Parametrin estimointi ja bootstrap-otanta

Käytännön asioista

- tiistaiaamun laskuharjoituksia ei jatkossa pidetä vähäisen osallistujamäärän vuoksi
- harjoitustyön 1. osan esitteleminen (10-15 min) luennolla 15.4. tuo 3 pistettä
- viimeisellä luentokerralla 24.4. vastaava tilaisuus, jossa jälleen halukkaat voivat ilmoittautua esittelemään aihettaan ja työnsä 2. osaa
  - noin 3 halukasta voidaan ottaa
  - näistä sovittava Mikon kanssa etukäteen

Luennon runko

- Tilastollisia käsitteitä (lähinnä kertaus)
  - otos ja populaatio, populaatiojakauma
  - otosjakauma
  - tunnusluku ja parametri
- Laskentaintensiivisiä menetelmiä
  - Monte Carlo ja satunnaistaminen (kertaus)
  - bootstrap
- Parametrin estimointi ja luottamusvälien määrittäminen
  - otosestimaattorit
  - luottamusvälien määrittäminen bootstrap-otannalla

Otos ja populaatio

- käytettävissä oleva aineisto on otos ilmiöstä
- koko ilmiön perusjoukkoa, josta otos on peräisin, nimitetään usein populaatioksi
- $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$
- satunnaismuuttujat $X_1, \ldots, X_n$ kuvaavat koko ilmiötä
- niiden jakaumasta käytetään usein nimitystä populaatiojakauma

Otosjakauma

- mielivaltainen tunnusluku t
- tietty otoksen koko n
- t:n otosjakauma kuvaa tunnusluvun arvojen jakauman kaikissa mahdollisissa n-kokoisissa otoksissa perusjoukosta
- teoreettinen jakauma
- otosjakauma **ei** viittaa otoksen arvojen jakaumaan (engl. *sampling distribution* = otosjakauma, *sample distribution* = (yksittäisen) otoksen arvojen jakauma
- sanat vaarallisen lähellä toisiaan, kysymys aivan eri asioista

Tunnusluku ja parametri

- populaatiojakauman tunnuslukuja kutsutaan usein parametreiksi, tällöin
  - tunnusluvun arvo on otoksen funktio (esim. otoskeskiarvo)
  - parametrin arvo on koko populaation funktio ("todellinen" keskiarvo = populaatiojakauman odotusarvo)
  - mitä voidaan sanoa koko populaatiosta, kun tunnetaan otos?
  - parametrin arvioiminen otoksen perusteella (tästä hiukan myöhemmin)

Monte Carlo

- populaatiojakauma tunnettu, tunnusluvun otosjakauma ei
- otosjakauman approksimointi simuloimalla otantaa populaatiosta
- generoidaan otoksia populaatiojakaumasta
- Monte Carlo approksimaatio: otoksista laskettujen tunnusluvun arvojen jakauma
- lisäämällä otosten määrää, päästään mielivaltaisen tarkkaan approksimaatioon (Suurten lukujen laki)

Satunnaistaminen

- satunnaistamismenetelmillä voidaan testata (kahta) otosta koskevia hypoteeseeja
- perusjoukko ei ole populaatio vaan käytettävissä oleva otos

$\Rightarrow$ approksoimoitava nollahypoteesin mukainen otosjakauma ei ole tunnusluvun jakauma populaatiossa, vaan otoksen kaikissa mahdollisissa permutaatioissa

- entä jos haluttaisiin tehdä päätelmiä koko populaatiosta, mutta ei tunneta populaatiojakaumaa?

Bootstrap-menetelmät

- tunnusluvun otosjakaumaa ei tunneta, vain otos
- mitään oletuksia populaatiosta ei tarvitse tehdä
- idea: ”kuvitellaan” että otos on koko populaatio ja tehdään otantaa otoksesta!
- jos otos on edustava, näin voidaan approksimoida minkä tahansa tunnusluvun otosjakaumaa
- engl. *bootstrap* = saappaan “hihna”; viittaa itsensä nostamiseen kiskomalla omasta saappaasta eli bootstrap-otannan intuitiiviseen mahdottomuuteen (ensikatsomalta)

Bootstrap ja satunnaistaminen

- bootstrap- ja satunnaistamisalgoritmit ovat hyvin samantapaisia
- satunnaistamisessa otanta (kahdesta) otoksesta tehdään ilman takaisinpanoa
- bootstrapissa otanta otoksesta tehdään takaisinpanolla

⇒ alkuperäisen otoksen alkio voi tulla valituksi useita kertoja samaan pseudo-otokseen

