5. NORMAALIJAKAUMAMALLIN PARAMETRIEN ESTIMOINNISTA

[vrt. Arjas - Sirén, jakso 2.6 (ja 2.4)]

Tark. mallia "riippumaton otos jakaumasta $N(\mu, \sigma^2)$":

$$\boxed{X_1, \ldots, X_n \sim N(\mu, \sigma^2) \quad \text{II}}$$

Mallin yhteistiheysf on (ks. s. 8) (merk. $\underset{\sim}{x} = (x_1, \ldots, x_n)$)

$$f(\underset{\sim}{x}; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

Tutkitaan parametrien μ ja σ^2 estimointia su-menetelmällä.

(A) Olet. aluksi: $\sigma^2 > 0$ on tunnettu luku ja parametrina vain μ, $\mu \in \mathbb{R}$. (harvoin realistinen tilanne)

Uskottavuusfunktio (vast. ainerstoa $\underset{\sim}{x} = (x_1, \ldots, x_n)$)

$$(*) \qquad L(\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}, \quad \mu \in \mathbb{R}$$

logaritmi:

$$\ell(\mu) = \log L(\mu) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\ell'(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[-2(x_i - \mu) \right] = \frac{1}{\sigma^2} \underbrace{\left[\sum_{i=1}^{n} x_i - n\mu \right]}_{\text{vaihtaa merkkinsä + :sta - :een}}$$

$$\ell'(\mu) = 0 \iff \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Saatu: μ:n su-estimaatti on

$$\boxed{\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

(otoskeskiarvo)

(B) Yleinen tilanne: parametri on 2-ulotteinen (μ, σ^2), jossa $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

Uskottavuusfunktio $L(\mu, \sigma^2)$ kuten edellä (lauseke (*)).
Pätee hajotelma (tarkista!)

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 ,$$

$$\left(\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{otoskeskiarvo} \right)$$

Logaritmoidaan ja käytetään tätä hajotelmaa:

$$\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2)$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 - \frac{n}{2\sigma^2} (\bar{x} - \mu)^2$$

Huom. tämä riippuu μ:stä vain viimeisen termin kautta!
Selvästi $-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2 \leq 0$ aina ja "=" pätee

$\iff \mu = \bar{x}$.

Tark. sitten (vain σ^2:sta riippuvaa!) funktiota

$$u(\sigma^2) = \ell(\bar{x}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 ,$$

$$u'(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Huom. Derivaatta σ^2:n suhteen, ei σ:n!

$$= \frac{1}{2\sigma^2} \left[-n + \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]$$

$$= 0 \iff \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

(tee merkkitarkastelu: tämä todella on maksimikohta!)

Saatu: parametrin (μ, σ^2) suurimman uskottavuuden estimaatti on

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i , \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Huomautuksia & lisätietoja:

* Tuntuu järkevältä estimoida jakauman odotusarvoa ("teoreettista keskiarvoa") otoskeskiarvolla $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;

* Vastaavalle sm:lle $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (eli estimaattorille) pätee

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Sanotaan: $\bar{X}$ on harhaton estimaattori μ:lle. Se siis tuottaa "odotusarvoisesti" oikean tuloksen! Toistetun ainerstonheruun näkökulmasta: jos otanta toistetaan yhä uudelleen ja uudelleen ja jokaisesta ainerstosta $\underset{\sim}{x}$ lasketaan $\hat{\mu} = \bar{x}$, niin saadut estimaatit osuvat keskimäärin oikeaan. (Vrt. HT 2/5.)

* Sen sijaan $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ ei ole harhaton estimaattori σ^2:lle, vaan sen odotusarvo on $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = (1-\frac{1}{n})\sigma^2$.

Tästä syystä σ^2:n estimaattina useimmiten käytetään luhua

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{(otosvarianssi)}$$

Sitä vastaavalle sm:lle S^2 pätee $E(S^2) = \sigma^2$. Ero $\hat{\sigma}^2$:een on merkityksetön, jollei n ole kovin pieni.

* Tarkemmin ottaen pätee seuraava lause:

Kun $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ $\perp\!\!\!\perp$, niin

i) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$

ii) $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$

iii) $\bar{X} \perp\!\!\!\perp S^2$

"khii-toiseen" jakauma, n-1 vapausastetta

[Todistus aineopintojen tn-laskennan kurssilla.]

6. HIEMAN YLEISTÄ VÄLIESTIMOINNISTA

vrt. Arjas-Sirén, luku 3

Tark. yleistä tilastollista mallia $f(\underset{\sim}{x}; \theta)$, jossa (tuntematon) parametri on θ ja aineisto $\underset{\sim}{x} = (x_1, \ldots, x_n)$.

