

# Biostatistiikkatiivistelmä

Sekalaisista matskuista koonnut Minna Pöntinen

## Peruskäsitteitä

- Perusjoukko on joukko joka on tutkimuksen kohteena
- Tilastoyksikkö on perusjoukon yksilö, jonka ominaisuuksia havainnot tai mittaukset koskevat
- Muuttujat kuvaavat perusjoukon tilastoyksiköiden niitä ominaisuuksia, joista ollaan kiinnostuneita
  - Diskreetti muuttuja voi saada vain äärellisen määrän erillisiä arvoja (tai numeroituvasti äärettömän) (esim. sukupuoli, lasten määrä)
  - Jatkuva muuttuja voi saada mitä tahansa arvoja tietyltä väliltä (esim. lämpötila, pituus)
  - Selittävä ja selitettävä muuttuja: selitettävän muuttujan vaihtelua pyritään selittämään selittävien muuttujien avulla
- Otos = perusjoukon osa, joka valitaan jollakin otantamenetelmällä
- Näyte = mielivaltainen perusjoukon osa, jota ei ole valittu todennäköisyysotannalla
- Parametri = mallin eri tekijöiden vaikutuksia kuvaava suure tai muuttujaan liittyvä tunnusluku, jonka arvo tutkimuksessa pyritään selvittämään

## Tutkimustyyppejä

- Kokonaistutkimus: hankitaan jokaisesta perusjoukon yksilöstä tiedot kiinnostuksen kohteena olevasta muuttujasta (esim. tilastollisia testejä ei tarvita)
- Otantatutkimus: jos perusjoukko on liian suuri tai ääretön niin tutkitaan osaa perusjoukosta
- Kokeellinen tutkimus: pyritään selvittämään "käsittelyjen" vaikutusta kiinnostuksen kohteena olevaan muuttujaan
- Ei-kokeellinen tutkimus: tutkija ei voi suorittaa jakoa vertailuryhmiin, eikä voi vaikuttaa siihen, esiintyykö selittävä tekijä tietyillä yksilöillä ja miten selittävät tekijät ovat jakaantuneet

## Mittaaminen

- Mittausvirheet: systemaattinen virhe (esim. nopeusmittari); satunnainen virhe (esim. inhimilliset erheet tai mittausvälineen tai -menetelmän huonous)
- Mittarin reliabiliteetin määrittävät satunnaiset mittausvirheet
- Mittarin validiteetti on mitta sille, mittaako mittari todella sitä mitä on tarkoitus mitata

## Mitta-asteikot

- Luokitteluasteikko: havaintoaineiston jokainen havainto kuuluu vain yhteen luokkaan (esim. siviilisääty)
- Järjestysasteikko: muuttujien arvojen välillä ei voida suorittaa aritmeettisiä laskutoimituksia, eikä järjestyslukujen välinen ero kuvaa muuttujan suuruuseroja (esim. sotilasarvo, vamman vaikeusaste)
- Välimatka-asteikko: muuttujalla on aina jokin mittayksikkö; muuttujan arvojen välillä on järjestys ja arvojen erotusten suuruudet ovat mielekkäästi tulkittavissa ja järjestettävissä (esim. lämpötila celsiusasteikolla, syntymävuosi)
- Suhdeasteikko: muuttuja toteuttaa välimatka-asteikon ehdot ja sillä on absoluuttinen nollapiste jossa mitattava ominaisuus häviää (esim. lämpötila Kelvin-asteikolla, pituus, paino)
- Luokittelu- ja järjestysasteikko ovat kvalitatiivisia suureita, välimatka- ja suhdeasteikko ovat kvantitatiivisia

## Tutkimuksen vaiheet

- Poimitaan perusjoukosta sopivankokoinen otos tai näyte
- Satunnaistetaan koeyksiköt käsittelyryhmiin, jonka jälkeen kunkin ryhmän koeyksilöille sovelletaan vain yhtä ja samaa käsittelyä
- Mitataan vastemuuttujan arvot ja analysoidaan tulokset

### Koejärjestelytyyppejä

- Täydellisesti satunnaistettu koe (koeyksilöt jaetaan koetekijän tasoille täysin satunnaisesti)
- Tekijäkoe (koe, jossa käsittelet ovat kahden tai useamman tekijän yhdistelmiä; voidaan tutkia useamman tekijän vaikutusta samanaikaisesti, ja saadaan tietoa myös yhdysvaikutuksista)

