5 Luottamusvälit ja luottamusjoukot

- 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 joskus tarkastellaan aivan muunlaisia välejä, esim. ennustevälejä.

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 (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

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

Luottamusväli

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

Määritelmä (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

$$P_\theta(L(\mathbf{Y}) \leq \tau \leq U(\mathbf{Y})) \geq 1 - \alpha \tag{2}$$

Huomautuksia luottamusjoukon määritelmästä

- Tässä 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.)

Huomautuksia luottamusjoukon määritelmästä (jatkoa)

- Tahtoisimme luottamusjoukon olevan jollakin tavalla pieni. Koko parametriavaruus $A(\mathbf{y}) = \Theta$ on minkä tahansa tason $1 - \alpha$ luottamusjoukko mallin parametrille θ, mutta tämä triviaali luottamusjoukko ei kiinnosta ketään.
- Kaikkein mieluiten konstruoisimme luottamusjoukon sillä tavalla, että 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 (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 saranasuureesta

Esimerkki 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 $\mathbf{Y}$:n arvo, mutta ei parametrinarvoa $\theta = \mu$ (tässä σ^2 on tunnettu luku). $\triangle$

Entä, jos varianssi on tuntematon?

- 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.
- Näemme t-jakauman määritelmän myöhemmin.

5.4 Ala- ja yläkvantiilit

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

Oletuksia kvantiilifunktion määrittelyä varten

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 $(0, \infty)$.

Kvantiilin ja kvantiilifunktion määritelmä

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

$$F_X(x) = u \tag{3}$$

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 $0 < u < 1$.

Kvantiilien merkitys

Kvantiilifunktio q saa arvon x, ts.

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

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.

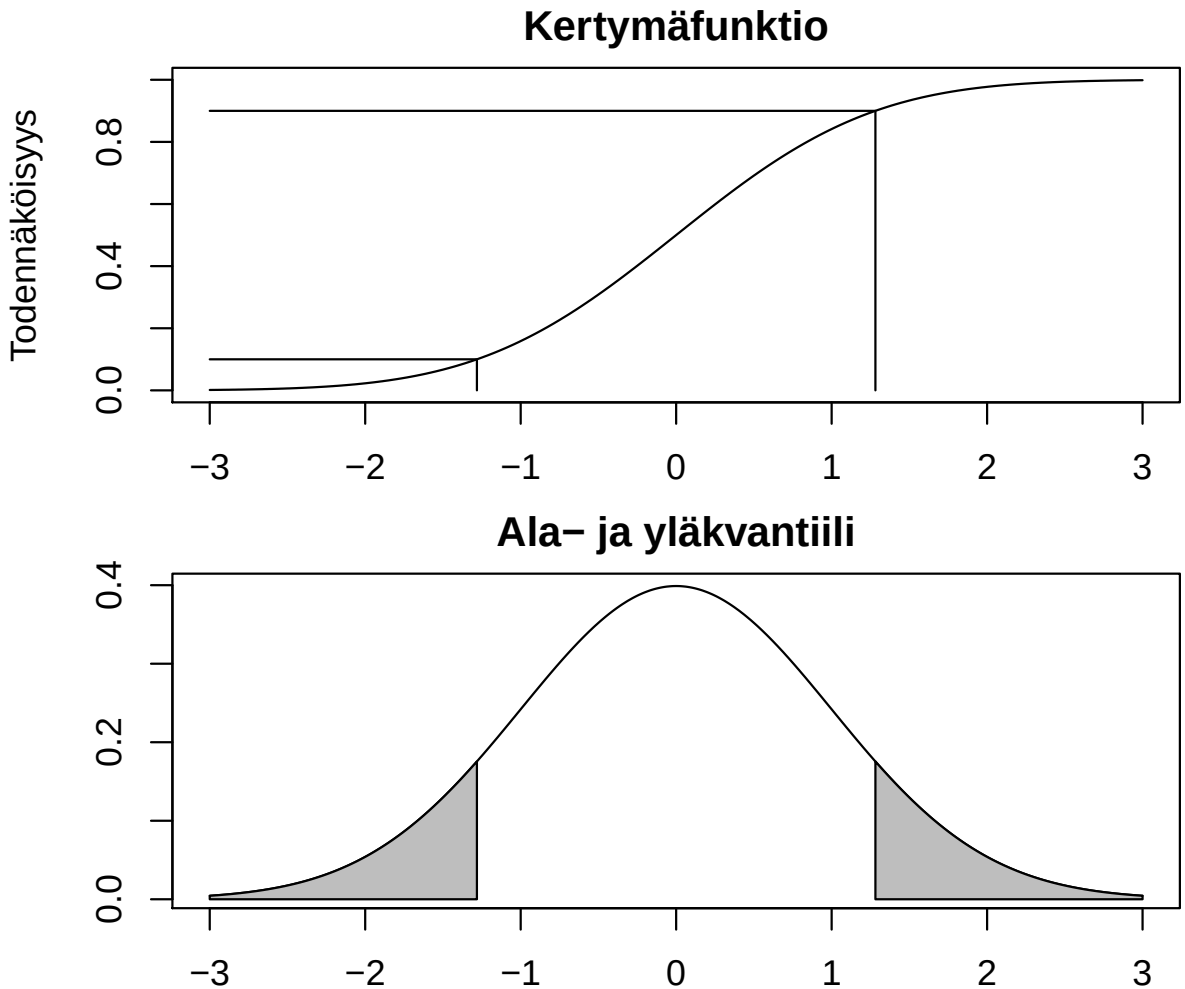
Ala- ja yläkvantiilit

Määritelmä (Ala- ja yläkvantiilit)