- otanta takaisinpanolla simuloi tilannetta, että otantaa tehdään koko populaatiosta

Bootstrap-algoritmi

- Olkoon S otos, jonka koko on N
  - Toista K kertaa
    - muodosta N:n alkion pseudo-otos S_i^* otoksesta S seuraavasti:
      - toista N kertaa: valitse satunnaisesti alkio S:stä ja lisää se otokseen S_i^*
    - laske pseudo-tunnusluvun θ_i^* arvo S_i^*:stä
- pseudo-tunnusluvun jakauma on bootstrap-otosjakauma

Bootstrapin intuitiivinen perustelu

- otoksen arvojen jakauma on paras arvio sille, millaisia arvoja populaatiossa on ja kuinka ne ovat jakautuneet
- bootstrap ei ole taikatemppu, joka toimii aina
- tuottaa huonon tuloksen, jos otos on huono tai harvinainen
- toisaalta tällöin on vaikea tehdä mitään järkevää (paitsi kerätä lisää/uutta dataa)
- vrt. nollahypoteesin hylkäämisen virhemahdollisuus eli p-arvo

⇒ joskus menee pieleen

Mihin bootstrap soveltuu

- bootstrapia voi käyttää lukuisiin erityyppisiin tehtäviin
- hypoteesin testaus tilanteissa joissa populaatiojakaumaa ei tunneta
- estimaatin luotettavuuden arvioiminen

Parametrin estimointi

- arvioi parametrin arvoa, kun tunnetaan otos
- arvioinnin eli *estimoinnin* tulos on arvio eli *estimaatti*
- termi *estimaattori* viittaa estimointimenetelmään
- hyvä estimaattori tuottaa estimaatteja, jotka ovat "mahdollisimman usein mahdollisimman lähellä" estimoitavan parametrin todellista arvoa
- luottamusväli (*confidence interval*) on tapa ilmaista arvion tarkkuutta

Otostunnusluvut estimaattoreina

- yksinkertaisin tapa estimoida parametrien arvoja on käyttää otoksesta laskettuja tunnuslukuja
  - käytetään odotusarvon arviona otoskeskiarvoa, varianssin arviona otosvarianssia, korrelaation arviona otoskorrelaatiokerrointa jne.
- muitakin menetelmiä on (ei käsitellä tällä kurssilla)
- populaatiojakaumasta saattaa olla lisätietoa, jota voidaan hyödyntää esim. ns. suurimman uskottavuuden (*maximum likelihood*) menetelmällä

Esimerkki

- oletetaan, että populaatiojakauman odotusarvoa μ ei tunneta
- kerätään otos ja lasketaan otoskeskiarvo $\bar{x} = 10.0$
- käytetään otoskeskiarvoa estimoimaan todellista μ:n arvoa
- 10.0 *saattaa* olla tarkka, oikea μ:n arvo
- luultavasti kuitenkin μ on "jossakin lähellä" arvoa 10.0, eli $\mu = \bar{x} \pm \epsilon$
- jos ϵ on "pieni", $\bar{x}$ on hyvä estimaatti μ:lle

Otostunnusluvut estimaattoreina

- *harhattomia* ja *tarkentuvia* estimaattoreita
  - harhattomalle (*unbiased*) estimaattorille $E\check{\theta} = \theta$

⇒ tuottaa "keskimäärin" oikean tuloksen
  - tarkentuvalle (*consistent*) estimaattorille

$$\lim_{n\to\infty} P\{|\check{\theta} - \theta| > \epsilon\} = 0$$

⇒ otoksen kokoa kasvattamalla päästään mielivaltaisen lähelle parametrin todellista arvoa

Arvion tarkkuus: luottamusväli

- jos otoksen koko on kiinnitetty, kuinka lähellä todellista arvoa estimaatti sitten on?

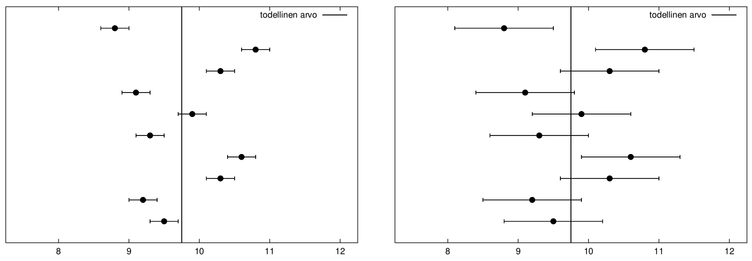
⇒ luottamusväli

- todennäköisyys sille, että todellinen arvo on tietyllä välillä (johon estimaatti sisältyy)
- esim. 95 %:n luottamusväli
- ei tarkoita sitä, että tiedämme 95 %:n varmuudella mikä todellinen arvo on

