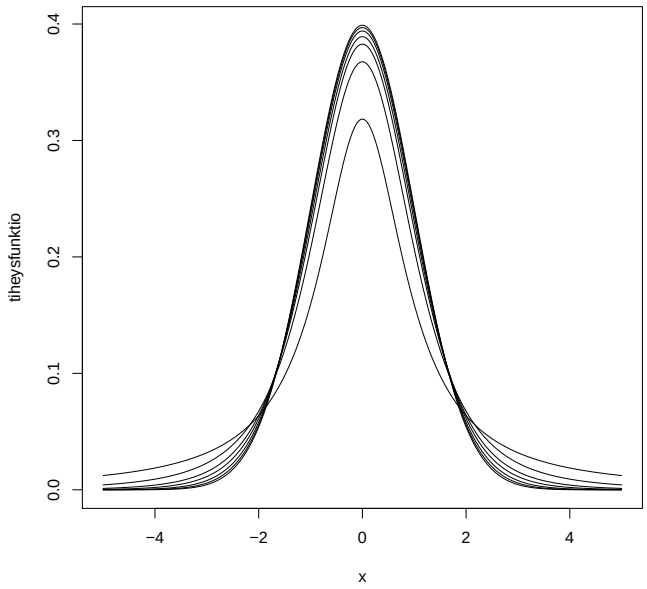
$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \tag{5.15}$$

osoittautuu saranasuureeksi. Sen jakauma on tietty $t$-jakauma.

**Määritelmä 5.5.** Jos $\nu > 0$ ja $Z \sim N(0,1)$ ja $X \sim \chi^2_\nu$ ja $Z$ ja $X$ ovat riippumattomia, niin satunnaismuuttujalla

$$Y = \frac{Z}{\sqrt{X/\nu}}$$

**Kuva 5.4** $t_\nu$-jakauman tiheysfunktioita vapausasteluvun $\nu$ arvoilla 1, 3, 6, 10, 20 ja 50. Vertailun vuoksi kuvassa on myös standardinormaalijakauman $N(0,1)$ tiheysfunktio, jota voidaan pitää $t$ jakaumana vapausasteluvulla $\infty$. Tiheysfunktion arvo pisteessä $x = 0$ on sitä suurempi mitä suurempi on vapausasteluku $\nu$. Häntäalueilla järjestys on päinvastainen.




on $t_\nu$-jakauma eli $t$-jakauma vapausasteluvulla $\nu$ (engl. *$t$ distribution with $\nu$ degrees of freedom*).

Määritelmän avulla on mahdollista johtaa $t_\nu$-jakauman tiheysfunktio, mutta tätä kaavaa ei tässä yhteydessä tarvita. Tiheysfunktio osoittautuu parilliseksi funktioksi, joten $t_\nu$-jakauma on symmetrinen. Kuvassa 5.4 esitetään $t_\nu$-jakauman tiheysfunktio muutamilla vapausasteluvun arvoilla. Kun $\nu$ kasvaa, jakauman tiheysfunktio lähestyy standardinormaalijakauman $N(0,1)$ tiheysfunktiota. $t$-jakaumaa kutsutaan myös Studentin $t$-jakaumaksi W. S. Gossetin v. 1908 julkaiseman artikkelin kunniaksi. Tilastotieteilijä W. S. Gosset (1876–1937) työskenteli tuohon aikaan Guinessin panimolla. Panimo oli kieltänyt liikesalaisuuksien suojelemiseksi työntekijöitään julkaisemasta mitään kirjoituksia omalla nimellään, minkä takia Gosset käytti julkaisussa salanimeä Student.

Seuraavaksi tarvitsemme jaksossa 4.4.2 kerrottua tietoa satunnaismuuttujaparin $(\bar{Y}, S^2)$ yhteisjakaumasta (kaavat (4.14), (4.15) ja (4.16)):

- $\bar{Y}$ ja $S^2$ ovat riippumattomia
- $\bar{Y} \sim N(\mu, \frac{1}{n}\,\sigma^2)$,
- $\dfrac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$.

Edellä mainittujen jakaumatulosten ja $t$-jakauman määritelmän perusteella satunnaismuuttujalla

$$\frac{(\bar{Y}-\mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{n-1}{\sigma^2}S^2/(n-1)}} = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$$

on $t$-jakauma vapausasteluvulla $n-1$, mutta sieventämällä nähtiin, että tämä satunnaismuuttuja on sama kuin kaavan (5.15) satunnaismuuttuja $T$.

Olkoon $q$ nyt $t_{n-1}$-jakauman kvantiilifunktio, ja olkoon $0 < \alpha < 1$. Todennäköisyydellä $1-\alpha$ pätee epäyhtälöt

$$q(\alpha_1) \leq \frac{\bar{Y}-\mu}{S/\sqrt{n}} \leq q(1-\alpha_2),$$

jossa $\alpha_1 > 0$ ja $\alpha_2 > 0$ ovat sellaisia lukuja, joiden summa on $\alpha$. Tästä saadaan ratkaistua väli odotusarvolle $\mu$ aivan samoilla vaiheilla kuin aikaisemmin, ja tulos on

$$\bar{Y} - q(1-\alpha_2)\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} - q(\alpha_1)\frac{S}{\sqrt{n}}$$

Jos tässä valitaan $\alpha_1 = \alpha_2 = \alpha/2$, ja huomataan, että

$$q(\alpha/2) = -t_{n-1}(\alpha/2) \quad \text{ja} \quad q(1-\alpha/2) = t_{n-1}(\alpha/2),$$

niin päädytään siihen, että

$$P_{(\mu,\sigma^2)}\left(\bar{Y} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}\right) = 1-\alpha, \tag{5.16}$$

kaikilla $\mu \in \mathbb{R}$ ja kaikilla $\sigma^2 > 0$.

