

Tutkimustiedonhallinnan peruskurssi

Hannu Toivonen, Marko Salmenkivi, Inkeri Verkamo

`hannu.toivonen, marko.salmenkivi, inkeri.verkamo@cs.helsinki.fi`

Helsingin yliopisto – Tietojenkäsittelytieteen laitos

Satunnaistamistestaus



Satunnaistamistestaus: sisältö

- hypoteesin testauksen perusteet
- p -arvon arvioiminen satunnaistamalla
- kahden joukon vertaaminen
- parittainen testaus
- riippumattomuuden testaaminen
- Lähde: Cohen, luvut 4.2, 4.3, 5.3

Hypoteesin testaus

- tarkastellaan aiempaa shakkiesimerkkiä
- shakkiohjelma A voitti B:n 10 ottelussa 15:stä eli osuuden $f = 0.67$
- onko A parempi kuin B?
- f on otoksen tunnusluku; olkoon π populaation parametri, joka kuvaa A:n todennäköisyyden voittaa satunnainen ottelu
- jos A ja B ovat yhtä hyviä niin $\pi = 0.5$
 - tämä on *nollahypoteesi*
- *vaihtoehtoinen hypoteesi*: A on parempi kuin B

Otosjakauma ja hypoteesin testaus

- nollahypoteesin vallitessa tunnusluku f noudattaa jotain tiettyä otosjakaumaa
 - (binomijakauma, mutta sillä ei tässä ole merkitystä)
- f :n otosjakauma määrää, millä todennäköisyydellä 15 pelin otoksessa havaitaan $f \geq 0.67$
- jos todennäköisyys $P(f \geq 0.67)$ on suuri, niin havaittu A:n kannalta suotuisa tulos voidaan saada sattumalta
 $\Rightarrow$ ei voida päätellä luotettavasti että A on parempi kuin B

-
- jos todennäköisyys $P(f \geq 0.67)$ on pieni, on epätodennäköistä että A:n kannalta niin suotuisa tulos on saatu sattumalta
 $\Rightarrow$ *hylätään* nollahypoteesi $\pi = 0.5$
 - $p = P(f \geq 0.67)$ on tuloksen *p-arvo*
 - mitä pienempi *p-arvo*, sitä tilastollisesti merkitsevämpi tulos
 - jos $p \leq 0.05$, niin A on parempi kuin B *merkitsevyystasolla* $\alpha = 0.05$
 - tyypillisiä merkitsevyystasoja (raja-arvoja merkitsevyydelle): 0.05, 0.01, 0.001

Satunnaistamistestaus

- nollahypoteesin mukainen tunnusluvun otosjakauma on usein tuntematon
 - p -arvon laskeminen analyttisesti voi olla vaikeaa tai mahdotonta
 - otosjakautumaa voidaan yleensä estimoida helposti Monte Carlo -menetelmin
 - esimerkiksi 15 ottelun turnausten simulointi
- ⇒ arvojen jakauma on estimaatti otosjakautumalle

Nollahypoteesi

Valinta, esim.

- kahden joukon tai otoksen vertaaminen
 - nollahypoteesi: joukot ovat samasta jakaumasta
- riippuvuus tai korrelaatio asioiden välillä
 - nollahypoteesi: yhteydet ovat satunnaisia

Testaaminen

- sekoitetaan data nollahypoteesin mukaisesti
 - = permutation test = (re)randomization test
 - generoidaan monta permutaatiota
- saadaan nollahypoteesin mukainen jakauma

Esimerkki

Sovellus: opiskelijavalinta

- opiskelijavalintaperusteita muutettiin
- käytössä myös erillinen, samana pysynyt testi osalle opiskelijoista
- hajonta testin pisteissä näyttää kasvaneen
- hajonnan mittana kvartiilivälin pituus:
 $\text{IQR} = 0.75\text{-fraktiili} - 0.25\text{-fraktiili}$
- testattava muuttuiko tulos merkittävästi
 - käytettävissä 25 opiskelijan tulokset viime vuodelta
 - 20 opiskelijan tulokset tältä vuodelta

