

# TILM 3701 - Tilastotiede ja Data 2022

Koonneet      Henri Nyberg<sup>1</sup>      Roope Rihtamo<sup>2</sup>

2022-08-25

<sup>1</sup>Turun Yliopisto, matematiikan ja tilastotieteen laitos, henri.nyberg@utu.fi

<sup>2</sup>Turun Yliopisto, matematiikan ja tilastotieteen laitos, roope.rihtamo@utu.fi



# Sisällys

<b>Kurssin rakenne</b>	<b>7</b>
<b>1 Johdantoa ja johdattelua tilastotieteeseen</b>	<b>9</b>
1.1 Tilastotiede ja kurssin idea . . . . .	9
1.2 Tilastotieteen asema tutkimusyhteisön ulkopuolella . . . . .	11
1.3 Kurssin luonne tilastotieteen opintojen esittelijänä . . . . .	12
<b>2 Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa</b>	<b>13</b>
2.1 Mitä on tiede? . . . . .	13
2.2 Tieteellinen menetelmä . . . . .	17
2.3 Tilastojen yleisestä roolista yhteiskunnassa . . . . .	20
2.4 Mitä on tutkimus? . . . . .	23
2.5 Tieteellisen tutkimuksen vaiheet ja tulosten julkaiseminen . . . . .	25
<b>3 Tilastotiede tieteenalana</b>	<b>27</b>
3.1 Lisää tilastotieteen perustermejä . . . . .	27
3.2 Mitä tilastotiede on ja mitä se ei ole? . . . . .	29
3.3 Tilastotieteen suhde lähitieteisiin . . . . .	33
3.4 Tilastotieteen osa-alueet . . . . .	36
<b>4 Sattuma ja satunnaisuus</b>	<b>45</b>
4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä . . . . .	46
4.2 Tilastotieteen suhde satunnaisuuteen ja todennäköisyksiin . . . . .	49
4.3 Tilastolliset mallit, jakaumat ja parametrit . . . . .	51

4.4	Odotusarvo ja varianssi . . . . .	54
4.5	Joitain jakaumia . . . . .	55
4.6	Sattuman rooli tieteenteossa: Vale-emäviale-tilasto? . . . . .	60
<b>5</b>	<b>Tilastolliset aineistot, niiden kerääminen ja mittaaminen</b>	<b>63</b>
5.1	Kertausta: Data eli aineisto . . . . .	64
5.2	Otannan idea . . . . .	67
5.3	Tilastollisten muuttujien mittaaminen ja mitta-asteikot . . . . .	70
5.4	Kontrolloidut kokeet ja suorat havainnot . . . . .	75
5.5	Otantamenetelmät . . . . .	78
5.6	Otantaesimerkkejä . . . . .	86
5.7	Otannan haasteita vielä kootusti . . . . .	87
<b>6</b>	<b>Otokset ja otosjakaumat: tilastollisen päätelyn näkökulma</b>	<b>89</b>
6.1	Satunnaisotos, yhteisjakauma ja tilastollinen malli . . . . .	89
6.2	Otosjakauma: Estimaattori ja estimaatti . . . . .	91
6.3	Otoskeskiarvo ja otosvarianssi (estimaattoreinta) . . . . .	94
6.4	Suhteellisen frekvenssin otosjakauma . . . . .	97
6.5	Muita tunnuslukuja . . . . .	99
6.6	Luottamusvälit . . . . .	100
6.7	Otoskoko . . . . .	106
<b>7</b>	<b>Tilastollinen riippuvuus ja korrelaatio</b>	<b>107</b>
7.1	Muuttujien väliset riippuvuudet tilastollisen tutkimuksen kohteena	107
7.2	Kahden muuttujan havaintoaineiston kuvaaminen . . . . .	109
7.3	Tunnusluvut . . . . .	111
7.4	Satunnaismuuttujien kovarianssi ja korrelaatio . . . . .	113
<b>8</b>	<b>Regressioanalyysi</b>	<b>121</b>
8.1	Johdatus regressioanalyysin ideaan . . . . .	121
8.2	Yhden selittäjän lineaarinen regressiomalli . . . . .	123
8.3	Muita regressiomalleja . . . . .	130

<b>SISÄLLYS</b>	<b>5</b>
<b>9 Tilastotieteen rooli uuden tiedon tuottamisessa</b>	<b>131</b>
9.1 Tilastollisen tutkimuksen yhteisiä elementtejä . . . . .	131
9.2 Tutkimusprosessi . . . . .	134
<b>10 Aineisto- ja tutkimustyyppit ja koeasetelmat</b>	<b>139</b>
10.1 Tutkimustyyppit . . . . .	140
10.2 Tutkimusstrategiat . . . . .	146
10.3 Eriisia aineistoja ja aineistolähteitä . . . . .	155
<b>11 Tilastollisesta ennustamisesta</b>	<b>169</b>
11.1 Tilastollinen selittäminen vs. ennustaminen . . . . .	169
11.2 Tilastolliseen ennustamiseen liittyviä huomioita . . . . .	171
<b>12 Tilastotieteen kehityksen nykytrendejä</b>	<b>175</b>



# Kurssin rakenne

- Tällä kurssilla tarkoituksena on melko yleisellä tasolla johdatella tilastotieteen ja aineistojen (datan) maailmaan pohtimalla myös näiden laajempia merkityksiä tieteellisen tutkimuksen hyvin keskeisinä osina.
- Kurssilla vältetään, mahdollisuksien mukaan, kovin teknistä matemaattista esitystapaa, mutta tarvittavissa määrin tullaan myös käyttämään tilastotieteen perusopinnoissa tarvittavia matemaattisia merkintöjä ja määritelmiä. Esim. todennäköisyyslaskennan ja tilastollisen päättelyn perusteita ei käydä vielä riittävällä matemaattisella tarkkuudella lävitse, vaan nämä tarkastelut jäävät tästä kurssia seuraavien kurssien (TILM3553 Todennäköisyyslaskennan peruskurssi tai TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille sekä TILM3555 Tilastollisen päättelyn peruskurssi) asiaksi. Nämä kurssit, yhdessä alkuvaiheen pakollisten matematiikan kurssin lisäksi, muodostavat siis tämän kurssin johdannon kanssa lähtökohdan tilastotieteen opinnoille.
- Luennot eivät suoraan perustu yhteen kirjaan tai lähteeseen. Käytettyjä lähdemateriaaleja luetellaan alapuolella oheislukemiston myötä.
- Oheislukemistoa (sopivilta osin):
  - Mellin, I. (2004). Johdatus tilastotieteesseen: Tilastotieteen johdantokurssi (1.kirja). Yliopistopaino, Helsingin yliopisto.
  - Mellin, I. (2000). Johdatus tilastotieteesseen: Tilastotieteen jatkokurssi (2.kirja). Yliopistopaino, Helsingin yliopisto.
  - Mellin, I. (2006). Tilastolliset menetelmät. Luentomoniste, Aalto yliopisto (TKK).
  - Holopainen, M. ja P. Pulkkinen (2008). Tilastolliset menetelmät. Sannoma Pro Oy.
  - Pahkinen, E. ja R. Lehtonen (1989). Otanta-asetelmat ja tilastollinen analyysi. Gaudeamus, Helsinki.
  - Pahkinen, E. ja R. Lehtonen (2004). Practical Methods for Design and Analysis of Complex Surveys. 2. painos, Wiley.
  - Sund, R. (2003). Tilastotiede käytännön tutkimuksessa -kurssi. Helsingin yliopisto.

- Silver, N. (2014). Signaali ja kohina: Miksi monet ennusteet epäonnistuvat mutta jotkin eivät? Terra Cognita. (Suomentanut Kimmo Pietiläinen)
    - \* Englanninkielinen teos: Silver, N. (2015). *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*. Penguin Books; Illustrated edition
  - Pesonen, M. (2017). Kurssimateriaali kurssille Aineistonhankinta ja tutkimusasetelmat, Turun yliopisto.
  - Vartia, Y. (1989). Tilastotieteen perusteet. Yliopistopaino, Helsinki. II painos.
- Muita taustamateriaaleja
    - Tilastokeskuksen tilastokoulu ([linkki](#))
    - Tilastotieteen sanasto suomi-englanti-suomi, ks. Juha Alho, Elja Arjas, Esa Läärä ja Pekka Pere (2021). Tilastotieteen sanasto. Suomen Tilastoseuran julkaisuja 8.

Suuret kiitokset Visa Kuntzelle ja Emil Lehdelle kommentteista ja avusta materiaalin työstämisessä. Kaikki jäljelle jääneet painovirheet ovat materiaalin koajien.

# Luku 1

## Johdantoa ja johdattelua tilastotieteeseen

*Ihmisellä on luontainen pyrkimys ymmärtää, mitä hänen ympärillään tapahtuu. Ymmärrys perustuu ihmisen tekemiin havaintoihin, joita luokittelemalla tai seuraamalla hän pyrkii löytämään säännönmukaisuuksia. Näiden säännönmukaisuuksien löytäminen vaatii loogisten johtopäätösten tekoa. Pelkän uteliaisuuden tyydyttämiseen ja älyllisen mielihyvän lisäksi ihminen pyrkii ennakoimaan tulevaa ja siten varautumaan tuleviin tapahtumiin... Edellä kuvattuja taitoja voi oppia.*

Holopainen ja Pulkkinen, 2008

### 1.1 Tilastotiede ja kurssin idea

- Tämän tilastotieteen ensimmäisen kurssin ideana on (ainakin)
  - Esitellä ja johdatella **tilastolliseen ja tieteelliseen ajatteluun** ja sen hyödyntämiseen eri tyypillisissä tutkimusongelmissa.
  - Esitellä tilastotieteen roolia **empiriisen tutkimusaineiston keräämisessä ja analyysissä** sekä tarkastella tieteentekemisen ja tilastotieteen suhdetta.
  - Pohtia **tilastotieteen olemusta tieteenalana** ja tarkastella tilastotieteen ja datatieteiden (data science) samankaltaisuksia ja eroja.
  - Pohtia **sattuman ja satunnaisuuden roolia** jokapäiväisessä elämässä ja erityisesti osana tieteellistä tutkimusprosessia.
  - Oppia tilastotieteen peruskäsitteitä ja (tilastollisen) tutkimuksenteon alkeita ja siihen liittyviä mahdollisia ongelmia esimerkiksi tilastollisten aineistojen keräämisessä.

## 10 LUKU 1. JOHDANTOA JA JOHDAATELUA TILASTOTIETEESEEN

- Oppia tilastollisten aineistojen **kuvaamisen ja käsittelyn** alkeita sekä tilasto(tieteellisen)llisen **mallintamisen ja koeasetelmien** peruskäsitteitä.
  
- Kurssilla käsitellään myös **tilastollisen päättelyn** peruskäsitteitä ja perusteita kuten
  - Mitä on **todennäköisyys** ja miten sen tulkitaan tilastotieteessä sekä laajemmin tieteessä. Erityisesti tilastotieteen osalta keskiössä on tämän kurssin osalta **satunnaismuuttujat** sekä niihin liitettävät käsitteet
    - \* **Odotusarvo, varianssi** ja kahden (tai useamman) satunnaismuuttujan **korrelatio**.
    - \* Satunnaismuuttujien **todennäköisyysjakaumien** perusteita ja niiden yhteyksiä mm. normaalijakaumaan ja muutamiin muihin keskeisiin jakaumiin.
    - \* Tilastollinen malli työkaluna satunnaismuuttujien formaalisissa mallintamisessa ja päättelyssä. Tilastollisen malliin liittyy (usein) **parametreja** joihin tilastollinen päättely kohdistuu.
    - \* Tilastollisten mallien **estimoinnin** perusidea, eli miten tilastollisen mallin parametreille muodostetaan arvot käytettäväissä olevan aineiston pohjalta. Esimerkiksi: mitä tarkoittaa tilastollisen mallin parametrin **estimaattori** ja sen **harhattomuus**?
    - \* Alustavia tarkasteluja tilastollisen mallin uskottavuuden käsitteelle ja **luottamusväleille** tilastollisen mallin estimoidulle parametreille.
  
- Toinen kurssin keskeisistä teemoista on tarkastella tieteellistä tutkimusprosessia teoriassa ja käytännössä. Tämä sisältää mm. seuraavia aiheita (joita siis käsitellään tällä kurssilla päällisin puolin ja varsin yleisestä näkökulmasta katsoen): tarkemmat yksityiskohdat jäävät tästä kurssia seuraavien tilastotieteen kurssien aihepiireiksi):
  - **Tutkimusongelman** asettaminen: mitä halutaan tutkia?
  - Tutkimusongelman täsmantäminen ja **tutkimusstrategian** laatiminen: millä keinoin asetettuun tutkimusongelmaan voidaan vastata?
  - **Tutkimusaineiston** (tai vain lyhyemmin **aineiston** eli **datan**) kerääminen
    - \* **Aineiston ennakkoehdot**: mitkä ehdot tulee täytyy, jotta asetettuun tutkimusongelmaan voidaan vastata?

## 1.2. TILASTOTIETEEN ASEMA TUTKIMUSYHTEISÖN ULKOPUOLELLA11

- \* **Otanta** (ja mittaaminen): miten tutkimusaineisto kerätään niin, että se täyttää aineiston ennakkoehdot? Erilaisissa tutkimuksissa käytetään erilaisia aineistoja kuten:
  - Survey- ja rekisteriaineistot
  - Havaintoarvojen välistä korrelatiota esiintyy mm. aikasarja-aineistojen tai pitkittäisaineistojen tapauksessa
- **Aineiston kuvaaminen:** minkälaisista aineistoa on kerätty ja vastaako se ennakkoehtoja?
- **Aineiston analyysin** lähtökohtia
  - Mitä tilastollista mallia/malleja käytetään?
  - Mitä tarkoitetaan mallien tuntemattomien parametrien arvojen estimoinnilla?
  - Tilastollinen päättely (estimointitulosten pohjalta)
- **Johtopäätelmien** tekeminen tilastollisen päättelyn pohjalta: saatiinko tutkimusongelmaan vastaus ja kuinka luotettava saatua vastaus on?

## 1.2 Tilastotieteen asema tutkimusyhteisön ulkopuolella

- Tilastotiede on oppiaineena usein varsin tuntematon toisen asteen opinnoista valmistuneelle, sillä sitä ei juurikaan opeteta lukioissa tai ammatti-kouluissa huolimatta sen keskeisestä ja kasvavasta roolista tiedemaailman kentillä.
- Tiedeyhteisön ulkopuolellakin **tilastotiedettä ja tilastotieteilijöitä arvostetaan laajalti**.
- **Tilastotiede onkin nostanut profiliaan viimeisten vuosikymmenien aikana** tietoteknisen kehityksen tuotua laajat tietoaineistot ja kehittyneet laskennalliset menetelmät lähes jokaisen kansalaisen saataville.
- Tämä “datavallankumous” näkyy tilastotieteilijöiden kysynnässä työmarkkinoilla: erilaisten aineistojen määrään lisääntyessä kasvaa myös kysyntä työntekijöistä, jotka osaavat ammatitaitoisen käsittellä, tulkita ja mallintaa tilastollisia aineistoja.
- Ei siis liene ihmekään, että erilaisten “data”-alkuisten työpaikkojen, kuten **datatieteilijä** (eng. **data scientist**) tai **data-analytikko** (**data-analyst**) määrä on kasvanut voimakkaasti jo pidempään. Kaikkia tieto- ja datainensiivisten ammattien tekijötä yhdistää yksi tekijä: **heidän tulee hallita ja osata tilastotiedettä!** Karkeistettuna mitä paremmin ja enemmän (laajemmin), sen parempi palkka ja monipuolisemmat työtehätävät!

### 1.3 Kurssin luonne tilastotieteen opintojen esiteltijänä

Kurssin mittaan esitellään tilastotieteen perusteiden lisäksi **miten TY:ssa tilastotieteen opinnoissa syvennytään** tällä kurssilla esiteltäviin menetelmiin, aineistotyypeihin ja mallinnuskokonaisuuksiin.

## Luku 2

# Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa

Tässä luvussa tarkastellaan tieteen ja tieteellisen tutkimusprosessin luonnetta erityisesti uuden **tutkitun** tiedon tuottamisen näkökulmasta. **Tiedelukutaidon** merkitys on kasvanut nyky-yhteiskunnassa, kun tiedejulkaisujen saavutettavuus ja tunnettiuus on lisääntynyt mm. tieteen popularisoinnin ja median laajemman tiedeuitisoinnin vuoksi. Tiedon, erityisesti tieteellisen tiedon, rooli korostuu yhä enemmän myös kaikilla elämän osa-alueilla: terveysteknologia (esim. sykemittarit tai Oura-sormus) perustuu lääke- ja terveystieteisiin läpimurtoihin, talouspoliittisia päätöksia edeltää entistä suurempi määrä asiantuntijoiden taloustiedeperusteista analyysia ja jopa peruskouluopetus on murroksessa kasvatustieteen saavutusten myötä. Voidakseen ymmärtää ja arvioida kriittisesti tiedeuitisia tulee lukijan olla tietoinen tieteellisen tutkimuksen luonteesta: miten tutkimusartikkeleja luetaan, mitä niiltä voidaan odottaa ja minkälaiset tulokset ovat uskottavia. **Tilastotiede näyttelee keskeistä roolia lähes kaikkessa tutkimuksessa ja erityisesti erilaisten tutkimuskysymysten ja niitä vastaavien hypoteesien testauksessa.** Aloitetaan kurssin varsinainen oppimateriaalia kunnianhimoisesti tarkastelemalla, että mitä tiede oikeastaan on.

### 2.1 Mitä on tiede?

- Annetaan tieteen määritelmälle ensin muutamia pohtivia suuntaviivoja:
  - *Tiede on järjestelmällistä ja järkiperäistä uuden tiedon hankintaan.*<sup>1</sup> Tiede (voidaan) siis ymmärtää toiminnaksi, jossa tavoitellaan

---

<sup>1</sup>Haaparanta ja Niiniluoto (1986). Johdatus tieteelliseen ajatteluun. Filosofian laitoksen julkaisuja 3/86. Helsingin yliopisto.

## 14LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

ja hankitaan **tietoa**.

- Tieteellinen tutkimus on tutkivan subjektiin ja tutkimusobjektiin välistä vuorovaikutusta.
  - Tiede pyrkii järjestämään tiedon yksinkertaisiksi kokonaisuksiksi ja pyrkii löytämään säännönmukaisuuksia.
- 
- Tiede on siis tiedon hankintaa, jonka kohteena on meitä ympäröivä todellinen maailma sen ilmiöineen ja tapahtumineen.
    - Tiedon hankinnalla tarkoitetaan kumulatiivista prosessia, jossa ympäröivän maailman ilmiötä ja niiden välisiä suhteita
      - \* i) selitetään,
      - \* ii) niitä koskevia käsityksiä vahvistetaan osoittamalla ne tosiksi sekä
      - \* iii) löydetään niistä uutta tietoa.
    - Tiede siis erottaa intuition ja ”arkitiedon” oikeasta, tutkitusta tiedosta esittämällä reaalimaailmaa koskevia väitteitä ja osoittamalla ne toteksi tieteellisin menetelmin.
    - Tiede käsitteää myös aiemman tutkimuksen ja se toimii kaiken tieteellisen tiedon jäsenneltyvä kokonaisuutena.
    - Tieteen tekemiseen liittyvä vaatimus **uudesta tiedosta** kuitenkin sulkee tieteen ulkopuolelle toiminnot, joissa on kyse vain aikaisemmin hankittujen tietojen omaksumisesta ja järjestämisestä (vrt. opiskelu, komitea/selvitystyöt).
      - \* Aikaisemmin hankittujen tietojen vahvistaminen ja todentaminen, eli uuden tutkimuksen tekeminen, on kuitenkin tiedettäsen tuottaessa uutta tietoa.
- 
- Tieteelle voidaan asettaa (ainakin) seuraavat kaksi sitä määrittelevää ominaisuutta.
    - **Järjestelmällisyys:** tieteellinen tiedonhankinta on yhteiskunnalliseksi organisoitu tutkimusta (ja opetusta) järjestävien instituutioiden tehtäväksi, joka kokoa tutkimustulokset systemaattisiksi tietojärjestelmiksi niin kansallisella kuin kansainvälisellä tasolla.
      - \* Näihin instituutioihin lukeutuu yliopistot, korkeakoulut ja tutkimuslaitokset ja vastaavasti tietojärjestelmiksi mm. tieteelliset julkaisut.
      - \* Tiede ylittää järjestelmällisyytensä vuoksi tiedostamisen ”arkitason” (vrt. aiemmat pohdinnat arkitiedon ja tieteellisen tiedon välillä).
    - **Järkiperäisyys:** Järkiperäisyyden vaatimus asettaa rajoitteita tieteelliselle ajattelutavalle.

- \* Tiede ei voi nojautua yksilölliseen vaistoon tai intuitioon
- \* Suostutteluun
- \* Propagandaan
- \* “Jumalalliseen ilmoitukseen” tai vastaavaan
  
- Tieteen keskiössä on todellista maailmaa koskevat teoriat ja niihin liitettyt hypoteesit.
  - Teoriat ovat hyvin perusteltuja kuvausia ja väittämää siitä, miten ympäröivä maailmamme toimii tai esimerkiksi siitä miten eri ilmiöt ovat yhteyksissä toisiinsa.
  - Teoriat kehittyvät vuorovaikutuksessa todellisen maailman kanssa kun tieteellisessä tutkimuksessa niitä ja erityisesti niihin liittyviä hypoteeseja testataan ja saatuja tuloksia tulkitaan vallitsevien teorioiden valossa.

### Hypoteesi

- Hypoteesi tarkoittaa (tausta)teorioista johdettua tai aikaisemman tutkimuksen perusteella esitettyä ennakoitua ratkaisua tai selitystä tutkittavaan ongelmaan.
- Hypoteesi ilmaistaan väitteenä, jonka paikkansapitäävyttä halutaan tutkia.
- Kokeelliset tiedot voivat osoittaa hypoteesin vääräksi
- Nollahypoteesi vastaa tavallisesti tyypillistä, odotettavissa olevaa tulosta, esimerkiksi ettei kahden mitatun ilmiön välillä ole yhteyttä tai että tietty hoito on tehotonta.
- Nollahypoteesia ei todisteta (“hyväksytä”), vaan voidaan ainoastaan sanoa, ettei aineisto tarjoa todistusaineistoa (“evidenssiä”) nollahypoteesin hylkäämiselle – ts. sille tulemalle, että emme hylkää nollahypoteesia.
- Vastahypoteesi sisältää usein mielenkiinnon kohteena olevan tapahtuman, kuten “on eroa” tai “on vaikutusta”.
- Tutkijoilla on usein taipumus jättää julkaisematta tutkimustuloksia, joissa nollahypoteesi jää voimaan. Yleensä tämä tilanne syntyy, kun lopputulos ei eroa jo aikaisemmin otaksutusta. (Toki ajoittain tilanne on myös toisinpäin eli “toivotaan” nollahypoteesin hylkäämistä).

- Uuden tieteellisen tiedon tuottaminen ja jo tuotetun tiedon ymmärtäminen vaatii **tieteellisen ajattelutavan** omaksumista, jonka **perustana on lähes aina tilastollinen päättely**.

## 16LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- Tieteelliselle ajattelulle ja tiedon tuottamiselle on tunnusomaista, että se pohtii ja kehittelee **paradigmojaan** eli oman toimintansa perusteita.

**Paradigma** on tietyn alan oman tieteellisen toiminnan oppirakennelma, ajattelutapa ja peruste, joka mm. ohjaa tutkimuskysymysten asettelua, käytettäviä menetelmiä ja tulosten tulkintoja. Paradigmat elävät jatkuvassa muutoksessa tieteen kehityksen myötä. Esimerkkinä toimii taloustieteen nk. “uskottavuusvallankumous”, jossa tilastollisten menetelmien myötä taloustieteellisen tutkimuksen painopiste tuntuu siirtyneen vahvemmin empiirisen kausaalitutkimuksen puolelle.

- Paradigmat siis ohjaavat uuden tieteellisen tiedon tuottamista asettamalla tutkimukselle yhteneväät raamatit, jotka ohjaavat sitä, miten tutkimuskysymisiä asetetaan ja miten niihin etsitään vastauksia sekä myös sitä miten saatuja tuloksia tulkitaan.
  - Tieteellinen tieto perustuu siis tutkimusalan tiedeyhteisön yhteiseen sopimukseen paradigmista ja täten siitä, minkälaisista tutkimusta, ja mistä ilmiöistä, kannattaa tehdä.
  - Paradigmojen ei pidä ajatella olevan kaavoihin kangistuneita ajattelu- ja menettelyapoja, jotka oikeuttavat vain tietynlaisen tutkimuksen tekemisen.
    - \* Päinvastoin, paradigmat ovat ajan myötä kumuloitunutta tietoa siitä, mitkä toimintatavat ja -menetelmät tuottavat uskottavaa, koko tiedeyhteisön hyväksymää tiedettä, joka täyttää hyvän tieteen kriteerit.
    - \* On kuitenkin mahdollista, ja käytännössä varmaa, että vallitsevat paradigmat myös estävät osaltaan uusien löytöjen syntymistä: liian vahvasti alan paradigmoiden kanssa ristiriidassa oleva tulos saattaa jäädä julkaisematta, mikäli tutkija ei pidä sitä lainkaan mahdollisena suhteessa vallitseviin paradigmoihin.
  - Tieteelliseen ajattelutapaan kuuluu olennaisesti juuri tiedon kumuloitumisen ymmärtäminen: yksittäinen vahva tulos on vasta alku ja vahvistettu tieto jostain ilmiöstä, yhteydestä tai vaikutuksesta syntyy monien mittausten ja tutkimusten jatkumona.
  - Tietoa ei siis voida johtaa siitä, miltä asiat näyttävät, kuten on tyypillistä ”arkiajattelussa”.
    - \* Tiede kehittää teorioita kriittisesti ja määriteltioiden rationaalisin ajattelun keinoin.
    - \* Teorioita ja niihin liittäviä hypoteeseja testataan tieteellisin menetelmin ja näin saadaan uutta tietoa tutkittavasta ilmiöstä.
  - Tiivistetysti voidaan sanoa että tiede on kumulatiivinen tutkimusprosessi, jossa hankitaan uutta tietoa ja samalla vahvistetaan vanhaa,

mutta epävarmaa tietoa tieteellisin menetelmin. Tieteellisten menetelmien käyttöä ohjaa tutkimusalakohtaiset paradigmat, jotka ovat suuntaviivoja ja viiteistöjä siitä, minkälainen tutkimus tuottaa uskottavia tuloksia.

### Arkitieto

- ▶ epäluotettavat havainnot
- ▶ epäjohdonmukaisuus
- ▶ omien kokemusten vaikutus
- ▶ logiikan puute
- ▶ lyhytjänteisyys
- ▶ valikoivat havainnot
- ▶ muistamattomuus
- ▶ irrallisuus asiayhdestä
- ▶ tytyminen ensimmäiseen selitykseen
- ▶ liiallinen yleistäminen

### Tieteellinen tieto

- ▶ perustuu tietoiseen opiskeluun, analyysiin ja yleistämiseen (otantateoria)
- ▶ muodostaa hierakkisen järjestelmän
- ▶ objektiivisuus
- ▶ etsii yleisiä lainmukaisuuksia ja periaatteita
- ▶ perusteltua
- ▶ julkista
- ▶ korjaantuvaa
- ▶ kriittisyyss
- ▶ olennaisen ja epäolennaisen erottaminen

Kuva 2.1: Arkitieto ja tieteellinen tieto

## 2.2 Tieteellinen menetelmä

- Milloin tutkimus sitten on tieteellistä? Tiede on tiedonhankintaa, jossa käytetään erityistä, mahdollisesti tilanteesta (sovelluksesta) riippuen, tieteellistä **menetelmää** eli **metodia**.

**Tieteellinen menetelmä:** Tieteellinen menetelmä on kullakin tieteen alalla vallitseva, ajan myötä kehittynyt ja nykyisten paradigmoiden muo-

vaama menettelytapa, jolla uutta tietoa tuotetaan ja vanhaa, mutta epävarmaa tietoa vahvistetaan. Se ei ole selkeä työvaiheiden luettelo tai menetelmähakemisto, vaan yleisesti hyväksytty ja hyväksi todettu tapa pyrkii totuuteen erilaisten tutkimusongelmien ratkaisussa. Hyvälle tieteelliselle menetelmälle voidaan lukea seuraavia kriteerejä.

- **Objektiivisuus ja loogisuus**

- Tutkimuskohteen ominaisuudet ovat tutkijan mielipiteistä riippumattomia.
- Tieteellinen tieto tutkimuskohteesta syntyy tutkijan ja tutkimuskohteen vuorovaikutuksen tuloksena.
- Tiedon lähteenä on tutkimuskohteesta saatava kokemus.
- Tutkimuskohteesta voidaan saada totuudellista tietoa, jonka laadusta myös tutkijayhteisö voi olla yhtä mieltä.

- **Kriittisyyys**

- Ilmenee niinä vaatimuksina, joita **hypoteesin** asettamiselle, testaamiselle ja hyväksymiselle on asetettu.
- Tieteellisten hypoteesien tulee olla intersubjektiivisesti testattavissa eli niillä täytyy olla yhdessä sopivien lisäoleustusten kanssa sellaisia seurauksia, joiden totuus tai virheellisyys voidaan julkisesti tarkistaa.

- **Autonomisuus**

- Tieteen tulosten arvioiminen on (tiukasti ottaen) tieteellisen yhteisön oma asia, johon tieteen ulkopuolella olevat ryhmät eivät saa vaikuttaa.
- Ei ole hyväksyttää vedota siihen, että väitteen totuus olisi toivottavaa tai epätoivottavaa esimerkiksi poliittisista, uskonollisista tai moraalista syistä.

- **Edistyyvyys**

- Tieteen edistyminen merkitsee kasvun eli tulosten määrällisen lisääntymisen ohella sitä, että virheellisiä hypoteeseja tai teorioita korvataan uusilla tuloksilla, jotka ovat toisia tai ainakin vähemmän virheellisiä kuin aikaisemmat.

- **Toistettavuus ja yleistättävyys**

- Tieteen tulokset tulee olla muiden tutkijoiden toistettavissa eli replikoitavissa. Toistettavuudelle (paikoin myös uusittavuudelle, joskin merkitys vaihtelee) on erilaisia määritelmiä.

- Tarkastellaan lähemmin erästä määritelmää erilaisille toistettavuuden la-

jeille. Esittemme tässä Hamermeshin (2007)<sup>2</sup> esittämän erilaisten taloustieteellisten tutkimusten replikointien jaottelun:

- **Puhdas replikointi:** toinen tutkija, käyttäen täysin samaa tutkimusaineistoa ja samaa tilastollista menetelmää kuin alkuperäisessä tutkimuksessa, saa täsmälleen samat tutkimustulokset.
- **Tilastollinen replikointi:** toinen tutkija, käyttäen eri tutkimusaineistoa (joka on kuitenkin poimittu populaatiosta, ks. Luku 5) mutta samaa menetelmää, saa vastaanvalaisia tuloksia, jotka vahvistavat alkuperäisen tutkimuksen perustulokset.
- **Tieteellinen replikointi:** toinen tutkija, käyttäen samoja asioita mittavaa tutkimusaineistoa, joka on kuitenkin kerätty eri populaatiosta, ja käyttäen samankaltaista, mutta ei identtistä menetelmää, saa vastaanvalaisia tuloksia, jotka vahvistavat alkuperäisen tutkimuksen perustulokset.

- Teorioiden sisältämiä väitteitä voidaan muotoilla tieteellisiksi malleiksi, joihin voidaan liittää hypoteeseja, joita testataan tieteellisin menetelmin käyttäen ilmiö(i)stä mitattua havaintoaineistoa.
  - Tieteelliset mallit ovat yksinkertaistuksia reaalimaailmasta ja ne kuvaavat tutkimuksen aihetta jostain näkökulmasta tarkasteltavana systeeminä.
  - Mallit hyödyntävät matemaattista esitystapaa, sillä se tarjoaa formulain ja objektiivisen tutkimusaiheen kuvauksen sekä mahdollistaa siihen liittyvän loogisen päättelyn havaitun, empiirisen aineiston pohjalta.
  - Tilastolliset mallit ovat käytännössä tieteellisten mallien formaaleja matemaattisia esityksiä, jotka lisäksi mahdollistavat mallia koskevan tilastollisen päättelyn esimerkiksi hypoteesien testaamisen avulla. Päättely perustuu tilastotieteen teoriaan, joka mahdollistaa päättelyn epävarman ja satunnaisen aineiston tapauksissa.
  - Hypoteesien testaamisen voidaan ajatella tutkittavaa ilmiötä koskeviksi ennusteiksi, joita verrataan havaittuun aineistoon. Mikäli havaittu aineisto ei sovi testattavaan teoriaan tai siihen liittyviin hypoteeseihin, voidaan teoriaa kehittää paremmaksi. Tämä vuoropuhelu vie tiedettä eteenpäin ja tuottaa lisää tutkittua tietoa ympäröivästä maailmasta.

---

<sup>2</sup>Hamer mesh, D. S. (2007). Replication in economics *Canadian Journal of Economics/Révue canadienne d'économique* 40 (3), 715–733.

- Hypoteesien testaaminen on yhtäältä tieteellisten teorioiden kehittämisen ja vahvistamisen ja toisaalta kritiikin keskiössä.
  - Metodologinen pluralismi: Kaikkia menetelmiä voi soveltaa hyvin tai huonosti, mutta niitä voi käyttää myös luovasti väärin.

## 2.3 Tilastojen yleisestä roolista yhteiskunnassa

- Ihminen ei voi toimia maailmassa järkevästi, ellei hän pysty muodostamaan oikeata kuvaaa maailmasta ja sen tilasta. Nykyäikana oikeaa kuvaaa varten tarvitaan maailmaa ja sen tilaa merkityksellisesti ja oikein kuvaavia, ajantasaisia (**tilasto**)tietoja.
- Yhteiskunnan kaikilla sektoreilla toiminnan seuranta, päätöksenteko ja ennakointi perustuvat eri sektoreita kuvaaviin (**tilasto**)tietoihin ja niiden analysoinnissa käytettäviin **tilastollisiin menetelmiin**.
  - Oikein todellisuutta kuvaavat, ajantasaiset (tilasto)tiedot ovat välttämättömiä modernin yhteiskunnan toiminnalle.
  - Esim. päätöksenteko sekä julkisella että yksityisellä sektorilla (elinkeinoelämässä) perustuu pitkälti yhteiskuntaa ja elinkeinoelämää kuvaaviin (tilasto)tietoihin ja tilastollisten menetelmien tuottamiin tuloksiin sekä niiden perusteella tehtäviin päätöksiin. Esimerkkejä ovat esim. tietyt konkreettiset (talous)poliittiset toimenpiteet (talous)tilastojen perusteella. Lisäksi tuotantoprosessien ohjaus ja laadunvalvonta teollisuudessa sekä markkinatutkimus kaupan alalla perustuvat tilastollisiin menetelmiin.
  - (Tilasto)tietojen saatavuutta voidaan pitää jopa toimivan demokratian edellytyksenä.
- Koska todellisuutta kuvaaviin (tilasto)tietoihin sisältyy (lähes) aina epävarmuutta ja satunnaisuutta, tilastotiede ja tilastolliset menetelmät luovat perustan tilastojen tuotannolle, jalostukselle ja analysoinnille.
  - Niinpä tilastojen tuotannon, jalostuksen ja analysoinnin menetelmien kehittäminen on keskeinen osa tilastotieteen tehtäväkenttää.
  - Samoin tilastotieteen menetelmien ymmärtämällä on keskeinen rooli tietoyhteiskunnassa toimimisessä ja vaikuttamisessa.

**Esimerkki (väite):** Naiset puhuvat enemmän kuin miehet.

- Lähtökohta väitteen (hypoteesin) tutkimiseen:
  - Uskomus on väärä kunnes toisin todistetaan.
  - Lähdetään liikkeelle olettamuksesta, että miehet ja naiset puhuvat yhtä paljon.
  - Olettamuksen tueksi tai kumoamiseksi täytyy kerätä todistusaineistoa
  - Jotta tutkimukseen saataisiin täysin varma vastaus, kaikki miesten ja naisten puheet ihmiskunnan olemassa olon ajalta pitäisi pystyä laskemaan = mahdotonta.
- Mitä siis tehdä?
  - Täytyy tyytyä tutkimaan osajoukkoja miehistä ja naisista (otos), mihin tarvitaan **otantamenetelmiä** (käsitellään tarkemmin myöhemmin luvussa 5).
  - Arvotaan satunnaisesti tutkimushenkilötä miesten ja naisten joukosta ja mitataan kuinka paljon he puhuvat.
  - Satunnaisuus tärkeää, sillä jos valikoitaisiin tarkoitukSELLA PUHELIAITA TAI VÄHÄSANAISIA TUTKIMUSHENKILÖITÄ, TULOKSET VÄÄRISTYISIVÄT.
- Jokaiseen mittaukseen liittyy virhe.
  - Täysin satunnainenkaan otos ei edusta täydellisesti koko väestöä. Joukkoon saattaa valikoitua puhtaasti sattumaltakin poikkeuksellisen puheliaita tai harvasanaisia naisia tai miehiä.
  - Millaisia sekoittavia tekijöitä tulee mieleen? Mitkä seikat voisivat vaikuttaa tutkittavaan asiaan?
  - Tosin mitä suurempi otos, sitä pienemmäksi sattuman osuus käy ja joudutaan turvautumaan todennäköisykyksiin: Kun aineisto on kerätty, halutaan tietää kuinka todennäköistä on, että uskomus pitää paikkaansa.
- Palataan takaisin esimerkkiimme: Yleisen uskomuksen mukaan naiset puhuvat kolme kertaa enemmän kuin miehet.
  - Tutkimuksen mukaan miehet vaikuttavat kuitenkin puhuvan yhtä paljon kuin naisetkin.
  - Laajemmat tutkimukset osoittavat, että tilanteella on puheen määrään paljon suurempi vaikutus kuin sukupuolella.
  - Kiitos tilastotieteen, väärä uskomus on korvautunut tiedolla!

## Are Women Really More Talkative Than Men?

Matthias R. Mehl<sup>1,\*</sup>, Simine Vazire<sup>2</sup>, Nairán Ramírez-Esparza<sup>3</sup>, Richard B. Slatcher<sup>3</sup>, James W. Pennebaker<sup>3</sup>

+ Author Affiliations

\* To whom correspondence should be addressed. E-mail: mehl@email.arizona.edu

Science 06 Jul 2007;  
Vol. 317, Issue 5834, pp. 82  
DOI: 10.1126/science.1139940

### Abstract

Women are generally assumed to be more talkative than men. Data were analyzed from 396 participants who wore a voice recorder that sampled ambient sounds for several days. Participants' daily word use was extrapolated from the number of recorded words. Women and men both spoke about 16,000 words per day.

Kuva 2.2: Are women really more talkative than men?

## 2.4 Mitä on tutkimus?

- Tiede tavoittelee tietoa, mutta mistä?
  - Jokaisen tutkimuksen lähtökohtana on (tai ainakin pitäisi useimmiten olla) tiedollisen uteliaisuuden, käytännön tarpeiden tai teorian kehittämisykyksen herättämä ongelma, johon tutkimuksen avulla etsitään vastausta. Tutkimus yrittää käsittää sekä tulkitun ilmiön, että sen tajunnassa synnyttämät spontaanit mielikuvat tai arkipäivän tiedot.
  - Tutkimus siis pyrkii löytämään täysin uutta tietoa, varmentamaan (mahd. aiempien tutkimusten myötä) syntyneitä vallitsevia mutta epävarmoja käsityksiä sekä tarkistamaan vakiintuneen tiedon paikansapitäävyyttä.
  - Valtaosa tieteestä asemoituu erityisesti kahden viimeisen kohdan alaisuuteen vaikka tieteen popularisoinnissa (mm. median toimesta) usein keskitytäänkin uusiin tiedemaailmaa järisyttäviin löydöksiin, jotka tosin voivat usein olla hyvin epävarmoja!
  - Lisää tieteen popularisoinnista ja jaksossa 4.6.
  
- Millaisia kysymyksiä **tutkimuksessa** asetetaan (voidaan asettaa)?
  - **Kuvaus:** Kuinka suuri on yli 65-vuotiaiden osuus Suomen väestöstä?
  - **Riippuvuuden kuvaus:** Ovatko paljon mainostavat yritykset kannattavampia kuin vähän mainostavat?
  - Kuvattujen ilmiöiden **selittäminen ja ymmärtäminen**. Miksi vanhempien sosioekonominen asema vaikuttaa ekonomien työhönsijoitumiseen? Tämän tutkimuskysymyksen tapauksessa pyrkimys on lännän selittää (ymmärtää) ilmiötä.
  - **Ennustaminen:** Jos kansantulon kasvu pienenee x%, työttömyyden ennustetaan kasvavan y tuhannella.
  - Kohdetta kuvavien käsitteiden ja teorioiden rakentaminen, teorioiden ansioiden ja puutteiden arviointi.
  
- Myöhemmin materiaalissa (luvussa 11) keskustellaan vielä tarkemmin miten tilastotieteessä ilmiön ymmärtäminen (selittäminen) ja ennustaminen eroavat toisistaan.
  
- **Tutkimuksen rajat?** Onko niitä?
  - Tutkimus antaa aina vajavaisen kuvan tutkimuskohteesta.
  - Ymmärtämiseen tarvittava havaintomaailman hahmotus (saattaa) tuottaa ideologisesti ja historiallisesti sitoutuneita yksinkertaistavia sekä luonteeltaan usein hyvin teoreettisia abstraktioita.

## 24LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- On mahdotonta päästä (täyneen) neutraaliutueen ja objektiivisuteen. Tästä huolimatta on hyvä ja tärkeää pystyä tunnistamaan tämä haaste.
  - \* Tutkimusta voi tehdä joistakin arvolähtökohdista, mutta sen tulisi olla näkyvää. Omien arvojen mahdollisimman selvä eksplikointi on yksi keino, jolla voi yrittää vähentää piiloarvojen vaikutusta tutkimukseen.
  - \* Käsitteet ovat harvoin arvovapaita. Useimmat käsitteet voidaan korvata toisilla, joilla on paikoin hyvin erilainen arvosisältö. Arvottava lataus saattaa olla paikoin tarkoituksellista (?)
  - \* Toisaalta arvoihin sitoutuminen väistämätöntä, sillä se on sosiaalisen olemassaolon sivutuote. Yhteiskunnan jäseninä meillä on tuskkin mahdollisuksia (täydellisesti) irroittautua arvoistamme kun pyrimme esim. ammatillisin päämääriin.
  - \* Arvpainotteisten valintojen tunnistaminen saattaa olla vaikeaa.
  - \* Myös pääinvastainen ongelma olemassa: Tutkimusta arvioidaan siihen perustellusti tai perusteettomasti kiinnitetyjen arvonäkökohtien mukaan.
- Tutkimukseen kuuluu olennaisesti myös oman tutkimustyön kuvaaminen, ts. kertomus siitä, miten esitettyihin tuloksiin on päästy.
  - Tämän myötä tieteelliselle ajattelulle on ominaista automaattinen **itsensä korjaaminen**.
  - Tutkimuskysmys, valitut menetelmät, käytetty aineisto ja tehdyt joh-topäätökset perataan auki tutkimusartikkeliissa/raportissa, joka siten lähetetään **vertaisarvioitavaksi** tietelliseen julkaisuun, jossa muut alan asiantuntijat arviovat sen ja päätävät hyväksytäänkö se julkistaavaksi.
- **Vertaisarvioinnissa** yksi tai useampi, tehdystä tutkimuksesta riippumaton, saman alan tutkija lukee ja tarkastaa tehdyn tutkimusartikkelin, arvoo sitä ja suosittaa tietellisen julkaisun editoivalle toimitajalle kyseisen artikkelin hyväksymistä tai hylkäämistä.
  - Vertaisarvointi ei aina takaa sitä, että julkaistu tutkimus olisi virheetön ja erinomaisesti tehty, vaan myös väärää tietoa pääsee välillä vertaisarvointiprosessin läpi.
  - Tämä ei kuitenkaan poista tieteellisen prosessin luotettavuutta, sillä uusi tieto varmentuu vasta usean samaa tutkimuskysymystä tutkineen ja vastaavat tulokset saaneen tutkimuksen myötä. Toisin sanone, tieteellisen prosessin voidaan ajatella konvergoituvan totuuteen, vaikka yksittäisiä virhearvointeja sattuisikin.

## 2.5. TIETEELLISEN TUTKIMUKSEN VAIHEET JA TULOSTEN JULKAISEMINEN25

- **Tutkimuksen vaatimukset**

- Tutkimus edellyttää arkikielä täsmällisempää kommunikaatiota. Ongelmaan liittyvien käsitteiden huolellinen määritteleminen ja erityyli on tarpeellista. Eivät korvaa empiiristä tietoa vaan vaikuttavat tiedon järjestymiseen ja sen perusteella tehtäviin päätelmiin.

**Esimerkki: Luonnontieteelliset vs. yhteiskunnalliset sovellutukset:**

- Luonnontieteiden lainalaisuuksia: Monet luonnontieteelliset ilmiöt ovat luonteeltaan varsin pysyviä.
  - Voidaan tehdä luotettavasti laajojakin yleistyksiä.
  - Selityksiä voidaan empiirisesti testata.
  - Luotettavia matemaattisia esityksiä voidaan kehittää.
- Yhteiskuntatieteissä (yhteiskuntatieteiden historiallisuuden myötä) erinäisiä lainalaisuuksia ja tyypillisiä piirteitä:
  - Usein tutkitaan yhteiskunnallisia **ilmiöitä**, jotka eivät suurelta osin ole toistettavissa.
  - Vaihtelevat huomattavasti ajan myötä (aiemmin voimassaolleet lainalaisuudet eivät välttämättä ole enää voimassa ja päinvastoin), mikä vaikeuttaa tilastollista analyysiä.
  - Yhteiskunnallisten ilmiöiden mittaaminen?
    - \* Yhteiskunnan rakenne ja toiminta on ehdollinen siinä käytettävän merkitysjärjestelmän suhteen. Kysymys **mittaamisesta** on asetettava suhteessa tähän käsitejärjestelmään. Joudutaan tekemään erilaisia kompromisseja eksaktisuus- ja systemaattisuusvaatimusten sekä arkikielessä monimerkityksellisyyden välillä.

## 2.5 Tieteellisen tutkimuksen vaiheet ja tulosten julkaiseminen

Tieteellinen tutkimus ja asiantuntijatyö tuottavat valtavan määän perusteltua, luotettavaa tutkimustietoa. Ks. tarkemmin tieteellisestä julkaisemisesta linkin tapauksessa erityisesti yhteiskuntatieteiden alalla, mutta perusperiaatteet pätevät myös muiden tieteenalojen tapauksessa

<https://blogs.uef.fi/tiedonhaku-yhteiskuntatiede/tieteelliset-julkaisut/>

## 26LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

Vastuullisen tieteen

<https://vastuullinentiede.fi/fi/julkaiseminen>

artikkelit tarjoavat tietoa siitä, kuinka tutkittua tietoa tuotetaan, julkaistaan ja arvioidaan luotettavasti ja yhteisesti hyväksyttyllä tavalla. Jotta tiete vaikuttaa koko yhteiskunnan hyväksi, toiminnan on oltava vastuullista tutkimuksen jokaisessa vaiheessa.

- Julkisuus ja avoimuus tekevät tutkimuksesta tiedettä. - Tiedeviestintä on tiedeyhteisöjen sisäistä ja ulkoista tiedonvälitystä ja vuorovaikutusta. Tutkimuksesta viestiminen ei ole vain tutkimustuloksista viestimistä. Vastuullinen tiedeviestintä lisää luottamusta tieteelliseen tietoon. - Tieteellinen julkaiseminen on tutkijoille tärkeä meritoitumisen tapa, ja siksi on tärkeää, että tekijyys määritellään niin, että se palkitsee tutkijat oikeudenmukaisesti.

## Luku 3

# Tilastotiede tieteenalana

Tässä luvussa hahmottelemme tilastotieteen piirteitä tieteenalana. Käymme läpi tilastotieteelle ominaisia piirteitä, jotka erottavat sen niin lähitieteistä, kuten matematiikasta ja tietojenkäsittelytieteestä, kuin myös sovellusaloista. Usein näkee tilastotieteen typistettävän vain työkaluksi eri sovellusalojen empiriseen tutkimukseen siitätäkin huolimatta että tilastotieteellä on oma rikas teoriapohjansa sekä kiistaton asema omana tieteenalanaan. Tieteenalan määritteleminen lyhyesti on aina hieman hankala. Tästä huolimatta seuraavassa yritämme osaltaan vastata seuraaviin kysymyksiin:

- Mitä tilastotiede on ja mitä se ei ole? Miksi tilastotiede ei ole vain sovellettua matematiikkaa tai matematiikalla höystettyä tietojenkäsittelyä?
- Mihin tilastotiedettä käytetään? Onko tilastotieteellä käyttöä ns. "akatemian" eli tutki- musyhteisön ulkopuolella?
- Tilastotieteelle tyyppillistä kriitikiä?

### 3.1 Lisää tilastotieteen perustermejä

Seuraavia tilastotieteen esittelyä ja karakterisointeja ajatellen määritellään seuraavassa lisää tilastotieteellisen tutkimuksen peruskäsitteitä. Näihin käsitteisiin paneudutaan osaltaan tarkemmin mm. luvussa ?? [otantaluku].

- Tilastotieteellinen tutkimus tarkastelee reaalimaailman ilmiöitä. Täten tutkimuskohteena on tavallisessa elämässä tavattavia asioita, ihmisiä tai tapahtumia. Tutkimuskohteita kutsutaan tilastoyksiköiksi ja niiden joukkoa kutsuaan populaatioksi (perusjoukoksi). Esimerkiksi jos tutkitaan kuntavaaleissa äänestävien tuloja niin jokainen äänestysikäinen muodostaa oman tilastoyksikkönsä (ks. alla) ja täten populaationa (perusjoukko-

na) toimii kaikki äänestysikäiset kansalaiset. Jos taas tutkitaan äänestysaktiivisuutta eri kunnissa, muodostaa jokainen kunta oman tilastoyksikönsä ja kaikki Suomen kunnat muodostavat populaation.

