Luku 4

Luottamusvälit ja luottamusjoukot

4.1 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).

4.2 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 4.1 (Luottamusjoukko)**.** Olkoon $0 < \alpha < 1$ jokin luku. Aineistosta riippuva Θ: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

$$P_\theta\left(\tau \in A(\mathbf{Y})\right) \geq 1-\alpha, \qquad \text{kaikilla } \theta \in \Theta. \tag{4.1}$$

Huomautuksia

- Tässä (kuten tilastollisissa testeissä) α 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 τ 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 θ 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 θ, mutta tämä triviaali luottamusjoukko ei kiinnosta ketään.

- Kaikkein mieluiten konstruoisimme luottamusjoukon sillä tavalla, että kaavassa (4.1) 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.

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

Määritelmä 4.2 (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

$$P_\theta(L(\mathbf{Y}) \leq \tau \leq U(\mathbf{Y})) \geq 1 - \alpha \tag{4.2}$$

4.3 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 4.3 (Saranasuure)**.** Parametrin $\tau = k(\theta)$ ja satunnaisvektorin $\mathbf{Y}$ funktiota, jonka jakauma ei riipu parametrinarvosta, kutsutaan *saranasuureeksi* (tai napamuuttujaksi) (engl. *pivotal quantity, pivot*) parametrille τ.

Esimerkki 4.1. Jos $Y_1, \ldots, Y_n$ on satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, ja varianssiparametri σ^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 σ^2 on tunnettu luku). △

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

4.4 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_ν tällainen väli on $(-\infty, \infty)$;
- khiin neliön jakaumalle χ^2_ν tällainen väli on $(0, \infty)$.

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

$$F_X(x) = u \tag{4.3}$$

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

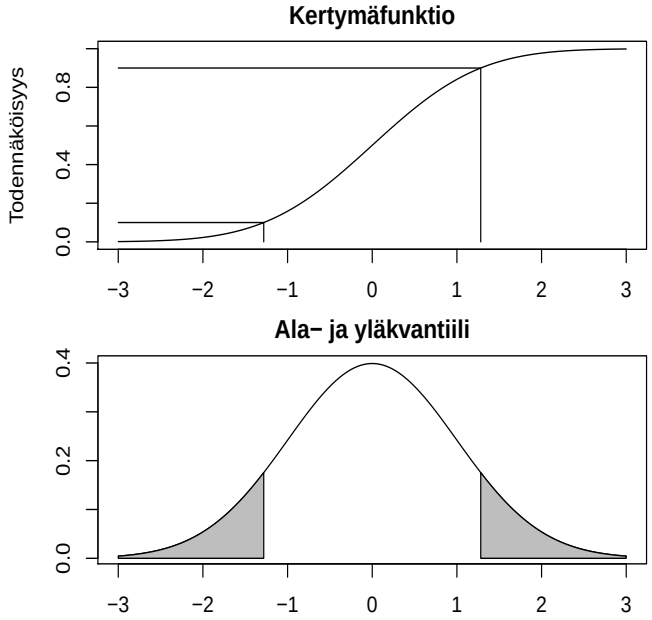
$$q(u) = x \quad \text{eli} \quad F_X(x) = u$$

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 4.4.

Kuva 4.1 Standardinormaalijakauman $N(0,1)$ kertymäfunktio sekä sen ala- ja yläkvantiilit, kun $u = 0.1$. Kummankin varjostetun häntäalueen pinta-ala on u.



Määritelmä 4.4 (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.

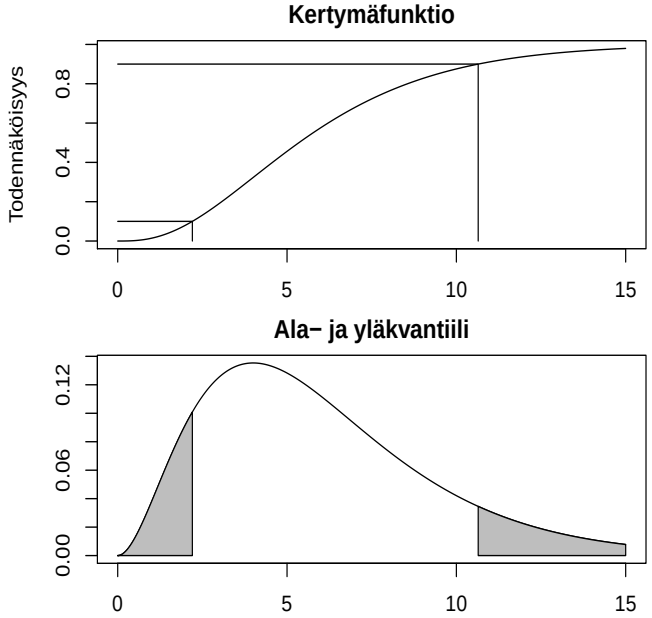
Huomautuksia:

- 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 4.4 havainnollistetaan ala- ja yläkvantiileja sekä vastaavia häntäalueita standardinormaalijakaumalle $N(0,1)$, ja kuvassa 4.4 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 4.2 Khiin neliön χ^2_ν 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] -1.281552 -1.644854 -1.959964 -2.326348 -2.575829
```

ja yläkvantiilit samoissa pisteissä komennolla

```
> qnorm(u, lower = FALSE)

[1] 1.281552 1.644854 1.959964 2.326348 2.575829
```

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

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

[1] -1.439756 -1.943180 -2.446912 -3.142668 -3.707428

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

[1] 1.439756 1.943180 2.446912 3.142668 3.707428
```

ja khiin neliön jakauman ala- ja yläkvantiilit komennoilla

```
> qchisq(u, df = nu)

[1] 2.2041307 1.6353829 1.2373442 0.8720903 0.6757268

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

[1] 10.64464 12.59159 14.44938 16.81189 18.54758
```

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

$$z_u \quad \text{on } N(0,1)\text{-jakauman } u\text{-yläkvantiili} \tag{4.4}$$
$$t_\nu(u) \quad \text{on } t_\nu\text{-jakauman } u\text{-yläkvantiili.} \tag{4.5}$$

Varoitus: Merkinnät ovat eri lähteissä erilaisia. Useissa kirjoissa z_α tarkoittaa $N(0,1)$ jakauman u-kvantiilia eikä u-yläkvantiilia. Vapausasteluvun merkintä t-jakauman yhteydessä on hyvin kirjavaa.

4.5 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 α_1 ja ylempään häntään osuus α_2. Tästä näemme, että

$$A(\mathbf{y}) = \{\tau : q(\alpha_1) \le h(\tau, \mathbf{y}) \le q(1-\alpha_2)\} \tag{4.6}$$

on parametrin τ 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 α jaetaan alemmalle ja ylemmälle saranasuureen jakauman hännälle riippuu siitä, minkälainen joukko parametrille saadaan ratkaisemalla ko. epäyhtälöt: epäyhtälöpari (4.6) 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.

4.6 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 μ kahdessa tilanteessa.

1) Kun varianssiparametri on tunnettu, jolloin mallin parametri on μ.

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

Lopuksi muodostamme vielä luottamusvälin varianssiparametrille σ^2.

4.6.1 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 4.1)

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \tag{4.7}$$

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

$$q(\alpha_1) \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq q(1 - \alpha_2), \tag{4.8}$$

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 (4.8) μ: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 μ.

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 (4.4) mukaan

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

joten

$$P_\mu\left(\bar{Y}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \quad (4.9)$$

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

$$[\bar{y}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] \quad (4.10)$$

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 (4.10) 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 μ: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 (4.10) 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 μ:n suhteen, nähdään että

$$P_\mu\left(\mu \geq \bar{Y}-z_\alpha\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \quad (4.11)$$

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

$$P_\mu\left(\mu \leq \bar{Y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{4.12}$$

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

$$[\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty) \tag{4.13}$$

$$(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}] \tag{4.14}$$

4.6.2 Aineistosta lasketun luottamusvälin tulkinta

Lasketaan nyt 95% luottamusväli (4.10) (eli kaksisuuntainen z-luottamusväli) populaation odotusarvolle μ käyttämällä kuvan 3.3 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 μ

- 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, -0.80,  0.27,  1.79,
+ 1.16)
> n <- length(y)
> z <- qnorm(0.05/2, lower = FALSE)
> sigma <- 1
> sigma/sqrt(n)