Esimerkki

- viime vuonna $IQR = 6$, tänä vuonna $IQR = 8.5$
- erotus $d = 6 - 8.5 = -2.5$
- millä todennäköisyydellä erotus $d = -2.5$ saatiin sattumalta?
- nollahypoteesi: vaihtelu ei tänä vuonna ole viime vuotta suurempaa, vaan opiskelijat ovat samasta (tuntemattomasta) populaatiojakaumasta
- kuka tahansa opiskelijoista voisi olla kummassa ryhmässä tahansa
- nollahypoteesin mukainen tilanne saadaan sijoittamalla opiskelijat ryhmiin satunnaisesti

Kahden joukon vertaaminen

- **Algoritmi:**
- Olkoon S_A ja S_B kaksi otosta, joiden koot ovat N_A ja N_B .
Olkoon $\theta = f(S_A, S_B)$ otoksista laskettu tunnusluku.
- Toista K kertaa ($i = 1, \dots, K$)
 - sekoita joukkoyhdisteen $S_{A+B} = S_A \cup S_B$ alkiot
 - sijoita N_A ensimmäistä pseudo-otokseen A_i^* ja loput N_B alkiota B_i^* :hin
 - laske $\theta_i^* = f(A_i^*, B_i^*)$
- Käytä θ_i^* :n jakaumaa sen testaamiseen, mikä on θ :n todennäköisyys nollahypoteesin vallitessa

p -arvon arvioiminen

- oletetaan, että vaihtoehtoinen hypoteesi on muotoa
“ θ on suurempi kuin voisi odottaa sattumalta”
- $p = P(\theta_i^* \geq \theta \mid \text{nollahypoteesi on voimassa}) \approx \frac{|\{\theta_i^* \geq \theta \mid 1 \leq i \leq K\}|}{K}$ ts. $p \approx \max_i \{(K - i)/K \mid \theta_{(i)}^* \geq \theta\}$
- vastaavasti ilmeisin muutoksin jos vaihtoehtoinen hypoteesi on, että θ on satunnaista pienempi
 - nämä ovat “yksihäntäisiä” testejä
- jos hypoteesi on muotoa θ poikkeaa sattumanvaraisesta, käytetään kaksihäntäistä testiä:
 - jos $\theta \geq \theta_i^*$:n mediaani: $p \approx 2 \frac{|\{\theta_i^* \geq \theta \mid 1 \leq i \leq K\}|}{K}$
 - jos $\theta \leq \theta_i^*$:n mediaani: $p \approx 2 \frac{|\{\theta_i^* \leq \theta \mid 1 \leq i \leq K\}|}{K}$

Merkitsevyystasot

- usein käytetään (ennalta asetettuja) kynnyksarvoja tilastollisen merkitsevyyden kriteereinä
- testi on merkitsevä tasolla α jos $p \leq \alpha$
- merkitsevyystasoa α vastaavat testisuureen kynnyksarvot ovat satunnaistamistestin järjestettyjen tulosten $\theta_{(1)}^*, \dots, \theta_{(K)}^*$
 - α -fraktiili ja $(1-\alpha)$ -fraktiili (yksihäntäiset testit)
 - $(1-\alpha)$ -fraktiiliväli (kaksihäntäinen testi)

Satunnaistamistestien tulkinta

- Satunnaistamistestien tulkinta aina mahdollisimman konservatiivinen

- Esim.

- $\theta_{(1)}^* = 3, \theta_{(2)}^* = 4, \dots, \theta_{(7)}^* = 4, \theta_{(8)}^* = 10, \dots, \theta_{(K)}^* = 231$
- $p = P(\theta_i^* \leq 4) \approx 7/10$
- $p = P(\theta_i^* \leq 1) \neq 0! \quad "p < 1/K"$

Parittainen testi

- kaksi robottia osallistuu robottien 10-otteluun
 - A saa tehtävistä keskimäärin enemmän pisteitä kuin B
 - onko A merkitsevästi parempi kuin B?
 - nollahypoteesi: robotit ovat yhtä hyviä
- kunkin ottelun tulos olisi voinut yhtä hyvin olla toisin päin
- satunnaistaminen: vaihda satunnaisesti tehtävien sisällä robottien tuloksia keskenään
 - huom: kukin pistemäärä liittyy tiettyyn tehtävään, pisteiden vaihto tehtävien kesken ei perusteltua