Vastaava aineistosta $\mathbf{y}$ laskettu väli

$$[\bar{y} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}] = \bar{y} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}, \tag{5.17}$$

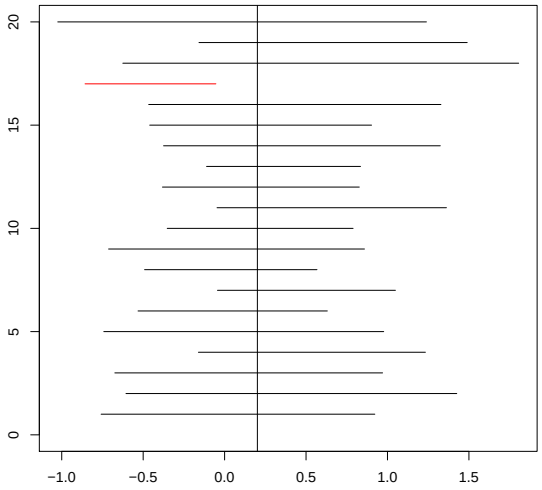
jossa $\bar{y}$ on otoskeskiarvo ja $s$ on otoskeskihajonta, on normaalijakauman odotusarvon $\mu$ luottamusväli luottamustasolla $1-\alpha$. Sitä kutsutaan usein $t$-luottamusväliksi (viitejakauman $t_{n-1}$ mukaan). Huomaa, että $\bar{y}$ on myös parametrin $\mu$ SU-estimaatti ja että $s/\sqrt{n}$ on tämän estimaatin keskivirhe.

Suure $t_{n-1}(\alpha/2)$ lähestyy lukua $z_{\alpha/2}$, kun otoskoko kasvaa. Esimerkiksi luottamustasoa 95% vastaa $\alpha = 0.05$, ja $z_{0.025} = 1.96$. Otoskokoja $n = 50, 100, 200, 500$ ja 1000 vastaavat seuraavat $t$-jakaumaperheen $\alpha/2$-yläkvantiilit

```
n <- c(50, 100, 200, 500, 1000)
qt(0.05/2, df = n - 1, lower = FALSE)

## [1] 2.010 1.984 1.972 1.965 1.962
```

Väli (5.17) on symmetrinen piste-estimaatin $\bar{y}$ suhteen. Toisin kuin $z$-luottamusvälin yhteydessä, $t$-luottamusvälin leveys vaihtelee aineistosta toiseen, koska välin leveys määräytyy aineiston otoskeskihajonnasta. Tämän $t$-luottamusvälin leveys riippuu luottamustasosta ja otoskoosta. Otoskoon nelinkertaistaminen karkeasti ottaen puolittaa kaksisuuntaisen $t$-luottamusvälin leveyden (mutta tämä ei pidä paikkaansa tarkalleen).

**Kuva 5.5** 20 kappaletta kaavalla (5.17) laskettua $t$-luottamusväliä jakaumasta $N(\mu, \sigma^2)$ simuloiduille, kokoa $n = 10$ oleville aineistoille. Todellinen odotusarvoparametrin arvo on merkitty pystyviivalla. Kun normaalijakauman varianssi on tuntematon, niin luottamusvälin leveys vaihtelee otoksesta toiseen.




Kuvassa 5.5 näytetään 20 kappaletta $t$-luottamusvälejä, jotka on laskettu aineistoista, jotka on generoitu tietystä normaalijakaumasta.

Kuten jaksossa 5.6.2 selitettiin, aineistosta lasketulla luottamusvälillä ei ole todennäköisyystulkintaa, vaan se joko sisältää todellisen parametrinarvon tai ei sisällä sitä, emmekä (todellisessa tilanteessa) tiedä kumpi tilanne on kyseessä. Todennäköisyystulkinta vaatii sitä, että tulkitsemme välin päätepisteet satunnaismuuttujiksi tai ajattelemme toistettua aineiston keruuta tai ajattelemme tilannetta, joka vallitsi ennen kuin aineisto kerättiin. Kaikkien luottamusvälin sisällä olevien arvojen voidaan ajatella olevan kohtuullisen hyvin sopusoinnussa aineiston kanssa. Paras arvauksemme on parametrin piste-estimaatti.

**Esimerkki 5.2** Kuvan 4.3 aineistolle

$$\bar{y} = 0.726, \quad s = 1.074, \quad n = 10, \quad t_9(0.025) = 2.262.$$

Parametrin $\mu$ piste-estimaatti on 0.73, sen keskivirhe on 0.34 (kaavalla $s/\sqrt{n}$), ja 95%:n luottamusväli on $[-0.04, 1.50]$.