[1] 0.3162278

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

[1] 0.106205

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

```
[1] 1.345795
```

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 (4.10), 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 4.3. Jos laskemme luottamusvälin (4.10) suurelle määrälle r normaalijakaumasta $N(\mu, \sigma^2)$ simuloituja kokoa n olevia otoksia (jossa σ^2 on tunnettu)

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

niin saamme r kappaletta luottamusvälejä

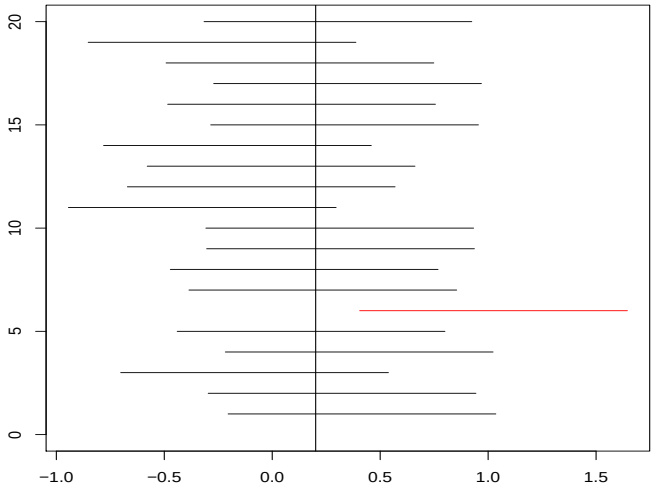
$$[L(\mathbf{y}_1), U(\mathbf{y}_1)], \ldots, [L(\mathbf{y}_r), U(\mathbf{y}_r)].$$

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).

Kuva 4.3 20 kappaletta kaavalla (4.10) 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.



4.6.3 Odotusarvon luottamusväli, kun varianssi on tuntematon

Tässä tilanteessa sekä μ että σ^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

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

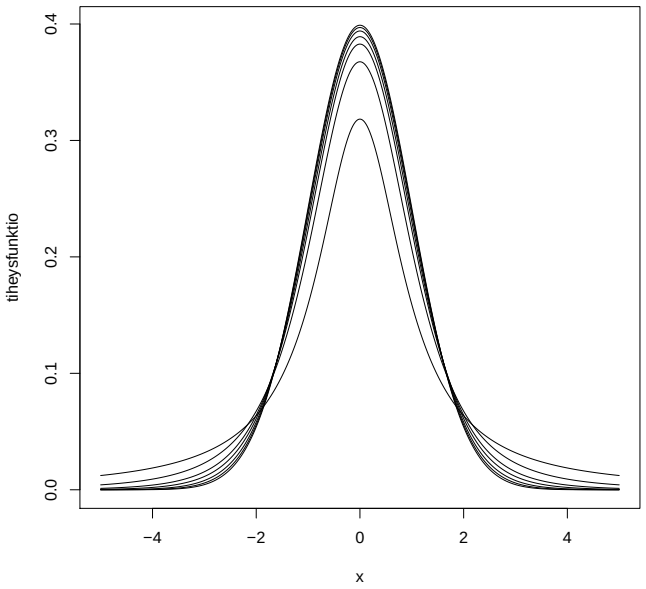
Kun varianssi on tuntematon, Z ei ole saranasuure, koska se riippuu paitsi aineistosta ja kiinnostusparametrista μ, myös haittaparametrista σ^2. Ajatuksena on kuitenkin matkia mahdollisimman tarkoin aikaisempaa konstruktiota. Koska populaation keskihajonta σ on tuntematon, sen tilalle sijoitetaan otoskeskihajontaa (3.29) vastaava satunnaismuuttuja S. Tässä mallissa satunnaismuuttuja

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \tag{4.15}$$

osoittautuu saranasuureeksi. Sen jakauma on tietty t-jakauma.

Määritelmä 4.5. Jos $\nu > 0$ ja $Z \sim N(0, 1)$ ja $X \sim \chi^2_\nu$ ja Z ja X ovat

Kuva 4.4 t_ν-jakauman tiheysfunktioita vapausasteluvun ν arvoilla 1, 3, 6, 10, 20 ja 50. Vertailun vuoksi kuvassa on myös standardinormaalijakauman $N(0,1)$ tiheysfunktio, jota voidaan pitää t jakaumana vapausasteluvulla ∞. Tiheysfunktion arvo pisteessä $x = 0$ on sitä suurempi mitä suurempi on vapausasteluku ν. Häntäalueilla järjestys on päinvastainen.



riippumattomia, niin satunnaismuuttujalla

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

on t_ν-jakauma eli t-jakauma vapausasteluvulla ν (engl. *t distribution with ν degrees of freedom*).

