

# TILM3701 - Tilastotiede ja data 2023

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

2023-09-05

<sup>1</sup>Turun yliopisto, matematiikan ja tilastotieteen laitos, [henri.nyberg@utu.fi](mailto:henri.nyberg@utu.fi)  
<sup>2</sup>Turun yliopisto, matematiikan ja tilastotieteen laitos, [roope.rihtamo@utu.fi](mailto:roope.rihtamo@utu.fi)



# Sisällys

<b>Kurssin rakenne</b>	<b>7</b>
Kurssimateriaali . . . . .	8
<b>1 Johdantoa ja johdattelua tilastotieteeseen</b>	<b>11</b>
1.1 Tilastotiede ja kurssin idea . . . . .	11
1.2 Tilastotieteen asema tutkimusyhteisön ulkopuolella . . . . .	13
1.3 Kurssin luonne tilastotieteen opintojen esittelijänä . . . . .	14
<b>2 Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa</b>	<b>15</b>
2.1 Mitä on tiede? . . . . .	15
2.2 Tieteellinen menetelmä . . . . .	20
2.3 Tilastojen yleisestä roolista yhteiskunnassa . . . . .	23
2.4 Mitä on tutkimus? . . . . .	25
2.5 Tutkimuksen vaiheet ja tulosten julkaiseminen . . . . .	28
2.6 Luvun 2 yhteenveto, keskeisiä termejä ja kokonaisuuksia. . . . .	29
<b>3 Tilastotiede tieteenalana</b>	<b>31</b>
3.1 Lisää tilastotieteen perustermejä . . . . .	31
3.2 Mitä tilastotiede on ja mitä se ei ole? . . . . .	33
3.3 Tilastotieteen suhde lähitieteisiin . . . . .	38
3.4 Tilastotieteen osa-alueet . . . . .	43
3.5 Tilastotieteen kritiikkiä . . . . .	48
3.6 Tilastotieteen sovellusalojen ja “rajamittateitä” . . . . .	53
3.7 Luvun 3 yhteenveto, keskeisiä termejä ja kokonaisuuksia. . . . .	54

<b>4 Sattuma ja satunnaisuus tilastotieteessä</b>	<b>57</b>
4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä . . . . .	58
4.2 Satunnaisuus ja todennäköisydyt . . . . .	61
4.3 Tilastolliset mallit, jakaumat ja parametrit . . . . .	64
4.4 Odotusarvo ja varianssi . . . . .	66
4.5 Joitain jakaumia . . . . .	68
4.6 Sattuman rooli tieteenteossa: Vale-emäviale-tilasto? . . . . .	73
4.7 Luvun 4 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	74
<b>5 Tilastolliset aineistot, niiden kerääminen ja mittaaminen</b>	<b>77</b>
5.1 Kertausta: Data eli aineisto . . . . .	78
5.2 Otannan idea . . . . .	82
5.3 Mittaaminen ja mitta-asteikot . . . . .	85
5.4 Kontrolloidut kokeet ja suorat havainnot . . . . .	90
5.5 Otantamenetelmät . . . . .	93
5.6 Otantaesimerkkejä . . . . .	101
5.7 Otannan haasteita vielä kootusti . . . . .	103
5.8 Luvun 5 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	103
<b>6 Otokset ja otosjakaumat: tilastollisen päättelyn näkökulma</b>	<b>105</b>
6.1 Satunnaisotos, yhteisjakauma ja tilastollinen malli . . . . .	105
6.2 Otojakama: Estimaattori ja estimaatti . . . . .	108
6.3 Otokeskiarvo ja otosvarianssi (estimaattoreina) . . . . .	111
6.4 Suhteellisen frekvenssin otojakama . . . . .	114
6.5 Muita tunnuslukuja . . . . .	116
6.6 Luottamusvälit . . . . .	117
6.7 Otokoko . . . . .	124
6.8 Luvun 6 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	129

<b>7 Tilastollinen riippuvuus ja korrelaatio</b>	<b>131</b>
7.1 Muuttujien väliset riippuvuudet . . . . .	131
7.2 Kahden muuttujan havaintoaineiston kuvaaminen . . . . .	133
7.3 Tunnusluvut . . . . .	136
7.4 Satunnaismuuttujien kovarianssi ja korrelaatio . . . . .	137
7.5 Luvun 7 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	143
<b>8 Regressioanalyysi</b>	<b>145</b>
8.1 Johdatus regressioanalyysin ideaan . . . . .	145
8.2 Yhden selittäjän lineaarinen regressiomalli . . . . .	147
8.3 Muita regressiomalleja . . . . .	154
8.4 Luvun 8 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	154
<b>9 Tilastotieteen rooli uuden tiedon tuottamisessa</b>	<b>157</b>
9.1 Tilastollisen tutkimuksen yhteisiä elementtejä . . . . .	157
9.2 Tutkimusprosessi . . . . .	160
9.3 Luvun 9 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	164
<b>10 Aineisto- ja tutkimustyyppit ja koeasetelmat</b>	<b>167</b>
10.1 Tutkimustyyppit . . . . .	168
10.2 Tutkimusstrategiat . . . . .	174
10.3 Erilaisia aineistoja ja aineistolähteitä . . . . .	183
10.4 Luvun 10 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	194
<b>11 Tilastollisesta ennustamisesta</b>	<b>195</b>
11.1 Tilastollinen selittäminen vs. ennustaminen . . . . .	195
11.2 Tilastolliseen ennustamiseen liittyviä huomioita . . . . .	196
11.3 Luvun 11 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	200
<b>12 Tilastotieteen kehityksen nykytrendejä</b>	<b>201</b>
12.1 Modernin tilastotieteen suuntaviivoja . . . . .	201
12.2 Luvun 12 yhteenvetö, keskeisiä termejä ja kokonaisuuksia. . . . .	202



# 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 välttetään, 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 tätä 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 tilastotieteeseen: Tilastotieteen johdantokurssi (1.kirja). Yliopistopaino, Helsingin yliopisto.
  - Mellin, I. (2000). Johdatus tilastotieteeseen: 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 Lääärä ja Pekka Pere (2021). [Tilastotieteen sanasto. Suomen Tilastoseuran julkaisuja 8.](#)

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

## Kurssimateriaali

Kurssin materiaali on koostettu em. lähteistä ja pyrkii paikoin pelkistettyyn esitysmuotoon mutta kuitenkin niin että materiaalin opiskelemalla kurssin osaamistavoitteet täytyvät kokonaisuudessaan. Osaamistavoitteet on listattu Turun yliopiston opinto-oppaassa matematiikan ja tilastotieteen laitoksen opintotarjonnasta [kurssikuvausen alta](#) ja ne löytyvät alta vielä laajemmin.

- Opintojakson suoritettuaan opiskelija:
  - On saanut kokonaiskuvan tilastotieteestä ja sen perusteista
  - Osaa hahmottaa tilastotieteen roolin omana tieteenalana ja eri sovellusalueiden yhteydessä
  - Tunnistaa erilaiset tutkimusasetelmat ja aineistotyyppit
  - On sisäistänyt tilastotieteen keskeisiä käsitteitä ja osaa niiden avulla tarkastella kriittisesti tieteellisiä tutkimuksia
  - Pystyy erottamaan edustavan otoksen ja näytteen

Kurssin sisältöä on listattu opinto-oppaassa ja laajemmin alla. Tämä listaus toimii hyväänä luettelona kurssin keskeisistä teemoista.

- Kurssin sisältöä:
  - Tilastotiede tieteenalana ja sen suhde lähitieteisiin, kuten datatieteeseen (data science)
  - Tilastotieteen rooli uuden tieteellisen tiedon tuottamisessa
  - Tilastolliset aineistot (data), niiden kerääminen ja mittaaminen
  - Tilastollisen päättelyn perusteita
  - Otannan perusteet
  - Tilastotieteen sovellusten ja sovellusalueiden esittelyä

Materiaalin seassa on erityltä väärinkodatuin tietolaatikoin erinäisiä tärkeitä tilastotieteellisiä konsepteja ja termejä sekä esimerkkejä tilastotieteen sovelliuksesta. Näistä ensin mainitut löytyvät Deltan violeteista laatikoista ja jälkimmäiset Statistikkan oransseista.<sup>1</sup> Alla esimerkkilaatikot.

**Konsepti tai termi**

Konseptin tai termin löyhä määritelmä.

**Esimerkki**

Aihetta koskeva esimerkki.

---

<sup>1</sup>Toim. Huom. värit eivät täysin alkuperäisten värien kanssa yhteneväisiä.



# 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 ihmisen 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 tyypissä tutkimusongelmissa.
  - Esitellä tilastotieteen roolia **empiirisen tutkimusaineiston keräämisessä ja analyysissä** sekä tarkastella tieteentekemisen ja tilastotieteen suhdetta.
  - Pohtia **tilastotieteen olemusta tieteenalana** ja tarkastella tilastotieteen ja datatieteiden (data sciencen) 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.

## 12 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 se 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 formaalissa mallintamisessa ja päättelyssä. Tilastolliseen 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 estimoiduille parametreille.
- Toinen kurssin keskeisistä teemoista on tarkastella tieteellistä tutkimusprosessia teoriassa ja käytännössä. Tämä sisältää mm. seuraavia aiheita (joita siis käsitellään tällä kurssilla päällisin puolin varsin yleisestä näkökulmasta katsoen ja tarkemmat yksityiskohdat jätetään 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 täyttyä, jotta asetettuun tutkimusongelmaan voidaan vastata?

## 1.2. TILASTOTIETEEN ASEMA TUTKIMUSYHTEISÖN ULKOPUOLELLA13

- \* **Otanta** (ja mittaaminen): miten tutkimusaineisto kerätään niin, että se täyttää aineiston ennakkohdot? Erilaisissa tutkimuksissa käytetään erilaisia aineistoja kuten:
  - Survey- eli haastatteluaineistot: aineisto kerätään haastattelemalla tutkimuskohteita
  - Rekisteriaineistot: aineisto on kerätty valmiiksi rekisteriin ja sitä käytetään tutkimukseen
  - Aikasarja-aineistot tai pitkittäisaineistot: useita mahdollisesti korreloituneita havaintoja samoista tutkimuskohteista
  - Ynnä muita, ks. luku 10
- **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 saatu 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 tieteenteossa.
- Tiedeyhteisön ulkopuolellakin **tilastotiedettä ja tilastotieteilijötä arvostetaan laajalti**.
- **Tilastotiede onkin nostanut profiliaan viimeisten vuosikymmenien aikana** tietoteknisen kehityksen tuotua laajat tietoaineistot ja kehitetyneet laskennalliset menetelmät lähes jokaisen kansalaisen saataville.
- Tämä “*datavallankumous*” näkyy tilastotieteilijöiden kysynnässä työmarkkinoilla: erilaisten aineistojen määrään lisääntyessä kasvaa myös kysyntä työntekijöistä, jotka osaavat ammatitaitoisesti 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 datainensivisten ammattien tekijöitä yhdistää yksi tekijä: **heidän tulee hallita ja osata tilastotiedettä!**
  - Karkeistettuna mitä paremmin ja enemmän (laajemmin), sen parempi palkka ja monipuolisemmat työtehtävät!

### 1.3 Kurssin luonne tilastotieteen opintojen esiteltijänä

Kurssin mittaan esitellään tilastotieteen perusteiden lisäksi **miten Turun yliopistossa tilastotieteen opinnoissa syvennytään** tällä kurssilla esiteltäviin menetelmiin, aineistotyyppeihin ja mallinnuskokonaisuuksiin. Tilastotieteen opintotarjontaan voi perehtyä [TY:n opinto-oppaan avulla!](#)

## 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 saavutetavuus ja tunnettiuus on lisääntynyt mm. tieteen popularisoinnin ja median laajemman tiede-uutisoinnin vuoksi. Tiedon, erityisesti tieteellisen tiedon, rooli korostuu yhä enemmän myös kaikilla elämän osa-alueilla: terveysteknologia (esim. sykemittarit tai Oura-sormus) perustuu lääke- ja terveystieteellisiin läpimurtoihin, talouspoliittisia päätöksiä edeltää entistä suurempi määrä asiantuntijoiden taloustiedeperusteista analyysia ja jopa peruskouluopetus on murroksessa kasvatustieteen saavutusten myötä.

Voidakseen ymmärtää ja arvioida kriittisesti tiede-uutisia tulee lukijan olla tietoinen tieteellisen tutkimuksen luonteesta: miten tutkimusartikkeleja luetaan, mitä niiltä voidaan odottaa ja minkälaiset tulokset ovat uskottavia. **Tilastotiede näyttelee keskeistä roolia lähes kaikessa tutkimuksessa ja erityisesti erilaisten tutkimuskysymyksien ja niitä vastaavien hypoteesien testauksessa.** Aloitetaan kurssin varsinainen oppimateriaali kunnianhimoisesti tarkastelemalla 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.

## 16LUKU 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) niitää koskevia käsityksiä vahvistetaan osoittamalla ne tosiksi se kä
      - iii) löydetaan niistä uutta tietoa.
    - Tiede siis erottaa intuition ja ”arkitiedon” oikeasta, tutkitusta tiedosta esittämällä reaalimaailmaa koskevia väitteitä ja osoittamalla ne toteksi tieteellisin menetelmin.
    - Tiede käsittää myös aiemman tutkimuksen ja se toimii kaiken tieteellisen tiedon jäsenneltyynä 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 tiedettä sen tuottaessa uutta tietoa.
- 
- Tieteelle voidaan asettaa (ainakin) seuraavat kaksi sitä määrittelevää ominaisuutta.
    - **Järjestelmällisyys:** tieteellinen tiedonhankinta on yhteiskunnalliseksi organisoitu tutkimusta tekevien (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 siis voi nojautua

- \* Yksilölliseen vaistoon tai intuitioon
  - \* Suostutteluun
  - \* Propagandaan
  - \* "Jumalalliseen ilmoitukseen" tai vastaavaan
- 
- Tieteen keskiössä on todellista maailmaa koskevat (tieteelliset) **teoriat** ja niihin liittävät **hypoteesit**.

### Tieteellinen teoria

Tieteelliset teoriat ovat hyvin perusteltuja kuvausia ja selityksiä siitä, miten ympäröivä maailmamme toimii tai esimerkiksi siitä miten eri ilmiöt ovat yhteyksissä toisiinsa. Ne ovat luotetuina, täsmällisin ja kattavin tieteellisen tiedon muoto. Teorian vahvuus riippuu siitä, kuinka laajoja ja erilaisia reaalimailman ilmiöitä sillä voidaan (yksinkertaisesti) selittää.

- Teoria muodostuu tieteellistä menetelmää käytämällä ja se on kehittynyt ajassa kumulatiivisesti kertyneen tiedon myötä. Teoria muodostuu siis toistuvien sitä vahvistavien uusien havaintojen ja tutkimuksen myötä.
- Tieteellisen teorian pyrkimys on selittää ja ennustaa sen kohteena olevaa ilmiötä tyylkkäästi sekä yksinkertaisesti. Se on luonteeltaan induktiivinen ja alisteinen muutokksille tai jopa hylkäämiselle empiirisen todistusaineiston ("evidenssin") osoittaessa sen olevan puutteellinen tai väärä.
  - Tieteellisen teorian tulee siis olla empiirisesti testattavissa ja sen tekemät ennusteet falsifioitavissa: teoriaan liittyvät ennustukset määrittelevät sen hyödyllisyden, sillä teoria joka ei tee testattavia ennustuksia on hyödytön.
  - Teoriat kehittyvät vuorovaikutuksessa todellisen maailman kanssa kun tieteellisessä tutkimuksessa niitä ja erityisesti niihin liittyviä hypoteeseja testataan ja saatuja tuloksia tulkitaan vallitsevien teorioiden valossa.
    - \* Jos tulokset ovat linjassa teorian tekemien ennustusten kanssa, teoria vahvistuu (se "verifioidaan") ja riittävän evidenssin myötä se voidaan hyväksyä, eli siitä on *tieteellinen konsensus*: paras mahdollinen selitys kys. ilmiölle.
    - \* Jos tulokset poikkeavat teorian ennustuksista, ne tulkitaan teorian empiiriseksi vastaväitteeksi ("falsifikaatioksi"). Tällöin voidaan ensin tarkastella onko tulokset saatu uskottavalla *tieteellisellä menetelmällä* ja mikäli näin on, ja seuraavatkin tutkimustulokset ovat vastaavia, teoriaa voidaan parantaa tai mahdolisesti muuttaa kokonaan.

## 18LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- Tämä tieteellisen tiedon kumuloituminen muokkaa teorioita vuosien saatossa täsmällisemmiksi ja paremmiksi kuvausksiksi ympäröivästä maailmasta.
  - \* On kuitenkin syytä huomauttaa että tieteellisetkään teoriat eivät ikinä ole (eikä niiden tarvitse olla) täydellisen täsmällisiä, jotta ne olisivat käyttökelpoisia ja hyödyllisiä.
- Teorianmuodostukseen liittyy keskeisesti tieteellinen menetelmä, johon taas liittyy teorioita koskevien *hypoteesien* testaaminen.

### Hypoteesi

- Hypoteesi tarkoittaa teorioista johdettua tai aikaisemman tutkimuksen perusteella esitettyä ennakoitua ratkaisua tai selitystä tutkittavaan ongelmaan.
- Hypoteesi ilmaistaan teoriaa koskevana väitteenä, jonka paikkansapitävyyttä halutaan tutkia.
- Hypoteeseja voidaan testata kokeellisesti ja näin saadut tiedot/tulokset voivat osoittaa hypoteesin vääräksi.
- **Nollahypoteesi** vastaa tavallisesti tyypillistä, odottavissa olevaa tulosta, esimerkiksi ettei kahden mitatun ilmiön välillä ole yhteyttä tai että tietty hoito on tehotonta.
  - Nollahypoteesia *ei todisteta* (“*hyväksytä*”), vaan voidaan ai-noastaan sanoa, ettei aineisto tarjoa todistusaineistoa nollahypoteesin hylkäämiselle.
- **Vastahypoteesi** sisältää usein mielenkiinnon kohteena olevan tapahuman, kuten “on eroa” tai “on vaikutusta”.
  - Tiedeyhteisöllä on usein taipumus jättää julkaisematta tutkimustuloksia, joissa nollahypoteesi jäädä voimaan. Yleensä tämä tilanne syntyy, kun lopputulos ei eroa jo aikaisemmin otaksumusta. (Toki ajoittain tilanne on myös toisinpäin eli “toivotaan” nollahypoteesin hylkäämistä).

- Tieteilijät yleensä perustavat hypoteesinsa aikaisemmin tehtyihin havaintoihin, joita ei voida selittää olemassa olevilla tieteellisillä teorioilla tyydyttävästi.
- Uuden tieteellisen tiedon tuottaminen ja jo tuotetun tiedon ymmärtäminen vaatii **tieteellisen ajattelutavan** omaksumista, jonka **perustana on lähes aina tilastollinen päättely**.
  - Tieteelliselle ajattelulle ja tiedon tuottamiselle on tunnusomaista, että se pohtii ja kehittelee **paradigmojaan** eli oman toimintansa perusteita.

**Paradigma** on tietyyn 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 yhtenevät raamatit, jotka ohjaavat sitä, miten tutkimuskysymyksiä asetetaan ja miten niihin etsitään vastauksia sekä myös sitä, miten saatuja tuloksia tulkitaan.
  - Tieteellinen tieto perustuu siis eri tutkimusalojen tiedeyhteisöjen paradigmoihin ja täten siihen, minkälaisista tutkimusta, ja mistä ilmiöistä, kannattaa tehdä.
  - Paradigmojen ei pidä ajatella olevan kaavoihin kangistuneita ajatteluja 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 paradigmojen kanssa ristiriidassa oleva tulos saattaa jäädä julkaisematta, mikäli tutkija ei pidä sitä lainkaan mahdollisena suhteessa vallitseviin paradigmoihin.
    - \* Samoin on käytännössä varmaa, että vallitsevat paradigmat muuttuvat ajan myötä uusien löytöjen myötä!

## 20LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- Tieteelliseen ajattelutapaan kuuluu ollenaisesti juuri tiedon kumuloitumisen ymmärtäminen: yksittäinen vahva tulos on vasta alkua 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 määritetietoisesti rationaalisen 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 käyttöö ohjaa tutkimusalakohtaiset paradigmat, jotka ovat suuntaviivoja ja viiteistöjä siitä, minkälainen tutkimus tuottaa uskottavia tuloksia.

Arkitieto	Tieteellinen tieto
<ul style="list-style-type: none"><li>▶ epäluotettavat havainnot</li><li>▶ epäjohdonmukaisuus</li><li>▶ omien kokemusten vaikutus</li><li>▶ logiikan puute</li><li>▶ lyhytjänteisyys</li><li>▶ valikoivat havainnot</li><li>▶ muistamattomuus</li><li>▶ irrallisuus asiyhteydestä</li><li>▶ tytyminen ensimmäiseen selitykseen</li><li>▶ liiallinen yleistäminen</li></ul>	<ul style="list-style-type: none"><li>▶ perustuu tietoiseen opiskeluun, analyysiin ja yleistämiseen (ontantateoria)</li><li>▶ muodostaa hierakkisen järjestelmän</li><li>▶ objektiivisuus</li><li>▶ etsii yleisiä lainmukaisuuksia ja periaatteita</li><li>▶ perusteltua</li><li>▶ julkista</li><li>▶ korjaantuvaa</li><li>▶ kriittisyys</li><li>▶ olennaisen ja epäolennaisen erottaminen</li></ul>

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) riippuvala, tieteellistä **menetelmää** eli **metodia**.