Tavallisesti luottamusväli lasketaan tietokoneella. R:llä tämä onnistuu seuraavasti

```
y <- c(1.38, -0.96, 1.08, 0.41, 0.48, 2.45, -0.8, 0.27, 1.79,
    1.16)
t.test(y)

##
##  One Sample t-test
```

```
##
## data:  y
## t = 2.137, df = 9, p-value = 0.0613
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.04244  1.49444
## sample estimates:
## mean of x
##     0.726
```

Valitettavasti tässä näkyy luottamusvälin selvittämisen kannalta tarpeetonta tietoa; pelkän välin saisi selville antamalla komennon `t.test(y)$conf.int`.

```
t.test(y)$conf.int
```

```
## [1] -0.04244  1.49444
## attr(,"conf.level")
## [1] 0.95
```

Jos tahdotaan käyttää muita luottamustasoja kuin 95%, kuten esim. luottamustasoa 99%, niin haluttu luottamustaso pitää antaa `t.test`-funktiolle tyyliin `t.test(y, conf.level = 0.99)`. Valitettavasti `t.test` ei raportoi pisteestimaatin keskivirhettä, mutta sen saa laskettua helposti erikseen seuraavasti.

```
sd(y)/sqrt(length(y))
```

```
## [1] 0.3397
```

Mikään ei pakota meitä laskemaan luottamusväliä vain yhdellä luottamustasolla 0.95. Kuvassa 5.6 näytetään luottamusvälin päätepisteet luottamustason funktiona. △

### 5.6.4 Varianssiparametrin luottamusväli

Oletamme, että sekä $\mu$ että $\sigma^2$ ovat tuntemattomia, jolloin mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$. Haluamme muodostaa luottamusvälin varianssiparametrille

$$\sigma^2 = k(\mu, \sigma^2).$$

(Nyt funktio $k$ palauttaa toisen argumenttinsa arvon.)

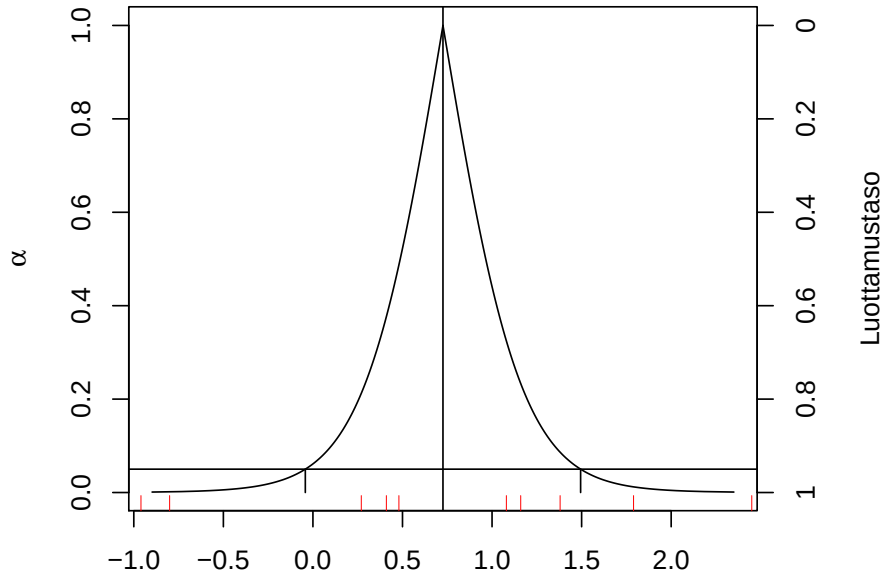
Käytämme saranasuureena sopivasti skaalattua otosvarianssia, sillä tiedämme, että

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}.$$

Jos $q$ on $\chi^2_{n-1}$-jakauman kvantiilifunktio, ja $0 < \alpha < 1$ sekä $\alpha_1 > 0$ ja $\alpha_2 > 0$ ovat lukuja siten, että $\alpha = \alpha_1 + \alpha_2$, niin todennäköisyydellä $1 - \alpha$ pätee

$$q(\alpha_1) \leq \frac{n-1}{\sigma^2} S^2 \leq q(1 - \alpha_2)$$

**Kuva 5.6** Kuvan 4.3 aineistolle lasketut parametrin $\mu$ kaksisuuntaiset $t$-luottamusvälit. Piste-estimaatti sekä 95%:n luottamusväli on korostettu pystyviivoilla. Aineisto on esitetty $x$-akselin yläpuolella olevilla pienillä viivoilla.




Kun tämä epäyhtälö ratkaistaan muuttujan $\sigma^2$ suhteen, saadaan väli

$$\frac{n-1}{q(1-\alpha_2)}\,S^2 \le \sigma^2 \le \frac{n-1}{q(\alpha_1)}\,S^2$$

Tässäkin on tapana valita $\alpha_1 = \alpha_2 = \alpha/2$, jolloin varianssiparametrille $\sigma^2$ saadaan kaksisuuntainen tason $1-\alpha$ luottamusväli

$$[\frac{n-1}{q(1-\alpha/2)}\,s^2, \frac{n-1}{q(\alpha/2)}\,s^2], \tag{5.18}$$

jossa $s^2$ on otosvarianssi (joka on varianssiparametrin piste-estimaatti) ja $q$ on $\chi^2_{n-1}$-jakauman kvantiilifunktio. Tämä väli ei ole symmetrinen piste-estimaatin suhteen.