### Populaatio

Konkreettinen tai hypoteettinen tutkimuskohteiden joukko, joka koostuu kaikista tilastoyksiköistä

- Populaation muodostavilta tilastoyksiköiltä tarkastellaan niiden ominaisuuksia, eli **tilastollisia muuttuja**. Edellisissä esimerkeissä nämä olisivat esim. äänestäjien tulot ja kuntien äänestysprosentti. Mielenkiannon kohteena olevia tilastollisia muuttuja kutsutaan **tutkimusmuuttujiksi** (tulot ja kuntien äänestysprosentti) ja niiden lisäksi voidaan kerätä yli-määriästä tietoa eli **taustamuuttuja** (näitä voisi olla esimerkiksi asuinpaikka ja kunnan väkiluku).
- Tilastoyksiköiden tilastollisilla muuttujilla on tietty mahdollisten arvojen joukko, ja näillä arvoilla on jokin **jakauma** populaatiossa. Esimerkiksi tulot voivat määritelmästä riippuen saada minkä tahansa positiivisen arvon mutta äänestysprosentti on luonnollisesti rajattu nollan ja sadan prosentin välillä.

### Tilastoyksikkö ja tilastollinen muuttuja

Populaation muodostavilta tilastoyksiköiltä (populaation alkioilta) tarkastellaan tilastollisia muuttuja, joita voidaan mitata tai havaita.

- Kun tarkasteltavien tilastoyksikön tilastollisten muuttujien (numeeriset) arvot havaitaan, kutsutaan näiden arvojen joukkoa **havainnoksi**

### Havainto

Havainto muodostuu tilastoyksikön tarkasteltavien tilastollisten muuttujien havaitusta arvoista.

- Populaatio koostuu tilastoyksiköistä, joilla on tilastollisia muuttuja. Tarkasteltavista tilastollisista muuttujista kerätään havaintoja, joiden pohjalta tutkitaan **populaation ominaisuuksia**.
- Kerättyjen havaintojen joukko muodostaa **havaintoaineiston**, eli **datan**.

**Havaintoaineisto/data**

Havaintoaineisto, data, on tilastoyksiköiden tilastollisista muuttujista kerrätty havaintojen joukko.

**Tiivistettynä:**

- Populaatio tutkimuksen kohteena olevia tilastoyksiköitä.
  - Havaitaan tilastoyksiköistä tutkimuksen kannalta mielenkiintoisia tilastollisten muuttujien numeerisia arvoja.
  - Nämä havainnot muodostavat havaintoaineiston, eli datan, jota voidaan käyttää tutkimuksessa.
- Terminologiaa (käydään vielä läpi tarkemmin jatkossa): - Tilastoala = Tilastotiede + Tilastotoimi - Tilastotiede = Teoreettinen tilastotiede + Soveltava tilastotiede - Tilastotoimi = Tilastojen tuotanto + Tilastojen hyödyntäminen

### 3.2 Mitä tilastotiede on ja mitä se ei ole?

- Aloitetaan tarkastelemalla erinäisiä tilastotieteen “karakterisointeja” eri tahojen ja tutkijoiden toimesta:
  - *Tilastotiede on tietotuotannon teknologiaa, jonka avulla voidaan suorittaa kvantitatiivisten tietojen joukkotuotantoa ja havaintoihin perustuvia tieteellisiä ja käytännöllisiä päätöksiä. Tilastotiede on siis yksikköjen muodostamaan joukkoon liittyvän numeerisen tiedoaineiston keräämistä, analysointia ja tulkintaa koskeva tiete*<sup>1</sup>.
  - *Tilastotiede on yleinen menetelmätiede, jota sovelletaan, jos reaalimaailman ilmiöstä halutaan tehdä johtopäätöksiä ilmiötä kuvaavien kvantitatiivisten tai numeeristen tietojen perusteella sellaisissa tilanteissa, joissa tietoihin liittyy epävarmuutta tai satunnaisuutta*<sup>2</sup>.
  - *Tilastotiede on yleinen menetelmätiede, jota sovelletaan, jos reaalimaailman ilmiöstä halutaan tehdä johtopäätöksiä ilmiötä kuvaavien kvantitatiivisten tai numeeristen tietojen perusteella sellaisissa tilanteissa, joissa tietoihin liittyy epävarmuutta tai satunnaisuutta.*
  - *Vale, emävale, tilasto*<sup>3</sup>.
  - *Statistics concerns what can be learned from data*<sup>4</sup>.
  - *“Maalaisjärjen tehostamista”*<sup>5</sup>.

<sup>1</sup>Leo Törnqvistin, Suomen ensimmäisen tilastotieteen professorin, esittämä luonnehdinta (Vartia, 1989).

<sup>2</sup>Mellin, (2005).

<sup>3</sup>Mark Twain popularisoi tämän lausahduksen teoksessaan *Chapters from My Autobiography* jo vuonna 1907.

<sup>4</sup>(A.C. Davison)

<sup>5</sup>(Sund, 2003)

- Tilastotiede siis **kehittää ja soveltaa menetelmiä** ja (tilastollisia) **malluja**, joiden avulla reaalimaailman ilmiöistä voidaan tehdä johtopäätöksiä ilmiötä kuvaavien numeeristen tai kvantitatiivisten tietojen perusteella tilanteissa, joissa tietoihin liittyy **epävarmuutta ja satunnaisuutta**.
  - Tilastollisten menetelmien avulla pyritään löytämään reaalimaailman satunnaisia ilmiötä kuvaavista numeerisista (eli kvantitatiivisista) tiedoista **systemaattisia piirteitä** joita jalostetaan sellaiseen muotoon, että ilmiöistä voidaan tehdä päätelmiä.
    - \* Vrt. signaalin ja kohinan erottaminen (ks. Silver, 2014).
  - Tilastolliset mallit perustuvat todennäköisyyslaskentaan ja niillä mallinnetaan reaalielämän ilmiöiden alla piileviä prosesseja tai mekanismeja. Näiden prosessien tuottamia tietoja (aineistoja) tiivistetään usein graafisiksi esityksiksi ja tunnusluvuiksi sekä tilastollisten mallien parametreiksi, joiden pohjalta johtopäätöksiä tehdään.
  - Tässä onnistuakseen tilastollisten menetelmien tuleekin pyrkii erottelemaan **sattuma ja systemaattisuus** tarkasteltavissa ilmiöissä tai, tarkemmin, niitä kuvaavissa aineistoissa, jotta johtopäätökset olisivat luotettavia.

**Voidaan sanoa, että saadakseen tarkemmin selville mitä tilastotiede on, pitää opiskella tilastotiedettä ja sen käyttöä!**

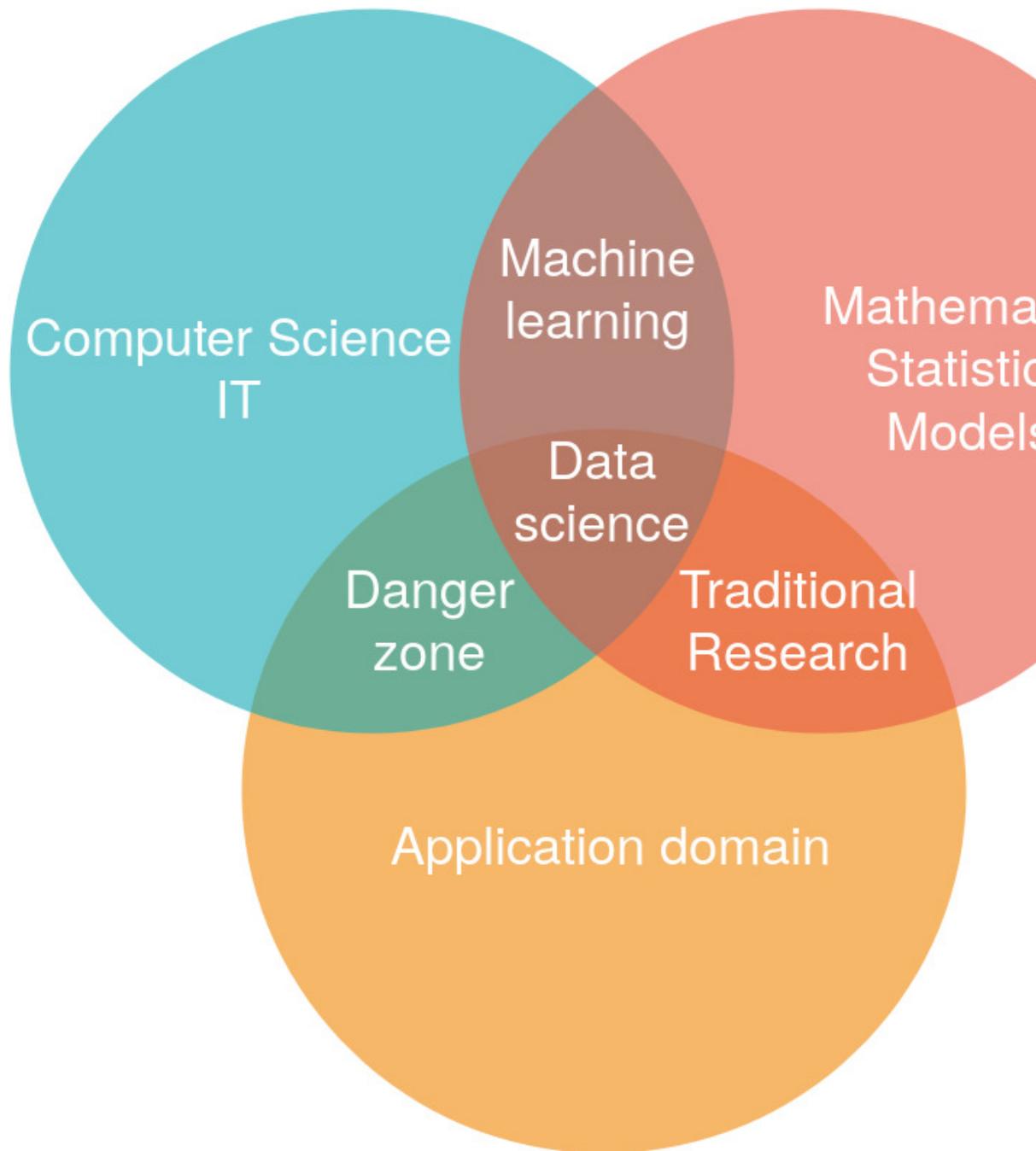
- Edellisten tilastotieteteen yleismaailmallisten luonnehdintojen jälkeen onkin sopivaa kysyä **mitä tilastotiede ei ole**.
  - Vaikka sana **tilasto** tuo useimille ensimmäisenä mieleen yhteiskuntaa ja sen toimintaa kuvaavat **numeeristen tietojen järjestelmälliset kokonaisuudet**, tilastotiede ei suinkaan ole ainoastaan tilastojen ja niiden tekemisen oppia.
    - \* Tämä siitäkin huolimatta, että niiden menetelmien konstruointi, joilla näitä tilastojen tuotetaan, jalostetaan ja analysoidaan on keskeinen osa tilastotiedettä. Tilastot ovat siis usein tilastotieteen soveltajan tutkimuskohteena ja tilastojen laadinnassa käytetään apuna tilastotieteen menetelmiä.
    - \* Suomessa Tilastokeskus toimii virallisena tilastoviranomaisena ja tilastotuottajana. Tätä **tilastotuotannon** kokonaisuutta nimitetään ajoittain **tilastotoimeksi**. **Tilastotieteen käytöalue on paljon tästä laajempi**.
  - Tilastotieteen kannalta mikä tahansa reaalimaailman ilmiötä kuvaava **numeeristen tai kvantitatiivisten tietojen järjestelmällinen kokonaisuus** voi muodostaa **tilastollisen aineiston** ja siten tilastollisen tutkimuksen mahdollisen kohteen.
  - Esimerkiksi kaikki **empiriisen** tai **kvantitatiivisen** tutkimuksen tutkimus- tai havaintoaineistot ovat tilastotieteen kannalta tilastollisia aineistoja.

- Tilastotiede sijoittuu tieteiden kentässä matematiikan, filosofian ja tietojenkäsittelytieteen rinnalle. Tästä huolimatta se ei kuitenkaan ole yksiselitteisesti minkään näiden osa-alue.
- **Tilastotiede ei ole matematiikan osa-alue**, sillä tilastotiede lähestyy tieteellistä ongelmanratkaisua eri tavoin: matematiikka on tiettyllä tavallaan aina eksaktia ja sen tulokset perustuvat formaaliin deduktioon ja loogisiin todistuksiin, johtuen usein ”eksaktiin” ratkaisuun tai matemaattisesti formaaliin ratkaisun esitystapaan. Tilastotiede sen sijaan on aina konteksti- ja aineistopohjaista ja perustuu induktiiviseen päätelyyn. Saadut tulokset ovat aina epävarmoja - koska ne kuvailevat epävarmaa tietoa generoivia prosesseja!
  - \* Tilastotiede on siis hyvä nähdä omana tieteenalanaan matematisesta esitystavastaan huolimatta. Eihän esimerkiksi myöskään fysiikkaa (sentäään) pidetä matematiikan osa-alueena!
- **Tilastotiede ei ole myöskään tietojenkäsittelytieteen osa-alue**, vaikkakin useiden laskennallisten menetelmien ja tehokkaan tietojenkäsittelyn rooli tilastollisissa analyyseissä on jatkuvasti kasvanut. Tietojenkäsittelytieteen teoria ei rakennu tilastotieteen tavoin ajatukselle epävarmoista ja satunnaisista reaalimaailman ilmiöistä.
- Vaikka nämä ja jotkin muut alat jakavat tilastotieteen kanssa useita piirteitä ja ominaisuuksia, on tilastotiede kuitenkin siis perustellusisti oma tieteenalansa. Tämä erottelun vaikeus jo itsessään todistaa kuinka keskeinen rooli tilastotieteellä on eri aloilla!
- Tilastotiede ei siis kuulu yksiselitteisesti sen lähitieteden alle, vaan muodostaa oman tieteenalan omine teorioineen ja tieteellisine premissineen. Käsittelemme myöhemmin tilastotieteen roolia matematiikan ja/tai datatieteiden (“data science”) kokonaisuudessa ja keskustelemme tarkemmin näiden erojen luonteesta.
- **Tilastotiede yleisenä menetelmätieteenä**
  - Tieteellistä tietoa ympäröivästä maailmasta hankitaan tieteellisillä **menetelmillä/metodeilla** (Ks. tieteellisen menetelmän kriteerit [Luku ?? 2]), joiden avulla tutkitaan jotaain ilmiötä tai sen generoimaa kvantitatiiivistä mutta epävarmaa tietoa sisältävää aineistoa.
  - Tilastotieteessä kehitetyt ja kehitettävät menetelmät antavat tutki-jolle yhtenevät ja tiedeyhteisön hyväksymät raamit, jotka mahdol-listavat (tilastollisen) päätelyyn ja päätöksenteon epävarman tiedon vallitessa. Näin voidaan uskottavasti ja luotettavasti tiivistää tietoa, jota erilaiset aineistot sisältävät, perustaa johtopäätöksiä näille tiivistyksille ja saavuttaa uusia tieteellisiä löytöjä.
    - \* Tilastotieteen menetelmien käyttö ja soveltaminen onkin siis aina alakohtaista. Tästä huolimatta tilastollisia menetelmiä sovelletaan aina johonkin **aineistoon!**

- Tilastotieteen nähdäänkin usein kuuluvan ns. **menetelmätieteisiin**, joissa mm.:
  - \* Kehitetään työkaluja muiden tieteiden tutkimusongelmien ratkaisuksi
  - \* On myös oma sovelluksista vapaa teorianmuodostuksensa
- Menetelmäkehityksen näkökulma tilastotieteeseen: *tilastotiede kehitää matemaattisia malleja satunnaisilmiötä kuvavia kvantitatiivisia tietoja generoiville prosesseille*. Koska tietoihin liittyy **epävarmuutta** tai **satunnaisuutta**, **tilastolliset mallit** perustuvat **todennäköisyyslaskentaan**.
- Juuri sattuman ja epävarmuuden huomioiminen tutkimusasetelmissa erottaa tilastotieteen muista menetelmätieteistä!
  
- **Aineisto:** Tilastotieteessä lähtökohtana ja ratkaisevassa asemassa on siis aina jonkin satunnaisilmiön generoima aineisto, josta haluamme oppia tai tietää lisää, kenties voidaksemme tehdä suuria yhteiskunnallisia päätöksiä sen pohjalta!
  - Tämä aineistokeskeisyys osaltaan erottaa tilastotieteen rajatieteistään ja osaltaan tuo sen lähemmäksi niitä ja sovellusaloojaan. (Näitä tarkastellaan myöhemmin luvussa ??).
  - Aineistoa analysoidaan, kuvillaan ja mallinnetaan tilastollisin menetelmin, joiden kehittäminen on keskeinen osa tilastotiedettä.
  - Pelkkä menetelmien kehittäminen kuuluu pitkälti matemaattisen tai teoreettisen tilastotieteen osa-alueelle.
  - Pelkkä aineestoon keskityminen ja (mekaaninen) analysointi voi sen sijaan olla joissain tilanteissa pitkälti tietojenkäsittelyä.
  - **Tilastollinen “mallintaminen”** löytyykin näiden välistä ja se sisältää eri alojen sovelluksista kumpuavan tarpeen uusien menetelmien kehittämiseen.
    - \* Tämä vuoropuhelu muodostaa tilastotieteelle luonnollisen “takaisinkytkenän” teoreettisen ja soveltavan puolen väillä: uudet teoreettiset menetelmät vastaavat soveltavan tilastotieteen ongelmiin mutta herättävät aina uusia kysymyksiä, jotka palautuvat taas teoreetikon pöydälle!
  - Luonnollisesti valtaosa tilastotieteilijöistä ja lähitieteiden harrastajista asettuvat näiden äärimmäisten luonnehdintojen välimaastoon eikä tarkkaa luokittelua ole sinänsä tarpeen tehdä ja korostaa.
  - Joka tapauksessa tilastotieteen kehityksen keskiössä ovat aina sovelusalakohtaiset ongelmat, joista useat palautuvat yleisemmälle tasolle teoreettisen tilastotieteen kehityspolkuihin.

### 3.3 Tilastotieteen suhde lähitieteisiin

- Kuvio 3.1 tarjoaa karkean yleistyksen tietojenkäsittelytieteen (Computer Science) ja sovellusalan (Application domain) sekä tilastotieteen (Statistics) ja matematiikan (Mathematics) välisistä yhteyksistä. On selvää että tilastotieteellä on paljon päälekäisyysksiä lähitieteiden kanssa ja joskus näkeekin (huolimatta edellä tehdyistä huomioista) että tilastotiede nipputaan yhteen matematiikan tai tietojenkäsittelytieteen kanssa.
- Yritetään siis vielä hahmotella tilastotiedettä lähimpänä olevaa (soveltaa) matematiikkaa.
  
- Tilastotieteessä olennaisen otantateorian (Luku ??) voisi ajatella olevan matemaattisesti määritellyt teoria, jossa myös on aineiston käsite, mutta se ei tee siitä vielä varsinaisesti tilastotiedettä.
- Matematiikassa kuvataan ongelma ja esitetään se teorian muodossa, malli on “parametreista havaintoihin”
- Tilastotieteessä ongelma on käänneinen, edetään “havainnoista parametreihin”, mutta ongelman matemaattinen kuvaus vaaditaan ensin
- Tilastotiede esittää menetelmiä ja käsittää käänteisen ongelman ratkaisemiseen
  - Karkeasti erotellen tilastotieteessä käsitteltävät ongelmat lähtevät aina havainnoista eli aineistosta ja matematiikassa suunta on teoriasta aineistoon.
  - Voidaan siis sanoa, että tilastotieteen erottaa puhtaasta matematiikasta se, että siinä tutkitaan metodeja, jotka mahdollistavat päättelyn/tiedon hankinnan puutteellisesta tai epävarmasta tiedosta.
  
- Ilmiöiden kuvaamiseen ja käyttäytymisen ennakoimiseen käytetään usein **mallia**. Mallit (matematiiset/tilastolliset mallit) voidaan jakaa **deterministisiin** ja **stokastisiin** malleihin.
  - Deterministisen mallin tapauksessa, tiettyjen alkuehtojen (alkuarvojen) vallitessa voidaan määrittää tarkaltevan ilmiön lopputulos. Esimerkkejä ovat esim. monet fysikan lait.
  - Stokastiset mallit perustuvat todennäköisyyslaskentaan. Stokastisia malleja käytetään kun alkuehtojen perusteella ei voida varmasti määrittää tarkasteltavan ilmiön lopputulosta. Tällöin eri vaihtoehtoihin liittyvä tietyt esiintymistodennäköisyydet. Esimerkkejä ovat esim. rahanteitto tai säänen ennustaminen.



Kuva 3.1: Tilastotieteen ja rajatieteiden yhteyksiä kuvaava Venn-diagrammi

- Kun jotain ilmiötä kuvataan stokastisen mallin avulla, voidaan käyttää (joudutaan käyttämään) tilastollisia menetelmiä. Vaikka käytännössä laskenta hoidetaan tietokoneohjelmien avulla, meidän tilastotieteen tutkijoina ja käyttäjinä on huolehdittava tutkimusprosessin onnistuneesta toteutuksesta muita osin.
- Tarkastellaan seuraavaksi tilastotieteen suhdetta viime vuosien aikana paljon suosiota keränneeseen datatieteeseen (data science)
  - Tilastollinen tietojenkäsittely
  - Data-analyysi
  - Koneoppiminen
- Tilastotiede = tietojenkäsittelytiede? vai Tilastotiede = datatiede (data science)?
  - Hyödyllisen tiedon survomista aineistosta. Suomen kielessä tietojenkäsittely ymmärretään kuitenkin laajemmassa mielessä ohjelmoitavissa olevaksi automatisoimiseksi, jota tilastotiede ei perusolemukseltaan suinkaan ole.
- “Danger zone”
  - Kuvan 3.1 “danger zone” kuvaaa tilannetta, jossa ilmiöiden/mallien tilastotieteellinen perusta unohdetaan.
  - Tilastotieteen näkökulman ohittava (laiminlyövä) soveltaja ei aina kykene ajattelemaan kriittisesti muodostuvaa ennustemallia, tai ennen muuta vain esilletulevaa ennustetulosta, kohtaan eikä päädy parhaisiin mahdollisiin (tarkimpia) ennustetuloksiin tilanteessa, jossa jokin toinen malli kuvasi ilmiötä annettua mallia paremmin.
  - Ko. soveltaja ottaa mallin sekä sen antaman ennustetuloksen annettuna, eikä mietti mistä kyseinen ennustetulos johtuu. Jotta tarkat ennustetulokset toteutuvat jatkossakin (kun uutta aineistoa, dataa, tulee saataville), on ennustajan oleellista huomioida mitkä tekijät johtivat tarkkaan ennustulokseen.
  - Eri menetelmät sopivat eri sovelluskohteisiin. Tilastotieteilijä osaa useimmiten tunnistaa eri sovelluskohteisiin sopivat menetelmät paremmin kuin tietojenkäsittelijä. Vastaavasti tehokkaan/onnistuneen ohjelointikoodin kirjoittamisessa tilanne on usein toisinpäin.

### 3.4 Tilastotieteen osa-alueet

- Tilastotiede jakautuu moniin osa-alueisiin. Osa-alueita on niin paljon, että alan huiputkaan eivät voi hallita niitä kaikkia!
- Tästä huolimatta tilastotiede voidaan karkeasti jakaa teoreettiseen ja soveltavaan osa-alueeseen, jotka toimivat alituissa vuoropuhelussa.

#### Soveltava tilastotiede

on nimensä mukaisesti teoreettisen tilastotieteen kehittämien menetelmien soveltamista jonkin tutkimusalan empiiriseen ongelmaan. Suurin osa tilastotieteen menetelmistä on alun perin kehitetty jonkin konkreettisen tutkimusongelman innoittamana.

- Yleisesti ottaen eri tieteenaloilla kohdattavat menetelmäsuuntaukset voidaan jakaa kahteen luokkaan tutkimusaineistojen tyypin perusteella:
- **Kvantitatiivinen:** eli määrällinen tutkimus on tutkimusta, jossa tutkimusongelma on muotoiltu tarkasti etukäteen ja tutkimuskysymyksiin vastataan käyttäen tilastollisia menetelmiä pyrkien **selittämään ja ennustamaan** tutkimuksen kohteena olevaa ilmiötä.
  - Täsmällisten ja laskennallisten tilastollisten menetelmien käytäminen on kvantitatiiviselle tutkimukselle ehkä ominaisin piirre.
  - Perustuu yleensä satunnaisotokseen (kts. luvut 4, 5 ja 6) ja tutkimusaineisto on tiivistetty numeeriseksi havaintomatriisiksi, jolle oleellinen vaatimus on sen totuudellisuus.
  - Kritiikki: määrällinen tutkimus on (paikoin) sokea tutkittavien ilmiöiden sellaiselle luonteelle, jota ei pystytä kvantifioimaan, eli muuntamaan numeeriseen muotoon. Näihin voidaan katsoa lukeutuvan mm. tunteet, merkitykset ja kokemukset, ellei tutkija keksi niiden numeeriselle mittamiselle uskottavaa keinoa.
- **Kvalitatiivinen:** eli laadullinen tutkimus on tutkimusta, jossa tutkimuksen kohteena olevaa ilmiötä ja sen merkitystä sekä tarkoitusta pyritään **ymmärtämään** kokonaisvaltaisella tavalla.
  - Laadullisessa tutkimuksessa annetaan usein tilaa tutkimuksen kohteena olevien ilmiöiden ja/tai ihmisten näkökulmille, vaikuttimille, kokemuksille ja tuntemuksille. Tutkimusyksikköjen otanta on täten usein harkinnanvaraista.
  - Laadullisessa tutkimuksessa tutkimusongelma muotoutuu tutkimuksen edetessä ja sille tyypillistä on hypoteesittomuus, eli tutkimus on tarkoitus aloittaa mahdollisimman vähin ennakkooletuksin. Ennakkooletuksista on kuitenkin mahdotonta täysin irtautua, joten

niiden ilmi tuominen esioletuksina tai ”tutkimushypoteeseina” eli arvauksina tuloksista on osa tutkimusta.

- Kritiikkiä: laadullinen tutkimus ei pysty vastaamaan kysymykseen miksi, sillä ilman määrellisiä (numeraalisia) aineistoja ei ilmiöiden välisiä riippuvuuksia kyetä tutkimaan: laadullisessa tutkimuksessa menetetäänkin mahdollisuus tutkia ilmiöiden todellisia syitä.
- Usein pyritään vastaamaan kysymyksiin ”miksi?”, ”miten?” ja ”miltä?”
- Yleisenä menetelmätieteenä tilastotiedettä voidaan (ja myös pitäisi) soveltaa kaikilla reaalimaailmaa tutkivilla tieteenaloilla, joiden tutkimusaineistot voidaan esittää kvantitatiivisessa muodossa.
- Tilastotiede on saanut alkunsa siitä, että yhteiskunnan modernisoitussa on tarvittu yhä enemmän tietoja erilaisiin hallinnollisiin tarpeisiin. Samalla on syntynyt tarve kehittää menetelmiä joiden avulla tilastojen luotettavuutta on voitu parantaa.
- Kehitys oli pitkään ns. ongelmasta menetelmään!
- Tilastollisia menetelmiä voidaan soveltaa tietojen keruun, jalostuksen ja analysoinnin jokaisessa vaiheessa. Päämäääränä on jalostaa tiedot muotoon, joka mahdollistaa tutkittavaa reaalimaailman ilmiötä koskevien johdotuustöiden tekemisen käytettyjen menetelmien pohjalta, eli ns. **tilastollisen päätelyn**.
  - Tutkimuksessa on pystyttää valitsemaan ja käyttämään menetelmiä, jotka antavat aineistosta vastauksia haluttuihin kysymyksiin. Tämä vaatii yhtä lailla sovellusalakohtaista osaamista (ns. substanssiosaamista) kuin myös kattavaa menetelmäosaamista.
- Menetelmien käytön tarkoituksesta on (voi olla) (*i*) **kuvailla ja tiivistää tietoa**, jota havaittu aineisto sisältää (*ii*) tilastotieteen oman ja jonkin toisen tieteenalan **teorian empiirinen testaus** tai (*iii*) edellisten pohjalta **tilastollinen päätely**.
  - **Deskriktiivinen eli kuvaileva tilastotiede** kehittää ja soveltaa menetelmiä, joiden avulla havaintoaineistosta voidaan esimerkiksi laskea tunnuslukuja, kuvata havaintomuuttujien jakaumia ja visualisoida aineiston generoimaa ilmiötä tai siitä johdettuja tunnuslukuja.
  - **Teorian testaaminen** voi johtaa joko teorian vahvistumiseen (*verifiointiin*) tai sen vääräksi osoittamiseen (*falsifioimiseen*).<sup>6</sup> On myös

---

<sup>6</sup>Lienee kuitenkin tieteenfilosofinen kysymys, onko falsifioinnin vastakohta juuri verifioin-

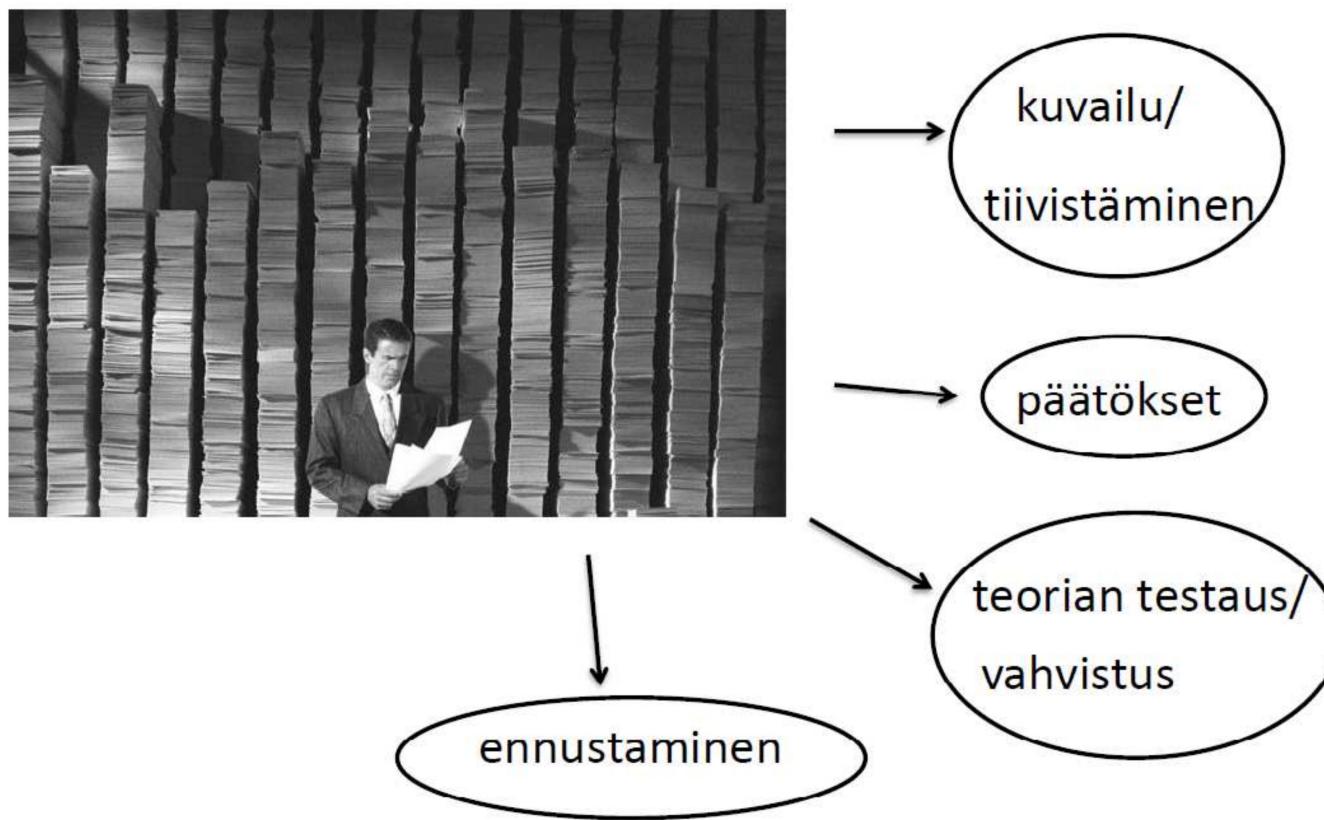
syytä muistaa, että yksi tutkimus ei vielä osoita teoriaa oikeaksi tai vääräksi vaan siihen tarvitaan useita tutkimuksia sekä erilaisia tutkimusasetelmia ja -menetelmiä.

- **Tilastollinen päättely** on sen sijaan aineiston tarkasteluun/kuvailuun sekä mallintamiseen perustuva päättöksenteko, jossa kvantitatiiviseen aineistoon kuuluva epävarmuus ja satunnaisuus on otettu huomioon. Tilastollisia päättely voidaan pohjata näiden mallien testaamiseen - hypoteesien testaaminen!
- Kuvaleva tilastotiede ja tilastollinen päättely kulkevat tilastollisessa tutkimuksessa käsi kädessä.
  
- Useimmiten kuitenkin ajatellaan, että on käytettävä niin yksinkertaisia menetelmiä kuin mahdollista, mutta ei yhtään yksinkertaisempia. Ns. **parsimoonisuusperiaate** eli **vähäparametrisuus-** tai **säästeliäisyysperiaate**.
  - Vähäparametrisuusperiaatteen voidaan nähdä perustuvan ns. Occamin partaveitsen -periaatteeseen, jonka mukaan *“ilmiötä selittävien tekijöiden määräն tulee olla mahdollisimman vähäinen”*, ts. tilastotieteessä menetelmien (mallien) tulee olla mahdollisimman yksinkertaisia, mutta silti riittäviä.
  - Tämä periaate ja sen suhde ns. **varianssin ja harhan väliseen kompromissiin** on erityisen tärkeä erityisesti tilastollisen ennustamisen ja viime vuosikymmeninä yleistyneen tilastollisen (kone)oppimisen sovellutuksissa (ks. tarkemmin alaluku 3.?? ja luku ?? (estimaattoriluku)).

**Teoreettinen tilastotiede** kehittää (tilasto)matemaattisia malleja kuvaamaan satunnaisilmiöitä- ja prosesseja, jotka generoivat reaalimaailman ilmiötä kuvaavia numeerisia tai kvantitatiivisia tietoja, joihin liittyy epävarmuutta ja satunnaisuutta.

- Teoreettinen tilastotiede luo pohjan tilastollisten menetelmien ymmärtämiselle, soveltamiselle ja kehittämiselle.
  - Ilman riittävää ymmärrystä tilastollisten menetelmien toimintaperiaatteista niiden soveltaja on vaarassa tehdä virhepäätelmiä! (Ks. alaluku 3.?? tilastotieteen kriitikistä)
- Mallit perustuvat todennäköisyyslaskentaan, ja niitä kutsutaan tilastollisiksi malleiksi, stokastisiksi malleiksi tai todennäköisyysmalleiksi.

ti tai että onko kumpikaan ylipääätään mahdollista. Jätämme kuitenkin semantiikan muille kursseille ja käytämme näitä termejä löyhästi, tavan kansalaisen ymmärtämin sisällöin.



Kuva 3.2: Soveltava tilastotiede

- Tilastolliset mallit perustuvat laajalti niin kutsuttuun uskottavuusfunktioon. Se on malli, joka riippuu havaintoaineiston lisäksi yhdestä tai useammasta parametrista.
- Uskottavuusfunktion arvo kertoo kuinka todennäköisenä voidaan havaittua aineistoa pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvoilla.
- Uskottavuuspäättelyn perusajatuksena on, että se tai ne parametritarvot, joilla uskottavuusfunktion arvo maksimoituu kuvaaa aineiston generoimaa prosessia parhaiten.
- Aineistoa koskevia hypoteeseja voidaan testata käyttäen uskottavuusfunktion maksimia vastaavaa tilastollista mallia! - *“Kaikki mallit ovat väärää, mutta jotkut ovat käyttökelpoisia.”*(Box, 1976).
  
- Uskottavuusfunktiot perustuvat aina satunnaisilmiöiden mahdollisia arvoja kuvaaviin nk. **tiheysfunktioihin** tai todennäköisyysfunktioihin.
  - Tiheysfunktiot kuvaavat jonkin satunnaismuuttujan (satunnaisilmiön) saamien arvojen jakaumaa.
  - Esimerkiksi kolikonheitto on satunnaisilmiö ja sillä on vain kaksi arvoa<sup>7</sup> ja kolikonheittoa voidaan kuvata nk. binomijakaumalla, merkitään Bin( $n, p$ ) missä  $n$  on heittojen lukumäärä ja  $p$  on kruunan todennäköisyys.
  - Esimerkki: heitetään kolikkoa 40 kertaa ja saadaan kruuna 40/40 tapauksessa. Onko tämän havaintoaineiston perusteella uskottavaa, että kolikonheitto noudattaa binomijakaumaa Bin(40, 0.5)? Eli kuinka uskottavan voidaan pitää että kyseinen kolikko on tavallinen, painottamatonta kolikko??
  
- Todennäköisyyslaskenta luo tilastotieteelliselle epävarmuuden mallintamiselle vahvan ja uskottavan matemaattisen perustan.
  - Todennäköisyyslaskentaa opetetaan tarkemmin (tätä kurssia seuraavilla) kursseilla <https://opas.peppi.utu.fi/fi/opintojakso/TILM3553/1734> (pääaineopiskelijoille, <https://opas.peppi.utu.fi/fi/opintojakso/TILM3568/3385> ja <https://opas.peppi.utu.fi/fi/opintojakso/SMAT5306/4400>.

---

<sup>7</sup>Kolikon kantilleen jäämistä ei tässä lasketa mahdolliseksi tapahtumaksi.

$$\begin{aligned}
 E[\sigma_y^2] &= E \left[ \frac{1}{n} \sum_{i=1}^n \left( y_i - \frac{1}{n} \sum_{j=1}^n y_j \right)^2 \right] \\
 &= \frac{1}{n} \sum_{i=1}^n E \left[ y_i^2 - \frac{2}{n} y_i \sum_{j=1}^n y_j + \frac{1}{n^2} \sum_{j=1}^n y_j \sum_{k=1}^n y_k \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \left[ \frac{n-2}{n} E[y_i^2] - \frac{2}{n} \sum_{j \neq i} E[y_i y_j] + \frac{1}{n^2} \sum_{j=1}^n \sum_{k \neq j} E[y_j y_k] + \frac{1}{n^2} \sum_{j=1}^n E[y_j^2] \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \left[ \frac{n-2}{n} (\sigma^2 + \mu^2) - \frac{2}{n} (n-1)\mu^2 + \frac{1}{n^2} n(n-1)\mu^2 + \frac{1}{n} (\sigma^2 + \mu^2) \right] \\
 &= \frac{n-1}{n} \sigma^2.
 \end{aligned}$$



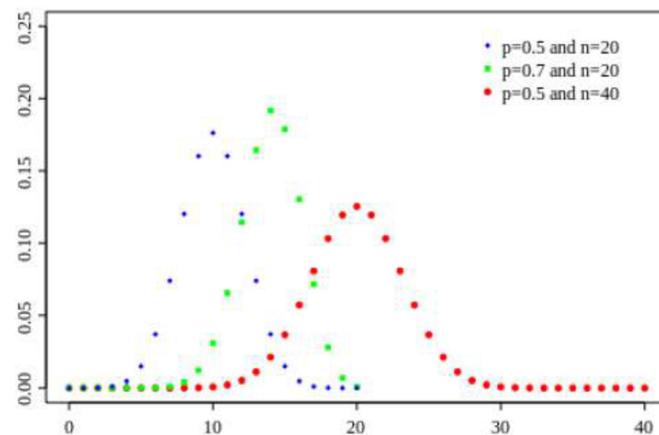
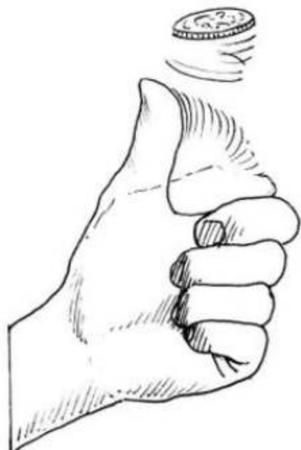
Pohja tilastollisten menetelmien ymmärtämiselle, soveltamiselle ja kehittämiselle

Kuva 3.3: Teoreettinen tilastotiede

$$\begin{aligned}
 & \text{Lataa } \ln(\frac{f(x)}{f(x_0)}) = \ln(\frac{\sqrt{2\pi} e^{-x^2/2}}{\sqrt{2\pi} e^{-x_0^2/2}}) = \ln(2e^{-(x-x_0)^2/2}) = \\
 & \frac{1}{2} \ln(2) - \frac{1}{2} \ln(e^{-(x-x_0)^2/2}) = \frac{1}{2} \ln(2) + \frac{1}{2} \ln(e^{-(x-x_0)^2}) = \frac{1}{2} \ln(2) - \frac{1}{2} \ln((x-x_0)^2) \\
 & \frac{1}{2} \ln(2) - \frac{1}{2} \ln((x-x_0)^2) = \frac{1}{2} \ln(2) - \frac{1}{2} \ln(\sum_{i=1}^n (x_i - \bar{x})^2) = \frac{1}{2} \ln(2) - \frac{1}{2} \ln(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2) \\
 & \ln(\frac{f(x)}{f(x_0)}) = \ln(\frac{f(x)}{f(\bar{x})}) = \frac{1}{2} (\ln(\frac{1}{S^2}) - \ln(\frac{1}{S^2 + (\bar{x} - \mu_0)^2})) \\
 & = \frac{1}{2} \frac{1}{S^2} \cdot \frac{1}{S^2 + (\bar{x} - \mu_0)^2} \cdot 2(\bar{x} - \mu_0)
 \end{aligned}$$

Kuva 3.4: Matemaattinen tilastotiede

Tilastotiede perustuu uskottavuksiin, jotka taas perustuvat todennäköisyyteen ja tiheysfunktioihin.

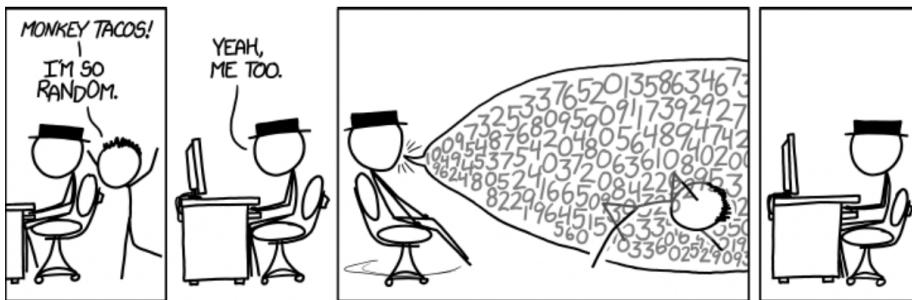


Kuva 3.5: Tilastotiede ja todennäköisyys



## Luku 4

# Sattuma ja satunnaisuus



Kuva 4.1: Hauska kuva satunnaisuudesta.

Tässä luvussa pohdimme sattuman ja satunnaisuuden roolia tilastotieteessä ja tieteessä ylipäätään. Satunnaisuudella tarkoitetaan yleensä säännönmukaisuuden puuttumista ja ennustamattomuutta ja kenties juuri siksi sitä voidaan pitää yhtenä maailman vaikuttavammista ilmiöistä. Jokainen haluaisi tietää mitä tulenan pitää ja siksi sattuma on myös filosofisesti mielenkiintoinen: se vaikuttaa ja muokkaa niin meitä itseämme kuin ympäröivää maailmaa mitä merkityksellisimmin tavoin - joskus jopa vasten tahtoamme ja usein vailla täyttä ymmärtämämme!

Ihmisen oma kokemus on kuitenkin altis kaikenlaisille virhepäätelmileille, joita kutsutaan myös kognitiivisiksi vinouumiksi. Haluamme löytää systematiikkaa ja tarkoitusta kaaoksesta sekä merkityksiä ja syy-seuraussuhdeita sellaisista tapahtumista, jotka kuuluvat normaalivaihtelun piiriin. Tällaisissa tilanteissa usein tilastollinen tarkastelu paljastaakin ilmiön todellisen, alkuperäisestä kuvitelmasesta poikkeavan luonteen. Osatakseen erottaa systemaattisen vaihtelon ja ymmärtääkseen oikeasti merkityksellisiä syy-seuraussuhdeita, on välttämöntä ymmärtää satunnaisuutta. Tämä välttämättömyys päätee erityisesti tiedeyhteisön jäseniin, jotka pyrkivät tutkimaan ympäröivän maailman satunnaisia ilmiöitä.

Tilastotiede perustuu satunnaisilmiöiden ja satunnaisen aineiston tutkimiseen, joten sen ymmärtäminen on keskeisessä roolissa tieteen ja maailman ymmärtämisessä.

## 4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä

- Edellisestä luvusta muistamme, että tilastotieteellisen tutkimuksen kohde on aina jokin tilastoyksikköjen tutkimusmuuttujista koostuva havaintoaineisto, jonka pohjalta tehdään päätelmiä perusjoukosta/populaatiosta.
- Nämä tilastolliset muuttujat tulkitaan satunnaisiksi, ja täten tilastollisen tutkimuksen tavoite on tutkia satunnaisilmiötä, joka on generoinut nämä havaitut eli toteutuneet arvot.
  - Yksi tilastotieteen olennainen tehtävä onkin kehittää **tilastollisia malleja**, joiden avulla satunnaisilmiötä voidaan kuvata, selittää ja ennustaa.
  - Tilastollisen mallin satunnaisten piirteiden kuvaus perustuu johonkin **todennäköisyysmalliin**.

### Satunnaisilmiö

Reaalimaailman ilmiö on satunnaisilmiö, jos seuraavat ehdot pätevät:

- Ilmiöllä on useita erilaisia tulosvaihoehtoja.
- Sattuma määräää mikä tulosvaihoehtoista toteutuu, eli yksittäistä tulosta ei voida tietää etukäteen.
- Vaikka tulos vaihtelee ilmiön toistuessa satunnaisesti, käytäytyy tulosvaihoehtojen suhteellisten osuuksien jakauma tilastollisesti stabilisti ihniön toistokertojen lukumäärän kasvaessa.

- **Tilastollisella stabiiliudella** tarkoitetaan sitä, että on mahdollista arvioida kuinka **todennäköisiä** erilaiset tapahtumat, eli satunnaisilmiön tulosvaihoehdot ovat.
  - Toisin sanoen satunnaisilmiön tulosvaihoehtoihin on liityttävä säännönmukaisuutta, jonka on tultava esille ilmiön toistuessa.

### Esimerkkejä satunnaisilmiöstä uudistettava...

- Kvanttimekaniikan ja hiukkasfysiikan ilmiöt ovat perusuonteellisia.

## 4.1. SATUNNAISILMIÖT JA SATUNNAISMUUTTUJAT TILASTOTIETEESSÄ

taan satunnaisia.

- Luonnontieteellisiin mittauksiin liittyvien mittausvirheiden syntymekanismit ovat (ainakin osittain) satunnaisprosesseja.
- Uhhapeleissä kuten arpajaisissa, lotossa, ruletissa, korttipeleissä ja noppapeleissä sattumalla on keskeinen rooli.
- Perinnöllisyys noudattaa sattuman lakeja.
- Eliöiden ominaisuuksien jakautuminen populaatiossa on satunnista.
- Ihmisten, ihmisryhmien ja ihmisten muodostamien organisaatioiden sosiaalisessa ja taloudellisessa käyttäytymisessä on monia satunnaisia elementtejä.
- Teknisten prosessien tuloksien ominaisuudet jakautuvat satunnaisesti.

### Satunnaismuuttujat

- Satunnaisilmiötä koskevan tutkimuksen kohteena olevat tilastolliset muuttujat tulkitaan **satunnaismuuttujiksi** ja havaimnot (havaintoarvot) voidaan näin ollen tulkita näiden satunnaismuuttujien realisoituneiksi arvoiksi. Satunnaismuuttuja siis kuvaaa tarkasteltavan mitattavan ominaisuuden (satunnais)vaihtelua tutkimuksen kohteiden, eli tilastoiksiiden joukossa.
  - Mitattavan ominaisuuden mahdolliset arvot määräväät satunnaismuuttujan luonteen. Yleisesti satunnaismuuttujat jaetaan kahteen luokkaan: **jatkuihin ja diskreetteihin**.
  - Satunnaismuuttujan **todennäköisyyssjakauma**, määräää erilaisten tulosvaihtoehtojen todennäköisyden ja mahdolistaa täten tilastollisen analyysin ja päättelyn.
    - \* Satunnaisuus eroaa mielivaltaisesta prosessista siinä, että satunnaista ilmiötä voidaan kuvata jollakin **tilastollisella lailla** kun taas mielivaltaista prosessia ei.

#### Satunnaismuuttuja

Satunnaismuuttuja (usein lyhyesti sm., englanniksi random variable, merkitään esim.  $Y$ , ja kutsutaan ajoittain myös stokastiseksi muuttujaksi) on todennäköisyyslaskennan peruskäsite, jolla tarkoitetaan satunnaisilmiön määräämää lukua.

- Satunnaismuuttujan  $Y$  realisoituvaa arvoa  $y$  kutsutaan realisatioksi tai toteumaksi.

- Tilastollinen aineisto muodostuu useiden satunnaismuuttujien (tilastoysiköiden tutkimusmuuttujien) realisoituneista arvoista.
- Realisoituneiden arvojen vaihtelua tilastoysiköiden välillä kutsutaan satunnaisvaihteluksi.

### Jatkuват ja diskreetit satunnaismuuttujat

- Satunnaismuuttuja  $Y$  on jatkuva, jos se voi saada ylinumeroituvan määän arvoja tai ts. minkä tahansa arvon joltain väliltä, kuten tyyppillisesti minkä tahansa arvon joltain reaalilukuväliltä.
- Satunnaismuuttuja  $Y$  on diskreetti, jos se voi saada vain joitain mahdollisia arvoja (vain yksittäisiä, äärellisen tai numeroituvasti äärettömän määän, arvoja). Yksinkertaisimmillaan diskreetti satunnaismuuttuja  $Y$  on kaksiarvoinen (binääriinen), jolloin sen mahdollisia arvoja tyyppillisesti merkitään  $y = 0$  tai  $y = 1$ .

### Esimerkki: satunnaismuuttuja

Ihmisen pituutta voidaan pitää (ennen mittaukseen tulemista) satunnaismuuttujana  $Y$  ja lopullista pituutta täten pituuden realisaationa  $y$ . Pituutta kohdellaan jatkuvana muuttujana senttimetreissä, mutta mikäli määritetään toteumaksi jonkin pituuden raja-arvon, esimerkiksi 170cm, ylittävä pituus, on kyseessä kaksiarvoinen (binääriinen) satunnaismuuttuja (pituus on joko yli tai alle 170 cm).