**Tieteellinen menetelmä:** Tieteellinen menetelmä on kullakin tieteen alalla vallitseva, ajan myötä kehittynyt ja nykyisten paradigmojen mukainen 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 ratkomisessa. Hyvälle tieteelliselle menetelmälle voidaan lukea seuraavia kriteerejä.

- **Objektiivisuus ja loogisuus**

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

- **Kriittisyys**

- 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 edistymisen 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 yleistettävyys**

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

## 22LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- Tarkastellaan lähemmin erästä määritelmää erilaisille toistettavuuden lajeille. Esittemme tässä Hamermeshin (2007)<sup>2</sup> esittämän erilaisten 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 (otosta), joka on kuitenkin poimittu samasta populaatiosta (ks. Luku 5), mutta samaa menetelmää, saa vastaavanlaisia tuloksia, jotka vahvistavat alkuperäisen tutkimuksen perustulokset.
  - **Tieteellinen replikointi:** toinen tutkija, käyttäen samoja asioita mittaavaa tutkimusaineistoa, joka on kuitenkin kerätty eri populaatiosta, ja käyttäen samankaltaista, mutta ei identtistä menetelmää, saa vastaavanlaisia 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 formaalin ja objektiivisen tutkimusaiheen kuvaukseen 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 ja niiden testaamisen avulla. Päättely perustuu tilastotieteen teoriaan, joka mahdollistaa päättelyn epävarman ja satunnaisen aineiston tapauksissa.
  - Hypoteesien asettamisen voidaan ajatella tutkittavaa ilmiötä koskeviksi ennusteiksi, joita verrataan havaittuun aineistoon. Mikäli havaittu aineisto ei sovi testattavaan teoriaan tai siihen liittyviin hypoteeseihin, voidaan (hieman yksinkertaistaen) teoriaa kehittää paremaksi. Tämä vuoropuhelu vie tiedettä eteenpäin ja tuottaa lisää tutkittua tietoa ympäröivästä maailmasta.

<sup>2</sup>Hamermesh, D. S. (2007). Replication in economics. *Canadian Journal of Economics/Revue 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.
  - Esimerkiksi 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 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 ymmärtämislä on keskeinen rooli tietoyhteiskunnassa toimimisessa ja vaikuttamisessa.

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

**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 tarkeimmin myöhemmin luvussa 5).
  - Arvotaan satunnaisesti tutkimushenkilötä miesten ja naisten joukosta ja mitataan kuinka paljon he puhuvat.
  - Satunnaisuus tärkeää, sillä jos valikoitaisiin tarkoituksella puheilaita tai vähäsanaisia tutkimushenkilöitä, tulokset väärityisivät.
- 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?
  - Otoskoolla, eli sillä kuinka monta tutkimushenkilöä tutkitaan, on keskeinen rooli tutkimuksen luotettavuudelle. Mitä suurempi otos, sitä pienemmäksi sattuman osuus käy ja vastaavasti mitä pienempi otos, sitä suurempi on yksittäisten sattumienvaikutus.
- Tilastolliset mallit turvautuvat todennäköisyyskuviin erottaakseen sattuman vaikutuksen: kun aineisto on kerätty, halutaan tietää kuinka todennäköistä on, että uskomus pitää paikkaansa.
  - \* Tilastolliset mallit turvautuvat todennäköisyyskuviin erottaakseen sattuman vaikutuksen: kun aineisto on kerätty, halutaan tietää kuinka todennäköistä on, että uskomus pitää paikkaansa.
- Palataan takaisin esimerkkiimme: Yleisen uskomuksen mukaan naiset puhuvat enemmän kuin miehet.
  - Tutkimuksen mukaan miehet vaikuttavat kuitenkin puhuvan yhtä paljon kuin naisetkin.
  - Laajemmat tutkimukset osoittavat, että **tilanteella** on puhelin määräään paljon suurempi vaiketus 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 useimmissa olla) tiedollisen uteliaisuuden, käytännön tarpeiden tai teorian kehittämispyrkimen 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 paikkansapitävyyttä.
  - Valtaosa tieteestä asemoituu kahden viimeisen kohdan alaisuuteen vaikka tieteen popularisoinnissa (mm. median toimesta) usein keskitytäänkin uusiin tiedemaailmaa ja joskus "käytännön" elämää järistyväin löydöksiin, jotka tosin voivat olla hyvin epävarmoja!
  - Lisää tieteen popularisoinnista jaksossa 4.6.

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

- 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 työhönsijointumiseen? Tämän tutkimuskysymyksen tapauksessa pyrkimys on lähtien selittää (ymmärtää) ilmiötä.
  - **Ennustaminen:** Jos kansantulon kasvu pienenee x%, työttömyyden ennustetaan kasvavan y tuhannella.
  - Kohdetta kuvaavien 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.
    - \* Kehittynytkin tieteellinen teoria tai malli on aina reaalimaailman yksinkertaistus: tutkimus on aina alisteinen käytetylle menetelmälle ja sen oletuksille!
  - Ymmärtämiseen tarvittava havaintomaailman hahmotus (saattaa) tuottaa ideologisesti ja historiallisesti sitoutuneita yksinkertaistavia sekä luonteeltaan usein hyvin teoreettisia abstraktioita.
    - \* Alakohtainen substanssitetous sekä sen vahvuusien ja puutteiden sekä historiallisen ja ideologisen kontekstin tiedostaminen on ensiarvoisen tärkeää kaikessa tutkimuksessa!
  - Joka tapauksessa täyneen neutraaliuteen ja objektiivisuuteen on mahdotonta päästä. 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.
    - \* Arvot ilmenevät esimerkiksi tutkimuksessa käytetyissä käsitteisissä, jotka harvoin ovat arvovapaita. Useimmat käsitteet voidaan korvata toisilla, joilla on paikoin hyvin erilainen arvosisältö joskin arvottava lataus saattaa myös olla paikoin tarkoituksellista! Joka tapauksessa arvpainotteisten valintojen tunnistaminen on vaikeaa.

- \* Toisaalta arvoihin sitoutuminen on väistämätöntä, sillä se on sosiaalisen olemassaolon sivutuote. Yhteiskunnan jäseninä meillä on tuskin mahdollisuksia (täydellisesti) irroittautua arvoistamme kun pyrimme esim. ammatillisii päämääriin.
  - Myös päinvastainen ongelma olemassa: Tutkimusta arviodaan 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**.
  - Tutkimuskysymys, valitut menetelmät, käytetty aineisto ja tehdyt johtopäätökset perataan auki tutkimusartikkelissa/raportissa, joka sitten lähetetään **vertaisarvioitavaksi** tietelliseen julkaisuun, jossa muut alan asiantuntijat arvioivat sen ja päättävät hyväksytäänkö se julkistarvaksi.
- **Vertaisarvioinnissa** yksi tai useampi, tehdystä tutkimuksesta riippumaton, saman alan tutkija lukee ja tarkastaa tehdyn tutkimusartikkelin, arvioi sitä ja suosittaa tietellisen julkaisun arvioinnista vastaavalle päätoimittajalle (editorille) kyseisen artikkelin hyväksymistä tai hylkäämistä.
  - Vertaisarviointi ei aina takaa sitä, että julkaistu tutkimus olisi virheellinen ja erinomaisesti tehty, vaan myös väärää tietoa pääsee välillä vertaisarviointiprosessin läpi.
  - Tämä ei kuitenkaan poista tieteellisen prosessin luotettavuutta, sillä uusi tieto varmentuu vasta usean samaa tutkimuskysymystä tutkineen ja vastaavat tulokset saaneen tutkimuksen myötä. Toisin sanoen, tieteellisen prosessin voidaan ajatella konvergoituvan totuuteen, vaikka yksittäisiä virhehearvointeja sattuisikin.
- **Tutkimuksen kieli**
  - Tutkimus edellyttää arkikieltä täsmällisempää kommunikaatiota.
  - Ongelmaan liittyvien käsitteiden huolellinen määritteleminen ja erityyli on tarpeellista.
    - \* Käsitteiden ja eri aloilla, osin samoista asioista käytettävien, toisistaan eroavien termien systemaattinen määrittely ja jäseni tely selkeyttää tiedeyhteisön välistä kommunikointia.

## 28LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- \* Eivät korvaa empiiristä tietoa vaan vaikuttavat tiedon järjestymiseen ja sen perusteella tehtäviin päätelmiin.

### Esimerkki: Luonnontieteelliset vs. yhteiskunnalliset sovellutukset:

- Luonnontieteiden lainsäädäntö: 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 lainsäädäntöjä ja tyypillisiä piirteitä:
  - Usein tutkitaan **yhteiskunnallisia ilmiöitä**, jotka eivät suurelta osin ole toistettavissa.
  - Vaihtelevat huomattavasti ajan myötä (aiemmin voimassaolleet lainsäädäntöjä eivät välttämättä ole enää voimassa ja päinvastoin), mikä vaikeuttaa tilastollista analyysiä.
  - Yhteiskunnallisten ilmiöiden mittaaminen?
    - \* Yhteiskunnan rakenne ja toiminta on ehdollinen siinä käytettäväni merkitysjärjestelmän suhteen. Kysymys **mittaamisesta** on asetettava suhteessa tähän käsitejärjestelmään. Joudutaan tekemään erilaisia kompromisseja eksaktisuus- ja systemaattisuusvaatimusten sekä arkikielessä monimerkityksellisyyden välillä.

## 2.5 Tutkimuksen vaiheet ja tulosten julkaiseminen

Tieteellinen tutkimus ja asiantuntijatyö tuottavat valtavan määrän 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/>

## 2.6. LUVUN 2 YHTEENVETO, KESKEISIÄ TERMEJÄ JA KOKONAISSUUKSIA.29

Vastuullisen tieteen

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

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

Helsingin yliopisto tarjoaa lisäksi [Tiedelukutaidon perusteet -kurssia](#) MOOC-toteutuksena (Massive Open Online Course). Keskustelethan ennen kurssin käymistä oman alasi koulutussuunnittelijan (tai vastaavan vastuuhenkilön) kanssa siitä, soveltuuko kyseinen kurssi sisällytettäväksi johonkin omaan opintokokonaisuuteesi.

- 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.

## 2.6 Luvun 2 yhteenveton, keskeisiä termejä ja kokonaisuuksia.

- Tieteellinen teoria
- Hypoteesi
- Tieteellinen menetelmä ja hyvän tieteellisen menetelmän kriteerit
- Epävarmuus
- Mitä on tutkimus

**30LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA**

## 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 empiiriseen tutkimukseen siitätäkin huolimatta että tilastotieteellä on oma rikas teoriapohjansa sekä kiistaton asema omana tieteenalanaan.

Tieteenalan määritteleminen lyhyesti on aina hieman hankalaa. 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 tutkimusyhteisön ulkopuolella?
- Minkälaisista on tyyppillinen tilastotiedettä kohtaan esitetty kriitikki?

### 3.1 Lisää tilastotieteen perustermejä

Seuraavia tilastotieteen esittelyä ja karakterisointeja ajatellen määritellään seuraavassa lisää tilastotieteellisen tutkimuksen peruskäsitteitä. Näihin käsitteisiin paneudutaan osaltaan tarkemmin mm. luvussa 5.

- Tilastotieteellinen tutkimus tarkastelee reaalimaailman ilmiöitä. Täten tutkimuskohteena on tavallisessa elämässä tavattavia asioita, ihmisiä tai tapahtumia. Tutkimuskohteita kutsutaan **tilastoyksiköiksi** ja niiden joukkoa kutsutaan **populaatioksi (perusjoukoksi)**.

**Esimerkki: vaalitutkimus:**

- Politiikan tutkimuksen alalla yksi mielenkiintoinen tutkimuskohde on tutkia kuntavaaleissa äänestävien ihmisten tulova.
- Tällöin jokainen äänioikeuttaan käyttävä muodostaa oman tilastoyksikkönsä.
- Vastaavasti populaationa (perusjoukkona) (ks. alla) toimii kaikki äänestysikäiset kansalaiset, jotka äänioikeuttaan käyttävät.
- Pohdi: miksi pelkästään äänioikeuttaan käyttävien tutkiminen saattaisi olla tutkimuksen tulosten luotettavuuden kannalta ongelmallista?
- Toinen tutkimuskysymys voisi käsitellä kuntien välistä äänestysaktiivisuutta.
  - Tällöin jokainen kunta muodostaa oman tilastoyksikkönsä ja vastaa vakiin kaikki Suomen kunnat muodostavat populaation.
  - Kuntien äänestysaktiivisuus saadaan kuitenkin tutkimalla kunnan sisäistä äänestysaktiivisuutta.
    - \* Toisin sanoen, voidaksemme mitata kuntien äänestysaktiivisuutta, tulee ensiksi selvittää kuntien äänestysikäiset kansalaiset ja äänioikeuttaan käyttävät.

**Populaatio**

Konkreettinen tai hypoteettinen tutkimuskohteiden joukko, joka koostuu kaikista tilastoyksiköistä

- Populaation muodostavilta tilastoyksiköiltä tarkastellaan niiden ominaisuuksia, eli **tilastollisia muuttuja**.
  - Edellissä 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 lisätietoja eli **taustamuuttuja** (näitä voisivat 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 välille.

**Tilastoyksikkö ja tilastollinen muuttuja**

Populaation muodostavilta tilastoyksiköiltä (populaation alkioilta) tarkastellaan tilastollisia muuttujia, 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.

- Kerättyjen havaintojen joukko muodostaa **havaintoaineiston**, eli **datan**.

**Havaintoaineisto/data**

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

Tiivistettynä:

- Populaatio koostuu tutkimuksen kohteena olevista tilastoyksiköistä.
- Havaitaan tilastoyksiköistä tutkimuksen kannalta mielenkiintoisia tilastollisten muuttujien numeerisia arvoja.
- Nämä havainnot muodostavat havaintoaineiston, eli datan, jota voidaan käyttää tutkimuksessa ja tutkia **populaation ominaisuuksia**.

## 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>.

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

- *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>.
- *Vale, emävale, tilasto*<sup>3</sup>.
- *Statistics concerns what can be learned from data*<sup>4</sup>.
- *“Maalaisjärjen tehostamista”*<sup>5</sup>.
- Tilastotiede siis **kehittää ja soveltaa menetelmiä ja (tilastollisia) mallia**, joiden avulla reaalimaailman ilmiöstä 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 ilmiöstä voidaan tehdä päätelmiä.
    - \* Vrt. signaalin ja kohinan erottaminen (ks. Silver, 2014)<sup>6</sup>.
  - Tilastolliset mallit perustuvat todennäköisyyslaskentaan ja niillä mallinnetaan reaalielämän ilmiöiden alla pileviä prosesseja tai mekanismeja. Näiden prosessien tuottamia tietoja (aineistoja) tiivistetään usein graafisiksi esityksiksi ja tunnusluvuksi sekä tilastollisten mallien parametreiksi, joiden pohjalta johtopäätöksiä tehdään.
  - Tässä onnistuakseen tilastollisten menetelmien tulekin pyrkiä 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!

### Mitä tilastotiede ei ole

---

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

<sup>3</sup>Mark Twain popularisoi tämän lausahduksen teoksessaan *Chapters from My Autobiography* jo vuonna 1907. Huomionarvoista toki on, että valtaosa “modernin” tilastotieteen, jolle nykytilastotiede pohjautuu, teoriakehityksestä on tapahtunut vasta Twainin teoksen julkaisun jälkeen. Esimerkiksi Ronald Fisher, jota pidetään modernin tilastotieteen isänä, julkaisi merkityksellisimmät työnsä vasta 1920- ja 30-lukujen aikana. Tällä lentävällä lausahduksella ei siis ole mitään tekemistä nykyisten tilastollisten menetelmien kanssa.

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

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

<sup>6</sup>Silver, N. (2014). Signaali ja kohina: Miksi monet ennusteet epäonnistuvat mutta jotkin eivät? Terra Cognita. (Suomentanut Kimmo Pietiläinen)

- **Tilastotiede ei ole vain tilastojen tuotantoa**

– Vaikka sana **tilasto** tuo useimmille ensimmäisenä mieleen yhteiskuntaa ja sen toimintaa kuvaavat **numeeristen tietojen järjestelmälliset kokoelmat**, tilastotiede ei suinkaan ole ainoastaan tilastojen ja niiden tekemisen oppia.

\* Tämä siitäkin huolimatta, että niiden menetelmien konstruointi, joilla näitä tilastoja 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 käyttöalue on paljon tätä laajempi**.

\* Terminologiaa:

· Tilastoala = Tilastotiede + Tilastotoimi

· Tilastotiede = Teoreettinen tilastotiede + Soveltava tilastotiede

· Tilastotoimi = Tilastojen tuotanto + Tilastojen hyödyntäminen

• Tilastotieteen kannalta mikä tahansa reaalimaailman ilmiötä kuvaava **numeeristen tai kvantitatiivisten tietojen järjestelmällinen kokoelma** 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 tiellä tavalla aina eksaktia ja sen tulokset perustuvat formaaliin deduktioon ja loogisiin todistuksiin, johdaen usein “eksaktiin” ratkaisuun tai matemaattisesti formaaliin ratkaisun loogiseen esitystapaan.

- \* Tilastotiede sen sijaan on aina konteksti- ja aineistopohjaista ja perustuu induktiiviseen päättelyyn. Saadut tulokset ovat aina epävarmoja - koska ne kuvailevat epävarmaa tietoa generoivia prosesseja!
  - Tilastotiede on siis hyvä nähdä omana tieteenalanaan matemaattisesta 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 perustellusti oma tieteenalansa. Tämä erottelun vaikeus jo itsessään todistaa kuinka keskeinen rooli tilastotieteellä on eri aloilla!
  - Tilastotiede ei siis kuulu yksiselitteisesti sen lähitieteiden 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.

### Mitä tilastotiede (ainakin) on

- **Tilastotiede yleisenä menetelmätieteenä**
  - Tieteellistä tietoa ympäröivästä maailmasta hankitaan tieteellisillä **menetelmillä/metodeilla** (ks. tieteellisen menetelmän kriteerit luku 2)), joiden avulla tutkitaan jotain ilmiötä tai sen generoimaa kvantitatiivista mutta epävarmaa tietoa sisältävää aineistoa.
  - Tilastotieteessä kehitetyt ja kehitettävät menetelmät antavat tutki-jolle yhtenevät ja tiedeyhteisön hyväksymät raamit, jotka mahdol-listavat (tilastollisen) päättelyn ja päätöksenteon epävarman tiedon vallitessa. Näin voidaan uskottavasti ja luotettavasti tiivistää tietoa, jota erilaiset aineistot sisältävät, perustaa johtopäätöksiä näille tii-vistyksille ja saavuttaa uusia tieteellisiä löytöjä.
  - \* Tilastotieteen menetelmien käyttö ja soveltaminen onkin siis ai-na alakohtaista. Tästä huolimatta tilastollisia menetelmiä sovel-leataan aina johonkin **aineistoon!**

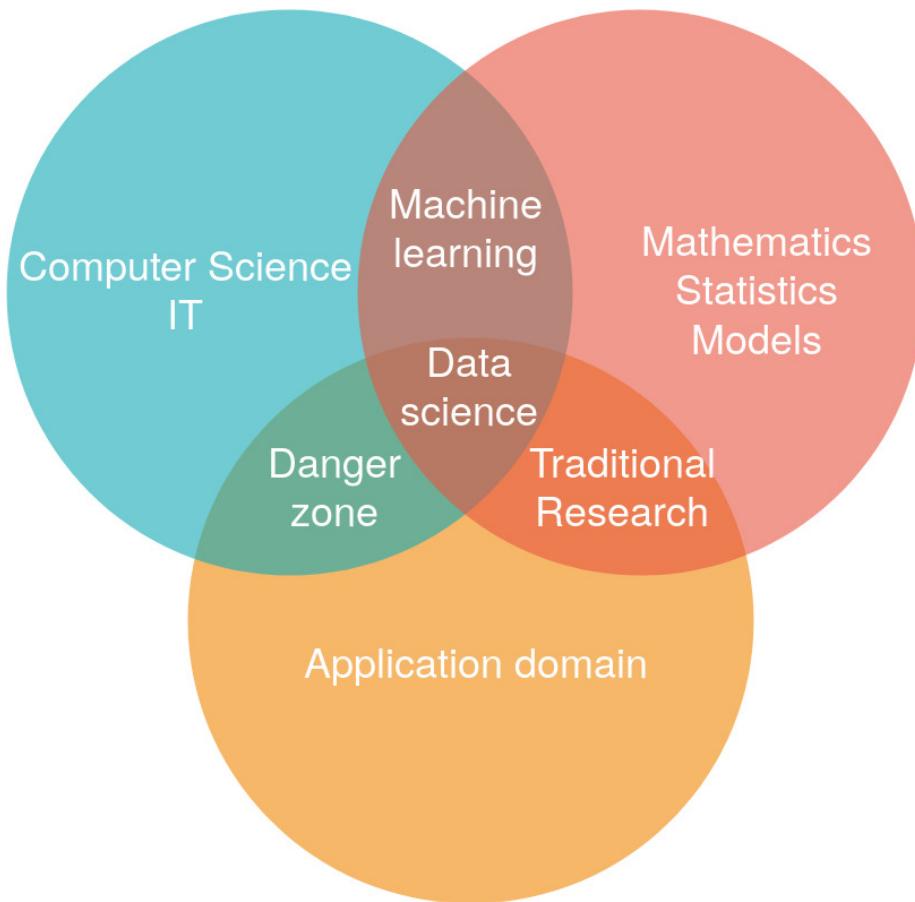
- Tilastotieteen nähdäänkin usein kuuluvan ns. **menetelmätieteisiin**, joissa mm.:
  - \* Kehitetään työkaluja muiden tieteiden tutkimusongelmien ratkaisuksi
  - \* On myös oma sovelluksista vapaa teorianmuodostuksensa
- Menetelmäkehityksen näkökulma tilastotieteeseen: *tilastotiede kehitää matemaattisia malleja satunnaisilmiötä kvaavia kvantitatiivisia tietoja generoiville prosesseille*. Koska tietoihin liittyy **epävarmuutta** tai **satunnaisuutta**, **tilastolliset mallit** perustuvat **todennäköisyyslaskentaan**.
  - \* Juuri sattuman ja epävarmuuden huomioiminen tutkimusasetelmissa erottaa tilastotieteen muista menetelmätieteistä!
- Tilastollisia menetelmiä voidaan soveltaa tietojen keruun, jalostuksen ja analysoinnin jokaisessa vaiheessa. Päämäärenä on jalostaa tiedot muotoon, joka mahdollistaa tutkittavaa reaalimaailman ilmiötä koskevien johtopäättösten tekemisen käytettyjen menetelmien pohjalta, eli ns. **tilastollisen päätelyn**.
  - Tutkimuksessa on pystyttävä valitsemaan ja käyttämään menetelmiä, jotka antavat aineistosta vastauksia haluttuihin kysymyksiin. Tämä vaatii yhtä lailla sovellusalakohtaista osaamista (ns. **substanssiosaamista**) kuin myös kattavaa **menetelmäosaamista**.
- Tilastotieteessä lähtökohtana ja ratkaisevassa asemassa on siis aina jokin satunnaisilmiön generoima **aineisto**, josta haluamme oppia tai tietää lisää, kenties voidaksemme tehdä suuria yhteiskunnallisia päätöksiä sen pohjalta!
  - Tämä aineistokeskeisyys yhtäältä erottaa tilastotieteen rajatieteistään ja toisaalta tuo sen lähemmäksi niitä ja sovellusalojaan.
  - Aineistoa analysoidaan, kuvallaan ja mallinnetaan tilastollisin menetelmin, joiden kehittäminen on keskeinen osa tilastotiedettä.
  - Pelkkä menetelmien kehittäminen kuuluu pitkälti matemaattisen/-teoreettisen tilastotieteen osa-alueelle.
  - Pelkkä aineiston keskittyminen 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 välillä: uudet teoreettiset menetelmät vastaavat soveltavan tilastotieteen ongelmiin mutta herättävät aina uusia kysymyksiä, jotka palautuvat taas teoreettisen tilastotieteilijän 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 sovelluslakohtaiset ongelmat, joista useat palautuvat yleisemmälle tasolle teoreettisen tilastotieteen kehityspolkuihin.