Kuvan 4.3 aineistolle

$$s^2 = 1.1539, \quad n = 10, \quad q(0.025) = 2.7004, \quad q(0.975) = 19.0228,$$

ja näistä luvuista laskettu varianssiparametrin piste-estimaatti on 1.15 ja 95%:n luottamusväli on $[0.55, 3.85]$. Tämä väli sisältää todellisen simuloinnissa käytetyn varianssin $\sigma^2 = 1$.

## 5.7 Likimääräinen luottamusväli

Jos otoskoko $n$ on suuri ja jos piste-estimaattorin $\hat{\tau}(\mathbf{Y})$ otantajakauma on osapuilleen $\tau$-keskinen normaalijakauma, niin tällöin osapuilleen todennäköisyydellä

$1-\alpha$ pätee epäyhtälö

$$-z_{\alpha/2} \leq \frac{\hat{\tau}(\mathbf{Y}) - \tau}{\sqrt{\mathrm{var}_\theta\, \hat{\tau}(\mathbf{Y})}} \leq z_{\alpha/2}.$$

Kun tämä epäyhtälöpari ratkaistaan parametrin $\tau$ suhteen, saadaan väli

$$\hat{\tau}(\mathbf{Y}) - z_{\alpha/2}\sqrt{\mathrm{var}_\theta\, \hat{\tau}(\mathbf{Y})} \leq \tau \leq \hat{\tau}(\mathbf{Y}) + z_{\alpha/2}\sqrt{\mathrm{var}_\theta\, \hat{\tau}(\mathbf{Y})}$$

Tässä estimaattorin otantajakauman keskihajonta $\sqrt{\mathrm{var}_\theta\, \hat{\tau}(\mathbf{Y})}$ on tavallisesti tuntematon. Jos se korvataan estimaatilla, eli estimaatin $\hat{\tau}$ keskivirheellä, niin päädytään nimellistä (engl. *nominal*) $1-\alpha$ luottamustasoa vastaavaan (kaksisuuntaiseen) likimääräiseen luottamusväliin

$$\hat{\tau} \pm z_{\alpha/2} \times \mathrm{se}, \tag{5.19}$$

jossa suure (se) on (jollakin järkevällä tavalla laskettu) estimaatin $\hat{\tau}$ keskivirhe.

Koska $z_{0.025} = 1.96$, niin suurella otoskoolla erityisesti

$$\hat{\tau} \pm 2 \times \mathrm{se},$$

on likimääräinen 95%:n luottamusväli. Koska $z_{0.16} = 0.994$, niin suurella otoskoolla

$$\hat{\tau} \pm \mathrm{se},$$

on likimääräinen 68%:n luottamusväli.

Esimerkiksi binomikokeessa onnistumistodennäköisyyden luottamusväli lasketaan tyypillisesti tällä periaatteella. SU-estimaattori

$$\hat{p}(\mathbf{Y}) = \bar{Y}$$

(eli onnistumisten suhteellinen osuus) on harhaton, ja sen varianssi on

$$\mathrm{var}_p\, \bar{Y} = \frac{1}{n}\, p(1-p).$$

Koska estimaattori on keskiarvo $n$ riippumattomasta ja samoin jakautuneesta satunnaismuuttujasta, sen jakaumaa voidaan suurella otoskoolla approksimoida normaalijakaumalla (todennäköisyyslaskennan keskeisen raja-arvolauseen perusteella). Kun keskivirheelle käytetään kaavaa

$$\mathrm{se} = \sqrt{\frac{1}{n}\, \hat{p}\,(1-\hat{p})},$$

saadaan binomikokeen onnistumistodennäköisyydelle $p$ likimääräinen $1-\alpha$ luottamusväli

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\, \hat{p}\,(1-\hat{p})}, \tag{5.20}$$

joka kerrotaan kaikissa tilastotieteen alkeisoppikirjoissa. Jos $0 < p < 1$ on kiinteä, ja otoskoko $n$ kasvaa rajatta, niin todennäköisyyslaskennan keinoilla voidaan osoittaa, että vastaavan satunnaisen luottamusvälin peittotodennäköisyys lähestyy arvoa $1-\alpha$, joten suurella otoskoolla tämän välin peittotodennäköisyys on suunnilleen $1-\alpha$.

Edellinen asymptoottinen perustelu jättää avoimeksi sen, milloin otoskoko on riittävän suuri. Tämän takia tarkastelemme lähemmin likimääräisen perinteisen likimääräisen luottamusvälin (5.20) ominaisuuksia. Se voi äärellisellä otoskoolla käyttäytyä kummallisella tavalla:

- Sen päätepisteet voivat olla parametriavaruuden ulkopuolella; käytännössä luottamusväliksi pitäisi otta välin (5.20) sekä parametriavaruuden leikkaus.
- Väli surkastuu yhdeksi pisteeksi, jos koesarjassa ei joko onnistuta yhtään kertaa tai jos ei epäonnistuta yhtään kertaa; parametriavaruuden reunojen lähellä tätä väliä ei kannata käyttää.