Määritelmän avulla on mahdollista johtaa t_ν-jakauman tiheysfunktio, mutta tätä kaavaa ei tässä yhteydessä tarvita. Tiheysfunktio osoittautuu parilliseksi funktioksi, joten t_ν-jakauma on symmetrinen. Kuvassa 4.4 esitetään t_ν-jakauman tiheysfunktio muutamilla vapausasteluvun arvoilla. Kun ν 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 3.8.2 kerrottua tietoa satunnaismuuttujaparin $(\bar{Y}, S^2)$ yhteisjakaumasta (kaavat (3.26), (3.27) ja (3.28)):

- $\bar{Y}$ 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 (4.15) satunnaismuuttuja T.

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

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

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

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

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 todennäköisyydellä $1-\alpha$ pätee

$$P_{(\mu,\sigma^2)}\left(\bar{Y} - t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}} \le \mu \le \bar{Y} + t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}}\right) = 1-\alpha, \qquad (4.16)$$

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

Vastaava aineistosta $\mathbf{y}$ laskettu väli

$$[\bar{y} - t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}] = \bar{y} \pm t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}, \qquad (4.17)$$

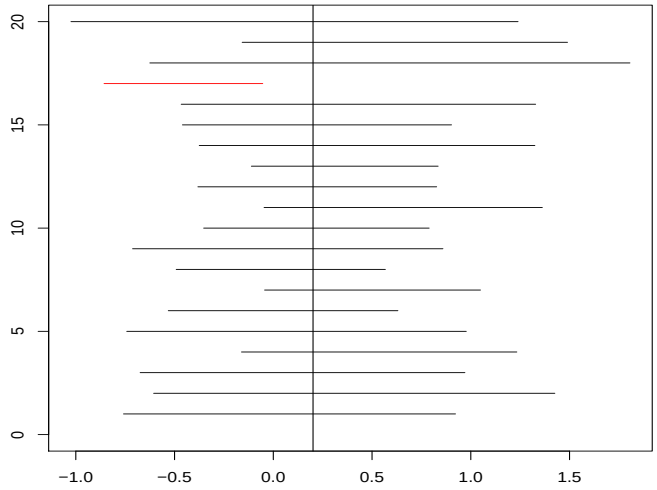
jossa $\bar{y}$ on otoskeskiarvo ja s on otoskeskihajonta, on normaalijakauman odotusarvon μ luottamusväli luottamustasolla $1-\alpha$. Sitä kutsutaan usein t-luottamusväliksi (viitejakauman t_{n-1} mukaan). Huomaa, että $\bar{y}$ on myös parametrin μ SU-estimaatti ja että $s/\sqrt{n}$ on tämän estimaatin 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] 2.009575 1.984217 1.971957 1.964729 1.962341
```

Kuva 4.5 20 kappaletta kaavalla (4.17) 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.



Väli (4.17) 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).

Kuvassa 4.5 näytetään 20 kappaletta t-luottamusvälejä, jotka on laskettu aineistoista, jotka on generoitu tietystä normaalijakaumasta.

Kuten jaksossa 4.6.2 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 4.2. Kuvan 3.3 aineistolle

$$\bar{y} = 0.726, \quad s = 1.074, \quad n = 10, \quad t_{0.025}(9) = 2.262.$$

Parametrin μ 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, -0.80,  0.27,  1.79,  1.16)
> t.test(y)

        One Sample t-test

data:  y
t = 2.1372, df = 9, p-value = 0.0613
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.04243969  1.49443969
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`. 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] 0.3396933
```

Mikään ei pakota meitä laskemaan luottamusväliä vain yhdellä luottamustasolla 0.95. Kuvassa 4.6 näytetään luottamusvälin päätepisteet luottamustason funktiona. △

4.6.4 Varianssiparametrin luottamusväli

Oletamme, että sekä μ että σ^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