- Muuttujat voidaan luokitella myös **kvalitatiivisiin** ja **kvantitatiivisiin** muuttuijiin.
  - Kvalitatiivisiin muuttuijiin liittyy luokittelut- tai järjestysasteikko
  - Kvantitatiivisiin muuttuijiin välimatka- ja suhdeasteikko.
- Tilastolliset menetelmät perustuvat todennäköisyyslaskennan<sup>1</sup> tuloksiin ja tarjoavat keinon hallita satunnaisuuden aiheuttamaa epävarmuutta sekä tavan erottaa systemaattinen ja satunnainen vaihtelu, eli signaali ja kohina, toisistaan.
- Tilastollisen aineiston **tilastollisella mallilla** tarkoitetaan täten niiden satunnaismuuttujien todennäköisyysjakaumaa, jonka ajatellaan generoivien havainnot.

<sup>1</sup>Todennäköisyyslaskentaa käsitellään väilläisesti tulevissa luvuissa mutta varsinaisesti tarkeimmin 2. periodin kurssilla TILM3553 Todennäköisyyslaskennan peruskurssi ja (erityisesti sivuaineopiskelijoille) TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille.

#### 4.2. TILASTOTIETEEN SUHDE SATUNNAISUUTEEN JA TODENNÄKÖISYYKSIIN49

- Yksinkertaisimillaan esimerkiksi yksinkertaiseen satunnaisotantaan takaisinpanolla perustuva satunnaismalli (palaamme tähän otantaa käsitlevää luvussa 5).
- Satunnaisuus perustuu siihen, että satunnaismuuttujien toteutuvat arvot (ja niistä lasketut tunnusluvut kuten keskiarvo) vaihtelevat satunnaisesti otoksesta toiseen.
- Todennäköisyyslaskennan tehtävä on tuottaa **matemaattisia ja tilastollisia malleja** satunnaisilmiöissä havaittavalle tilastolliselle stabilitetille.

## 4.2 Tilastotieteen suhde satunnaisuuteen ja todennäköisyysyksiin

- Tilastotieteessä **tutkimusaineiston keräämistä** voidaan pitää hyvänä esimerkinä satunnaisilmiöstä.
  - Voimme ajatella, että tilastollisen tutkimuksen kohteet on aina valittu arpomalla.
  - Arvonta on mainio esimerkki satunnaisilmiöstä, sillä siihen liittyy aina ennustamattomuutta: vaikka yksittäisen arvonnan tulosta ei voi tietää etukäteen, noudattaa se kuitenkin todennäköisyden lakeja.
  - Koska arvonnan tulos vaihtelee satunnaisesti arvontakerrasta toiseen, myös tutkimuksen kohteita kuvaavat tiedot vaihtelevat satunnaisesti arvontakerrasta toiseen.
  - Tutkimuksen kohteita kuvaavien tietojen käytäytymisessä havaitaan kuitenkin arvontaa toistettaessa juuri sitä säännönmukaisuutta, jota kutsutaan tilastolliseksi stabilitetiksi. **Tämä säännönmukaisuus on tilastollisen tutkimuksen kohde.**
- Esimerkkejä tilastollisten aineistojen keräämisen menetelmistä, jotka perustuvat arvontaan:
  - **Satunnaistetut kokeet:** Kokeellisessa tutkimuksessa tavoitteena on vertailla erilaisten käsittelyiden vaikutuksia kokeen kohteisiin. Eri-laisten virhelähteiden kontrolloimiseksi käsittelyt on sytytä arpoa koh-teille.
  - **Satunnaisotanta:** Otannalla<sup>2</sup> tarkoitetaan laveasti tutkimusaineistojen keräämisen menetelmiä. Erilaisten virhelähteiden kontrolloimiseksi tutkimuksen kohteet on sytytä valita arpomalla. (Ks. Luku 5)
- Kerätyn (tai havaitun) aineiston pohjalta tehdään päätelmiä sen generoivista satunnaisilmiöstä esimerkiksi testaamalla erilaisia siihen liittyviä hypoteeseja.

---

<sup>2</sup>Erityisesti erilaisten otantamenetelmien yhteydessä, joita tarkastellaan tarkemmin luvussa 5.

- Tilastotiede voidaan jakaa kahteen suureen paradigmaan sen muukaan, miten tilastolliseen päätelyyn, ml. hypoteeseihin ja niiden testaamiseen, suhtaudutaan. Näitä ovat **klassinen eli frekventistinen tilastotiede sekä Bayesilainen tilastotiede**. Tarkastellaan seuraavaksi minkälaisia eroja ja yhtäläisyyskiä näiden koulukuntien välillä on.

### **Frekventistinen tilastotiede**

- Klassisessa eli frekventistisessä tilastotieteessä ajatellaan että hypoteesien testaaminen tulee perustua yksinomaan havaittuun aineistoon ja siihen liittäävään tilastolliseen malliin.
- Nimi ”frekventistinen” juontuu siitä, että tämä todennäköisyysjakauma määritää satunnaismuuttujan mahdollisten arvojen todennäköisydeksi niiden suhteellisen osuuden äärettömästä määrästä realisaatioita, ts. niiden suhteellisen frekvenssin.
- Klassisessa tilastotieteessä havaittuun aineistoon *sovitetaan* sitä kuvaavaan todennäköisyysjakaumaan perustuva tilastollinen malli, joka vastaa saatua aineistoa parhaiten.
  - Tämä tilastollinen malli perustuu nk. **uskottavuusfunktioon**, joka on *aineiston* sekä yhden tai useamman *parametrin* funktio ja joka saavuttaa suurimman arvonsa nk. ”suurimman uskottavuuden pistessä”.
  - Uskottavuusfunktio kertoo kuinka todennäköisenä havaittua aineistoa voidaan pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvolla. Täten ne parametriarvot, joilla uskottavuusfunktion arvo maksimoituu, kuvaavat aineiston generoimaa prosessia parhaiten, annettuna malli- eli jakaumaoletus.
  - Uskottavuusfunktioista, tilastollisten mallien estimoinnista ja parametreista lisää seuraavassa alaluvussa sekä luvussa 6.
- Perusjoukko koskevia hypoteeseja testataan tämän tilastollisen mallin avulla: havaittu aineisto määritää uskottavuusfunktion perusteella sellaiset hypoteesit, jotka jäävät joko voimaan tai tulevat hylätyiksi.
- Klassisessa tilastotieteessä hypoteesien testaus perustuu siis vain aineistoon eli tilastollinen päätely on induktiivista: aineiston avulla otosta koskeva päättelmä voidaan yleistää koskemaan perusjoukkoa.
  - Toki kaikki päätely on alisteista tehdynille oletuksille koskien käytetävää tilastollista mallia.

### Bayesilainen tilastotiede

- Bayesilainen tilastotiede on tilastotieteen toinen suuri paradigma ja on saanut nimensä englantilaiselta harrastelijamatemaatikko ja presbyteeri-pappi Thomas Bayesiltä, jota pidetään Bayesilaisen tilastotieteen isänä.
- Bayesilainen tilastotiede ulottaa todennäköisyyskäsityksen, eli tnjakauman, myös aineistoa koskevien hypoteesien puolle: kuinka todennäköisenä joitain hypoteesia voidaan pitää jo ennen tutkimusaineiston keräämistä?
  - Myös Bayesilaisessa tilastotieteessä hyödynnetään uskottavuusfunktioita, mutta hypoteesien testaus ei perustu niinkään frekventistiseen ajatuksen todennäköisyysistä suhteellisina osuuksina äärettömässä sarjassa.
  - Bayesilaiset perustavat sen sijaan hypoteesien testaamisen tutkimuskysymystä koskevien ennakkokäsitysten päivittämiselle sen jälkeen, kun aineiston on havaittu.
  - Nämä ennakkokäsitykset voidaan kuvata todennäköisyysjakaumana, priorijakaumana, jota päivitetään ns. posteriorijakaumaksi kun aineisto havaitaan. Näin päättely perustuu priorijakauman ja aineiston uskottavuusfunktion väliselle kompromissille!
- Ajatusta ennakkokäsityksistä todennäköisyysinä käytetään niin Bayesilaisen tilastotieteen kriitikkinä kuin puolustuksena.
  - Lopulta olemme kaikki Bayesilaisia: jokaisella on sisäisiä ennakkokäsityksiä, myös tutkijoilla! Nämä ennakkokäsitykset voivat perustua esimerkiksi aiempaan tutkittuun tietoon, mutta myös uskomuksiin.
  - Prioritedon hyödyntäminen tilastollisessa tutkimuksessa on usein perusteltua.
  - Bayesilaista tilastotiedettä tarkastellaan tarkemmin esimerkiksi kursseilla TILM3577 Bayes-päättely sekä TILM3601 Bayes-laskenta.

## 4.3 Tilastolliset mallit, jakaumat ja parametrit

- Tilastolliset mallit perustuvat satunnaismuuttujan mahdollisten tulosvaihtoehtojen todennäköisyksiä kuvaavalle **todennäköisyysjakaumalle**, joka määräät millä todennäköisydellä satunnaismuuttuja saa erilaisia arvoja.
- Toisaalta ajoittain tietyn suureen/ilmiön mallinnuksessa voidaan perustellusti käyttää molempien luokkiin kuuluvien satunnaismuuttuja- ja tilastollisen mallityypin vaihtoehtoja.

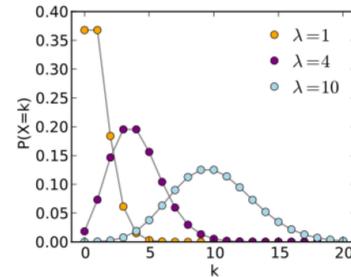
- Esimerkki: Esimerkiksi COVID19-tartuntatapausten lukumäärä Suomessa on periaatteessa diskreetti satunnaismuuttuja, joka saa yksittäisen (kokonaisluku)arvon joka kuukausi, mutta käytännössä luku-määrität ovat tässä tapauksessa sen verran suuria, että niitä (saate-taan) mallintaan jatkova-arvoisena muuttujana.
- Vastaavasti esimerkiksi potilaan jonotusaika päivystyksessä voi peri-aatteessa saada minkä tahansa arvon tietyltä reaalilukuväliltä (tällöin käytettäisiin jatkoviin sm;jiin perustuvia tilastollisia menetel-miä).
- Satunnaismuuttujan mahdolliset arvot, ja täten todennäköisyysjakauma, määräväät myös käytettävän tilastollisen mallin.
  - **Diskreetin satunnaismuuttujan** jakauma voidaan usein esit-tää taulukkomuodossa. Eri arvojen todennäköisyydet muodos-tavat kyseisen satunnaismuuttujan todennäköisyysjakaumanan (**pistetodennäköisyysfunktion**), jota voidaan havainnollistaa esimerkiksi pylväsdiagrammillä.
  - Jatkuvan satunnaismuuttujan  $Y$  arvot muodostavat jonkin reaaliak-selin välin, joka sisältää äärettömän määren lukuja. Tämän vuok-si jatkuvan satunnaismuuttujan jakaumanan esittäminen taulukossa ei ole luontevaa, vaan jakauma esitetään yleensä satunnaismuuttujan **tiheysfunktion** avulla.
    - \* Pistetodennäköisyys- ja tiheysfunktio siis määräväät satunnaismuuttujan mahdollisille arvoille todennäköisyydet väliltä  $[0, 1]$  ja näin voidaan arvioida havaitun aineiston uskottavuutta ja testata siihen liitettäviä hypoteeseja suhteessa estimoituun suuriman uskottavuuden estimaattiin.
- Tilastolliset mallit approksimoivat “todellista” aineiston generoinutta il-miötä. Tilastolliset mallit riippuvat **parametreista** ja keskeinen oletus erityisesti klassisessa tilastotieteessä on, että aineiston generoinutta satun-naisilmiotä kuvaaa jokin vakioinen mutta tuntematon parametriarvo (tai niiden joukko).

### Parametrien estimointi ja niiden testaus

- Satunnaismiötä kuvaava tilastollinen malli perustuu siis johonkin pa-rametriseen todennäköisyysjakaumaan, joka yhdessä havaintojen kanssa määrittää uskottavuusfunktion.
  - Aineistoa kuvaavan tilastollisen mallin uskottavuus pyritään maksimoimaan, mikä tarkoittaa valitun todennäköisyysjakauman sovitta-mista havaintoaineistoon mahdollisimman hyvin.

- Hevosen potkuun kuolleiden Preussin armeijan sotilaiden lukumäärä 20 vuoden aikana
- Guinnes -oluen valmistusprosessin hiivasolujen lukumäärä
- Bakteerien lukumäärä litrassa järvivettä
- Viimeisen 10 vuoden lento-onnettomuuksien lukumäärä

- Kaikille yhteistä: lasketaan **harvinaisten tapahtumien lukumäärä** tietyssä ajassa tai tilavuudessa
- Jakaumalla **parametrit**, joiden arvot vaihtelevat ja jotka halutaan estimoida



Kuva 4.2: Esimerkki: Poisson-jakauman sovelluskohteita ja sen pistetodennäköisyysfunktio eri parametrin arvoilla. Poisson-jakaumaa esitellään tarkemmin alaluvussa 4.5.

- Tässä nk. “suurimman uskottavuuden estimoimissa” aineiston generoiman (oletetun) todennäköisyysjakauman parametriarvot **estimoidaan** (eli arvioidaan) käytettäväni otoksen/aineiston avulla.
- Perusjoukkoa parhaiten kuvaavan (eli “aineiston generoineen”) parametrin arvo pyritään siis estimoimaan aineiston perusteella.
- Parametrien estimoinnin lisäksi usein **testataan** parametreja koskevia oletuksia (eli hypoteeseja).
- Estimointi ja testaus ovat tilastolliseen tutkimukseen liittyvän **tilastollisen päättelyn** keskeisiä välineitä, joiden avulla tutkittavasta ilmiöstä pyritään tekemään johtopäätöksiä siitä kerätyn havaintoaineiston perusteella.
  - Estimoitujen parametrien testaus voi vastata esimerkiksi seuraavalaisiin kysymyksiin:
    - \* Onko suomalaisten miesten keskipituus 180cm?
    - \* Vaikuttaako yliopistokoulutus tulevaisuuden ansioihin?
    - \* Auttaako tietty lääkeaine jonkin sairauden hoidossa?
    - \* Voiko osakemarkkinoiden tuottoja ennustaa?
- Parametrien testaus on osa tilastollista päättelyä, johon palataan tarkemmin luvussa 6

## 4.4 Odotusarvo ja varianssi

- Satunnaismuuttujan todennäköisyysjakauman tietoa voidaan tiivistää tunnuslukuihin, joista keskeisimpäät ovat **odotusarvo**, **varianssi** ja **keskihajonta**.

### Odotusarvo

Satunnaismuuttujan  $Y$  odotusarvo  $E(Y)$  kuvaa satunnaismuuttujan odottavissa olevaa arvoa.

- Muodostamalla satunnaiskokeen tulosten **painotettu keskiarvo**, jossa kunkin tuloksen painona on vastaavan tapauksen todennäköisyys, niin saatua arvoa sanotaan odotusarvoksi  $E(Y)$ .
- Odotusarvo kuvaa jakauman painopistettä.
- Merkinnän  $E(Y)$  käyttö juontaa juurensa englannin kielen sanoihin “odotus”, expectation, ja ‘odotusarvo’, expected value.

### Esimerkki: Odotusarvo

tähän joku esimerkki toisiaan.

- Odotusarvon lisäksi kiinnostuksen kohteena on usein jakauman keskityneisyys (hajaantuneisuus). Ts. kun halutaan puolestaan kuvata satunnaismuuttujan arvojen vaihtelua, tutkitaan todennäköisyysjakauman **varianssia ja keskihajontaa**.

### Varianssi

Satunnaismuuttujan  $Y$  hajontaa voidaan mitata varianssilla

$$\text{Var}(Y) = E\left[\left(Y - E(Y)\right)^2\right],$$

tai sen neliöjuuren eli **keskihajonnan** avulla

$$D(Y) = \sqrt{\text{Var}(Y)}.$$

- Mitä lähempänä nolla keskihajonta ja varianssi ovat, sitä todennäköisempää on, että satunnaismuuttujan arvo on lähellä odotusarvoa. - Merkintöjen  $\text{Var}(Y)$  ja  $D(Y)$  taustalla on englannin kielen sanat variance (varianssi) ja deviation, joka tarkoittaa poikkeamaa, hajontaa.

- Odotusarvon ja varianssin (keskijajonnan) tavanomaiset estimaattorit ovat otoskeskiarvo ja otosvarianssi (otoshajonta), joihin palataan vielä myöhemmin.

## 4.5 Joitain jakaumia

Tarkastellaan seuraavassa muutamia keskeisiä tilastollisia jakaumia. Esittelemme ensin keskeisintä jatkuvien satunnaismuuttujien jakaumaa, normaalijakaumaa, ennen muutamien diskreettien satunnaismuuttujien jakaumia.

### 4.5.1 Normaalijakauma

- Jos satunnaismuuttuja  $Y$  noudattaa **normaalijakaumaa** odotusarvolta  $E(Y) = \mu$  ja varianssilla  $\text{Var}(Y) = \sigma^2$ , niin tällöin merkitään  $Y \sim N(\mu, \sigma^2)$ .
- $Y$ :n tiheysfunktio on muotoa (ks. kuva alla)

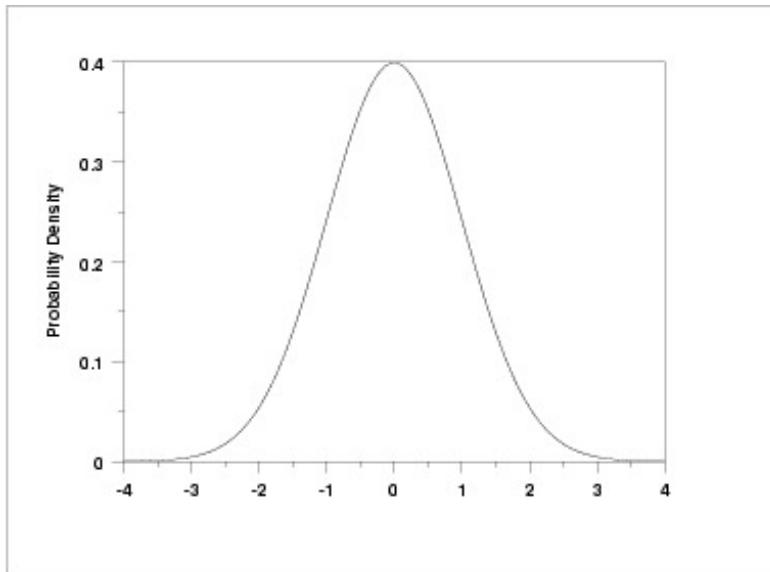
$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2},$$

jossa  $e$  viittaa Neperin lukuun  $e \approx 2,71828$

- Ylläoleva tf. määrittelee parven normaalijakaumia kun parametreille (vakioille)  $\mu$  ja  $\sigma^2$  annetaan erilaisia arvoja. Nämä kaksi parametria määrävät normaalijakauman tarkemman muodon.

#### Esimerkki: Miesten pituus

- Tutkitaan miesten pituutta hyvin määritellyssä joukossa, kuten varusmiespalvelusta tietynä vuonna suorittavien joukossa.
  - Pituus on ominaisuus, jonka voidaan nähdä määrytyväksi monista perintö- ja ympäristötekijöistä. Pituutta voidaan siis pitää satunnaismuuttujana.
  - Oletetaan, että pituus noudattaa normaalijakaumaa. Näin ollessa  $Y$  on valitun miehen pituus ja  $Y \sim N(\mu, \sigma^2)$ .
- Tuntemattomien parametrien  $\mu$  ja  $\sigma^2$  tulkinta:
  - Odotusarvo  $\mu = E(Y)$  on satunnaisesti valitun miehen pituuden odottettavissa oleva arvo.
  - Varianssi  $\sigma^2 = \text{Var}(Y) = E[(Y - \mu)^2]$  kuvaa valitun miehen pituuden odotusarvostaan määritetyn poikkeaman (kes-



Kuva 4.3: Normaalijakauma

kihajonnan) neliön odotettavissa olevaa arvoa (kuvaten ts. pituksien jakauman keskityneisyyttä/hajaantuneisuutta pituksien odotusarvon ympärillä).

#### 4.5.2 Bernoulli-, binomi- ja Poisson-jakauma

- **Bernoulli-jakauma** on todennäköisyysjakauma, jossa satunnaismuuttujalla  $Y$  on kaksi mahdollista tulosvaihtoehtoa  $Y = 1$  tai  $Y = 0$ .
  - Yleensä  $Y = 0$  tarkoittaa, että jokin tapahtuma ei tapahdu ja  $Y = 1$  että tapahtuu.
  - Todennäköisyys tapahtumalle  $Y = 1$  on  $P(Y = 1) = p$  ja vastaavasti vastatodennäköisyys  $P(Y = 0) = 1 - p$ .
  - Bernoulli-jakaumaa merkitään  $Y \sim B(p)$ , jossa siis  $0 < p < 1$ .
  - Bernoulli-jakauman **pistetodennäköisyysfunktio** on muotoa

$$f(y; p) = P(Y = y) = p^y(1 - p)^{(1-y)},$$

jossa  $y$  on sm:n  $Y$  realisaatio (havaittu arvo) ja parametri  $p$  on tuntematon (voidaan estimaoida otoksen avulla).

- Bernoulli-jakauman odotusarvo  $E(Y) = p$  ja varianssi  $\text{Var}(Y) = p(1 - p)$ .

- **Binomijakauma**

- Olkoon  $Y_1, \dots, Y_n$  riippumattomia satunnaismuuttujia ja  $Y_i \sim B(p)$ ,  $i = 1, \dots, n$ .
- Jos  $X = Y_1 + Y_2 + \dots + Y_n$ , niin  $X \sim \text{Bin}(n, p)$ . Ts. sm.  $X$  noudattaa **binomijakaumaa** parametrein  $n$  ja  $p$ .
- Pistetodennäköisyydfunktio:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{(n-k)}.$$

- Jakauman odotusarvo  $E(X) = np$  ja varianssi  $\text{Var}(X) = np(1 - p)$ .
- Binomijakaumalla kyetään vastaamaan mm. kysymykseen millä todennäköisyydellä  $n$ :n kokoisessa otoksessa tapahtuu  $k$  onnistumista.

**Esimerkki: Miesten lukumäärä Saksin osavaltion perheissä 1876–1885<sup>a</sup>**

Vuosien 1876–1885 aikana Saksin osavaltiossa rekisteröitiin yli neljä miljoonaa syntynytä lasta. Tällöin vanhempien tuli ilmoittaa lapsen suku puoli (mies tai nainen) heidän syntymätodistuksensa. Myöhemmässä tutkimuksessa tutkittiin tarkemmin 6115 perhettä, joissa asui 12 lasta ja tarkemmin miesten (poikien) lukumäärää näissä perheissä.

Seuraavassa taulukossa taulukoidaan miesten (poikien) lukumäärät näissä 12 lapseen perheissä:

Miesten lkm.	( $k$ )	0   1   2   3   4   5   6   7   8   9   10   11   12	Perheiden lkm.
( $n_k$ )	3   24   104   286   670   1033   1343   1112   829   478   181   45   7		

Tarkasteltava jakauma esitetään vielä erikseen allaolevassa kuviossa.

Tässä tilantessa mielenkiinnon kohteena saattaisi olla hypoteesi, jonka mukaan pojан (miehen) syntymätodennäköisyys  $P(\text{mies}) = p$  on  $p = 0.5$ .

<sup>a</sup>Ks. tarkemmin esimerkki 3.2 kirjassa (s. 67-68) Friendly, M., ja D. Meyer (2015). *Discrete Data Analysis with R. Visualization and Modeling Techniques for Categorical and Count Data*. Chapman & Hall/CRC.

### Poisson-jakauma

- Jos satunnaismuuttuja  $Y$  on Poisson-jakautunut, merkitään  $Y \sim P(\lambda)$ , jossa parametri  $\lambda > 0$  on Poisson-jakauman parametri, jota kutsutaan myös ajoittain intensiteettiparametriksi.
- Poisson-jakaumaa voidaan käyttää tilanteissa, joissa sm.  $Y$  on jokin lukumäärä ja sen pistetodennäköisyysfunktio on muotoa

$$P(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

- Odotusarvo ja varianssi ovat Poisson-jakauman tapauksessa samat:  $E(Y) = \text{Var}(Y) = \lambda$ .

#### Esimerkki: Poisson-jakauma

Tarkastellaan Englannin Valioliigakauden 1995–1996 otteluissa tehtyjä maalimääriä. Valioliiga (The F.A. Premier League) on korkein Englannin jalkapalloliigan sarjataso, jossa ensi kerran juuri kaudella 1995–1996 20 joukkueita (aiemmin Valioliigan perustamisen kauden 1992–1993 alussa 22 joukkueita) pelasivat keskenään kerran toisiaan vastaan koti- ja vieraskentällä. Otteluita oli siis yhteensä 380.

Tämä esimerkki perustuu edellä mainittuun Friendlyn ja Meyerin (2015) kirjan esimerkkiin 3.9 (s. 78–79), joka vastaavasti perustuu Alan J. Leen (1997) artikkeliin<sup>a</sup>, jonka esittämään kysymykseen (hypoteesiin) vastaus on tietenkilä ilmeinen! Näin ollen seuraavassa tarkastellaankin kotijoukkueiden ja vierasjoukkueiden maalintekointensiteettiä Poisson-jakaumaan perustuen. Seuraavassa emme siis pyri mallintamaan tietyn spesifin ottelun loppulokosta vaan tarkastelemme “keskimäärisen” kotijoukkueen ja vierasjoukkueen “edustavaa” ottelua.

Seuraava taulukko raportoi tehtyjen maalimäärien jakaumat pelatuissa 380 ottelussa. Neljän tai yli neljän maalin tapaukset kirjataan 4+:nä maalina. Ts. esim. kys. kauden loppulokset *Blackburn Rovers - Nottingham Forest* 7-0 ja *Bolton Wanderers - Manchester United* 0-6 tulevat aineistoon tuloksina 4+ vs. 0 ja 0 vs. 4+.

Kotij. maalien lkm.	Vierasj. maalien lkm.					Yht.
	0	1	2	3	4+	
0	27	29	10	8	2	76
1	59	53	14	12	4	142
2	28	32	14	12	4	90
3	19	14	7	4	1	45
4+	7	8	10	2	0	27
Yht.	140	136	55	38	11	380

Olettamalla, että koti- ja vierasjoukkueen todennäköisyys tehdä maali ottelun aikana on vakio, niin tällöin koti- ja vierasjoukkueen ottelun aikana tekemien maalien lukumäärää (ilman edellä käytettyä maalimäärien ”katkaisua” neljään) voidaan melko hyvin approksimoida oletuksella, että nämä lukumäärität ovat Poisson-jakautuneita. Ts.  $Y_i^H \sim P(\lambda_H)$  on sm., joka kuvailee  $i$ :n ottelun kotijoukkueen tekemien maalien lukumääriä ja intensiteettiparametrin  $\lambda_H$  arvon määrittäminen kuuluu tilastollisen päättelyn ja erityisesti estimointiteorian piiriin. Vastaavasti vierasjoukkueen maalimäärität:  $Y_i^A \sim P(\lambda_A)$ .

Osoittautuu, että parametreille  $\lambda_H$  ja  $\lambda_A$  saatavat estimaatit ovat  $\lambda_H = 1,49$  ja  $\lambda_A = 1,06$  ja ne vastaavat tässä yksinkertaistetussa tilanteessa koti- ja vierasjoukkueen keskimääriäisiä maalimääriä:

	kotijoukkue (home)	vierasjoukkue (away)	Yht.
keskiarvo	1,486	1,063	2,550
varianssi	1,316	1,172	2,618

Tuloksista voidaan siis päätellä, että kotijoukkueen (odottavissa oleva) maalimääriä on vierasjoukkuetta korkeampi (osoittaen osaltaan kotiedun merkitystä jalkapallossa). Lisäksi edellä todetun Poisson-jakauman teoreettisten ominaisuuksien mukaisesti keskimääriäiset maalimäärität ovat läheillä niiden variansseja, mikä osoittaa osaltaan tarkasteltavassa yksinkertaisessa tilanteessa, että Poisson-jakaumaan perustuva jakaumaoletus on kelvollinen.

On syytä todeta lopuksi, että tämän vahvasti yksinkertaistetun tilanteen sijaan tilastotieteessä on laaja ja kasvava kirjallisuuden haara jalkapalloa ja muuta urheilua koskevien tilastollisen menetelmien saralla. Nämä vaativat kuitenkin syvällisemmän ymmärryksen kannalta jälleen huomattavasti laajempia tilastotieteen (aine- ja syventäviä) opintoja.

<sup>a</sup>Alan J. Lee (1997). Modeling Scores in the Premier League: Is Manchester United Really the Best? *Chance* 10(1), 15-19.

## 4.6 Sattuman rooli tieteenteossa: Vale-emäväle-tilasto?

Eriyisesti nykypäivänä ei-tieteellinen tieto ja tarkoituksellinen disinformaatio, joita perustellaan heppoisin havainnoin, leviävät internetissä kulovalkean tavoin. On tiedeyhteisön ja tutkijoiden moraalinen vastuu taistella näitä uskomuksia vastaan popularisoimalla tutkimustietoa, mikä saattaa ajoittain jopa pahentaa ongelmaa, sillä popularisoinnissa päteviltäkin tutkijoilta voi unohtua satunnaisuuden voima.<sup>3</sup>

- Ei-tieteellinen tieto on juurtunut syvälle ja tutkijat pyrkivät taistelemaan tästä vastaan **popularisoimalla tiedettä**.
- Kuten todettua, tilastollisessa tutkimuksessa mielenkiannon kohteena on satunnaisilmiöiden tutkiminen ja erityisesti systemaattisen ja satunnaisen vaihtelun (signaalin ja kohinan) erottaminen sekä muuttujien välisen riippuvuuden tutkiminen.
  - Kiinnostuksen kohteena on siis hyvin harvoin vain jokin yksittäinen tunnusluku, kuten keskiarvo, varianssi tai korrelaatio (palaamme näihin myöhemmin luvussa 6).
  - Tieteen popularisointi on yksi tutkijoiden ja yliopistojen tiedeyhteisön tärkeimmistä yhteiskunnallisista tehtävistä, mutta valitettavan usein se typistyy yksittäisen viimeisimmän tutkimustuloksen esitteelyksi.
- Yliopistoyhteisössä kuitenkin luonnollisesti luotamme kumuloituneeseen tutkittuun tietoon ja tiedämme, että **yksittäinen tutkimus on vasta hyvä alku**.
  - Ihmistieteitä, kuten ilmeisesti erityisesti psykologiaa sekä osin myös muiden ohella lääke- ja taloustiedettä, on viimeisen vuosikymmenen ajan puhuttanut paljon niin sanottu **replikaatiokriisi**, sillä useaa arvostettuakaan tutkimusta ei ole saatu **toistettua eli replikoitua**.
  - On ymmärettäväää, että replikaatiokriisi, varsinkin jos se on (alakohtaisesti) laajalle levinyttä, murentaa kansalaisten luottamusta tieteellisiin tuloksiin.
  - Toistettavuus on yksi tutkimuksen peruskriteereistä, joka erottaa tieteellisen tiedon muista tietolähteistä, joten sen puuttuminen herättää ymmärettävästi huolta tieteellisen prosessin toimivuudesta.
  - Replikaatiokriisiin voi kuitenkin myös tulkita toisin: ilman kriittisyyttä omia (ja muiden) tuloksia kohtaan, ei mitään kriisiä olisikaan, joilla silkkä sen olemassaolo on osoitus tieteellisen prosessin toimivuudesta.

---

<sup>3</sup>Tämä jakso perustuu osin psykometriikan yliopisto-opettajan Jari Lipsasen blogiin vuodelta 2021.

#### *4.6. SATTUMAN ROOLI TIETEENTEOSSA: VALE-EMÄVALE-TILASTO?61*

- Kun tuntee ja tunnistaa sattuman voiman ja ymmärtää kaikki mahdolliset satunnaisuuden lähteet, jotka altistavat tutkimusprosessin virheille, tulee samalla ymmärtäneeksi että eri tavoin koeteltu, useassa tutkimuksessa kumuloitunut tieto tulisi olla kaiken tieteen popularisoinnin keskiössä yksittäisten, mahdollisesti uusien ja yllättävien tutkimustulosten sijaan.
  - Tähän mennessä olemme jo oppineet, että tälle on myös vahvat tilastolliset perustelut: satunnaisen tiedon maailmassa mikään ei ole täysin varmaa, ei edes kaikkein edistyneimpien tilastomenetelmien avulla!



## Luku 5

# Tilastolliset aineistot, niiden kerääminen ja mittaaminen

Edellisessä luvussa käsiteltiin tilastotieteen suhtautumista satunnaisilmiöihin. Tässä luvussa tarkastelemme lähemmin miten reaalimaailman satunnaisilmiöstä kerätään tietoa ja miten niitä voidaan mitata. Tilastotieteen perusoppimääärä rakentuu ajatukselle ilmiöiden tutkimisesta rajallisen ja epävarman tiedon valitessa. Käytännössä tämä tarkoittaa sitä, että tutkimuksen kohteena ovat rajalliset aineistot sisältävät niin systemaattista kuin satunnaisuudesta johtuvaa vaihtelua. Tilastollisten menetelmien avulla pyrimme erottamaan systemaattisen vaihtelon satunnaisesta sekä tekemään tilastollista päättelyä aineiston generoimasta mekanismista. Lyhyesti tämä tarkoittaa aineiston systemaattisen vaihtelon tilastollista mallintamista ja sen parametrien estimointia otoksesta, joka kattaa vain (pienien) osajoukon koko populaation (perusjoukon) tilastoyksiköisän.

Voidaksemme tehdä uskottavaa päättelyä “havainnoista parametreihin”, tulee otoksen olla riittävä **edustava**. Tämän luvun keskeisin oppi onkin, että miten otanta tulisi suorittaa, jotta havaintoaineisto olisi **edustava otos** populaatiosta, silloin kun aineisto kerätään otannalla. Vaikka aineiston hankinta vaatii yleensä runsaasti käytännön työtä, kannattaa se tehdä huolellisesti, sillä huonosti toteutetun otannan vuoksi tutkimusongelman kannalta keskeisiä johtopäätöksiä ei voida tehdä!

## 5.1 Kertausta: Data eli aineisto

- **Tilastollinen tutkimus** aloitetaan tutkimusaineiston keruun suunnitellulla.
- Kertauksen vuoksi: tilastollinen tutkimusaineisto (havaintoaineisto) koostuu tilastoyksiköiden populaatiosta havaittuista tilastomuuttujien arvoista.
- Havaintoaineisto voidaan koota taulukoksi, johon listataan tilastoyksiköt riveille ja tilastomuuttujat sarakkeisiin. Jos havaintoaineisto koostuu  $n$  tilastoyksiköstä, joista jokaisesta on kerätty esim.  $m$  tilastomuuttujasta havainnot, niin havainnot voidaan kirjoittaa taulukon muotoon

	tilastomuuttuja 1	tilastomuuttuja 2	...	tilastomuuttuja $m$
tilastoyksikkö 1	$x_{1,1}$	$x_{1,2}$		$x_{1,m}$
tilastoyksikkö 2	$x_{2,1}$	$x_{2,2}$		$x_{2,m}$
...	...	...	...	...
tilastoyksikkö $n$	$x_{n,1}$	$x_{n,2}$		$x_{n,m}$

Tässä siis rivillä  $i$  on  $i$ . **tilastoyksikön** havainto ja sarakkeessa  $j$  on  $j$ . tilastollisesta muuttujasta havaitut arvot  $x_{i,j}$ . Ts. yhdellä rivillä on yhden tilastoyksikön tiedot kaikista tilastomuuttujista ja yksi sarake on kaikkien tilastoyksiköiden tiedot yhdestä tilastomuuttujasta.

- Usein (varsinkin parhaillaan kiihtyvällä vauhdilla) kerättävät havaintoaineistot ovat niin suuria, ettei edellisenkaltaisesta havaintotaulukosta voida usein suoraan tarkastelemalla nähdä aineiston pääpiirteitä.
  - Tällöin voi olla tarpeen luokitella aineistoa taulukon muodostamiseksi.
  - Luokittelussa on kysymys aineiston tiivistämisestä kohtuullisen koiseksi ja havainnollisempaan muotoon. Luokittelussa tilastomuuttujan arvot sijoitetaan eri luokkiin siten, että yhden tilastomuuttujan arvo voi kuulua vain yhteen luokkaan. Luokka ilmoitetaan yleensä luokkavälinä, kuten reaalilukuvälinä. Esimerkiksi henkilön ikä on tapania luokitella ikäjakauman kuvaamisessa 10-vuotislukkiin (15-24, 25-34, ...), vaikka periaatteessa ikä voitaisiin ilmoittaa minuutinkin tarkkuudella.
  - Luokkien lukumäärään vaikuttavat muun muassa tilastomuuttujan arvojen vaihteluväli ja havaintoaineiston laajuus. Luokittelussa pyritään siihen, että luokkien lukumäärä saadaan tarvittaessa luokkia yhdistämällä kohtuulliseksi ja että luokat valitaan tasavälistä eli

siten, että kahden peräkkäisen luokan alarajojen erotus on vakio. Kun aineistoa luokitellaan, aineiston luettavuus paranee mutta toisaalta osa tiedoista menetetään eivätkä yksittäiset havaintoarvot ole enää tiedossa.

- Emme vielä tällä kurssilla etene tämän pidemmälle tilastografiikan esittämisessä ja siihen liittyvissä pohdinnoissa. Muun muassa tilastollisen päättelyn peruskurssi (TILM3555) vastaa näihin kysymyksiin tarkemmin. Graafiset menetelmät ovat joka tapauksessa erittäin tärkeä osa aineiston havainnollistamista. Kuvat helpottavat aineiston tulkitsemista ja toimivat usein perusteltuna lähtökohtana monimutkaisempien tilastollisten mallien (ja algoritmien) sovittamiselle.

- Kvantitatiivisen tutkimuksen aineistoksi kelpaa periaatteessa kaikki havaintoihin perustuva informaatio, joka on **mittauksen** avulla muutettavissa numeeriseen muotoon.
  - Havaintoyksiköiden tilastollisten muuttujien numeerisia arvoja kutsutaan **havaintoarvoiksi** tai **havainnoiksi**.
  - Kaikki havaitut tilastolliset muuttujat eivät ole aina mielenkiintoisia. Tutkimuksen kannalta mielenkiintoisia muuttuja kutsutaan **tutkimusmuuttujiksi**, joiden lisäksi havaintoaineisto pitää mahdollisesti sisällään **taustamuuttujia**.
    - \* Esimerkiksi, jos tutkimuksella halutaan tietoa suomalaisen aikuisväestön mielipiteistä, havaintoyksikköinä ovat aikuisväestöön kuuluvat henkilöt. Jos halutaan tietoa suomalaisista kunnista, havaintoyksikköinä ovat Suomen kunnat jne.
    - \* Ensimmäisessä tapauksessa tilastollisina muuttujina on aikuisväestön mielipiteet, joita voidaan selvittää esimerkiksi kyselytutkimuksella. Toisaalta voidaan myös kerätä taustamuuttujiksi haastatellusta muita tietoja, kuten asuinpaikka, ikä ja ammatti.
  - Kaikkia mielenkiintoisia muuttuja ei kuitenkaan välttämättä voida havaita, eli niille ei voida määrittää numeerista arvoa.
  - Tällöin puhutaan nk. **latenteista muuttujista**, eli muuttujista joita ei suoraan havaita mutta joiden oletetaan vaikuttavan havaittavien muuttujien taustalla. Latentteja muuttuja voidaan rakentaa tilastollisten mallien avulla käyttäen hyödyksi niihin liittyviä havaittuja muuttuja.
  - Latentteja muuttuja ovat esimerkiksi elämänlaatu, onnellisuus, konservatiivisuus, yms.

## 66LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Tilastollinen tutkimus voi olla joko **kokonaistutkimus** tai **otantatutkimus**.

### Kokonaistutkimus

Kokonaistutkimus on tutkimus, jossa tutkitaan kaikki tutkimuksen kohde-teenaa olevan perusjoukon alkiot, ts. kaikki ajateltavissa olevat kohteet tutkitaan.

- Kokonaistutkimus on yleinen tutkimustapa silloin, kun kohdeperusjoukko on selvästi määritelty ja sen alkioita koskevat tilastolliset muuttujat ovat helposti mitattavissa.
- Esimerkiksi jos tutkitaan Suomen kuntia, niin kokonaistutkimuksessa tutkitaan kaikki kunnat, joista on helppoa kerätä mielenkiinnon kohteena olevia tilastollisia muuttuja useimmissa tilanteissa.
- Toisaalta jos tutkitaan jonkin lääkeaineen vaikutuksia ihmisiin, niin kokonaistutkimuksessa tutkittaisiin jokainen ihminen erikseen. Selvää on, että tällainen kokonaistutkimus olisi liian vaikeaa toteuttaa.

### Otantatutkimus

**Otantatutkimuksessa** tutkimus kohdistetaan johonkin (populaation-/perusjoukon) osajoukkoon, joka pootimitaan sopivaa **otantamenetelmää** käyttäen (ks. alaluku 5.5) ja populaatiota/perusjoukkoa koskevat johtopäätelmät tehdään tähän otokseen perustuen.

- Otantatutkimus on usein luonnollinen valinta, sillä koko populaation tutkiminen ei useinkaan ole mahdollista tai kannattavaa.
  - Esimerkiksi aseiden patruunoita valmistava tehtailija ei voi tutkia toimivatko kaikki ammukset. Myöskään valaisimien valmistaja tuskin tekee kokonaistutkimuksia valmistamiensa tuotteiden kestoajan selvittämiseksi.
- Perusjoukosta otokseen poimittuja alkioita kutsutaan **otosyksiköiksi** ja niiden muodostama osajoukko, eli **otos**, on se osa perusjoukkoa, joka tutkitaan tutkimusaineiston keräämisen jälkeen.
  - Lääketutkimusta tehdäänkin poikkeuksella otantatutkimuksena (ja kontrolloituina kokeina, ks. alempaa), jolloin lääketä testataan vain osajoukolla koko ihmispopulaatiosta ja tämän osajoukon alkiot ovat otosyksiköitä.
  - Näin toimimalla, ja riittävän edustavalla otoksella, saadaan kuitenkin tarpeeksi tietoa lääkeaineen vaikutuksista ja tulokset voidaan yleistää populaatiotasolle ja lääke ottaa käyttöön.

- Otantatutkimus on halvempi kuin kokonaistutkimus ja tulokset saadaan nopeammin!

- Otantatutkimuksessa keskitytään siis perusjoukkoa edustavan pienemmän, mieluusti satunnaisesti valitun otoksen tutkimiseen.
  - Otantatutkimuksissa tiedot kerätään useimmiten haastattelemella, kirjallisella/sähköisellä kyselyllä tai suoraan tietorekistereistä. Tiedonkeruun toteuttaminen (eri sovelluksissa) määräää osaltaan käytetään otantamenetelmän.
  - Teoriassa äärelliseen perusjoukkoon kohdistuvat kokonaistutkimukset voidaan aina tulkita otantatutkimuksiksi (perusjoukko tulkitaan otokseksi hypoteettisesta äärettömästä perusjoukosta)!
    - \* Esimerkiksi Galilein tekemät painovoiman vaikutusta kappaleiden putoamisaikaan liittyneet mittaukset. Koetuloksia (mittauksia) voidaan pitää otoksenä äärettömästä mahdollisten koetulosien joukosta. Tällöin ainoa mahdollisuus ilmiön tutkimiseen on käyttää otantaa.
- Otantatutkimuksen tulokset voivat olla luotettavampia kuin kokonaistutkimuksen.
  - Otantatutkimuksessa voidaan panostaa enemmän huolelliseen ja tarkkaan mittaamiseen sekä valitun otoksen tavoittamiseen.
  - Kokonaistutkimuksessa vastauskato ja tarkasteltavan populaation valintavirhe ovat mahdollisia siinä kuin otantatutkimuksessakin.
- Otantateoria on yksi tilastotieteen keskeisimpää oppeja ja tarjoaa teoreettisen kehikon empiiristen tutkimusten tulosten yleistämiseen. Tarkasteluaan siis tarkemmin otannan ideaa ja toteuttamista seuraavassa alaluvussa.

## 5.2 Otannan idea

- Otantatutkimuksen (karkeat) suunnittelu- ja työvaiheet ovat seuraavat:
  1. Tavoitteiden asettaminen
  2. Perusjoukon (populaation) asettaminen
  3. Kehikko
  4. Kerättävän informaation sisältö (mitä tietoa todella tarvitaan, mitä voidaan jättää pois, suunnitellaan kysymykset ja mahdollinen kyselylomake)

## 68LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

5. Otoskoon määrittäminen
  6. Suoritetaan otoksen poiminta, tietojen keräys ja tarkastus
  7. Aineiston taulukointi ja analysointi
  8. Raportin laatiminen
- Otantatutkimuksessa ajatuksena on siis poimia **edustava otos** siitä populaatiosta (perusjoukosta), joka on mielenkiinnon kohteena eli jota halutaan tutkia ja josta halutaan tietoja.
    - **Tavoiteperusjoukko** on joukko, johon otannan myötä saatavat tutkimustulokset halutaan yleistää. Toisin sanoen, se mistä haluamme tietoja määrään populaation.
    - **Kohdeperusjoukko** on joukko, jota koskevia tietoja halutaan kerätä.
      - \* Esimerkiksi äänestysikäiset Suomen kansalaiset.
      - \* Usein tavoiteperusjoukko = kohdeperusjoukko.
      - \* Tavoiteperusjoukko voi joskus olla laajempi (esim. "ihmiset" vs. "suomalaiset").
  - Tutkimuksessa (edustavaan) otokseen poimitut tilastoyksiköt, näiden tilastolliset muuttujat ja niiden arvot muodostavat **otosaineiston** eli siis tutkimus- tai havaintoaineiston (datan).
    - Tutkimuskysymykseen vastatakseen tutkija valitsee sopivan tilastollisen mallin ja estimoii sen parametrit tähän otokseen perustuen.
    - Perusoletuksena on otoksen ja valitun tilastollisten mallin pohjalta suoritettavan tilastollisen päätelyn **yleistettävyys koko populaatioon**.
    - Otos valitaan **otantaa** ja erilaisia **otantamenetelmiä** hyödyntäen pyrkien varmistamaan otoksen **edustavuus** (perusjoukko pienoisessa, ks kuva ??(fig:otanta)).

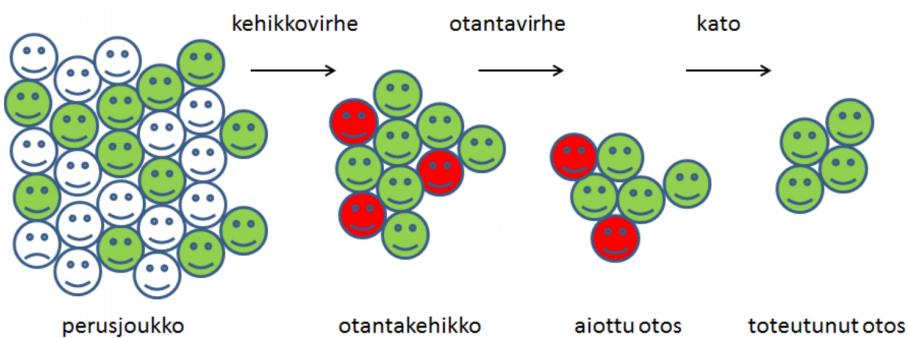
### Edustavuus

Tutkimukseen valitut yksiköt edustavat koko populaatiota, ts. tutkimukseen valittu osajoukko kuvailee perusjoukon ominaisuuksia kattavasti.

- Keskeistä tutkimuksen ja sen edustavuuden kannalta on, että tutkija osaa kerätä sisällöllisesti ja määrellisesti **sopivan kokoisen** aineiston.
- Tietyn otoksen edustavuutta arvioidessa voi käyttää apuna seuraavia kysymyksiä:
  - Miksi päädyttiin tämän kokoiseen otokseen?
  - \* **Otoskoko** vaikuttaa siihen miten hyvin otoksesta tehdyt johdotpääökset voidaan yleistää koskemaan koko perusjoukkoa, ts. kuinka luotettavia ne ovat. Tämä johtuu siitä että yksittäisten

otosyksiköiden ominaisuudet saattavat vaihdella suuresti ja kasvattamalla otoskokoa perusjoukon systemaattiset piirteet tulevat otoskoon kasvaessa yhä paremmin esille. Kun otoskoko vastaa populaation kokoa, on kyseessä tieteenkin kokonaistutkimus, joka kertoo kaiken perusjoukosta. Otoskoon valintaan ja määräämiseen palataan myöhemmin luvussa 6.

- Käytettiinkö apuna tilastotieteellisesti vankkaa suunnittelua otoskoon määrittämiseksi ja/tai miten pyrittiin varmistamaan tutkimuksen kannalta tärkeisiin analyysiryhmiin kuuluvien riittävä määrä ai-neistossa?
- Harkittiinko muita otantamenetelmiä ja miksi päädyttiin juuri käytössä olleeseen menetelmään?
- Edustavuuteen vaikuttaa keskeisesti se, millä tavoin otanta pystytään suorittamaan, ts. mihin kohdeperusjoukkoon otanta kohdistetaan.
  - **Kehikkoperusjoukko** on rekisterin, luettelon tms. peittämä osa kohdeperusjoukkoa. Kyseessä on siis se osa kohdeperusjoukkoa, josta otanta ylipäänsä pystytään suorittamaan.
  - **Otantakehikon alipeitto** esiintyy, kun otantakehikosta puuttuu osa kohdeperusjoukon alkioista (esim. tutkimus suoritetaan puhelin-haastattelulla, mutta osa aiottuun otokseen kuuluvista haastateltavista ei omista puhelinta).



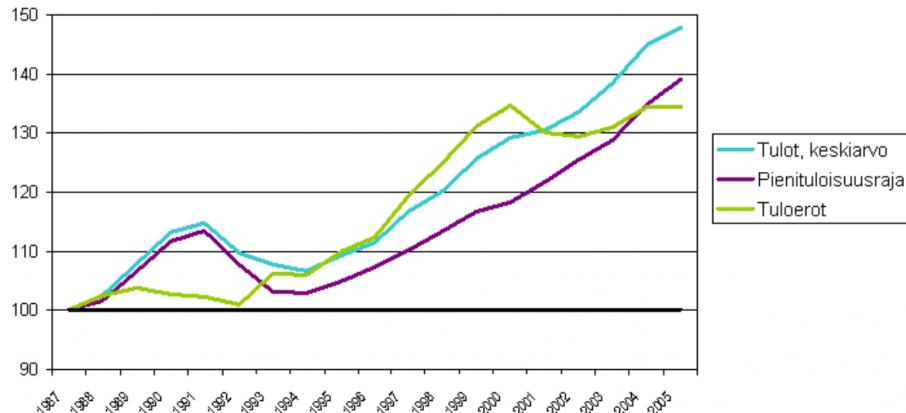
Kuva 5.1: Otannan idea.