Kuvassa 5.7 otoskoko on $n = 20$. Siinä esitetään eri onnistumisten lukumäärää $0 \leq k \leq n$ vastaavat $n+1$ mahdollista luottamusväliä laskettuna kaavalla (5.20). Kuvassa on myös piirretty luottamusvälin todellinen peittotodennäköisyys

$$P_p(L(\mathbf{Y}) \leq p \leq U(\mathbf{Y})).$$

Kuvasta näemme, että tällä pienehköllä otoskoolla tämän luottamusvälin todellinen peittotodennäköisyys on melkein koko parametriavaruudessa paljon pienempi kuin nimellinen peittotodennäköisyys. Ainakaan otoskoolla $n = 20$ tätä perinteistä likimääräistä luottamusväliä ei pitäisi käyttää.

## 5.8 Muita luottamusvälejä binomikokeessa

Likimääräisen luottamusvälin (5.20) todellinen peittotodennäköisyys (kun väli tulkitaan satunnaiseksi) käyttäytyy millä tahansa äärellisellä otoskoolla $n$ huonosti joissakin parametriavaruuden pisteissä. Parametriavaruuden reunojen lähellä tämän välin peittotodennäköisyys romahtaa nollaan, koska itse väli surkastuu kummallakin rajalla pisteeksi. Tämän lisäksi todellinen peittotodennäköisyys voi olla selvästi nimellistä tasoa pienempi muuallakin vielä suurehkolla otoskoolla, ks. artikkelia Brown, Cai ja DasGupta [1]. Nämä kirjoittajat toteavat tästä luottamusvälistä seuraavaa:

> ... the performance of this standard interval is persistently chaotic and unacceptably poor. Indeed its coverage properties defy conventional wisdom. The performance is so erratic and the qualifications given in the influential texts are so defective that the standard interval should not be used.

Newcombe [2] vertaa empiirisesti seitsämää erilaista mentelmää luottamusvälin laskemiseksi, ja hän käyttää vertailussa peittotodennäköisyyden lisäksi muitakin kriteereitä. Newcombe kommentoi tätä traditionaalista luottamusväliä (ja sen parannusta, jossa käytetään jatkuvuuskorjausta) seuraavasti,

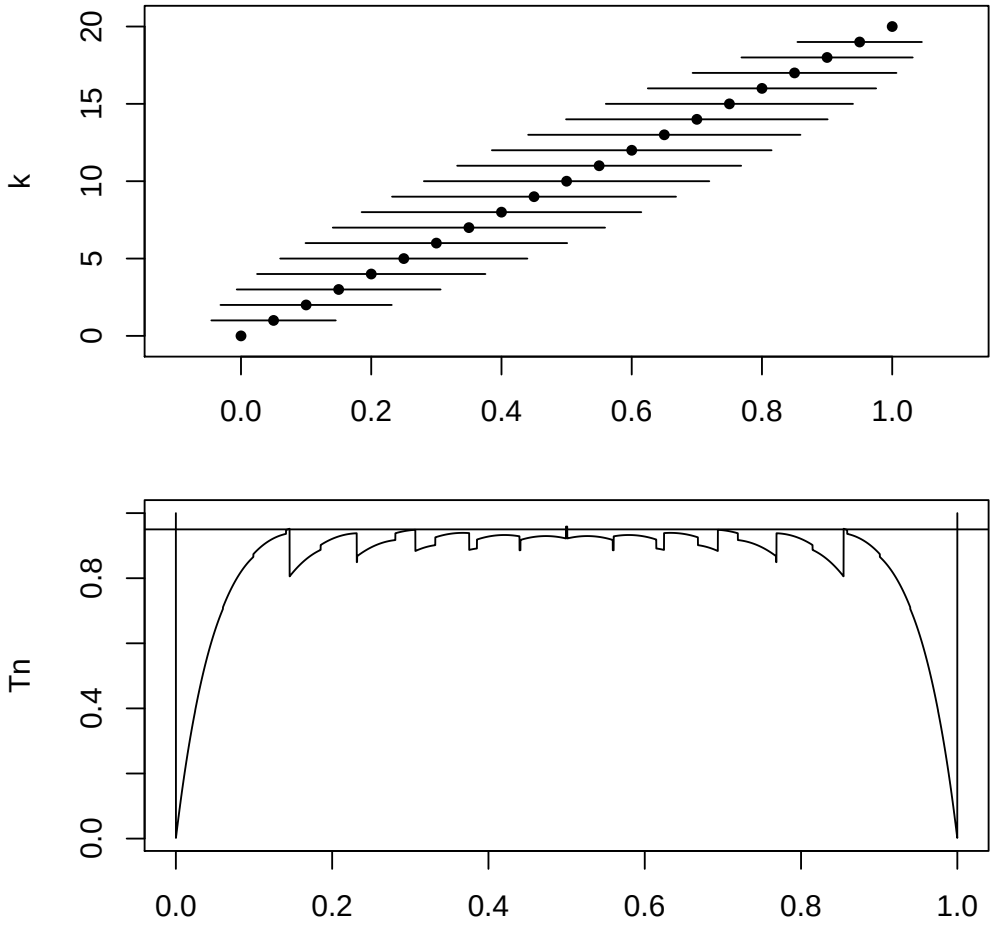
> ... it is strongly recommended that intervals calculated by these methods should no longer be acceptable for the scientific literature

Nämä neuvot on syytä ottaa huomioon. Älkää käyttäkö perinteistä likimääräistä luottamusväliä (5.20) omissa töissänne.