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

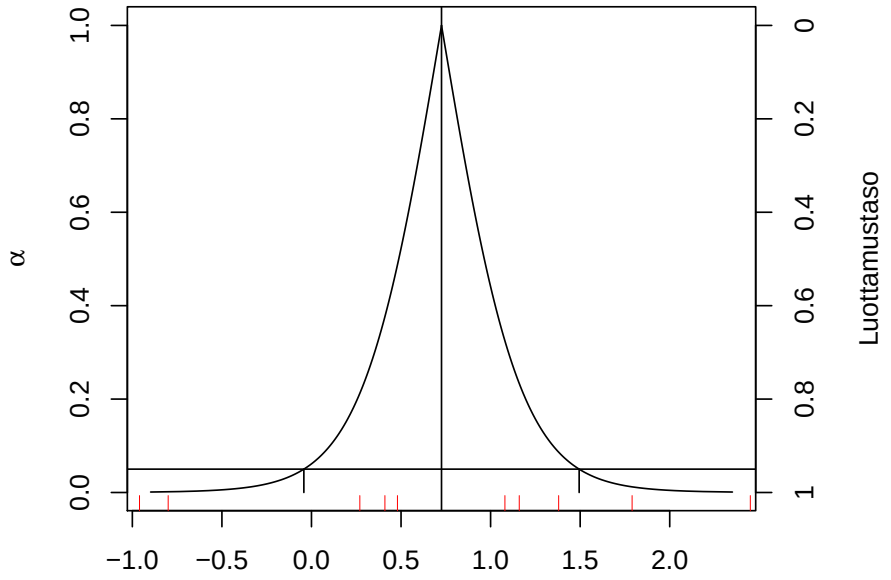
Jos q on χ^2_{n-1}-jakauman kvantiilifunktio, ja $0 < \alpha < 1$ sekä $\alpha_1 > 0$ ja $\alpha_2 > 0$ ovat lukuja siten, että $\alpha = \alpha_1 + \alpha_2$, niin todennäköisyydellä $1 - \alpha$ pätee

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

Kun tämä epäyhtälö ratkaistaan muuttujan σ^2 suhteen, saadaan väli

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

Kuva 4.6 Kuvan 3.3 aineistolle lasketut parametrin μ kaksisuuntaiset t-luottamusvälit. Piste-estimaatti sekä 95%:n luottamusväli on korostettu pystyviivoilla. Aineisto on esitetty x-akselin yläpuolella olevilla pienillä viivoilla.



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

$$[\frac{n-1}{q(1-\alpha/2)}\, s^2, \frac{n-1}{q(\alpha/2)}\, s^2], \tag{4.18}$$

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

Kuvan 3.3 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 $\sigma^2 = 1$.

4.7 Likimääräinen luottamusväli

Jos otoskoko n on suuri ja jos piste-estimaattorin $\hat{\tau}(\mathbf{Y})$ otantajakauma on osapuilleen τ-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 τ suhteen, saadaan väli

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

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

$$\hat{\tau} \pm z_{\alpha/2} \times \mathrm{se}, \tag{4.19}$$

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

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\,\hat{p}\,(1-\hat{p})}, \tag{4.20}$$

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 (4.20) ominaisuuksia. Se voi äärellisellä otoskoolla käyttäytyä kummallisella tavalla:

- Sen päätepisteet voivat olla parametriavaruuden ulkopuolella; käytännössä luottamusväliksi pitäisi otta välin (4.20) 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 4.7 otoskoko on $n = 20$. Siinä esitetään eri onnistumisten lukumäärää $0 \leq k \leq n$ vastaavat $n + 1$ mahdollista luottamusväliä laskettuna kaavalla (4.20). 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.