Riippumattomuustesti

- tarkastellaan esim. kahta aineiston attribuuttia
- onko niillä merkitsevää tilastollista riippuvuutta?
- **Algoritmi:**
- Olkoot S_A muuttujan A saamien arvojen jono ja S_B muuttujan B saamien arvojen jono. Olkoon $\theta = f(S_A, S_B)$ jonoista laskettu tunnusluku (esim. korrelaatio)
- Toista K kertaa ($i = 1, \dots, K$):
 - sekoita S_B :n järjestys satunnaiseksi
 - laske $\theta_i^* = f(S_A, S_B)$
- Käytä θ_i^* :n jakaumaa sen testaamiseen, mikä on θ :n todennäköisyys nollahypoteesin vallitessa

Satunnaistamistesti

- ei tarvitse oletuksia jakaumista ja niiden parametreista
- ei estimoi populaation parametreja
- käyttää vain saatavilla olevaa dataa
- tarvitsee tyypillisesti kaksi aineistoa tai kaksi muuttujaa, joita vertaillaan
- tuottaa likimääräisen p -arvon
 - täydellinen permutointi tuottaisi tarkan arvon, mutta harvoin mahdollinen
- voi olla laskennallisesti vaativaa

Esimerkki

- sovellus: geenikartoitus eli tietylle perinnölliselle sairaudelle altistavan geenin paikantaminen
- aineisto koostuu sairaista ja terveistä ihmisistä
- lisäksi tietoja kunkin ihmisen perimän mutaatioista (geenimerkeistä)
- kartoitusmenetelmän ensimmäisessä vaiheessa muodostetaan kullekin mutaatiolle malli, joka ennustaa mutaation perusteella onko henkilö sairas vai terve
 - taudille altistavaa mutaatiota käyttävä malli erottelee sairast hyvin terveistä
 - mutta hyvä malli voi löytyä sattumalta, jos testattavia mutaatioita on paljon

Esimerkki (2)

- mallin hyvyyden arviointi satunnaistamistestillä
 - riippumattomuustesti: sairas/terve-luokittelu ei riipu perimästä
 - satunnaistetaan luokat
 - kuinka hyviä malleja menetelmä onnistuu löytämään nollahypoteesin vallitessa?
- geenin ennustetaan sisältävän sen mutaation, joka saa parhaan p -arvon

Hypoteesin testaus: yhteenveto

- mitä testattava hypoteesi koskee?
 - usein eroavuus (kahden otoksen/ aineiston välillä) tai tilastollinen riippuvuus (kahden muuttujan välillä)
 - usein jokin “rakenteellinen” asia: kahden joukon suhde, kahden muuttujan suhde, asioiden järjestykseen liittyvä asia
- nollahypoteesia vastaa yleensä rakenteen satunnainen sekoittaminen
- vaihtoehtoinen ja nollahypoteesi ovat yksinkertaisia
 - “hypoteesissa”, että A:n todennäköisyys voittaa B on 0.67, on itse asiassa kyse parametrin estimoinnista, josta puhutaan erikseen

-
- hypoteesin testaaminen satunnaistamalla
 - toista satunnainen sekoittaminen monta kertaa
 - p -arvo on niitten satunnaistamisten osuus, jotka ovat vähintään yhtä “hyviä”
 - saatu p -arvo on likimääräinen
 - karkea nyrkkisääntö: kohtuullinen tarkkuus $K = 50/\alpha$ satunnaistamisella, missä α on merkitsevyystaso (tai p -arvo)

Hypoteesin testaus: kertausta

- hypoteesin testaus: nollahypoteesi, vaihtoehtoinen hypoteesi
- p -arvo: todennäköisyys saada havaittu tai voimakkaampi tunnusluku nollahypoteesin vallitessa
 - pieni p -arvo $\rightarrow$ tilastollisesti merkitsevä tulos
 - p -arvo perustuu tunnusluvun otosjakauman tarkasteluun
- otosjakaumaa voidaan estimoida satunnaistamalla
- miten valitaan nollahypoteesi?
- millaisella satunnaistamisella mallitetaan nollahypoteesin mukaista tunnusluvun jakaumaa?