Mainituissa artikkeleissa käydään läpi monta vaihtoehtoehtoista tapaa muodostaa luottamusväli onnistumistodennäköisyydelle. Esimerkiksi Wilsonin v. 1927 ehdottama luottamusväli osoittautuu edellistä selvästi paremmaksi. Myös se perustuu siihen approksimaatioon, että suureella

$$\frac{\hat{p}(\mathbf{Y}) - p}{\sqrt{\mathrm{var}_p(\hat{p}(\mathbf{Y}))}} = \frac{\hat{p}(\mathbf{Y}) - p}{\sqrt{\frac{1}{n}\, p(1-p)}}$$

**Kuva 5.7** Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat nimellistä luottamustasoa 95% vastaavat luottamusvälit (5.20), kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti $k/n$ on merkitty pisteellä. Alemmassa kuvassa on esitetty (satunnaiseksi ymmärretyn) luottamusvälin peittotodennäköisyys todellisen onnistumistodennäköisyyden $p$ funktiona. Nimellinen luottamustaso on osoitettu vaakaviivalla.



on osapuilleen standardinormaalijakauma $N(0,1)$, mutta tällä kertaa tätä tietoa käytetään hyväksi hienostuneemmalla tavalla. Nyt luottamusväli muodostetaan ratkaisemalla epäyhtälöpari

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

muuttujan $p$ suhteen toisen asteen yhtälön ratkaisukaavan avulla. Tuloksena saadaan Wilsonin luottamusväli

$$\frac{\hat{p}+\frac{1}{2n}z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})+\frac{1}{4n^2}z^2_{\alpha/2}}}{1+\frac{1}{n}z^2_{\alpha/2}}, \tag{5.21}$$

joka on luottamusväliä (5.20) selkeästi parempi. (Luottamusväliä kutsutaan myös nimellä *Wilson score interval*, sen takia, että se voidaan johtaa invertoimalla tässä tilanteessa ns. pistemäärätesti, engl. *score test*.) Myös Wilsonin luottamusväli on likimääräinen, sillä luottamusvälin määritelmän epäyhtälö (5.2) ei sille toteudu. Kuvassa 5.8 esitetään Wilsonin luottamusvälin toiminta, kun $n = 20$. Tämä luottamusväli ei surkastu pisteeksi, jos onnistumisia on nolla tai $n$.

Clopper ja Pearson esittivät v. 1934 erään tavan muodostaa ns. tarkka (engl. *exact*) luottamusväli onnistumistodennäköisyydelle. Termi tarkka tarkoittaa tässä sitä, että kyseinen luottamusväli ei ole likimääräinen, vaan määritelmän (ks. kaava (5.2)) mukainen, eli

$$P_p(L(\mathbf{Y}) \leq p \leq U(\mathbf{Y})) \geq 1-\alpha, \quad \text{kaikilla } 0 < p < 1.$$

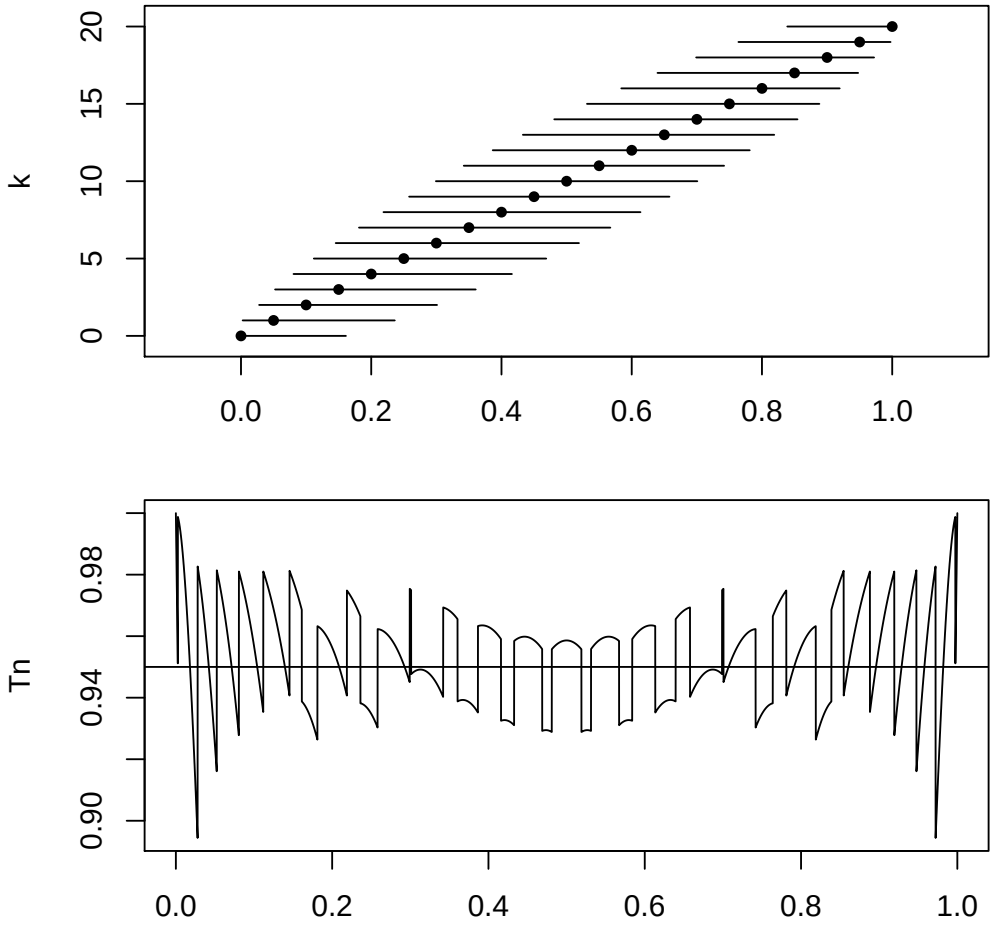
Lisäksi alarajaa $1-\alpha$ ei voida yhtään suurentaa ilman, että epäyhtälö rikkoontuisi jollakin otoskoolla $n$ ja jollakin $0 < p < 1$ . Muualla väli on turhan konservatiivinen, eli sen todellinen peittotodennäköisyys on aidosti lukua $1-\alpha$ suurempi, kuten kuvasta 5.9 nähdään, kun otoskoko $n = 20$.