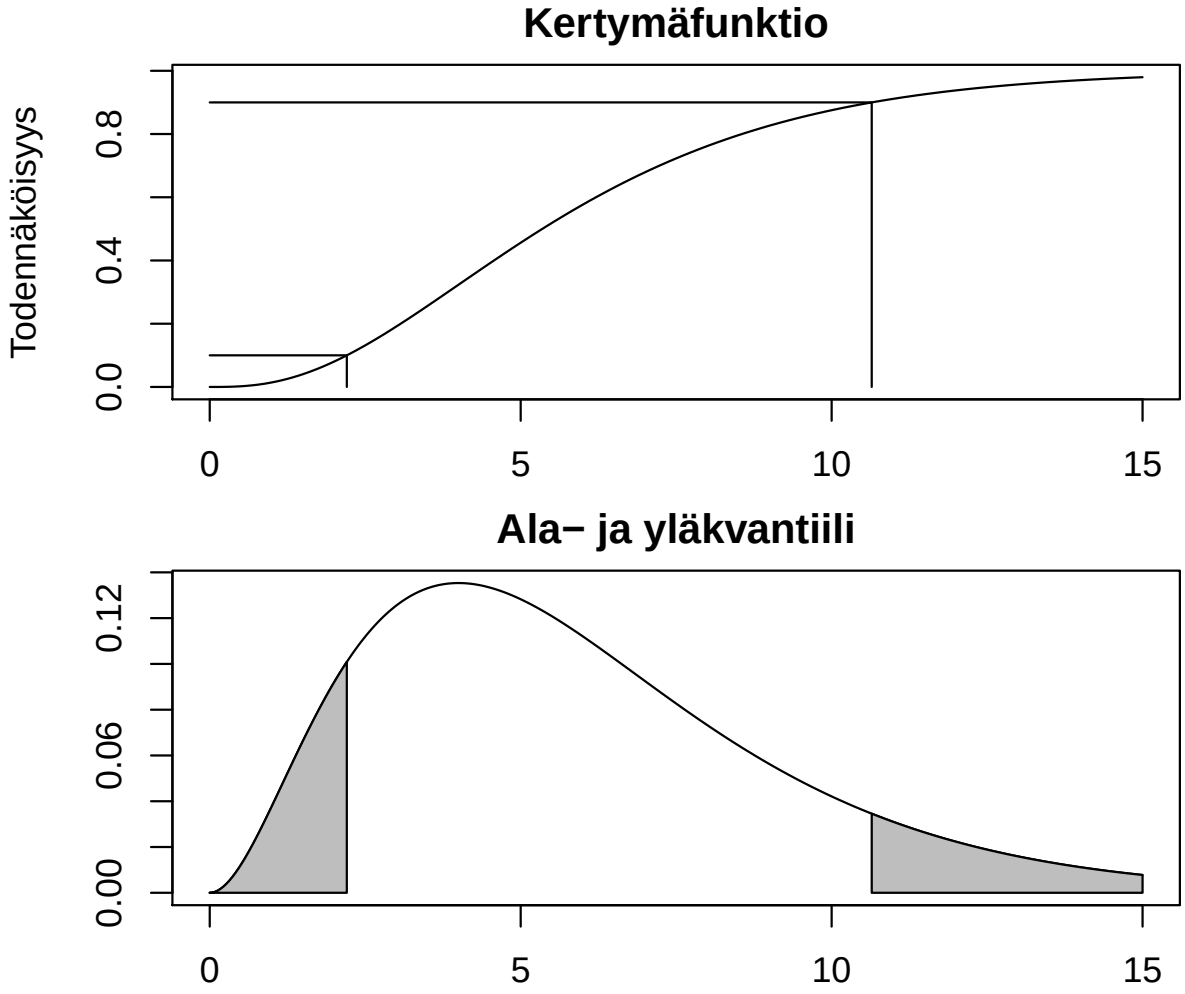
Sellaista pistettä, josta oikealle jää satunnaismuuttujan todennäköisyysmassasta osuus $0 < u < 1$ kutsutaan ko. jakauman *u*-yläkvantiiliksi (engl. *upper u quantile*).
Sellaista pistettä, josta vasemmalle jää satunnaismuuttujan todennäköisyysmassasta osuus $0 < u < 1$ kutsutaan ko. jakauman *u*-kvantiiliksi tai *u*-alakvantiiliksi.

- 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.

$N(0,1)$-jakauman kertymäfunktio ja sen ylä- ja alakvantiilit, kun $u = 0.1$



Khiin neliön χ^2_ν kertymäfunktio sekä sen ala- ja yläkvantiilit, kun $u = 0.1$ ja vapausasteluku $\nu = 6$



Tilastolliset taulukot ovat nykyään tarpeettomia

- 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.

$N(0,1)$-jakauman kvantiilit R-ohjelmistolla

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
```

t-jakauman kvantiilit R-ohjelmistolla

Vastaavasti t-jakauman ala- ja yläkvantiilit saadaan laskettua (annetulla ν: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
```

Khiin neliön jakauman kvantiilit R-ohjelmistolla

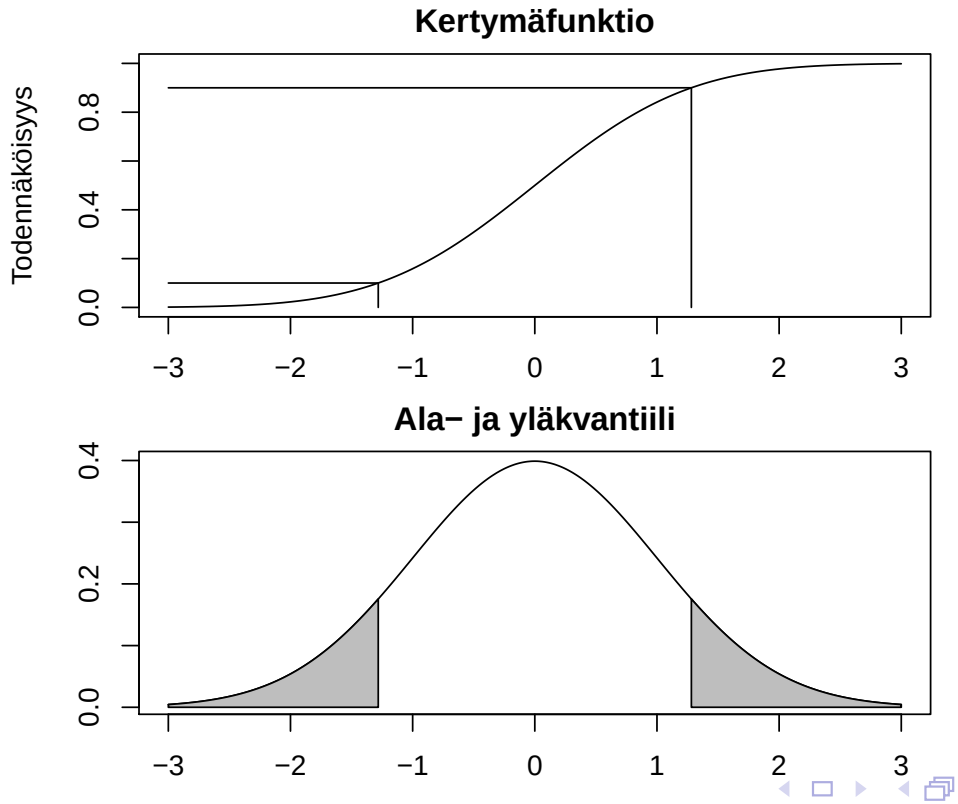
Khiin neliön jakauman ala- ja yläkvantiilit saadaan selville komennoilla

```
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
```

Kvantiilit symmertriselle jakaumalle

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

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



Merkintöjä $N(0,1)$- ja t-jakaumille

Symmetrisille jakaumille ei tarvita kuin toista jakauman häntää vastaavat kvantiilit. Näille käytetään usein lyhyitä merkintöjä. Tässä monisteessa

$$z_u \quad \text{on } N(0,1)\text{-jakauman } u\text{-yläkvantiili} \tag{4}$$

$$t_\nu(u) \quad \text{on } t_\nu\text{-jakauman } u\text{-yläkvantiili.} \tag{5}$$

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

Luottamusjoukon konstruointi

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

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

sillä alempaan jakauman häntään jää saranasuureen jakauman todennäköisyysmassasta osuus α_1 ja ylempään häntään osuus α_2. Tästä näemme, että

$$A(\mathbf{y}) = \{\tau : q(\alpha_1) \leq h(\tau, \mathbf{y}) \leq q(1 - \alpha_2)\} \tag{6}$$

on parametrin τ luottamusjoukko (luottamus-)tasolla $1 - \alpha$.

Lisää konstruktioita

Rajankäynnillä ($\alpha_1 \to 0$ tai $\alpha_2 \to 0$) saadaan vielä seuraavat luottamusjoukot

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

Se miten virhetodennäköisyys α jaetaan alemmalle ja ylemmälle saranasuureen jakauman hännälle riippuu siitä, minkälainen joukko parametrille saadaan ratkaisemalla ko. epäyhtälöt.
Yleisin valinta on

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

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

Onko konstruktio järkevä?

- 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 $N(\mu, \sigma^2)$.
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.