Parametrin estimointi ja bootstrap-otanta



Bootstrap: sisältö

- satunnaistaminen vs. bootstrap
- bootstrap-algoritmi
- parametrin estimointi
- luottamusvälit
- lähde: Cohen, luvut 4.6, 5.2, 5.4

Kertaus: satunnaistaminen

- satunnaistamismenetelmillä voidaan testata otosta koskevia hypoteeseja
- perusjoukko josta otantaa tehdään, ei ole populaatio vaan käytettävissä olevan otoksen permutaatiot
- ⇒ approksoimitava 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 vs. satunnaistaminen

- bootstrap- ja satunnaistamisalgoritmit ovat hyvin samantapaisia
- *satunnaistaminen matkii nollahypoteesin mukaista tilannetta, bootstrap ei pyri muuttamaan/tekemään oletuksia*
- satunnaistamisessa otanta otoksesta (tai kahdesta) 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 ($i = 1, \dots, K$)
 - muodosta N :n alkion pseudo-otos S_i^* otoksesta S seuraavasti:
 - toista N kertaa: valitse (takaisinpanolla) satunnainen alkio S :stä ja lisää se otokseen S_i^*
 - laske pseudo-tunnusluvun θ_i^* arvo S_i^* :stä
- pseudo-tunnusluvun θ_i^* jakauma on tunnusluvun θ bootstrap-otosjakauma

Bootstrapin intuitio

- 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
- parametrin estimaatin luottamusvälin arvioiminen

Tunnusluku vs. parametri

- populaatiojakauman tunnuslukuja kutsutaan usein *parametreiksi*, tällöin
 - tunnusluvun arvo on otoksen funktio
 - esim. otoskeskiarvo $\bar{x}$, otoskeskihajonta σ_x
 - parametrin arvo on koko populaation funktio (“todellinen” tunnusluku)
 - esim. satunnaismuuttujan odotusarvo $EX = \mu_X$ ja keskihajonta $VX = \sigma_X$
- käsitteellisesti siis kaksi eri asiaa, vaikkakin lähellä toisiaan

Jakaumien parametreja

- muuttujan *odotusarvo* EX on sen keskimääräinen arvo:

$$EX = \sum_x xP(X = x) \text{ tai } \int_x x f(x) dx$$

- *varianssi* σ_X^2 kuvaa jakauman leveyttä:

$$\sigma_X^2 = E(X - EX)^2$$

- *hajonta* eli standardipoikkeama:

$$\sigma = \sqrt{\sigma_X^2}$$

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 $\mu = \bar{x} \pm \varepsilon$
- jos ε on “pieni”, $\bar{x}$ on hyvä estimaatti μ :lle

Otostunnusluvut estimaattoreina

■ *harhattomia* ja *tarkentuvia* estimaattoreita

- harhattomalle (*unbiased*) estimaattorille $E\hat{\theta} = \theta$
 - $\hat{\theta}$ on parametrin θ estimaatti

⇒ tuottaa "keskimäärin" oikean tuloksen

- tarkentuvalle (*consistent*) estimaattorille

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| > \varepsilon\} = 0$$

⇒ otoksen kokoa kasvattamalla päästään mielivaltaisen lähelle parametrin todellista arvoa