### 3.3 Tilastotieteen suhde lähitieteisiin

- Kuvio 3.1<sup>7</sup> tarjoaa karkean yleistyksen tietojenkäsittelytieteen (Computer Science) ja sovellusalan (Application domain) sekä tilastotieteen (Statistics) ja matematiikan (Mathematics) välisistä yhteyksistä. On selvää että tilastotieteellä on paljon päälekäisyksiä lähitieteiden kanssa ja joskus näkeekin (huolimatta edellä tehdystä huomioista) että tilastotiede nipputaan yhteen matematiikan tai tietojenkäsittelytieteen kanssa.
- Yritetään siis vielä hahmotella tilastotieteen suhdetta sitä lähimpänä olevaan (soveltavaan) matematiikkaan.
  - Tilastotieteessä olennaisen otantateorian (Luku 5) voisi ajatella olevan matemaattisesti määritelty teoria, jossa myös on aineiston käsite, mutta se ei tee siitä vielä varsinaisesti tilastotiedettä.
  - Matematiikassa kuvataan ongelma ja esitetään se teorian muodossa, eli malli on *“parametreista havaintoihin”*.
  - Tilastotieteessä ongelma on käänneinen, edetään *“havainnoista parametreihin”*, mutta ongelman matemaattinen kuvaus vaaditaan ensin.
  - Tilastotiede esittää menetelmiä ja käsitteitä tämän käänneisen 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 menetelmiä eli metodeja, jotka mahdollistavat päättelyn/tiedon hankinnan puutteellisesta tai epävarmasta tiedosta.
- Ilmiöiden kuvaamiseen ja käyttäytymisen ennakoimiseen käytetään usein **mallia**. Mallit (matemaattiset/tilastolliset mallit) voidaan jakaa **deterministisiin** ja **stokastisiin** malleihin.

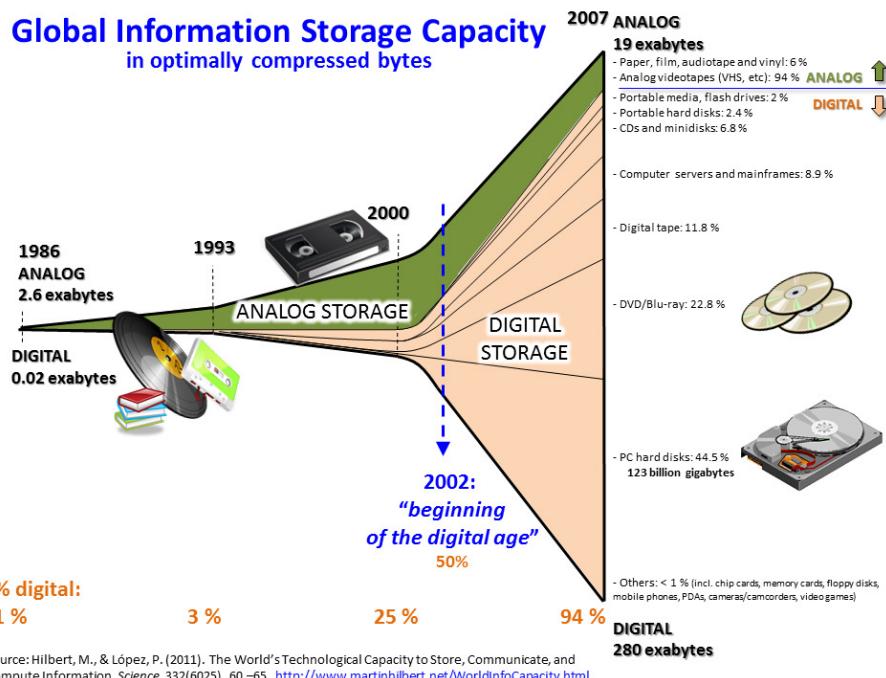
<sup>7</sup>Kuvan lähde: Duchesnay (2020)



Kuva 3.1: Tilastotieteen ja rajatieteiden yhteyksiä kuvaava Venn-diagrammi.  
(Duchesnay, 2020)

- Deterministisen mallin tapauksessa, tiettyjen alkuehtojen (alkuarvojen) vallitessa voidaan määrittää tarkasteltavan ilmiön lopputulos. Esimerkkejä ovat esim. monet fysiikan lait.
  - Stokastiset mallit perustuvat todennäköisyyslaskentaan. Stokastisia malleja käytetään kun alkuehtojen perusteella ei voida varmasti määrittää tarkasteltavan ilmiön lopputulosta. Tällöin eri vaihtoehtoihin liittyvät tiettyt esiintymistodennäköisydet. Esimerkkejä ovat esim. kolikonheitto tai sään ennustaminen.
  - Kun joitain ilmiötä kuvataan stokastisen mallin avulla, voidaan käyttää (joudutaan käyttämään) tilastollisia menetelmiä. Vaikka käytänössä laskenta hoidetaan tietokoneohjelmien avulla, meidän tilastotieteen tutkijoina ja käyttäjinä on huolehdittava tutkimusprosessin onnistuneesta toteutuksesta muita osin.
- 
- Tilastotiede ei myöskään ole puhtaasti tietojenkäsittelyä, vaikka tilastotiede onkin luonteeltaan aineistopohjaista ja aineistojen sisältämää tietoa on käsitelty osin samoin kuin tietojenkäsittelyssä siitä asti kun se on ollut mahdollista (tietokoneen keksimisen myötä).
    - Tilastotieteen ja tietojenkäsittelytieteen ero on lähitieteistä selvin: tilastotieteellä on “mekaanisesta” tai teoreettisesta tietojenkäsittelystä selkeästi erillinen ja oma teoriaapohjansa.
      - \* Siinä missä tilastotieteen teoria perustuu aineiston stokastiselle mallintamiselle, tietojenkäsittely on enemmänkin algoritmista ajattelua, missä aineistolla on ratkaisevalla tavalla erilainen rooli.
    - Lisäksi suomen kielessä tietojenkäsittely ymmärretään laajemmassa mielessä ohjelmoitavissa olevaksi automatisoimiseksi, jota tilastotiede ei perusolemukseltaan suinkaan ole.
  - Tarkastellaan seuraavaksi tilastotieteen suhdetta viime vuosien aikana paljon suosiota keränneeseen **datatieteeseen (data science)** johon voidaan katsoa lukeutuvan mm.
    - Tilastotiede ja matematiikka
      - \* Erityisesti tilastollinen data-analytiikka ja satunnaisen aineiston mallintaminen sekä soveltuvat soveltavan matematiikan osa-alueet.

- Tietojenkäsittely
  - \* Tietoteknologian kehityksen myötä taitavien tietojenkäsitteljöiden kysyntä on kasvanut merkittävästi. Lähes jokaisella alalla kerätään entistä enemmän dataa lähes kaikesta, ja jonkin pitääsi osata myös käsitellä sitä!
  - \* Datatieteen voidaan osaltaan katsoa syntyneen tästä elinkeinoelämän tarpeesta asiantuntijoille, jotka osaavat käsitellä suuria tietoaineistoja (dataa) sekä mallintaa niitä hyödyllisellä tavalla.
- Sovellusala
  - \* Datatiede on luonteeltaan pääosin soveltaavaa ja sen alaan lukeutuvia menetelmiä sovelletaan aina johonkin tosielämän ongelmaan. Tästä syystä substanssiosaaminen sovellusalalta on datatieteilijälle erityisen tärkeää ja nykypäivänä datatieteilijän rooli onkin pirstaloitunut yhä enemmän eri sovellusalojen datatieteisiin.
  - \* Tästä huolimatta datatieteilijöiden käyttämät mallinnusmenetelmät ovat usein varsin samanlaisia, sillä ne pohjautuvat edelleen tilastotieteen ja matematiikan teoriapohjaan. Ilman jälkimmäisten riittävää osaamista, liikutaan datatieteen osalta vaarallisilla vesillä! (Ks. oheinen kuva ja keskustelu alla).
- Datatieteellä ei usein nähdä olevan omaa historiallisen tieteellisen prosessin luomaa teoriapohjaa vaan sen voidaan katsoa olevan kokoelma eri alojen tieteellisiä menetelmiä ja tuloksia, jotka voidaan yhdistää tavalla, jonka “datavallankumous” (ks. kuva 3.2) mahdollistaa ja jotka ovat keskeisessä roolissa dataintensiivisissä sovellutuksissa.
- “Danger zone”
  - Kuvan 3.1 “danger zone” ([Duchesnay, 2020](#)) kuvailee tilannetta, jossa ilmiöiden/mallien tilastotieteellinen perusta unohdetaan.
  - Tilastotieteen näkökulman ohittava (laiminlyövä) soveltaja ei aina kykene suhtautumaan kriittisesti muodostuvaa ennustemallia, tai ennustetulosta, kohtaan eikä tätä päädy parhaisiin mahdollisiin (tarkimpiin) ennustetuloksiin tilanteessa, jossa jokin toinen malli kuvaisi ilmiötä annettua mallia paremmin.
  - Ko. soveltaja ottaa mallin sekä sen antaman ennustetuloksen annettuna, eikä mietti *mistä kyseinen ennustetus 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 pa-



Kuva 3.2: Datavallankumous (Hilbert, M. ja Lopez, P. (2011) The Worlds Technological Capacity to Store, Communicate and Compute Information. *Science*, 332(6025), 60-65.

remmin kuin tietojenkäsittelijä. Vastaavasti tehokkaan/onnistuneen ohjelmointikoodin kirjoittamisessa tilanne on usein toisinpäin.

## 3.4 Tilastotieteen osa-alueet

- 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 ja tutkimusalojen erilaisuudesta johtuen myös tilastotiede on kehittynyt vastaamaan monipuolisesti erilaisiin menetelmällisiin ongelmuihin!
  - Tämä on johtanut osaltaan siihen, että 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 alituisessa vuoropuhelussa.

### Soveltava tilastotiede

#### 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 käyttäminen numeeriseen aineistoon on kvantitatiiviselle tutkimukselle ominaisin piirre.
    - \* Perustuu yleensä satunnaisotokseen (kts. luvut 4, 5 ja 6) ja tutkimusaineisto on tiivistetty numeeriseksi havaintomatriisiksi, jolle oleellinen vaatimus on sen totuudellisuus.
    - \* **Kriitikki:** määrällinen tutkimus on (paikoin) sokea tutkittavien ilmiöiden sellaiselle luonteeelle, jota ei pystytä kvantifioimaan, eli

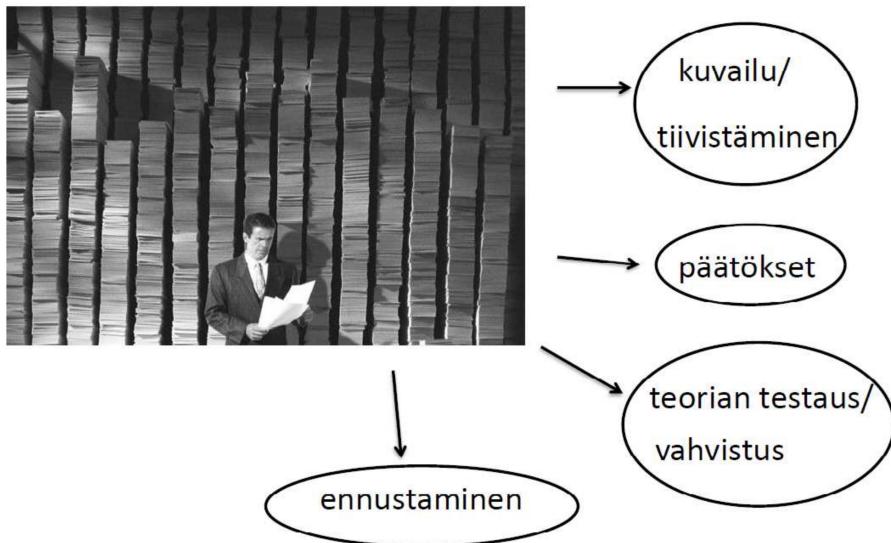
muuntamaan numeeriseen muotoon. Näihin voidaan katsoa lu-keutuvan mm. tunteet, merkitykset ja kokemukset, ellei tutkija keksi niiden numeeriselle mittaamiselle uskottavaa keinoa.

- **Kvalitatiivinen** eli laadullinen tutkimus on tutkimusta, jossa tutkimuksen kohteena olevaa ilmiötä ja sen merkitystä sekä tarkoitusta pyritään **ymmärtämään** kokonaivaltaisella 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 sillä tyypillistä on hypoteesittomuus, eli tutkimus on tarkoitus aloittaa mahdollisimman vähin ennakkooletuksin. Ennakko-oletuksista on kuitenkin mahdotonta täysin irtautua, joten niiden ilmi tuominen esioletuksina tai ”tutkimushypoteeseina” eli arvauksina tuloksista on osa tutkimusta.
  - \* Kriittiä: laadullinen tutkimus ei pysty vastaamaan kysymykseen miksi, sillä ilman määrällisiä (numeraalisia) aineistoja ei ilmiöiden välisiä riippuvuuksia kyötä tutkimaan: **laadullisessa tutkimuksessa menetetäänkin mahdollisuus tutkia ilmiöiden todellisia syitä**.
    - Laadullinen tutkimus nähdään usein vähemmän objektiiviseksi ja sen otosta koskevia tuloksia ei useinkaan voida yleistää koskemaan perusjoukkoa.

- **Yleisenä menetelmätieteenä tilastotiedettä voidaan (ja myös pitääsi) soveltaa kaikilla reaalimailmaa tutkivilla tieteenaloilla, joiden tutkimusaineistot voidaan esittää kvantitatiivisessa muodossa.**
  - Tilastollisten menetelmien käyttö on siis huomattavan paljon yleisempää määrällisessä kuin laadullisessa tutkimuksessa.
- Menetelmien soveltamisen tarkoituksena on (voi olla): **i) kuvalla ja tii-vistää tietoa**, jota havaittu aineisto sisältää **ii) sovellusalan oman teorian empiirinen testaus** tai **iii) edellisten pohjalta tehtävä tilastollinen päätely**.
  - **Deskriktiivisellä eli kuvalevalla tilastotieteellä** tarkoitetaan sellaisten menetelmien soveltamista, joiden avulla havaintoaineistos-

ta voidaan esimerkiksi laskea tunnuslukuja, kuvata havaintomuuttujien jakaumia ja visualisoida aineiston generoimaa ilmiötä tai siitä johdettuja tunnuslukuja.

- **Tilastollinen päättely** on sen sijaan aineiston tarkasteluun/kuvailuun sekä mallintamiseen perustuva päätöksentekoa, jossa kvantitatiiviseen aineistoon kuuluva epävarmuus ja satunnaisuus on otettu huomioon.
  - \* Keskeinen tilastollisen päättelyn käyttötarkoitus soveltajille on usein **teorian ja siihen liittävien hypoteesien testaaminen**, joka voi johtaa joko teorian vahvistumiseen (*verifiointiin*) tai sen vääräksi osoittamiseen (*falsifiointiin*) (ks. luku 2.1).
  - \* On myös syytä muistaa, että yksi tutkimus ei vielä osoita teoriaa oikeaksi tai vääräksi vaan siihen tarvitaan useita tutkimuksia sekä erilaisia tutkimusasetelmia ja -menetelmiä.
- Kuvaleva tilastotiede ja tilastollinen päättely kulkevat soveltuvaltaan tilastollisessa tutkimuksessa käsi kädessä.



Kuva 3.3: Soveltava tilastotiede

### Teoreettinen tilastotiede

#### Teoreettinen tilastotiede

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

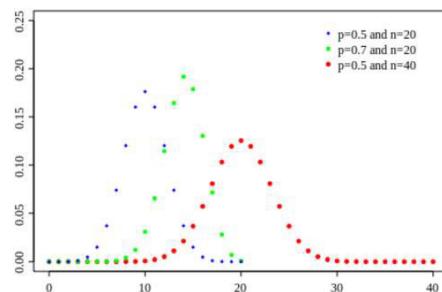
- Teoreettinen tilastotiede luo pohjan tilastollisten menetelmien ymmärtämiselle, soveltamiselle ja kehittämiselle.
  - Ilman riittävää ymmärrystä tilastollisten menetelmien toimintaperiaatteista niiden soveltaja on vaarassa tehdä virhepäätelmiä! (ks. ala-luku 3.5 tilastotieteen kriitikistä)
- Mallit perustuvat todennäköisyyslaskentaan, ja niitä kutsutaan tilastollisiksi malleiksi, stokastisiksi malleiksi tai todennäköisyysmalleiksi.
  - Tilastolliset mallit perustuvat laajalti niin kutsuttuun uskottavuusfunktioon. Se on malli, joka riippuu havaintoaineiston lisäksi yhdestä tai useammasta parametrista. (ks. tarkemmin luku 6)
  - Uskottavuusfunktion arvo kertoo kuinka todennäköisenä voidaan havaittaa aineistoa pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvoilla.
  - Uskottavuuspäättelyn perusajatuksena on, että se tai ne parametriarvat, joilla uskottavuusfunktion arvo maksimoituu kuvaa aineiston generoinutta prosessia parhaiten.
  - Aineistoa koskevia hypoteeseja voidaan testata käyttäen uskottavuusfunktion maksimia vastaavaa tilastollista mallia!
  - “*Kaikki mallit ovat vääräitä, mutta jotkut ovat käyttökelpoisia.*” (Box, 1976).
- Uskottavuusfunktiot perustuvat aina satunnaisilmiöiden mahdollisia arvoja kuvaaviin nk. **tiheysfunktioihin** tai **pistetodennäköisyysfunktioihin**.
  - Nämä funktiot kuvaavat jonkin satunnaismuuttujan (satunnaisilmiön) saamien arvojen jakaumaa.
  - Esimerkiksi kolikonheitto on satunnaisilmiö ja sillä on vain kaksi arvoa<sup>8</sup> ja kolikonheittoa voidaan kuvata nk. binomijakaumalla, jota merkitään  $\text{Bin}(n, p)$ , jossa  $n$  on heittojen lukumäärä ja  $p$  on kruunun todennäköisyys.

<sup>8</sup>Kolikon kantilleen jäämistä ei tässä lasketa mahdolliseksi tapahtumaksi.

**Esimerkki: kolikonheitto:**

- Eräs klassinen yksinkertainen todennäköisyyslaskennassa ja tilastotieteessä käytettävä esimerkki käsittelee kolikonheittoa.
- Kuvitellaan että olemme heittäneet kolikkoa 40 kertaa ja saatu kruuna 40/40 tapauksessa.
  - Kolikonheittoa seuranneet havainnot muodostavat nyt havaintoaineiston, jonka pohjalta voidaan perustellusti kysyä, että onko uskottavaa että kolikonheitto noudattaa binomijakaumaa  $\text{Bin}(40, 0.5)$ ?
  - Toisin sanoen, kuinka uskottavana voidaan pitää sitä että kyseinen kolikkon on tavallinen, painottamaton kolikko?