### Otantamenetelmät

- Yksinkertainen satunnaisosanta

- 1) numeroidaan perusjoukon tilastoyksiköt luvuilla  $1, 2, \dots, N$
- 2) poimitaan  $n$  lukua satunnaisesti joukosta  $\{1, N\}$
- 3) valitaan otokseen saatujen satunnaislukujen osoittamat yksilöt

- Systemaattinen otanta: käyttökelpoinen kun perusjoukko on järjestetty jonkin ominaisuuden suhteen ja järjestysominaisuus ei vaikuta tutkittaviin muuttujiin

- 1) Määrätään poimintaväli  $k = [N/n]$
- 2) Valitaan satunnaisesti luku  $p$  joukosta  $\{1, \dots, K\}$
- 3) Ensimmäinen otokseen tuleva tilastoyksikkö on luvun  $p$  määräämä  $a_p$ . Seuraavat mukaan tulevat tilastoyksiköt ovat  $p+k, p+2k, \dots, p+(n-1)k$ .

- Ositettu otanta: käyttökelpoinen kun perusjoukko on hyvin heterogeeninen jonkin tutkimuksen kannalta merkityksellisen tekijän suhteen

- 1) Bla bla blaa.
- 2) Ositteesta poimitaan erillinen otos satunnais- tai systemaattisella otannalla.

- Ryväotanta: hyödynnetään perusjoukon jakautumista luonnollisiin osajoukkoihin eli ryppäisiin

- 1) Suoritetaan otanta ryppäiden joukosta jollakin yllämainituista tavoista
- 2) Suoritetaan mittaukset kunkin ryppään sisällä tai jokaisen poimitun ryppään sisältä poimitaan uusi satunnaisotos. (Tällöin menetelmää kutsutaan kaksivaiheisotannaksi.)

### Yksiulotteiset empiiriset jakaumat

- Frekvenssi on tietyn havaintoarvojen esiintymiskertojen lukumäärä tilastoaineistossa
- Summafrekvenssi kertoo niiden havaintojen lukumäärän, jotka ovat pienempiä tai yhtäsuuria kuin ko. luokan todellinen yläraja. Vastaavasti suhteellinen summafrekvenssi kertoo näiden havaintojen suhteellisen osuuden kaikista havainnoista
- Viiksilaatikko on graafinen tiivistelmä aineistosta, joka perustuu tunnuslukuihin  $\min, Q_1, M_d, Q_1$  ja  $\max$ . Se kuvaa tiiviissä muodossa sijainnin, hajonnan ja vinouden.

### Tunnusluvut

- Keskiluvut pyrkivät kuvaamaan tarkasteltavan muuttujan arvojen suuruutta
- Fraktiilit antavat informaatiota jakaumaan liittyvästä hajonnasta ja jakauman vinoudesta
- Moodi = muuttujan se arvo jolla on suurin frekvenssi muuttujan frekvenssijakaumassa
- Mediaani = sellainen muuttujan arvo, jota pienempiä tai yhtä suuria arvoja on 50%
- Entropia: luokitteluasteikollisen muuttujan hajontaa kuvataan tällä (satunnaisasteella)
- Vaihteluväli = pienimmän ja suurimman havaintoarvon väli
- Kvartiiliväli = väli ( $Q_1, Q_3$ )
- Kvartiilipoikkeama =  $Q = \frac{1}{2}(Q_3 - Q_1)$

### Hypoteesit

Nollahypoteesi ( $H_0$ ) on väite/oletus joka koskee perusjoukon jakauman parametrejä.

Nollahypoteesista pidetään kiinni, mikäli havainnot eivät todista nollahypoteesia vastaan kyllin voimakkaasti. Vaihtoehtoinen hypoteesi ( $H_1$ ) on oletus joka astuu voimaan silloin jos  $H_0$  hylätään

### Merkitsevyystaso

Merkitsevyystaso  $\alpha$  = sen todennäköisyys että testattavana olevan suureen havainnoista tietty, määrätty arvo joutuu hylkäysalueelle mikäli  $H_0$  pätee; jos havainnoista määrätty arvo joutuu ( $H_0$ :n pätiessä) hylkäysalueelle,  $H_0$  on virheellinen. Alfaksi valitaan pieniä lukuja (tavalliset merkitsevyystasot ovat 0,05, 0,01 ja 0,001).