Odotusarvon luottamusväli, kun varianssi on tunnettu

Tässä tapauksessa luottamusvälin muodostaminen on helpointa ymmärtää. Valitettavasti tätä tapausta ei käytännössä tarvita juuri koskaan.
Käytämme saranasuuretta (vrt. esimerkki 1)

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \tag{7}$$

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

$$q(\alpha_1) \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq q(1 - \alpha_2), \tag{8}$$

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

Epäyhtälöparin ratkaiseminen

Merkitään väliaikaisesti

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

ja ratkaistaan kaksoisepäyhtälö (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.

Loppusilaus

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) mukaan

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

joten

$$P_\mu\left(\bar{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{9}$$

z-luottamusväli

Tämä on parametrin μ luottamustason $1-\alpha$ luottamusväli, kun normaalijakaumaa noudattavan populaation varianssi σ^2 on tunnettu luku.

$$[\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] \tag{10}$$

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.

Huomioita z-luottamusvälistä

- Luottamusväli (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.
- Tämä luottamusväli on kaksisuuntainen (eli kaksitahoinen) (engl. *two-sided*).
- Luottamusvälin leveys riippuu luotamustasosta ja otoskoosta. Otoskoon nelinkertaistaminen puolittaa tämän luottamusvälin leveyden.

Yksisuuntaisten z-luottamusvälien johtaminen

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

$$P_\mu\left(\mu \geq \bar{Y} - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{11}$$

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

$$P_\mu\left(\mu \leq \bar{Y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha, \qquad \text{kaikilla } \mu \in \mathbb{R}. \tag{12}$$

Yksisuuntaiset välit

Seuraavat aineistosta **y** lasketut yksisuuntaiset välit ovat luottamustason $1 - \alpha$ luottamusvälejä normaalijakauman odotusarvoparametrille, kun sen varianssi on tunnettu.

$$[\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty) \tag{13}$$

$$(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}] \tag{14}$$

5.6.2 Aineistosta lasketun luottamusvälin tulkinta

Lasketaan nyt 95% luottamusväli (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.

Kysymys

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 $[0.10, 1.35]$.

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

Vastaus:

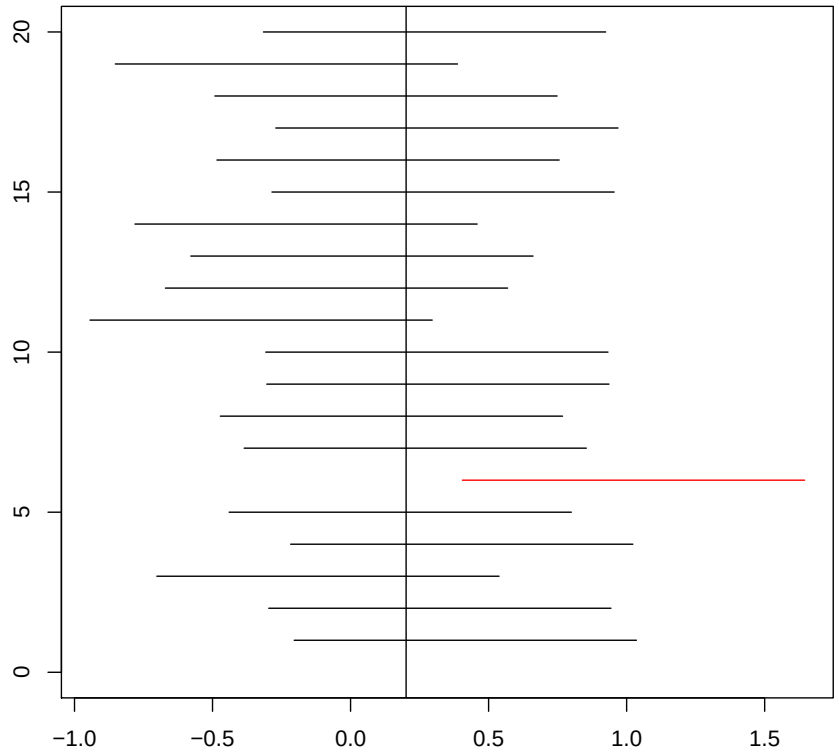
- Aineistosta laskettu luottamusväli joko sisältää todellisen parametrinarvon tai tai ei sisällä sitä. Emme voi pelkästään aineistoa tarkastelemalla sanoa mitään sen enempää, vaan tätä varten pitäisi tuntea todellinen parametrinarvo.
- Frekventistisessä tilastotieteessä parametri on tuntematon, mutta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

Onko z-luottamusvälissä jotakin vikaa?

- Tulkinnallinen vaikeus ei liity kaavaan (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 $[L(\mathbf{Y}), U(\mathbf{Y})]$.

Luottamusvälejä toistetusta otannasta

20 kappaletta kaavalla (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.



Luottamusvälit toistetusta ainestonkeruusta

Jos laskemme luottamusvälin (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.

Lasketun luottamusvälin tulkinta

En saa sanoa, että $\mu \in [0.10, 1.35]$ todennäköisyydellä 0.95. Miten saan tulkita aineistosta lasketun luottamusvälin?

- Aineiston perusteella paras arvauksemme parametrinarvolle on piste-estimaatti 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

- Oletus: satunnaisotos $Y_1, \ldots, Y_n$ normaalijakaumasta $N(\mu, \sigma^2)$. Nyt sekä μ että σ^2 ovat tuntemattomia, joten mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- Haluamme muodostaa luottamusvälin odotusarvoparametrille

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

- Kun varianssi oli tunnettu, käytimme saranasuuretta

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

- Pulma: kun varianssi on tuntematon, Z ei ole saranasuure, koska se riippuu paitsi aineistosta ja kiinnostusparametrista μ, myös haittaparametrista σ^2.

Pulman ratkaisu

- Ajatuksena on matkia mahdollisimman tarkoin aikaisempaa konstruktiota.
- Koska populaation keskihajonta σ on tuntematon, sen tilalle sijoitetaan otoskeskihajontaa vastaava satunnaismuuttuja S.
- Tässä mallissa satunnaismuuttuja

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \tag{15}$$

osoittautuu saranasuureeksi. Sen jakauma on tietty t-jakauma.

t-jakauma

Määritelmä

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

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

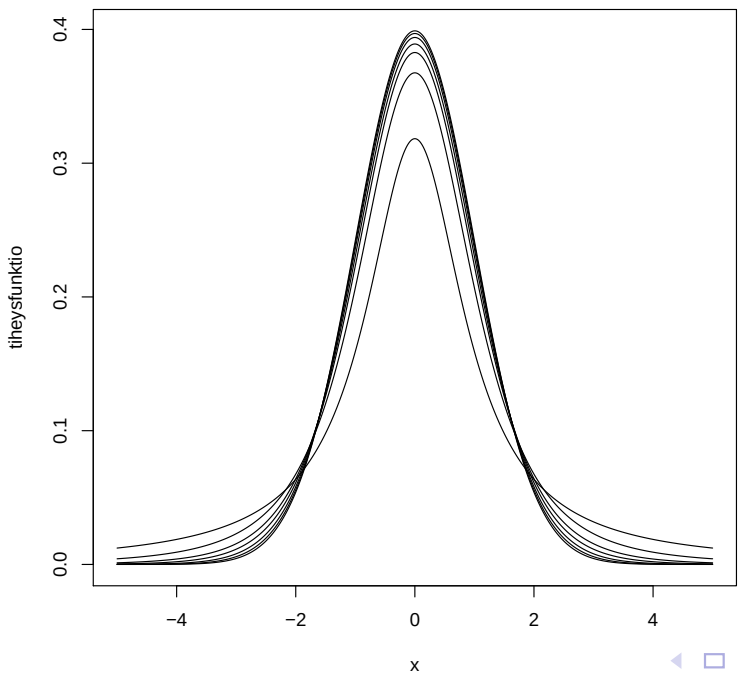
on t_ν-jakauma eli t-jakauma vapausasteluvulla ν (engl. *t distribution with ν degrees of freedom*).

t-jakauman ominaisuuksia

- 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.
- Kun ν kasvaa, jakauman tiheysfunktio lähestyy standardinormaalijakauman $N(0,1)$ tiheysfunktiota.
- t-jakaumaa kutsutaan myös Studentin t-jakaumaksi W. S. Gossetin v. 1908 käyttämän salanimen mukaan.

t-jakauman tiheysfunktioita

t_ν-jakauman tiheysfunktioita vapausasteluvun ν arvoilla 1, 3, 6, 10, 20 ja 50. 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.