**Tilastotiede perustuu uskottavuuksiin, jotka taas perustuvat todennäköisyyteen ja tiheysfunktioihin.**



Kuva 3.4: Tilastotiede ja todennäköisyys

- **Todennäköisyyslaskenta** luo tilastotieteelliselle epävarmuuden mallintamiselle vahvan ja uskottavan matemaattisen perustan.
  - Todennäköisyyslaskentaa opetetaan tarkemmin (tätä kurssia seuraavilla) kursseilla **TILM3553 Todennäköisyyslaskennan peruskurssi pääaineopiskelijoille**, **TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille** ja **SMAT5306 Todennäköisyyslaskennan jatkokurssi**.

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

Pohja tilastollisten menetelmien ymmärtämiselle, soveltamiselle ja kehittämiselle

Kuva 3.5: Esimerkki teoreettisesta tilastotieteestä ja tilastollisesta päättelystä.

### 3.5 Tilastotieteen kritiikkiä

- Tilastotieteen rooli tiedeyhteisössä on niin tärkeä että sitä kohtaan on ymmärrättävästi esitetty myös paljon kritiikkiä. Valtaosa kritiikistä kohdistuu joko tilastotieteen matemaattisuuteen tai siten siinä tarvittaviin oletuksiin, jotka mahdollistavat esimerkiksi hypoteesien testaamisen.
  - Usein kritiikki on aiheetonta ja johtuu sen esittäjän puutteellisesta tilastotieteen ymmärryksestä. Perusteettoman kritiikin esittäminen toista tieteenalaa kohtaa ei kuitenkaan ole vieraas ilmiö juuri millään alalla.
- Tässä alaluvussa käymme läpi yleisimpiä kritiikin muotoja, joita tilastotiedettä kohtaan esitetään ja pyrimme tarjoamaan vastauksia/vastineita silloin kun niitä voidaan antaa.

#### “Yleismaailmallinen” kritiikki

- Aloitetaan yleismaailmallisella kritiikkillä, jota tilastollista tutkimusta vastaan on esitetty:
  - Tilastotieteessä käytettävien tunnuslukujen, kuten keskiarvon, reaalimaailman vastineet ovat joskus mielivaltaisia. Esimerkiksi keskiarvo

on ajoittain ongelmallinen tunnusluku, sillä lienee varsin selvää, että keskimääräistä ihmistä ei ole olemassa vaikka tilastotieteessä näitä tunnuslukuja usein lasketaankin.

- \* Esimerkiksi puhekielessä yleinen nk. "Keskiarvo-Kalle", eli 1,8 lapsen vanhempi ja 1,5 auton omistaja on tietenkin täysin kuvitteellinen.
- \* Lisäksi joskus kuulee tilastotieteilijöitä kritisoidavan lausumalla "*Jos toinen jalka on jääkylmässä vedessä ja toinen kiehuvassa vedessä, niin tilastotieteilijän mielestä ihmisen tällöin keskimäärin hyvä olla*"

- **Korrelaatio on tunnusluku**, joka kuvaaa kahden muuttujan välistä riippuvuutta (palaamme tähän tarkemmin luvussa 7). Se ei kuitenkaan kuva millään tavoin **kausaalisuutta**, eli sitä kumpi aiheuttaa kumman, jos kumpikaan.<sup>9</sup>
  - Esimerkiksi "jäätelön syönti ja hukkumiskuolemat" -tapauksessa havainnollisesti todetaan jäätelönlukulukseen ja hukkumiskuolemien lukumäärän korreloivan keskenään, mutta taustalla vaikuttava tekijä onkin lämmin kesä, joka vaikuttaa molempia.
- Vaikkei näiden esimerkkien oikeellisuutta ole syytä kiistää, niin tilastollisen tiedon arvioinnissa on kuitenkin syytä päästää syvemmälle.

### Kritiikki matemaattisuutta kohtaan

- Ehkä merkittävin kritiikki tilastollisia menetelmiä kohtaan kohdistuu kritiikan näkökulmasta perusteettomaan, tai ainakin liian vahvaan, matemaattisuuden tuomaan itsevarmuuteen. Vidaankin siis perustellusti kyseä, että **onko tieteellisyys = matemaattisuus?**
  - Useat tieteenalat käyttävät tutkimuksessaan edistyneitakin tilastollisia menetelmiä siitä huolimatta, että tutkijoiden tilastomatematisen pohjakoulutus ei välittämättä ole riittävällä tasolla kyseisten menetelmien kokonaisvaltaiseen ymmärtämiseen.
    - \* Helppokäytöisistä tilasto-ohjelmistoista on riittävät perustaidot omaaville käyttäjille erittäin paljon hyötyä mutta koneiden ja ohjelmien käytön opettelu ei kuitenkaan ole varsinaista tilastotiedettä (tarvitaan enemmän tilastotieteen opintoja).

---

<sup>9</sup>Tyler Vigen on kerännyt [verkkosivulleen](#) (ks. [linkki](#)) mitä moninaisimpia esimerkkejä kahdenvälisistä nk. *näennäisistä* korrelaatioista.

- \* Laskentatehon ja modernin tietojenkäsittelyteknologian ansiosta monimutkaisiakin tilastollisia analyysejä on kuitenkin mahdollista tehdä vaikka tutkijalla olisi tilastotieteestä vain perustiedot, jos sitäkään.
- \* Pahimillaan tämä saattaa johtaa siihen, että analyyseja tehdään ymmärtämättä mitä itse asiassa ollaan tekemässä.
- Tilastollisten analyysien hyödyllisyyden ja järkevyyden ehtona on kuitenkin käytettävien menetelmien, aineiston ja tutkittavan ilmiön pintaa syvemmälle ulottuva tuntemus.
  - \* Käytettävien tilastollisten menetelmien oletukset on osattava ottaa huomioon ja toisaalta odottamattomien tulosten syyt on pysyttää jäljittämään.
    - Teknistä esitystä käyttää tutkijaa saatetaan pitää erityisen uskottavana, koska hän kykenee käyttämään vaikeita menetelmiä. Tästä huolimatta tutkimusongelma ei saisi päästää unohtumaan.
    - Tutkijan tulisikin varmistua siitä, että käytettävät menetelmät todella vastaavat asetettuihin tutkimuskysymyksiin ja että tutkimusongelma on ratkaistavissa käytettävillä menetelmillä.
    - Tekninen esitys ei takaa onnistunutta tilastollista tutkimusta eri näkökulmista katsoen. Monet tilastolliset menetelmät ovat vaikeita ja vaativat soveltajiltaan paljon.
    - Lisäksi on hyvä muistaa, että käytettävien menetelmien lähtökohdat ja oletukset eivät matemaattisudestaan huolimatta ole välttämättä neutraaleja!
  - \* Kaikkia tieteentekijöitä ei voida velvoittaa opiskelemaan edistynytä abstraktia tilastotieteen teoriaa (tilastomatematiikkaa), mutta menetelmien oikeaoppinen käyttö kuitenkin vaatii riittävää ymmärrystä.

### Kritiikki yksinkertaistuksia kohtaan

- Edellisiä kohtia yleisemmin tilastotiedettä on kritisoidu siitä, että se ei kykene riittävällä tasolla huomioimaan reaalimaailman kompleksisuutta.
  - Merkittävässä osassa tilastollisia analyyseja lähtökohtana on usko ”todellisen” maailman ja näin ollen aineistoa generoivien mekanismien olemassaoloon.
    - \* Tätä saatetaan usein pitää kuitenkin kyseenalaisena: voiko ”toisielämän stokastiikasta” muka todella löytyä säännönmukaisuuksia?

- \* Tämä kysymys on kuitenkin pitkälti tieteenfilosofinen ja palautuu lopulta sovellusalaan sekä tutkimusongelmaan ja -kysymykseen: tilastollisten menetelmien toimivuutta voidaan helposti testata esimerkiksi simulaatiokokeilla.
- Tilastotiedettä on myös kritisoidu sen “sokeudesta” sosiaaliseen vuorovaikutukseen liittyviin subjektiivisiin kokemuksiin kuten tunteisiin, kokemuksiin ja ei-numeeriisiin havaintoihin.
  - \* Tämä kritiikki ei kuitenkaan suoranaisesti ole tilastotieteen kritiikkiä, vaan jälleen sovellusalakohtainen ja erityisesti tutkimuskysymyksien asettelua koskeva ongelma.
    - Tuntemuksia ja kokemuksia voidaan hyvin testata tilastollisen menetelman, mikäli tutkija osaa uskottavasti määritellä niille numeerisen mittauksen kriteeristöt!
    - Tämä on kuitenkin vaikeaa, sillä aivan kaikkea ei voida kvantifioida: kirjoitetun tekstin tai sosiaalisten merkitysten tulkinnan sekä elämysten kuten musiikin ja taiteen aiheuttamien mielikuvien ja tunteiden voidaan perustellusti nähdä olevan hyvin haastavia kvantifioida.
  - \* Näiden aiheiden tulkinta, ymmärtäminen ja tutkiminen ulottuu kvantitatiivisen tutkimuksen ulkopuolelle.
- Mikäli tutkittavasta ilmiöstä pystyy kvantitatiivilla mittauksilla saada relevanttia tietoa, tulisi aineiston analyysin apuna joka tapauksessa aina käyttää tilastollisia menetelmiä!
- Vaikka kvantitatiivisia aineistoja ei voi pitää objektiivisina faktoina asioiden tilasta, se ei tarkoita, etteivätkö tulokset voisi olla käytökeloisia.

### Temppukokoelmakritiikki

- Eräs ehkä osin implisiittinen kritiikki tilastotiedettä kohtaan on sen pitäminen nk. **“temppukokoelmana”**.
  - Tilastotieteen voi nähdä koostuvan numeeristen tietojen jalostamisen menetelmistä. Tämä näkemys, joka on usein tahaton, pelkistää tilastotieteen *vain menetelmäkokoelmaksi*, vailla omaa teoriaa.
  - Eri tutkimusalojen empiirisessä työssä (liian) usein vain kerätään aineisto ja vasta sitten mietitään mitä sillä voitaisiin tehdä.
  - Usein apuun haetaan tilastotieteilijä, jonka odotetaan loihtivan (tilastollisen) ratkaisun ongelmaan kuin ongelmaan.
  - \* Joskus tämä toki onnistuukin, mutta useimmiten ei.

- \* Tilastotiede ei siis ole ”työkalupakki”, josta valitsemalla oikeanlainen menetelmän voi vastata mihin tahansa tutkimuskysymykseen!
- Tilastolliset menetelmät tulee ymmärtää ja niitä tulee soveltaa kaikesta soveltavan tutkimuksen vaiheissa, jotta tutkimusongelmaan kyetään vastaamaan eikä turhaa työtä tule tehdynksi.
- Karkeasti luokitellen tilastotieteilijät kehittävät menetelmiä, joita soveltajat käyttävät.
  - \* Soveltavia tilastotieteilijöitä löytyy kuitenkin yhä kiihtyvissä määrin! Erityisesti eri rajatieteiden alueilla, kuten alaluvussa [3.6](#) lyhyesti esitellään.

### Tilastotieteen väärinkäytö

- Tilastotiedettä on myös mahdollista käyttää väärin monin eri tavoin, joka edelleen altistaa koko tieteenalan (perusteettomalle) kritiikille!
  - Tilastoja ja tilastotiedettä käytetään paljon väärin, mutta tämä on usein tahatonta (esim. puutteellisesta koulutuksesta johtuva).
  - \* Joskus kuitenkin näkee tarkoituksellista tilastojen vääristelyä tai tahallista tilastollisten menetelmien väärinkäytöä!
  - \* Kansalaisten tiedelukutaidon ja tilastollisten menetelmien tuntemuksen merkitys on kasvanut viime vuosikymmeninä ja kasvanee jatkossa yhä, kun esimerkiksi erilaiset ”vaihtoehtotieteet” ovat nousseet suosittumiaksi.
  - \* Tilastotieteen ymmärrys auttaa itse kutakin tunnistamaan virheellisiä tai puutteellisia tiedoja tehtyjä päätelmiä ja täten helpottaa tietoyhteiskunnassa toimimista ja kriittistä ajattelua!
- Yleisiä tilastollisten menetelmien väärinkäytötapoja ovat esimerkiksi seuraavat:
  - **“Kolmannen tyypin virhe”**: kun tilastollisia menetelmiä käytetään saadaan oikeita vastauksia, mutta väärin kysymyksiin! Esimerkiksi jos tutkija ei täysin ymmärrä minkälaisia vastauksia käytetävissä olevasta aineistosta ja valitulla menetelmällä voidaan saada, voi hän syyllistyä kolmannen tyypin virheeseen. Tällöin voi nimittäin käydä niin, että hän tulkitsee tilastolliset testit täysin oikein, mutta huilee väärin niiden vastaavaan eri kysymykseen kuin on esitetty.
  - Black-box ilmiö: saadaan *ehkä* oikeita vastauksia, mutta ei tiedetä *miksi* ja *mihin* kysymyksiin.
    - \* Totaalinen tilastollisen päättelyn osaamattomuus saattaa johtaa tutkijan täysin väärille urille ja esimerkiksi jokseenkin epäoleelliseen tekniseen näpertelyyn monimutkaisten mallien kanssa.

**Esimerkki: Kolmannen tyypin virhe**

- Oletetaan että haluat tutkia onko kahden eri ikäryhmän ihmisten pituuksissa eroja ja sinulla on käytettäväissä edustava otos molempien ikäluokkien edustajista.
- Pääät tutkia *yksisuuntaisesti* onko toisen ryhmän, ryhmän A, keskipituus *pienempi* kuin ryhmän B.
  - Testitulos osoittaa, että voit hylätä nollahypoteesin, jonka muukaan ryhmien *keskipituus olisi sama*.
  - Kolmannen tyypin virhe syntyy silloin, jos tosiasiallisesti testin hylkääminen johtui siitä, että ryhmän A keskipituus olikin *suurempi* kuin ryhmän B keskipituus, mutta tästä et testin tuloksen perusteella voi tietää!

## 3.6 Tilastotieteen sovellusaloja ja “rajatieteitä”

- Yleisenä menetelmätieteenä tilastotiedettä sovelletaan useilla eri tieteenoilla.
  - Jokaisella sovellusalalla on oma erillinen teoriapohjansa sekä empiiriset käytänteet, joten substanssitetous on sovellettaessa erityisen tärkeää.
    - \* Huolimatta vaihtelevista empiirisistä käytännöistä sovellusmenetelmän taustalla on (lähes aina) kuitenkin tilastotieteen alalla kehitetty menetelmä.
    - \* Sovellusalilla ongelmanratkaisussa yhdistetäänkin metodiseen osaamiseen välttämättä myös substanssitetoutta. Tämän myötä soveltavan tilastollisen tutkimuksen kenttä on laaja ja rikas.
  - Osa näistä sovelluskentistä on kehittynyt vahvassa yhteisvaikutuksessa tilastotieteen ja lähitieteiden (viime aikoina erityisesti koneoppimisen) yhteydessä.
- Usein on pystyttävä arvioimaan ongelmanasettelun ja tulosten tarkoituksenmukaisuutta ja pyrkiä välttymään siltä että tutkijan tieteelliset ja yhteisölliset sitoumuukset heijastuisivat tutkimuksen kulkun.
- Tilastotieteen pääaineopiskelun osalta substanssitetous saavutetaan usein sivuaineopintojen perusteella. Vastaavasti toisinpäin muiden aineiden pääaineopiskelijoiden kohdalla, jolloin tilastotiete voi yhtä hyvin toimia (laajalti opiskeltuna) vahvana sivuaineena.

- Jokaisella tieteenalalla, jonka tutkimusaineistot voidaan esittää numeerisessa tai kvantitatiivisessa muodossa voi soveltaa/voisi soveltaa/pitäisi soveltaa tilastollisia menetelmiä sekä tutkimusaineistoja kerättäessä että niitä analysoitaessa.
  - Siten jokainen empirisen tutkimuksen havaintoaineisto on tilastollisen tutkimuksen mahdollinen kohde.
  - Esim. kokeellinen tutkimus käyttää apunaan tilastollisia menetelmiä.
- Koska tilastotieteellä on sovelluksensa miltei kaikilta tieteenhaaroilla, on syntynyt nk. "rajatieteitä":
  - Sovellusalojen, joilla tilastotieteen soveltaminen on muodostunut omaksi tutkimuskohteekseen/tieteenlajikseen (ks. linkit):
    - \* [Psykologia: psykometriikka](#),
    - \* [Sosiaalitieteet: sosiometria](#),
    - \* [Taloustiede: ekonometria](#),
    - \* [Kemia: kemometria](#),
    - \* [Bio- ja lääketiede: biometria](#),
    - \* [Epidemiologia](#),
  - Soveltavan matematiikan tutkimusalojen, jotka ovat osaltaan päälekkäisiä tilastotieteen kanssa
    - \* [Informaatioteoria](#),
    - \* [Matemaattinen tilastotiede](#),
    - \* [Todennäköisyyslaskenta](#),
    - \* [Operaatioanalyysi](#)
  - Tietojenkäsittelytieteen alaan (osittain) lukeutuvia tutkimusalajoja
    - \* [Laskennalliset menetelmät](#),
    - \* [Data mining](#),
    - \* [Knowledge discovery](#),
    - \* [Hahmontunnistus](#),
    - \* [Tekoäly](#),
    - \* [Koneoppiminen](#)
- Ja paljon muita!

### 3.7 Luvun 3 yhteenvetö, keskeisiä termejä ja konaisuuksia.

Tilastotieteen perustermejä: - Populaatio - Tilastoyksikkö ja tilastollinen muuttuja - Havainto - Havaintoaineisto eli data

Mitä tilastotiede on ja mitä se ei ole? - Tilastotieteen suhde lähitieteisiin, kuten matematiikkaan, tietojenkäsittelytieteeseen ja datatieteeseen.

### *3.7. LUVUN 3 YHTEENVETO, KESKEISIÄ TERMEJÄ JA KOKONAISSUUKSIA.55*

Tilastotieteen osa-alueet - Soveltava tilastotiede (ml. kvantitatiivinen ja kvalitatiivinen tutkimus) - Teoreettinen tilastotiede (tilastollisen mallin ajatus ja merkitys)

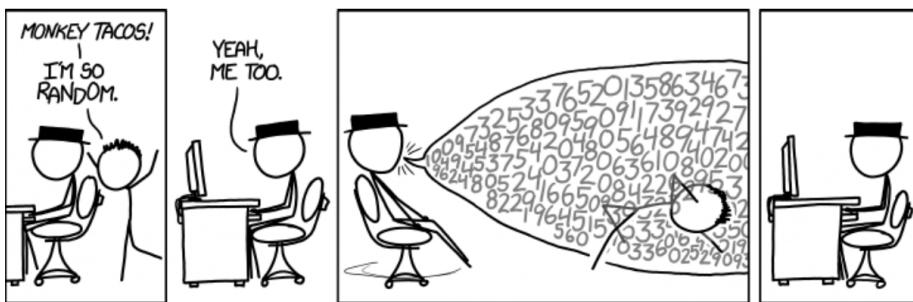
Tilastotieteen sovellusalat ja “rajatieteet” (ks. tarkemmin luentomoniste ja sieltä löytyvät linkit!)

Tilastotieteen kritiikki



## Luku 4

# Sattuma ja satunnaisuus tilastotieteessä



Kuva 4.1: Hauska kuva satunnaisuudesta.

Tässä luvussa pohdimme sattuman ja satunnaisuuden roolia tilastotieteessä ja tieteessä ylipäättäään. Satunnaisuudella tarkoitetaan yleensä säännönmukaisuuden puuttumista ja ennustamattomuutta ja kenties juuri siksi sitä voidaan pitää yhtenä maailman vaikuttavimmista ilmiöistä. Jokainen haluaisi tietää mitä tuleman pitää ja siksi sattuma tekee elämästä mielenkiintoista: se vaikuttaa ja muokkaa niin meitä itseämme kuin ympäröivää maailmaa mitä merkityksellisimmin tavoin - joskus jopa vasten tahtoamme ja usein vailla täytyy ymmärtystämme!

Ihmisen oma kokemus on kuitenkin altis kaikenlaisille virhepäätelmille, joita kutsutaan myös kognitiivisiksi vinoumiksi. Haluamme löytää systematiikkaa ja tarkoitusta kaaksesta sekä merkityksiä ja syy-seuraussuhteita sellaisista tapahtumista, jotka kuuluvat normaalivaihtelun piiriin. Tällaisissa tilanteissa usein tilastollinen tarkastelu paljastaakin ilmiön todellisen, alkuperäisestä kuvitelmasesta poikkeavan luonteen. Erotaakseen sistemaattinen vaihtelu satunnaisesta ja

ymmärtääkseen oikeasti merkityksellisiä syy-seuraussuhteita, satunnaisuutta on välttämätöntä ymmärtää. Tämä välttämättömyys päätee erityisesti tiedeyhteisön jäseniin, jotka pyrkivät tutkimaan ympäröivän maailman satunnaisia ilmiöitä. Tilastotiete perustuu satunnaisilmiöiden ja satunnaisen aineiston tutkimiseen, joten sen ymmärtäminen on keskeisessä roolissa niin tilastotieteen kuin muidenkin tieteiden sekä lopulta maailman ymmärtämisessä.

## 4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä

- Edellisestä luvusta muistamme, että tilastotieteellisen tutkimuksen kohdeena 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 **todennäköisyyslaskentaan**.

### Satunnaisilmiö

Reaalimailman ilmiö on satunnaisilmiö, jos seuraavat ehdot pätevät:

- Ilmiöllä on useita erilaisia tulosvaihtoehtoja.
- Sattuma määräää mikä tulosvaihtoehtoista toteutuu, eli yksittäistä tulosta ei voida tietää etukäteen.
- Vaikka tulos vaihtelee ilmiön toistuessa satunnaisesti, käytätyy tulosvaihtoehtojen suhteellisten osuksien jakauma tilastollisesti stabiiliisti ilmiön toistokertojen lukumäärän kasvaessa.