Arvion tarkkuus: luottamusväli

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

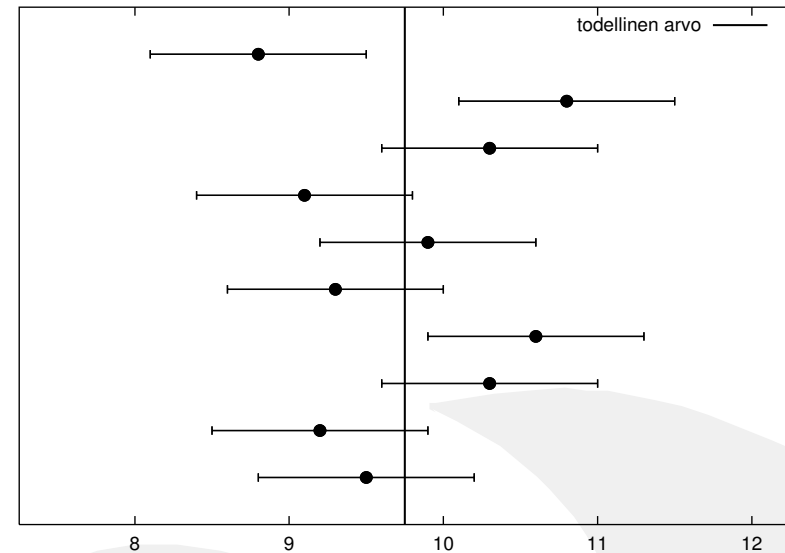
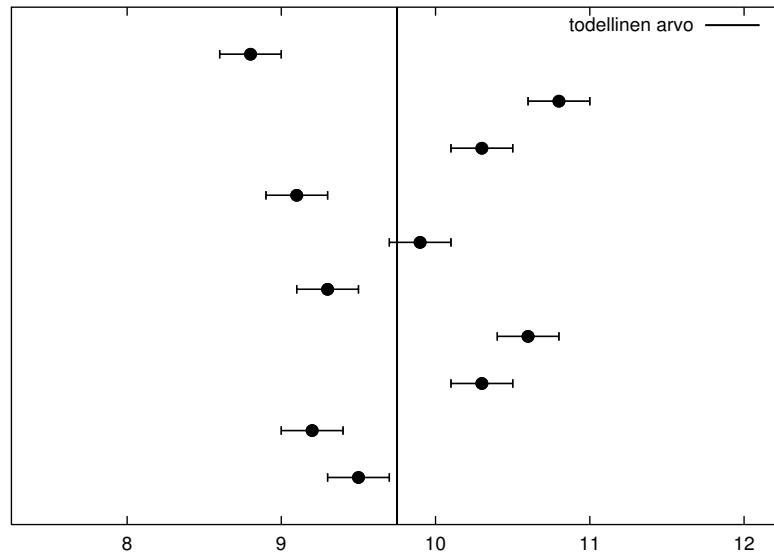
⇒ luottamusväli

- 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 $\bar{x} \pm \varepsilon$, kun ε on jokin kiinnitetty arvo?
- oletetaan että tunnemme populaation odotusarvon, olkoon se $\mu = 9.75$
- kerätään 10 uutta otosta
- asetetaan ensin $\varepsilon = \varepsilon_1 = 0.1$, sitten $\varepsilon = \varepsilon_2 = 0.6$

Luottamusväli, esimerkki



- $\varepsilon = \varepsilon_1 = 0.1 \Rightarrow$ vain yhdessä tapauksessa arvio on välillä $[\mu - \varepsilon, \mu + \varepsilon] \Rightarrow \mu \pm \varepsilon$ on noin 10 %:n luottamusväli
- $\varepsilon = \varepsilon_2 = 0.6 \Rightarrow$ yhdeksässä tapauksessa arvio on välillä $[\mu - \varepsilon, \mu + \varepsilon] \Rightarrow \mu \pm \varepsilon$ on noin 90 %:n luottamusväli

Luottamusväli bootstrapilla

- Olkoon S otos ja $\hat{\theta}$ otostunnusluku, jota käytetään estimaattina parametrille θ
- 1. muodosta K kappaletta bootstrap-otoksia S :stä ja laske niistä pseudo-tunnusluku θ_i^* , $1 \leq i \leq K$
- 2. X prosentin luottamusväli on $\theta_{(i)}^*$:n $X\%$ -fraktiili, eli
 - lajittele pseudo-tunnusluvut
 - $X = 100(1 - 2\alpha)$ -prosentin luottamusvälin approksimaation alaraja on lajiteltujen pseudo-tunnuslukujen $K\alpha$:s arvo $\theta_{(K\alpha)}^*$ ja yläraja $K(1 - \alpha)$:s arvo $\theta_{K(1-\alpha)}^*$

Esimerkki

- alla alkuperäinen otos (korrelaatiokerroin -0.552), sen 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
id	1	2	3	4	5	6	7	8	9	10

id	1	2	7	1	8	6	3	5	9	1
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

Luottamusvälien muodostaminen

- luottamusvälien muodostaminen ilman bootstrapia 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 $\bar{x} = 10.0 = \hat{\mu}$
- 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 N
 - otosjakauman varianssista
 - estimaattorista
 - esim. bootstrap-luottamusväli vs. normaaliusoletus
- VS. ...

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)

Yhteenveto

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