- Edustavan otoksen avulla on mahdollista tehdä perusjoukkoa koskevaa tilastollista päätelyä, sillä otos kuvailee perusjoukon ominaisuuksia riittävän hyvin. Tämä on yksi tilastotieteen keskeisimpia oppeja mutta myös kriittisen tiedelukutaidon ja arkijärjen kannalta tärkeää.

## 70LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

### Esimerkki: Kotitalouksien tulot, tuloerot ja pienituloisuusajan kehitys 1987-2005 (Tilastokeskus)

- Tilastotyksikkö kotitalous, joten kaikkien kotitalouksien tutkiminen (kokonaisutkimus, ks. alla) olisi vaikeaa ja aikaavievää.
- Tutkittavaksi valitaan vain muutama tuhat kotitaloutta (ts. otantatutkimus) ja selvitetään näiden tulot.
- On mahdollista tehdä **kaikkia** suomalaisia kotitalouksia koskevia johtopäätöksiä, jos tutkitut yksiköt olivat **edustava otos** suomalaisista kotitalouksista. Ts. osajoukko koskevat päätelmat voidaan yleistää koskemaan perusjoukkoa, mikäli osajoukko on edustava otos perusjoukosta.



Kuva 5.2: Tuloerot.

### 5.3 Tilastollisten muuttujien mittaaminen ja mitta-asteikot

#### Mittaaminen

- Tilastotieteellinen tutkimus perustuu aina mitattaviin satunnaisilmiöihin: tavoitteena on mittamalla liittää jokin luku ilmiötä kuvaavaan ominaisuuteen, ts. mitata kyseisen satunnaismuuttujan havaittua arvoa.
- Kumpaa tahansa tutkimusotetta (kokonais- tai otantatutkimus) noudattaessa tietojen keräämisessä on olennaisena osana kohteiden ominaisuuksien **mittaaminen**.

### 5.3. TILASTOLLISTEN MUUTTUJIEN MITTAAMINEN JA MITTA-ASTEIKOT71

- Mittaaminen vaatii aina mittauksen kohteen, hyvin määritellyn mitattavan ominaisuuden ja **mittarin**, joka liittää mielekkääät lukuvat mitattavaan ominaisuuteen.
- Eriaiset mittarit heijastavat ilmiön ominaisuuksia eri tavoin ja eri tarkkuudella
  - \* Esimerkiksi jos tutkitaan opiskelijoiden pituuden kehitystä niin mitataan pituutta eri aikoina. Pituudet voidaan mitata senttimetreissä, metreissä, kilometreissä tai vaikkapa tuumissa.
  - \* Mittari on hyvä jos sen antama mittaus on
    - (i) **validi** eli mittaus esittää oikein mitattavaa ominaisuutta (senttimetri mittaa pituutta, gramma ei) ja
    - (ii) **luotettava** eli mittaus on **harhaton ja toistettavissa**.
  - \* Määritellään nämä termit vielä erikseen, sillä ne ovat keskeisiä tilastotieteessä.

#### Harhattomuus

Mittari on harhaton, jos se ei systemaattisesti ali- tai yliarvioi mitattavan ominaisuuden määräää.

- Harhaton mittari siis antaa keskimäärin oikeita mittauksia mitattavasta ominaisuudesta.
- Harhattomuutta pidetään myös hyvänä ominaisutena tilastollisten malleiden parametrien estimaattoreille. Tähän palataan myöhemmin luvussa 6.

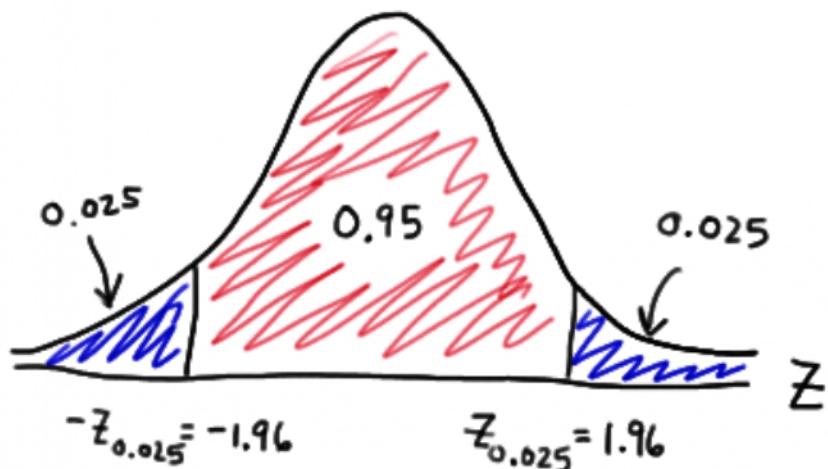
#### Toistettaavuus

Mittari on toistettava, jos se tuottaa keskimäärin samanlaisia mittauksia samanlaisista otoksista eli se on johdonmukainen ja mittausvirheet ovat pieniä.

- Huonosti toistettava mittari antaa tilastoysiköiden samankaltaisille ominaisuuksille hyvin erilaisia arvoja riippuen otoksesta.
- **Mittausten reliabilitettiluotettavuutta** arvioidessa voidaan pohdita esimerkiksi seuraavia kysymyksiä:
  - Kuinka hyvin mittaustulokset ovat toistettavissa, kuinka paljon niissä on ei-sattumanvaraisuutta?
  - Mittausten validiteetti: kuinka hyvin pystyytiin mittaamaan sitä, mitä oli tarkoitus mitata?

## 72LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Kun mittaaminen on luotettavaa ja validia, tutkimusaineisto on **sisäisesti luotettavaa**.
- Aineiston **ulkoinen luotettavuus** toteutuu silloin, kun tutkittu otos edustaa perusjoukkoa eli on edustava. Validi mittaaminen ei pelasta epäedustavaa otosta!
- Jokaisen tutkimuksen tulosten luotettavuuden perusteena on käytetty aineisto, kuinka se on hankittu ja mistä lähteestä. Kun käytetään luotettavaksi havaittuja mittareita, voidaan kustakin aineistosta laskea erikseen tunnuslukuja mittauksen luotettavuudelle. Esimerkkinä **luottamusväli**:
  - Luottamusväliä käytetään määrittämään estimaatin luotettavuutta.
  - Väli, joka vaihtelee otoksesta toiseen ja joka usein sisältää mielenkiinnon kohteena olevan parametrin, kun otantakoetta toistetaan!



Kuva 5.3: Normaalijakauman luottamusväli. Väliestimointia tarkastellaan tarkemmin seuraavassa luvussa.

- Luotettavuudella voidaan tarkoittaa myös tutkimuksen **objektiivisuutta / puolueettomuutta**
  - **Objektiivinen totuus**, tutkimustulokset ovat samat riippumatta siitä kuka pätevä tutkija tutkimuksen on tehnyt.
  - Tulosten tulisi olla luotettavia, mutta luotettavatkin havainnot voivat olla puolueellisia siinä mielessä, että ne tarkastelevat asiaa vain yhdeltä näkökannalta!
  - Esim. tarkastellaan yrityksen henkilöstökykyisyyksiä, työn organisointia ja työmoraalia, ongelmien tarkastelu johdon vs. henkilöstön näkökulmasta.

### 5.3. TILASTOLLISTEN MUUTTUJIEN MITTAAMINEN JA MITTA-ASTEIKOT73

#### Esimerkki: C-vitamiinin vaikutus syövän hoidossa

- Annettiin C-vitamiinia 100 terminaalivaiheen syöpäpotilaalle ja seurattiin kuolleisuutta (Cameron and Pauling, 1976).
  - Pyrittiin luomaan tärkeiden ominaisuuksien suhteen samanlaisia verrokkiryhmiä ja valittiin kutakin potilasta kohden 10 verrokkia, jotka olivat samanlaisia iän, sukupuolen, primääri-kasvaimen sijaintipaikan ja histologisen kasvaintyyppin suhteesta.
  - Seuranta-aika: aika hetkestä, jolloin todettiin tavanomaisten hoitojen olevan tehottomia, kuolinhetkeen saakka.
  - Tulos: C-vitamiinia saaneet käsittelyryhmän potilaat elivät 4 kertaa kauemmin ( $p < 0.0001$ ).
- Ristiriitaista evidenssiä saatiin tutkimuksessa, jossa vastaava tutkimusongelma, mutta toteutettu satunnaistettuna kokeena (Moertel et al. 1985).
  - Satunnaistettiin potilaat, joilla pitkälle edennyt paksunsuolen tai peräsuolen syöpää, C-vitamiinia saavien ja lumelääketä saavien ryhmiin.
  - Tulos: kontrolliryhmän potilaat elivät keskimäärin hieman pidempään, mutta ero ei merkitsevä.
- Mistä kahden tutkimuksen erot johtuivat? Huonolla tuurilla kallistetut verrokkit erosivat käsittelyryhmän potilaista joillakin merkittävillä tavoilla, joita ei oltu mitattu! Miten kvantifioida “huonoa tuuria”?
- Tilastolliset menetelmät tekevät juuri tämän: “Mikä on todennäköisyys, että havaittu tulos (tai sitä enemmän nollahypoteesista poikkeava tulos) olisi syntynyt vain sattumalta?”
  - Ilman satunnaistamista tuota kenties merkittävää ei-mitattua eroa ei pystytä varmuudella kontrolloimaan.
  - Todellisuudessa ero johtui siitä, että ensin mainitun tutkimuksen kontrollit valittiin jo kuolleista syöpäpotilaista, eikä heihin liittyneet enää mitään satunnaisuutta!

#### Mitta-asteikot

- Kuten satunnaismuuttuja koskeneessa luvussa 4 opittiin, satunnaismiöillä on erilaisia tulosvaihtoehtoja jotka kantavat satunnaismuuttujien to-

## 74LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

dennäköisyysjakaumia.

- On syytä huomauttaa, että vaikka mitattava ilmiö ei olisikaan numeroinen, se voidaan aina ”koodata” eli muuntaa numeeriseksi. Esimerkiksi perinteinen kaksiarvoinen mies-nainen -muuttujan tapauksessa voidaan käyttää tunnuksia 0 ja 1.
- Ilmiön luonteesta riippuen voidaan näille tulosvaihtoehdolle käyttää erilaisia **mitta-asteikkoja**.
  - **Laatueroasteikko/luokitteluasteikko** (nominaaliasteikko): Muuttujan mittaustaso on tällöin sellainen, että sen arvot voidaan luokittaa toisistaan eroaviin luokkiin. Ts. mihiin luokkaan kohde kuuluu mitattavan ominaisuuden perusteella?
    - \* Tilastoyksiköt luokitellaan ennaltamääriteltyihin luokkiin. Luokkien järjestyksellä ei ole merkitystä.
    - \* Kukin tilastoyksikkö kuuluu vain yhteen luokkaan. Tällöin kahdesta tilastoyksiköstä/havainnosta voidaan päättää vain kuuluvatko ne saamaan luokkaan vai eivät.
    - \* Emme pysty määrittelemään empiirisesti mielekästä järjestystä havaintoarvojen välillä.
    - \* Esimerkkejä: Sukupuoli, veriryhmä tai kotikunta.
  - **Järjestysasteikko** (ordinaaliasteikko): Tällöin muuttujan arvot voidaan luokittelun lisäksi asettaa empiirisesti mielekkääseen järjestykseen. Tällöin siis mittauksen kohteella on ”enemmän mitattavaa ominaisuutta” kuin jollakin toisella kohteella
    - \* Tilastoyksiköt luokitellaan ennalta määrätyihin luokkiin, joilla on yksikäsitteinen järjestys.
    - \* Esimerkkejä: Sotilasarvo, sosiaaliryhmä, kilpailun tulos tai sairauksien tarttuvuus.
  - **Välimatka-asteikko** (intervalliasteikko): Luokittamisen ja järjestysken asettamisen lisäksi havaintoarvojen välimatkalla on empiirisesti mielekäs tulkinta. Ts. intervalliasteikon tasaisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi).
    - \* Välimatka-asteikolla pystytään mittaamaan yksittäisten luokkien tai havaintoarvojen ero. Esimerkiksi: Lämpötilan mittaaminen esim. celcius-asteina. Pystymme numeroarvoina ilmoittamaan onko tänään lämpimämpi, yhtä lämmin vai kylmempä sää kuin eilen ja kuinka monta astetta muutos on.
    - \* Kuinka paljon kahden mittauksen koteen ominaisuudet eroavat toisistaan.
    - \* Intervalliasteikon tasaisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi). Mitäkin nollapiste on kuitenkin ”keinotekoinen” ja siten vapaasti

valittavissa. Samoin voidaan valita käytettävä mittayksikkö vapaasti. Oleellista on vain se, että havaintojen välisellä välimatkalla on aina empiirisesti mielekäs tulkinta.

\* Yhteen- ja vähenyslasku ovat sallittuja.

– **Suhdeasteikko:** Jos intervalliasteikon ominaisuuksien lisäksi on määriteltyä yksikäsiteinen mittalukujen absoluuttinen nollapiste.

\* Esimerkiksi kuuden euron hintainen tuote on kaksi kertaa niin kallis kuin kolmen euron tuote.

\* Kunnan veroäyri tai henkilön pituus: Absoluuttinen nollapiste on 0.

\* Nollapisteen ollessa absoluuttinen, se “pysyy paikallaan” ja mittalukujen suhteet pysyvät samoina.

- Mitta-asteikot voidaan jakaa kahteen luokkaan: **Luokittelu- ja järjestysasteikkoja kutsutaan kvalitatiivisiksi asteikoiksi.** Tällöin muuttujien arvot kuvaavat vain tilastoyksiköiden laadullisia piirteitä.
- Vastavasti **välimatka- ja suhdeasteikkoja kutsutaan kvantitatiivisiksi asteikoiksi**, koska tällöin mittaluvut kuvaavat jonkin ominaisuuden määräää.
- Tilastollisen analyysin kannalta mitta-asteikkojen merkitys on siinä, ettei tilastollisten (matemaattisten) operaatioiden sallittavuus määräytyy muuttujan mitta-asteikon mukaan. Mitä korkeampi mitta-asteikko, sitä enemmän on käytettäväissä olevia analyysimenetelmiä. Esimerkiksi keskiarvon laskeminen on eräs tilastollinen operaatio, ja se ei ole sallittu kvantitatiivisille muuttujille.
- **Aineistotyyppejä:** Käsitellään tarkemmin vielä myöhemmin (Luvussa 10), joiden yhteydessä mitattavat muuttujat voivat olla kvalitatiivisia tai kvantitatiivisia.
  - Poikkileikkausaineisto: Tietoja yhdeltä ajanhetteltä tai aikaväliltä
  - Aikasarja-aineisto: Tietoja samasta tutkimuskohteesta eri ajanhettilä
  - Paneeliaineisto: Tietoja useilta ajanhettiltä useista tutkimuskohteista
  - Tapahtumahistoria-aineisto: Tietoja tapahtumahetkiltä

## 5.4 Kontrolloidut kokeet ja suorat havainnot

- Tilastollinen tutkimusaineisto voidaan kerätä:
  - **Kontrolloiduilla kokeilla**, joissa tutkimuksen kohteet altistetaan suunnitelmallisesti erilaisiin koeolosuhteisiin selvittääkseen miten kohteet reagoivat muutoksiin.

## 76LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- **Suoria havaintoja** tehtäessä koeolosuheteita ei pyritä aktiivisesti muuttamaan vaan ainoastaan seurataan miten erilaiset olosuhteet ja niissä tapahtuvat muutokset vaikuttavat kohteisiin.
- Näistä tutkimusasetelmista kontrolloidut kokeet ovat tieteenkin ihanteellisempia tutkimuksen tekemiselle, sillä tutkijan on mahdollista tarkastella tutkittavaa asiaa koeolosuhteissa “eristyksissä”.
- Kontrolloidut kokeet eivät kuitenkaan ole aina mahdollisia, jolloin on käytettävä suoria havaintoja.
  - Tällöin tutkimuskohdetta ei suunnitelmallisesti altisteta koeolosuhteille (“käsittelyille”) vaan muuttuvien olosuhteiden vaikutuksia tilastoyksikköihin seurataan passiivisesti.
  - Toisin sanoen tutkimuksen kohteena olevat tilastoyksiksöt eivät välttämättä edes tiedä osallistuvansa tutkimukseen.
- Lisäksi usein tehdään hoito/käsittelyvastetta koskevia vertailuja erilaisissa olosuhteissa, joka osaltaan vaikuttaa tulosten uskottavuuteen, sillä tutkitavien tilastoyksikköihin voi vaikuttaa olosuhteiden muutosten lisäksi muut ulkopuoliset tekijät.
  - Näiden **selittävien ja sekoittavien tekijöiden** vaikutusten kontrollointi on suoria havaintoja tehtäessä vaativa tehtävä.
  - Mikäli ulkopuolisia tekijöitä ei havaita ja/tai pystytä mittamaan, tai muuten jostain syystä olla lisätty ja käytetty käytettävässä tilastollisessa mallissa, voi kyseeseen tulla ns. **puuttuvien selittäjien harha**, joka tarkoittaa sitä että havaittuihin tuloksiin vaikuttaa joakin havaitsematon tekijä, mutta jonka vaikutusta ei kyötä kvantifioimaan puutteellisten havaintoarvojen vuoksi.
- Suoria havaintoja tehtäessä ei voida (usein) selvittää vasteen ja olosuhteiden **kausaalista** yhteyttä. Suorilla havainnoilla voidaan lähinnä saada selville onko vasteella ja olosuilla jokin yhteys (korrelaatio) (ks. luku 7).
- Suorien havaintojen keräämiseen liittyy olennaisesti joitain riskejä ja toisaalta rajoituksia. Riskit liittyvät käytännössä otoksen harhaisuuteen (erit. valikoitumisharha)
  - Esimerkiksi jos havaintoja tehtäessä suositaan systemaattisesti joitakin tulosvaihoehtoja. Tämä suosiminen voi olla tahallista tai tahaton.
  - Tämä tilastoyksiköiden **valikoituminen** otokseen aiheuttaa harhaa, sillä otokseen valikoituvia osajoukko saattaa ylikorostaa perusjoukon joitain ominaisuuksia.

### Valikoituminen

Valikoitumista tapahtuu, jos otokseen poiminta ei ole riippumatonta tilastoyksikön ominaisuuksista. Tätä kutsutaan valikoitumisharaksi.

- Esimerkiksi verrattaessa sydän- ja verisuonitautipotilaiden hoito-toimenpiteitä potilaat eivät mahdollisesti ole valikoituneet yhtä todennäköisesti pallolaajennukseen, ohitusleikkaukseen tai lääkehoitoryhmään, sillä taudin vakavuus saattaa jo määritellä mikä hoi-totoimenpide valitaan.
- Valikoituminen on iso ongelma seurantatutkimuksissa, sillä har-haisten havaintotulosten, eli harhaisen otoksen, perusteella ei voi-da tehdä luotettavia johtopäätöksiä perusjoukosta!

- Harhan syntymistä pyritään välttämään valitsemalla havaintojen kohteet perusjoukosta satunnaisesti (ellei tavoitteena ole tutkia kaikkia perusjoukon alkioita). Tämä merkitsee satunnaisotannan soveltamista havaintojen kohteiden valintaan, eli otokseen poimittavien tilastoyksiköiden valintaan sovelletaan **satunnaistamista**, jolloin sattuma määräää mitkä perusjoukon alkioista tulevat poimituksi otokseen (tutkimuksen kohteiksi)!

### Satunnaistaminen

Tilastoyksiköiden poimimista populaatiosta otokseen riippumatta mui- den yksiköiden poiminnasta tai kyseisten (poimittavien) yksiköiden omi-naisuuksista.

- Satunnaistaminen takaa sen, että mahdolliset sekoittavat tekijät ovat jakaantuneet tasaisesti tutkittavassa joukossa. Tällöin sekoittavat tekijät eivät aiheuta harhaa otokseen ja tutkimuksen tulokset voidaan yleistää koko populaatioon.
- Satunnaistaminen poistaa otannasta valikoitumisharhan, sillä otokseen poiminta suoritetaan riippumatta tilastoyksiköiden omi-naisuuksista. Satunnaistaminen on ainoa puolueeton tapa poimia otos (ei suosi mitään perusjoukon osaa)!

- Satunnaistaminen (osaltaan) mahdollistaa **tilastollisen päättelyn**, jonka avulla otoksesta saatuja tietoja voidaan hyödyntää tehtäessä päätelmiä koko perusjoukosta.
  - Tilastollisen päättelyn avulla voidaan muodostaa esimerkiksi jakau-mien ja tilastollisten mallien tuntemattomille parametreille arviot

## 78LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

(piste-estimaatit) ja arvioida niiden epävarmuutta (keskivirheet ja luottamusvälit) sekä testata tarkasteltavaan ilmiöön liittyviä hypotheseja (ks. luku 6 ).

- Johtopäätelmien pätevyys riippuu mm. siitä, kuinka hyvin otanta on suoritettu. Tämän vuoksi on tärkeää ymmärtää otannan perusperiaatteet ja erilaisten otantamenetelmien luonne.
- Kontrolloiduissa kokeissa satunnaistaminen jakaa yksilöt **riippumatta yksilön omista ilmiöön vaikuttavista muuttujista joko käsittely- tai kontrolliryhmään** (eng. treatment ja control).
  - Se takaa, ettei valikoitumista jonkin käsittelyä edeltävän ominaisuuden mukaan esiinny
  - Tämä tarkoittaa **altisteen** (käsittely / “treatment”) antamista (täysin) satunnaisesti kokeeseen valituille yksilöille, riippumatta näiden taustamuuttujien arvoista.
  - Nämä yksilöt sinänsä voivat olla satunnaisosot jostain populaatiosta ( tai ainakin niiden toivotaan olevan), mutta satunnaistaminen tarjoittaa siis käsittelyn kohdentamista koeyksilöille, ei satunnaisotantaa sinänsä
  - Esimerkiksi tutkittavat voidaan satunnaistaa lääkehoito- ja placebo-ryhmiin, jotta mahdolliset erot tutkittavien iässä, sukupuolessa ja muissa taustamuuttujissa eivät aiheuta systemaattista harhaa, kun tutkitaan lääkehoidon vaikutusta.

### 5.5 Otantamenetelmät

- Tässä jaksossa tarkastellaan erilaisia **otantamenetelmiä**. Näiden menetelmien tarkoitus on suorittaa otosaineiston (tutkimusaineiston) kerääminen niin, että se huomioi aiemmin esitellyt hyvän otannan kriteerit, ts. että sen tuottama otos on edustava ja luotettava. Näin ollen otos kuvailee koko perusjoukkoa.
  - Otantamenetelmän, joskus myös **otanta-asetelman**, valinta on tietenkin vahvasti sovellusalakohtainen: käytettäväät aineistot ja täten otantamenetelmät määrätytyvät pitkälti tehtävän tutkimuksen luonteen perusteella. Ts. käytännön tilanteet poikkeavat toisistaan lopulta varsinkin paljon ja eri tilanteisiin tarvitaan omat menetelmänsä.
  - Otanta-asetelmalla tarkoitetaan erityisesti otoksen poimintaan käytettyä **satunnaistuksen menetelmää**.
- Otannan tavoitteena on tietenkin edustava otos. Otoksen edustavuuteen vaikuttaa käytännön otannassa se, miten todennäköistä kullakin perusjoukon alkiolla (populaation tilastoyksiköllä) on tulla poimituksi otokseen. Tätä kutsutaan **sisältymistodennäköisyydeksi**.

### Sisältymistodennäköisyys

Sisältymistodennäköisyys kuvailee sitä (tunnettua) todennäköisyyttä, jolla perusjoukon alkio tulee poimituksi otokseen.

- Käytännössä otoksen poiminta suoritetaan niin, että  $n$ :n alkion otos ( $n$  on otoskoko) poimitaan jollakin satunnaisotannan menetelmällä  $N$ :n alkion perusjoukosta ( $N$  on siis perusjoukon koko).
- Perusjoukon yksittäinen alkio (tilastoyksikkö)  $k$  tulee poimituksi  $n$ :n alkion otokseen (tutkimusaineistoon) tunnetulla **sisältymistodennäköisyydellä**  $\pi_k$ ,

$$0 < \pi_k \leq 1, \quad k = 1, \dots, N,$$

jossa siis  $N$  on perusjoukon alkioiden lukumäärä. Toisin sanoen, kaikilla perusjoukon alkioilla on oma nollaa suurempi todennäköisyytensä (voi olla 1),  $\pi_k$ , tulla poimituksi otokseen.

- Sisältymistodennäköisyys voi olla sama kaikille perusjoukon alkioille tai vaihdella perusjoukon eri osajoukkojen (alkioryhmien) välillä. Tämä tulee huomioida otantamenetelmän valinnassa, jotta saadun otoksen edustavuus ei vaarannu.
- Sisältymistodennäköisyyttä voidaan käyttää monimutkaisemmassa otantateoriassa **asetelma-** ja **analyysipainojojen** muodostamisessa sekä uudelleenpainotuksessa (vastauskadon korjaus).
- Tässä luvussa käsitellään erilaisia perinteisiä otantamenetelmiä sekä siitä, minkälaisten perusjoukkojen tilanteissa mikäkin otantamenetelmä on sopivin.
  - **Yksinkertainen satunnaisotanta** (YSO): perinteisin otantamenetelmä, jossa jokaisella tietyn kokoisella otoksella sama mahdollisuus tulla valituksi.
  - **Systemaattinen otanta** (SYS):
  - **Osittelu otanta**: perusjoukko (populaatio) jaetaan ominaisuuksiltaan yhtenäisiin eli homogenisiin **ositteisiin**, joista jokaisesta poimitaan erillinen otos.
  - **Ryvästonta** tai joskus myös **moniasteinen otanta**: Hyödynnetään perusjoukossa esiintyvää kerroksellisuutta, eli hierarkkisuutta otannassa.

#### 5.5.1 Yksinkertainen satunnaisotanta

- **Yksinkertaisessa satunnaisotannassa** (YSO) jokaisella tilastoyksikölle (perusjoukon alkiolla) on nollasta poikkeava todennäköisyys tulla valituksi otokseen.

## 80LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Otanna satunnaisuus tulee siis siitä, että jokainen tilastoyksikkö poimitaan otokseen *satunnaisesti!* (Ks. luku 4)
- YSOa pidetään otannan perusmuotonäkymänä, jossa jokaisella perusjoukon alkioilla on lähtökohtaisesti yhtä suuri todennäköisyys tulla valituksi otokseen.
  - \* Yksinkertainen satunnaisotanta on periaatteeltaan intuitiivinen ja helppo ymmärtää. Lisäksi se on tietyissä tilanteissa usein helppo toteuttaa.
- Tällöin on selvää että myös jokaisella perusjoukon samankokoisella osajoukolla on sama todennäköisyys tulla valituksi.
- Toisin sanoen, todennäköisyys tulla poimituksi ei riipu tilastoyksikön ominaisuuksista tai siitä minkälaisia ominaisuuksia jo poimituilla otosyksiköillä on.
- Satunnaisotanta siis selvästi korjaa valikoitumisharhaa (viittaus aiempaan lukuun) satunnaistamalla otokseen valikoitumisen täysin! YSO voidaan aina tulkita arvonnaksi. Käytännön työssä arvonta onkin oiva satunnaistamisen keino.

### • YSO:n toteuttaminen

- Käytännössä yksinkertainen satunnaisotanta etenee vaiheittain:
  - \* Tutkimuksen alussa tutkijalla tulisi olla käytettäväänään (ts. tulisi koostaa) lista kaikista perusjoukon havaintoyksiköistä (alkioista). Tämä muodostaa tutkimuksen **otantakehikon**.
  - \* Tämän jälkeen jokaiseen perusjoukon alkioon voidaan liittää numeriset tunnukset.
  - \* Sitten valitaan haluttu otoksen koko. Otoskoon määrittäminen on keskeinen osa koesuunnittelua, ks. luku 6.6
  - \* Otantakehikosta arvotaan perusjoukon alkiot otokseen yksi kerrallaan.
  - \* Käytännössä arvonta voidaan toteuttaa satunnaislukuja generoimalla (tuottamalla) niin että jokaisen otantakehikon alkion sisältymistodennäköisyys on yhtä suuri.<sup>1</sup>

### • YSO:n **poimintastrategiat**: Käytännössä yksinkertainen satunnaisotanta voidaan suorittaa kahdella eri tavalla: **palauttaen** tai **palauttamatta**.

- Tarkastellaan, aiemman mukaisesti, **äärellistä populaatiota** (perusjoukkoa), jossa on  $N$  alkioita ja tarkoituksesta on poimia  $n$ :n alkion kokoinen otos (huom.  $n < N$ ). Olkoon  $i$  yksittäisen alkion indeksiluku (ts. jokainen alkio on numeroitu esimerkiksi tavalla  $i = 1, \dots, N$ ).

### YSO:n poiminta palauttaen

---

<sup>1</sup>Satunnaislukujen generointia käsitellään ja opetellaan mm. R-kurssilla ja kurssilla TILM3705 Johdatus laskennalliseen tilastotieteen.

- Kun poiminta suoritetaan **palauttaen**, niin poimittu alkio palautetaan aina ennen uuden alkion arpomista takaisin perusjoukkoon, jolloin alkio voi tulla poimituksi otokseen useita kertoja.
  - Kyseessä on siis otanta **takaisinpanolla** (with replacement).
  - Tällöin alkioiden arvonnat ovat riippumattomia: alkion todennäköisyys tulla poimituksi otokseen ei riipuu siitä kuinka monta alkiota otokseen on jo poimittu.
  - Alkion  $i$  sisältymistodennäköisyys on tällöin selvästi

$$\pi_i = \frac{1}{N}, \quad \forall i$$

- Otantaan palauttaen liittyviä todennäköisyyksiä hallitaan **binomijakau-man** avulla (ks. luku 4), joka johtaa yksinkertaiseen **tilastolliseen malliin** YSO:a käytettäessä.
- Poiminta palauttaen, tai otanta takaisinpanolla, on toisaalta varsin epärealistinen otantamenetelmä useassa tutkimuksessa. Esimerkiksi lienee mahdotonta testata samaa lääkettä useaan otteeseen samaan aikaan yhdellä koehenkilöllä.

#### YSO:n poiminta palauttamatta

- Kun poiminta suoritetaan **palauttamatta**, poimittua alkiota ei palauteta perusjoukkoon poiminnan jälkeen eikä se täten voi tulla poimituksi otokseen kuin kerran.
  - Kyseessä on siis otanta **ilman takaisinpanoa** (without replacement).
  - Tällöin alkioiden arvonnat eivät enää ole riippumattomia: alkion todennäköisyys tulla poimituksi otokseen riippuu siitä kuinka monta alkiota otokseen on jo poimittu.
  - Alkion  $i$  sisältymistodennäköisyys on tällöin vastaavasti

$$\pi_i = \frac{1}{N - A_i},$$

- Tässä  $A_i$  on jo poimittujen alkioiden lukumäärä ennen kyseistä **otosite-raatiota**: ensimmäisen poiminnan kohdalla  $A_i = 0$ , toisen kohdalla  $A_i = 1$  ja niin edespäin.
  - Ilman takaisinpanoa populaatiosta voidaan poimia  $\binom{N}{n}$  erilaista otosta.<sup>2</sup>

---

<sup>2</sup>Kun otosyksiköiden järjestyksellä ei ole merkitystä.  $\binom{N}{n}$  on ns. binomikerroin, joka saadaan kaavasta  $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ , jossa  $N! = N \cdot (N-1) \cdot (N-2) \cdots 1$  on  $N$ :n kertoma.

## 82LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Otantaan palauttamatta liittyviä todennäköisyyksiä hallitaan **hypergeometrisen jakauman** avulla, joka johtaa (melko) yksinkertaiseen **tilastolliseen malliin** YSO:a käytettäessä.

### Esimerkki: Yksinkertaisen satunnaisotannan poimintastrategiat

- Esimerkki: Poimitaan palloja kulhosta satunnaisesti.
  - Jos yksittäinen pallo (alkio) voi tulla poimituksi useammin kuin kerran, eli pallo palautetaan kulhoon sen poiminnan jälkeen, on kyseessä yksinkertainen satunnaisotanta takaisinpanolla.
  - Vastaavasti jos pallo voi tulla valituksi vain kerran, eli pallo poistetaan kulhosta sen poiminnan jälkeen, on kyseessä otanta ilman takaisinpanoa.

### Otoskoon vaikutus YSO:n

- Yksinkertaisen satunnaisotannan erot takaisinpanolla ja ilman takaisinpanoa riippuvat otantakehikon (tai yleisemmin perusjoukon) koosta. Mikäli poimittava otos muodostaa suuren osan perusjoukosta (ts.  $\frac{n}{N}$  on “suuri”, eli lähellä yhtä) menetelmät poikkeavat olennaisesti.
- Toisaalta, jos perusjoukko on ääretön niin menetelmillä ei ole käytännössä eroa (ts. kun  $N \rightarrow \infty$  niin  $\frac{n}{N} \rightarrow 0$  eli todennäköisyys että sama alkio poimittaisiin otokseen useammin kuin kerran lähestyy nollaa otoskoon lähestyessä ääretöntä).
  - Monesti onkin (teoreettiselta) kannalta järkevää olettaa että otos poimitaan äärettömästä perusjoukosta vaikka perusjoukko tosiasiallisesti olisikin äärellinen (mutta riittävän “iso”).
  - Tällöin voidaan olettaa käytettävä otantaa takaisinpanolla, sillä siinä käytettävät tilastolliset mallit ovat yksinkertaisempia kuin otanassa ilman takaisinpanoa ja tämä helpottaa tilastollisessa päätelyssä käytettäviä kaavoja.

### YSO: Potentiaaliset ongelmat

- Monissa tapauksissa ei kuitenkaan ole helppoa saada lista kaikista perusjoukon havaintoyksiköistä (jolloin menetelmän käyttö on mahdotonta).
- Kyselytutkimuksissa perusjoukko on usein suuri ja laajalle alueelle hajaantunut. Henkilökohtaisten, kasvotusten toteutettavien, haastattelujen tekeminen vaatii suuria resursseja (haastattelijat joutuisivat esim. matkustamaan ympäri Suomea satunnaisotokseen valikoituneiden henkilöiden asuinpaikkojen mukaan).

- Tällaisissa tutkimustilanteissa käytetäänkin usein muunlaisia otantamenetelmiä.

### 5.5.2 Systemaattinen otanta

- Systemaattisessa, eli tasavälisessä, otannassa poimintakehikkoon (perusjoukkoon) kuuluvat alkioit järjestetään jonoon ja siitä poimitaan otokseen joka *k.* alkio.
  - Esimerkiksi jos oletetaan että perusjoukkoon kuuluu 1000 tilastoyksikköä ja valittu otoskoko on 100, niin otos voidaan poimia perusjoukon alkioiden järjestetyistä listasta poimimalla siitä joka kymmenes yksikkö.
  - Systemaattinen otanta ei oikeastaan kuulu satunnaisotannaksi laskettaviin menetelmiin, koska siinä ei sovelleta arvontaa.
  - Yksinkertainen satunnaisotanta voidaan kuitenkin nähdä systemaattisen otannan erikoistapauksena (eli systemaattinen otanta voidaan toteuttaa satunnaisotantana), missä perusjoukon alkioit järjestetään jonoon **satunnaistamalla**. - Ts. jonon järjestys on satunnainen, eli joka *k.* jonon alkio on “satunnaisotos” otantakehikosta.
  - Systemaattinen otanta tuottaa tällöin samat johtopäätelmät kuin yksinkertainen satunnaisotanta, jos perusjoukon alkioiden järjestys on tutkittavan ilmiön kannalta satunnainen! Toisin sanoen, harhaa ei synny mikäli perusjoukon alkioiden järjestys ei riipu sellaisesta ominaisuudesta, jota tutkitaan.
  - Systemaattisen otannan suhteen potentiaaliseksi ongelmaksi muotoutuu havaintoyksikkölistan mahdollinen säänöllinen jaksollisuus, jota se ei havaite ja jolloin satunnaisotanta toimisi (kenties) paremmin.
    - \* Ongelmaa syntyy esimerkiksi silloin, jos tiedot perusjoukosta koostuvat pariskunnista ja poimintaintervalli on parillinen luku. Tällöin seurausena voi olla, että otokseen saattaisi valikoitua ainoastaan joko miehiä tai naisia.
- Myös systemaattisessa otannassa tarvitaan siis lista tai rekisteri kaikista perusjoukon havaintoyksiköistä ja sitä sovellettaankin tavallisesti YSO:n sijasta silloin, kun perusjoukon alkioista on käytettäväissä tietorekisteri, luettelo tai havaintoja kerätään ajassa tai tilassa.
  - Esimerkiksi mielipidekyselyn kohteet poimitaan (voitiin poimia) puheelinluettelosta (tai vastaavasta rekisteristä) valitsemalla haastateltavaksi jokaiselta aukeamalta ensimmäisenä esiintyvä henkilö tai joitain tuotetta valmistavan tehtaan laaduvalvonnassa valitsemalla laatuvarvointiin joka sadas tuote, joka hihnalta valmistuu. Muita esimerkkejä ovat esim. liikenteen, jäsenrekisteriin tai kassajonossa seisovien otantayksiköiden poiminta otokseen.

### 5.5.3 Ositettu otanta

- Ositettu otanta on sopiva menetelmä tilanteisiin, joissa perusjoukko koostuu jonkin ominaisuuden suhteen homogeenisista ryhmistä, ts. alkioryhmistä (osista). Ositettu otanta pyrkii varmistamaan, että tutkittava otos on edustava kaikkien (tutkimuksen kannalta) olennaisten ryhmien osalta.
  - Esimerkiksi jos tavoitteena on tutkia jonkin maan erilaisten ja usein hyvin eri kokoisten kieliryhmien taloudellista asemaa. Kaikista ryhmistä tulisi saada edustava otos.
  - Tällöin maan koko populaatioon kohdistettu yksinkertainen satunnaisotanta ei olisi järkevä, sillä otoskoon pitäisi olla (todennäköisesti) hyvin suuri, että jokaisesta kieliryhmästä saataisiin poimittua edustava otos.
  - Ositetun otannan avulla otos voitaisiin kerätä niin, että jokaisesta ryhmästä (ositteesta) poimitaan osaotos yksinkertaisella satunnaisotannalla tai systemaattisella otannalla ja nämä osaotokset yhdistetään yhdeksi otokseksi.
- Ositettu otanta voi (oikein toteutettuna ja sopivassa asetelmassa) tuottaa paljon tarkempaa tietoa kuin yksinkertainen satunnaisotanta samaa otoskokoa käytettäessä! Voidaan esimerkiksi käyttää tietoa siitä, että otosyksiköt ovat joka ositteessa keskenään samankaltaisia.
- Ositetun otannan käyttöön suurissa kyselytutkimuksissa liittyy samoja ongelmia kuin yksinkertaiseen ja systemaattiseen satunnaisotantaan.
  - Otokseen valikoituneet vastaajat voivat olla mm. levittäyneinä suulle maantieteelliselle alueelle. Näin ollen otannan suorittaminen vaatii suuria kustannuksia.
  - Onko (järkevä) osittaminen ylipäättäään mahdollista toteuttaa tarkasteltavassa sovelluskohteessa?

### 5.5.4 Ryvästotanta

- Ryvästotanta soveltuu tilanteisiin, joissa perusjoukko on “ryvästeistä” eli se voidaan jakaa luonnollisiin ryhmiin eli rypäisiin (eng. *clusters*).
- Rypäkset indikoivat aineiston luontaisista hierarkkista, eli monitasoista- tai asteista rakennetta.
  - Esimerkkejä tällaisista ryhmistä ovat erilaiset yritykset tai koululuokat. Esimerkiksi yritykset muodostavat luonnollisesti eri ryppääitä, joiden alkiot ovat työntekijöitä ja koululuokat muodostavat koulun sisällä omia luonnollisia ryppääitä ja opiskelijat ovat alkioita näissä ryppäissä.

- Huomionarvoista onkin, että toisin kuin ositetussa otannassa, ryväätannassa ryppäiden oletetaan olevan toistensa kanssa riittävän samankaltaisia, että jokaista rypäästä ei tarvitse erikseen tutkia.
  - Tämä onkin yksi ryväätannan tärkeimpää motivoointeja, sillä sitä usein perustellaan kustannustehokkuudella: sen sijaan että poimitaan satunnaisia koululaisia mahdollisesti suuresta määristä kouluja, voidaan poimia satunnaisia ryppääitä (kouluja), joista tutkimusyksiköt eli koululaiset poimitaan.
  - Lisäksi koulun sisällä koululuokat muodostavat aliryppääitä, joista voidaan edelleen poimia satunnaisotos, jotta päästään tutkimaan perusjoukon alkioita eli koululaisia esim. haastattelututkimuksen muodossa.
  - Tavoitteena on vähentää tietojen keruun aiheuttamia kustannuksia samalla varmistaen, että otos on kuitenkin mahdollisimman edustava!
- Ryväätannan voi suorittaa **yksi-** tai **kaksivaiheisena** (**yksiasteinen/-kaksiasteinen ryväätanta** ).
  - **Kaksivaiheisessa ryväätannassa**
    - \* **Ensimmäisessä vaiheessa** poimitaan joukko ryppääitä kaikien ryppäiden joukosta, eli vain osa ryppäästä on mukana lopullisessa otoksessa.
    - \* **Toisessa vaiheessa** poimitaan ensimmäisessä vaiheessa poimitusta ryppäästä alkiotason otokset.
  - **Yksivaiheisessa ryväätannassa** toisessa vaiheessa valitaan kaikki ensimmäisen vaiheen otosryppäiden alkiot, jolloin toisen vaiheen otanta typistyy ensimmäisen vaiheen ryppäiden alkioiden kokonaistutkimukseksi.
  - Poiminnan eri vaiheissa voidaan soveltaa yksinkertaista satunnaisottantaa tai systemaattista otantaa.
- Ryväätantaa käytetään usein suuria haastattelututkimuksia tehtäessä. Erityisesti, ryväätantaa voidaan hyödyntää myös silloin, kun tutkijalla ei ole käytettävissään kattavaa listaa kaikista havaintoyksiköistä, mutta näiden muodostamat ryppäät on määritettävissä.
- Ryväätannan heikkoutena pidetään sitä, ettei aina ole helppoa muodostaa ryppäät, jotka ovat toistensa kaltaisia. Tulosten tarkkuus myös riippuu moninpaikoin siitä, kuinka hyvin ryppäisiin jako onnistuu.

#### Esimerkkejä ryväätannasta

- Esimerkki 1:

## 86LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Poimitaan oppilaitoksen opiskelijoista otos arpomalla ensin otos luokkahuoneista (=rypäistä).
- Arvotuissa luokkahuoneissa käydään sitten suorittamassa kysely.
  - \* Esim. Oppilaitoksen opiskelijoista voidaan poimia otos arpomalla ensin otos luokkahuoneista, jolloin luokkahuoneet ovat nk. ryppääitä.
  - \* Mahdollisia ongelmia? Miten huomoida päivä- ja iltaopiskelijat? Tämän voisi toteuttaa arpomalla otos luokkahuoneista päiväsaikaan ja toinen otos ilt-aikaan. Tässä yhdistetään ryvästontaan ositettu otanta, jolla taataan päivä- ja iltaopiskelijoiden edustus.
- Esimerkki 2: Tutkittaessa tänä vuonna peruskoulun aloittavia voidaan ensin poimia otos kouluista, jolloin koulut ovat ryppääitä. Tämän jälkeen arvotaan kustakin otokseen tulleesta koulusta tietty määrä tutkimuksen kohderyhmään kuuluvia oppilaita.

### 5.6 Otantaesimerkkejä

#### Esimerkki: Työllisyys ja työttömyys, Tilastokeskuksen työvoimatutkimus

- Työvoimatutkimus on otostutkimus, jonka avulla tilastoidaan 15–74-vuotiaan väestön työmarkkinoille osallistumista, työllisyyttä, työttömyyttä ja työaikaa (yhden viikon aikana) kuukausittain, neljännesvuosittain ja vuosittain.
  - Työvoimatilastoja käytetään työvoimapolitiikan ennusteiden ja suunnitelmien laadinnassa, toimien seurannassa ja päätöksenteon tukena.
  - Työmarkkina-aseman perusluokittelussa väestö jaetaan työllisiin, työttömiin ja työvoiman ulkopuolisiin.
    - \* Työlliset ja työttömät muodostavat työvoiman.
  - Työvoimatutkimuksen **perusjoukon** muodostavat Suomessa vakinaisesti asuvat 15–74-vuotiaat henkilöt.
  - Työvoimatutkimuksen otos poimitaan **ositetulla satunnaisotannalla** väestön keskusrekisteriin perustuvasta Tilastokeskuksen väestötietokannasta kahdesti vuodessa.
- Ositetun satunnaisotoksen poiminta:

- Tutkimus on paneelitutkimus, jossa samaa henkilöä haastellaan viisi kertaa.
- Joka kuukauden otokseen kuuluu noin 12 000 henkilöä, keskimäärin noin joka 300. henkilö perusjoukosta.
- Yhden tutkimuskaukuden otos koostuu viidestä rotaatioryhmästä, jotka ovat tulleet tutkimukseen mukaan eri aikoina. Otos vaihtuu asteittain siten, että kolmena peräkkäisenä kuukautena vastaamisvuorossa ovat eri henkilöt.
- Julkisuudessa seurataan useimmiten kuukausittain työllisyiden ja työttömyyden muutoksia edellisen vuoden vastaavasta kuukaudesta. Vaihtoehtoisesti voidaan käyttää kausitasoitettuja lukuja, jolloin tilannetta voidaan verrata edelliseen kuukautteen.

**Esimerkki: Terveys 2000**

- Terveys 2000 -tutkimuksen tavoite oli tuottaa ajankohtainen kattava kuva työikäisen ja iäkkään väestön terveydestä ja toimintakyvystä selvittämällä tärkeimpien terveysongelmien yleisyyttä ja syitä sekä niihin liittyvän hoidon, kuntoutuksen ja avun tarvetta.
- Tutkimus koskee (koski) 18 vuotta täyttänyttä Suomen aikuisväestöä (perusjoukko), josta valitaan valtakunnallisesti edustava 10 000 henkilön otos.
- Poimittiin kaksivaiheinen ryväatos terveyskeskuspiireistä.
  - Ositus perustui yliopistosairaaloiden vastuualueiden väestömäärään suhteutettuun kiintiöintiin.
  - Suurimmat 15 terveyskeskuspiiriä poimittiin otokseen ja lopuista 65:stä piiristä poimittiin loppuotos kussakin ositteessa systemaattisella (PPS) otannalla (sisältymistodennäköisyys suhteessa alkion kokoon).

## 5.7 Otannan haasteita vielä kootusti

- Poimintaharha: Otos ei edusta populaatiota. Vaarana varsinkin silloin, kun otokseen tulleet populaation alkiot ovat valikoituneet tai ovat itse valinneet itsensä otokseen. Vastaavasti toisinaan otoksen peitto ei ole hyvä eli tällöin otanta ei kata koko perusjoukkoa tai se kattaa perusjoukon ja vähän muutakin.

## 88LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Jos poimitaan tutkimukseen ne perusjoukon alkiot, jotka ovat tutkimuksen tekemishetkellä ‘saatavilla’, niin kyseessä on **näyte**. Näyte ei siis kata ilmiön koko vaihtelua edustavan satunnaisotoksen tapaan.
  - Esimerkiksi perinteiset katukyselyt eivät ole hyvä otantatapa, sillä kadulla liikkujat eivät välittämättä kovin hyvin edusta tutkittavaa perusjoukkoa, ellei perusjoukkona ole kyseisellä kadulla kyseiseen aikaan liikkuvat ihmiset.
  - Jos television ajankohtaisohjelma pyytää katsojia twiittaamaan mielipiteensä ajankohtaisesta asiasta, kyseessä on itse valikoituvä näyte (osallistujat valitsevat itse itsensä).
- Vajaapeittävyys: Populaation alkioista ei ole välittämättä täydellistä lueteloa
- Vastauskato: Tutkimuksen kohteita ei tavoiteta tai he kieltäytyvät vastaamatta. Kadon vuoksi lopullinen otoskoko saattaa jopa karsiutua pois tai jokin osajoukko on alioidustettuna.
- Vastausharha: Kysymykset voivat olla huonosti muotoiltuja tai vastaajat voivat antaa väärää tietoja.

## Luku 6

# Otokset ja otosjakaumat: tilastollisen päätelyn näkökulma

Tarkastellaan seuravaaksi otoksia ja otosjakaumia “tilastollisemmin” mitä edellisten lukujen erityisesti otantaa koskevan johdannon yhteydessä. Tilastollinen päätely on keskeinen osa tilastotiedettä, sillä se mahdollistaa päättelmiens yleistämisen otoksesta populaatioon/perusjoukkoon. Tämä luku toimii esimerkkinä formaaliin matemaattiseen esitykseen perustuvan tilastollisen päätelyn perusteista (otannan ja otantajakaumien näkökulmasta), jonka ideana on yleisesti tehdä luotettavia johtopäätöksiä perusjoukosta otoksen perusteella. Tällä kursilla käydään läpi (vain) tarvittavia yksityiskohtia sekä rakennetaan pohjia tnlaskennan kurssin jälkeiselle tilastollisen päätelyn peruskurssille (TILM3555).

### 6.1 Satunnaisotos, yhteisjakauma ja tilastollinen malli

- Luvusta 4 muistamme, että tilastollisen tutkimuksen kohteena on satunnaisilmiöt, joita kuvataan satunnaismuuttujilla. Satunnaismuuttujilla on todennäköisyysjakaumat, joita tilastotieteessä kuvataan todennäköisyyseli tiheysfunktion avulla.
  - Merkitään satunnaismuuttuja isolla kirjaimella,  $Y$ , ja yksittäisen satunnaismuuttuja realisaatiota pienellä kirjaimella  $y$ . Otoskokoa, eli otokseen osallistuvien tilastoyksiköiden määrää merkitään  $n$ :llä ja tilastoyksiköitä indeksöidään alaindeksillä  $i = 1, \dots, n$ .

## 90LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Otoksen poimimisen jälkeen satunnaismuuttujat  $Y_1, \dots, Y_n$  saavat havaituksi arvoikseen havaintoarvot  $y_1, \dots, y_n$  (ts.  $Y_1 = y_1, \dots, Y_n = y_n$ ).
- Näin havaintoaineisto on siis **satunnaisotos**, joka voidaan määritellä tarkemmin seuraavasti.