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

## 4.1. SATUNNAISILMIÖT JA SATUNNAISMUUTTUJAT TILASTOTIETEESSÄ 59

### Esimerkkejä satunnaisilmiöistä

- Helpoin esimerkki on uhkapelit, kuten kortti- ja noppapelit, arpajaiset, lotto tai ruletti: näitä käytetäänkin usein todennäköisyyslaskennan peruskursseilla satunnaisilmiöiden esittelyyn.
- Lukion biologian tunneilta muistetaan, että perinnöllisyykskin on osaltaan sattumaa: se määräää kummalta vanhemmalta perittävä geenikopio on peräisin.
  - Vastaavasti populaatiotasolla eri ominaisuuksien jakautuminen yksilöiden ja populaatioiden välillä on satunnaista.
  - Populaatiotaso voi tässä tarkoittaa esimerkiksi erilaisten eliöiden eri alueilla eläviä populaatioita, joiden välisiä eroja pyritään tutkimaan ja selittämään.
  - Vastaavasti ihmisten, ihmisyryhmien ja ihmisten muodostamien organisaatioiden sisäisessä ja välisessä käyttäytymisessä on useita satunnaisia elementtejä.
- Jopa deterministiseen toimintaperiaatteeseen tähtäävissä tehdastuotannossa käy satunnaisia virheitä tuotteiden valmistusprosesseissa, jotka ilmenevät esimerkiksi viallisina tuotteina.
- Vastaavasti luonnontieteellisiin mittauksiin liittyy mittausvirheitä, jotka kuuluvat satunnaisvaihtelun piiriin. Esimerkiksi varhaisissa valonnopeusmittauksissa mittausvirheet saattoivat olla suuriakin!
- Myös kvanttimekaniikan ja hiukkasfysiikan tutkimat ilmiöt ovat perusuonteeltaan satunnaisia.

### Satunnaismuuttujat

- Tilastollista vaihtelua ilmentävät tilastolliset muuttujat tulkitaan **satunnaismuuttujiksi** ja havainnot (havaintoarvot) voidaan näin ollen tulkitaan näiden satunnaismuuttujien realisoituneiksi arvoiksi. Tällöin tilastollisen tutkimuksen kohteena on nämä havainnot generoinut *satunnaisilmiö*.
  - Satunnaismuuttuja siis kuvailee tarkasteltavan mitattavan ominaisuuden (satunnais)vaihtelua tutkimuksen kohteiden, eli tilastoiksiiden joukossa.
  - Mitattavan ominaisuuden mahdolliset arvot määrääävät satunnaismuuttujan luonteen. Yleisesti satunnaismuuttujat jaetaan kahteen luokkaan: **jatkuihin** ja **diskreetteihin**.
  - Satunnaismuuttujan **todennäköisyysjakauma**, määräää erilaisten tulosvaihtojen todennäköisyden ja mahdollistaa 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 realisaatioksi tai toteumaksi.
- Tilastollinen aineisto muodostuu useiden satunnaismuuttujien (tilastoiksiköiden tutkimusmuuttujien) realisoituneista arvoista.
- Realisoituneiden arvojen vaihtelua tilastoiksiköiden välillä kutsutaan satunnaisvaihteluksi.

### Jatkuvat ja diskreetit satunnaismuuttujat

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

### Esimerkki: satunnaismuuttuja

- Ihmisen pituutta voidaan pitää (ennen mittaukseen tulemista) satunnaismuuttujana  $Y$  ja lopullista pituutta täten pituuden realisaationa  $y$ .
  - Yleensä pituutta kohdellaan jatkuvana muuttujana senttimetreissä.
  - Mikäli kuitenkin määritetään toteumaksi jonkin pituuden raja-arvon, esimerkiksi 170 cm, ylittävä pituus, on kysessä kaksiarvoinen (binäärisen) muuttuja (pituus on joko yli tai alle 170 cm).

- Muuttujat voidaan luokitella myös **kvalitatiivisiin** ja **kvantitatiivisiin** muuttuijiin (ks. @ref{alaluku53}).
  - Kvalitatiivisiin muuttuijiin liittyy luokittel- 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.
  - Yksinkertaisimillaan esimerkiksi yksinkertaiseen satunnaisotantaan takaisinpanolla perustuva satunnaismalli (palaamme tähän otantaa käsittelevässä luvussa 5).
  - Satunnaisuus perustuu siihen, että satunnaismuuttujien toteutuvat arvot (ja niistä lasketut tunnusluvut kuten keskiarvo) vaihtelevat satunnaisesti otoksesta toiseen.
- Todennäköisyyslaskennan ja tilastotieteen tehtävä on tuottaa **tilastollisia malleja** satunnaisilmiöissä havaittavalle tilastolliselle stabiliteetille.

## 4.2 Satunnaisuus ja todennäköisyyydet

- 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 laki ja.
  - Koska arvonnan tulos vaihtelee satunnaisesti arvontakerrasta toiseen, myös tutkimuksen kohteita kuvaavat tiedot vaihtelevat satunnaisesti arvontakerrasta toiseen.
  - Tutkimuksen kohteita kuvaavien tietojen käyttäytymisessä havaitaan kuitenkin arvontaa toistettaessa juuri sitä säännönmukaisuutta, jota kutsutaan tilastolliseksi stabiliteetiksi. **Tämä säännönmukaisuus on tilastollisen tutkimuksen kohde.**

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

- Esimerkkejä tilastollisten aineistojen keräämisen menetelmistä, jotka perustuvat arvontaan:
  - **Satunnaistetut kokeet:** Kokeellisessa tutkimuksessa tavoitteena on vertailla erilaisten käsittelyiden vaikuttuksia kokeen kohteisiin. Erilaisten virhelähteiden kontrolloimiseksi käsittelyt on syytä arpoa kohteille.
  - **Satunnaisotanta:** Otannalla<sup>2</sup> tarkoitetaan laveasti tutkimusaineistojen keräämisen menetelmiä. Erilaisten virhelähteiden kontrolloimiseksi tutkimuksen kohteet on syytä valita arpomalla. (Ks. Luku 5)
- Kerätyn (tai havaitun) aineiston pohjalta tehdään päätelmiä sen generoivesta satunnaisilmiöstä esimerkiksi testaamalla erilaisia siihen liittyviä hypoteeseja.
  - Tilastotiede voidaan jakaa kahteen merkittävään paradigmaan sen mukaan, miten **tilastolliseen päättelyyn**, ml. hypoteeseihin ja niiden testaamiseen, suhtaudutaan. Näitä ovat **klassinen eli frekventistinen tilastotiede** sekä **Bayesilainen tilastotiede**. Tarkasteluaan 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 liitettävään tilastolliseen malliin.
- Nimi “frekventistinen” juontuu siitä, että tilastollisen mallin perustana oleva todennäköisyysjakama määrittää satunnaismuuttujan mahdollisten arvojen todennäköisyystekijäniiden suhteellisen osuuden äärettömästä määristä realisaatioita, ts. niiden suhteellisen frekvenssin.
- Klassisessa tilastotieteessä havaittuun aineistoon *sovitetaan* tilastollinen malli, joka vastaa saattua aineistoa parhaiten.
  - Tämä tilastollinen malli voidaan (useimmiten) perustaa nk. **uskottavuusfunktioon**, joka on *aineiston* sekä yhden tai useaman *parametrin* funktio ja joka saavuttaa suurimman arvonsa nk. “suurimman uskottavuuden pisteessä”.
  - Uskottavuusfunktio kertoo kuinka todennäköisenä havaittua aineistoa voidaan pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvolla.

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

- \* Täten ne parametriarvot, joilla uskottavuusfunktion arvo maksimoituu, *kuvavat aineiston generoimaa prosessia parhaiten*, annettuna malli- eli jakaumaoletus.
- Uskottavuusfunktioista, tilastollisten mallien estimoinnista ja parametreista lisää seuraavassa alaluvussa sekä luvussa 6.
- Perusjoukkoa koskevia hypoteeseja testataan tilastollisen mallin avulla: havaittu aineisto määrittää uskottavuusfunktion perusteella sellaiset hypoteesit, jotka jäävät joko voimaan tai tulevat hylätyiksi.
- Klassisessa tilastotieteessä hypoteesien testaus perustuu siis vain aineistoon eli tilastollinen päättely on induktiivista: aineiston avulla otosta koskeva päätelmä voidaan yleistää koskemaan perusjoukkoa.
  - Toki kaikki päättely on alisteista tehdyille oletuksille koskien käytetään tilastollista mallia.

### **Bayesilainen tilastotiede**

- Bayesilainen tilastotiede on tilastotieteen toinen suuri paradigma ja on saanut nimensä englantilaiselta harrastelijamatemaatikko ja presbyteeri-pappi **Thomas Bayesilta**, jota pidetään Bayesilaisen tilastotieteent isänä.
- Bayesilainen tilastotiede ulottaa todennäköisyyskäsityksen, eli tajauksen, myös aineistoa koskevien hypoteesien puollelle: kuinka todennäköisenä jotain hypoteesia voidaan pitää jo ennen tutkimusaineiston keräämistä?
  - Myös Bayesilaisessa tilastotieteessä hyödynnetään uskottavuusfunktioita, mutta hypoteesien testaus ei perustu niinkään frekventistiseen ajatukseen todennäköisyysistä suhteellisina osuuksina äärettömässä sarjassa.
  - Bayesilaiset perustavat sen sijaan hypoteesien testaamisen tutkimuskysymystä koskevien ennakkokäsitysten päivittämiselle sen jälkeen, kun aineisto 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 kritiikkinä kuin puolustuksena.
  - Lopulta olemme kaikki Bayesilaisia: jokaisella on sisäisiä ennakkokäsityksiä, myös tutkijoilla! Nämä ennakkokäsitykset voivat perustua esimerkiksi aiempaan tutkittuun tietoon, mutta myös uskomuksiin.

- Prioritiedon hyödyntäminen tilastollisessa tutkimuksessa on usein perusteltua.
- Bayesilaista tilastotiedettä tarkastellaan tarkemmin esimerkiksi kursseilla [TILM3577 Bayes-päättely](#) sekä [TILM3601 Bayes-laskenta](#).

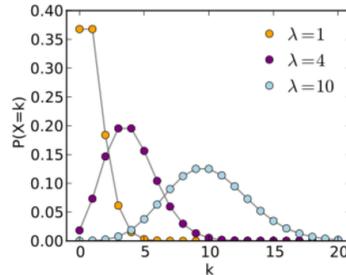
### 4.3 Tilastolliset mallit, jakaumat ja parametrit

- Tilastolliset mallit perustuvat satunnaismuuttujan mahdollisten tulosvaihtoehtojen todennäköisyksiä kuvaavalle **todennäköisyysjakaumalle**, joka määräät millä todennäköisydellä satunnaismuuttuja saa erilaisia arvoja.
  - Kuten aiemmin todettiin, satunnaismuuttujat jaetaan kahteen luokkaan: diskreetteihin ja jatkuviin.
- 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 lukumäärät ovat tässä tapauksessa sen verran suuria, että niitä mallinneitaan tyypillisesti jatkuva-arvoisena muuttujana.
  - Vastaavasti esimerkiksi potilaan jonotusaika päivystyksessä voi periaatteessa saada minkä tahansa arvon tietyltä reaalilukuväliltä ( $[0, \infty)$ , ts. mikä vain positiivinen arvo) ja tällöin käytettäisiin jatkuviin sm:jiin perustuvia tilastollisia menetelmiä.
- Satunnaismuuttujan mahdolliset arvot määräväät myös mahdollisen todennäköisyysjakauman ja täten myös käytettävän tilastollisen mallin.
  - **Diskreetin satunnaismuuttujan** jakauma voidaan usein esittää taulukkomuodossa. Eri arvojen todennäköisydet muodostavat kyseisen satunnaismuuttujan todennäköisyysjakauman, **pistetodennäköisyysfunktion**, jota voidaan havainnollistaa esimerkiksi pylväsdiagrammilla.
  - Jatkuvan satunnaismuuttujan  $Y$  arvot muodostavat jonkin reaalialakselin välin, joka sisältää äärettömän määren lukuja. Tämän vuoksi jatkuvan satunnaismuuttujan jakauman esittäminen taulukon kautta ei ole luonteva, vaan jakauma esitetään yleensä satunnaismuuttujan **tiheysfunktion** avulla.
    - \* Pistetodennäköisyys- ja tiheysfunktioit siis määräväät satunnaismuuttujan mahdollisille arvoille todennäköisydet väliltä  $[0, 1]$  ja näin voidaan arvioda havaitun aineiston uskottavuutta ja testata siihen liitettäviä hypoteeseja suhteessa estimoituun suurimman uskottavuuden estimaattiin.

- Tilastolliset mallit approksimoivat “todellista” aineiston generoinutta ilmiötä. Tilastolliset mallit riippuvat **parametreista** ja keskeinen oletus erityisesti klassisessa tilastotieteessä on, että aineiston generoinutta satunnaisilmiötä kuvaaa jokin vakiainen mutta tuntematon parametriarvo (tai niiden joukko).
  - Kuviossa 4.2 on kuvattu Poisson-jakauman sovelluskohteita ja sen pistetodennäköisyysfunktion muotoa eri parametrin  $\lambda$  arvoilla. Poisson-jakaumaa esitellään tarkemmin alaluvussa 4.5.

- 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  $\lambda$  arvoilla.

### Parametrien estimointi ja niiden testaus

- Satunnaisilmiötä kuvaava tilastollinen malli perustuu siis johonkin parametriseen todennäköisyysjakaumaan, joka yhdessä havaintojen kanssa määrittää uskottavuusfunktion.
  - Aineistoa kuvaavan tilastollisen mallin uskottavuus pyritään maksimimaan, mikä tarkoittaa valitun todennäköisyysjakauman sovittamista havaintoaineistoon mahdollisimman hyvin.
  - Tässä nk. “suurimman uskottavuuden estimoinnissa” aineiston generoiman (oletetun) todennäköisyysjakauman parametriarvot **estimoidaan** (eli arvioidaan) käytettäväni otoksen/aineiston avulla.

- Perusjoukko parhaiten kuvaavan (eli “aineiston generoineen”) parametrin arvo pyritään siis estimoimaan aineiston perusteella.
- Parametrien estimoinnin lisäksi usein **testataan** parametreja koskevia oletuksia (eli hypoteeseja).
- Estimoointi 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 180 cm?
  - \* 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**

Perinteikäs esimerkki odotusarvosta on tavallisen kuusitahoisen nopan silmäluvun odotusarvo. Nopanheitto on diskreetti satunnaisilmio ja tavallisen painottamattoman nopan tapauksessa jokaisen silmäluvun todennäköisyys on yhtä suuri. Merkitään nopan silmälukua (sm)  $Y$  ja sen realisaatiota  $y$ . Nopan silmäluvun realisaatioiden mahdolliset arvot ovat  $Y = \{1, 2, 3, 4, 5, 6\}$  ja niiden todennäköisyydet ovat  $P(Y = y) = \frac{1}{6}$ . Nopanheitton silmäluvun odotusarvo määritetään siis painotettuna keskiarvona

$$E(Y) = \sum_{i=1}^6 y \cdot P(Y = y) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2} = 3.5$$

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

**Varianssi**

Satunnaismuuttujan  $Y$  hajontaa voidaan mitata varianssilla

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

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 (keskihajonnan) tavanomaiset **estimaattorit** ovat otoskeskiarvo ja otosvarianssi (otoshajonta), joihin palataan vielä myöhemmin.

## 4.5 Joitain jakaumia

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

### 4.5.1 Normaalijakauma

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

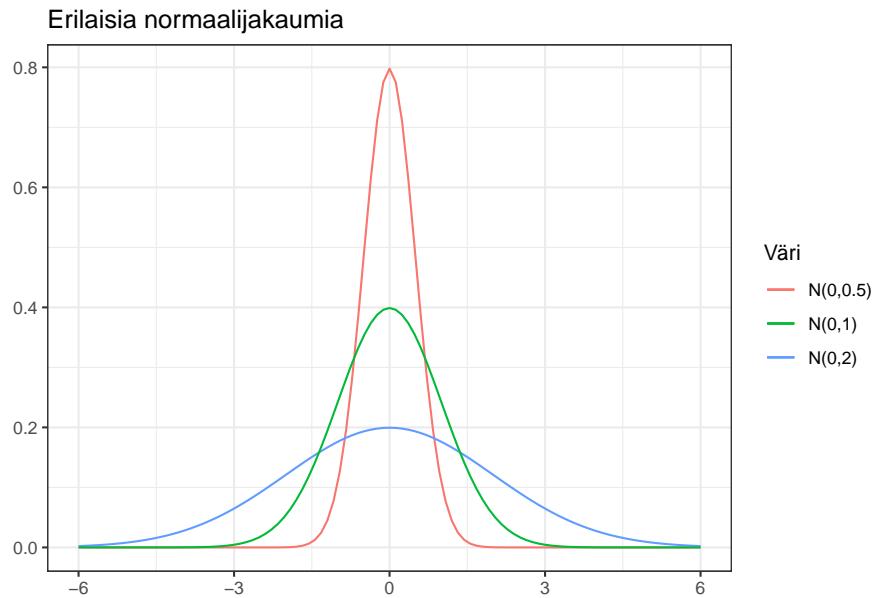
$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2},$$

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

- Ylläoleva tiheysfunktio määrittelee parven normaalijakaumia kun parametreille (vakioille)  $\mu$  ja  $\sigma^2$  annetaan erilaisia arvoja. Nämä kaksi parametria määrävät normaalijakauman tarkemman muodon.
  - Alla olevassa kuvassa 4.3 on kuvattu erilaisia normaalijakauman tiheysfunktion muotoja eri parametriarvoille.

#### Esimerkki: Miesten pituus

- Tutkitaan miesten pituutta hyvin määritellyssä joukossa, kuten vanusmiespalvelusta tietynä vuonna suorittavien joukossa.
  - Pituus on ominaisuus, jonka voidaan nähdä määrätyvän 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 odotettavissa oleva arvo.
  - Varianssi  $\sigma^2 = \text{Var}(Y) = E[(Y - \mu)^2]$  kuvailee valitun miehen pituuden odotusarvostaan määritetyn poikkeaman (keskihajonnan) neliön odotettavissa olevaa arvoa (kuvaten ts.



Kuva 4.3: Normaalijakaumien muotoja eri parametriarvoilla.

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, kuten myöhemmin tullaan näkemään).

- 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  $p$ .
- Pistetodennäköisyysfunktio:

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

- 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 syntynyttä lasta. Tällöin vanhempien tuli ilmoittaa lapsen sukupuoli (mies tai nainen) heidän syntymätodistuksessaan. Myöhemmässä tutkimuksessa tutkittiin tarkemmin 6115 perhettä, joissa asui 12 lasta ja tarkemmin miesten (poikien) lukumäärää näissä perheissä.

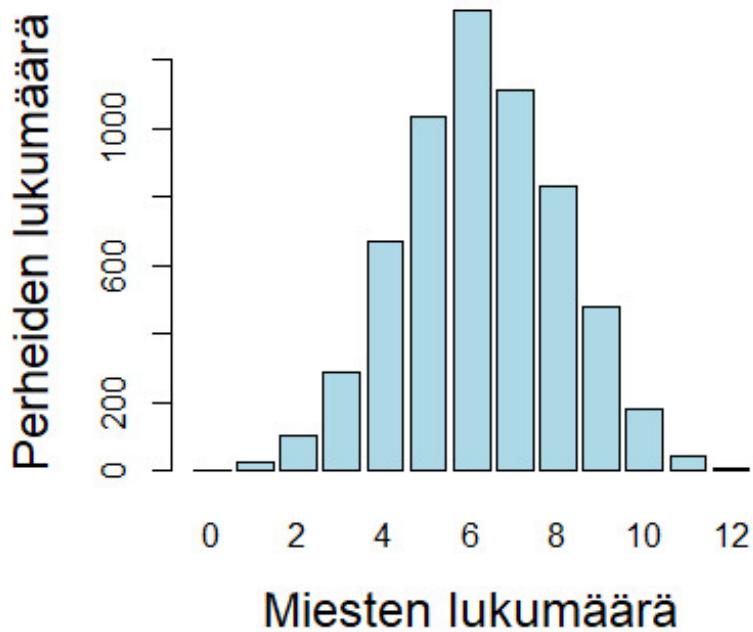
Oheisessa taulukossa taulukoidaan miesten (poikien) lukumäärät näissä 12 lapseen perheissä. Tarkasteltava jakauma esitetään vielä erikseen oheisessa kuviossa 4.4.

Tässä tilantessa mielenkiannon 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.

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



Kuva 4.4: Miesten lukumäärä Saksin osavaltiossa 12:n lapsen perheissä.

### 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 tietenkin ilmeinen! Näin ollen seuraavassa tarkastellaankin kotijoukkueiden ja vierasjoukkueiden maalintekointensiteettiä Poisson-jakaumaan perustuen. Seuraavassa emme siis pyri mallintamaan tietyn spesifin ottelun lopputulosta vaan tarkastelemme ”keskimääräisen” 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 lopputulokset *Blackburn Rovers - Nottingham Forest* 7-0 ja *Bolton Wanderers - Manchester United* 0-6 tulevat aineistoon tuloksina 4+ vs. 0 ja 0 vs. 4+.