Tunnusluvun T jakauma

Palautetaan mieleen satunnaismuuttujaparin $(\bar{Y}, S^2)$ yhteisjakauma:

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

Tämän sekä 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 (15) satunnaismuuttuja T.

t-luottamusvälin johto

- 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 α.
- Tästä saadaan ratkaistua väli odotusarvolle μ aivan samoilla vaiheilla kuin aikaisemmin, ja tulos on

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

Loppusilaus t-luottamusvälin johdolle

- Valitaan $\alpha_1 = \alpha_2 = \alpha/2$, ja huomataan, että

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

- Nyt pätee

$$P_{(\mu,\sigma^2)}\left(\bar{Y} - t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{n-1}(\alpha/2)\,\frac{S}{\sqrt{n}}\right) = 1-\alpha, \tag{16}$$

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

t-luottamusväli

- Aineistosta **y** laskettu väli

$$[\bar{y} - t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}] = \bar{y} \pm t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}, \tag{17}$$

  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.

Huomioita kaavasta $\bar{y} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}$

- 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

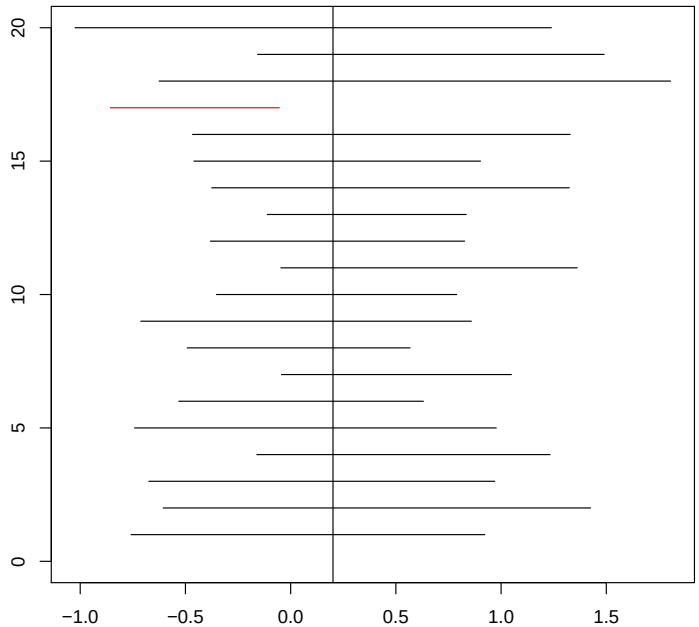
```
n <- c(50, 100, 200, 500, 1000)
qt(0.05/2, df = n - 1, lower = FALSE)
```

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

- Väli on symmetrinen piste-estimaatin $\bar{y}$ suhteen.
- t-luottamusvälin leveys vaihtelee aineistosta toiseen, koska välin leveys määräytyy aineiston otoskeskihajonnasta.
- t-luottamusvälin leveys riippuu luottamustasosta ja otoskoosta. Otoskoon nelinkertaistaminen karkeasti ottaen puolittaa kaksisuuntaisen t-luottamusvälin leveyden.

t-luottamusvälejä toistetusta otannasta

20 kappaletta kaavalla (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.



t-luottamusväli kuvan 3.3 aineistolle

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

- Parametrin μ piste-estimaatti on 0.73,
- sen keskivirhe on 0.34 (kaavalla $s/\sqrt{n}$),
- 95%:n luottamusväli on $[-0.04, 1.50]$ (kaavalla $\bar{y} \pm t_{n-1}(\alpha/2)\,\frac{s}{\sqrt{n}}$)
- Kaikkien luottamusvälin sisällä olevien arvojen voidaan ajatella olevan kohtuullisen hyvin sopusoinnussa aineiston kanssa. Paras arvauksemme on parametrin piste-estimaatti.

t-luottamusväli R:llä

```
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
```

Kommentteja: R-funktio `t.test()`

- Valitettavasti tulostuksessa näkyy meille tarpeetonta tietoa; pelkän välin saisi selville antamalla komennon

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

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

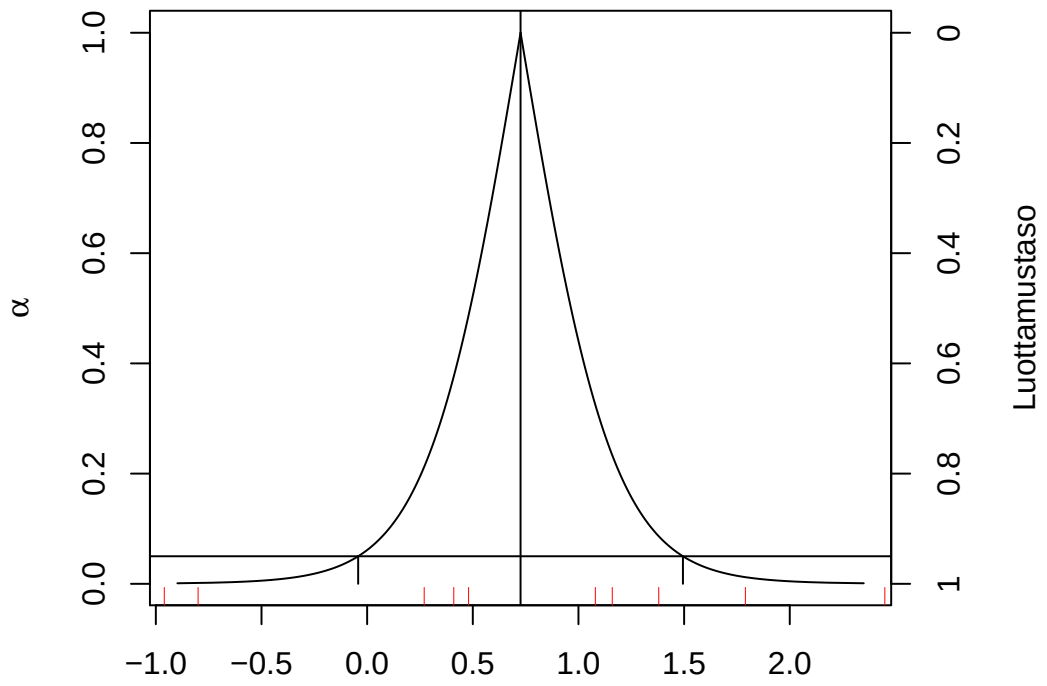
- Jos tahdotaan käyttää muita luottamustasoja kuin 95%, kuten esim. 99%, niin haluttu luottamustaso pitää antaa `t.test`-funktiolle: `t.test(y, conf.level = 0.99)`.
- `t.test` ei raportoi piste-estimaatin keskivirhettä, mutta sen saa laskettua helposti:

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

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

Luottamusvälin päätepisteet luottamustason funktiona

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.