Edellä tutkimme piste-estimointia, jossa aineiston perusteella piti määrittää sellainen parametrinarvo, joka olisi (tavalla tai toisella) hyvä ja perusteltu "arvaus" todelliselle parametrinarvolle. Su-menetelmä oli (eräs) ratkaisu tähän tehtävään.

Ongelma: piste-estimaatti harvoin osuu juuri oikeaan (emmehä tiedä milloin niin käy).

Haluaisimme määrittää sellaisen joukon parametrinarvoja (esim. $\hat{\theta}$:n ympäriltä), josta voitaisiin "suurella varmuudella" sanoa, että se sisältää oikean parametrinarvon.

1-ulotteisessa tapauksessa (ts. kun θ on 1-ulotteinen tai halutaan tehdä päätelmiä vain sen yksittäisestä komponentista) tällainen joukko on yleensä väli ja puhutaan väli-estimoinnista.

KAKSI RATKAISUA tähän tehtävään:

(1.) Uskottavuusvälit tai -joukot (melko harvoin käytetty mutta helppo menetelmä uskottavuusfunktion pohjalta)

Olk. $L(\theta) = L(\theta; \underset{\sim}{x}) = f(\underset{\sim}{x}; \theta)$ aineistoa $\underset{\sim}{x}$ vast. uskottavuusfunktio.

Mursta: Su-estimaatti $\hat{\theta}$ on piste, jossa $L(\theta)$ saa suurimman arvonsa.

Määr. Kun $0 < c < 1$, niin joukko

$$U_c = U_c(\underline{x}) = \{\theta \in \Omega \mid L(\theta) \geq c\, L(\hat{\theta})\}$$

on $100 \cdot c$ %:n uskottavuusjoukko θ:lle.
(1-ulott. tapauksessa se on yleensä väli)

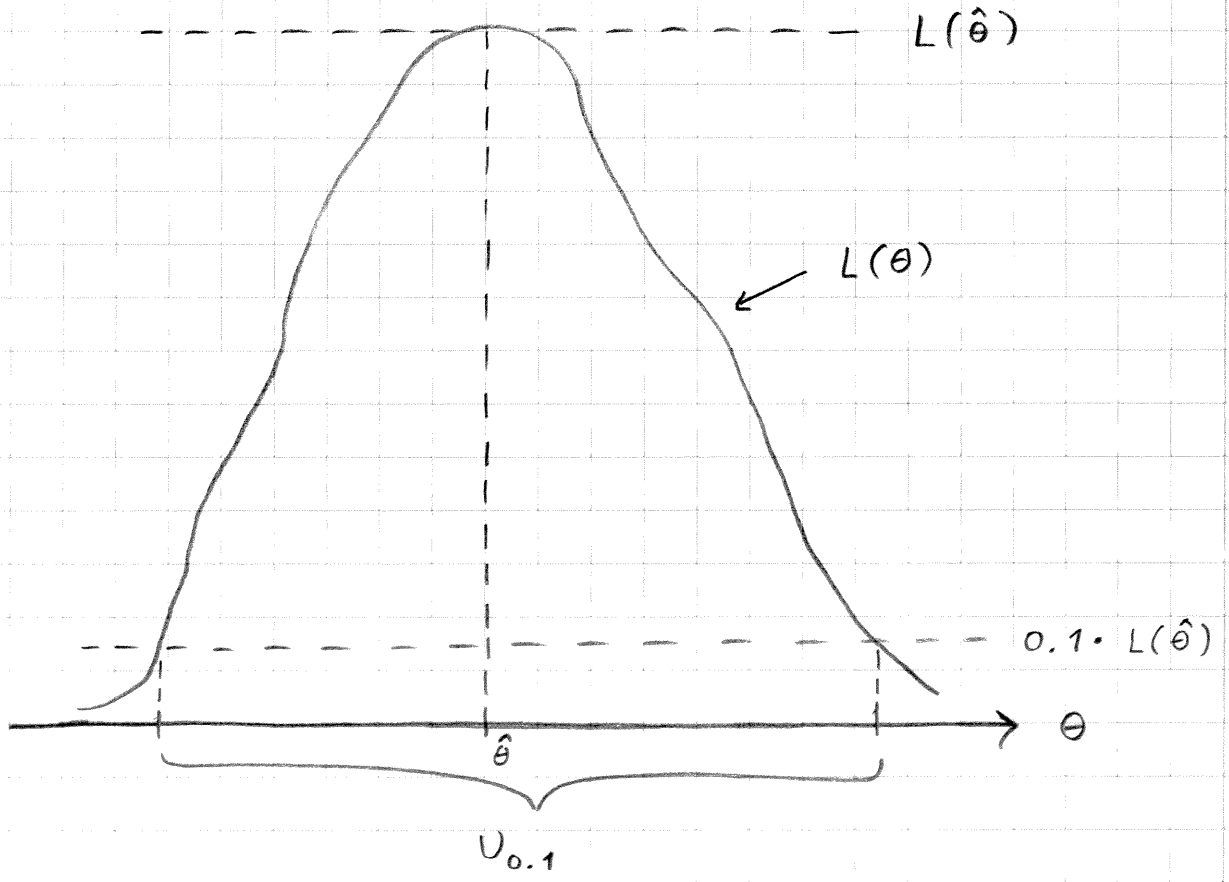
Ω = kaikkien mahdollisten parom. arvojen joukko