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

Kotij. maalien lkm.	Vierasj. maalien lkm.					Yht.
	0	1	2	3	4+	
						Yht.
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

**Esimerkki (jatkuu): Poisson-jakauma**

Olettamalla, että koti- ja vierasjoukkueen todennäköisyys tehdä maali ottelun aikana on vakio, niin tällöin koti- ja vierasjoukkueen ottelun aikana tekemien maalien lukumääriä (ilman edellä käytettyä maalimäärien ”katkaisua” neljään) voidaan melko hyvin approksimoida oletuksella, ettei 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 intensiteetiparametrin  $\lambda_H$  arvon määrittäminen kuuluu tilastollisen päättelyyn.

telyn ja erityisesti estimointiteorian piiriin. Vastaavasti vierasjoukkueen maalimääritä:  $Y_i^A \sim P(\lambda_A)$ .

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 keskimääräisiä 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 kotiedun merkitystä jalkapallossa). Lisäksi edellä todetun Poisson-jakauman teoreettisten ominaisuuksien mukaisesti keskimäärität maalimäärität ovat lähellä niihin variansseja, mikä osoittaa osaltaan (tässä yksinkertaistetussa 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 saavuttamiseksi jälleen huomattavasti laajempia tilastotieteen (aine- ja syventäviä) opintoja.

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

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

- Kuten todettua, tilastollisessa tutkimuksessa mielenkiinnon kohteena on satunnaisilmiöiden tutkiminen ja erityisesti systemaattisen ja satunnaisen vaihtelun (signaalin ja kohinan) erottaminen sekä muuttujien välisten riippuvuuksien tutkiminen.

<sup>3</sup> Tämä jakso perustuu osin psykometriikan yliopisto-opettajan Jari Lipsasen [blogiin](#) vuodelta 2021.

- 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 esitellyksi.
- 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 arvostettuaan tutkimusta ei ole saatu **toistettua eli replikoitua**.
  - On ymmärrettävä, 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, jotien sen puuttuminen herättää ymmärrettävästi huolta tieteellisen prosessin toimivuudesta.
  - Replikaatiokriisin voi kuitenkin myös tulkita toisin: ilman kriittisyyttä omia (ja muiden) tuloksia kohtaan, ei mitään kriisiä olisikaan, jojen silkka sen olemassaolo on osoitus tieteellisen prosessin toimivudesta.
- Kun tuntee ja tunnistaa sattuman voiman ja ymmärtää kaikki mahdolliset satunnaisuuden lähteet, jotka altistavat tutkimusprosessin virheille, tulee samalla ymmärtääneksi 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!

## 4.7 Luvun 4 yhteenvetö, keskeisiä termejä ja konaisuuksia.

- Satunnaisilmiö
- Satunnaismuuttuja
- Jatkuват ja diskreetit satunnaismuuttujat ja niihin liittyvät pistetodennäköisyysfunktio ja tiheysfunktio
- Todennäköisyysjakauma

#### *4.7. LUVUN 4 YHTEENVETO, KESKEISIÄ TERMEJÄ JA KOKONAISSUUKSIA.75*

- Tilastollinen malli (todennäköisyysmalli)
- Kvalitatiiviset ja kvantitatiiviset muuttujat
- Frekventistinen ja Bayesiläinen tilastotiede
- Odotusarvo ja varianssi
- Yleisempiä jakaumia: Normaalijakauma, Bernoulli-jakauma, binomijakama ja Poisson-jakauma
- Tieteen popularisointi ja sen suhde (yksittäiseen) tutkimukseen



## Luku 5

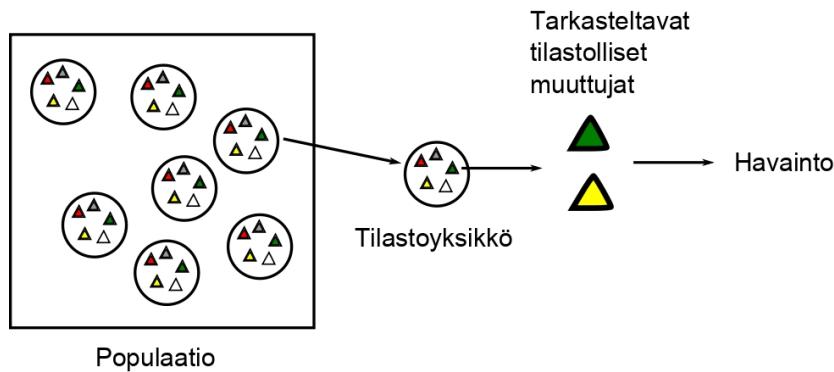
# Tilastolliset aineistot, niiden kerääminen ja mittaaminen

Edellisessä luvussa käsiteltiin tilastotieteen suhtautumista satunnaisilmiöihin. Tässä luvussa tarkastelemme lähemmin miten reaalimaailman satunnaisilmiöistä 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 olevat 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) tilastoysiksiötä.

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** populatiosta, 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 havaituista tilastomuuttujien arvoista.



Kuva 5.1: Populaatiosta havaintoon.

- Havaintoaineisto voidaan koota taulukoksi, johon listataan tilastoyksiköt riveille ja tilastomuuttujat sarakkeisiin. Jos havaintoaineisto koostuu  $n$  tilastoyksiköstä, joista jokaisesta on kerätty esim.  $m$ :stä tilastomuuttujasta havainnot, niin aineisto 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  $i$ . **tilastoyksikon** havainto ja sarakkeessa  $j$  on  $j$ . tilastollisesta muuttujasta havaitut arvot  $x_{i,j}$ . Ts. yhdellä rivillä on yhden tilastoyksikon tiedot kaikista tilastomuuttujista ja yksi sarake on kaikkien tilastoyksiköiden tiedot yhdestä tilastomuuttujasta.

- Usein (varsinkin parhaillaan kiihyväällä 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 koikiseksi 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 tavan luokitella ikäjakauaman kuvaamisessa 10-vuotisluokkiin (15-24, 25-34, ...), vaikka periaatteessa ikä voitaisiin ilmoittaa minuutinkin tarkkuudella.
  - Luokkien lukumääärään vaikuttavat muun muassa tilastomuuttujan arvojen vaihteluväli ja havaintoaineiston laajuus. Luokittelussa pyritään siihen, että luokkien lukumäärä saadaan tarvittaessa luokkia yhdistämällä kohtuulliseksi ja että luokat valitaan tasavälisesti 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 käsittele tilastografiikan esittämistä tarkemmin. 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 **taustamuuttuja**.
      - \* Esimerkiksi, jos tutkimuksella halutaan tietoa suomalaisen aikuisväestön mielipiteistä, havaintoyksikköinä ovat aikuisväestöön kuuluvat henkilöt. Jos halutaan tietoa suomalaisista kunnista, havaintoyksikköinä ovat Suomen kunnat jne.

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

- \* Ensimmäisessä tapauksessa tilastollisina muuttujina on aikuisväestön mielipiteet, joita voidaan selvittää esimerkiksi kyselytutkimuksella. Toisaalta voidaan myös kerätä taustamuuttujiksi haastatelluista muita tietoja, kuten asuinpaikka, ikä ja ammatti.
  - Kaikkia mielenkiintoisia muuttuja ei kuitenkaan välttämättä voida havaita, eli niille ei voida määritellä numeerista arvoa. Tällöin puhutaan nk. **latenteista muuttujista**, eli muuttujista joita ei suoraan havaita mutta joiden oletetaan vaikuttavan havaittavien muuttujien taustalla. Latenteja muuttuja voidaan rakentaa tilastollisten mallien avulla käyttäen hyödyksi niihin liittyviä havaittuja muuttuja.
    - \* Latenteja muuttuja ovat esimerkiksi elämänlaatu, onnellisuus, konservatiivisuus, yms.
- 
- Tilastollinen tutkimus voi olla joko **kokonaistutkimus** tai **otantatutkimus**.

### Kokonaistutkimus

Kokonaistutkimus on tutkimus, jossa tutkitaan kaikki tutkimuksen kohteena 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 kokonaistutkimussa tutkitaan kaikki kunnat. Kunnista on useimmissa tilanteissa mahdollista kerätä mielenkiinnon kohteena olevia tilastollisia muuttuja.
- Toisaalta, jos tutkitaan jonkin lääkeaineen vaikutuksia ihmisiin, niin kokonaistutkimussa tutkittaisiin jokainen ihminen erikseen. Selvää on, että tällainen kokonaistutkimus olisi liian vaikeaa toteuttaa.

### Otantatutkimus

Otantatutkimuksessa tutkimus kohdistetaan johonkin (populaatio/perusjoukon) osajoukkoon, joka poimitaan 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 ammuksit. Myöskaän 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 poikkeuksetta otantatutkimuksena (ja kontrolloituina kokeina, ks. alempaa), jolloin lääkettä 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 otoksena äärettömästä mahdollisten koetulosten joukosta. Tällöin ainoa mahdollisuus ilmiön tutkimiseen on käyttää otantaa.
- Otantatutkimuksen tulokset voivat olla luotettavampia kuin kokonaistutkimuksen.

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

- Otantatutkimuksessa voidaan panostaa enemmän huolelliseen ja tarkkaan mittamiseen sekä valitun otoksen tavoittamiseen.
- Kokonaistutkimuksessa vastauskato ja tarkasteltavan populaation vaalintavirhe ovat mahdollisia siinä missä otantatutkimuksessakin.
- Otantateoria on yksi tilastotieteen keskeisimpä oppeja ja tarjoaa teoreettisen kehikon empiiristen tutkimusten tulosten yleistämiseen. Tarkastellaan siis tarkemmin otannan ideaa ja toteuttamista seuraavassa alaluvussa.

### 5.2 Otannan idea

- Otantatutkimuksen (karkeat) suunnittelua- 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)
  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 määräää 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 **yleistävyys koko populaatioon**.

- Otos valitaan erilaisia **otantamenetelmiä** hyödyntäen pyrkien varmistamaan otoksen **edustavuus** (perusjoukoon pienoiskoossa, ks. kuvaa [5.2](#)).

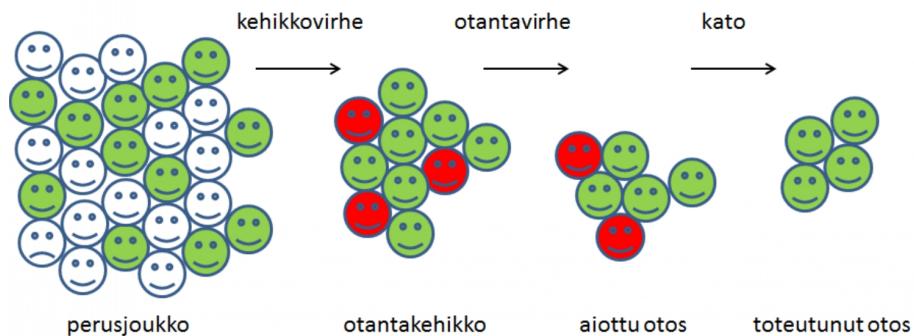
### **Edustavuus**

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 määrällisesti **sopivan kokoinen** aineiston.
- Tietyn otoksen edustavuutta arviodessa voidaan 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 tietenkin 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 tärkeisiin analyysiryhmiin kuuluvien riittävä määrä aineistossa?
  - 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, luetteloon tms. peittämä osa kohdeperusjoukkoa. Kyseessä on siis se osa kohdeperusjoukkoa, josta otanta ylipäänsä pystytään suorittamaan eli **otantakehikko**.
  - **Otantakehikon alipeitto** esiintyy, kun otantakehikosta puuttuu osa kohdeperusjoukon alkioista (esim. tutkimus suoritetaan puhelinhaastattelulla, mutta osa aiottuun otokseen kuuluvista haastateltavista ei omista puhelinta). Vastaavasti **otantakehikon ylipeittoa** esiintyy, kun otantakehikkoon kuuluu kohdeperusjoukkoon kuulumattomia alkioita.
    - \* Nämä ovat nk. **kehikkovirheetä**. Lisäksi esimerkiksi kyselytutkimuksissa tai rekisteriaineistoissa saattaa esiintyä **katoa**, eli osa vastauksista jää uupumaan tai niitä ei jostain syystä mitata.

## 84LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- \* **Otantavirhe** taas on satunnaisuudesta johtuvaa tilastollisten muuttujien vaihtelua otoksesta toiseen ja se onkin ainoa virhelaji, jonka suuruutta voidaan tilastollisin menetelmin arvioida.

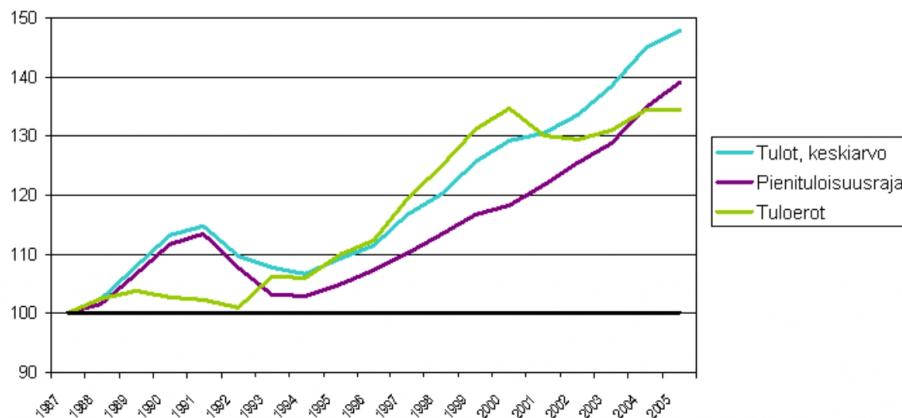


Kuva 5.2: Otannan idea.

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

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

- Tilastoysikkö on kotitalous, joten kaikkien kotitalouksien tutkiminen (kokonaistutkimus, ks. alla) olisi vaikeaa ja aikaavievää.
- Tutkittavaksi valitaan vain muutama tuhat kotitaloutta (ts. otatututkimus) ja selvitetään näiden tulot.
  - Tuloja, pienituloisuusrajaa ja tuloeroja on havainnollistettu kuvassa 5.3.
- On mahdollista tehdä **kaikkia** suomalaisia kotitalouksia koskevia johtopäätöksiä, jos tutkitut yksiköt ovat **edustava otos** suomalaisista kotitalouksista. Ts. osajoukko koskevat päätelmat voidaan yleistää koskemaan perusjoukkoa, mikäli osajoukko on edustava otos perusjoukosta.



Kuva 5.3: Tuloerot.

### 5.3 Mittaaminen ja mitta-asteikot

#### Mittaaminen

- Kumpaa tahansa tutkimusotetta (kokonais- tai otantatutkimus) noudatetaessa tietojen keräämisessä on olennaisena osana kohteiden ominaisuuksien **mittaaminen**.
- Tilastotieteellinen tutkimus perustuu aina mitattaviin satunnaisilmiöihin: tavoitteena on mittamalla liittää jokin luku ilmiötä kuvavaan ominaisuuteen, ts. mitata kyseisen satunnaismuuttujan havaittua arvoa.
  - Mittaaminen vaatii aina mittauksen kohteen, hyvin määritellyn mitattavan ominaisuuden ja **mittarin**, joka liittää mielekkäät lukuarvat 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

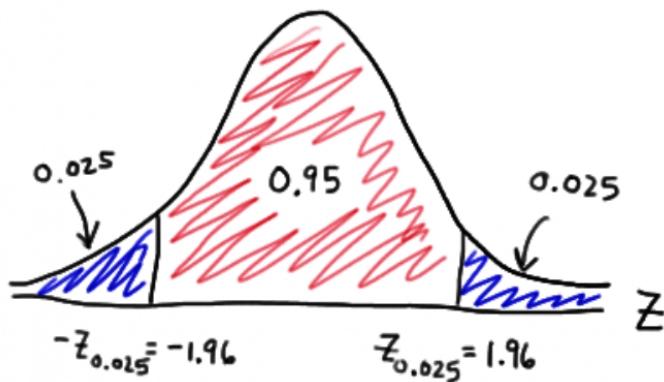
Mittari on harhaton, jos se ei systemaattisesti ali- tai yliarvioi mitattavan ominaisuuden määrää.

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

### Toistettavuus

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 reliabiliteettia/luotettavuutta** arvioidessa voidaan pohjata esimerkiksi seuraavia kysymyksiä:
  - Kuinka hyvin mittaustulokset ovat toistettavissa? Kuinka paljon niissä on ei-sattumanvaraisuutta?
  - Mittausten validiteetti: kuinka hyvin pystyttei mittamaan sitä, mikä oli tarkoitus mitata?
- 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 otosta, jos se ei ole edustava!
- 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. Esimerkinä **luottamusväli**:
  - Väli, joka vaihtelee otoksesta toiseen ja joka usein sisältää mielenkiinnnon kohteena olevan parametrin, kun otantakoetta toistetaan!
  - Luottamusväli käytetään määrittämään estimaatin luotettavuutta.
  - Väliestimointia tarkastellaan tarkemmin luvussa 6.



Kuva 5.4: Normaalijakaumaan perustuva 95% luottamusväli.

- 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 tulokset voivat olla puolueellisia siinä mielessä, että ne tarkastelevat asiaa vain yhdestä näkökannalta!
  - Esim. tarkastellaan yrityksen henkilöstökysymyksiä, työn organisointia ja työmoraalia, ongelmien tarkastelua johdon vs. henkilöstön näkökulmasta.

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

- Annettiin C-vitamiinia 100:lle 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, primäärikasvaimen sijaintipaikan ja histologisen kasvaintyyppin suhteiden.
  - Seuranta-aika: aika hetkestä, jolloin todettiin tavanomaisten hoitojen olevan tehotonta, kuolinhetkeen saakka.
  - Tulos: C-vitamiinia saaneet käsittelyryhmän potilaat elivät 4 kertaa kauemmin ( $p < 0.0001$ ).
- Ristiriitaista evidenssiä saatettiin tutkimuksessa, jossa vastaava tutkimusongelma, mutta toteutettu satunnaistettuna kokeena (Moertel et al. 1985).

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

- Satunnaistettiin potilaat, joilla pitkälle edennyt paksunsuolen tai peräsuolen syöpää, C-vitamiinia saavien ja lumelääketää saavien ryhmiin.
- Tulos: kontrolliryhmän potilaat olivat keskimäärin hieman pidempään, mutta ero ei tilastollisesti merkitsevä.
- Mistä kahden tutkimuksen erot johtuivat?
  - Huonolla tuurilla kaltaistetut verrokit 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 liittynyt enää mitään satunnaisuutta!

### Mitta-asteikot

- Kuten satunnaismuuttujia koskeneessa luvussa 4 opittiin, satunnaismiljöillä on erilaisia tulosvaihtoehtoja, jotka kantavat satunnaismuuttujien todennäköisyysjakaumia.
  - On syytä huomauttaa, että vaikka mitattava ilmiö ei olisikaan numeerinen, 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/luokitteluateikko** (nominaaliasteikko): Muuttujan mittautaso on tällöin sellainen, että sen arvot voidaan luokittaa toisistaan eroaviin luokkiin. Ts. mihin luokkaan kohde kuuluu mitattavan ominaisuuden perusteella?
    - \* Tilastoysiköt 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 väillä.
- \* Esimerkkejä: Sukupuoli, veriryhmä tai kotikunta.
- **Järjestysasteikko** (ordinaaliasteikko): Tällöin muuttujan arvot voidaan luokitteluun lisäksi asettaa empiirisesti mielekkääseen järjestykseen. Tällöin siis mittauksen kohteella on “enemmän mitattavaa ominaisuutta” kuin jollakin toisella kohteella
  - \* Tilastoyksiköt luokitellaan ennalta määritettyihin luokkiin, joilla on yksikäsitteinen järjestys.
  - \* Esimerkkejä: Sotilasarvo, sosiaaliryhmä, kilpailun tulos tai sairauksien tarttuvuuus.
- **Välimatka-asteikko** (intervalliasteikko): Luokittamisen ja järjestyskseen 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 mittamaan yksittäisten luokkien tai havaintoarvojen ero. Esimerkiksi: Lämpötilan mittäminen esim. celcius-asteina. Pystymme numeroarvoina ilmoittamaan onko tänään lämpimämpi, yhtä lämmin vai kylmempi 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). Mittarin 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 vähennyslasku ovat sallittuja.
- **Suhdeasteikko:** Jos intervalliasteikon ominaisuuksien lisäksi on määriteltyä yksikäsitteinen 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 asteikkoiksi.** Tällöin muuttujien arvot kuvaavat vain tilastoyksiköiden laadullisia piirteitä.

## 90LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Vastavasti **välimatka- ja suhdeasteikko kutsutaan kvantitatiiviseksi asteikoiksi**, koska tällöin mittaluvut kuvaavat jonkin ominaisuuden määräätä.
- Tilastollisen analyysin kannalta mitta-asteikkojen merkitys on siinä, että 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 kvalitatiivisille muuttujille.