5.6.4 Varianssiparametrin luottamusväli

- Oletus: satunnaisotos $Y_1, \ldots, Y_n$ normaalijakaumasta $N(\mu, \sigma^2)$. Sekä μ että σ^2 ovat tuntemattomia, joten mallin parametrivektori on $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- Haluamme muodostaa luottamusvälin varianssiparametrille

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

- Käytämme saranasuureena sopivasti skaalattua otosvarianssia, sillä tiedämme, että

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

Luottamusvälin johto

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

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

Loppusilaus luottamusvälin johdolle

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

$$[\frac{n-1}{q(1-\alpha/2)}\, s^2, \frac{n-1}{q(\alpha/2)}\, s^2], \tag{18}$$

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.

Luottamusväli aineistosta

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 $\sigma^2 = 1$.

5.7 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

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

Kun tämä epäyhtälöpari ratkaistaan parametrin τ suhteen, saadaan väli

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

Tässä estimaattorin otantajakauman keskihajonta $\sqrt{\operatorname{var}_\theta \hat{\tau}(\mathbf{Y})}$ on tavallisesti tuntematon.

Likimääräinen luottamusväli

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

$$\hat{\tau} \pm z_{\alpha/2} \times \mathrm{se}. \tag{19}$$

- 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.

Likimäärinen luottamusväli binomikokeessa

- SU-estimaattori
$$\hat{p}(\mathbf{Y}) = \bar{Y}$$
(onnistumisten suhteellinen osuus) on harhaton, ja sen varianssi on
$$\operatorname{var}_p \bar{Y} = \frac{1}{n} p(1-p).$$
- Kun keskivirheelle käytetään kaavaa
$$\operatorname{se} = \sqrt{\frac{1}{n} \hat{p}\,(1-\hat{p})},$$
saadaan binomikokeen onnistumistodennäköisyydelle p likimääräinen $1-\alpha$ luottamusväli
$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{1}{n} \hat{p}\,(1-\hat{p})}, \tag{20}$$
joka kerrotaan kaikissa tilastotieteen alkeisoppikirjoissa.

Likimääräisen luottamusvälin oletukset ovat kunnossa

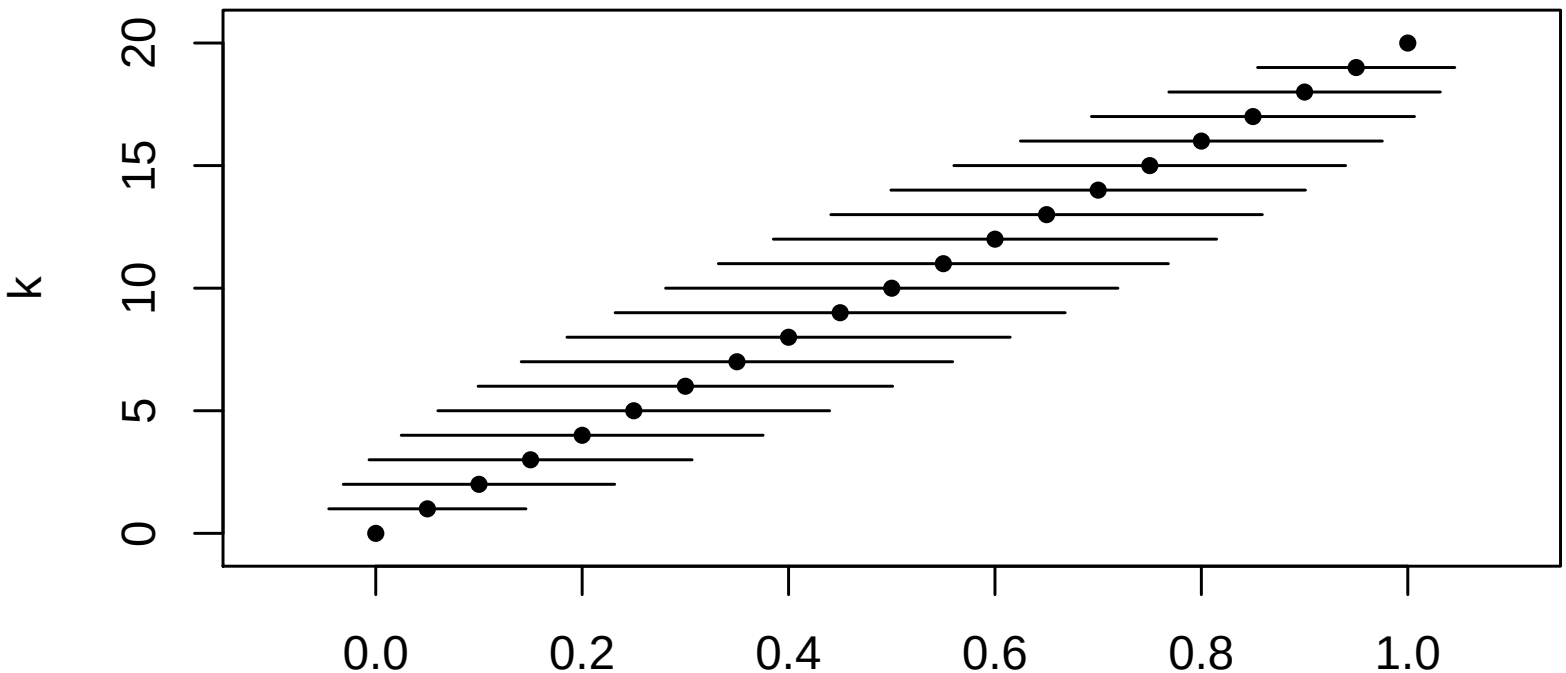
- Koska estimaattori $\bar{Y}$ on keskiarvo n riippumattomasta ja samoin jakautuneesta satunnaismuuttujasta, sen jakaumaa voidaan suurella otoskoolla approksimoida normaalijakaumalla, todennäköisyyslaskennan keskeisen raja-arvolauseen nojalla.
- Jos $0 < p < 1$ on kiinteä, ja otoskoko n kasvaa rajatta, niin voidaan osoittaa, että vastaavan satunnaisen luottamusvälin peittotodennäköisyys lähestyy arvoa $1 - \alpha$, joten suurella otoskoolla tämän välin peittotodennäköisyys on suunnilleen $1 - \alpha$.
- Milloin otoskoko on niin suuri, että approksimaatio on hyvä?

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

- Asymptoottinen perustelu jättää auki sen kysymyksen, milloin otoskoko on riittävän suuri. Tarkastelemme tämän perinteisen luottamusvälin ominaisuuksia kiinteällä n.
- Välin päätepisteet voivat olla parametriavaruuden ulkopuolella; käytännössä luottamusväliksi pitäisi ottaa tämän välin 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.

Luottamusvälit $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{1}{n} \hat{p}\,(1-\hat{p})}$, kun $n = 20$

Nimellistä luottamustasoa 95% vastaavat likimääräiset luottamusvälit, kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti k/n on merkitty pisteellä.