Esim. $U_{0.1} = \{\theta \in \Omega \mid L(\theta) \geq 0.1 \cdot L(\hat{\theta})\}$

on 10 %:n uskottavuusjoukko.

Sen ulkopuolelle jäävät ne θ:n arvot, joiden uskottavuus on alle 10 % uskottavuuden suurimmasta arvosta.

(ks. kuva alla)



2. Luottamusvälit (laajassa käytössä tilastollisessa tutkimuksessa)

Kiinnitetään jokin (pieni) luku $0 < \alpha < 1$.

Pyritään aineiston $\underset{\sim}{x}$ perusteella määrittämään joukko $A(\underset{\sim}{x})$ parametriavaruudessa Ω siten, että vastaavalle "satunnaiselle joukolle" $A(\underset{\sim}{X})$ pätee

$$P\left(\theta \in A(\underset{\sim}{X})\right) \geq 1-\alpha.$$

Tällöin $A(\underset{\sim}{x})$ on θ:n luottamusjoukko (luottamus) tasolla $1-\alpha$.

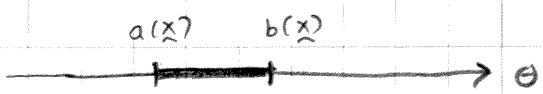
Tyypillisesti esim. $\alpha = 0.05$, jolloin

$$P\left(\theta \in A(\underset{\sim}{X})\right) \geq 0.95$$

ja kyse on 95 %:n (eli tason 0.95) luottamusjoukosta. Satunnainen (toistetun aineistonkeruun mielessä !) joukko $A(\underset{\sim}{X})$ siis peittää todellisen parametrinarvon θ (ainakin) tn:llä 0.95.

1-ulotteisen parametrin tapauksessa $A(\underset{\sim}{x})$ on yleensä väli reaaliakselilla : voimme siis kirjoittaa esim. $A(\underset{\sim}{x}) = \left(a(\underset{\sim}{x}), b(\underset{\sim}{x})\right)$, jossa päätepisteillä on ominaisuus

$$P\left(a(\underset{\sim}{X}) < \theta < b(\underset{\sim}{X})\right) \geq 1-\alpha$$



Tutkimme seuraavassa jaksossa luottamusvälien muodostamista normaalijakaumamallin tapauksessa. Ylempiin asioa käsitellään aineopintojen til. päättelyn kurssilla.

6. LUOTTAMUSVÄLIT NORMAALIJAKAUMALLE

[vrt. Arjas-Siren, luku 3]

Tark. mallina $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ ⫫ .

Tavoite: Muodostetaan luottamusväli odotusarvolle μ annetulla luottamustasolla $1-\alpha$ (esim. 0.95)

Pal. mieleen, että μ:n (piste-)estimaattina käytetään

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Vastaavasti σ^2:n estimaattina (useimmiten) käytetään

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

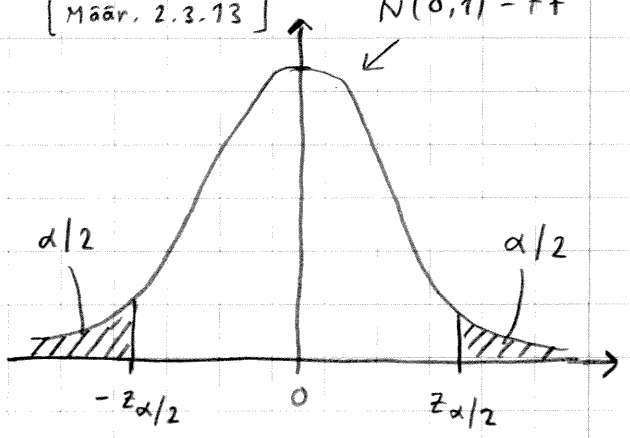
(A) Olet. että $\sigma^2 > 0$ on tunnettu luku (ei siis parametri)