### Aineistotyypejä

- Käsitellään tarkemmin vielä myöhemmin (Luvussa 10), joiden yhteydessä mitattavat muuttujat voivat olla kvalitatiivisia tai kvantitatiivisia.
  - Poikkileikkausaineisto: Tietoja useista tutkimuskohteista yhdeltä ajanhettelta tai aikaväliltä
  - Aikasarja-aineisto: Tietoja samasta tutkimuskohteesta eri ajanhettilta
  - Paneeliaineisto: Tietoja useilta ajanhettilta useista tutkimuskohteista
  - Tapahtumahistoria-aineisto: Tietoja tapahtumahetkiltä

## 5.4 Kontrolloidut kokeet ja suorat havainnot

- Tilastollinen tutkimusaineisto voidaan kerätä:
  - **Kontrolloidulla kokeilla**, joissa tutkimuksen kohteet altistetaan suunnitelmallisesti erilaisiin koeolosuhdeisiin selvittääkseen miten kohteet reagoivat muutoksiin.
  - **Suoria havaintoja** tehtäessä koeolosuhdeita ei pyritä aktiivisesti muuttamaan vaan ainoastaan seurataan miten erilaiset olosuhteet ja niissä tapahtuvat muutokset vaikuttavat kohteisiin.
- Näistä tutkimusasetelmista kontrolloidut kokeet ovat tietenkin ihanteellisempia tutkimuksen tekemiselle, sillä tutkijan on mahdollista tarkastella tutkittavaa asiaa koeolosuhdeissa ”eristyksissä”.
- Kontrolloidut kokeet eivät kuitenkaan ole aina mahdollisia, jolloin on käytettävä suoria havaintoja.
  - Tällöin tutkimuskohdetta ei suunnitelmallisesti altisteta koeolosuhdeille (“käsitellyille”) vaan muuttuvien olosuhteiden vaikutuksia tilastoyksikköihin seurataan passiivisesti.

- Toisin sanoen tutkimuksen kohteena olevat tilastoyksiköt eivät välttämättä edes tiedä osallistuvansa tutkimukseen.
- Lisäksi usein tehdään hoito/käsittelyvastetta koskevia vertailuja erilaisissa olosuhteissa, joka osaltaan vaikuttaa tulosten uskottavuuteen, sillä tutkitavien tilastoyksiköihin voi vaikuttaa olosuhteiden muutosten lisäksi muut ulkopuoliset tekijät.
  - Näiden **selittävien ja sekoittavien tekijöiden** vaikutusten kontrollointi on suoria havaintoja tehtäessä vaativia tehtävä.
  - Mikäli ulkopuolisista tekijöitä ei havaita ja/tai pystytä mittamaaan, 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, jonka vaikutusta ei kyötä kvantifioimaan puutteellisten havaintoarvojen vuoksi.
- Suoria havaintoja tehtäessä ei voida (usein) selvittää vasteen ja olosuhteiden **kausaalista** yhteyttä. Suorilla havainnoilla voidaan lähiinä saada selville onko vasteella ja olosuilla jokin yhteys (korrelatio) (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 tulosvaihtoehtoja. Tämä suosiminen voi olla tahallista tai tahaonta.
  - 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 valikoitumisharhaksi.

- Esimerkiksi verrattaessa sydän- ja verisuonitautipilaiden hoito-toimenpiteitä potilaat eivät mahdollisesti ole valikoituneet yhtä todennäköisesti pallolaajennukseen, ohitusleikkaukseen tai lääkehoitoryhmään, sillä taudin vakavuus saattaa jo määritellä mikä hoito-toimenpide valitaan.
- Valikoituminen on iso ongelma seurantatutkimuksissa, sillä harhais-ten havaintotulosten, eli harhaisen otoksen, perusteella ei voida 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 määrää mitkä perusjoukon alkioista tulevat poimituksi otokseen (tutkimuksen kohteiksi)!

### Satunnaistaminen

Tilastoyksiköiden poimimista populaatiosta otokseen riippumatta muiden yksiköiden poiminnasta tai kyseisten (poimittavien) yksiköiden ominaisuuksista.

- 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 ominaisuuksista. Satunnaistaminen on ainoa puolueeton tapa poimia otos (ei suosi mitään perusjoukon osaa)!
- Satunnaistaminen (osaltaan) mahdollistaa **tilastollisen päättelyn**, jolla avulla otoksesta saatuja tietoja voidaan hyödyntää tehtäessä päätelmiä koko perusjoukosta.
  - Tilastollisen päättelyn avulla voidaan muodostaa esimerkiksi jakaukien ja tilastollisten mallien tuntemattomille parametreille arviot (piste-estimaatit) ja arvioida niiden epävarmuutta (keskivirheet ja luottamusväli) sekä testata tarkasteltavaan ilmiöön liittyviä hypoteeseja (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 ominaisuden mukaan esiinny.

- Tämä tarkoittaa **altisteen** (käsittely / “treatment”) antamista (täyssin) satunnaisesti kokeeseen valituille yksilöille, riippumatta näiden taustamuuttujien arvoista.
- Nämä yksilöt sinänsä voivat olla satunnaisotos jostain populaatiosta (tai ainakin niiden toivotaan olevan), mutta satunnaistaminen tarkoittaa 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. Nämä ollen otos kuvailee koko perusjoukkoa.
  - Otantamenetelmän, joskus myös **otanta-asetelman**, valinta on tietenkin vahvasti sovellusalakohtainen: käytettävä aineistot ja täten otantamenetelmät määrytyvät pitkälti tehtävän tutkimuksen luonteen perusteella. Ts. käytännön tilanteet poikkeavat toisistaan lopulta varsin 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 **sisältymistodennäköisyystekijää**.

### 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**  $\pi_k$ ,

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

## 94LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

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

- Sisältymistodennäköisyys voi olla sama kaikille perusjoukon alkioille tai vaihdella perusjoukon eri osajoukojen (alkioryhmien) välillä. Tämä tulee huomioida otantamenetelmän valinnassa, jotta saadun otoksen edustavuus ei vaarannu.
- Sisältymistodennäköisyystä voidaan käyttää monimutkaisemmassa otantateoriassa **asetelma-** ja **analyysipainojoen** muodostamisessa sekä uudelleenpainotuksessa (vastauskadon korjaus).
- Tässä luvussa käsitellään erilaisia perinteisiä otantamenetelmiä sekä siitä, minkälaisista perusjoukojen tilanteissa mikäkin otantamenetelmä on sopivin.
  - **Yksinkertainen satunnaisotanta** (YSO): perinteisin otantameneelmä, jossa jokaisella tietyn kokoisella otoksella sama mahdollisuus tulla valituksi.
  - **Systemaattinen otanta** (SYS): eli tasavälisessä, otannassa poimin-takehikkoon (perusjoukkoon) kuuluvat alkiot järjestetään jonoon ja siitä poimitaan otokseen joka k. alkio.
  - **Oositettu otanta**: perusjoukko (populaatio) jaetaan ominaisuuksiltaan yhtenäisiin eli homogenisiin **ositteisiin**, joista jokaisesta poimitaan erillinen otos.
  - **Ryväsatanta** 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.
  - Otannan satunnaisuus tulee siis siitä, että jokainen tilastoyksikkö poimitaan otokseen *satunnaisesti!* (Ks. luku 4)
  - YSOa pidetään otannan perusmuotona, jossa jokaisella perusjoukon alkiolla on lähtökohtaisesti yhtä suuri todennäköisyys tulla valituksi otokseen.
    - \* YSO 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 (ks. aiempi luku 5.4) satunnaistamalla otokseen valikoitumisen täysin! YSO voi daankin 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 havaintoysiköistä (**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  $N$  alkiota 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

- 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 riipu siitä kuinka monta alkioita otokseen on jo poimittu.
  - Alkion  $i$  sisältymistodennäköisyys on tällöin selvästi

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

---

<sup>1</sup>Satunnaislukujen generointia käsitellään ja opetellaan mm. kursseilla [TILM3517 R-kielen alkeet](#) ja [TILM3705 Johdatus laskennalliseen tilastotieteeseen](#).

## 96LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- 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 epäre- listinen otantamenetelmä useassa tutkimuksessa. Esimerkiksi lienee mahdotonta testata samaa läkettä 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 riipumattomia: 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 otos- ta.<sup>2</sup>
  - Otantaan palauttamatta liittyviä todennäköisyyksiä hallitaan **hy- pergeometrisen jakauman** avulla, joka johtaa (melko) yksinker- taiseen **tilastolliseen malliin** YSO:a käytettäessä.

---

<sup>2</sup>Kun otosyksiköiden järjestyksestä 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.

**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 todennäköisyyys 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 käytettävän otantaa takaisinpanolla, sillä siinä käytettävät tilastolliset mallit ovat yksinkertaisempia kuin otannassa ilman takaisinpanoa ja tämä helpottaa tilastolisessa päättelyssä 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 haajaantunut. Henkilökohtaisten, kasvotusten toteutettavien, haastattelujen tekeminen vaatisi suuria resursseja (haastattelijat joutuisivat esim. matkustamaan ympäri Suomea satunnaisotokseen valikoituneiden henkilöiden asuinpaikkojen mukaan).
- Tällaisissa tutkimustilanteissa käytetäänkin usein muunlaisia otantameneitä.

### 5.5.2 Systemaattinen otanta

- Systemaattisessa, eli tasavälisessä, otannassa poimintakehikkoon (perusjoukkoon) kuuluvat alkiot 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 alkiot 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 omaisuudesta, jota tutkitaan.
  - Systemaattisen otannan suhteen potentiaaliseksi ongelmaksi muotoutuu havaintoyksikkölistan mahdollinen säädöllinen jaksollisuus, jota se ei havaitse ja jolloin satunnaisotanta toimisi (kenties) paremmin.
    - \* Ongelmaa syntyy esimerkiksi silloin, jos tiedot perusjoukosta koostuvat heteropariskunnista ja poimintaintervalli on parilleen 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 sovelletaan tavallisesti YSO:n sijasta silloin, kun perusjoukon alkioista on käytettäväissä tietorekisteri, luettelo tai havaintoja kerätään ajassa tai tilassa.
  - Esimerkiksi mielipidekyselyn kohteet poimitaan (voitiin poimia) puhelinluettelosta (tai vastaavasta rekisteristä) valitsemalla haastateltavaksi jokaiselta aukeamalta ensimmäisenä esiintyvää henkilö tai jotain tuotetta valmistavan tehtaan laaduvalvonnassa valitsemalla laatuarviointiin joka sadas tuote, joka hihnalta valmistuu. Muita esimerkkejä ovat esim. liikenne-, jäsenrekisteri- tai kassajonossa seisovien ottayksiköiden 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.
  - Ositetut 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.
- Ositetut 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*).
- Rypäät indikoivat aineiston luontaista hierarkkista, eli monitasoista- tai asteista rakennetta.
  - Esimerkkejä tällaisista ryhmistä ovat erilaiset yritykset tai koululuokat. Esimerkiksi yritykset muodostavat luonnollisesti eri rypääitä, joiden alkiot ovat työntekijöitä ja koululuokat muodostavat koulun sisällä omia luonnollisia rypääitä ja opiskelijat ovat alkioita näissä rypäissä.

## 100LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERÄÄMINEN JA MITTAAMINEN

- Huomionarvoista onkin, että toisin kuin ositetussa otannassa, ryvästannassa rypäiden oletetaan olevan toistensa kanssa riittävän samankaltaisia, että jokaista rypästä ei tarvitse erikseen tutkia.
  - Tämä onkin yksi ryvästannan tärkeimpiä motivointejä, sillä sitä usein perustellaan kustannustehokkuudella: sen sijaan että poimitaan satunnaisia koululaisia mahdollisesti suuresta määristä kouluja, voidaan poimia satunnaisia rypääitä (kouluja), joista tutkimusyksiköt eli koululaiset poimitaan.
  - Lisäksi koulun sisällä koululuokat muodostavat alirypääitä, 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!
- Ryvästannan voi suorittaa **yksi-** tai **kaksivaiheisena (yksiasteinen/kaksiasteinen ryvästanta)**.
  - **Kaksivaiheisessa ryvästannassa**
    - \* **Ensimmäisessä vaiheessa** poimitaan joukko rypääitä kaikkien rypäiden joukosta, eli vain osa rypäistä on mukana lopullisessa otoksessa.
    - \* **Toisessa vaiheessa** poimitaan ensimmäisessä vaiheessa poimitusta rypäistä alkiotason otokset.
  - **Yksivaiheisessa ryvästannassa** toisessa vaiheessa valitaan kaikki ensimmäisen vaiheen otosrypäiden alkiot, jolloin toisen vaiheen otanta typistyy ensimmäisen vaiheen rypäiden alkioiden kokonaistutkimukseksi.
  - Poiminnan eri vaiheissa voidaan soveltaa yksinkertaista satunnaisottantaa tai systemaattista otantaa.
- Ryvästantaa käytetään usein suuria haastattelututkimuksia tehtäessä. Erityisesti, ryvästantaa voidaan hyödyntää myös silloin, kun tutkijalla ei ole käytettävissään kattavaa lista kaikista havaintoyksiköistä, mutta näiden muodostamat rypät on määritettävissä.
- Ryvästannan heikkoutena pidetään sitä, ettei aina ole helppoa muodostaa rypääitä, jotka ovat toistensa kaltaisia. Tulosten tarkkuus myös riippuu moninpakoin siitä, kuinka hyvin rypäisiin jako onnistuu.

**Esimerkkejä ryvästotannasta**

- Esimerkki 1:
  - Poimitaan oppilaitoksen opiskelijoista otos arpomalla ensin otos luokkahuoneista (=rypäästä).
  - Arvotuissa luokkahuoneissa käydään sitten suorittamassa kysely.
    - \* Esim. Oppilaitoksen opiskelijoista voidaan poimia otos arpomalla ensin otos luokkahuoneista, jolloin luokkahuoneet ovat nk. rypääitä.
    - \* Mahdollisia ongelmia? Miten huomoida päivä- ja iltaopiskelijat? Tämän voisi toteuttaa arpomalla otos luokkahuoneista päiväsaikaan ja toinen otos ilta-aikaan. Tässä yhdistetään ryvästontaan osittettu otanta, jolla taataan päivä- ja iltaopiskelijoiden edustus.
- Esimerkki 2: Tutkittaessa tänä vuonna peruskoulun aloittavia voidaan ensin poimia otos koulusta, jolloin koulut ovat rypääitä. 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 työvoimapoliittisten ennusteiden ja suunnitelmien laadinnassa, toimien seurannassa ja päätöksenteon tukena.
  - Työmarkkina-aseman perusuokittelussa väestö jaetaan työlliisiin, työttömiin ja työvoiman ulkopuolisiin.
    - \* Työlliset ja työttömät muodostavat työvoiman.
  - Työvoimatutkimuksen **perusjoukon** muodostavat Suomessa vakinaisesti asuvat 15–74-vuotiaat henkilöt.
  - Työvoimatutkimuksen otos poimitaan **ositetulla satunnaisuudella**.

## 102LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERÄÄMINEN JA MITTAAMINEN

**sotannalla** väestön keskusrekisteriin perustuvasta Tilastokeskuksen väestötietokannasta kahdesti vuodessa.

- Ositetun satunnaisotoksen poiminta:
  - Tutkimus on paneelitutkimus, jossa samaa henkilöä haastatellaan viisi kertaa.
  - Joka kuukauden otokseen kuuluu noin 12 000 henkilöä, keskimäärin noin joka 300. henkilö perusjoukosta.
  - Yhden tutkimuskuukauden 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 työllisyyden ja työttömyyden muutoksia edellisen vuoden vastaavasta kuukaudesta. Vaihtoehtoisesti voidaan käyttää kausitasoitettuja lukuja, jolloin tilannetta voidaan verrata edelliseen kuukauteen.

### Esimerkki: Terveys 2000

- Terveys 2000 -tutkimuksen tavoite oli tuottaa ajankohtainen kattava kuva työikäisen ja iäkkään väestön terveydestä ja toimintakyvystä selvittämällä tärkeimpien terveysongelmien yleisyyttä ja sijitä 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 ryvästotos terveyskeskuspisteistä.
  - Ositus perustui yliopistosairaaloiden vastuualueiden väestömäärään suhteutettuun kiintiöintiin.
  - Suurimmat 15 terveyskeskuspisteitä 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 valkoituneet 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.
  - Jos television ajankohtaisohjelma pyytää katsojia twittaamaan mieleipiteensä ajankohtaisesta asiasta, kyseessä on itse valikoituva näyte (osallistujat valitsevat itse itsensä).
- Jos poimitaan tutkimukseen ne perusjoukon alkiot, jotka ovat tutkimuksen tekemishetkellä ‘saatavilla’, niin kyseessä on **näyte**. Näyte ei siis kata ilmiön koko vaiotelua 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 aiakaan liikkuvat ihmiset.
- **Vajaapeittävyys:** Populaation alkioista ei ole välittämättä täydellistä luetteloa
- **Vastauskato:** Tutkimuksen kohteita ei tavoiteta tai he kieltäytyvät vastaamatta. Kadon vuoksi lopullinen otoskoko saattaa jopa karsiuua pois tai jokin osajoukko on aliédustettuna.
- **Vastausharha:** Kysymykset voivat olla huonosti muotoiltuja tai vastaajat voivat antaa väärää tietoja.

## 5.8 Luvun 5 yhteenvetö, keskeisiä termejä ja konkaisuuksia.

**Termejä** - Otanta - Perusjoukko/populaatio - Otos/otosaineisto/data - Havainno/havaintoarvo - Edustavuus ja edustava otos - Tilastoysiikkö ja tilastomuuttuja - Tutkimusmuuttuja ja taustamuuttuja(t) - Kokonaistutkimus vs. otanta-tutkimus - Otoksokko - Otantakehikko - Mittaaminen - Harhattomuus - Toistetavuus - Kontrolloidut kokeet vs. suorat havainnot - Selittävät ja sekoiittavat tekijät - Puuttuvien selittäjien harha - Valikoituminen - Satunnaistaminen - Käsittely- ja kontrolliryhmät

**Mitta-asteikot** - Laatueroasteikko/luokitteluaasteikko (nominaaliasteikko) - Järjestysasteikko (ordinaaliasteikko) - Välimatka-asteikko (intervalliasteikko) - Suhdeasteikko - Kvalitatiiviset ja kvantitatiiviset muuttujat

## **104LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERÄÄMINEN JA MITTAAMINEN**

**Otantamenetelmät** - Yksinkertainen satunnaisotanta (YSO) - Systemaattinen otanta (SYS) - Ositettu otanta - Ryvästonta - Sisältymistodennäköisyys - YSO takaisinpalauttaen ja YSO takaisin palauttamatta - Poimintaharha - Vajaapeittävyys - Vastauskato - Vastausharha - Näyte

## Luku 6

# Otokset ja otosjakaumat: tilastollisen päätelyn näkökulma

Tarkastellaan seuraavaksi otoksia ja otosjakaumia “tilastollisemmin” mitä edellisten lukujen erityisesti otantaa koskevan johdannon yhteydessä. Tilastollinen päätely on keskeinen osa tilastotiedettä, sillä se mahdollistaa päätelmien 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 satunnaismuuttuja käyttäen. Satunnaismuuttujilla on todennäköisyysjakaumat, joita tilastotieteessä kuvataan todennäköisyys- eli tiheysfunktion (tai pistetodennäköisyysfunktion) avulla.
  - Merkitään satunnaismuuttujaa isolla kirjaimella,  $Y$ , ja satunnaismuuttujan realisaatiota pienellä kirjaimella  $y$ . Otoskokoa, eli otokseen osallistuvien tilastoyksiköiden määrää merkitään  $n$ :llä ja tilastoyksiköitä indeksöidään alaindeksillä  $i = 1, \dots, n$ .

## 106LUKU 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 pistetodennäköisyysfunktiota (ptnf)) merkitään  $f(y, \theta)$ :llä, jossa  $y$ :n 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 arvo lopulta käytettävässä olevasta aineistosta.

### 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 tuloksesta saatuihin havaintoarvoihin, vaan otoksen poimintaan.
- Satunnaismuuttujien  $Y_1, \dots, Y_n$  **yhteisjakauma** muodostaa (tiettyjen liäätusten jälkeen) **tilastollisen mallin** havaintoarvojen satunnaiselle vaihtelulle eri otoksissa.
  - Koska tällä kurssilla satunnaismuuttujat  $Y_1, \dots, Y_n$  oletetaan **riippumattomiksi toisiinsa nähden**, niiden yhteisjakauma on tulomuotoa  $f(y_1, \dots, y_n; \theta) = f(y_1; \theta) \times \dots \times f(y_n; \theta)$ .
- Tässä  $f(y_1, \dots, y_n; \theta)$  on siis tilastollinen malli: sen muodon määrästä tutkijan tekemä aineistoa koskeva jakaumaoletus, mikä voi paikoin olla hyvin monimutkainen. Tilastollisen mallin monimutkaisuus ilmenee sen parametrien määristä: mitä useampi parametri (erit. suhteessa havaintojen määrään), sitä monimutkaisempi malli.
  - Useimmista kuitenkin ajatellaan, että on käytettävä niin yksinkertaisia menetelmiä kuin mahdollista, mutta ei yhtään yksinkertaisempia. Tämä on ns. **parsimonisuusperiaate** eli **vähäparametrisuus- tai sääteliäisyysperiaate**.
  - Vähäparametrisuusperiaatteen voidaan nähdä perustuvan ns. **Occam partaveitsen -periaatteeseen**, jonka mukaan “*ilmiötä selittävien*

## 6.1. SATUNNAISOTOS, YHTEISJAKAUMA JA TILASTOLLINEN MALLI107

*tekijöiden määrän 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 3.3 ja luku 6).
- Oletetaan, että  $Y_1, \dots, Y_n$  ovat aiempien oletusten pätissä riippumattomia sm:jia ja että ne muodostavat satunnaisotoksen jakaumasta, jonka odotusarvo on  $\mu$  ja varianssi on  $\sigma^2$ .
  - Ts. oletamme

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

- Tässä tapauksessa mielenkiinnon kohteena olevat parametrit ovat siis  $\mu$  ja  $\sigma^2$  eli  $\theta = (\mu \ \ \sigma^2)$ .
- Tilastollisten mallien tehtäväänä on siis estimoida nämä todennäköisyysjakaumien parametrit havaitun aineiston perusteella, joten keskeinen tilastollinen kysymys on että miten estimoointi suoritetaan luotettavasti.