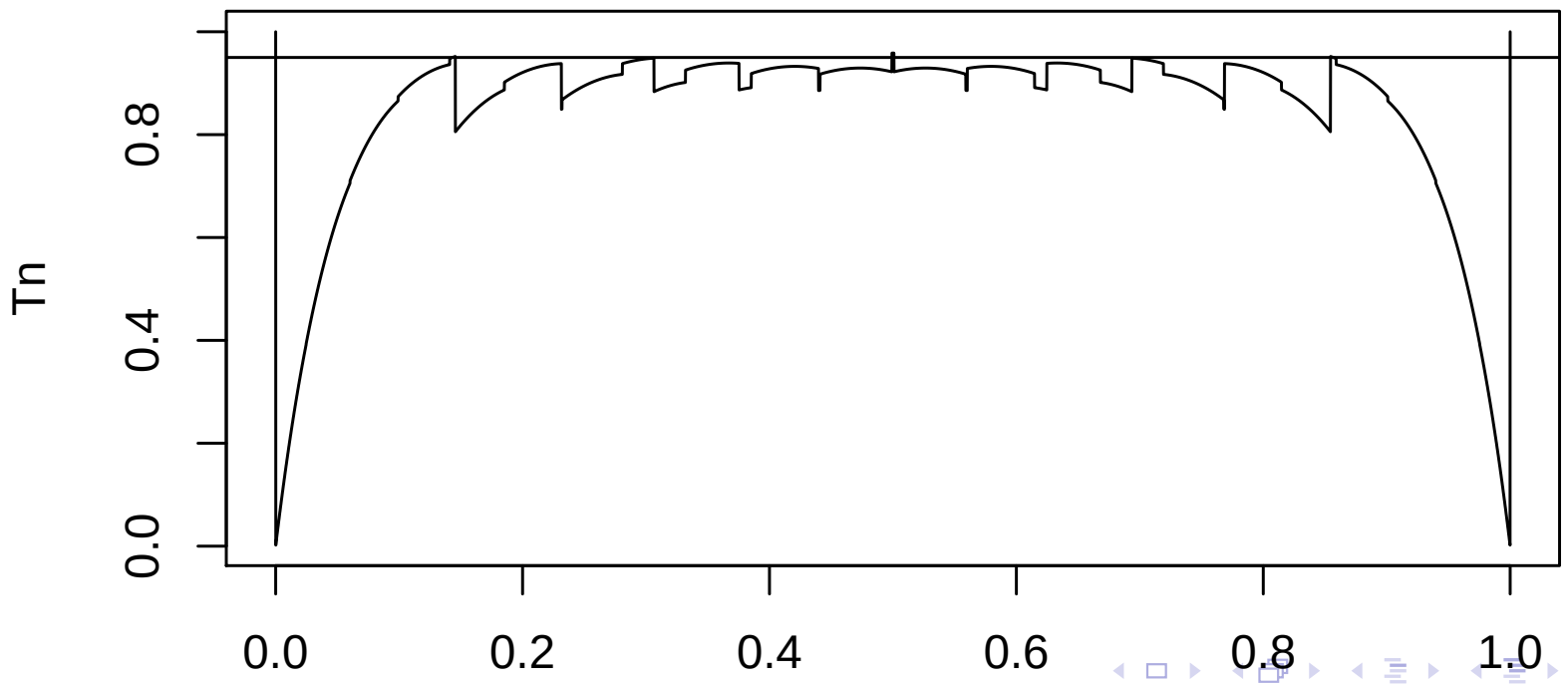


Likimääräisen luottamusvälin peittotodennäköisyys

Kuvassa on likimääräisen luottamusvälin $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{1}{n} \hat{p}\,(1-\hat{p})}$ todellinen peittotodennäköisyys

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

kun $n = 20$ ja $\alpha = 0.05$.



5.8 Muita luottamusvälejä binomikokeessa

- Likimääräisen luottamusvälin (20) $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{1}{n} \hat{p}\,(1-\hat{p})}$ 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.
- Lisäksi todellinen peittotodennäköisyys voi olla selvästi nimellistä tasoa pienempi muuallakin vielä suurehkolla otoskoolla, ks. artikkelia Brown, Cai ja DasGupta 2001.

Johtopäätös perinteisestä likimääräisestä välistä

Brown, Cai ja DasGupta toteavat likimääräisestä luottamusvälistä $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{1}{n} \hat{p}\,(1 - \hat{p})}$ 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.*

Tutkimustuloksia perinteisestä likimääräisestä välistä

- Newcombe (1998) 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 tätä perinteistä likimääräistä luottamusväliä omissa töissänne.
- Mainituissa artikkeleissa käydään läpi monta vaihtoehtoehtoista tapaa muodostaa luottamusväli onnistumistodennäköisyydelle.

Wilsonin luottamusväli

Wilsonin v. 1927 ehdottama luottamusväli on perinteistä selvästi parempi. Myös se perustuu siihen approksimaatioon, että suureella

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

on osapuilleen standardinormaalijakauma $N(0,1)$, mutta 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.

Wilsonin likimääräinen luottamusväli

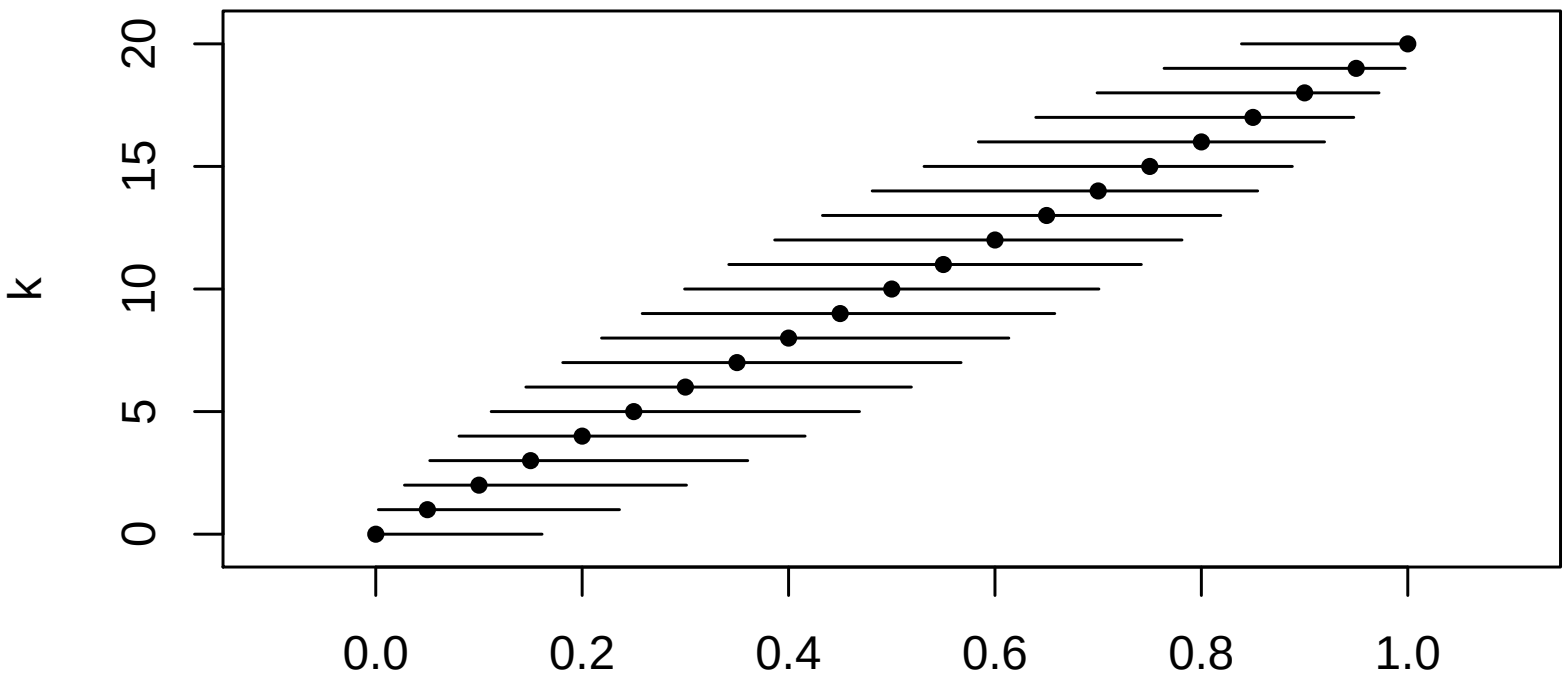
- Tuloksena saadaan Wilsonin luottamusväli