### Peruskäsitteitä

- Parametri: perusjoukkoa kuvaava arvo
- Estimaattori: tunnusluku, jota käyttäen saadaan otoksesta arvio parametrille esim. otoskeskiarvo on estimaattori keskiarvolle perusjoukossa
- Estimaatti: otoksesta saatu arvio parametrille
- Keskivirhe: estimaattorin keskihajonta kaikissa mahdollisissa n:n suuruisissa otoksissa kuinka paljon n:n suuruisista otoksista lasketut estimaatit keskimäärin vaihtelevat oikean arvon ympärillä käytetään estimaatin luottamusvälin laskentaan

### Luottamustaso ja luottamusväli

- Luottamustaso = todennäköisyys sille, että luottamusväli sisältää parametrin yhteiskuntatieteellisessä tutkimuksessa useimmin käytetty luottamustaso on 95 prosenttia, muita yleisiä ovat 99 ja 99,9 %
- Luottamusväli = otoksesta laskettu arvoväli, joka sisältää valitun luottamustason ilmoittamalla todennäköisyydellä parametrin esim. jos käytetään 95 prosentin luottamustasoa, niin otettaessa samasta perusjoukosta suuri määrä samansuuruisia riippumattomia otoksia luottamusväli sisältää 95 prosentissa otoksista parametrin mitä suurempi luottamustaso, sitä suurempi luottamusväli
- Luottamusväli liittyy ainoastaan otannan aiheuttamaan epävarmuuteen!

### Tilastolliset testit

- Testisuure esim. ero muuttujan keskiarvossa kahden ryhmän välillä (t-testi)
- Nollahypoteesi ja vastahypoteesi  
nollahypoteesi on yleensä muotoa: testisuure = 0  
esim. "keskiarvoissa ei ole eroa" tai "muuttujien välillä ei ole tilastollista yhteyttä"  
tällöin vastahypoteesina on yleensä: testisuure  $\neq$  0
- Tilastollisen testin lopputuloksena joko hylätään nollahypoteesi tai jos sitä ei riittävän suurella varmuudella voida hylätä, se jää voimaan
- Testien tulokset kertovat tilastollisista yhteyksistä ja eroista, eivät kausaalisista suhteista!
- Merkitsevyytaso / riskitaso: kuinka suuri riski ollaan valmiita ottamaan sille, että hylätään nollahypoteesi, vaikka se todellisuudessa pitäisi paikkansa
- 95 prosentin luottamustaso = 5 prosentin merkitsevyytaso

## Biostatistiikan kaavat

Kaava	Merkitys
$S = \sqrt{\frac{pq}{n}}$	<p>S = tunnusluvun keskivirhe = otantajakauman keskihajonta                      p = positiiviset vastaukset prosentteina                      q = negatiiviset vastaukset prosentteina                      n = otoskoko</p>
$S = \frac{s}{\sqrt{n}}$	<p>S = keskiarvon keskivirhe                      s = otoksesta laskettu muuttujan keskihajonta                      n = otoskoko</p>
<p>p = ka ± z * kv</p> <p>p<sub>95</sub> -&gt; z = 1.96                      p<sub>99</sub> -&gt; z = 2.58                      p<sub>99,9</sub> -&gt; z = 3.29</p>	<p>p = luottamusväli                      ka = keskiarvo                      kv = keskivirhe                      z = kerroin</p> <p>z luetaan x<sup>2</sup>-jakaumataulukosta mikäli otoskoko on alle 30;                      tällöin vapausaste f = n - 1</p>
$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$	<p>s = σ = keskihajonta                      n = lukujen määrä  <math>\bar{x}</math> = lukujen keskiarvo</p>
$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	<p>s<sup>2</sup> = σ<sup>2</sup> = varianssi                      n = lukujen määrä  <math>\bar{x}</math> = lukujen keskiarvo</p>
$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}$	<p>r = korrelaatiokerroin                      n = lukuparien x<sub>i</sub>, y<sub>i</sub> lukumäärä                      s<sub>x</sub>, s<sub>y</sub> = muuttujien x ja y keskihajonnat                      x ja y yläviivoilla = muuttujien x ja y keskiarvot</p>
$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$	<p>C = kontingenssikerroin                      X<sup>2</sup> = khii toiseen -testi                      N = havaintojen lukumäärä</p>
<p>Sk = [3(<math>\bar{x}</math> - Md)]/s</p> <p>Sk = 0 -&gt; symm. jakauma;                      Sk &lt; 0 -&gt; vasemmalle vino;                      Sk &gt; 0 -&gt; oikealle vino</p>	<p>Sk = Pearsonin vinousmitta  <math>\bar{x}</math> = lukujen keskiarvo                      Md = mdiaani                      s = keskihajonta</p>

r<sup>2</sup> \* 100% = selityssaste