### Satunnaisotos

Olkoot  $Y_1, \dots, Y_n$  riippumattomia ja samoinjakautuneita satunnaismuuttuja, joiden tiheysfunktiota (tf., tai pistetoden-näköisyydfunktiota (ptnf)) merkitään  $f(y, \theta)$ :llä, jossa  $y$  on yksittäisen sm:jan  $Y$  reaalisaatio ja  $\theta$  on jokin jakauman muodon määrävä parametri (tai parametrit). Parametrin  $\theta$  arvoa ei yleensä tunneta ja tavoitteena onkin päättää, **estimoida**, sen arvoa lopulta käytettävään aineistoa käyttäen.

### Satunnaisotoksen tilastollinen malli

- Havaintoarvot  $y_1, \dots, y_n$  ovat kiinteitä lukuja, mutta ne vaihtelevat satunnaisesti otoksesta toiseen. Satunnaisotannassa **satunnaisuus liittyy siis havaintoarvojen vaihteluun satunnaisesti otoksesta toiseen**.
  - Satunnaisuus ei siis liity otannan tuloksena saatuihin havaintoarvoihin, vaan otoksen poimintaan.
- Satunnaismuuttujien  $Y_1, \dots, Y_n$  **yhteisjakauma** muodostaa (tiettyjen läädeusten jälkeen) **tilastollisen mallin** havaintoarvojen satunnaiselle vaiotelulle eri otoksissa.
  - Koska tällä kurssilla satunnaismuuttujat  $Y_1, \dots, Y_n$  oletetaan **riippumattomiksi toisiinsa nähdön**, niiden yhteisjakauma on tulomuotona  $f(y_1, \dots, y_n; \theta) = f(y_1; \theta) \times \dots \times f(y_n; \theta)$ .
- Oletetaan, että  $Y_1, \dots, Y_n$  ovat aiempien oletusten pätiessä riippumattomia sm:jia ja, että ne muodostavat satunnaisotoksen jakaumasta, jonka odotusarvo on  $\mu$  ja varianssi on  $\sigma^2$ .
  - Ts. oletamme

$$E(Y_i) = \mu, \quad i = 1, \dots, n, \quad \text{Var}(Y_i) = \sigma^2, \quad i = 1, \dots, n.$$

- Tässä tapauksessa mielenkiinnon kohteena olevat parametrit ovat siis  $\mu$  ja  $\sigma^2$  eli  $\theta = (\mu \quad \sigma^2)$ .

- Tilastollisten mallien tehtäväänä on siis estimoida nämä todennäköisyysjakaumien parametrit havaitun aineiston perusteella, joten keskeinen tilastollinen kysymys on etä miten estimointi suoritetaan luotettavasti.

**Esimerkki: satunnaisotos normaalijakaumasta**

Normaalijakautuneiden satunnaismuuttujien satunnaisotokselle pätee  $Y_1, \dots, Y_n \perp\!\!\!\perp, Y_i \sim N(\mu, \sigma^2), i = 1, \dots, n$ .

- Esimerkiksi R-ohjelmassa voidaan generoida 10 havainnon ( $n = 10$ ) satunnaisotos standardoidusta normaalijakaumasta (ts.  $Y_i \sim N(0, 1), i = 1, \dots, 10$ ) komennolla rnorm(10).

**Esimerkki: miesten pituus**

- Kerätään havaintoja miesten pituuksista yksinkertaisella satunnaisotannalla (takaisinpalauttaen)  $n$  kappaletta.
- Tällöin havaintoarvoja  $Y_1, \dots, Y_n$  voidaan pitää riippumattomina satunnaismuuttujina, joista jokainen noudattaa tehdyн jakaumaoletuksen mukaan normaalijakaumaa  $N(\mu, \sigma^2)$ .
- Estimoinnin tehtäväänä on muodostaa parhaat mahdolliset arviot parametreille  $\mu$  ja  $\sigma^2$ , ja mahdollisesti testata esimerkiksi odotusarvolle  $\mu$  asetettua hypoteesia.

## 6.2 Otosjakauma: Estimaattori ja estimaatti

- Erityisesti klassisessa tilastotieteessä päättely pohjautuu aineiston tilastollisen mallin kuvamalle tilastolliselle stabiliteetille, joka ilmenee ajatuksena aineiston keruun toistamisesta.
  - Oletetaan, että tarkasteltavan aineiston on tuottanut satunnaisotanta tai satunnaiskoe, joka noudattaa tilastollista mallia  $f(y_1, \dots, y_n; \theta)$  (aiemmin merkinnöin).
  - Toistetaan aineiston keruu samoissa olosuhteissa yhä uudelleen ja uudelleen.
  - Saatava aineisto (numeeriset arvot)  $y_1, \dots, y_n$  vaihtelevat näin ollen valitun tilastollisen mallin jakauman kuvamalla tavalla.
- Satunnaisotoksesta voidaan laskea erilaisia **tunnuslukuja/otossuureita**, joita merkitään  $T$ :llä, ts. ne ovat aineiston funktioita

$$T = g(Y_1, \dots, Y_n).$$

## 92LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Tunnusluvut ovat satunnaismuuttujien funktioina myös satunnaismuuttuja.
  - Tunnusluvulla on nk. todellinen arvo,  $g(\theta)$ , joka vastaa tunnusluvun arvoa perusjoukon tasolla ja jota pyritään aineistoa käyttäen estimoimaan.
  - Esimerkkinä tunnusluvusta on keskiarvo  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ .
  - Tunnusluvun havaittu arvo (realisaatio) pisteessä  $(y_1, \dots, y_n)$  eli havaitussa aineistossa on

$$t = g(y_1, \dots, y_n).$$

- Otoksen poimimisen jälkeen, havaintoarvoja käyttäen, voidaan laskea tunnuslukujen havaitut arvot (jolloin ne ovat siis ei-satunnaisia).
- Esimerkiksi keskiarvo on havaittujen arvojen keskiarvo, kun se lasketaan kerätystä aineistosta.
- Jos tunnuslukua  $T$  käytetään tilastollisen mallin parametrin (parametreiden)  $\theta$  estimointiin, niin tästä sanotaan tällöin parametrin **estimaattoriksi**.
  - Estimaattorin otoskohtaisia arvoja, kuten yllä  $t$ , kutsutaan **estimaatteiksi**.
  - Toivottavaa olisi, että estimaatit  $t = g(y_1, \dots, y_n)$  osuisivat mahdollisimman läheille tunnusluvun todellista arvoa  $g(\theta)$ . Ts. satunnaismuuttujan eli tässä tapauksessa estimaattorin  $T = g(Y_1, \dots, Y_n)$  jakauman tulisi keskittyä mahdollisimman tiiviisti  $g(\theta)$ :n ympärille.
- Koska tunnusluku/estimaattori  $T$  on satunnaismuuttuja, sillä on todennäköisyysjakauma, jota kutsutaan tunnusluvun  $T$  **otosjakaumaksi**.
  - Ototjakauma muodostaa (tilastollisen mallin) todennäköisyysmallin tunnusluvun  $T$  arvojen satunnaismuuttujan.
  - Ototjakaumat riippuvat tuntumattomista **parametreista**, joiden arvoja ei yleensä tunnetta ja niitä pyritään estimoimaan kerättyä otosta ja sopivaa tunnuslukua käyttäen.
  - Parametri on (usein) perusjoukon tunnusluku, jota halutaan arvioida. Parametrit **estimoidaan** havaintoaineistoa käyttäen.

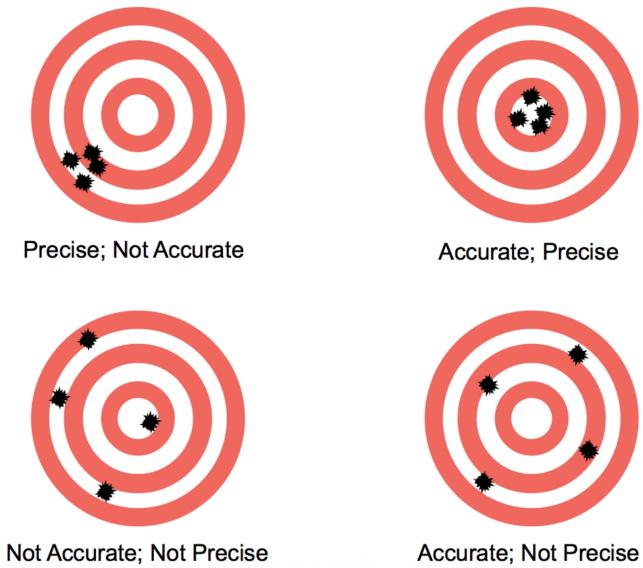
### Estimaattorin ominaisuudet

- Merkitään seuraavassa parametrin  $\theta$  estimaattoria  $\hat{\theta}$ :lla ja siltä voidaan toivoa seuraavia ominaisuuksia:

**Harhattomuus**

Estimaattorin odotettavissa oleva arvo yhtyy tuntemattoman parametrin  $\theta$  todelliseen arvoon eli  $E(\hat{\theta}) = \theta$ .

- Harhaton estimaattori tuottaa keskimäärin oikean kokoisia arvoja (estimaatteja) estimoitavalle parametrille
- Estimaattorin tuottama arvo parametrille saattaa tietylle otokseen poiketa paljonkin parametrin todellisesta arvosta, mutta odotusarvon frekvenssitulkinnan mukaan estimaattorin tuottamat otoskohdaiset arvot parametrille jakautuvat otantaa toistettaessa (symmetrisesti) parametrin todellisen arvon ympärille



Kuva 6.1: Harhaton estimaattori

**Tyhjentävyys**

Tyhjentävä estimaattori käyttää kaiken otokseen sisältyvän parametria  $\theta$  koskevan informaation.

**Tehokkuus**

Kahdesta saman parametrin  $\theta$  estimaattorista tehokkaampi on se, jonka varianssi on pienempi. Ts.  $\hat{\theta}^{(1)}$  on tehokkaampi kuin  $\hat{\theta}^{(2)}$ , jos  $\text{Var}(\hat{\theta}^{(1)}) \leq \text{Var}(\hat{\theta}^{(2)})$ .

**Tarkentuvuus**

Tarkentuvan estimaattorin  $\hat{\theta}$  arvot lähestyvät parametrin  $\theta$  oikeaa arvoa otoskoon kasvaessa.

- Voidaan osoittaa (yksityiskohdat sivuutetaan tällä kurssilla), että esimerkiksi yksinkertaisen satunnaisotoksen tapauksessa tavanomaisilla binomijaa normaalijakauman parametreiden estimaattoreilla on kaikki edellä mainitut hyvysominaisuudet.
  - Näin ei ole yleisesti monimutkaisemmissa otantatilanteissa ja tilastollisissä malleissa.
  - Estimaattoreiden kehittäminen erilaisten tilastollisten mallien tapauksessa kuuluu teoreettisen tilastotieteen alaan.

### 6.3 Otoskeskiarvo ja otosvarianssi (estimaattoreinta)

- Tarkastellaan seuraavaksi tarkemmin kahta kenties useimmiten tarkasteltua tunnuslukua ja niiden otosjakaumia:
  - Aritmeettisen keskiarvon otosjakaumaa
  - Suhteellisen osuuden (frekvenssin) otosjakaumaa

#### Otoskeskiarvo

- Oletetaan, kuten aiemmin, että  $Y_1, \dots, Y_n$  ovat riippumattomia sm:jia ja että ne muodostavat satunnaisotoksen jakaumasta jonka odotusarvo on  $\mu$ , ts.  $E(Y_i) = \mu$  ja varianssi on  $\sigma^2$ , ts.  $\text{Var}(Y_i) = \sigma^2$ .
  - Havaintojen (satunnaismuuttujien)  $Y_1, \dots, Y_n$  **otoskeskiarvo** on

$$\bar{Y} = \frac{1}{n}(Y_1 + \dots + Y_n) = \frac{1}{n} \sum_{i=1}^n Y_i$$

### 6.3. OTOSKESKiarvo ja otosvarianssi (estimaattoreinta) 95

- Yksittäisen otoksen otoskeskiarvo on tällöin sm:jien realisaatioiden aritmeettinen keskiarvo

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

- Otoskeskiarvo on satunnaismuutuja, jonka saama arvo vaihtelee satunnaisesti otoksesta toiseen johtuen satunnaisotannasta. - Kun satunnaismuuttujat ovat samoin jakautuneet odotusarvonaan  $\mu$ , on otoskeskiarvo jakauman odotusarvon harhaton estimaattori, ts.

$$E(\bar{Y}) = \mu$$

- Täten otoskeskiarvo kuvaaa aineiston perusjoukon tilastollisen mallin odotusarvoa, ts.  $E(\bar{Y}) = E(\hat{\theta}) = \theta = \mu$

**Aritmeettisen keskiarvon ominaisuuksia** - Aiempien oletusten pätiessä aritmeettisella keskiarvolla  $\bar{Y}$  on seuraava odotusarvo ja varianssi:

$$E(\bar{Y}) = \mu, \text{Var}(\bar{Y}) = \frac{\sigma^2}{n}.$$

- Aritmeettisen keskiarvon  $\bar{Y}$  standardipoikkeama

$$D(\bar{Y}) = \sqrt{\text{Var}(\bar{Y})} = \frac{\sigma}{\sqrt{n}}.$$

- Standardipoikkeamaa kutsutaan myös **keskiarvon keskivirheeksi** ja se kuvaaa otoskeskiarvon otosvaihtelua odotusarvon  $\mu$  ympärillä.
- Aritmeettisen keskiarvon otosjakauma keskittyy yhä voimakkaammin ha-vaintojen yhteen odotusarvon  $\mu$  ympärille, kun otoskoko  $n$  kasvaa.
  - Ts. otoskoon  $n$  kasvaessa  $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$  pienenee.

**Otosvarianssi** - Aineiston sisältämää vaihtelua kuvataan **otosvarianssilla**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

- Vastaavasti sm:jien vaihtelua perusjoukon tasolla kuvataan **populaatiovarianssilla**

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^N (Y_j - \mu)^2,$$

jota otosvarianssi harhattomasti estimoi. - Huomioi, että **otosvarianssi** on eri asia kuin **otoskeskiarvon varianssi**.

## 96LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Otoskeskiarvo  $\bar{Y}$  ja otosvarianssi  $S^2$  ovat siis satunnaismuuttuja, joiden saamat arvot vaihtelevat satunnaisesti otoksesta toiseen.

### Normaalijakautunut otos

- Muodostakoot havainnot  $Y_1, \dots, Y_n$  satunnaisotoksen normaalijakaumasta  $N(\mu, \sigma^2)$ .
- Tällöin voidaan osoittaa, että havaintojen  $Y_1, \dots, Y_n$  keskiarvo  $\bar{Y}$  noudattaa normaalijakaumaa odotusarvolla  $\mu$  ja varianssilla  $\sigma^2/n$ . Merkitään

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- Itse asiassa ns. **asymptoottiseen teoriaan** vedoten (suurten otosten tapauksessa) voidaan osoittaa, että edellämainittu tulos pätee myös ilman normaalisuusoletusta.
  - Nämä tarkastelut vaativat jälleen selvästi enemmän käytyjä tilastotieteen (ja matematiikan) opintoja.

### Standardoidun aritmeettisen keskiarvon otosjakauma

- Tarkastellaan **standardoitua** satunnaismuuttuja

$$Z = \frac{\bar{Y} - E(\bar{Y})}{D(\bar{Y})} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right).$$

- Tällöin  $Z$ :n odotusarvo  $E(Z) = 0$  ja varianssi  $Var(Z) = 1$ .
- Jos  $Y_i \sim N(\mu, \sigma^2), i = 1, \dots, n$ , niin tällöin  $Z$  noudattaa standardoitua normaalijakaumaa:

$$Z \sim N(0, 1).$$

- Jälleen voidaan osoittaa, että tämä tulos pätee asymptoottisesti (suurissa otoksissa) myös ilman yllä tehtyä normaalisuusoletusta.

## 6.4 Suhteellisen frekvenssin otosjakauma

### Frekvenssi ja suhteellinen frekvenssi

- Oletetaan, että tapahtuman  $A$  todennäköisyys on

$$\mathrm{P}(A) = p,$$

jolloin tapahtuman  $A$  komplementtitapahtuman (vastatapahtuman)  $A^c$  todennäköisyys on

$$\mathrm{P}(A^c) = 1 - p = q.$$

- Poimitaan satunnaisotos, jonka koko on  $n$ . Tällöin  $A$ -tyyppisten alkioiden frekvenssi eli lukumäärä kyseisessä otoksessa on  $f$ .
- Suhteellinen frekvenssi eli osuus on tällöin

$$\hat{p} = \frac{f}{n}.$$

- Sekä frekvenssi (lukumäärä)  $f$  ja (täten myös) suhteellinen frekvenssi  $\hat{p}$  ovat satunnaismuuttuja, joiden saamat arvot vaihtelevat satunnaisesti otoksesta toiseen.

### Frekvenssin otosjakauma

- Frekvenssillä  $f$  on odotusarvo

$$\mathrm{E}(f) = np,$$

ja varianssi

$$\mathrm{Var}(f) = npq = np(1 - p).$$

- Frekvenssi  $f$  noudattaa binomijakaumaa parametrein  $n$  ja  $p$ :

$$f \sim \mathrm{Bin}(n, p).$$

### Suhteellinen frekvenssi: Odotusarvo ja varianssi

- Suhteellisen frekvenssin  $\hat{p}$  odotusarvo

$$\mathrm{E}(\hat{p}) = \mathrm{E}\left(\frac{f}{n}\right) = p,$$

ja varianssi

$$\mathrm{Var}(\hat{p}) = \frac{pq}{n} = \frac{p(1 - p)}{n}.$$

## 98LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Suhteellisen frekvenssin  $\hat{p}$  standardipoikkeamaa

$$D(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{pq}{n}}$$

voidaan kutsua **suhteellisen frekvenssin keskivirheeksi** ja se kuvaa suhteellisen frekvenssin otosvaihtelua odotusarvon  $p$  ympärillä.

### Suhteellisen frekvenssin otosjakauma

- Koska  $E(\hat{p}) = p$  ja  $\text{Var}(\hat{p}) = \frac{pq}{n}$ , niin suhteellisen frekvenssin otosjakauma keskittyy yhä voimakkaammin tapahtuman A todennäköisyyden  $P(A) = p$  ympärille, kun otoskoko  $n$  kasvaa.
- Jälleen suurten otosten tapauksessa voidaan osoittaa, että suhteellinen frekvenssi noudattaa em. oletusten pätiessä normaalijakaumaa:

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right).$$

- Aritmeettisen keskiarvon tapaan standardoitutu sm.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0, 1)$$

noudattaa suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa.

### EU-kansanäänestys

- Suomen EU-kansanäänestyksessä vuonna 1994 jäsenyyttä kannataneiden suhteellinen osuus oli 0,54 (54%).
- Mikä olisi ollut tällöin tn., että ennen äänestystä 200 havainnon otoksessa kyllä-osuuus olisi ollut alle 50%?
- Suhteellisen frekvenssin otosjakauman perusteella kyllä-kannatusosuuden jakauma olisi

$$\hat{p} \sim N\left(0.54, \frac{0.54 \times (1 - 0.54)}{200}\right),$$

$$\text{jossa } \frac{0.54 \times (1 - 0.54)}{200} = 0.0352^2.$$

- Näin ollen haluttu todennäköisyys (ts. saada sellainen satunnaismuuttujan  $Z \sim N(0, 1)$  arvo että suhteellinen osuus on pienempi

kuin 0.5)

$$P\left(Z < \frac{0.5 - 0.54}{0.0352}\right) = P(Z < -1.14) \approx 0.127.$$

## 6.5 Muita tunnuslukuja

Tilastollisia analyysejä tehtäessä johtopäätösten ja objektiivisten tulkintojen tueksi tarvitaan tunnuslukuja, joita muodostetaan tarkasteltavasta jakaumasta ja mm. otoskeskiarvon osalta jo sivuttiin edellä. Tunnuslukuja on paljon, ja jokainen niistä valottaa muuttujan jakaumaa eri näkökulmista.

Jakaumien tunnusluvut voidaan jakaa sijaintilukuihin, hajontalukuihin ja muihin tunnuslukuihin. Kahdesta ensimmäisestä esimerkkejä ovat keskiarvo ja varianssi tai keskihajonta (välimatka- ja suhdeasteikon havaintojen tapauksessa). Esitellään seuraavassa vielä lyhyesti muutamia muita tunnuslukuja.

- **Moodi:** Moodi eli tyypiarvo on havaintoaineiston yleisin muuttujan arvo tai se on luokka, jolla on suurin frekvenssi.
- **Mediaani:** Mediaani on järjestetyn havaintoaineiston keskimmäinen arvo (jos havaintoarvoja on pariton määrä, parillisessa tapauksessa esitetään jompikumpi keskimmäisistä arvoista). Mediaani siis jakaa järjestetyn havaintoaineiston kahteen osaan siten, että puolet arvoista on mediaania pienempiä ja puolet arvoltaan mediaania suurempia.
  - Luokittelusteknologilla mitattaville muuttujille ei ole olemassa luontevia sijaintilukuja keskilukujen yhteydessä pl. moodi.
- Järjestysasteikolla mitatuille muuttujille voidaan mediaanin lisäksi määrittää **fraktiileja**: pp%:n fraktiili jakaa tilastoaineiston kahteen osaan siten, että kyseistä fraktiilia pienempiä havaintoarvoja on pp%.
  - Eniten käytettyjä fraktiileja ovat **kvartiilit**. **Alakvartiili**  $Q_1$  on 25%:n fraktiili, ja **yläkvartiili**  $Q_3$  on 75% fraktiili.
  - Tietyistä fraktiileista käytetään nimitystä **desili**. Ensimmäinen desili  $D_1$  on 10% fraktiili ja esim. yhdeksäs fraktiili  $D_9$  on 90% fraktiili.
- Hajontalukuja: Varianssin/keskihajonnan lisäksi, jos muuttuja on mitattu vähintään järjestysasteikolla, sille voidaan määrittää vaihteluväli ja kvartiliväli. **Vaihteluväli** kuvaa aineiston kokonaispeittoa ja siinä ilmoitetaan aineiston pienin havainto ja suurin havainto. Ts. vaihteluväli=(pienin havainto, suurin havainto). **Kvartiliväli** =  $(Q_1, Q_3)$ .

- Muita tunnuslukuja: Tilastollisen päätöksenteon yhteydessä käytettäviä tunnuslukuja ovat **vinous** ja **huipukkuus**. Vinous ja huipukkuus voidaan määritää välimatka- ja suhdeasteikon muuttujille. Vinous ja huipukkuus mittaaavat kumpikin omalla tavallaan jakauman poikkeamaa normaalijakaumasta. Normaalijakauman vinous on 0 ja huipukkuus on 3.

## 6.6 Luottamusvälit

- Satunnaisesti saadusta aineistosta laskettujen tunnuslukujen luotettavuus on tilastollisen mallin parametrien estimoinnissa keskeinen tilastollinen kysymys.
  - Otoksen poimintaan liittyvän satunnaisvaihtelon vuoksi emme voi varmuudella tietää onko saatu otokseen perustuva parametristeista "lähellä" vai "kauhana" sen todellisesta arvosta.
  - Tätä tarvitaan jokin tapa, jolla saadun parametristimaatin luotettavuutta voidaan arvioda.

### Luottamusväli

Luottamusväli on otoksen perusteella määritty väli, joka tutkijan valitsemalla todennäköisyydellä (luottamustasolla) peittää tarkasteltavan tilastollisen mallin  $f(y; \theta)$  parametrin  $\theta$  tuntemattoman todellisen arvon. Se perustetaan otostunnusluvun, estimaattorin, otosjakaumaan.

- Otoskoko on luottamusvälejä koskevissa tarkasteluissa keskeinen ja luottamusväleihin palataankin otoskoon käsittelyyn yhteydessä.
- Valittua luottamustasoa merkitään usein  $1 - \alpha$ :lla, jossa **merkitsevyys-taso (riskitaso)**  $\alpha$  on esimerkiksi  $\alpha = 0.05$ .
- Tulkinta: Jos **otantaa** jakaumasta  $f(y; \theta)$  toistetaan, niin keskimäärin  $100 \times (1 - \alpha)\%$  otoksista kontstruloiduista luottamusväleistä peittää parametrin  $\theta$  todellisen arvon.
- Oletetaan, että olemme tehneet johtopäätöksen, että konstruloitu luottamusväli peittää parametrin  $\theta$  tuntemattoman todellisen arvon.
  - Tällöin otantaa toistettaessa luottamusvälin konstruktiosista seuraa, että tehty johtopäätös on oikea keskimäärin  $100 \times (1 - \alpha)\%$  tapauksista.
  - Vastaavasti taas  $100 \times \alpha\%$  ei peitä parametrin todellista arvoa.

- Luottamusväli on kenties tunnetumpi kansankieliseltä nimitykseltään **virhemarginaali**, joka on itseasiassa luottamusvälin puolikas: todellinen parametriarvo kuuluu saadun estimaatin ja virhemarginaalien sisään jäävälle osuudelle.
  - Normaalisti mm. otoskoon kasvu pienentää virhemarginaalia.
  - Kuten jatkossa tullaan havaitsemaan, virhemarginaalin suuruuteen vaikuttavat otosasetelma, otoskoko, luottamustaso ja tutkittavan tilastollisen tunnusluvun jakauma.
- Luottamusvälissä ei kuitenkaan varsinaisesti ole kyse “virheestä” vaan saadun/muodostetun tiedon tarkkuudesta.
  - Luottamusvälit, eli virhemarginaalit, siis (yleisesti) riippuvat valitavasta luottamustasosta  $1 - \alpha$  ja näin ollen samasta aineistosta on saatavissa useita virhemarginaaleja.
    - \* Täten on tarkalleen ottaen virheellistä sanoa, että “tutkimuksen virhemarginaali on 3,5 puoleen tai toiseen”.
    - \* Oikeammin olisi sanoa esimerkiksi “tutkimuksessa saadun kannatuksen virhemarginaali on 3,5 puoleen tai toiseen 95% luottamustasolla.”
    - \* Virhemarginaali kasvaa, kun aineistoa lohkotaan: jos tuhannen hengen otoksesta esitetään tietoja, jotka kuvavat erikseen miesten ja naisten ominaisuuksia, sukupuolittain lasketut ovat estimaatit epävarmempia kuin koko otoksesta esitettyt.
  - Vastaavasti on virheellistä sanoa että tutkimuksella olisi virhemarginaali, sillä virhemarginaali liittyy aina vain tutkimuksen antamiin numeerisiin arvoihin.
  - Aitoja virhelähteitä ovat mm. otantatutkimukseen liittyvien kysymysten muotoilu, käsitteiden monitulkintaisuus, vastaajien valikointuminen ja vastauskato.

### Normaalijakauaman odotusarvon luottamusväli

- Käsittelemme seuraavassa (normaalijakauaman) odotusarvon  $\mu$  luottamusvälejä ja jatkossa oletetaan (ellei toisin mainita), että taustalla oleva populaatio,  $N$ , on ”iso” (ääretön).
  - Näin ollen ns. äärellisyyskorjausta ei käytetä (yksinkertaisuuden vuoksi).
- Tarkastellaan satunnaisotosta normaalijakauumasta  $Y_1, \dots, Y_n \perp\!\!\!\perp, Y_i \sim N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ .
  - Merkintä  $\perp\!\!\!\perp$  tarkoittaa, että sm:t  $Y_1, \dots, Y_n$  ovat riippumattomia ja samoin jakautuneita (toisinaan myös lyhyesti *iid*, joka tulee englannin kielen ilmaisusta “independent and identically distributed”).

## 102LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Tarkastellaan normaalijakauman odotusarvon  $\mu$  luottamusvälin määräämistä otannan avulla olettaen että jakauman varianssi  $\sigma^2$  on tunnettu.
  - Muistetaan että normaalijakauman odotusarvoparametrin  $E(Y_i) = \mu$  harhaton estimaattori} on aritmeettinen keskiarvo

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

- Valitaan **luottamustasoksi**  $1 - \alpha$ , eli  $\alpha$  määräää todennäköisyyden, jolla luottamusväli peittää odotusarvon  $\mu$  todellisen arvon: yleinen valinta ihmistieteissä on  $\alpha = 0.05$  tai  $\alpha = 0.1$  vastaten 95% ja 90% prosentin luottamustasoa. Luonnontieteissä  $\alpha$  on usein paljon pienempi.
- Määräätään **luottamuskertoimet**  $-z_{\alpha/2}$  ja  $z_{\alpha/2}$  (luottamusväli on kaksi-suuntainen), joille pätee

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha,$$

jossa standardoitu satunnaismuuttuja

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \left( \frac{\bar{Y} - \mu}{\sigma} \right),$$

(ks. aiemmat jaksot 6) noudattaa  $N(0, 1)$ -jakaumaa.

- $P(\cdot)$ :llä merkitään todennäköisyyssjakaumaa, ts. se on normaalijakauman jakaumafunktio ja  $z_{\alpha/2}$  on jakaumafunktion arvo pisteessä  $\alpha/2$ .
- Tällöin etsitään odotusarvoparametrille  $\mu$  sellainen arvo, jolla oheinen epäyhtälö pätee ja päädytään luottamusväliin.
- Nyt epäyhtälöketju voidaan kirjoittaa muodossa

$$-z_{\alpha/2} \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}.$$

- Joka voidaan kirjoittaa uudelleen muodossa

$$\bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

kertomalla nimittäjällä puolittain ja vähentämällä sm:jien keskiarvo molemminkin puolin.

- Normaalijakauman odotusarvon  $(1 - \alpha) \times 100\%$  luottamusväli on siis

$$\left( \bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

- Luottamusväli on symmetrinen keskipisteensä  $\bar{Y}$  suhteen. Siksi luottamusväli esitetään usein

$$\bar{Y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- Luottamusvälin pituus

$$2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- **Virhemarginaali** on luottamusvälin pituuden puolikas eli

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- Edellä tiettyyn otokseen liittyvä luottamusväli perustetaan realisoituneeseen otoskeskiarvoon  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ .
- Olisi toivottavaa pystyä konstruoimaan parametrille  $\mu$  mahdollisimman lyhyt luottamusväli, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman korkea. Molempien vaatimusten samanaikainen täyttäminen ei ole kuitenkaan mahdollista, jos otoskoko  $n$  pidetään kiinteänä:
  - Luottamustason kasvattaminen pidentää luottamusväliä, jolloin tieto parametrin  $\mu$  todellisesta arvosta tulee epätarkemmaksi.
  - Luottamusvälin lyhtenäminen pienentää luottamustasoa, jolloin tieto parametrin  $\mu$  todellisesta arvosta tulee epävarmemmaksi.

### Normaalijakauman odotusarvon luottamusväli ( $\sigma^2$ tuntematon)

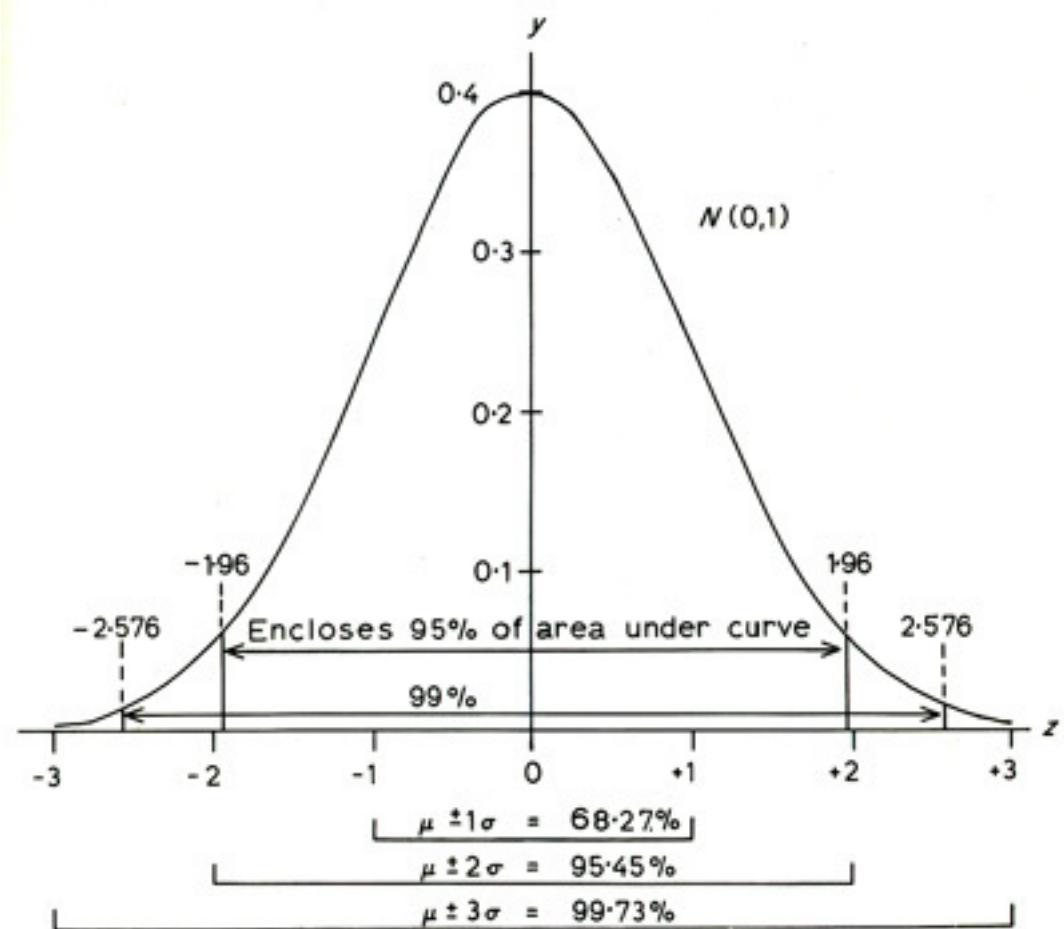
- Tarkastellaan edelleen satunnaisotosta normaalijakaumasta, mutta oletetaan nyt että varianssi  $\sigma^2$  tuntematon.
- Normaalijakauman odotusarvon  $(1 - \alpha) \times 100\%$  luottamusväli:

$$\left( \bar{Y} - t_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{Y} + t_{\alpha/2} \frac{S}{\sqrt{n}} \right),$$

jossa **luottamuskertoimet**  $-t_{\alpha/2}$  ja  $t_{\alpha/2}$  saadaan nyt  $t$ -jakaumasta}  $t_{n-1}$ , jossa  $S^2$  on varianssin  $\sigma^2$  harhaton estimaattori ja vapausasteiden lukumäärä on  $n - 1$ .

- (Studentin)  $t$ -jakauma muistuttaa silmämäärisesti normaalijakaumaa, mutta se on paksuhäntäisempi. Vapausteluvun kasvaesssa  $t$ -jakauma lähestyy normaalijakaumaa.

104LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA



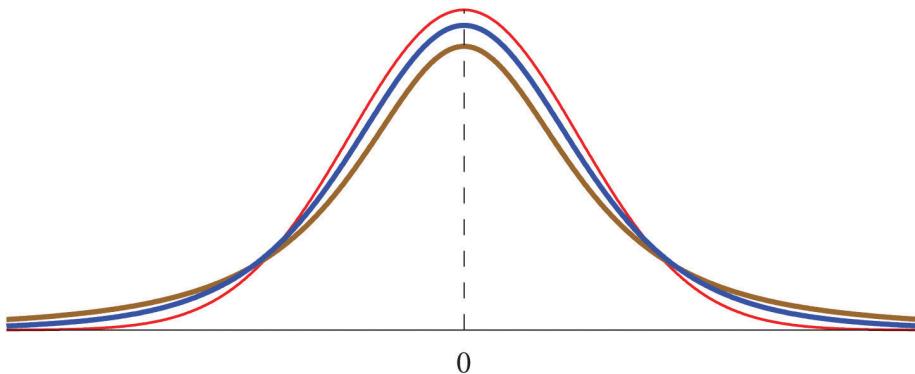
Kuva 6.2: Standardoitu normaalijakauma: Virhemarginaaleja

- Suurissa otoksissa ( $n$  iso) luottamuskertoimet voidaan poimia (appproksimatiivisesti) myös normaalijakaumasta eli korvata edellä kertoimet  $t_{\alpha/2}$  aiemmin käytetyillä kertoimilla  $z_{\alpha/2}$ .
- Normaalijakauman odotusarvon luottamusväli ( $\sigma^2$  tuntematon),  $t$ -jakauma eri vapausastein  $df$

**Standard normal**

**$t$ -distribution with  $df = 5$**

**$t$ -distribution with  $df = 2$**



Kuva 6.3: Standardoitut normaalijakauma: Virhemarginaaleja

### **Luottamusväli: Suhteellisen osuuden odotusarvo**

- Käsittelemme seuraavassa suhteellisen osuuden  $p$  luottamusvälejä.
- Tarkastellaan satunnaisotosta Bernoulli-jakaumasta  $Y_1, \dots, Y_n \perp\!\!\!\perp, Y_i \sim B(p)$ ,  $i = 1, \dots, n$ , jossa merkitään  $Y_i = 1$  jos tapahtuma A tapahtuu ja  $Y_i = 0$  jos tapahtuma A ei tapahdu.
- Bernoulli-jakauman odotusarvoparametrin  $p = E(Y_i)$  harhaton estimaattori on tapahtuman A suhteellinen otosfrekvenssi

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

## 106LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Bernoulli-jakauman (vrt. binomijakauma) ominaisuuksien perusteella  $E(Y_i) = p$  ja  $\text{Var}(Y_i) = pq$ , jossa  $q = 1 - p$ .
- Näin ollen voimme normaalijakauman odotusarvoparametrin luottamusvälin konstruloinnin tapaan määritellä satunnaismuuttujan  $Z$ :

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \sqrt{n} \left( \frac{\hat{p} - p}{\sqrt{p(1-p)}} \right),$$

joka noudattaa (suurissa otoksissa)  $N(0, 1)$ -jakaumaa.

- Suhteellisen frekvenssin hajonnan estimaattori on siis

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

jossa tuntematon  $p$  on korvattu sen estimaattorilla (otosvastineella)  $\hat{p}$ .

- Luottamuskertoimet määräätään aiempaan tapaan:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha,$$

- Näin ollen odotusarvoparametrin (suhteellisen osuuden)  $p$   $(1 - \alpha)\%$  luottamusväliksi saadaan

$$\left( \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

- Luottamusväli voidaan kirjoittaa

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

ja luottamusvälin pituus on

$$2 \times z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- Ks. kuva 6.2

## 6.7 Otoskoko

## Luku 7

# Tilastollinen riippuvuus ja korrelaatio

- Tarkastelemme tässä luvussa tilastollisia tutkimusasetelmia, joissa on muun kaksi tai useampia **muuttuja**.
- Pyrimme vastaamaan tässä ja seuraavissa luvuissa (ainakin) seuraaviin kysymyksiin:
  - Miten kahden (tai useamman) muuttujan samanaikainen tarkastelu vaikuttaa tilastolliseen analyysiin?
  - Mitä tarkoitetaan kahden muuttujan tilastollisella riippuvuudella ja miten se eroaa eksaktista riippuvuudesta?
  - Mitä tarkoitetaan korrelatiolla?
  - Mikä on korrelaation ja riippuvuuden suhde?
  - Miten korrelatiota ja sen voimakkuutta voidaan estimoida?
- Käsittelemme myös jatkossa regressioanalyysia yhden selittäjän lineaariselle regressiomallille tapauksessa. Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään koko tilastotieteen opinto-ohjelman lävitse, kuten perusteellisesti Lineaariset ja yleistetyt lineaariset mallit -kurssin myötä.

### 7.1 Muuttujien väliset riippuvuudet tilastollisen tutkimuksen kohteena

- Tieteellisen tutkimuksen tärkeimmät ja mielenkiintoisimmat kysymykset liittyvät tavallisesti **tutkimuksen kohteena olevaa ilmiötä kuvaavien muuttujien väliin riippuvuuksiin**.

- Jos tilastollisen tutkimuksen kohteena olevaan ilmiöön liittyy useampia kuin yksi muuttuja, yhden muuttujan tilastolliset menetelmät antavat tavallisesti vain rajoittuneen kuvan ilmiöstä.
- Sovellusten kannalta ehkä merkittävin osa tilastotiedettä käsittelee kahden tai useamman muuttujan välisten riippuvuuksien kuvaamista ja määrittämistä.

#### Esimerkkejä riippuvuustarkasteluista

- Miten työttömyysaste Suomessa (% työvoimasta) riippuu BKT:n (bruttokansantuotteen) kasvuvauhdista Suomessa, Suomen viennin volyymista sekä BKT:n kasvuvauhdista muissa EU-maissa ja USA:ssa? Taloustieteilijät pyrkivät yleisesti löytämään muitakin lainalaisuuksia. Esimerkkejä tällaisista ovat riskin ja tuoton välinen suhde osakesijoittamisessa, hajauttaminen pienentää riskiä ja/tai alhainen korkotaso suosii sijoittamista pörssiin.
- Miten alkoholin kulutus (l per capita vuodessa) riippuu alkoholi-juomien hintatasosta, ihmisten käytettäväissä olevista tuloiista ja alkoholin saatavuudesta?
- Miten todennäköisyys sairastua keuhkosyöpään riippuu tupakointin määristä ja kestosta?
- Miten vehnän hehtarisato (t/ha) riippuu kesän keskilämpötilasta ja sademääristä sekä maan muokkauksesta, lannoituksesta ja tuholaisien torjunnasta?
- Miten betonin lujuus (kg/cm<sup>2</sup>) riippuu sen kuivumisajasta?
- Miten kemiallisen aineen saanto (%) riippuu valmistusprosessissa käytettävästä lämpötilasta?

- **Eksakti vs. tilastollinen riippuvuus**

- Tarkastelemme tässä esityksessä yksinkertaisuuden vuoksi pääasiassa kahden muuttujan välistä riippuvuutta:
  - \* (i) Muuttujien välinen riippuvuus on **eksaktia**, jos toisen arvot voidaan ennustaa tarkasti toisen saamien arvojen perusteella.

- \* (ii) Muuttujien välinen riippuvuus on **tilastollista**, jos niiden välillä ei ole eksaktia riippuvuutta, mutta toisen muuttujan arvoja voidaan käyttää apuna toisen muuttujan arvojen ennustamisessa.
- Tilastollinen riippuvuus ja **korrelaatio**
  - Kahden muuttujan välistä (lineaarista) tilastollista riippuvuutta kutsutaan tilastotieteessä (tavallisesti) **korrelatioksi**.
  - Korrelaation eli (lineaarisen) tilastollisen riippuvuuden voimakkuutta mittaavia tilastollisia tunnuslukuja kutsutaan korrelatiokertoimiksi.
  - Korrelatiot muodostavat perustan muuttujien välisten riippuvuuksien ymmärtämiselle.
  - Vaikka korrelatiot muodostavat perustan muuttujien välisten riippuvuuksien ymmärtämiselle, riippuvuuksia halutaan tavallisesti analysoida myös tarkemmin.
  - **Regressioanalyysi** on tilastollinen menetelmä, jossa jonkin, ns. selittävän muuttujan tilastollista riippuvuutta joistakin toisista, ns. selittävästä muuttujasta pyritään mallintamaan regressiomalliksi kutsutulla tilastollisella mallilla. Käsittelemme johdatusta regressioanalyysin vielä myöhemmin luvussa 8.

## 7.2 Kahden muuttujan havaintoaineiston kuvaaminen

- Kuten yhden muuttujan havaintoaineistojen tapauksessa, lähtökohdan kahden tai useaman muuttujan havaintoaineistojen kuvaamiselle muodostaa tutustuminen havaintoarvojen jakaumaan.
- Havaintoarvojen jakaumaa voidaan kuvalla ja esitellä tiivistämällä havaintoarvoihin sisältyvä informaatio sopivan muotoon:
  - Havaintoarvojen jakaumaa kokonaisuutena voidaan kuvata sopivasti valituilla graafisilla esityksillä.
  - Havaintoarvojen jakauman karakteristisia ominaisuuksia voidaan kuvata sopivasti valituilla otostunnusluvuilla (ks. otostunnuslukuja ja otosjakaumat luvussa 6 ).
- Koska useampi- kuin kaksiulotteisten kuvioiden tekeminen ei ole usein kovin mielekästä, kolmen tai useaman muuttujan havaintoaineistoja havainnollistetaan tavallisesti niin, että muuttuja tarkastellaan pareittain.
- Kahden järjestys-, välimatka- tai suhdeasteikkoillisen muuttujan havaittujen arvojen pareja havainnollistetaan tavallisesti graafisella esityksellä, jota kutsutaan hajontakuvioksi tai pistediagrammiksi (“pistekaavio” engl. scatter plot).

- Usean muuttujan havaintoaineistojen karakteristisia ominaisuuksia voidaan kuvata muuttujakohtaisilla otostunnuslukuilla.
- Muuttujakohtaiset otostunnusluvut eivät kuitenkaan voi antaa informaatiota muuttujien välisistä riippuvuuksista.
- Muuttujien pareittaisia tilastollisia riippuvuuksia voidaan kuvata sopivasti valitulla korrelaation mitalla.

### Pistediagrammi (hajontakuvio)

- Tarkastellaan tilannetta, jossa tutkimuksen kohteina olevista havaintoyksiköistä on mitattu kahden järjestys-, välimatka- tai suhdeasteikollisen muuttujan  $X$  ja  $Y$  arvot.
- Muuttujien  $X$  ja  $Y$  arvojen samaan havaintoyksikköön liittyvien parien  $(X, Y)$  muodostamaa havaintoaineistoa voidaan kuvata graafisesti pistediagrammilla.
- Pistediagrammi sopii erityisesti kahden muuttujan välisen riippuvuuden havainnollistamiseen. Se on keskeinen työväline korrelaatio- ja regressioanalyysissä.

#### Pistediagrammi

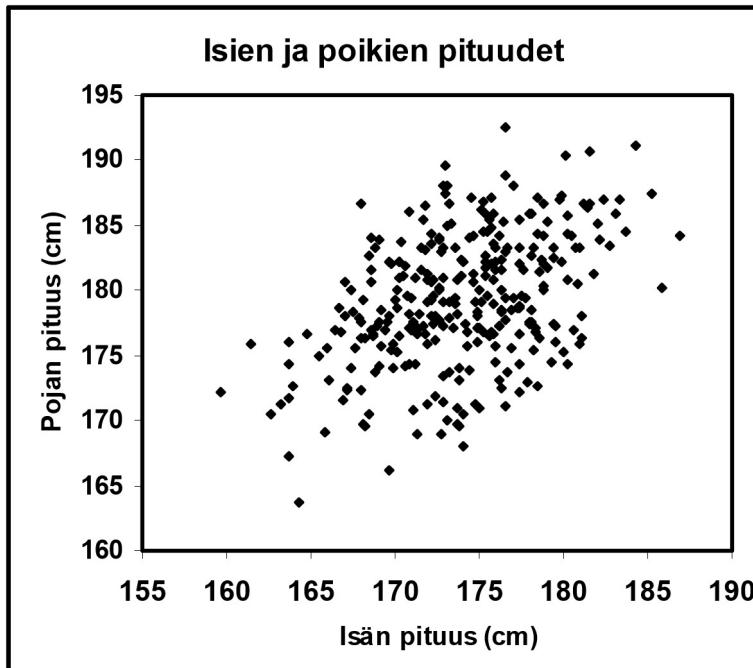
Olkoot  $X$  ja  $Y$  järjestys-, välimatka- tai suhdeasteikollisia muuttuja, joiden havaitut arvot ovat  $x_1, x_2, \dots, x_n$  ja  $y_1, y_2, \dots, y_n$ . Oletetaan lisäksi, että havaintoarvot  $x_i$  ja  $y_i$  liittyvät samaan havaintoyksikköön kaikille  $i = 1, 2, \dots, n$ . Havaintoarvojen parien  $(x_i, y_i)$  pistediagrammi saadaan esittämällä lukuparit niiden määrittelemien pisteen tasokoordinaatistossa.

#### Esimerkki: Isän ja pojantien pituus

- Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.
- Periytyykö isän pituus heidän pojilleen?
- Havaintoaineisto koostuu 300:n isän ja heidän poikiensa pituksien

muodostamasta lukuparista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, 300$ , jossa  $x_i$  = isän  $i$  pituus ja  $y_i$  = isän  $i$  pojien pituus.

- Yhtä pitkillä isillä näyttää olevan monen mittaisia poikia.
- Mutta: Lyhyillä isillä näyttää olevan keskimäärin lyhyempiä poikia kuin pitkillä isillä ja pitkillä isillä näyttää olevan keskimäärin pittempiä poikia kuin lyhyillä isillä.
- Tällaisten tilastollisten riippuvuuksien analysoimista lineaaristen regressiomallien avulla tarkastellaan myöhemmin luvussa 8 Yksinkertainen lineaarinen regressiomalli.



Kuva 7.1: Isien ja poikien pituudet. Lähde: Mellin (2006).

### 7.3 Tunnusluvut

- Kahden välimatka- tai suhdeasteikollisen muuttujan havaintoarvojen parien muodostamaa jakaumaa voidaan karakterisoida seuraavilla tunnuslukuilla:
  - Havaintoarvojen keskimääräistä sijaintia kuvataan aritmeettisilla keskiarvoilla.

- Havaintoarvojen hajaantuneisuutta tai keskityneisyyttä kuvataan keskihajonnoilla tai (otos-) variansseilla.
- Havaintoarvojen (lineaarista) riippuvuutta kuvataan otoskovariansilla ja otoskorrelatiokertoimella.
- Ts. oletetaan seuraavassa, että meillä on käytettävissä välimatka- tai suhdeasteikollisten muuttujien  $x$  ja  $y$  havaittuja arvoja  $x_1, x_2, \dots, x_n$  ja  $y_1, y_2, \dots, y_n$ . Oletetaan lisäksi, että havaintoarvot  $x_i$  ja  $y_i$  liittyvät samaan havaintoyksikköön kaikille  $i = 1, 2, \dots, n$ . Havaintoarvojen parien  $(x_i, y_i)$
- Käsitellään seuraavassa otoskeskiarvoa ja otosvarianssia. Olemme käsitelleet vastaavia estimaattoreita jo aiemmin luvussa 6.
- Havaintoarvojen  $y_1, y_2, \dots, y_n$  aritmeettinen keskiarvo on

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Vastaavalla tavalla voidaan määritellä havaintojen  $x_1, x_2, \dots, x_n$  (aritmeettinen) keskiarvo  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ .

- Havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  laskettujen aritmeettisten keskiarvojen, otoskeskiarvojen,  $\bar{x}$  ja  $\bar{y}$  muodostama lukupari  $(\bar{x}, \bar{y})$  on havaintoarvojen parien muodostamien pisteen painopiste.
- Havaintoarvojen aritmeettinen keskiarvo kuvaava havaintoarvojen keskimääräistä sijaintia.
- Osoittautuu, että (aritmeettinen) keskiarvo toimii tilastollisessa mielessä hyvänen estimaattorina satunnaismuuttujan  $y$  odotusarvolle.