4.8 Muita luottamusvälejä binomikokeessa

Likimääräisen luottamusvälin (4.20) 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,

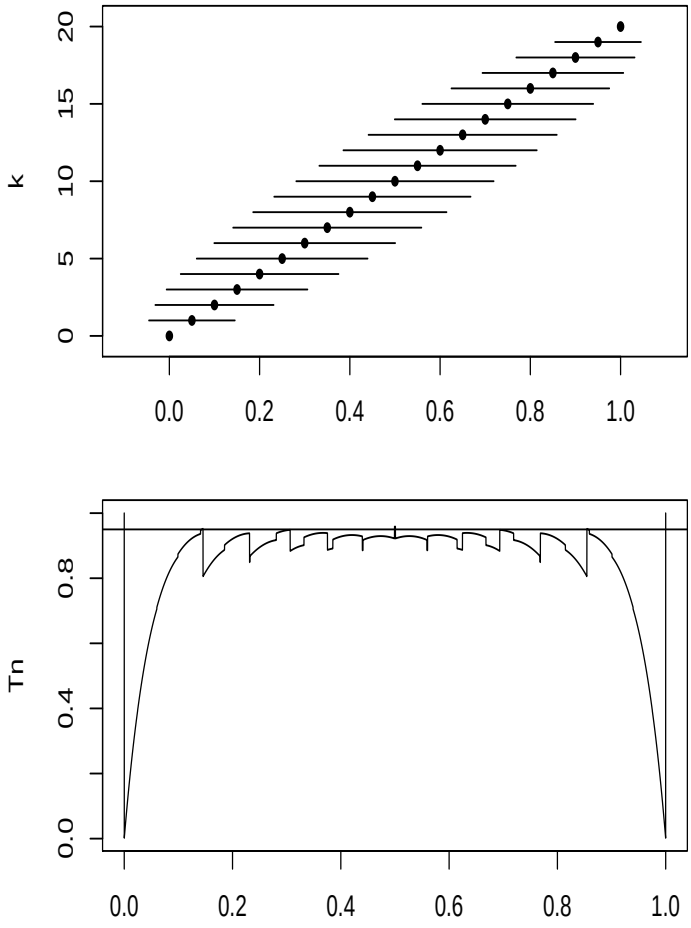
> ... it is strongly recommended that intervals calculated by these methods should no longer be acceptable for the scientific literature

Nämä neuvot on syytä ottaa huomioon. Älkää käyttäkö perinteistä likimääräistä luottamusväliä (4.20) 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 4.7 Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat nimellistä luottamustasoa 95% vastaavat luottamusvälit (4.20), 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

$$\frac{\hat{p}+\frac{1}{2n}z_{\alpha/2}^2 \pm z_{\alpha/2}\sqrt{\frac{1}{n}\,\hat{p}(1-\hat{p})+\frac{1}{4n^2}\,z_{\alpha/2}^2}}{1+\frac{1}{n}z_{\alpha/2}^2}, \tag{4.21}$$

joka on luottamusväliä (4.20) 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 (4.2) ei sille toteudu. Kuvassa 4.8 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 (4.2)) 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 4.9 nähdään, kun otoskoko $n = 20$.

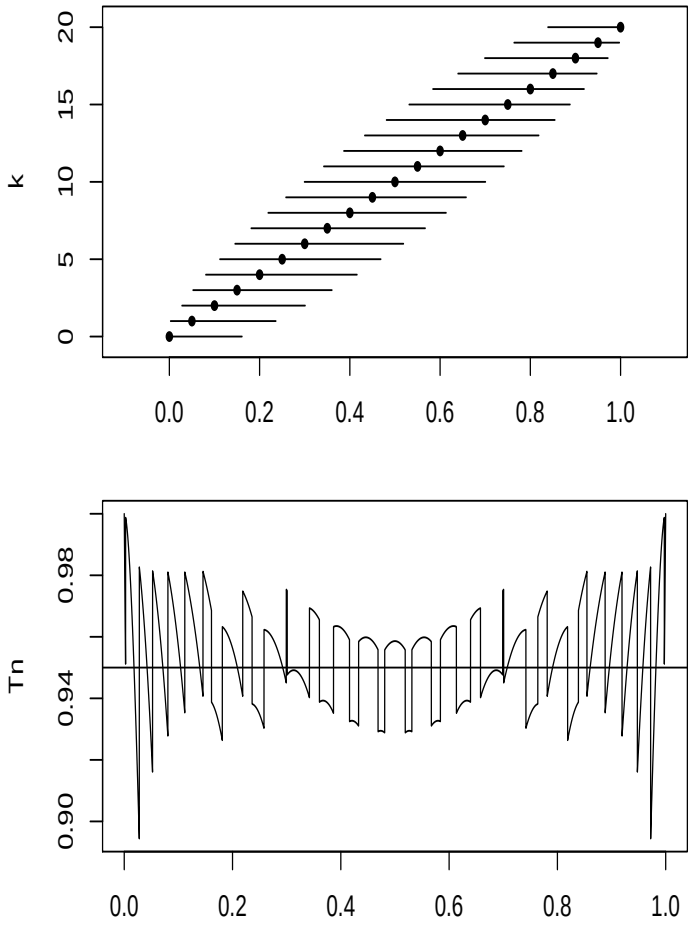
Silloin kuin havaintosatunnaisvektorin jakauma on diskreetti, niin yleensä aina joudutaan tekemään luottamusvälien kanssa samantapaisia kompromisseja. Joko käytetään likimääräisiä luottamusvälejä, joiden todellinen peittotodennäköisyys on joskus pienempi kuin niiden nimellinen peittotodennäköisyys, tai 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.)