Esimerkki, jatkoa

- kuinka todennäköistä on, että todellinen populaation keskiarvo μ sisältyy välille $\mu = \bar{x} \pm \epsilon$, kun ϵ on jokin kiinnitetty arvo?
- oletetaan että tunnemme populaation odotusarvon, olkoon se 9.75
- kerätään 10 uutta otosta
- asetetaan ensin $\epsilon = \epsilon_1 = 0.1$, sitten $\epsilon = \epsilon_2 = 0.6$

Luottamusväli, esimerkki



- $\epsilon = \epsilon_1 = 0.1$ ⇒ vain yhdessä tapauksessa arvio on välillä $[\mu - \epsilon, \mu + \epsilon]$
- $\epsilon = \epsilon_2 = 0.6$ ⇒ yhdeksässä tapauksessa arvio on välillä $[\mu - \epsilon, \mu + \epsilon]$

Luottamusväli bootstrapilla

- Olkoon S otos ja $\check{\theta}$ otostunnusluku, jota käytetään estimaattina parametrille θ

1. muodosta K kappaletta bootstrap-otoksia S:stä ja laske niistä pseudo-tunnusluku $\theta_i^*, 1 \leq i \leq K$
2. lajittele pseudo-tunnusluvut
3. $100(1-2\alpha)$-prosentin luottamusvälin approksimaation alaraja on lajiteltujen pseudo-tunnuslukujen $K\alpha$:s arvo ja yläraja $K(1-\alpha)$:s arvo

Luottamusvälien muodostaminen

- luottamusvälien muodostaminen ilman simulaatiota perustuu **keskeiseen raja-arvolauseeseen**
  - tarkastellaan keskiarvon otosjakaumaa otoksen koon n funktiona
  - otoksen koon kasvaessa otosjakauma ”muistuttaa enemmän ja enemmän” normaalijakaumaa
  - jos populaatiojakauman odotusarvo on μ ja keskihajonta σ,normaalijakauman odotusarvo on μ ja keskihajonta $\sigma/\sqrt{N}$
- tuloksen merkittävyys: pätee kaikille mahdollisille populaatioille

Esimerkki, jatkoa

- otoskeskiarvo $\bar{x} = 10.0 = \check{\mu}$
- jos otoskoko N on ”suuri”, arvo 10.0 on peräisin jakaumasta, joka on likimäärin normaalijakauma odotusarvona μ

$\Rightarrow$ $\bar{x} = \mu \pm 1.96\sigma/\sqrt{N} \approx 95$ %:n varmuudella

- 95 %:n luottamusväli siis kyseisen normaalijakauman 95%-fraktiiliväli
- σ ei välttämättä ole tunnettu

$\Rightarrow$ joudutaan käyttämään otoskeskihajontaa

Estimaatin luottamusväli riippuu

- siitä, kuinka suuri varmuus halutaan
- otoksen koosta
- otoksen varianssista
- estimaattorista

Luottamusvälit

- voidaan muodostaa perinteisin menetelmin otostunnusluvuille, joiden otosjakauma on (asymptoottisesti) normaalijakauma
- ehto ei kuitenkaan päde edes kaikille tavallisimmille tunnusluvuille
- esim.korrelaatiokertoimen otosjakauma ei ole asymptoottisesti normaali
- bootstrapilla luottamusväli voidaan muodostaa otoskorrelaatiokertoimelle (todellisen korrelaation estimaatille)

Esimerkki

- yllä alkuperäinen otos (korrelaatiokerroin -0.552), alla yksi bootstrap-otos (-0.545)

x	5	1.75	0.8	5	1.75	5	1.75	1	5	1.75
y	27.8	20.82	44.12	29.41	31.19	28.68	29.53	34.62	20	41.54

x	5	1.75	1.75	5	1	5	0.8	1.75	5	5
y	27.8	20.82	29.53	27.8	34.62	28.68	44.12	31.19	20	27.8

Yhteenveto

- bootstrap perustuu otantaan otoksesta takaisinpanolla
- simuloi otantaa koko populaatiosta
- bootstrap-otosjakauma approksimoi tunnusluvun otosjakaumaa koko populaatiossa
- ei oletuksia populaatiosta
- otoksen täytyy vain olla edustava
- bootstrapia voidaan käyttää parametrin estimoinnissa luottamusvälien konstruoimiseen