**Otosvarianssi:** Havaintoarvojen  $y_1, y_2, \dots, y_n$  (otos-) varianssi (on todettu jo aiemmin) on muotoa

$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2,$$

jossa  $\bar{y}$  on  $y$ -havaintoarvojen aritmeettinen keskiarvo. - Jälleen vastaavalla tavalla voidaan määritellä  $x$ -havaintoarvojen (otos-) varianssi  $S_x^2$ . - Havaintoarvojen varianssi mittaa havaintoarvojen hajaantuneisuutta tai keskityneisyyttä havaintoarvojen aritmeettisen keskiarvon suhteen.

- **(Otos-) keskihajonta:** Havaintoarvojen  $y_1, y_2, \dots, y_n$  (otos-) keskihajonta

$$s_y = \sqrt{s_y^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2},$$

jossa  $\bar{y}$  on on y-havaintoarvojen aritmeettinen keskiarvo. Huomaa suhde (otosvarianssiin.

- Jälleen vastaavalla tavalla voidaan määritellä x-havaintoarvojen (otos-) keskihajonta  $s_x$ .
- Havaintoarvojen keskihajonta mittaa havaintoarvojen hajaantuneisuutta tai keskityneisyyttä havaintoarvojen aritmeettisen keskiarvon suhteen.

## 7.4 Satunnaismuuttujien kovarianssi ja korrelaatio

- Tarkastellaan välimatka- tai suhdeasteikollisten satunnaismuuttujien  $X$  ja  $Y$  Pearsonin (tulomomentti-) korrelatiokerrointa  $\rho_{XY}$  ja sen estimointia.
- Tällä kurssilla emme tarkastele tarkemmin tilastollisia testejä korrelatiokertoimelle  $\rho_{XY}$ , kuten: -Yhden otoksen testi korrelatiokertoimelle - Korrelatiokertoimien vertailutesti -Korreloimattomuuden testaaminen
- Jälleen kerran, lisätietoja ja tarkempia yksityiskohtia moniulotteisista satunnaismuuttujista ja jakaumista tarkastellaan todennäköisyyslaskennan kursseilla.

### Satunnaismuuttujien kovarianssi ja korrelaatio

Olkoon  $(X, Y)$  satunnaismuuttujien  $X$  ja  $Y$  muodostama järjestetty pari.  
Olkoot

$$\mu_X = E(X) \quad \text{and} \quad \mu_Y = E(Y)$$

satunnaismuuttujien  $X$  ja  $Y$  odotusarvot ja

$$\sigma_X^2 = \text{Var}(X) = D^2(X) = E[(X - \mu_X)^2] \sigma_Y^2 = \text{Var}(Y) = D^2(Y) = E[(Y - \mu_Y)^2]$$

satunnaismuuttujien  $X$  ja  $Y$  varianssit.

Määritellään satunnaismuuttujien  $X$  ja  $Y$  kovarianssi  $\sigma_{XY}$  kaavalla

$$\sigma_{XY} = \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Määritellään satunnaismuuttujien  $X$  ja  $Y$  korrelaatio  $\rho_{XY}$  kaavalla

$$\rho_{XY} = \text{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

jossa siis  $\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\text{D}^2(X)}$  ja  $\sigma_Y = \sqrt{\text{Var}(Y)} = \sqrt{\text{D}^2(Y)}$

- Satunnaismuuttujien  $X$  ja  $Y$  korrelatiota

$$\rho_{XY} = \text{Cor}(X, Y)$$

kutsutaan ajoittain **Pearsonin korrelatiokertoimeksi** (tulomomentti-korrelatiokertoimeksi).

- Pearsonin korrelatiokerroin  $\rho_{XY}$  mittaa satunnaismuuttujien  $X$  ja  $Y$  lineaarisen riippuvuuden voimakkuutta. Ts. sm:jen välistä (lineaarista) yhteyttä.

- Pearsonin (tulomomentti-) korrelatiokerroin voidaan estimoida vastavalla Pearsonin **otoskorrelatiokertoimella**

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Estimaattori  $r_{XY}$  voidaan johtaa sekä momenttimenetelmällä että suurimman uskottavuuden menetelmällä, jotka ovat tyypillisesti estimointimenetelmiä tilastotieteessä ja tarkemmin tilastollisessa päättelyssä.

#### Pearsonin otoskorrelatiokerroin

Havaintoarvojen  $(x_i, y_i)$  pareista laskettu **otoskovarianssi** on

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

jossa  $\bar{x}$  ja  $\bar{y}$  ovat havaintoarvojen  $x$  ja  $y$  aritmeettiset keskiarvot.

Otoskovarianssin  $s_{xy}$  avulla voidaan määritellä  $x$ - ja  $y$ -havaintoarvojen lineaarisen tilastollisen riippuvuuden voimakkuuden mittari, jota kutsutaan Pearsonin otoskorrelatiokertoimeksi. Pearsonin otoskorrelatiokerroin  $r_{xy}$  saadaan otoskovarianssista  $s_{xy}$  **normeerausoperaatiolla**, jossa otoskovarianssi  $s_{xy}$  jaetaan  $x$ - ja  $y$ -havaintoarvojen keskihajonnoilla  $s_x$  ja  $s_y$ .

Ts. havaintoarvojen pareista  $(x_i, y_i), i = 1, 2, \dots, n$  laskettu Pearsonin otoskorrelatiokerroin on siis

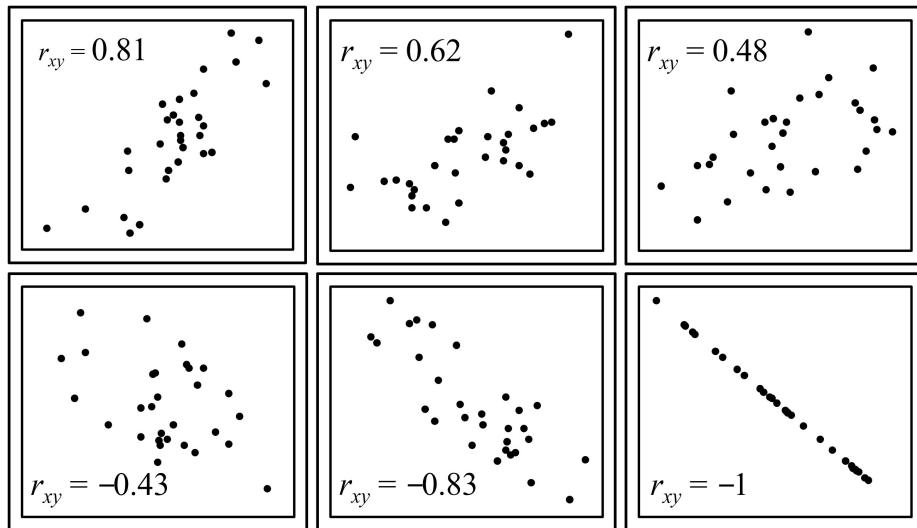
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

jossa  $s_{xy}$  on  $x$ - ja  $y$ -havaintoarvojen otoskovarianssi,  $s_x$  on  $x$ -havaintoarvojen keskihajonta ja  $s_y$  on  $y$ -havaintoarvojen keskihajonta.

- Otoskovarianssi:
  - Huomaa, että  $x$ - ja  $y$ -havaintoarvojen otoskovarianssit niiden itsensä kanssa ovat niiden variansseja.
  - Otoskovarianssi  $s_{xy}$  mittaa  $x$ - ja  $y$ -havaintoarvojen yhteisvaihtelua niiden aritmeettisten keskiarvojen ympärillä.
  - Otoskovarianssilla on taipumus saada positiivisia (negatiivisia) arvoja, jos havaintopisteiden muodostama ”pistepilvi (pisteparvi)” näyttää nousevalta (laskevalta) oikealle mentäessä; ks. pistediagrammin ilmeen ja Pearsonin otoskorrelatiokertoimen yhteys, jota käsitellään seuraavaksi.
- Pearsonin otoskorrelatiokertoimella  $r_{xy}$  on seuraavat ominaisuudet:
  - (i)  $-1 \leq r_{xy} \leq 1$
  - (ii)  $r_{xy} = \pm 1$ , jos ja vain jos  $y_i = \alpha\beta x_i$ , jossa  $\alpha$  ja  $\beta$  ovat reaalisia vakiota ja  $\beta \neq 0$
  - (iii) Korrelatiokertoimella  $r_{xy}$  ja kovarianssilla  $s_{xy}$  on aina sama etumerkki
- Pearsonin otoskorrelatiokerroin  $r_{xy}$ : Tulkinta/tulkintoja:
  - Havaintoarvojen pareista  $(x_i, y_i), i = 1, 2, \dots, n$  laskettu Pearsonin otoskorrelatiokerroin  $r_{xy}$  mittaa  $x$ - ja  $y$ -havaintoarvojen lineaarisen tilastollisen riippuvuuden voimakkuutta.
  - Jos  $r_{xy} = \pm 1$ , niin  $x$ - ja  $y$ -havaintoarvojen välillä on eksakti eli funktionaalinen lineaarinen riippuvuus, mikä merkitsee sitä, että kaikki havaintopisteet  $(x_i, y_i)$  asettuvat samalle suoralle.
  - Jos  $r_{xy} = 0$ , niin  $x$ - ja  $y$ -havaintoarvojen välillä ei voi olla eksaktia lineaarista riippuvuutta.
  - Vaikka  $r_{xy} = 0$ ,  $x$ - ja  $y$ -havaintoarvojen välillä saattaa silti olla jopa eksakti epälineaarinen riippuvuus.
- **Havainnollistus:** Kuviot alla havainnollistavat kahden muuttujan havaittujen arvojen ( $n = 30$ ) pistediagrammin ilmeen ja korrelaation välistä yhteyttä.
  - Toinen havainnollistus: Ks. seuraavasta linkistä lisää havainnollistuksia. Arvio korrelaation voimakkuutta erilaisissa simuloiduissa tilanteissa:

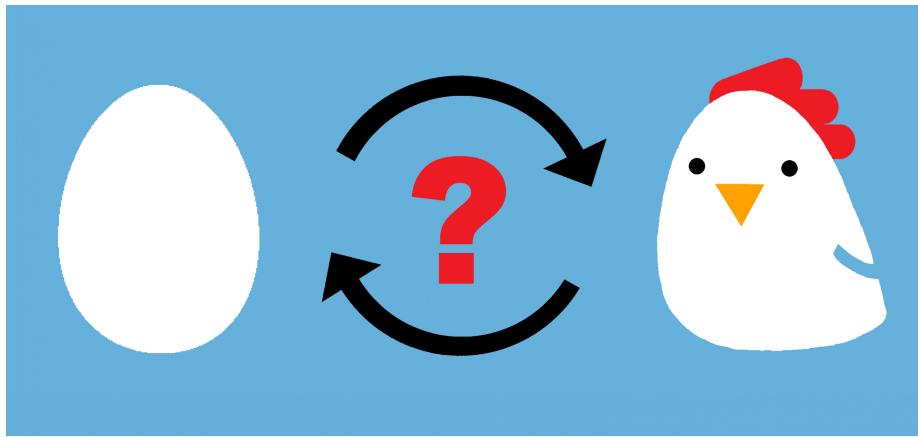
```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed,
```

- **Kausaalisuus**

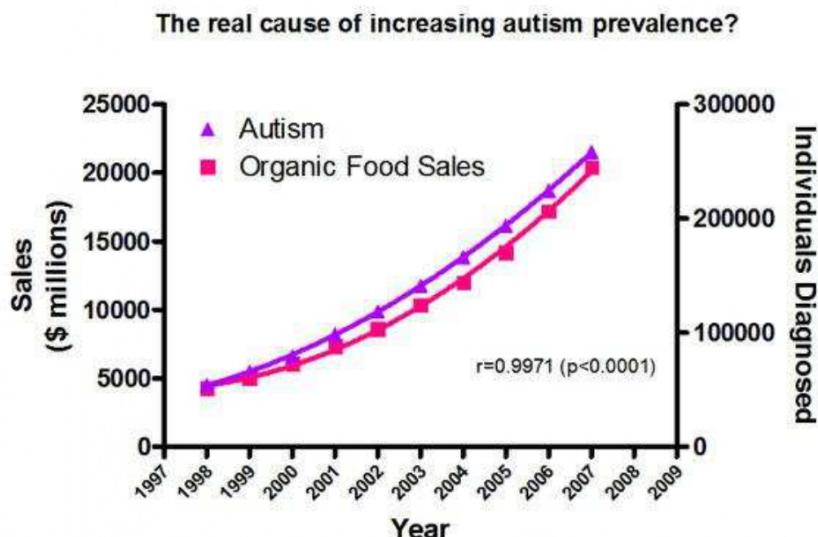


Kuva 7.2: Havainnollistuksia Pearsonin otoskorrelatiokertoimen arvosta ja erilaisista  $xy$ -pisteparvista. Lähde: Mellin (2006).

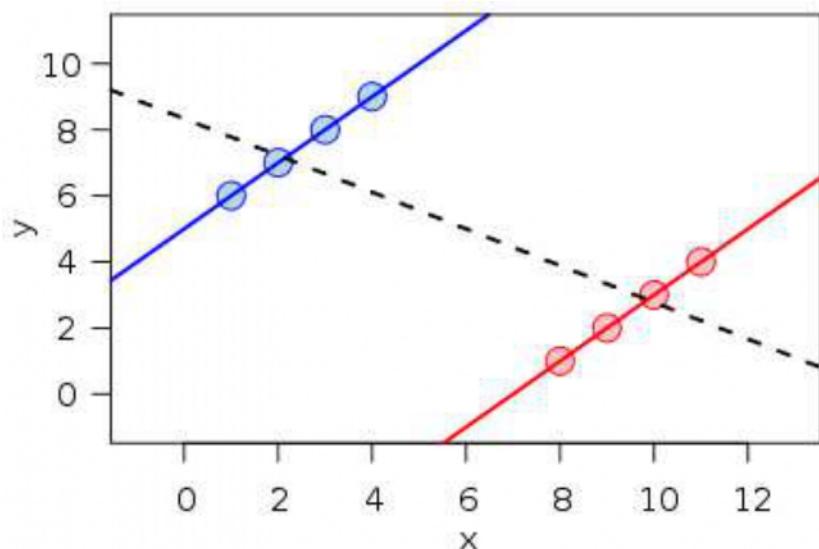
- Muuttujan  $x$  arvojen muutos vaikuttaaa muuttujan  $y$  arvoihin (syvaikutussuhde), jos seuraavat kolme ehtoa täytyvät:
  - \* muuttujan  $x$  muutos esiintyy ajallisesti ennen  $y$ :n muutosta
  - \* muuttujissa  $x$  ja  $y$  tapahtuvien muuttujien välillä on riippuvuutta
  - \* muuttujassa  $y$  tapahtunutta muutosta ei voida selittää millään mulla tekjöillä
- Kausaalisuhteita selvitettäessä on tunnettava etukäteen ilmiötä koskevat aiemmat teoriat ja tutkimukset tarkasti, jotta voidaan ottaa huomioon ilmiöön vaikuttavat tekijät
- Todellisuus on usein monimutkaisempi, kuin mitä kausaalishdke kuvaa: **kahden muuttujan yhteisvaihtelu ei riitä todisteeksi siitä, että kyseessä olevien muuttujien välillä on kausalistä yhteyttä**
- Yhteisvaihtelu voi johtua myös kolmannen muuttujan vaikutuksesta molempien muuttuihin tai virheellisestä otannasta, vaikka muuttujat olisivatkin perusjoukossa toisistaan riippumattomia
- Simpsonin paradoksi
  - Simpsonin paradoksi syntyy, kun kahden muuttujan välinen korrelaatio muuttuu päinvastaiseksi, otettaessa huomioon jokin kolmas muuttuja, joka korreloii molempien muuttujien kanssa



Kuva 7.3: Kausaalisuus: kumpi tuli ensin?



Kuva 7.4: Esimerkkejä: luomuruoka syypää lisääntyneisiin autismitapauksiin?



Kuva 7.5: Simpsonin paradoksi

**Esimerkki: Berkeleyn sukupuolisyrjintä**

Yksi tunnetuimmista esimerkeistä Simpsonin paradoksista on Berkeleyn yliopiston sukupuolisyrjintätapaus. Yliopisto haastettiin oikeuteen vuonna 1973 sukupuolisyrjinnästä. Väitettiin, että yliopistoon olisi miesten helpompi päästää kuin naisten. || Hakijat | Hyväksytyt | |---|---| | Miehet|8442| 44% | Naiset|4321| 35% |

Taulukosta nähdään, että mieshakijoista on päässyt 9 prosenttiyksikköä enemmän sisälle kuin naisista.

- Tarkasteltaessa erikseen eri tiedekuntia huomataan, että itseissä useammassa tiedekunnissa naisia on päässyt sisälle isompi osuus hakijoista. Aineisto kuudesta isoimmaasta tiedekunnasta on listattu alla olevaan taulukkoon. || Miehet | | | Naiset | | | Tdk|Hakijat|Hyväksytyt%|Hakijat|Hyväksytyt%| |---|---| |---|---| |A|825|62|108| **82**| |B|560|63|25| **68**| |C|325| **37**|593|34| |D|417|33|375| **35**| |E|191| **28**|393|24| |F|373|6|341| **7**|

- Vielä tiivistäen korrelatiokertoimen tulkintavirheitä aiheuttavat useimmiten seuraavat seikat:
  - Riippuvuudesta ei välttämättä seuraa syy-seuraussuhdetta.
  - Kolmas muuttuja eli kahden muuttujan välinen yhteys selittyy yhtisestä syystä (esimerkiksi lämpimästä kesäästä).
  - Muuttujien välinen yhteys ei ole lineaarinen.
  - Poikkeavien havaintojen vaikutus.
- Puutteita: Korrelatiokertoimella on kaksi puutetta:
  - Se mittaa vain lineaarista riippuvutta.
  - Se ei ole (tilastollinen) malli, jonka avulla nähtäisiin, miten toinen muuttuja vaikuttaa toiseen muuttujaan.



## Luku 8

# Regressioanalyysi

Tilastollinen riippuvuus ja korrelaatio -jakson laajennuksena pyrimme tässä luvussa vastaamaan seuraavaan kysymykseen: *Miten jonkin selitettävän muuttujan tilastollista riippuvuutta joistakin toisista, selittäviksi muuttujiksi kutsutusta muuttujista voidaan mallintaa?* Muuttujien välisen riippuvuuksien, eli erilaisten tosielämän asioiden ja ilmiöiden välisen yhteyksien analysointi on tavallisesti keskeinen kysymys tieteellisessä tutkimuksessa. Regressioanalyysi ja -mallintaminen on yksi tunnetuimpia ja eniten sovellettuja **tilastollisia menetelmiä** kuvaamaan kahden muuttujan **tilastollista riippuvuutta**.

Jos tilastoaineistossa on havaittavissa säädönmukaisuutta ja muuttujien välillä näyttäisi olevan järkevä (asioinen) yhteys, niin päästään “malliajatteluun”. Ts. pyritään rakentamaan tilastollista mallia kys. aineistolle. Pyritään siis muodostaa tilastollinen malli että se valitun kriteeristön perusteella parhaiten kuvaaa analysoitavaa pistejoukkoa.

### 8.1 Johdatus regressioanalyysin ideaan

- Regressioanalyysi pyrkii siis havaintoaineiston perusteella **mallintamaan tilastoyksikköjen tilastollisten muuttujien välistä riippuvuutta**.
  - Regressiomallissa tilastollisia muuttujia on kahdenlaisia: selitettyä muuttuja, jonka tilastollista vaihtelua pyritään selittämään selittävän muuttujan vaihtelulla.
  - Toisin sanoen, pyritään erottamaan se selitettävän muuttujan arvojen vaihtelu, joka voidaan selittää selittävän muuttujan arvojen vaihtelulla siitä vaihtelusta, joka on täysin satunnaista.
    - \* Esimerkiksi voitaisiin tutkia selittääkö vaaleissa puolueiden/ehdokkaiden vaalimainontabudjetti heidän äänimääriään, ja jos, niin kuinka paljon?

- \* Jos **tilastollisesti merkitsevä osa** selitettävän muuttujan havaittujen arvojen vaihtelusta voidaan selittää selittävien muuttujien arvojen vaihtelun avulla, sanomme, että selitettävä muuttuja **riippuu tilastollisesti** selittäjinä käytetyistä muuttujista.
- Yleisemmin regressioanalyysi pyrkii vastaamaan seuraaviin kysymyksiin koskien tilastollisten muuttujien välistä riippuvuutta:
  - Muuttujien välisen **riippuvuusien kuvaaminen**. Millainen on riippuvuuden muoto? Kuinka voimakasta riippuvuus on?
  - Muuttujien välisen **riippuvuusien selittäminen**. Tilastollisen riippuvuuden luonteen selittäminen.
  - Selitettävän muuttujan käyttäytymisen **ennustaminen**.
- **Lineaarin regressioanalyysi** siis (teknisesti) rajoittuu muuttujien *lineaaristen* riippuvuusien kuvaamiseen. Kuitenkin, laajemmin asiaa pohdittaessa, lineaaristen regressiomallien suuri käytökelpoisuus muuttujien välisen riippuvuusien tilastollisessa analyysissa perustuu (ainakin) seuraaviin seikkoihin:
  - Lineaarillisella regressiomallilla voidaan usein vähintään kohtuullisella (riittävällä) tarkkuudella approksimoida epälineaarisiakin muuttujien väisiä riippuvuuksia!
  - Muuttujien välinen epälineaarinen riippuvuus voidaan usein myös linearisoida käytäen sopivia muunnoksia alkuperäisiin muuttuijiin.
  - Epälineaariset regressiomallit muodostavat oman tilastollisten (regressio)mallien luokkansa (joita ei käsitellä tällä kurssilla, mutta kylläkin myöhemmissä tilastotieteen opinnoissa).
- Regressiomalleja käytetään apuvälineinä monilla tilastotieteen osa-alueilla. Esimerkkejä regressiomallien käyttökohteista tilastotieteessä:
  - Varianssianalyysi
  - Koesuunnittelu
  - Monimuuttujamenetelmät
  - Biometria/biostatistiikka
  - Aikasarja-analyysi ja ennustaminen
  - Ekonometria
- Regressioanalyysissä sovellettavat tilastolliset mallit voidaan luokitella usealla eri periaatteella.
  - Luokittelun regressiomallin funktionaalisen muodon mukaan:
    - \* Lineaariset regressiomallit
    - \* Epälineaariset regressiomallit

- Luokittelu regressiomallin yhtälöiden lukumäärän mukaan:
  - \* Yhden yhtälön regressiomallit
  - \* Moniyhtälömallit

Tällä kurssilla käsitellään vain **lineaarisia yhden yhtälön regressiomalleja**. Kuitenkin luvussa 8.3 [alapuolella] esitellään lyhyesti minkälaisia laajennuksia tälle regressioanalyysin perustilanteelle tyypillisesti käsitellään.

## 8.2 Yhden selittäjän lineaarinen regressiomalli

- Yhden selittäjän lineaarinen regressiomalli pyrkii selittämään selitettävän muuttujan havaittujen arvojen vaihtelun yhden selittävän muuttujan havaittujen arvojen vaihtelon avulla. Se on siis yksinkertaisin esimerkki yhden yhtälön lineaarisista regressiomalleista, sillä se sisältää vain yhden selittävän muuttujan useamman sijaan.
  - Selitettävää muuttuja kutsutaan usein myös *vastemuuttujaksi, riippuvaksi muuttujaksi tai tulosmuuttuja*
  - Vastaavasti selittävää muuttuja kutsutaan paikoin *selittäjäksi, riippumattomaksi muuttujaksi tai ennustavaksi muuttujaksi*.
- Tässä luvussa tarkastellaan seuraavia yhden selittävän muuttujan lineaarisesta regressiomallin soveltamiseen liittyviä kysymyksiä:
  - Miten malli formuloidaan?
  - Mitkä ovat mallin osat ja mitkä ovat osien tulkinnat?
  - Mitkä ovat mallia koskevat oletukset?
  - Miten mallin parametrit estimoidaan?
  - Miten mallin parametreja koskevia hypoteeseja testataan?
  - Miten mallin hyväyyttä mitataan?
  - Miten mallilla ennustetaan?
- Oletetaan, että selitettävän muuttujan  $Y$  havaittujen arvojen vaihtelu halutaan selittää selittävän muuttujan eli selittäjän  $x$  havaittujen arvojen vaihtelon avulla. Tulkitaan selitettävä muuttuja tässä kohtaa kiinteäksi eli sen arvot oletetaan tunnetuksi.<sup>1</sup>
- Tehdään siis seuraavat oletukset:
  - (i) Selitettävä muuttuja  $Y$  on suhdeasteikollinen satunnaismuuttuja.
  - (ii) Selitettävä muuttuja  $x$  on kiinteä eli ei-satunnainen muuttuja.

---

<sup>1</sup>Kyseinen muuttuja voidaan myös tulkita satunnaismuuttujana eikä seuraavat tarkastelut muutu ratkaisevasti tämän seurausena. Tätä pohditaan vielä tarkemmin alempana.

- Olkoot  $y_1, y_2, \dots, y_n$  selitettävän muuttujan  $Y$  ja  $x_1, x_2, \dots, x_n$  selittävän muuttujan  $x$  havaittuja arvoja. Oletetaan lisäksi, että havaintoarvot  $x_i$  ja  $y_i$  liittyvät samaan havaintoyksikköön kaikille  $i = 1, 2, \dots, n$ .
  - Matemaattisemmin tämä tarkoittaa sitä, että tällöin havaintoarvot  $x_i$  ja  $y_i$  muodostavat pisteitä 2-ulotteisessa avaruudessa.

- Oletetaan seuraavaksi, että havaintoarvojen  $y_i$  ja  $x_i$  välillä on **lineaarinen tilastollinen riippuvuus**, joka voidaan ilmaista yhtälöllä

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

- Tämä yhtälö määrittelee yhden selittäjän lineaarisen regressiomallin, jossa
  - $y_i$  on selitettävän muuttujan  $Y$  satunnainen ja havaittu arvo havaintoyksikölle  $i$
  - $x_i$  selittävän muuttujan eli selittäjän  $x$  ei-satunnainen ja havaittu arvo havaintoyksikölle  $i$
  - $\varepsilon_i$  on virtermi (ajoittain myös jäännöstermi) ja sen satunnainen ja ei-havaittu arvo havaintoyksikölle  $i$
- Yhden selittäjän lineaarisessa regressiomallissa on seuraavat regressiokerroimet:
  - $\beta_0$  on vakioselittäjän regressiokerroin;  $\beta_0$  on ei-satunnainen ja tuntematon vakio. Kerrointa  $\beta_0$  kutsutaan myös vakioselittäjän regressiokerimeksi. Nimitys johtuu siitä, että kerrointa  $\beta_0$  vastaa keino-tekoinen selittäjä, joka saa kaikille havaintoyksiköille  $i = 1, 2, \dots, n$  vakioarvon 1.
    - \* Huomautus: Jatkossa esitettävät kaavat eivät välittämättä pädeesitettävässä muodossa, jos mallissa ei ole vakiota (vakioselittäjää), joka yleensä automaatisesti lisätään mukaan malliin.
    - \* Oletamme jatkossa, että mallissa on aina vakioselittäjä.
  - $\beta_1$  on selittäjän  $x$  regressiokerroin;  $\beta_1$  on ei-satunnainen ja tuntematon vakio
    - \* Huomautus: Regressiokerotimet  $\beta_0$  ja  $\beta_1$  on oletettu samoiksi kaikille havaintoyksiköille  $i$ .
- Virhetermeistä  $\varepsilon_i$  tehtävät ns. standardioletukset ovat seuraavat:
  - (i)  $E(\varepsilon_i) = 0, i = 1, 2, \dots, n$
  - (ii) Virhetermeillä on vakiovarianssi eli ne ovat homoskedastisia:  $\text{Var}(\varepsilon_i) = \sigma^2, i = 1, \dots, n$ . Virhetermi  $\varepsilon_i$  tässä yhteiseksi oletettua varianssia kutsutaan jäännösvarianssiksi.

- (iii) Jäännöstermit ovat korreloimattomia:  $\text{Cov}(\varepsilon_i, \varepsilon_l) = 0, i \neq l$
  - (iv) Lisäksi tehdään ajoittain normaalisuusoleitus eli että virhetermit ovat normaalisti jakautuneita:  $\varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \dots, n$ .
- \* Huomautus: Oletus (iv) sisältää oletukset (i) ja (ii).
- Lineaarisen regressiomallin perusoletuksiin kuuluu se, että selittävien muuttujien arvot ovat ei-satunnaisia. On kuitenkin syytä korostaa (jo tässä vaiheessa), että selittävän muuttujan arvojen satunnaisuus ei kuitenkaan vaikuta mallin estimoinnissa ja testauksessa käytettäviin menetelmiin seuraavissa tilanteissa:
    - Tavanomaiset mallista tehdyt oletukset pätevät (sopivasti modifioituina), kun siirrytään tarkastelemaan selittävän muuttujan ehdollista odotusarvoa selittäjien suhteen.
    - Voidaan (ajoittain) olettaa, että selitettävä muuttuja ja selittäjät noudattavat yhdessä **multinormaalijakamaa** eli aiemmin esitelyn yksiuotteisen normaalijakauman moniuotteista laajennusta.
  - Regressioanalyysille voidaan esittää kaksi asialoogisesti varsin erilaista lähtökohtaa, joilla on kuitenkin myös monia yhtymäkohtia:
    - (i) Ongelmat determinististen mallien sovittamisessa havaintoihin: Havainnoille postuloitu malli ei sovi täsmällisesti kaikkiin havaintoihin. Tämä onkin osaltaan tilastollisen mallinnuksen yksi ominaispiirteistä: Täydellistä sopivuutta aineiston kanssa ei käytännössä koskaan saavuteta.
    - (ii) Tavoitteena on moniuotteisen todennäköisyysjakauman regressiofunktion parametrien estimointi.
      - \* Vaikka moniuotteisten todennäköisyysjakaumien regressiofunktiot ovat yleisesti epälineaarisia, lineaariset regressiomallit muodostavat tärkeän ja paljon sovelletun malliluokan.
  - Koska regressiokertoimet  $\beta_0$  ja  $\beta_1$  sekä jäännösvarianssi  $\sigma^2$  ovat tavallisesti tuntemattomia, niiden arvot on **estimoitava** muuttujien  $x$  ja  $Y$  havaittuja arvoja  $x_i$  ja  $y_i, i = 1, 2, \dots, n$  käytäen.
    - Regressiomallien parametrien estimointiin käytetään tavallisesti **pienimmän neliösumman (PNS) menetelmää**. Tämän estimointimenetelmän tarkemmat yksityiskohdat ovat myöhempien tilastotieteen kurssien asioita, mutta seuraavassa kuitenkin muutamia lähtökohtia mihin PNS-menetelmä perustuu yhden selittäjän mallin tapauksessa.

- Edellä esitellyn yhden selittäjän lineaarisen regressiomallin regressiokertoimien  $\beta_0$  ja  $\beta_1$  estimaattorit määritetään minimoimalla virhetermien  $\varepsilon_i$  neliösummaa

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  suhteen.

- Tämä minimointi tapahtuu tavanomaiseen tapaan derivoimalla funktio  $S(\beta_0, \beta_1)$  kertoimien  $\beta_0$  ja  $\beta_1$  suhteen ja merkitsemällä derivaatat nolliksi:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0.$$

- Nämä ns. normaaliyhtälöt johtavat lopulta pienen sieventämisen jälkeen regressiokertoimien  $\beta_0$  ja  $\beta_1$  pienimmän neliösumman (PNS-) estimaatoreihin (ja lopulta käytännössä analysoitavasta aineistosta laskettaviin PNS-estimaatteihin)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}.$$

- Huomaa siis yhteys aiemmin keskusteltuihin  $x$ :n ja  $y$ :n otoskeskiarvioihin, keskihajontoihin sekä otoskovarianssiin ja korrelaatioon  $x$ :n ja  $y$ :n välillä.

- PNS-estimaattorit (estimaatit)  $\hat{\beta}_0$  ja  $\hat{\beta}_1$  määrittelevät suoran (matematisesti katsoen avaruudessa  $\mathbb{R}^2$ ):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

jossa

- $\hat{\beta}_0$  on estimoidun regressiosuoran ja pistekuvion y-akselin leikkauspiste
- $\hat{\beta}_1$  on estimoidun regressiosuoran kulmakerroin

- Tämän suoran tuottamat arvot  $\hat{y}_i$  ovat käytännössä eri havainnoille  $y$  saatavat **sovitteet** lineaariseen malliin perustuen.

- Sijoitetaan regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattoreiden lausekkeet estimoidun regressiosuoran lausekkeeseen. Tällöin estimoidun regressiosuoran yhtälö voidaan kirjoittaa muodossa:

$$y = \bar{y} + r_{xy} \frac{s_y}{s_x} (x - \bar{x})$$

– Yhtälöstä nähdään, että estimoitu regressiosuora kulkee havaintopisteiden  $(x_i, y_i), i = 1, 2, \dots, n$  painopisteen kautta. Voidaan siis nähdä, että estimoidulla regressiosuoralla on seuraavat ominaisuudet:

- \* (i) Jos  $r_{xy} > 0$ , suora on nouseva.
- \* (ii) Jos  $r_{xy} < 0$ , suora on laskeva.
- \*(iii) Jos  $r_{xy} = 0$ , suora on vaakasuorassa.
- \*(iv) Suora jyrkkenee (loivenee), jos
  - korrelaation itseisarvo  $|r_{xy}|$  kasvaa (pienenee)
  - keskihajonta  $s_y$  kasvaa (pienenee)
  - keskihajonta  $s_x$  pienenee (kasvaa)

- Tarkastellaan vielä estimoituun lineaariseen malliin liittyvät sovitteet ja residuaalit.
  - Estimoidun mallin **sovitteet** saadaan siis kaavalla

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n.$$

- Vastaavasti **residuaalit** saadaan havaintojen ja sovitteiden erotuksena

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n.$$

- Sovite on estimoidun regressiosuoran yhtälön selitettyvälle muuttujalle antama arvo havaintopisteessä  $x_i$ . Vastaavasti residuaali on selitettyvän muuttujan havaitun arvon  $y_i$  ja sovitteen  $\hat{y}_i$  eli estimoidun regressiosuoran yhtälön selitettyvälle muuttujalle havaintopisteessä  $x_i$  antaman arvon erotus. - Estimoitu regressiomalli selittää selitettyvän muuttujan havaittujen arvojen vaihtelun sitä paremmin mitä lähempänä estimoidun mallin sovitteet  $\hat{y}_i$  ovat selitettyvän muuttujan havaittuja arvoja  $y_i$ . - Yhtäpitävästi edellisen kanssa: Estimoitu regressiomalli selittää selitettyvän muuttujan havaittujen arvojen  $y_i$  vaihtelun sitä paremmin mitä pienempiä ovat estimoidun mallin residuaalit  $\hat{\varepsilon}_i$ .

- Liittyen vielä estimoidun mallin sopivuuden tarkasteluun, estimoidun regressiomallin hyväyyttä mitataan (tavanomaisesti) mm. **selitysasteella** ( $R^2$ ).
  - Selitysasteen määritelmä perustuu ns. varianssianalyysihajotelmaan, jossa selittävän muuttujan havaittujen arvojen vaihtelua kuvaava neliösumma on jaettu kahdeksi neliösummaksi, joista toinen kuva mallin ja havaintojen yhteensopivuutta ja toinen mallin ja havaintojen yhteensopimattomuutta.
  - Selitysaste saa arvoja nollan ja ykkösen väliltä (kun lineaarisessa regressiomallissa on mukana vakiotermi). Arvo 0 tarkoittaa, että malli (yhden selittäjän mallissa käytännössä siis selittäjä  $x$ ) ei selitä  $y$ :n lineaarista vaihtelua yhtään (yli vakiotermin). Ts. määritelty malli ei ollenkaan selitä selittävän muuttujan havaittujen arvojen vaihtelua.
  - Vastaavasti arvo  $R^2 = 1$  tarkoittaa, että malli sopii täydellisesti aiheistoon. Ts. selitysaste mittaa regressiomallin selittämää osuutta selittävän muuttujan havaittujen arvojen kokonaisvaihtelusta.
  - Korkea selitysasteen arvo on siis sinänsä usein toivottava lopputulos lineaarisen mallin käytön yhteydessä. Tämän liian mekaaninen tavoittelut johtaa kuitenkin ajoittain muihin ongelmisiin, kuten **ylisovittamiseen** usean selittäjän lineaarisia malleja käsiteltäessä.

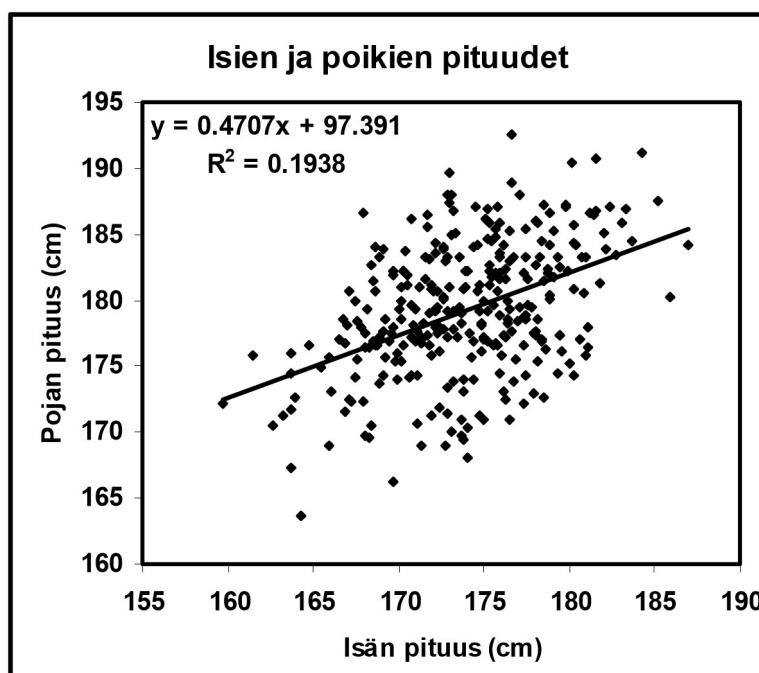
**Esimerkki: isän ja poikien pituus, tarkemmin**

Jatketaan isän ja heidän poikiensa pituutta koskevan aineiston tarkastelua. Periytyykö isän pituus heidän pojilleen? Käytännössä jo aiemmin tarkastelimme 300 havainnon havaintoaineistoa isän ja heidän poikiensa pituksien muodostamista lukupareista.

Estimoidun regressiosuoran yhtälö on (ks. oheinen kuva 8.1)

$$y = 97.391 + 0.4707x$$

Suoran kulmakertoimen  $\hat{\beta}_1 = 0.4707$  tulkinta on siis, että jos isä A on 1 cm pitempi kuin isä B, isä A:n poika on keskimäärin 0.4707 cm pitempi kuin isä B:n poika.



Kuva 8.1: Isien ja poikien pituudet: regressiosuoran sovite

### 8.3 Muita regressiomalleja

- Yksinkertaista lineaarista regressiomallia voidaan laajentaa monin tavoin monenlaisiin erilaisiin tilanteisiin.
  - Usean selittäjän lineaarinen regressiomalli: Yhden selittäjän sijaan käytetään useita selittäviä muuttujia.
  - Lineaarisen mallin sijaan malli voi olla myös epälineaarinen (epälineaarinen regressiofunktio).
- Erityisen tärkeitä laajennuksia ilmenee kun **vastemuuttuja on muuta muotoa** mitä edellä oletetaan lineaarisissa regressiomalleissa, joissa käytännössä oletetaan että vaste on reaaliarvoinen (jokin reaaliluku).
  - Vaste voi olla myös **diskreettiarvoinen**, kuten **binääriinen** ( $Y_i = 0$  tai  $Y_i = 1$ ) tai **lukumäärä** ( $Y_i \in \{0, 1, 2, 3, \dots\}$ )
  - Mikäli vaste on binääriinen, niin tällöin tyypillinen tarkasteltava ja täsmennettävä tilastollinen malli on **logistinen regressiomalli** (tunnetaan myös **logistisena regressiomallina** tai **logit-mallina**).
  - Jos vaste on lukumäärä, niin tällöin yksi mahdollinen malliliuokka on ns. **Poisson-regressiomalli**. Tässä yhteydessä oletetaan siis, ettei sm.  $Y$  noudattaa Poisson-jakaumaa ja regressiomalli rakennetaan tämän oletuksen ympärille.
- **Vastemuuttujan roolin/luonteen selvittäminen on hyvin keskeistä tilastollista mallia rakennettaessa.** Tässä pätee samat eroavaisudet mitkä tulevat tutuksi todennäköisyyslaskennan kursseilla kun käsitellään diskreettien ja jatkuva-arvoisten satunnaismuuttujien jakaumia ja näihin liittyviä yksityiskohtia.
- Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään useilla myöhemmillä tilastotieteen kursseilla.
  - Perusopintojen jälkeen aineopintojen tilastollisen päättelyn kurssien TILM3561- TILM3562 jälkeinen TILM3588 Lineaariset ja yleistetyt lineaariset -mallit kurssilla. Näistä jälkimmäisessä tarvitaan myös lineaarialgebran ja matriisilaskennan tietoja, joita tilastotieteen yhteydessä käydään läpi TILM3574 Matriisilaskenta tilastotieteessä - kurssilla.
  - Tämän jälkeen regressiomallien käsitteily jatkuu useilla eri aineopintojen ja syventävien opintojen erikoiskursseilla.

## Luku 9

# Tilastotieteen rooli uuden tiedon tuottamisessa

Tilastotieteen yhteiskunnallisesta roolista keskusteltiin luvuissa 2 ja 3. Tilastotieteen keskeinen yhteiskunnallinen rooli liittyy keskeisesti juuri uuden tieteellisen tiedon tuottamiseen: tilastotiede liittyy olennaisesti kaikkeen tieteesseen, joten ei liene yllätys että tilastotiede on jossain määrin tuttua kaikille tieteentekijöille. Tilastotiede tarjoaa pohjan uuden tiedon tuottamiselle, mutta toisaalta voitaisiin myös ajatella teoreettisen tilastotieteen ja siellä luotujen menetelmien ylipäätään mahdollistaneen uskottavan tieteenteon. Tässä luvussa emme kuitenkaan takerru tähän ”muna vai kana?”-ongelmaan, vaan tarkastelemme yleisemmällä tasolla tilastotieteen roolia tieteenteossa.

Ensiksi tarkastelemme kaikista tilastollisia menetelmiä hyödyntävistä ongelmienasetteluista löydettäviä yhteisiä elementtejä. Nämä elementit ovat niin yleisiä että niitä voidaan tarkastella ja kuvata ilman yhteyttä mihinkään yksittäiseen ongelmaan. Tämän jälkeen tarkastelemme tilastollisia menetelmiä hyödyntävän tieteellisen tutkimusprosessin eri vaiheita yleisesti. On kuitenkin mahdotonta koostaa yleisiä ”tee se näin”-listoja tilastollisen tutkimuksen toteuttamiseksi, joten tarkastelemme tähän asti kurssilla käsitellyjä asioita ja yleisiä elementtejä, jotka jokaisen tieteentekijän tulee osata ja muistaa.

### 9.1 Tilastollisen tutkimuksen yhteisiä elementtejä

#### 1. Satunnaisvaihtelu

- Satunnaismieloiden generoima havaintoaineisto on aina tilastollisen tutkimuksen tutkimuskohde. Täten kaikki tieteellinen tutkimus, joka koskee

## 132 LUKU 9. TILASTOTIETEEN ROOLI UUDEN TIEDON TUOTTAMISESSA

satunnaisvaihtelua ilmentäävä aineistoa on (tai tulisi olla) tilastotieteellistä.

- Tilastollisen tutkimuksen tavoitteena on (useimmiten) pyrkiiä erottamaan satunnaisilmiön systemaattinen ja satunnainen vaihtelu. Tämä vaatii substanssiosaamisen lisäksi menetelmäosaamista sekä hyvää tilastotieteellistä intuiota.
- Satunnaisvaihtelun ”välttämättömyys” satunnaisilmiöiden tutkimuksessa on tiedostettava ja ymmärrettää. Tämä on tärkeää niin luotettavan tiedontuotannon kuin tutkijan oman uskottavuuden vuoksi. Tilastollisten menetelmien huonon osaamisen vuoksi tehty (ja mahdollisesti julkaistu) tutkimus voi pahimillaan asettaa kyseisen aiheen tutkimuksen vuosiksi väärille raiteille!

### 2. Ilmiön ja ongelman hahmottaminen järjestelmäksi

- Tutkimusongelman substanssiosaaminen on erityisen tärkeää tilastollisessa tutkimuksessa: on osattava tunnistaa kaikki satunnaisilmiöön mahdollisesti vaikuttavat osatekijät, jotka muodostavat satunnaisen järjestelmän.
- Järjestelmä on joukko toisiinsa liittyviä asioita tai osia, jotka toimivat yhdessä tai ovat jonkinlaisessa yhteydessä siten, että niiden voidaan ajatella muodostavan eriteltävissä olevan kokonaisuuden.
  - Tarvitaan kuvaus järjestelmään liittyvistä oliontta, ilmiöstä ja toisaalta myös rajoituksista.
  - Lisäksi tutkimusongelman holistinen käsittely on tilastollisen tutkimuksen kannalta tärkeää: ilmiöön liittyvien tärkeiden ominaisuuksien unohtuminen tarkastelusta saattaa johtaa esimerkiksi puuttuvan muuttujan harhaan!
- Tilastolliset menetelmät auttavat tutkijaa vastaamaan kysymyksiin siitä, mitkä tilastolliset muuttujat ovat tutkimuskysymyksen kannalta oleellisia.
  - Varsinkin nykypäivänä kun datan määrä kasvaa alati kiihyvällä tahdilla, olemme ihmiskuntana ahdistavan informaatiotulvan edessä pakkoin aseettomia: mitkä ympäröivistä ilmiöstä liittyvät toisiinsa ja miten?
  - Erityisesti teoreettisen tilastotieteen kentällä on viimeisten vuosikymmenien aikana kehitetty lukuisia edistyksellisiä menetelmiä nk. dimension pienennyksen alalla. Nämä menetelmät pyrkivät löytämään yhdenmukaisuuksia hyvin korkeaulotteisesta aineistosta, eli aineistosta jossa jokaiselta tutkimusyksiköltä mitataan jopa miljoonia eri muuttujia, kuten DNA-tutkimuksessa genomitie-toa. <sup>7</sup>[Tilastotieteessä näitä menetelmiä kutsutaan monimuuttuja-menetelmiksi ja niitä käsitellään tarkemmin kursseilla TILM3704 Monimuuttujamenetelmät sekä TILM3611 Monimuuttujamenetelmien jatkokurssi]
- **Hahmottamisen vaiheet:**

- “Todellisen” järjestelmän operationalisointi kvantitatiiviseksi kuvaukseksi järjestelmästä.
- Tilastollisen mallin ja järjestelmästä mitattavissa olevan aineiston yhteensovittaminen
- Mallin antamien tulosten muotoilu sellaiseen muotoon, että ne auttavat ymmärtämään mitä aineisto kertoo todellisesta ilmiöstä

### **3. Tilastollisen mallin muodostaminen ja siihen perustuva päätteily**

- Muistetaan aiempi George Boxin sitaatti: Kaikki mallit ovat väärää, mutta jotkut ovat käyttökelpoisia.
  - Tilastollinen malli on vain kuvaus aineiston sisältämästä vaihtelusta: se ei ikinä täydellisesti ja tyhjentävästi vastaa aineiston generoinutta prosessia, mutta sitä voidaan silti käyttää kyseisen ilmiön kuvaamiseen.
- Kuinka saada malliin mukaan kaikki ongelmanasettelun kannalta keskeiset tekijät sellaisella tavalla, ettei oletuksiin ja abstraktioihin liittyvä informaation häviäminen kyseenalaista saatavia tuloksia?
  - Tutkimuskysymyksen kohteena olevan ilmiön taustateoria ja aiheen aiemman tutkimuskirjallisuuden hyvä osaaminen auttaa tässä.
- Vaikutusten erittelyminen on vaikeata, mutta tilastollinen malli on yksi tapa ajatella, kuinka erittely voidaan tehdä. Esimerkkinä tällaisesta mallista mm. edellä käsitelty yksinkertainen lineaarinen regressiomalli.

### **4. Synteesi**

- Tilastollisia tarkasteluja tehdään, koska substanssitetous ei aina riitä haluttuun käyttöön. Yhdistämällä tilastotieteen keinoja sekä substanssitetoutta saadaan ongelma ratkaistua vakuuttavalla ja perustellulla tavalla.
- Tilastollisen (soveltavan) tutkimuksen tavoitteena on tuottaa substanssitetoon perustuen ja tilastotieteen menetelmiä hyödyntäen uutta tietoa: lopputulos on menetelmä- ja substanssiosaamisen synteesi, joka tuottaa uutta substanssitetoutta (sekä joskus myös uusia ongelmia teoreettisen tilastotieteen menetelmäkehitykselle).
- **Jokaisen tutkijan tulisi olla tilastotieteilijä ja jokaisen tilastotieteilijän tutkija. Järkevä yhteistyö!**

### **5. Muita osatekijöitä:**

- Rikas mielikuvitus. Ilman mielikuvitusta uusia yhteyksiä ei keksi etsiä.
- Kriittinen ajattelu: Miksi tämä olisi nyt se oikea vastaus?