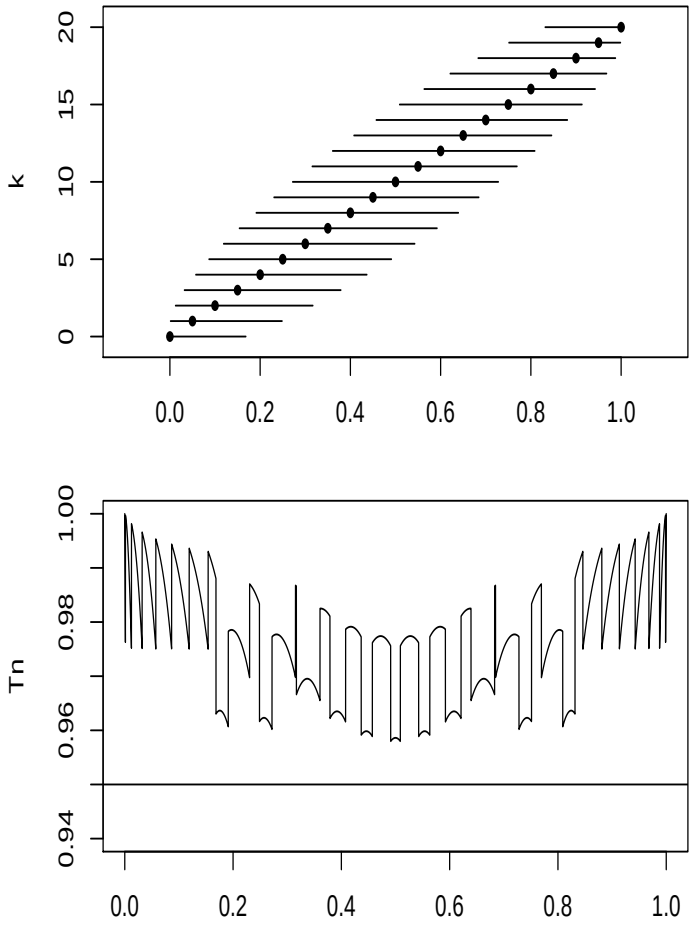
```
> n <- 20
> k <- 4
> library(Hmisc)
> binconf(k, n, method = 'asymptotic')

 PointEst      Lower     Upper
      0.2 0.02469549 0.3753045
```

Kuva 4.8 Ylemmässä kuvassa on esitetty otoskokoa $n = 20$ vastaavat nimellistä luottamustasoa 95% vastaavat Wilsonin luottamusvälit (4.21), 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 4.9 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.



```
> binconf(k, n, method = 'wilson')

 PointEst      Lower     Upper
      0.2 0.08065766 0.4160174

> binconf(k, n, method = 'exact')

 PointEst    Lower    Upper
      0.2 0.057334 0.436614
```

4.9 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 4.6 (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

$$P_\theta\left(L(\mathbf{Y}) \le Y_{n+1} \le U(\mathbf{Y})\right) \ge 1-\alpha, \quad \text{kaikilla } \theta \in \Theta. \tag{4.22}$$

Esimerkki 4.3. 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 μ, joten tätä vastaava aineistosta laskettu väli on tason $1-\alpha$ ennusteväli.

Huomaa, että uuden havainnon ennusteväli on *paljon leveämpi* kuin odotusarvon μ kaksisuuntainen luottamusväli (4.10).

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.