Silloin kuin havaintosatunnaisvektorin jakauma on diskreetti, niin yleensä aina joudutaan tekemään luottamusvälien kanssa samantapaisia kompromisseja. Joko käytetään likimääräisiä luottamusvälejä, joiden todellinen peittotodennäköisyys on joskus pienempi kuin niiden nimellinen peittotodennäköisyys, tai sitten käytetään tarkkaa luottamusväliä (mikäli sellainen sattuu olemaan saatavilla), joka on useimmilla parametrinarvoilla turhan konservatiivinen.

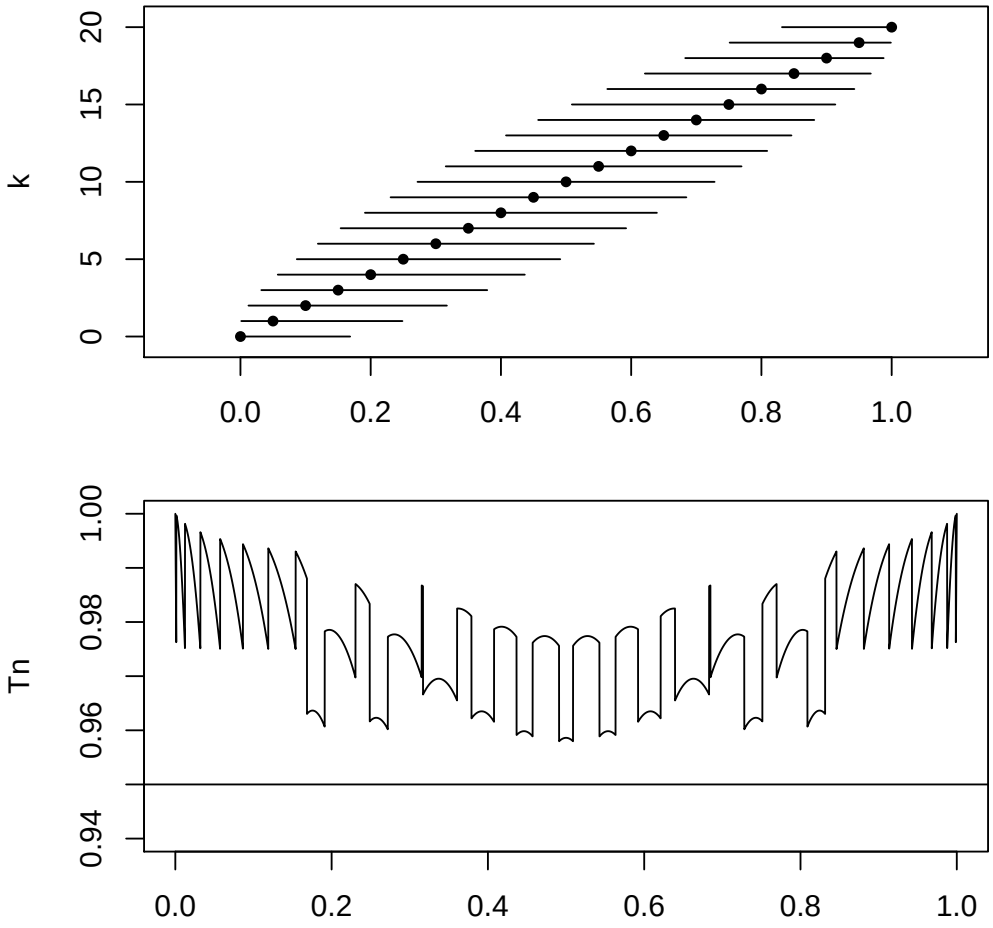
Tietokoneella minkä tahansa edellä mainitun binomijakauman luottamusvälin laskeminen on yhtä helppoa. Esim. R-ohjelmistossa nämä luottamusvälit on helppo laskea `Hmisc`-kirjaston funktiolla `binconf`. Nimellistä luottamustasoa 95% vastaavat välit saadaan laskettua seuraavalla tavalla. (Myös funktio `binom.test` laskee Clopperin ja Pearsonin tarkan luottamusvälin. Funktio `prop.test` laskee erään luottamusvälin, joka on sukua Wilsonin luottamusvälille. Valitettavasti tämän funktion dokumentaatiosta on vaikea saada selvää.)

```
n <- 20
k <- 4
library(Hmisc)
binconf(k, n, method = "asymptotic")
```