## 9.2 Tutkimusprosessi

- Soveltavassa tilastotieteessä tutkimusongelman asettelulla on erityisen tärkeää rooli.<sup>1</sup>
- Tutkimusta ei yleensä ole mahdollista jakaa täysin selvästi erillisin ja ajalliseksi toisiaan seuraaviin vaiheisiin.
  - Tutkimusprosessin vaiheet toistuvat vuorotellen ja limittäin, sillä tutkimuksen aikana tehdyt havainnot muokkaavat tutkimuksen kulkua.
  - Tutkimuksen tekeminen vaikuttaa lopulta saataviin johtopäätelmiin. Aineiston ja itse ilmiön tuntemus kasvaa tutkimuksen kuluessa.
  - Päättelmiin tieteellisyyden (periaatteellinen) tarkistusmahdollisuus, ja nykyään yhä useammin jo toistettavuus, on tärkeää.
- Usein saattaa kuitenkin olla järkevää jäsentää tutkimuksessa kohdattavia tehtäviä ja vaiheita sekä niiden välisiä suhteita osana tutkimusprosessia.
  - Tutkimuksen lähtökohtana on jokin ongelma, johon tutkimuksen avulla etsitään vastausta.
  - Ilmiön ymmärtäminen:
  - Tieto ei voi ylittää historiallisia rajojaan, joten tieteelliset teoriat ovat vain loogisia apuvälineitä, joita voidaan käyttää ilmiön tutkimuksen välineenä tai keinona sillä ehdolla, että sekä ilmiö että teoria asemoihin ja tulkitaan suhteessa vallitseviin olosuhteisiin ja tieteelliseen keskusteluun.
- Määritelmät:
  - Ilmiötä ei voida tutkia sellaisenaan, vaan vain niiden ilmentymien kautta käsitteiden avulla
  - Tutkimus edellyttää arkikieletä täsmällisempää kommunikaatiota, joten ongelmaan liittyvien käsitteiden huolellinen määritteleminen ja erittely on tarpeellista.
  - Määritelmät eivät korvaa empiiristä tietoa, mutta ne vaikuttavat tiedon järjestymiseen ja sen perusteella tehtävien päättelmiin tekemiseen.
- Havaittava tieto
  - Yleensä ajatellaan, että todellisuudesta saadaan tietoa tavalla taikka toisella havaintoja tekemällä.
  - Havaittava tieto ei mitenkään pysty kattamaan kaikkea tutkimuskohteeseen liittyvää ja toisaalta ymmärtämiseen tarvittava havaintomaailman hahmotus tuottaa ideologisesti ja historiallisesti sitoutuneita yksinkertaistavia sekä luonteeltaan usein hyvin teoreettisia abstraktioita.
- Operationalisointi: Siirrytään teoriasta empiriaan

<sup>1</sup>Yksi soveltavan tilastotieteen osa-alue onkin TILM3579 Kokeiden suunnittelu ja analyysi!

- Havainnoiminen ja mittaaminen joudutaan suhteuttaamaan valitseen käsitejärjestelmään.
- Joudutaan tekemään kompromisseja mittauksen eksaktisuus- ja systemaattisuusvaatimusten ja arkielen monimerkityksellisyyden välillä.
- On operationalisoitava tutkimusasetelma sellaiseksi, että tutkittavasta ilmiöstä pystytään tuottamaan ongelmaratkaisun kannalta tarkoituksenmukaista tietoa.
- Aineiston käsittely on tavallaan operationalisoinnin II vaihe. Tiedon (aineiston) muuttaminen hyödylliseksi.
- Näkökulman kiinnittäminen:
  - \* Operationalisoinnin avulla siirtyää teorian tasolta empirian tasolle ja samalla tulee määritellyksi näkökulma, josta ongelmaa tarkastellaan.
  - \* Käsitteet ja niiden yhteyksistä esitettävät näkemykset voivat vaihtua tutkimuksen kuluessa, kunnes lopulta saavutetaan käsitteiden kylläännytymispiste
- Numeerinen mittaus
  - \* Numeerisen mittauksen onnistumiseksi käsitteen muotoilu on kiinnitettävä mittariksi.
  - \* Numeeristen mittaustenkin tulkita edellyttää, että niitä on tulittava siinä kontekstissa, josta ne ovat peräisin.
  - \* On esim. mahdollista, että esitetty kysymys ei vältämättä vastaa tutkimuskohteteen ominaisuuksia.
- Aineisto
  - Aineisto edustaa tutkimuksessa empiiristä maailmaa ja se valitaan ongelmanasettelun perusteella
  - Tarvitaan systemaattinen aineisto, jonka avulla on mahdollista vastata tutkimuskysymyksiin.
  - Aineiston tuottamiseen liittyy useita valintoja, jotka implisiittisesti määräväät myös mahdolliset analyysimenetelmät.
  - Aineiston esikäsittely:
    - \* Aineisto ei ole keräämiseen jälkeen yleensä koskaan suoraan käytettävissä vaan vaatii erinäistä käsittelyä
    - \* Esikäsittely on operationalisoinnin II vaihe, jossa aikaisemmin tehtyjen valintojen aineistossa esiintyvät ilmentyvät sovitetaan vastaamaan ongelmankäsittelyä.
- Analyysi ja tulkinta
  - Analyysivaiheessa sopivasti käsitelty aineisto ja ongelmia pyritään sovittamaan yhteen siten, että ongelmaan saataisiin perusteltu ratkaisu (selitys ja lopulta tulkinta).
  - Keskeistä on, että tehtävät oletukset sisältävät ongelmanratkaisun kannalta keskeiset tekijät sellaisella tavalla, ettei oletuksiin liittyvä informaation häviäminen kyseenalaista saatavia tuloksia.

## 136 LUKU 9. TILASTOTIETEEN ROOLI UUDEN TIEDON TUOTTAMISESSA

- Analyysien tulokset on tulkittava eli käännettävä ne takaisin empirian kieletä teorian kielelle. Tavoitteena on siis substanssitetoutseen perustuen tuottaa uutta tietoa siten, että se lisää myös substanssitetoutta
  - Tulkinnan voi ajatella olevan operationalisoinnin käänteistapahtuma: Tutkimuksen läpiviennin sekä tulkinnan kannalta onnistunut operationalointi ovat loppujen lopuksi yksi ja sama asia.
- Raportointi
    - Parhaimmillaan tutkimusraportti on vakuuttava, ja periaatteessa (ja toivottavasti) toiston mahdollistava, kuvaus tutkimusprosessin kaikesta vaiheesta, jolloin tutkija voi itse päätää haluaako uskoa saatuihin tuloksiin vai ei.
    - Keskeistä on tuoda esille, mitä uutta kyseessä oleva tutkimus on paljastunut ilmiöstä ja suhteuttaa se olemassa olevaan tietoon.
    - Tulosten perustelu: Tutkimuksen pätevyyttä ja yleistävyyttä ja analyysin arviontavuutta ja uskottavuutta tulisi pohtia raportissa. Tutkimuksen kuluessa tehdyt valinnat tulisi perustella tiedostaen mukaan myös omat arvopainotteiset valinnat (ja ehkä oletuksetkin).

**Esimerkkejä tilastollisista tutkimusaselmista:** Näitä käsitellään vielä tarkemmin myöhemmin luvussa 10.

- **Kyselytutkimukset**

- Päättöksentekijät ja tiedotusvälineet kartoittavat säännöllisin välein suomalaisten mielipiteet erilaisista yhteiskuntata koskevista kysymyksistä.
- Esimerkkejä:
  - \* Miten suomalaiset suhtautuvat NATO-jäsenyyteen?
  - \* Miten suomalaiset suhtautuvat ydinvoiman lisärakentamiseen (osana vihreää siirtymää)?
  - \* Mitkä ovat poliittisten puolueiden kannatusosuudet?
- Mielipiteet selvitetään kyselytutkimuksilla, joiden kohteeksi poimitaan tyypillisesti esim. noin 1000-2000 suomalaista.
- Kyselytutkimuksen tavoitteena on tehdä kyselyn tulosten perusteella johtopäätöksiä mielipiteiden jakautumisesta kaikkien suomalaisten joukossa.
- Miten 1000-2000 suomalaiseen kohdistetun kyselyn tulokset voidaan yleistää koskemaan kaikkia suomalaisia?
  - \* Kyselyn tulokset voidaan yleistää, jos kyselyn kohteiksi poimitujen suomalaisten joukko muodostaa edustavan pienoiskuvan Suomen kansasta (onnistuneen otannan idea)

- \* Pienoiskuva on edustava, jos mielipiteet jakautuvat kyselyn kohteiksi poimittujen joukossa samalla tavalla kuin kaikkien suomalaisten muodostamassa perusjoukossa
- \* Kyselyn kohteiden poiminta arpomalla on ainoa menetelmä, joka mahdollistaa edustavan pienoiskuvan saamisen
- \* Kyselyn kohteiden poimintaa kaikkien suomalaisten muodostamasta perusjoukosta arpomalla voidaan nähdä satunnaisotantana ja tutkimuksen kotheeksi poimittua perusjoukon osa on tässä tapauksessa (satunnais)otos
- Arvonnan käyttö kyselyn kohteiden poiminnassa merkitsee sitä, että kyselyn tulokset ovat satunnaisia seuraavassa mielessä: Jos arvontaa toistettaisiin, kysely tuottaisi (suurella todennäköisyydellä) joka keran (ainakin jonkin verran) erilaiset tulokset, koska eri arvontoissa kyselyyn poimittaisiin (suurella todennäköisyydellä) eri henkilöt.
- Kysymyksiä:
  - \* Miten yhdestä otoksesta saadut ja satunnaiset kyselytulokset voidaan yleistää koskemaan koko sitä perusjoukkoa, josta otos poimitaan?
  - \* Miten luotettava tällainen yleistys on?
- Vastauksia:
  - \* Jos kyselyn kohteiden poiminnassa on käytetty satunnaisotantaa, kyselyn tuloksiin sisältyvälle epävarmuudelle ja satunnaisuudelle voidaan muodostaa tilastollinen malli, joka mahdollistaa sekä kyselyn tulosten yleistämisen että yleistyksen luotettavuuden arvioimisen.
  - \* Yleistyksen luotettavuutta ei pystytä arvioimaan, ellei otoksen poiminnassa ole käytetty satunnaisotantaa.
  - \* Kyselytutkimusten suunnittelussa, toteutuksessa ja tulosten analysoinnissa sovelletaan mm. seuraavia tilastollisia menetelmiä: otanta, estimointi ja testaus.
- Laadunvalvonta
  - Tehdas valmistaa korkealuokkaisia sulkimia kameroihin. Tehdas pyrkii siihen, että yli 90% sulkimista kestää vähintään 100 000 kameran laukaisua.
  - Sulkimien laadun valvonta on toteutettu seuraavalla tavalla:
    - \* (i) Tuotantolinjalta poimitaan arpomalla joukko sulkimia rasiuskokeeseen.
    - \* (ii) Rasituskokeessa määräätään vähintään 100 000 laukaisua keskätienvi sulkimien suhteellinen osuus.

## 138 LUKU 9. TILASTOTIETEEN ROOLI UUDEN TIEDON TUOTTAMISESSA

- Kokeen tavoitteena on tehdä kokeen tulosten perusteella yleisiä joh-topäätöksiä sulkimien kestävyydestä.
- Miten vain osaan sulkimista kohdistetun rasituskokeen tulokset voidaan yleistää koskemaan kaikkia sulkimia?
  - \* Kokeen tulokset voidaan yleistää, jos rasituskokeen kohteiksi poimittujen sulkimien joukko muodostaa edustavan pienoiskuvan kaikista valmistetuista sulkimista.
  - \* Pienoiskuva on edustava, jos sulkimien kesto jakautuu rasituskokeeseen poimittujen sulkimien joukossa samalla tavalla kuin kaikkien valmistettujen sulkimien muodostamassa perusjouksa.
  - \* Rasituskokeen kohteiden poiminta arpomalla on ainoa menetelmä, joka mahdollistaa edustavan pienoiskuvan saamisen.
  - \* Rasituskokeen kohteiden poiminta kaikkien valmistettujen sulkimien muodostamasta perusjoukosta arpomalla merkitsee satunnaisotannan soveltamista ja tutkimuksen kohteeksi poimittua perusjoukon osa toimii muodostettavana (satunnais)otokseksena.
- Arvonnan käyttö rasituskokeen kohteiden poiminnassa merkitsee sitä, että koetulokset ovat satunnaisia seuraavassa mielessä: Jos arvontaa toistetaisiin, kokeesta saataisiin (suurella todennäköisyydellä) joka kerran (ainakin jonkin verran) erilaiset tulokset, koska eri arvontoissa kokeeseen poimittaisiin (suurella todennäköisyydellä) eri sulkimet.
- Kysymyksiä:
  - \* Miten yhdestä kokeesta saadut ja satunnaiset koetulokset voidaan yleistää koskemaan kaikkia sulkimia?
  - \* Miten luotettava tällainen yleistys on?
- Vastauksia:
  - \* Jos rasituskokeen kohteiden poiminnassa on käytetty satunnaisotantaa, kokeen tuloksiin sisältyvälle epävarmuudelle ja satunnaisuudelle voidaan muodostaa tilastollinen malli, joka mahdollistaa sekä koetulosten yleistämisen että yleistyksen luotettavuuden arvioimisen.
  - \* Yleistyksen luotettavuutta ei pystyä arvioimaan, ellei kokeen kohteiden poiminnassa ole käytetty satunnaisotantaa.
  - \* Kokeen suunnittelussa, toteutuksessa ja tulosten analysoinnissa sovelletaan mm. seuraavia tilastollisia menetelmiä: koesuunnittelu, otanta, estimointi ja testaus.

## Luku 10

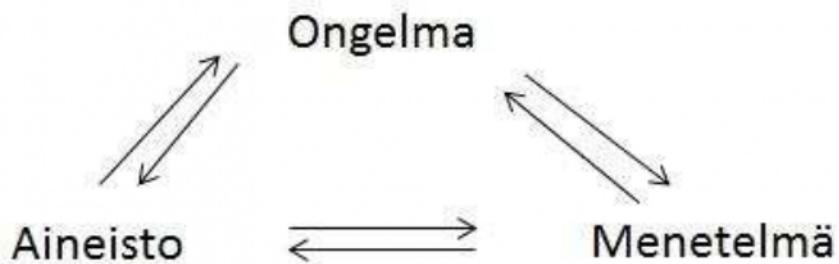
# Aineisto- ja tutkimustyyppit ja koeasetelmat

Tässä luvussa käsitellään erilaisia tapoja toteuttaa tilastollista tutkimusta. Emppiirisen tutkimuksen lähtökohtana on aina tutkimusongelma, joka sisältää kysymyksen tai kysymyksiä, joihin tutkimuksella haetaan vastauksia. Tilastotieteen näkökulmasta tutkimusongelman keskiössä on kuitenkin aineisto ja se, miten käytettäväissä olevasta aineistosta saadaan vastauksia tutkimuskysymyksiin - täten tarkastelemme tässä luvussa myös tutkimuksenteon käytäntöä käsittelemällä erilaisia aineistotyyppejä. Käymme läpi eri alojen ja tutkimusongelmien käytännön tutkimustyyssä kohdattavia aineistoja ja erittelemme pidemmälle eri tutkimuskysymysten käytännön haasteita aineistojen osalta sekä sitä, minkälaisia ongelmia erilaisiin tutkimuskysymyksiin käytännössä liittyy ja miten eri tutkimusasetelmat pyrkivät niitä ratkomaan.

Aineistotarpeen ja sen analysoinnin lähtökohdat määritetään tutkimusongelma. Tutkimus voi olla esimerkiksi kuvalevaa, vertailevaa, selittävää tai kokeellista ja aineistolle sekä menetelmille asetetaan kussakin tapauksessa erilaiset vaatimukset ja odotukset. Erilaisiin tutkimuskysymyksiin ja niihin vastausta etsivisiin **koeasetelmiin** liittyvien esimerkkien avulla pyrimme löytämään vastauksia esimeriksi seuraaviin kysymyksiin:

- Miten tilastotiede liittyy tiedon keruuseen?
- Miten aineisto generoituu?
- Millaisiin kysymyksiin saadaan kussakin asetelmassa vastauksia?
- Tarkemmat asetelmiin ja analyyseihin liittyvät yksityiskohdat käsitellään Kokeen suunnittelua ja analyysi -kurssilla

Luvussa käsiteltävät asiat kuuluvat tilastotieteelle ominaisesti kvantitatiivisen tutkimussuuntaukseen alaisuuteen (ks. luku 3). Luvussa esiteltävät karkeat



Kuva 10.1: Tutkimusasetelma

tutkimustyyppien- strategioiden ja aineistojen jaot ovat vain yksi jaottelutapa ja todennäköisesti poikkeaa eri oppikirjoissa ja lähteissä esitetyistä.

## 10.1 Tutkimustyypit

Tarkastellaan ensin erilaisia tutkimustyyppejä yleisellä tasolla. Erilaiset tutkimukset voidaan karkeasti jakaa neljään eri luokkaan: **kuvailleva**, **vertaileva**, **kokeellinen** ja **havainnoiva** tutkimus.

- **Kuvailleva tutkimus**

- Tarkoituksena on kuvata jonkin ilmiön, tilanteen tai tapahtuman luonnetta, yleisyyttä, historiallista kehitystä tai muita tunnuspiirteitä mahdollisimman todennäköisesti ja tarkasti.
- Keskeistä **tiedon lisääminen** ja pyrkimys vastata kysymyksiin **mitä, millainen tai miten**.
  - \* Yleisesti ottaen kuvailleva tilastollinen tutkimus perustuu aineistosta lasketulle tunnusluvuille, jotka kuvaavat aineiston ominaisuuksia. Esimerkkinä toimivat keskiarvon lisäksi sen kaltaiset keskimääräistä havaintoa mittavaat suuret kuten mediaani ja moodi tai vaihtelua kuvaavat eri muuttujien vaihteluvälit ja keskihajonnat.
  - Saadakseen luotettavia tunnuslukuja, tulee otoksen olla edustava ja havaintojen luotettavia ja päteviä eli saatujen mittausten pitää kuvata kohteena olevaa ilmiötä ilman virheitä.
  - Kuvilevassa tutkimuksessa ei tutkita muuttujien väisiä yhteyksiä tai riippuvuuksia eikä täten yleensä tehdä jakaoa selittäviin ja selitetäviin muuttujuihin vaan muuttujat ovat asetelmallisesti samantasoisia.
    - \* Vastaavasti kuvilevassa tutkimuksessa ei välittämättä testata hypoteeseja, ei tehdä ennusteita, ei anneta selityksiä tai pohdita seuraauksia: kyseessä on vain aineiston kuvailua ilman sen

merkityksellisempää sisältöä kuten havaintojen taustalla olevien ilmiöiden tutkimista tai perusjoukon ominaisuuksien päättelyä otoksen perusteella.

- **Vertaileva tutkimus**

- Vertaileva tutkimus voidaan jakaa kahteen luokkaan
  1. Ryhmäeroja selittävään tutkimukseen
  2. Korrelatiotutkimukseen
- **Ryhmäeroja selittävässä tutkimuksessa** pyritään selvittämään, mitkä tekijät liittyvät tutkittaviin ilmiöihin, jotka aiheuttavat ryhmissä ilmeneviä eroja.
- **Korrelatiotutkimuksissa** pyritään löytämään ilmiöiden välistä yhteyksiä tutkimalla kohdejoukkoa kokonaisuutena, jolloin mitattavien muuttujien joukkoon otetaan selittäviä muuttuja.
- Selittäviä muuttuja hyödynnetään molemmissa luokissa. Niiden avulla pyritään löytämään yhteyksiä verrattavien kohteiden välillä ja niiden voidaan ajatella olevan myös mahdollisia syitä selittäville muuttujille, seuraauksille.
  - \* Syy-seuraussuhteita ei kuitenkaan vertailevassa tutkimuksessa pohdita, ts. vertaileva tutkimus ei ole suoranaisesti kiinnostunut kohtena olevien ilmiöiden/ryhmienvertailussa löydettyjen erojen syistä vaan mielenkiinnon kohtena on kys. erot itsessään.
- Vertailevaa tutkimusta tehdessä on tarpeen pohtia:
  - \* Miksi jotakin tutkimuskohdetta vertaillaan eli mitä tutkimuskohteesta halutaan nimenomaan saada selville.
  - \* Mitkä ja minkälaisia tilastoyksiköitä vertailuun kannattaa ottaa mukaan, jotta tutkimuksen tavoitteet saavutetaan.
  - \* Tyypillistä se, että kontrolli on puutteellista ja ns. väliin tulevia muuttuja ei voida aina eliminoida.
  - \* Tutkimuksessa on hyväksyttävä myös muuttuihin liittyvä luonnollinen vaihtelu.

- **Kokeellinen tutkimus**

- Tarkastellaan syy-seuraussuhteita sellaisissa olosuhteissa, joissa tutkija pystyy kontrolloimaan tutkimusyksikköihin vaikuttavia tekijöitä, eli nk. **"käsittelytekijöitä"**.
- Tavallisesti kokeellisella tutkimuksella viitataan sellaiseen tutkimukseen, jossa aineiston on kerätty valvotussa ja kontrolloidussa ympäristössä, kuten laboratoriossa tai sairaalan koehuoneissa, jotta mittaukset ja käsittelytekijät on tutkimuksen tekijän puolesta kontrolloitu ja täten halutunlaisia.

- \* Tutkimusasetelman kontrollointi vähentää mittauksiin ja käsitteilytöihin liittyvien virhelähteiden mahdollisuksia ja täten jättää vähemmän sijaa epäilyksille.
- \* Lisäksi tutkimuksen toistettavuus ja objektiivisuus paranevat, kun koejärjestelyt tehdään tarkasti ja huolellisesti.
- Kokeelliset tutkimukset tuottavat yleensä nopeammin riittävään näyttöön perustuvaa evidenssiä kuin havainnoivat tutkimukset.
- Kokeellinen tutkimusasetelma ei kuitenkaan ole mahdollinen kaikissa tilanteissa.
  - \* Esimerkiksi erilaisten politiikkatoimien arvioimisessa olisi hyödyllistä, mikäli se voitaisiin satunnaisesti kohdistaa esimerkiksi vain osaan kansasta tai kunnista. Tällaisten kokeilujen ehdotukset ovat kuitenkin usein kaatuneet joko perustuslaillisiin ongelmiin tasavertaisesta kohtelusta tai muihin lainsääädännölliin ongelmii tai niitä ei ole toteutettu riittävän hyvin, jotta asetelma riittäisi kokeelliseen analyysiin.<sup>1</sup>
- Kontrolloitujen kokeiden yleisenä kritiikkinä ja heikkoutena voidaan kuitenkin pitää niiden vähäistä yleistäväyyttä: liian pitkälle kontrolloidut ja pelkistetyt koeolosuhteet eivät toimi kaikkien tutkimuskysymysten kannalta yleistäväyyden osalta.
  - \* Ihmiset käyttäytyvät eri tavalla laboratorio-olosuhteissa kuin normaalissa ympäristössä!

#### Esimerkki: kasvien kasvatus eri hiilidioksidipitoisuksissa

- Hiilidioksidipitoisuuden kasvu tehostaa kasvien yhteyttämistä
- Kasvit eroavat toisistaan siinä, millä tavalla ne sitovat hiilidioksidia ilmasta yhteyttämistä varten → muutokset vaikuttavat eri tavalla eri kasveihin
- Vaikuttaako ilmastonmuutos sadonmuodostukseen? Onko vaikutus suurempi joillain tietyillä kasveilla?

#### Esimerkki: Lääketieteelliset kokeet

- Erään tappavan taudin hoitoon on kehitetty uusi lääke, jonkaトイ votaan parantavan enemmän potilaita kuin kauan käytössä ollut vanha lääke. Miten saadaan varmuus siitä, että uusi lääke on parempi kuin vanha lääke?
- Paranemistulosten vertailemiseksi järjestetään tilastollinen koe:
  - (i) Jaetaan joukko potilaita arpomalla kahteen ryhmään:

<sup>1</sup> Esimerkki: Jeremias Nieminen avaa vuonna 2020 alkaneesta työllisyyden kuntakokeilusta koeasetelman tärkeydestä politiikkatoimien arvioinnissa.

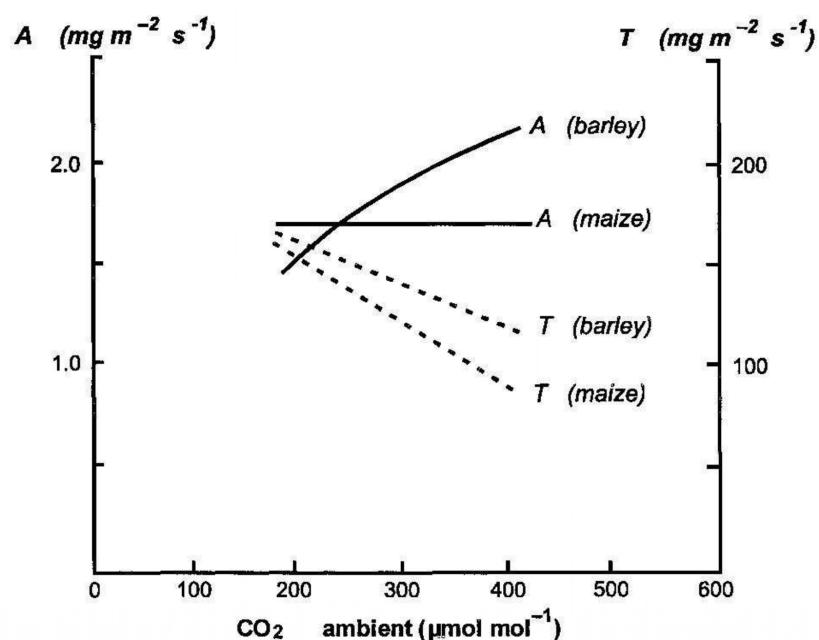


Figure 3: Measured canopy responses to ambient  $\text{CO}_2$  of photosynthesis (A) and transpiration (T) in barley (C3) and maize (C4) (from: Goudriaan and Unsworth, 1990).

Kuva 10.2: Hiilidioksidipitoisuuden kasvun vaikutus satomääriin.

\* Ryhmälle 1 annetaan uutta lääkettä.  
\* Ryhmälle 2 annetaan vanhaa lääkettä.

– (ii) Verrataan parantuneiden suhteellisia osuuksia ryhmissä 1 ja 2.

- Kokeen tavoitteena on tehdä kokeen tulosten perusteella yleisiä johtopäätöksiä uuden lääkkeen tehokkuudesta. Miten yhdestä kokeesta saadut tulokset voidaan yleistää koskemaan kaikkia tautia sairastavia potilaita?
  - Kokeen tulokset voidaan yleistää, jos kokeessa uutta ja vanhaa lääkettä saavien potilaiden ryhmät ovat samankaltaisia kaikissa muissa suhteissa paitsi siinä, että niihin kohdistetaan kokeessa erilainen käsitteily.
    - \* Tällöin mahdolliset erot parantuneiden suhteellisissa osuuksissa on oltava seurausta erilaisista käsitteilyistä.
    - \* Kokeen kohteiden jakaminen ryhmiin arpomalla on ainoa menetelmä, joka mahdolistaan samankaltaisten ryhmien saamisen.
    - \* Kokeen kohteiden jakamista erilaisen käsitteelyn kohteiksi joutuviin ryhmiin arpomalla kutsutaan tilastotieteessä **satunnaistamiseksi**.
  - Arvonnan käyttö ryhmiin jaossa merkitsee sitä, että koetulokset ovat satunnaisia seuraavassa mielessä: Jos arvontaa toistettaisiin, kokeesta saataisiin (suurella todennäköisyydellä) erilaiset ryhmäjat.
- Kysymyksiä:
  - Miten yhdestä kokeesta saadut ja satunnaiset koetulokset voidaan yleistää koskemaan kaikkia ko. tautia sairastavia potilaita?
  - Miten luotettava tällainen yleistys on?
- Vastauksia:
  - Jos potilaiden jaossa ryhmiin on käytetty satunnaistamista, kokeen tuloksiin sisältyvälle epävarmuudelle ja satunnaisuudelle voidaan muodostaa tilastollinen malli, joka mahdolistaan sekä koetulosten yleistämisen että yleistyksen luotettavuuden arvioimisen.
  - Yleistyksen luotettavuutta ei pystytä arvioimaan, ellei ryhmiin jaossa ole käytetty satunnaistamista.
  - Tilastollisen kokeen suunnittelussa, toteutuksessa ja tulosten analysoinnissa sovelletaan mm. seuraavia tilastollisia menetelmiä: koe-suunnittelu, estimointi ja testaus.

- **Havainnoiva tutkimus**

- Kuten edellä mainittiin, kokeellisia tutkimusasetelmia ei useinkaan ole mahdollista järjestää. Tällaisia kysymyksiä voidaan kuitenkin tutkia havainnoivassa tutkimuksessa, jossa syy-seuraussuhteita tarkastellaan tilanteissa, joissa tutkijalla ei ole välttämättä mitään kontrollia (tai syytä sille) tutkimusyksikköihin tai heihin vaikuttaviin muuttuuihin (käsittelytekijöihin).
  - \* Esimerkiksi tutkimusasetelmat, joissa tutkimuksen kohteena olevia yksiköitä (esim. ihmiset, kunnat, valtiot) ei voida satunnaisataa kuuluvaksi osaksi joukkoa, joka altistetaan jollekin käsittelylle.
- Tällöin tutkijan on tydyttää havainnoimaan sitä mitä tapahtuu luonnostaan tietyssä (mahdollisesti satunnaisesti poimitussa) tutkimusjoukossa tietyssä tilanteessa.
- Havainnoivan tutkimuksen aineistoa voidaan analysoida samoin menetelmin kuin kokeellisen tutkimuksenkin, mutta mitattujen tekijöiden vaikutusta ei voida erottaa kokonaisuudesta samalla tarkkuudella kuin kokeellisessa tutkimuksessa
- Havainnoivan tutkimuksen tilastollinen teoria muodostuu periaatteisesti ja menetelmistä, joiden avulla aineiston tuottaman evidenssin painoarvoa voidaan arvioida mahdollisimman ”puhtaasti”
- **Havainnoivan tutkimuksen edut**
  - \* Saadaan välitöntä ja suoraa tietoa yksilöiden, ryhmien ja organisaatioiden toiminnasta ja käyttäytymisestä.
  - \* Tutkija voi havainnoida tutkittavia luonnollisessa ympäristössä.
  - \* Sopii sekä määrellisen että laadullisen aineiston hankkimiseen.
  - \* Erinomainen menetelmä muun muassa vuorovaikutuksen tutkimisessa, ja silloin kun tilanteet ovat vaikeasti ennakoitavia ja nopeasti muuttuvia.
  - \* Sopii myös silloin, kun tutkittavilla on kielessiä vaikeuksia (kuten lapset) tai kun halutaan saada selville sellaista tietoa, jota tutkittavat eivät halua suoraan kertoa tutkijalle.
- **Havainnoivan tutkimuksen haitat**
  - \* tutkija saattaa häiritä tilannetta tai muuttaa sen kulkua.
  - \* Tutkija saattaa sitoutua emotionaalisesti tutkittavaan ryhmään tai tilanteeseen.

**Esimerkki: raskauden keskeytyksen ja rintasyövän välinen kausaalihyhteys**

- Kokeellinen asetelma: Poimitaan satunnaiseksi  $n$  kappaletta raskaana olevia naisia ja heistä  $n_1$  kappaletta satunnaistetaan käsitellyryhmään (raskauden keskeytys) ja  $n_2$  kontrolliryhmään. Kaikki naiset kävät muutaman seuraavan vuoden ajan syöpäseulonnoissa.
- Kokeelliseen asetelma ei selvästiikään ole eettisistä syistä mahdollinen, eikä sitä olisi mahdollista suorittaa sokkoutettuna kokeena
- Aiheesta julkaistut tutkimukset aloittavat yleensä naisista, joille on jo tehty raskauden keskeytys
- Käsittelyryhmään kuuluminen ei siis ole tutkijan kontrollissa

**Esimerkki: Lääkityksen aiheuttama harvinainen sivuvaikutus**

- Harvinaisen ilmiön tarkastelu satunnaistetulla kokeella on epäkäytännöllistä, sillä saattaa olla, että isossakaan tutkimusjoukossa sivuvaikutusta ei esiinny yhdelläkään tutkittavalla
- Havainnoiva tutkimus aloittaisi tässä tapauksessa etsimällä ensin sivuvaikutuksesta kärsivät potilaat ja sen jälkeen selvittäisi ketkä heistä ovat saaneet kyseistä lääkettä (ja saaneet sivuoireet lääkitynksen aloittamisen jälkeen)

## 10.2 Tutkimusstrategiat

Erilaiset tutkimusasetelmat voidaan jakaa edelleen kahteen **tutkimusstrategiaan** sen mukaan, miten niissä ryhmitellään tilastoysiksiötä: **poikkileikkaus- ja pitkittäistutkimuksiin**. Tilastoysikköjen erilainen ryhmittely tuottaa erilaisia aineistotyyppejä, jotka voidaan jaotella karkeasti kolmeen eri typpiin - Poikkileikkausaineistot: havaintoaineisto kattaa yhden ajankohdan ja mahdollisesti useita tilastollisia muuttuja - Aikasarja-aineistot: havaintoaineisto kattaa vain yhden tilastollisen muuttujan mitattuna useana ajanhetkenä - Paneeliaineistot: havaintoaineisto kattaa mahdollisesti useita tilastollisia muuttuja mitattuna useana ajanhetkenä

Eri tutkimusstrategiat hyödyntävät eri aineistotyyppejä sen mukaan, miten ne sopivat tutkimuskysymykseen ja valittuun menetelmään. Tarkastellaan seuraavasti erilaisia strategioita.

vaksi mitä em. kaksi tutkimuststrategiaa tarkoittavat, miten ne eroavat ja min-kälaisia tutkimustyyppejä-, asetelmia- ja aineistoja kumpaankin kuuluu.

### 10.2.1 Poikittaistutkimus tai poikkileikkaustutkimus

- Poikittaistutkimukseksi kutsutaan tutkimusstrategiaa, jossa tarkoitukse-na on tutkia kohdetta tai ilmiötä laaja-alaisesti tietynä ajankohtana käyt-täen poikkileikkausaineistoja.
  - Voidaan tarkastella useita ryhmiä, joissa on esimerkiksi eri-ikäisiä henkilöitä ja ryhmistä saatua tietoa vertailaan toisiinsa.
  - Voidaan käyttää kuvailemaan riskisuheteita (odds ratio) tai kuvaile-maan tiettyyn populaation osaan kohdistuvaa ilmiötä tai riskiä (esi-merkiksi sydän- ja verisuonitaudit).
    - \* Esimerkiksi tutkittaessa sydän- ja verisuonitauteja binäärисellä vastemuuttujalla käytäen aineistoa, joka koostuu eri ikäisistä ja kuntoisista ihmisiästä voidaan arvioida iän ja muiden muuttujien vaikuttuksia sydän- ja verisuonitauteihin sairastumisen riskiteki-jöinä.
  - Poikittaistutkimuksessa ei saada tietoa tilastoysikön mielenkiinnon kohteena olevien muuttujien arvojen muutoksesta yli ajan mutta tut-kimuksessa voidaan kuitenkin kerätä tietoa menneisyyteen liittyen.
  - Eri ikäryhmiä vertailtaessa ongelmana on myös niin sanottu kohort-tivaikutus: tietynä aikana syntyneiden, eli tietyn kohortin, elinolo-suhteet saattavat olla täysin erilaiset kuin jonakin toisena aikana syntyneiden, minkä vuoksi ikäryhmiä väliset erot saattavat johtua esimerkiksi yhteiskunnallisia olosuhteista.
  - Poikittaistutkimukseen osallistutaan vain yhden kerran, jolloin tietoa saadaan kerralla paljon.
    - \* Tämä on kuitenkin usein työlästä ja suuren poikkileikkausaines-ton kerääminen voi olla kallista.
    - \* Poikittaistutkimuksessa hyödynnetäänkin usein rutiinitoimenpi-teinä kerättyjä aineistoja (esimerkiksi tietyn ikävuoden terveys-tarkastuksista)
    - \* Näin voidaan selvittää korrelaatioita ilmiöiden välillä (esimerkiksi alkoholin käyttö ja maksakirroosi) ja siten luoda hypoteeseja tarkemmille jatkotutkimuksille
    - \* Tällöin on kuitenkin taas vaara sekoittavista tekijöistä, jos ai-neistoa ei ole kerätty varta vasten tätä tarkoitusta varten

## Beer and obesity: a cross-sectional study

M Bobak<sup>1\*</sup>, Z Skodova<sup>2</sup> and M Marmot<sup>1</sup>

<sup>1</sup>Department of Epidemiology and Public Health, International Centre for Health and Society, University College London, UK; and  
<sup>2</sup>Department of Preventive Cardiology, Institute of Clinical and Experimental Medicine, Prague, Czech Republic

**Objective:** There is a common notion that beer drinkers are, on average, more ‘obese’ than either nondrinkers or drinkers of wine or spirits. This is reflected, for example, by the expression ‘beer belly’. However, the few studies on the association between consumption of beer and abdominal obesity produced inconsistent results. We examined the relation between beer intake and waist–hip ratio (WHR) and body mass index (BMI) in a beer-drinking population.

**Design:** A cross-sectional study.

**Settings:** General population of six districts of the Czech Republic.

**Subjects:** A random sample of 1141 men and 1212 women aged 25–64 y (response rate 76%) completed a questionnaire and underwent a short examination in a clinic. Intake of beer, wine and spirits during a typical week, frequency of drinking, and a number of other factors were measured by a questionnaire. The present analyses are based on 891 men and 1098 women who where either nondrinkers or ‘exclusive’ beer drinkers (ie they did not drink any wine or spirits in a typical week).

**Results:** The mean weekly beer intake was 3.1 l in men and 0.3 l in women. In men, beer intake was positively related to WHR in age-adjusted analyses, but the association was attenuated and became nonsignificant after controlling for other risk factors. There appeared to be an interaction with smoking: the relation between beer intake and WHR was seen only among nonsmokers. Beer intake was not related to BMI in men. In women, beer intake was not related to WHR, but there was a weak inverse association with BMI.

**Conclusion:** It is unlikely that beer intake is associated with a largely increased WHR or BMI.  
*European Journal of Clinical Nutrition* (2003) 57, 1250–1253. doi:10.1038/sj.ejcn.1601678

---

**Keywords:** beer; alcohol; obesity; body mass index; waist-hip ratio; epidemiology

Kuva 10.3: Esimerkki poikkileikkaustutkimuksesta

### 10.2.2 Pitkittäistutkimus

- Pitkittäistutkimuksessa seurataan usein samoja tilastoyksiköitä “yli ajan”, eli mittauspisteitä on useita ja pitkältä aikaväliltä.
  - Hyödyntää niin aikasarja- kuin paneelaineistojakin.
  - Yleinen tutkimuskysymys pitkittäistutkimuksessa on jonkin **käsitelyn vaikuttuksen arvointi**. Tällaisia ovat esimerkiksi lääkeaine-tutkimus, poliittisten päätösten arvointi tai markkinointitutkimus.
    - \* Pitkittäistutkimuksessa voidaan siis tarkastella **muutosta** mutta on tärkeää muistaa, että pitkittäistutkimuksen eri mittauskerat eivät ole toisiaan **riippumattomia** ja tämä tulee ottaa tilastollisessa mallissa huomioon!
  - Pitkittäistutkimuksen hyvänen puolena on **ryhmien homogeenisyys**
    - \* Tutkittavan ryhmän henkilöt ovat eläneet saman historiallisen ajan sekä käyneet läpi samat yhteiskunnalliset muutokset, jolloin muutoksen tutkiminen on luotettavaa, sillä tutkimusta vääristää-vät tilastoyksiköiden ominaisuuksista erilliset ympäristön haittamuuttujat ovat kaikille samat.
    - \* Pitkittäistutkimuksen pitkän keston vuoksi tutkittavien määrää kuitenkin yleensä vähenee ja tutkimuksen valmistumisessa kestää kauan, jopa vuosikymmeniä.

## Breathing-Based Meditation Decreases Posttraumatic Stress Disorder Symptoms in U.S. Military Veterans: A Randomized Controlled Longitudinal Study

Emma M. Seppälä,<sup>1</sup> Jack B. Nitschke,<sup>2,3</sup> Dana L. Tudorascu,<sup>4</sup> Andrea Hayes,<sup>5</sup> Michael R. Goldstein,<sup>6</sup>  
Dong T. H. Nguyen,<sup>1</sup> David Perlman,<sup>2,5</sup> and Richard J. Davidson<sup>2,3</sup>

<sup>1</sup>Center for Compassion and Altruism Research and Education, School of Medicine, Stanford University, Stanford, California, USA

<sup>2</sup>Department of Psychology, University of Wisconsin-Madison, Madison, Wisconsin, USA

<sup>3</sup>Department of Psychiatry, University of Wisconsin-Madison, Madison, Wisconsin, USA

<sup>4</sup>Department of Internal Medicine, Biostatistics and Geriatric Psychiatry Neuroimaging Lab, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

<sup>5</sup>Waisman Laboratory for Brain Imaging and Behavior, University of Wisconsin-Madison, Madison, Wisconsin, USA

<sup>6</sup>Department of Psychology, University of Arizona, Tucson, Arizona, USA

Given the limited success of conventional treatments for veterans with posttraumatic stress disorder (PTSD), investigations of alternative approaches are warranted. We examined the effects of a breathing-based meditation intervention, Sudarshan Kriya yoga, on PTSD outcome variables in U.S. male veterans of the Iraq or Afghanistan war. We randomly assigned 21 veterans to an active ( $n = 11$ ) or waitlist control ( $n = 10$ ) group. Laboratory measures of eye-blink startle and respiration rate were obtained before and after the intervention, as were self-report symptom measures; the latter were also obtained 1 month and 1 year later. The active group showed reductions in PTSD scores,  $d = 1.16$ , 95% CI [0.20, 2.04], anxiety symptoms, and respiration rate, but the control group did not. Reductions in startle correlated with reductions in hyperarousal symptoms immediately postintervention ( $r = .93, p < .001$ ) and at 1-year follow-up ( $r = .77, p = .025$ ). This longitudinal intervention study suggests there may be clinical utility for Sudarshan Kriya yoga for PTSD.

Kuva 10.4: Esimerkki pitkittäistutkimuksesta

### Esimerkki: poikittais- ja pitkittäistutkimus epidemiologiassa

- Epäkokeelliset epidemiologiset tutkimukset voivat olla joko poikittäistutkimuksia tai pitkittäistutkimuksia
- Poikittäistutkimus on tiettyyn ajankohtaan rajoittuva tutkimus, jossa mitataan sairauksien vallitsevuutta eli prevalenssia
  - Vallitsevuus eli prevalenssi kuvailee sairauden tai haitan omaavien henkilöiden määrää tarkasteltavasta väestöstä tietynä ajankohdalla
  - Usein mitataan vallitsevuustiheyttä eli sairaiden lukumäärää tietynä ajanhetkenä / väkiluku samana ajankohdalla
- Pitkittäistutkimussa mitataan sairauksien ilmaantuvuutta eli insidenssiä
  - Tutkimuksessa seurataan väestössä ilmaantuvien uusien sairaustabausten lukumäärää tietyn ajanjakson aikana
  - Useimmiten mitataan ilmaantuvuustiheyttä, joka ilmoittaa uusien sairastapausten määrän henkilöaikaa kohden
  - Henkilöaika muodostuu tarkasteltavan henkilöryhmän yhteenlasketusta seuranta-ajasta ennen sairastumista, esimer-

kaksi 100 henkilövuotta muodostuu seurattaessa 100 henkilöä vuoden ajan tai 10 henkilöä 10 vuoden ajan

### 10.2.3 Kohorttitutkimus

- Kohorttitutkimus on altistelähtöinen tapa toteuttaa pitkittäistutkimus.
  - Kohortti on suljettu väestö (syntymäkohortti, tietyn työpaikan työtekijät, yms.), jota tutkimuksessa seurataan ja joka on valittu jonkin yhteisen ominaisuuden perusteella (syntymävuosi, työpaikka, yms.).
    - \* Kohortti voidaan jakaa myös alakohortteihin, mikäli se on alkuperäisestä suuresta kohortista.
    - \* Valitun kohortin tilastoyksiköst pyritään pitämään täysin samana koko tutkimuksen ajan, ts. kohortit ovat kiinteitä.
  - Tutkimuksessa voidaan valita esimerkiksi jollekin käsittelymuuttujalle altistunut ja altistumattomien ryhmä.
    - \* Kohortin seuranta-aikana tutkitaan ryhmien välillä ilmeneviä eroja mielenkiannon kohteena olevassa muuttujassa.
    - \* Näitä voi olla esimerkiksi ryhmien väliset erot sairastuvuudessa, kun käsittely on ollut jokin lääke kuten rokote tms. Vastaava esimerkki olisi em. työllisyyden kuntakokeilun osalta työllisyysaste eri kunnissa ja erojen tutkiminen kuntakokeilun osallistuvien ja osallistumattomien välillä.
  - Kohorttitutkimus voi olla **taannehtiva**: tutkija määrittelee kohortin menneisyydessä, ja seuraa olemassaolevien rekisterien avulla, mitä kohortin jäsenille on tapahtunut myöhemmin.
  - Kohorttitutkimuksessa voidaan yleensä tutkia kerrallaan vain **yhtä altistetta/käsittelyä**, mutta **useita tilastollisia muuttuja**.
  - Tutkimukset saattavat olla hyvin pitkäkestoisia, jos tutkitaan ilmiötä, joka ilmenee vasta pitkä ajan kuluttua altistuksesta (kuten sairaus tai työllisyyden paraneminen)
  - Kohorttitutkimus voi vastata kysymykseen: "Mitkä ilmiöt johtuvat tästä altisteesta?"

#### Esimerkki kohorttitutkimuksesta

Toisen maailmansodan aikana räjäytettiin Japanissa kaksi atomipommia. Tämän traagisen tapahtuman jälkeen tutkijat alkoivat selvittää, mitä terveysvaikutuksia ionisoiva säteily aiheuttaisi altistuneille. Tutki-

mukseissa seurattiin altistuneiden ja altistumattomien sairastumista vuodesta 1945 vuoteen 1970. Tutkimuksen mukaan ionisoiva säteily aiheutti etupäässä monenlaisia kasvaimia; mm. keuhkosyöpää, rintasyöpää ja kilpirauhasen syöpää.



Kuva 10.5: Yhdysvaltain räjäyttämä atomipommi Japanin Hiroshimassa aiheutti mittaamattomia tuhoja.

#### 10.2.4 Tapaus-verrokkitutkimus

- On **retrospektiivinen havainnoiva** pitkittäistutkimusmenetelmä, jossa tutkimukseen valitaan esimerkiksi tutkittavaan sairauteen (tai muulle altisteelle/käsittelylle altistuneita) potilaita (**tapaukset**) ja lisäksi henkilöitä, jotka eivät ole altistuneita tähän sairauteen (**verrokit**) (tai altistuneet altisteelle/käsittelylle).
  - Tavoitteena on tutkia miten tutkimusyksiköt reagoivat altistuttuaan

jollekin altisteelle tai käsittelylle. Soveltuu erityisesti harvinäisten ilmiöiden aiheuttajien selvittämiseen.

- \* Esimerkkinä altistuminen Covid-19 virukselle: retrospektiivisesti (jälkikäteen) voidaan tarkastella viruksen kantajan kanssa samassa tilassa olleita (virukselle altistuneita (virus on altiste)) kysimessä tilassa olleet olisivat tapauksia ja hypoteettinen toinen tila ilman virusta toimisi verrokkina (ts. ei yhtäkään tartuntapausta). Mielenkiinnon kohteena olisi tarkastella kuinka monta henkilöä sai tartunnan (ja minkälaiset olosuhteet olivat).
- \* Toisena esimerkkinä voitaisiin jälleen pitää em. työllisyden kuntakokeilua: ne kunnat jotka (satunnaisesti) valikoituisivat työllisyyskokeiluun tulisivat altistetuksi politiikkamuutokselle eli olisivat tapauskuntia. Näitä kuntia voidaan sitten verrata verrokikuntiin, joissa kys. politiikkamuutosta ei toteutettaisi. Mielenkiinnon kohteena olisi työllisyden kehitys altistumisen jälkeen.
- Käsittelyn tai altistuksen seuraauksia, esimerkiksi sairauden, syitä etiään vertaamalla tapausten ja verrokkien aikaisempaa altistumista erityisesti mielenkiinnon kohteena oleville altisteille.
- Tapaus-verrokitutkimus eroaa kohorttitutkimuksesta siten että siinä voidaan tutkia **yhtä tilastollista muuttuja** (kuten sairastumista), mutta **useita altisteita**: mistä altisteesta sairaus on seuraus, ts. mikä on taudinaiheuttaja?
  - \* Altistumishistoriaa voidaan selvitettää mm. mittauksilla, malleilla tai kyselylomakkeilla.
  - \* Esimerkki: tapauksien ja verrokkien altistumiseroista saadaan epäsuora arvio altistuneiden riskistä sairastua kyseiseen sairauksen suhteessa altistumattomien riskiin.
- Tapaus-verrokitutkimukset ovat yleensä suhteellisen yksinkertaisia ja halpoja toteuttaa niiden retrospektiivisestä luonteesta johtuen: tutkimuskysymys määrittelee aineistotarpeen, jonka jälkeen se tarvitsee vain kerätä.
  - \* Verrokin valinta kuitenkin kriittinen, sillä valitsemalla verrokkit/kontrollitapaukset väärin mikään tilastollinen testi tai metelmä ei korjaa tai kvantifioi tästä virhettä!
  - \* Esimerkki verrokkiryhmän epäkelvosta valinnasta on huonosti mitattu aiempi altistuminen ja/tai jos jokin tutkimuksen kannalta keskeinen taustamuuttuja sivuutetaan: mitä jos tauti tai sen vakavuus riippuukin sairastuneen muusta terveydentilasta?
- Eroaa poikittaistutkimuksesta siinä, että poikittaistutkimus pyrkii yleistämään tulokset koko kohdepopulaatioon, kun taas tapaus-verrokitutkimus keskittyy hyvin spesifiin populaation osaan

**Esimerkki tapaus-verrokkitutkimuksesta**

Länsi-Saksassa tuotiin 50-luvun lopulla markkinoille talidomidi-niminen uni- ja rauhoittava lääke. Varsin pian markkinoille tulon jälkeen tietyntyyppisten synnynnäisten epämuodostumien määrä alkoi lisääntyä rajuisti. Talidomidin ja lasten raajojen muodostumishäiriöiden yhteyks paljastettiin tapaus-verrokkitutkimuksilla. Tutkimuksissa selvitettiin sekä sairaiden lasten (tapaikset) että terveiden lasten (verrokki) äitien altistuminen talidomidille raskauden kriittisten viikkojen aikana. Melkein kaikki sairaiden lasten äidit olivat saaneet talidomidia ensimmäisten raskausviikkojen aikana (talidomidin oli myös havaittu helpottavan odottavien äitien raskauspahoinvointia). Talidomidi poistettiin markkinoilta ja epämuodostumatapausten määrä putosi jyrkästi.