Sivulla 16 todettiin, että sm:lle $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ pätee

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$ [vrt. Tuominen, TN I, Esim. 3.7.9]

eli

$$(*) \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$ [TN I, Määr. 2.3.13]



Valitaan $N(0,1)$ - jakaumasta kohta $z_{\alpha/2}$, josta oikealle "häntätn" $= \alpha/2$

Symmetria $\Rightarrow$ $-z_{\alpha/2}$:sta vasemmalle "häntätn" on samoin $\alpha/2$

Välille $(-z_{\alpha/2}, z_{\alpha/2})$ jäävä tn-massa $= 1-\alpha$.

Siten

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

$\Rightarrow$

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \qquad (**)$$

Siis (satunnaisten) pisteiden $\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ ja $\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ rajoittama väli sisältää μ:n tn:llä $1-\alpha$.

Tämä merkitsee:

> Kun $\underset{\sim}{x} = (x_1, \ldots, x_n)$ on havaittu aineisto ja $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, niin väli
>
> $$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \qquad (\#)$$
>
> on μ:n luottamusväli luottamustasolla $1-\alpha$.

Esim. Kun $\alpha = 0.05$, niin

$$z_{\alpha/2} = z_{0.025} \simeq 1.96 \ (\simeq 2)$$

joten 95 %:n luottamusväli μ:lle on

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Huom. Lukujen $z_{\alpha/2}$ arvoja on normaalijakauman taulukoissa!

(B) Oletetaan, että myös σ^2 on (tuntematon) parametri

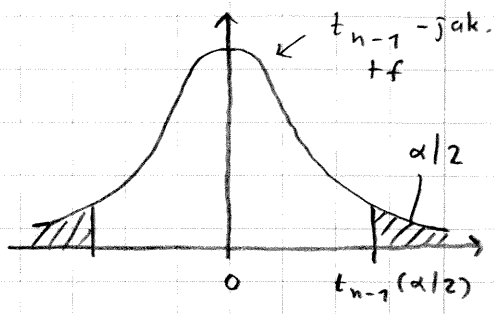
Ongelma: väli (#) yllä ei käytettävyssä, koska σ ei tunnettu $\Rightarrow$ joudumme estimoimaan sen aineistosta. Estimaattina käytämme

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \left(\frac{\text{otos-}}{\text{keskihajonta}}\right)$$

Olkoon S vastaava satunnaismuuttuja.
Tällöin pätee (vrt. A-tapauksen jakaumatulos (*) !)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \qquad \left(\begin{array}{l} \text{Studentin } \underline{t\text{-jakauma}}, \\ n-1 \ \underline{\text{vapausosteTta}} \end{array} \right)$$

Menetellään kuten A-kohdassa:
Valitaan piste $t_{n-1}(\alpha/2)$, josta oikealle t_{n-1}-jakauman häntätn $= \alpha/2$. Päädytään (**):n sijasta tulokseen



$$P\left(\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \right) = 1 - \alpha$$

Saatu:

Kun $\underset{\sim}{X} = (X_1, \ldots, X_n)$ on havaittu arvossa ja $\bar{x}$ = otoskeskiarvo sekä s = otoskeskihajonta (ks. edell. sivu), niin väli

$$\left(\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} \, , \, \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} \right) \qquad (\#\#)$$

on μ:n luoTamusväli luoTamustasolla $1-\alpha$.

Huom. * VerraTuna lv:iin (#) tässä lasketaan ainerstosta $\bar{x}$:n lisäksi myös keskihajonta s. Lisäksi kerroin $t_{n-1}(\alpha/2)$ riippuu otoskoosta n.

* t_{n-1}-jakaumalla on "paksummat hännät" kuin $N(0,1)$-jakaumalla $\Rightarrow$ kerroin $t_{n-1}(\alpha/2)$ on suurempi kuin $z_{\alpha/2}$. Käytännössä ero on merkityksellinen vain pienillä n:n arvoilla ($\sim$ alle 50), sillä t_{n-1} lähestyy $N(0,1)$-jakaumaa kun $n \to \infty$.

* Lukuja $t_{n-1}(\alpha/2)$ löytyy tauluhaista ja tietokoneohjelmista, joten t-jakaumien tiheysfunktioita ei taruitse tietää.