$$\frac{\hat{p} + \frac{1}{2n} z_{\alpha/2}^2 \pm z_{\alpha/2} \sqrt{\frac{1}{n} \hat{p}(1-\hat{p}) + \frac{1}{4n^2} z_{\alpha/2}^2}}{1 + \frac{1}{n} z_{\alpha/2}^2}. \tag{21}$$

- (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 (2) ei sille toteudu.

Wilsonin luottamusvälit, kun $n = 20$

Otoskokoa $n = 20$ vastaavat nimellistä luottamustasoa 95% vastaavat Wilsonin luottamusvälit (21), kun $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti k/n on merkitty pisteellä. Tämä väli ei surkastu pisteeksi, jos onnistumisia on nolla tai n.

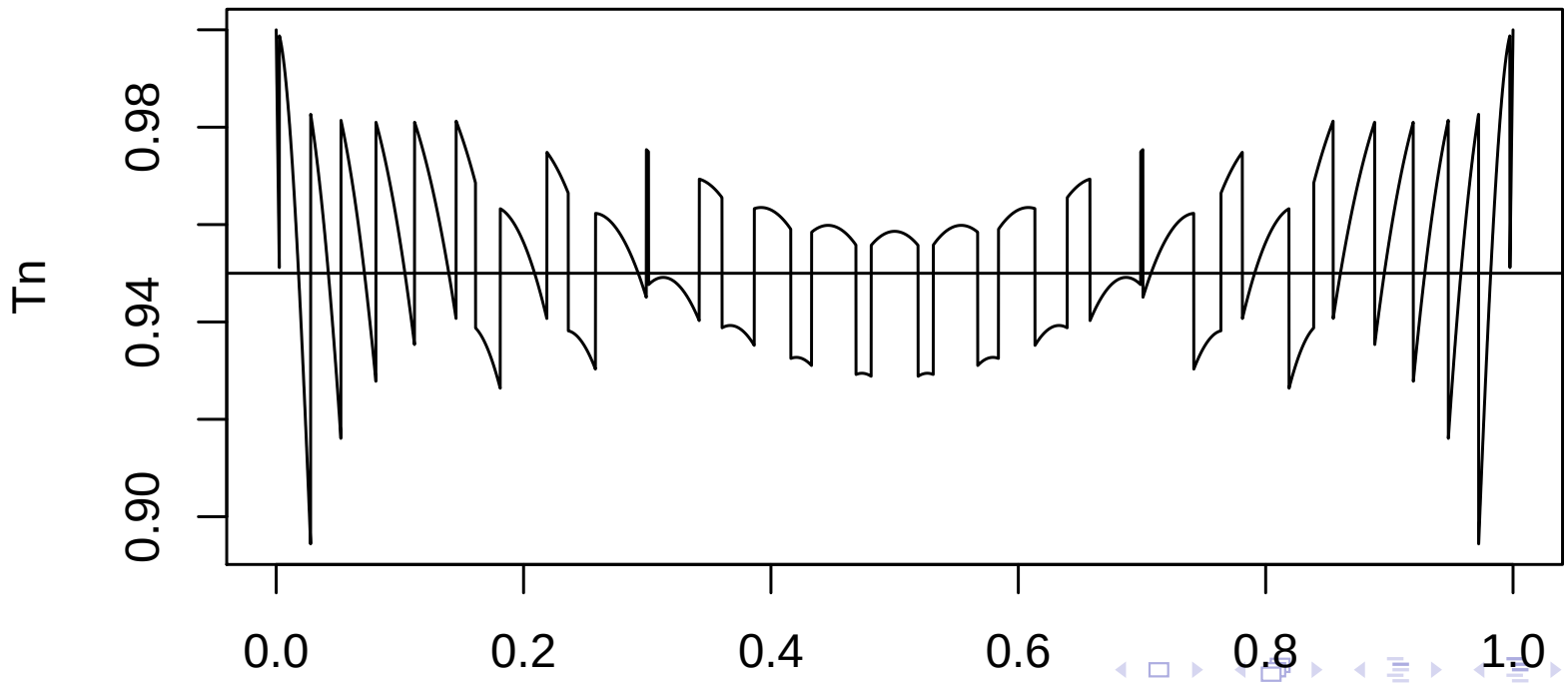


Wilsonin luottamusvälin peittotodennäköisyys

Kuvassa on Wilsonin likimääräisen luottamusvälin todellinen peittotodennäköisyys

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

kun $n = 20$ ja $\alpha = 0.05$.