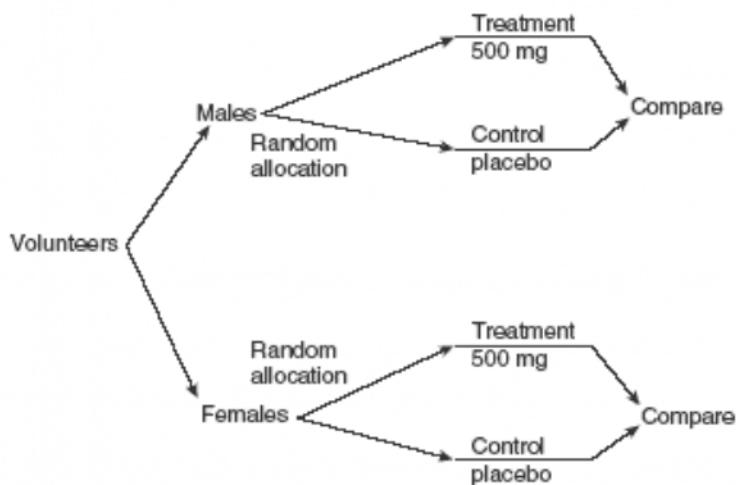


Kuva 10.6: Esimerkki tapaus-verrokkitutkimuksesta

- **Lohkot/osittaminen**

- Lohkomisella / osittamisella tarkoitetaan tutkimusyksiköiden järjestämistä ryhmiin niin, että ryhmät ovat mahdollisten sekoittavien tekijöiden suhteeseen mahdollisimman samankaltaiset

- Lohkotekijä on yleensä muuttuja, jonka aiheuttama vaihtelu vasteesseen ei ole tutkijan päämieniinnoon kohteena
- Lohkotekijän kontrollointi johtaa usein tarkkuudeltaan parempiin tuloksiin
- “Block what you can, randomize what you cannot.”



Kuva 10.7: Esimerkki ositetusta koeasetelmosta

#### • Sokkouttaminen

- Tutkimushoitojen sokkoutuksen tavoitteena on vähentää sekä potilaan että tutkimushenkilökunnan mahdollisten ennakkokäsitysten vaikutusta tutkimuksen tuloksiin.
- Yksöissokkotutkimussa (single-blind) tutkittava ei tiedä, mihin hoitoryhmään hän kuuluu.
- Kaksoissokkotutkimussa (double-blind) ei tutkittava eikä tutkija ja muukaan tutkimushenkilökunta saa tutkimuksen aikana tietää mihin hoitoryhmään tutkittava kuuluu.
- Kolmoissokkoutetussa tutkimussa edes havaintojen analysoija (esim. tilastotutkija) ei tiedä, miten kokeellisen tutkimuksen käsitteily on koodattu.
- Avoimessa tutkimussa (open) sekä tutkittava että tutkimushenkilökunta tietävät, mihin hoitoryhmään tutkittava kuuluu. Myös avoimessa tutkimussa tulisi yleensä käyttää satunmaistamista.

## 10.3 Erilaisia aineistoja ja aineistolähteitä

Käydään seuraavaksi läpi erilaisia aineistotyyppejä, joita käytännön tutkimuksissa Suomessa (ja maailmalla) usein käytetään. Emme käsitlee tässä erikseen itse otannalla koottavia aineistoja tai otannan järjestämistä, ks. luku 5.

### 10.3.1 Rekisteriaineistot

- Rekisteriperusteinen tutkimus hyödyntää aineistoinaan valmiiksi kerättyjä **tietokantoihin** tallennettuja aineistoja, joita kutsutaan rekisteriaineistoiksi.
  - Yleensä **hallinnollisia tarpeita** varten kerättyjä tietoja.
- Rekisteriaineistojen eduksi voidaan lukea mm. seuraavat seikat
  - Aineiston muodostaminen/kerääminen on verrattain helppoa ja Suomessa on paljon korkealaatusisia rekisteriaineistoja. Tätä edesauttaa tietotekniikan nopea kehittyminen, joka on mahdollistanut erittäin suuren aineistojen rutuininomaisen keräämisen.
  - Ei tarvetta erikseen tuottaa tutkimusaineisto: vältetään mahdollisesti kallis aineiston keräysvaihe.
  - Suomalainen henkilötunnusjärjestelmä mahdollistaa tietojen tehokkaan käytön ja laadukkaan tutkimuksen.
  - Rekisteriaineistoissa on paljon potentiaalia ja niiden tarjoamia tutkimusmahdolisuksia ei vielä täysimääräisesti tunneta tai ole hyödynnetty tutkijoiden toimesta.
- Rekisteriaineistojen ongelmina ja haittoina voidaan pitää mm. seuraavia
  - Mikäli tutkimuksessa lähtökohtaisesti käytetään rekisteriaineistoja, määräväät ne väilläisesti myös mahdolliset tutkimuskohteet: rekisteriaineistot kerätään eri tarkoitusta varten eivätkä ne täten välttämättä sisällä kaikkea haluttua informaatiota.
    - \* Tutkimuksen ongelmalahöösyys saattaa unohtua helpommin, kun tutkimusongelman aihiota asetellaan sopimaan rekisteriaineistojen tarjoamiin mahdolisuksiin.
    - \* Rekisteriaineistoilla on myös omat rajansa: tutkimuskysymysten kannalta väärin mitattua muuttujaa ei useinkaan voida millään tavalla muuntaa täydellisesti haluttuun muotoon.
  - Rekisteriaineisto pitää usein esikäsittellä sopivan muotoon laadullista tutkimusta muistuttavalla tavalla.
  - Rekisteriaineistojen analyysista ja niihin soveltuista tilastollisista menetelmistä on vähänäisesti metodologisia oppikirjoja ja/tai esimerkkitutkimuksia.

## 156 LUKU 10. AINEISTO- JA TUTKIMUSTYYPIT JA KOEASETELMAT

- “Ulkopuolisille” tutkijoille aineistojen käyttö saattaa olla hankalaa mm. korkeiden pääsykustannusten (rekisterien ylläpitäjien, viranomaisten ja tutkimuslaitosten ulkopuolella), tietosuojakysymysten tai teknisten hankaluksien takia.
  - \* Rekisteriaineiston käyttö vaatii tutkimussuunnitelman ja tutkimussuunnitelman perusteella myönnetyn käyttöluvan rekisterin ylläpitäjältä.
- Tietotekniikan kehittymisen vuoksi kasvaneet rekisteriaineistot tekevät käyttökelpoisentiedon esin seulomisesta haastavaa. Tämä näyttää esimerkiksi eri rekistereiden tietojen linkkaamista yhteen, jolla saattaa olla tutkimuksen kannalta ratkaiseva merkitys ja joka edelleen korostaa substanssitietoutta.
  - \* Eri rekisterejä ei aina saadakaan linkattua tehokkaasti yhteen esimerkiksi jos ne ovat mitanneet mielenkiinnon kohteina olevia muuttujia eri tilastoyksikön tasolla (vrt. kunnan vs kaupunginosan työllisyys)
- Erilaisia rekisterejä Suomessa
  - Verorekisterit (Verohallinnon rekisterit)
  - Kuolemansyyrekisterit
  - Eläkerekisterit
  - Väestölaskennat (väestörekisteri)
  - Syöpärekisteri
  - Lääkeostorekisteri
  - Sosiaali- ja terveydenhuollon rekisterit Kelan etuusrekisterit
  - Osoiterekisterit
  - Etukorttirekisterit
  - Opintosuoritusrekisteri
- Näiden lisäksi tulevat myös aikaisempien tutkimusten aineistot.
- Rekisteriaineiston käyttämisen **tilastollisia haasteita** tutkimuksessa
  - Rekisteriaineistot ovat usein kokonaisaineistoja, joten otantavirheseen perustuvan tilastollisen päätelyn oletukset eivät välttämättä päde.
    - \* Isoissa aineistoissa käytännössä merkityksettömistäkin eroista tulee helposti tilastollisesti merkitseviä!
  - Rekisteriaineistoja saadaan “valmiina” ja niiden kokonaistutkimukseen soveltuvalta luonteesta huolimatta niitä on arvioitava samojen periaatteiden mukaisesti kuin itse kerättäviäkin aineistoja.
    - Tutkimusongelman pitäisi aina olla keskeinen lähtökohta myös rekisteriaineiston käytössä.

- Itse kerätessä aineisto on mahdollista rääältöiden tuottaa vastamaan juuri tutkimuskysymykseen kun taas rekisteriaineisto on ”toisen käden” aineistoa ja ohjaa täten tutkimusta niin käsitteiden määrittelystä kuin tutkimuskysymysten asettelusta lähtien.
- **Tietosuojalaki:** Lain mukaan henkilötietoja voidaan kerätä ja tallettaa vain, jos rekisterinpitää ja rekisteröitävän henkilön välillä on asiallinen yhteys. Olennaista lain soveltamisessa on, voidaanko käytössä olevan aineiston tiedot tosiasiassa tavalla tai toisella liittää tiettyyn tunnistettavissa olevaan henkilöön.
  - Lain merkitys on paljolti siina, että se ohjaa suunnitelmissuuteen ja huolellisuuteen henkilötietojen käsittelyssä.
  - Sääntelyjen yleisenä tavoitteena on ettei tarpeettomasti kerätä ja tallenteta henkilötietoja ettei rekisteröityjen yksityisyys ja oikeuksia perusteettomasti loukata ja että rekisteri, siihen liittyvät tiedot ja niiden käsittely suojataan kaikissa vaiheissa.
- Rekisteriaineistot tietosuojalain takaaman yksityisyyslainsuojan näkökulmasta
  - Suomessa rekisteriaineistot pyritään luovuttamaan käyttöön vain tunnisteettomina yksityisyyslainsuojan säilyttämiseksi.
  - Tieteellinen tutkimustarkoitus luokitellaan kuitenkin poikkeustapaaksi tietosuojalaissa ja ensisijaisena tavoitteena on aina se että tietoja, joista henkilö voidaan tunnistaa käytetään ainoastaan silloin kun tutkimusta ei voida muutoin toteuttaa.
  - Tiedot tulee ensisijaisesti kerätä tutkimusyksiköiden suostumuksella ja siten, että henkilö saa halutessaan riittävästi informaatiota tietojen käyttötarkoituksesta ja -tavasta.
  - Rekisteritietojen hyödyntäjien etujen mukaista on, että tietosuoja koskevat säännökset ovat niin selkeitä ja kattavia, että yleisössä ei synny epäilyksiä tietojen väärinkäytön mahdollisuudesta.
  - Rekisteritietojen käyttöön on aina haettava lupaa. Erityisesti eri rekistereitä yhdistettäessä on hankittava myös tietosuojaavaltuutetun lausunto suunnitteilla olevan tutkimuksen laillisuudesta ja käytöehdoista.
- Rekisteriaineiston ymmärtämisessä ja käyttämisessä kannattaa huomioida ainakin seuraavat seikat
  - Mitkä tekijät ovat johtaneet alkuperäisen aineiston ja sen koonneen/-tuottaneen informaatiojärjestelmän syntymiseen?
    - \* Nimellisesti oikealta kuulostava muuttuja ei aina vastaa tutkijan käsitystä siitä muuttujasta, mitä kyseisen rekisteriaineiston ylläpitäjä/tuottaja on ajatellut.

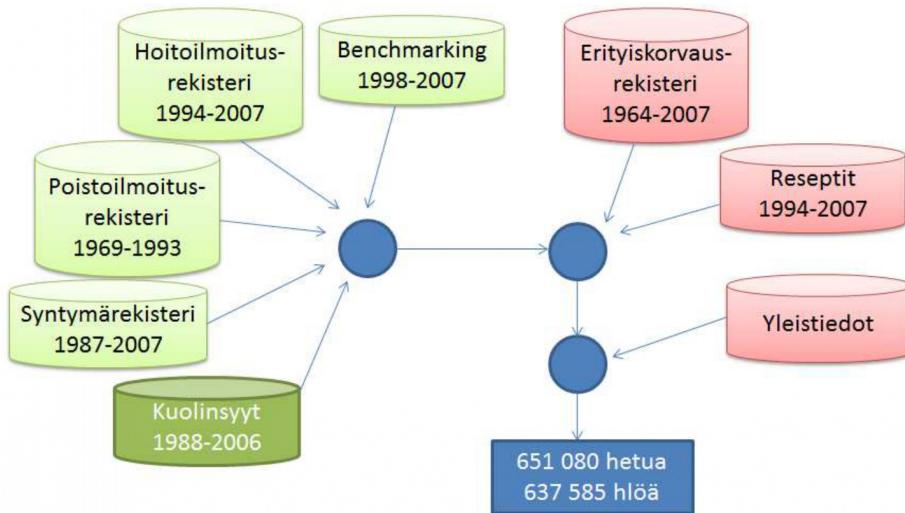
## 158 LUKU 10. AINEISTO- JA TUTKIMUSTYYPIT JA KOEASETELMAT

- Miten järjestelmän sisältämät tiedot on mitattu ja miten tämä ilmoitetaan eli miten tietojärjestelmän sisältämien tietojen informaatioarvo on dokumentoitu?
  - \* Rekisterin tuottajan ja sen käyttäjien näkemykset mitatuista muuttujista ja niistä johdetut tulkinnot eivät välttämättä aina kohtaa: [esimerkki?]
- Minkälaisia tietorakenteita aineistossa käytetään ja miten se vaikuttaa eri muuttujien tallentamiseen tietojärjestelmään?
  - \* Tutkimuskysymyksen kannalta on voi olla merkitystä esimerkiksi sillä, onko rekisterinpitää kerännyt henkilöiden ikätoea vuoden vai kymmenen vuoden tarkkuudella.

### **Esimerkki: Diabeteksen ja sen lisäsairauksien esiintyvyyden ja ilmaantuvuuden rekisteriperusteinen mittaaminen**

- Vaihe 1: Diabeteskohtortin identifiointi (Tilastokeskus: kuolinsyyt, THL: diagnoosit, Kela: erityiskorvaukset, reseptit)
  - Vaihe 2: Seurantatiedot (syöpärekisteri, sairaspäivärahat, eläkerekisteri...)
- 
- Ongelma: kuinka monta henkilöä diabeteskohortista on kuollut seuranta-aikana?
    - Kuolema on vakaa käsite, johon ei liity mittausvirhettä tai subjektiivisuutta
    - Katsotaan aineistosta ”yhden rivin säällöllä” kuinka monelle löytyy tieto kuolemasta
  - Kysymys: Kuinka moni diabeteskohorttiin kuuluvista sairastaa tyypin 1 diabetesta?
    - Tyypin 1 diabetes johtuu insuliinia tuottavien beetasolujen tuhoutumisesta autoimmuniprosessin seurauksena
    - Tyypin 1 diabeetikko tarvitsee jatkuvasti insuliinia, mutta ei hyödy haiman omaa insuliinineritystä tehostavista lääkkeistä
    - Rakennetaan algoritmi, jolla identifioidaan tyypin 1 diabeetikot lääkeostojen luokkien ja säällöllisyyden perusteella

### **Esimerkki: rekisteritutkimus pitkääikaisen laitoshoivan käytöstä**

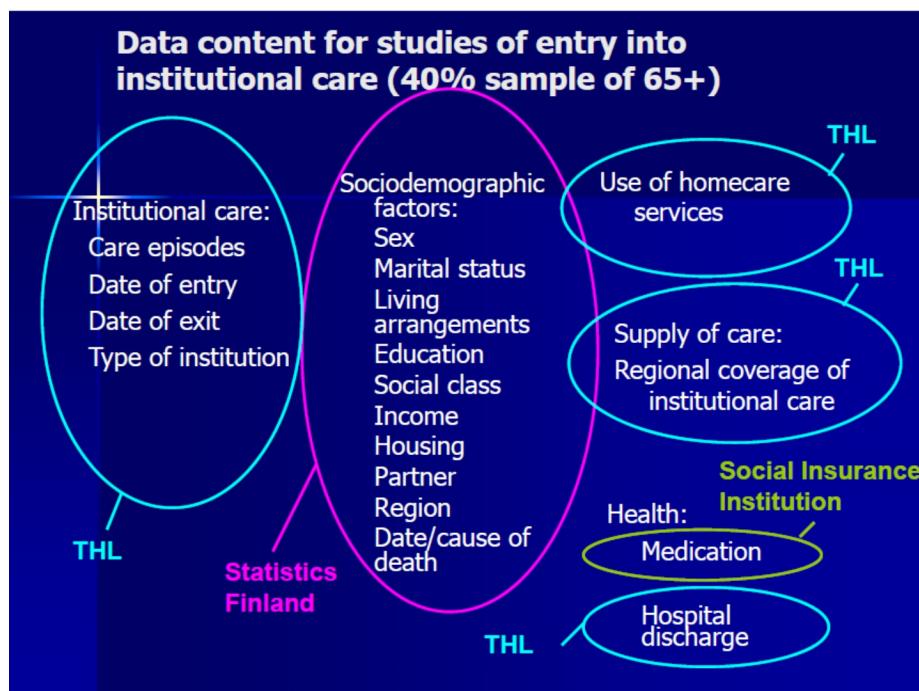


Kuva 10.8: Vaihe 1: Diabeteskohortin identifiointi

- Miten sosiaaliset tekijät, kuten sosioekonominen asema ja perherakenne, vaikuttavat laitoshoivan käyttöön?
- Kolme erityistä tutkimusintressiä
  - Laitoshoivaan siirtymisen riskit
  - Laitoksissa vietetty aika
  - Laitoshoivan käyttö elämän loppupäässä
- Aiemmat tutkimukset samasta aiheesta
  - Perustuvat potilasaineistoihin
  - Eivät sisällä laitostumis- ja poistumistietoja samassa aineistossa
  - Kärsivät vastauskadrosta
  - Kärsivät seurantakadrosta ja seurannan puutteellisuudesta
  - Perustuvat pieniin aineistoihin
  - Eivät mahdollista perhevaikutusten tutkimista

### 10.3.2 Aikasarjat ja paneeliaineistot

- **Aikasarjat**
  - Aikasarjaksi kutsutaan havaintojen jonoa, jossa aika määräää jostain tilastollisesta muuttujasta tehtyjen havaintojen järjestyksen.
  - Havainnot ovat tavallisesti peräkkäisiä, ja mittaukset on tehty tasaisin aikavälein, mutta väliaikojen tasaisuus ei kuitenkaan ole välttää



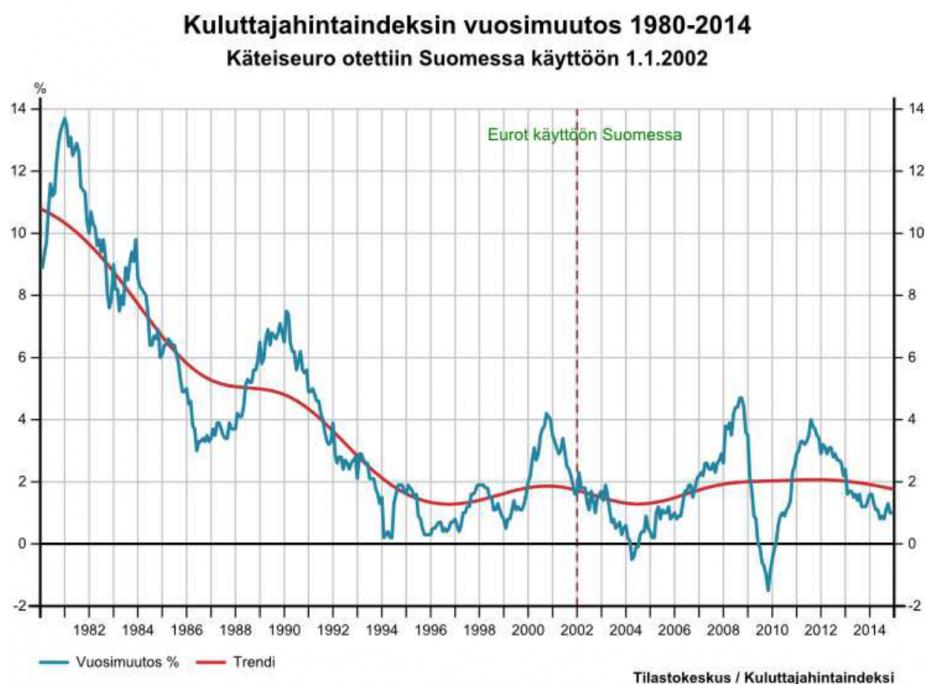
Kuva 10.9: Vaihe 1: Diabeteskohortin identifiointi

- mätöntä ja monissa tutkimusasetelmissa kohdeaikasarjasta voidaan poimia havaintoja jatkuvasti tai mielivaltaisen pienin aikavälein.
- Yksittäinen aikasarja on pitkittääsaineiston erikoistapaus, jossa tarkastellaan vain yhtä aikasarjaa. Pitkittääsaineistoon nähden toistot eivät välttämättä ole suunniteltuja, vaan niitä havaitaan jatkuvasti ajassa.
    - \* Vuotuinen bruttokansantuote Suomessa
    - \* Suomalaisten lukumäärä kunkin vuoden lopussa
    - \* Vuorokautinen sademäärä Helsingin Kaisaniemessä
  - Jotkut aikasarjat ovat suunnitelmallisesti muodostettu tiettyjen laskennallisten menetelmien avulla muista aikasarjoista. Tälläisiä tilastollisia suureita kutsutaan **indekseiksi** ja ne sisältävät tiivistettyä tietoa yhteiskunnasta, kuten esimerkiksi inflaation mittarina käytetty kuluttajahintaindeksi.

#### Esimerkki: kuluttajahintaindeksit

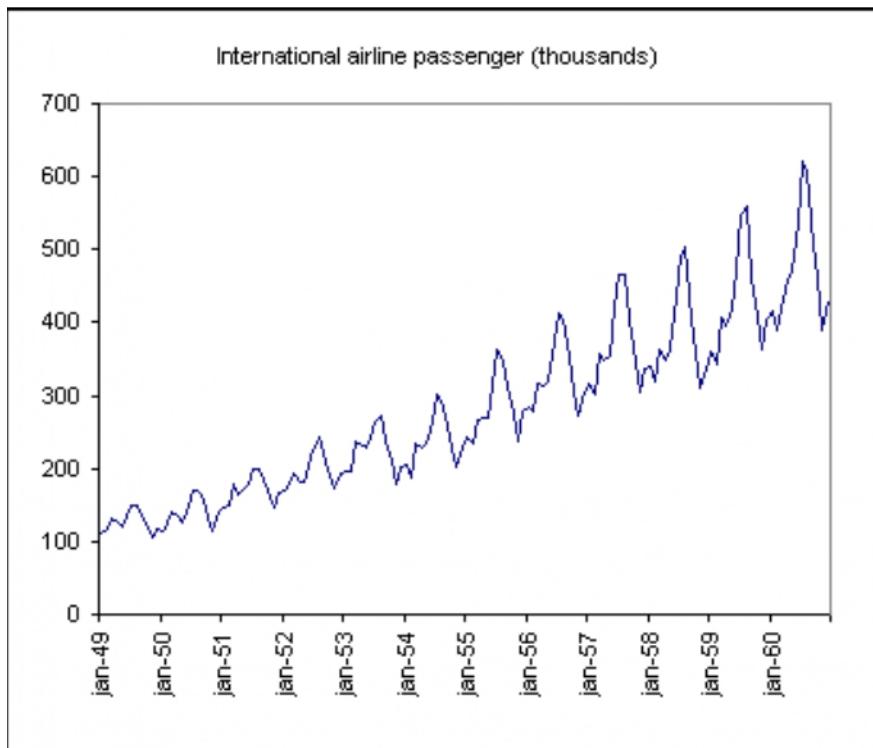
- Hintaindeksi on normalisoitu keskiarvo tarkasteltavan tuote- tai palvelukorin hinnoista, joka lasketaan säännöllisin väliajoin tavoitteena helpottaa hintojen muutosten seurantaa eri ajankohtien tai alueiden välillä. Laajimmat indeksit mittavat hintojen kehitystä koko talouden tasolla ja niitä hyödynnetään monin tavoin talouspolitiikassa.
  - Tilastokeskuksen haastattelijat keräävät indeksiä varten kaiken kaikkiaan noin 50000 hintatietoa lähes 500 hyödykkeestä noin 2700 (nämä joitain vuosia sitten) liikkeestä aina kuukauden puolivälissä, lisäksi noin 1000 hintatietoa kerätään keskitetysti.
  - Tilastokeskuksen laskema kuluttajahintojen vuosimuutos oli marraskuussa 2014 oli 1,0 prosentti
  - Marraskuussa kuluttajahintoja nostivat viime vuodesta eniten vuokrankorotukset, tupakkatuotteiden ja alkoholijuomien vähittäishintojen ja ravintola- ja kahvilopalvelujen sekä kerrostalojen kunnossapitolvelujen kallistuminen
  - Kuluttajahintojen nousua vuoden takaisesta hillitsi marraskuussa eniten elintarvikkeiden, poltonesteiden ja viihdeelektronikaan halpeneminen

- **Aikasarjojen tilastollinen analyysi** perustuu siihen, että sarja tulkitaan jonkin **stokastisen prosessin** eli satunnaisprosessin realisaatioksi
  - Jos aikasarjan generoinut prosessi saadaan selville, voidaan tietoja prosessista käyttää aikasarjan käyttäytymisen kuvaamiseen ja selittämiseen sekä aikasarjan tulevan käyttäytymisen ennustamiseen.



Kuva 10.10: Esimerkki: kuluttajahintaindeksit

- Aikasarja-aineistot ilmentävät ilmiöstä riippuen ns. autokorrelaatiota, eli ajallisesti toisiaan lähellä olevat havainnot ovat korreloitu-neempia kuin ajallisesti kaukana toisistaan olevat.
  - \* Aikasarja-analyysi on yksi tilastotieteen osa-alue, jossa on rikas ja pitkälle kehittynyt teoriapohja.
  - \* Aikasarja-analyysia opetetaan tarkemmin kursseilla TILM3541 Aikasarja-analyysi, TILM3589 Epälineaarinen aikasarja-analyysi ja TILM3586 Moniulotteinen aikasarja-analyysi.
- Aikasarjoja analysoimalla voidaan selvittää esimerkiksi
  - Onko aikasarjassa **trendejä** eli aikasarjan tason systemaattisia muutoksia?
  - Onko aikasarjassa **syklistä vaihtelua** kuten **suhdanne-** ja/tai **kausivaihtelua**?



Kuva 10.11: Esimerkki: Kansainvälisten lentomatkustajien lkm vuosina 1949-1960

- **Paneeli- eli pitkittäisaineisto**

- Paneeliaineistolla tarkoitetaan aineistoa, jossa tilastoiksiöistä on useita havaintoja ja aika määräät havaintojen järjestyksien (kuten aikasarjoissa) ja lisäksi jokaisena ajanhetkenä mitataan useampi kuin yksi tilastollinen muuttuja (kuten poikkileikkausaineistossa).
  - \* Paneeliaineisto on terminä käytetympä yhteiskuntatieteissä kun taas pitkittäisaineisto esimerkiksi lääketieteessä.
  - \* Havaintoiksiot voivat olla esimerkiksi yrityksiä, ihmisiä, kuntia tai kouluja. Ns. ”täydellisessä” paneeliaineistossa kaikista havaintoiksiöistä on havaittu kaikki muuttujat kaikkina ajanhetkinä. ”Kiertävä” paneeli on vastaavasti sellainen, jossa osa havaintoiksiöistä vaihtuu ajan kuluessa.
    - Tyypillisesti havaintoja kerätään tasaisin väliajoin, kuten kuukausittain tai vuosittain, ja yksittäisen ajanhetken havainto on poikkileikkausaineisto ja kustakin havaintoiksiöistä on oma usean muuttujan aikasarjansa.
- Paneeliaineisto mahdollistaa vastaamisen kysymykseen miksi? Yleisesti ottaen paneeliaineistoja käytetäänkin erityisesti ns. kausaalipäättelyyn tähtäävissä malleissa.<sup>2</sup>

### 10.3.3 Survey- eli haastattelu- tai kyselytutkimus

- Survey-tutkimus on ei-kokeellinen tutkimus, jonka lähtökohtana on tiettyjen ilmiöiden, ominaisuuksien tai tapahtumien yleisyyden tai jakautumisen selvittäminen, joka toteutetaan kysely- tai haastattelumenetelmällä.
  - Havaintoiksiöt pyritään valitsemaan satunnaisotannalla, sillä myös survey-tutkimuksessa pyritään yleistämään tulokset otoksesta koko perusjoukkoon.
  - Kyselytutkimukset muodostava kokonaan oman tutkimustapansa, joka mahdollistaa hyvin erilaisen informaation keräämisen kuin tavallisesti kvantitatiivissa aineistoissa.
- Survey-tutkimus koostuu seuraavista vaiheista
  - Kohdepopulaation määrittely ja otannan suunnittelu
  - Kyselylomakkeen rankentaminen ja testaaminen
  - Kyselymetodin määrittely (puhelinhaastattelu, elektroninen kysely...)
  - Mahdollisten haastattelijoiden koulutus
  - Aineiston keräys
  - Aineiston yhteneväisyyden tarkistaminen (muuttujat tallennettu oikein jne)

---

<sup>2</sup>Kausaalimalleja opetetaan kurssilla TILM3529 Kausaalipäättely havainnoivissa tutkimuksissa

- Tulosten adjustointi mahdollisten identifioitujen virhelähteiden mu-kaan
- Survey-tutkimusta käytetään erityisesti asenteiden, mielipiteiden ja käyt-täytymisen tutkimiseen.
  - Esimerkkejä ovat mm. poliittiset mielipidekyselyt, markkinointitut-kimukset, alkoholinkulutustottumukset ja terveyspalveluiden tyyty-väisyyskyselyt, joihin voi kaikkiin uskoa liittyvän vastausharhaa eri syistä.
  - Esimeriksi vastausharhaa arkaluontoisiin kysymyksiin voidaan vä-hentää avoimilla kysymyksillä tai alustamalla kysymystä johdannolla, jossa suvaitaan / ymmärretään kaikenlaiset vastaukset.
  - “Randomized response”: vastaukseen lisätään jonkin todennäköisyys-mallin mukaisesti harhaa todellisen vastauksen salaamiseksi.
- Kerätty aineisto on siten altisteinen tehdylle kyselylle ja täten sen käyt-tökelpoisuus perustuu hyvin pitkälti etukäteissuunnittelun, kunnolliseen toteutukseen ja kyselylomakkeen oikeaoppiseen rakentamiseen.
  - Lisäksi aineiston käyttökelpoisuus riippuu myös vastaajaotoksen poi-minnasta (edustavuudesta) ja siitä, kuinka totuudenmukaista infor-matiota vastaajat ovat kyselyssä antaneet. Tässäkin hyvä etukäteis-suunnitelu on keskeistä.
  - Tilastollisten menetelmien avulla pyritään arvioivan otoksen, kyse-lyn suunnittelun ja kerätyn vastaajaotoksen sisältämää (tai aiheut-tamaa) harhaa.

## The item count method for sensitive survey questions: modelling criminal behaviour

Jouni Kuha and Jonathan Jackson

*London School of Economics and Political Science, UK*

[Received July 2012. Revised January 2013]

**Summary.** The item count method is a way of asking sensitive survey questions which protects the anonymity of the respondents by randomization before the interview. It can be used to estimate the probability of sensitive behaviour and to model how it depends on explanatory variables. We analyse item count survey data on the illegal behaviour of buying stolen goods. The analysis of an item count question is best formulated as an instance of modelling incomplete categorical data. We propose an efficient implementation of the estimation which also provides explicit variance estimates for the parameters. We then suggest specifications for the model for the control items, which is an auxiliary but unavoidable part of the analysis of item count data. These considerations and the results of our analysis of criminal behaviour highlight the fact that careful design of the questions is crucial for the success of the item count method.

**Keywords:** Categorical data analysis; EM algorithm; List experiment; Missing information; Newton–Raphson algorithm; Randomized response

Kuva 10.12: Esimerkki survey-tutkimuksesta. (Kuva 1)

**Table 1.** The item count question on buying stolen goods, as included in the Euro-Justis survey

'I am now going to read you a list of five [six] things that people may do or that may happen to them. Please listen to them and then tell me how many of them you have done or have happened to you in the last 12 months. Do not tell me which ones are and are not true for you. Just tell me how many you have done at least once.'

[Items included in both the control and treatment groups]

1. Attended a religious service, except for a special occasion like a wedding or funeral.
2. Went to a sporting event.
3. Attended an opera.
4. Visited a country outside [your country]?
5. Had personal belongings such as money or a mobile phone stolen from you or from your house.

[Item included in the treatment group only]

6. Bought something you thought might have been stolen.

Kuva 10.13: Esimerkki survey-tutkimuksesta. (Kuva 2)

**Table 2.** Numbers of respondents with different reported totals for the item count question in the Euro-Justis survey

<i>Group</i>	<i>Item count</i>						<i>Total</i>	
	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	
Control	269	472	257	133	54	21	—	1206
Treatment	279	446	281	124	53	20	9	1212

Kuva 10.14: Esimerkki survey-tutkimuksesta. (Kuva 3)



## Luku 11

# Tilastollisesta ennustamisesta

Kuten olemme jo tähän menneessä nähneet, tilastollinen analyysi ja sen erottamattomana osana tilastollinen päättely on keskeinen vaihe tieteellistä tutkimusta. Vielä ennen tilastollisen selittämisen ja ennustamisen välistä eroja koskevia pohdintoja muistutetaan jo aiemmin käsitellystä **kuvalevasta tilastotieteestä**. Tämä voidaan nähdä vielä (ainakin) kolmantena yleisenä tilastotieteen tavoitteena mallintamisen/selittämisen ja ennustamisen lisäksi. Yksinkertaisin tilastollisen päättelyn muoto on hyödyntää aineistoa kuvalevia tunnuslukuja, kuten kesiarvoja ja keskijahontalukuja. Niistä voidaan kuitenkin tehdä vain melko rajoittuneita päätelmiä. Varsinkin havainnoivassa tutkimuksessa sen selvitämiseksi, miten selittävät muuttujat ovat yhteydessä selittävään vaste-muuttujaan, käytetään esim. lineaarista tai logistista regressiota (ja niiden monenmonia laajennuksia) tai esim. aikasarja-analyysiä aikasarjoja analysoitaessa. Näiden pohjalta voidaan arvioida muuttujien yhteyksiä ja riippuvuussuhteita.

Käytännössä tilastotieteen ja sen sovellusalueiden tutkimuksessa tulisi osata erottaa (tilastollinen) **selittäminen ja ennustaminen**. Tätä eroa koskevat tarkeimmat yksityiskohdat ovat jälleen selvästi tämän kurssin ulkopuolella myöhemmissä tilastotieteen opinnoissa, mutta seuraavassa kuitenkin tähän liittyviä keskeisiä huomioita.

### 11.1 Tilastollinen selittäminen vs. ennustaminen

- (Tilastollinen) **selittäminen** tarkoittaa esim. kahden muuttujan välisen yhteyden tutkimista (tämän kurssin yksinkertainen tilanne lineaarisesti)

regressiomallin yhteydessä, jota voidaan laajentaa useisiin muuttuijiin). Tutkijaa saattaa kiinnostaa esimerkiksi tupakoinnin vaikutus sepelvaltimotautikuolleisuuteen tai ylipainon vaikutus leikkauksen jälkeisiin infekcioihin.

- Tällöin **pyrkimyksenä** on rakentaa ”**selitysmalli**”, jossa on perustellut syy-seuraussuhteet selittävästä (selittävistä) muuttujista selittävään muuttujaan. % Tarvittaessa mukaan otetaan sekoittavia tekijöitä, joilla tiedetään olevan syy-seuraus-suhde kumpaankin.
- (Tilastollinen) **ennustaminen** tarkoittaa, että tietyillä selittävän tai selettävien (tai ”ennustavien”) muuttujien yhdistelmillä tai arvoilla voidaan ennustaa ennustettavan muuttujan arvoa.
  - Ts. siis ennustettavana muuttujana toimii tilastollisen mallin näkökulmasta katsoen vastemuuttujan arvo, jota pyritään ennustemallin avulla ennustamaan.
  - Ennustemalleja tutkitaessa varsinaisilla selityssuhteilla ei välttämättä ole merkitystä. Tärkeintä on mallin ennustekyky, ei niinkään esim. yksittäisen regressiokertoimen arvo ja siihen liittyvät tarkemmat tulkinnat. Tilastollisesti merkitsevä regressiokerroin ei tarkoita, että muuttujalla olisi välttämättä ennustekykyä.
  - Ennustekyky tutkittava erikseen. Esimerkiksi lineaarisen mallin perinteiset tunnusluvut, kuten selitysaste, eivät vielä kerro mallin todellisesta ennustekyvystä paljoakaan. Tästä huolimatta melko usein ennustemallin rakentaminen perustetaan pitkälle samoihin tilastollisen päättelyn ja estimointiteorian lähtökohtiin mitä olemme jo sivunneet tällä kurssilla.
  - Hyvin usein tutkimuksissa raportoidaan, että tietty muuttuja ”ennustaa” (predicts) toista. Usein kuitenkin taustalla on tällöin usean muuttujan selitysmalli, jonka regressiokertoimien tilastollista merkitsevyyttä on tulkittu. Yleensä tässä yhteydessä on kuitenkin siis kyse selittämisestä, ja kuten todettua, mallin ennustekyky pitää tutkia erikseen.
- Erityisesti aikasarja-analyysissä ennustaminen on perinteisesti ollut yksi kaikkein keskeisimmistä tavoitteista.
  - Nyt kuitenkin koneoppimisen (tilastollisen oppimisen) kasvatettua hurjasti suosioitaan viime vuosina ennustaminen on levinnyt, ja levää jatkossakin, voimaakkasti myös hyvin keskeiseksi osaksi muutakin tilastollista analyysiä.

## 11.2 Tilastolliseen ennustamiseen liittyviä huomioita

- Ennustamista on kaikki! Se rooli on paljon keskeisempi osa meidän kaikkien arkea mitä ensiajatukselta saattaa tulla mieleen.
  - Ennustaminen on elämässämme korvaamatonta.<sup>1</sup>
  - Kun valitsemme reitin työmatkalle, päättämekö menemmekö toisille trelleille tai säästämme huonompia aikoja varten, teemme ennusteen tulevaisuuden kehityksestä ja siitä, miten suunnitelmamme vaikuttavat suotuisan tuloksen todennäköisyyteen.
- Arkiset ongelmat eivät aina vaadi ankaraa ajattelua ja pohdiskelua erilaisista vaihtoehtojen välillä niihin käytettävissä olevan ajan ollessa rajallinen. Tästä huolimatta teet ennusteita tiedostaen ja useimmiten tiedostamatta monta kertaa päivässä!
- **Ennustevirhe:**
  - Ennusteita verrataan toteutuneeseen kehitykseen. Näiden erotuksena muodostuu ennustevirhe.
  - Lähtökohtana on (luonnollisesti) minimoida ennustevirheet. Käytännössä useinmiten mm. vastemuuttujan luonteen perusteella valitaan sopiva ennustevirheitä summarisoiva tunnusluku, kuten keskineliöennustevirhe (jatkuvat vastemuuttujat) tai luokitteluvasteiden tapauksessa väärin (tai oikein) ennustettujen luokitteluiden suhteellinen osuus.
  - Ajoittain ennustetarkkuutta on helpompi ja toisaalta sitten vaikeampi tarkkailla. Esim. taloustieteessä on paljon helpompi arvioida työttömyyttä koskevaa ennustetta kuin esimerkiksi ennustetta (jopa väitettyä) velkaelvytyksen tehokkuudesta. Toisaalta valtio-opissa voidaan arvioida vaalitulosta koskevia ennusteita suoraviivaisesti, mutta saattaa kulua vuosikymmeniä nähdä miten poliittisten instituutioiden ennusteisiin perustuvat ennakoidut muutokset vaikuttavat poliittisten päätösten tuloksiin.

**Esimerkki:** Naten kirjan luvun 1 pohdintaa ennustevirheestä *finanssikriisiin (rahoituskriisiin)* vuoden 2008 aikana (finanssikriisiin voidaan katsoa koskeneen lopulta vuosia 2007-2009). Pörssikurssien voimakas lasku, Lehman Brothersin kaltaisia aikoinaan arvostettuja yhtiöitä meni vararikkoon, luottomarkkinat olivat käytännössä ”jäätyneet”, Las Vegasissa asuntojen hinnat las-

<sup>1</sup>Tämän alaluvun pohdinnat, kuten tämä lainaus, perustuvat pitkälle kirjan Naten kirjan Signaali ja kohina (suom. Kimmo Pietiläinen) huomioihin.

kivat 40 prosenttia (osoittaen osaltaan vallinnutta laajempaan ”**asuntokuplaa**” (perusteettoman korkeita asuntojen hintoja), työttömyys kasvoi räjähdyksimäisesti jne.

**Ennustevirheen yhteisiä ja tyypillisiä piirteitä** (tässä tapauksessa), jotka laajentuvat moniin muihinkin tilanteisiin ja sovelliuksiin: 1. Asunnonomistajat ja sijoittajat ajattelivat, että nousevat hinnat viittasivat siihen, että asuntojen hinnat jatkaisivat nousuaan, kun todellisuudessa historia viittasi siihen, että sen takia niillä oli taipumus laskea (näissä olosuhteissa). 2. Luottoluokistuslaitokset (samoin kuin Lehman Brothersin kaltaiset pankit) eivät ymmärtäneet, miten riskialtiita asutovakuudelliset arvopaperit olivat. Ongelma ei varsinaisesti ollut siinä, että luokituslaitokset eivät nähneet asuntokuplaa. Sen sijaan niiden ennustemallit olivat täynnä huonoja oletuksia ja väärää ”itseluottamusta” mahdollisten asuntojen hintojen romahduksen riskeistä. 3. Laajasti ei enakoitu, miten asuntokriisi laukaisee globaalini rahoituskriisiin. Se johtui suurelta osin liiallisesta velkaantumisesta markkinoilla, jossa lyötiin erinäisten instrumenttien myötä vetoa yhdysvaltalaisen halukkuuden puolesta sijoittaa uuteen kotiin. 4. Rahoituskriisin välittömässä jälkimainungeissa ei osattu ennustaa, miten laajoja taloudellisia ongelmia se aiheuttaa. Rahoituskriisit tyypillisesti tuottavat erittäin syviä ja pitkäkestoisia taloudellisia taantumajaksoja.

Näissä ennustamisen epäonnistumisissa on **yhteinen piirre**. Kussakin tapauksessa aineistoa arviodessaan ihmiset jättivät keskeisen asiayhteyden palan huomiotta:

1. Asunnonomistajien luottamus asuntojen hintoihin johtui ehkä siitä, että lähimenneisyydessä Yhdysvalloissa asuntojen hinnat eivät olleet laskeneet merkittävästi. **Kuitenkaan** koskaan aikaisemmin Yhdysvaltojen asuntojen hinnat eivät olleet nousseet niin laajalla alueella kuin romahdusta edeltävällä kaudella.
2. Pankkien luottamus luottoluokituslaitosten (kuten Moody'siin ja S&P'siin) kykyyn luokittaa asutovakuudellisia arvopapereita ehkä perustui siihen, että laitoksina ne olivat onnistuneet pätevästi luokittamaan muunlaista rahoitusomaisuutta. **Kuitenkaan** luottoluokituslaitokset eivät olleet koskaan aikaisemmin luokittaneet yhtä uusia ja monimutkaisia arvopapereita mitä tuolloin (kuten ns. luotonvaihto-optioita).
3. (Taloustietelijöiden) luottamus rahoitusjärjestelmän kykyyn kestää asuntokriisi syntyi ehkä siitä, että aikaisemmin asuntojen hintojen heilahtelulla ei yleensä ollut suuria vaikuttuksia rahoitusjärjestelmässä. **Kuitenkaan** rahoitusjärjestelmä ei luultavasti koskaan aikaisemmin ole ollut yhtä vekkaantunut eikä vedonlyöntiä asunto-

jen hinnoista ollut tehty vastaavassa mittaluokassa.

4. Poliittisten päättäjien luottamus siihen, että talous toipuu nopeasti rahoituskriiseistä syntyi ehkä viime vuosikymmenten taantumista saaduista kokemuksista. Useampia niitä oli seurannut nopea "V-muotoinen" toipuminen (kuten nyt myös myöhemmin mm. koronapandemian aikaan). **Kuitenkaan** nämä taantumat eivät olleet liittyneet rahoituskriiseihin ja rahoituskriisit ovat (yleensä) erilaisia.

Jokaista edellistä kohtaa yhdistää ennustamiseen hyvin keskeisesti liittyvä termi: Ennustajien pohtimat ilmiöt olivat ns. **otoksen ulkopuolella** (engl. **out-of-sample**). Kun ennustaminen epäonnistuu merkittävällä tavalla, tämä ongelma jättää yleensä runsaasti sormenjälkiä rikospaikalle. Miten tämä ongelma näyttääsi siis oheisen esimerkin tapauksessa? - Luottoluokituslaitos (kuten Moody's) arvioi, missä määrin asuntolainojen hoitamatta jättämiset liittyivät toisiinsa, rakentamalla (luultavasti ainakin osin) tilastollisen mallin menneisyyden aineiston perusteella. Oletettavasti he käyttivät mallin rakentamiseen noin 1980-luvulle ulottuva Yhdysvaltain asuntomarkkina-aineistoa. - Ongelmana oli, että 1980-luvulta 2000-luvun alkuvuosiin saakka asuntojen hinnat olivat aina vakaat tai nousevat Yhdysvalloissa. Tässä tilanteessa oletus, että asunnonomistajien asuntolainat eivät juuri kaan liittyneet toisiinsa oli luultavasti perusteltu ja riittävän hyvä tilastollisen mallintamisen pohjaksi. - Kuitenkaan menneessä aineistossa mikään ei olisi kuvannut mitä tapahtuu kun asuntojen hinnat alkavat laskea kauttaaltaan samanaikaisesti. Ts. asuntoromahdus oli **otoksen ulkopuolin tapahtuma** ja tässä tilanteessa luottoluokituslaitosten mallit olivat arvottomia (huonoja) lainojen hoitamatta jättämisen riskiä arvioitaessa.

- Rahoituskriisiä koskevan esimerkin tilanteessa otoksen ulkopuolisia ilmiöitä koskeva ongelma realisoitui siten, että muodostettu tilastollinen malli, kuten vaikkapa lineaarisen regressiomallin sopiva laajennus, **estimoitiin**, tai koneoppimisesta tutussa kielenkäytössä **opetettiin**, aineistolla, joka ei lopulta ollut relevantti juuri myöhemmin tapahtunutta kriisivaihetta ajatellen.
  - Onkin tärkeää ymmärtää, että "todellisessa" ennustetilanteessa joudumme käyttämään aiempaa aineistoa mallien ja algoritmien rakentamiseen. Näin ollen näiden ennustekykyä arvioitaessa onkin mentävä otoksen ulkopuolelle, koska "otoksen sisällä" voimme opettaa kyseisiä malleja (ääritilanteessa) niin, että ne ovat periaatteessa ääret-

tömän tarkkoja. Ne eivät kuitenkaan takaa missään mielessä hyvää ennustekykyä tulevia tapahtumia ennustettaessa.

- Vastaavalla tavalla karakterisoidaan nykyään hyvin suosittujen koneoppimismenetelmien keskeinen piirre: Ne ja tarkasteltavat sovellukset perustuvat käytännössä (vielä) yksinomaan ennustesovelluksiin. Tällöin mallien ja algoritmiien opettaminen ns. **opetusaineistolla** (edellä olevassa esimerkissä aiemmassa asuntomarkkina-aineistolla) ja ennustekyvyn arviointi **ennusteotoksen** avulla pitää erottaa toisistaan.
  - **Otoksen sisäiseen sovittamiseen** (engl. **in-sample** tai **training sample** estimation, ajoittain prediction) liittyy siis (ennustamisen näkökulmasta) katsoen ns. **ylisovittamisen** vaara. On mahdollista, että yritämme puristaa lähes puhtaasta kohinasta signaaleja, jotka eivät missään mielessä tule olemaan valideja otoksen ulkopuolisessa ennustamistilanteessa.
  - Jälleen kerran näistä teemoista keskustellaan tarkemmin myöhemmin tilastotieteen aine- ja syventävien opintojen erikoiskursseilla, kuten TILM3587 Regressioanalyysi ja tilastollinen oppiminen ja TILM3592 Tilastollisen oppimisen jatkokurssi.
- Huolimatta edellä käydystä, kriittisestäkin, keskustelusta ennustamiseen liittyen, monet ennusteet ovat varsin tarkkoja ja samalla vapaita ylisovitamisen vaaroista!

## Luku 12

# Tilastotieteen kehityksen nykytrendejä

Seuraavassa vielä joitain tilastotiedettä parhaillaan koskevia kehitystrendejä, joita olemme myös osaltaan sivunneet tämän kurssimateriaalin myötä: - **Aineistot kasvavat ja monimutkaistuvat.** Näin ollen tilastotitedettä ja tilastollisten menetelmien kehitystyötä tullaan tarvitsemaan (yhä suuremmin) jatkossakin. - Informaatiota on tarjolla (paljon) enemmän kuin osaamme sitä hyödyntää. **Informaatio ei ole enää niukka hyödyke.** Keskeistä on kuitenkin, että (useimmiten) suhteellisen pieni osa informaatiosta on hyödyllistä. - Havaitsemme informaatiota (osin) ja valikoivasti subjektiivisesti. Luulemme haluavamme lisää informaatiota, kun todellisuudessa haluamme tietoa (ts. signaalajeja kun vastaavasti kasvava määrä kohinaa yrittää vaikeuttaa tästä signaalin erottamista kohinasta) - **Laskennallisuus** kasvaa. Tietokoneiden laskentakapasiteetti nousee vaikkakin suhteellinen kasvu ei ehkä olekaan enää niin suurta mitä muutamat viime vuosikymmenet. - Osin laskennallisuuteen liittyvä **koneoppimisen** kehittyminen ja sen "rajatieteiden" yhteyteen integroituneet käytännöt ja menetelmäkehitystyö tulee jatkumaan. - **Analyysien automatisointi:** Tilastolliset ohjelmat alkanevat jatkossa tulkitta tuloksia osin automaattisesti. Mihin tarvitaan tilastotieteilijää? Kokonaisvaltaiseen tutkimusprosessin valvontaan (?) - Osin edelliseen liittyen jo nyt ja jatkossa luultavasti yhä enemmän korostuu se, että melkein kuka vaan voi tehdä tilastollisia analyysejä. Niin valmiita paketteja jne. on jo saatavilla. "Tilastotieteellinen faktantsekkaus" noussee vahvemmin esille eli tilastollisten menetelmien käyttäjän on sittenkin edelleen kyettävä arvioimaan ovatko tulokset uskottavia ja vapaita ilmeisistä hankaluksista. - **Poikkitieteellisyys** tulee entisestäänkin vahvistumaan. Ts. substanttietouden ja tilastotieteen yhdistäminen ja sen tärkeys ei tule ainakaan vähenemään. Sivuaineopintoja kannattaa siis ottaa. - Tämän lisäksi kokonaisvaltaisesti tilastotieteen ytimen osaajien osaamista tulla kysymään jatkossakin.

- Oheisten huomioiden ohella lopuksi on syytä korostaa tilastotieteen opiskelun näkökulmasta, että oikotietä ei ole! **Oikaista siis ei voi:** Ensin on rakennettava vahva tilastollisen ajattelun ja menetelmien perusta (**alkaen todennäköisyyslaskennasta ja tilastollisesta päättelystä**), jotta myöhemmin voi kehittyä ja omata todellisia kykyjä ottaa vastaan monia jo varsin monimutkaisia tilastollisia menetelmiä!