**Kuva 5.8** Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat nimellistä luottamustasoa 95% vastaavat Wilsonin luottamusvälit (5.21), kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti $k/n$ on merkitty pisteellä. Alemmassa kuvassa on esitetty (satunnaiseksi ymmärretyn) luottamusvälin peittotodennäköisyys $p$:n funktiona. Nimellinen luottamustaso on osoitettu vaakaviivalla.



**Kuva 5.9** Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat luottamustasoa 95% vastaavat Clopperin–Pearsonin tarkat luottamusvälit, kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti $k/n$ on merkitty pisteellä. Alemmassa kuvassa on esitetty (satunnaiseksi ymmärretyn) luottamusvälin peittotodennäköisyys $p$:n funktiona.



```
##  PointEst  Lower  Upper
##       0.2 0.0247 0.3753

binconf(k, n, method = "wilson")

##  PointEst   Lower Upper
##       0.2 0.08066 0.416

binconf(k, n, method = "exact")

##  PointEst   Lower  Upper
##       0.2 0.05733 0.4366
```

## 5.9 Ennusteväli

Luottamusvälien lisäksi (tai sijasta) usein on mielekästä tarkastella aivan toisentyyppisiä välejä, ks. esim. Vardeman [3]. Käsittelemme tässä vain ennusteväliä. Vardeman esittelee myös ns. toleranssivälin.

Tarkastelemme yksinkertaisuuden vuoksi teoreettista populaatiota, jossa satunnaismuuttujat $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$ ovat riippumattomia ja samoin jakautuneita satunnaismuuttujia pistetodennäköisyysfunktiolla tai tiheysfunktiolla $g(y;\theta)$. Välit pitää muodostaa $n$ ensimmäisen satunnaismuuttujan $Y_1, \ldots, Y_n$ arvojen avulla, ja tavalliseen tapaan,

$$\mathbf{Y} = (Y_1, \ldots, Y_n).$$

Satunnaismuuttujan $Y_{n+1}$ ajatellaan olevan tulevaisuudessa saatava havainto tästä samasta jakaumasta.

**Määritelmä 5.6** (Ennusteväli)**.** Aineistosta laskettu väli $[L(\mathbf{y}), U(\mathbf{y})]$ on tason $1-\alpha$ *ennusteväli* (engl. *prediction interval*) satunnaismuuttujalle $Y_{n+1}$, jos vastaava satunnainen väli $[L(\mathbf{Y}), U(\mathbf{Y})]$ toteuttaa vaatimuksen

$$P_\theta\left(L(\mathbf{Y}) \leq Y_{n+1} \leq U(\mathbf{Y})\right) \geq 1-\alpha, \quad \text{kaikilla } \theta \in \Theta. \tag{5.22}$$

**Esimerkki 5.3** Jos normaalijakaumaa $N(\mu, \sigma^2)$ noudattavan populaation varianssi on tunnettu luku, ja $\bar{Y}$ on $n$ ensimmäisen satunnaismuuttujan otoskeskiarvo, niin

$$Y_{n+1} - \bar{Y} \sim N(0, (1+\frac{1}{n})\sigma^2)$$

Tästä nähdään helpoilla laskuilla, että todennäköisyydellä $1-\alpha$

$$Y_{n+1} \in \bar{Y} \pm z_{\alpha/2}\sqrt{1+\frac{1}{n}}\,\sigma$$

kaikilla $\mu$, joten tätä vastaava aineistosta laskettu väli on tason $1-\alpha$ ennusteväli.

Huomaa, että uuden havainnon ennusteväli on *paljon leveämpi* kuin odotusarvon $\mu$ kaksisuuntainen luottamusväli (5.10).

Jos myös varianssiparametri olisi tuntematon, niin ennusteväliä lähdetään konstruoimaan sillä perusteella, että

$$Y_{n+1} - \bar{Y} \sim N(0, (1 + \frac{1}{n})\sigma^2)$$
$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1},$$

jossa otosvarianssi $S^2$ lasketaan satunnaismuuttujista $Y_1, \ldots, Y_n$. Yllä nämä kaksi satunnaismuuttujat ovat lisäksi riippumattomia. Tästä havainnosta saadaan yksinkertaisilla laskuilla aikaan ennusteväli uudelle havainnolle $Y_{n+1}$ käyttämällä $t$-jakauman kvantiileja. △

# Kirjallisuutta

[1] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–116, 2001.

[2] Robert G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17:857–872, 1998.

[3] Stephen B. Vardeman. What about the other intervals? *The American Statistician*, 46:193–197, 1992.