Tarkka luottamusväli onnistumistodennäköisyydelle

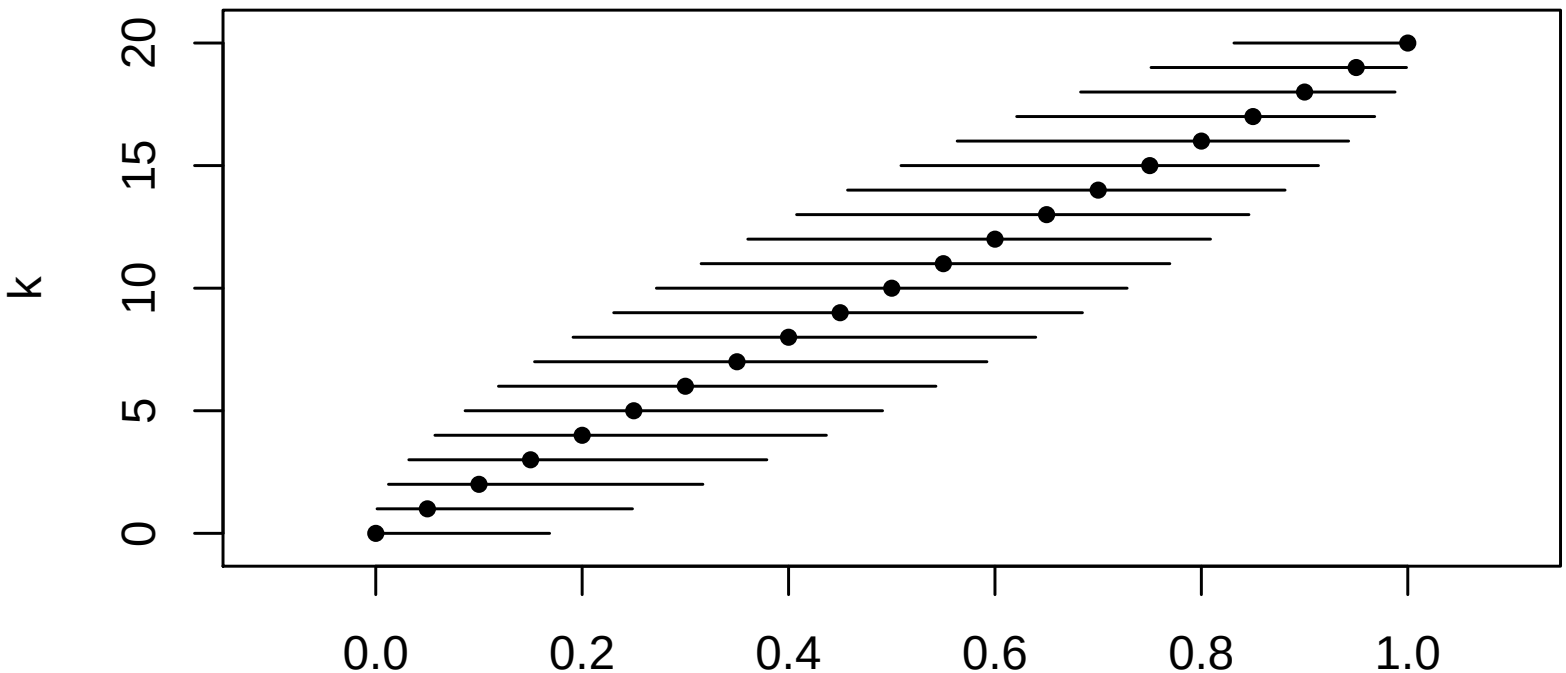
- Clopper ja Pearson esittivät v. 1934 erään tavan muodostaa ns. tarkka (engl. *exact*) luottamusväli onnistumistodennäköisyydelle.
- Termi tarkka tarkoittaa tässä sitä, että kyseinen luottamusväli ei ole likimääräinen, vaan määritelmän (ks. kaava (2)) mukainen, eli

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

- 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.

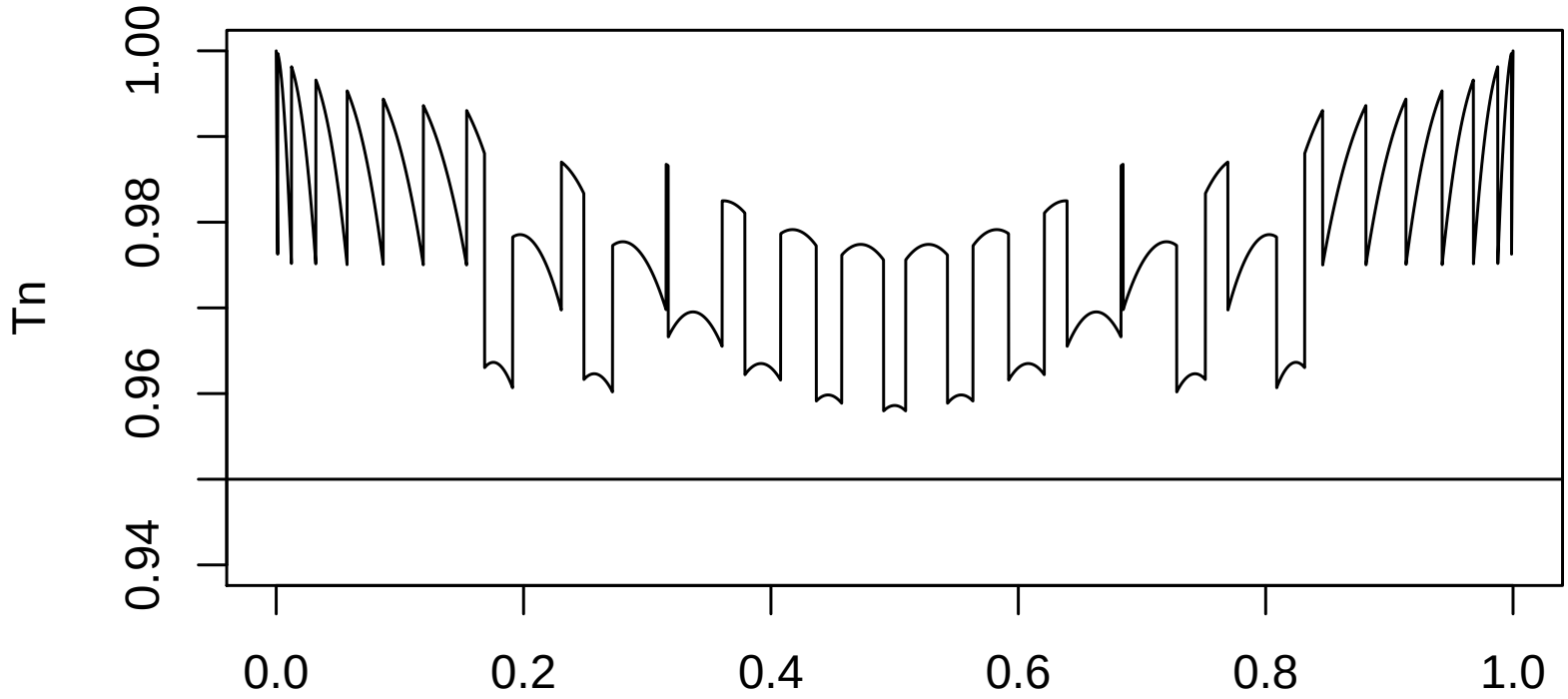
Clopperin ja Pearsonin luottamusvälit, kun $n = 20$

Clopperin–Pearsonin tarkat 95 %:n luottamusvälit, kun otoskoko $n = 20$, ja $0 \leq k \leq n$ on onnistumisten lukumäärä. SU-estimaatti k/n on merkitty pisteellä.



Clopperin ja Pearsonin luottamusvälin peittotodennäköisyys

Clopperin–Pearsonin tarkan 95 %:n luottamusvälin peittotodennäköisyys, kun otoskoko $n = 20$.



Diskreettien jakaumien ongelmallisuus

- Silloin kuin havaintosatunnaisvektorin jakauma on diskreetti, niin yleensä aina joudutaan tekemään luottamusvälien kanssa 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.

Binomijakauman luottamusvälit tietokoneella

- Tietokoneella minkä tahansa edellä mainitun binomijakauman luottamusvälin laskeminen on yhtä helppoa.
- R-ohjelmistossa nämä luottamusvälit on helppo laskea `Hmisc`-kirjaston funktiolla `binconf`. (Muitakin vaihtoehtoja on.)
- Seuraavassa lasketaan luottamustason 0.95 välit, kun $n = 20$ ja onnistumisia on $k = 4$.

Binomijakauman luottamusvälit R:llä

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

##  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 on olemassa aivan toisentyyppisiä välejä, ks. esim. Vardeman 1992. Käsittelemme tässä vain ennusteväliä. Vardeman esittelee myös ns. toleranssivälin.
- Tarkastelemme 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

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

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

Ennustevälin määritelmä

Määritelmä (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

$$P_\theta\left(L(\mathbf{Y}) \leq Y_{n+1} \leq U(\mathbf{Y})\right) \geq 1-\alpha, \quad \text{kaikilla } \theta \in \Theta. \tag{22}$$

Ennusteväli normaalijakaumalle, kun varianssi on tunnettu

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

Ennusteväli normaalijakaumalle, kun varianssi on tuntematon

- Jos myös varianssiparametri on 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


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


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


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