### Esimerkki: satunnaisotos normaalijakaumasta

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

- Merkintä  $\perp\!\!\!\perp$  tarkoittaa, että sm:t  $Y_1, \dots, Y_n$  ovat riippumattomia ja samoin jakautuneita (toisinakin myös lyhyesti *iid* tai *i.i.d.*, joka tulee englannin kielen ilmaisusta “independent and identically distributed”). Merkintä soveltuu käytettäväksi muidenkin jakaumien tapauksessa.
- 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 tehdyn 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 tilastollinen päättely pohjautuu aineiston tilastollisen mallin kuvaamalle 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 kuvaamalla tavalla.
- Satunnaisotoksesta voidaan laskea erilaisia **tunnuslukuja/otossuureita**, joita merkitään  $T$ :llä, ts. ne ovat aineiston funktioita

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

- **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 (parametrien)  $\theta$  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 keskityä mahdollisimman tiiviisti  $g(\theta)$ :n ympärille.
- Koska tunnusluku/estimaattori  $T$  on satunnaismuuttuja, sillä on todennäköisyysjakauma, jota kutsutaan tunnusluvun  $T$  **otosjakaumaksi**.
  - Otojakama muodostaa (tilastollisen mallin) todennäköisyysmallin tunnusluvun  $T$  arvojen satunnaisvaihtelulle otoksesta toiseen.
  - Otojakamat riippuvat tuntemattomista **parametreista**, joiden arvoja ei yleensä tunneta ja niitä pyritään estimoimaan kerättyä otosta ja sopivaa tunnuslukua käyttäen.
  - Parametri on (usein) perusjoukon tunnusluku, jota halutaan arvioida. Parametrit **estimoidaan**, kuten yllä jo todettiin, havaintoaineistoa käyttäen.

### Estimaattorin ominaisuudet

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

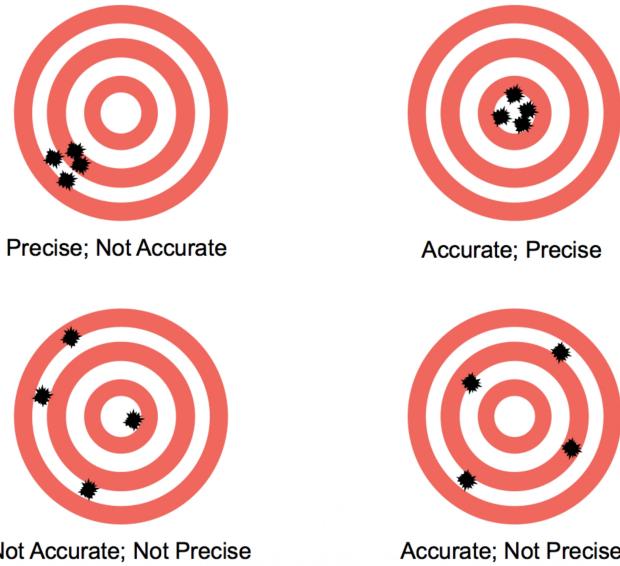
#### Harhattomuus

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 otokselle poiketa paljonkin parametrin todellisesta arvosta, mutta odotusarvon frekvenssitulkinnan mukaan estimaattorin tuottamat otoskohdaiset arvot parametrille jakautuvat otantaa toistettaessa (symmet-

## 110LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

risesti) parametrin todellisen arvon ympärille.



Kuva 6.1: Havainnollistuksia estimaattoreiden ominaisuuksista.

### 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 binomijakaumien parametreiden estimaattoreilla on kaikki edellä mainitut hyväät ominaisuudet.
  - Näin ei ole yleisesti monimutkaisemmissa otantatilanteissa ja tilastollisissa malleissa.
  - Estimaattoreiden kehittäminen erilaisten tilastollisten mallien tapauksessa kuuluu teoreettisen tilastotieteen alaan.
- Seuraavaksi perehdytään tarkemmin kahteen kenties useimmiten tarkasteltavaan tunnuslukuun ja niiden otosjakaumiin:
  - Aritmeettisen keskiarvon otosjakaumaan [6.3](#)
  - Suhteellisen osuuden (frekvenssin) otosjakaumaan [6.4](#)

### 6.3 Otoskeskiarvo ja otosvarianssi (estimaattoreina)

#### 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

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

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

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

## 112LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

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

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

- Täten otoskesiarvo kuvaaa aineiston perusjoukon tilastollisen mallin odotusarvoa.

### Aritmeettisen keskiarvon ominaisuuksia

- Aiempien oletusten pätissä aritmeettisella keskiarvolla  $\bar{Y}$  on seuraava odotusarvo ja varianssi:

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

- Aritmeettisen keskiarvon  $\bar{Y}$  **standardipoikkeama**

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

- Standardipoikkeamaa kutsutaan myös **keskiarvon keskivirheksi** ja se kuvaaa otoskesiarvon otosvaihtelua odotusarvon  $\mu$  ympärillä.
- Aritmeettisen keskiarvon otosjakauma keskittyy yhä voimakkaammin haavaintojen 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**.
- Otoskeskiarvo  $\bar{Y}$  ja otosvarianssi  $S^2$  ovat siis satunnaismuuttuja, joiden saamat arvot vaihelevat 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 käytäjää tilastotieteen (ja matematiikan) opintoja.

### Standardoidun aritmeettisen keskiarvon otosjakauma

- Tarkastellaan **standardoitua** satunnaismuuttujaa

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

- 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

$$P(A) = p,$$

jolloin tapahtuman  $A$  komplementtitalpahtuman (vastatahutman)  $A^c$  todennäköisyys on

$$P(A^c) = 1 - p = q.$$

- 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

- Frekvensillä  $f$  on odotusarvo

$$E(f) = np,$$

ja varianssi

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

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

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

### Suhteellinen frekvenssi: Odotusarvo ja varianssi

- Suhteellisen frekvenssin  $\hat{p}$  odotusarvo

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

ja varianssi

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

- Suhteellisen frekvenssin  $\hat{p}$  standardipoikkeamaa

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

voidaan kutsua **suhteellisen frekvenssin keskivirheeksi** ja se kuvailee 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 päissä normaalijakaumaa:

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

- Aritmeettisen keskiarvon tapaan standardoituu 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-osuus olisi ollut alle 50 %?
- Suhteellisen frekvenssin otosjakauman perusteella kyllä-kannatusosuuden jakauma olisi

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

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. Mm. otoskeskiarvoa tunnushukuna tarkasteltiin jo 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 tyyppiarvo 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.

- Luokittelusteikolla mitattaville muuttujille ei ole olemassa luontevia sijaintilukuja keskilukujen yhteydessä pl. moodi.
- Järjestysasteikolla mitatuille muuttujille voidaan mediaanin lisäksi määritetään **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.
  - Tietystä fraktiileista käytetään nimitystä **desiili**. Ensimmäinen desiili  $D_1$  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** kuvailee 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 määrittää välimatka- ja suhdeasteikon muuttujille. Vinous ja huipukkuus mittavat 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 parametriestimaatti ”lähellä” vai ”kaukana” sen todellisesta arvosta.
  - Täten tarvitaan jokin tapa, jolla saadun parametriestimaatin luotettavutta voidaan arvioda.

### **Luottamusväli**

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

## 118LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Otoskoko on luottamusvälejä koskevissa tarkasteluissa keskeinen ja luottamusväleihin palataankin otoskoon käsittelyn yhteydessä.
- Valittua luottamustasoa merkitään usein  $1 - \alpha$ :lla, jossa **merkitsevyystaso (riskitaso)**  $\alpha$  on esimerkiksi  $\alpha = 0.05$ .
- Tulkinta: Jos **otantaa** jakaumasta  $f(y; \theta)$  toistetaan, niin keskimäärin  $100 \times (1 - \alpha)\%$  otoksista konstruloiduista luottamusväleistä peittää parametrin  $\theta$  todellisen arvon.
- Oletetaan, että olemme tehneet johtopäätöksen, että konstruloitu luottamusväli peittää parametrin  $\theta$  tuntemattoman todellisen arvon.
  - Tällöin otantaa toistettaessa luottamusvälin konstruktiosta 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 itse asiassa 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.
- Luottamusväleissä ei varsinaisesti ole kyse “virheestä” vaan saadun/muodostetun tiedon tarkkuudesta.
  - Luottamusvälit, eli virhemarginaalit, siis (yleisesti) riippuvat valitusta 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 kananatuksen virhemarginaali on 3,5 puoleen tai toiseen 95 % luottamustasolla.”
    - \* Virhemarginaali kasvaa, kun aineistoa lohkotaan: jos tuhannen hengen otoksesta esitetään tietoja, jotka kuvaavat 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.

### Normaalijakauman odotusarvon luottamusväli

- Käsittelemme seuraavassa (normaalijakauman) 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 normaalijakaumasta  $Y_1, \dots, Y_n \perp\!\!\!\perp, Y_i \sim N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ .
- 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 alaluvut tässä luvussa 6) noudattaa  $N(0, 1)$ -jakaumaa.

- $P(\cdot)$ :llä merkitään todennäköisyyttä, joka tässä tapauksessa liittyy normaalijakaumaan, ja  $z_{\alpha/2}$  on jakaumafunktion arvo pisteessä  $\alpha/2$ .
- Tällöin etsitään odotusarvoparametrille  $\mu$  sellainen arvo, jolla oheinen epäyhtälö pätee ja päädytään luottamusväliin.

## 120LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Nyt epäyhtälöketju voidaan kirjoittaa muodossa

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

- 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  $\bar{Y}$  suhteeseen. 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

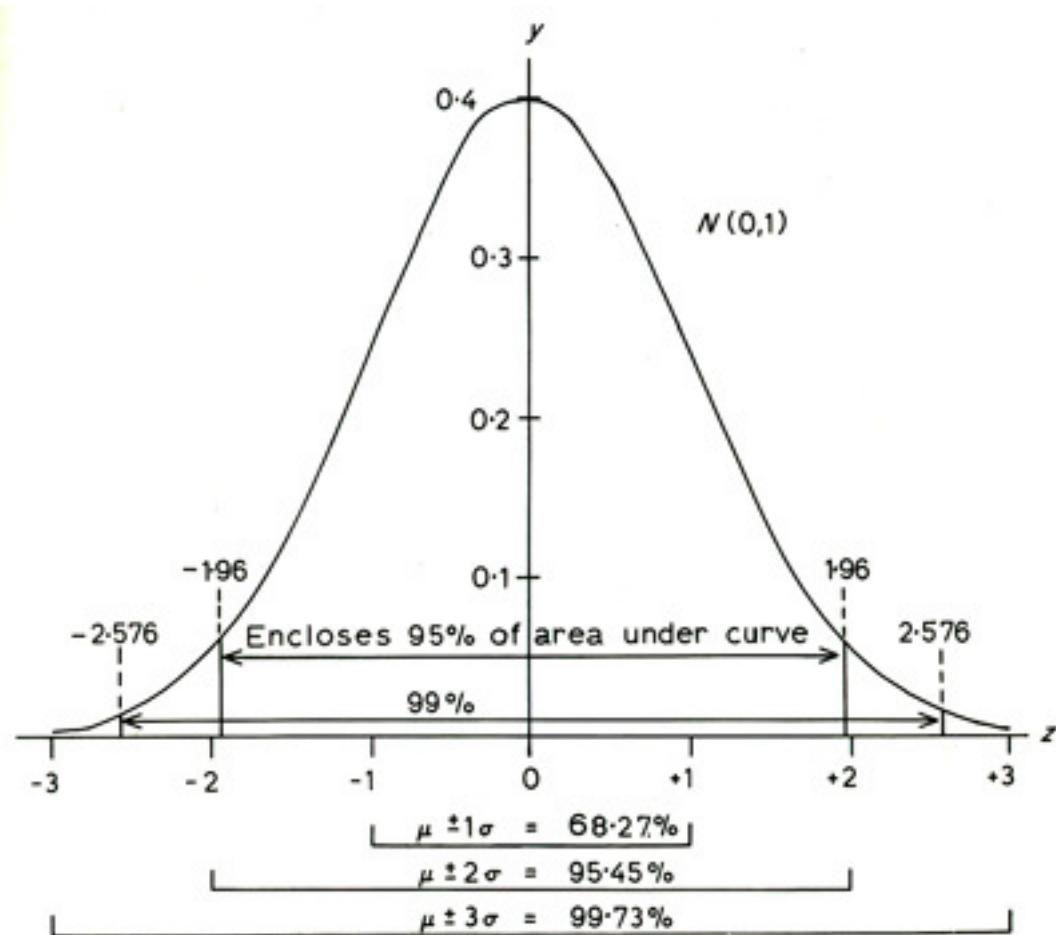
$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- 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)**



Kuva 6.2: Standardoitu normaalijakauma: Virhemarginaaleja

## 122LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

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

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

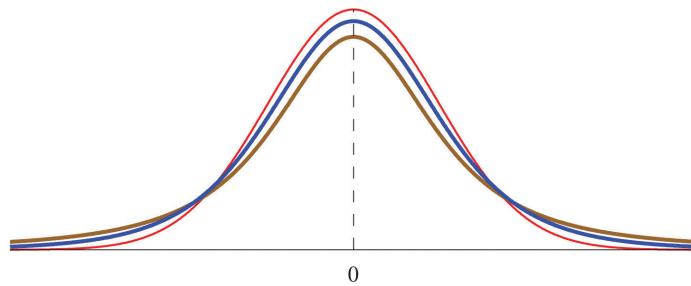
jossa **luottamuskertoimet**  $-t_{\alpha/2}$  ja  $t_{\alpha/2}$  saadaan nyt ***t*-jakaumasta**  $t_{n-1}$ , jossa  $S^2$  on varianssin  $\sigma^2$  harhaton estimaattori ja vapausasteiden lukumäärä on  $n - 1$ .

- (Studentin) *t*-jakauma muistuttaa silmämääräisesti normaalijakamaa, mutta se on paksuhäntäisempi. Vapausteluvun kasvaessa *t*-jakauma lähestyy normaalijakaumaa.
- Suurissa otoksissa ( $n$  iso) luottamuskertoimet voidaan poimia (appproksimatiivisesti) myös normaalijakaumasta eli korvata edellä kertoimet  $t_{\alpha/2}$  aiemmin käytettyillä 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: *t*-jakauman (ja standardoidun normaalijakauman) tiheysfunktioita

### 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

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

- 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 määritetään 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

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

## 6.7 Otoskoko

- Yksi tilastollisen päättelyn keskeisiä tavoitteita on yleistää otoksen pohjalta tehty päättely koskemaan koko perusjoukkoa. Seuraavaksi käymme läpi seikkoja, jotka tulee ottaa huomioon otoskokoa miettiessä.
- Kun on päättetty, millainen tutkimusaineisto halutaan kerätä, on päättetävä, kuinka suuri otoksen on oltava, jotta se edustaa tutkittavaa joukkoa kattavasti.
  - Liian pieni **otoskoko**, eli pieni määrä otokseen poimittuja tilastoyksiköitä, voi **sattumalta** poiketa paljonkin perusjoukosta.
    - \* Tämä niin kutsuttu **otantavirhe** on sitä suurempi mitä pienemppää otosta käytetään.
    - \* Liian pienien otoskosten vuoksi muuten hyvin toteuttu tutkimus ja otanta-asetelma saattaa epäonnistua vastaamaan tutkimuksen mielenkiinnon kohteena olevaan kysymykseen.
  - Todella suuren otoksen koostaminen voi olla **työlästä**, **kallista** tai joskus jopa **täysin mahdotonta** esimerkiksi siksi että käytettävissä olevat tutkimusyksiköt eivät ole käytettävissä ajallisten rajoitteiden vuoksi (kuten harvinaisten tautien kantajat).
    - \* Toisaalta perusjoukon systemaattiset piirteet tulevat otoskoon kasvaessa paremmin esille, vaikka yksittäisten otosyksiköiden tilastolliset muuttujat saattavat vaihdella suuresti.
  - **Otoskoko** siis vaikuttaa keskeisesti siihen, miten hyvin otoksesta tehdyt johtopäätökset voidaan yleistää koskemaan koko perusjoukkoa!
- Optimaalinen, tai ainakin tutkimusongelmaan vastaamisen kannalta vähintään riittävä arvio, otoskoosta voidaan usein määräätä etukäteen.

### Perusjoukon rooli otoskoon määrittämisessä

- Ensiksi tulee kuitenkin pohtia käsillä olevaa tutkimusongelmaa esimerkiksi kysymällä: **Millainen on perusjoukkosi?**
  - Onko tutkittavan muuttujan arvoissa paljon vaihtelua? Jos on, niin tämä täytyy huomioida kasvattamalla otoskokoa.
    - \* Esimerkiksi otosten keskiarvot alkavat käyttäytyä riittävän sisistä vasta noin otoskoosta  $n = 30$  alkaen.
    - \* Kyseinen otoskoko ei kuitenkaan ole missään tapauksessa yleinen ja pätevä peukalosääntö otoskoon koolle, vaan se tulee aina päättää tutkimusongelmakohtaisesti.

- Kuuluuko tutkimukseesi esimerkiksi otoksen sisällä olevien ryhmien keskiarvojen vertailua tai muita otoksen osajoukkujen tunnuslukujen vertailua? Jos kuuluu, niin otoskoko tulee valita pienimmän ryhmään mukaan, jotta siitäkin saadaan tarpeeksi edustajia.
  - \* Mitä isompaa otosta käytetään, sitä pienempi perusjoukossa esiintyvä ryhmien välinen ero pystytään otoksella tunnistamaan.

### Tulosten vaaditun tarkkuuden vaikutus otoskokoon

- Tarkastelemme pian esimerkin avulla, kuinka tarvittavaa otoskokoa voidaan approksimoida tulosten halutun tarkkuuden avulla.
- Tarkastellaan kuitenkin ensin minkälaiset kysymykset liittyvät otoskoon pohdintaan tulosten tarkkuuden osalta.
  - Kuinka varma sinun on oltava, että tulokset vastaavat joukon mielipiteitä? Tämä on virhemarginaali.
    - \* Esimerkiksi puoluekannatuksen arvioimiseen 2 % virhemarginaalilla riittää huomattavasti pienempi otoskoko kuin 0.2 % virhemarginaalilla. Politiikan tutkija voisikin kasvattaa otoskokoa vähän lähestyessä, mikäli mielii saada tarkempia tuloksia.
  - Kuinka varma haluat olla, että otos edustaa joukkoa oikein? Tämä on luottamustaso.
    - \* Luottamustaso on todennäköisyys sille, että valitsemasi otos on tulosten kannalta oleellinen.
    - \* Jos joukosta poimitaan 30 otosta sattumanvaraisesti, kuinka usein yhdestä otoksesta saadut tulokset eroavat merkittävästi muista 30 otoksesta? Jos luotettavuustaso on 95 %, samat johtopäätelmät saadaan 95 prosentissa tapauksista.

### Odotetun vastauskadon vaikutus otoskokoon

- Kuinka suuri vastauskato tulee mahdollisesti olemaan?
  - Yleensä osa kyselytutkimukseen valituista jättää vastaamatta. Tätä kutsutaan kadoksi. Kato vinouttaa otosta, jos vastaamatta jättäneet ovat mielipiteiltään erilaisia kuin vastanneet.
  - Otoskoon kasvattaminen ei paranna kadon aiheuttamaa vinoutumista.
- Esimerkki: Jos Alkon myymälän asiakastutkimus suoritetaan ovensukselyynä maanantaina aamupäivällä, niin vastaajat eivät luultavasti edusta myymälän koko asiakaskuntaa. Otantakehikko on tässä liian suppea ja seurauksena on todennäköisesti vinoutunut otos. Vinoutuma ei korjaannu vaikka otosta kasvatetaan maanantai-aamupäivän asiakkaille.

**Esimerkki: otoskoko normaalijakauman odotusarvon estimoinnissa**

- Palautetaan mieleen normaalijakauman  $N(\mu, \sigma^2)$  odotusarvon luottamusvälin määräämisen (kun varianssi  $\sigma^2$  oletetaan tunnetuksi).
- Luottamusväliksi saatiin

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

ja luottamusvälin symmetrisyydestä johtuen luottamusvälin pituus

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

- Oletetaan, että normaalijakauman odotusarvoparametrille  $\mu$  halutaan konstruoida luottamusväli, jonka toivottu pituus on  $2d$  (huom. symmetrisyys).
- Luottamusvälin lausekkeesta saadaan täten järjestelemällä

$$n = \left( \frac{z_{\alpha/2} \sigma}{d} \right)^2.$$

- Jos varianssi  $\sigma^2$  on tuntematon, se voidaan kaavassa korvata havaitulla otosvarianssilla  $s^2$ , jolloin

$$n = \left( \frac{z_{\alpha/2} s}{d} \right)^2.$$

- Pitäydytään (yksinkertaisuuden vuoksi) luottuskertoimissa  $z_{\alpha/2}$  vaikka varianssi  $\sigma^2$  olisikin tuntematon.

**Esimerkki: otoskoko**

- Oletetaan, että haluamme määräätä otoskoon niin, että otoskesiarvo poikkeaa populaatiokesiarvosta korkeintaan yhden yksikön ( $d = 1$ ) todennäköisyydellä 0.05. Oletetaan, että varianssi on aiemmissa tutkimuksissa ollut  $\sigma^2 = 5$ . Oletetaan lisäksi, että taustallaoleva perusjoukko on iso (ääretön).
- Tällöin otoskoon tulisi olla

$$n \geq \left( \frac{z_{\alpha/2} \sigma}{d} \right)^2 = \left( 1.96 \sqrt{5} \right)^2 \approx 19.2.$$

- Tarvittavan otoskoon tulisi siis olla tässä tapauksessa noin 20.

### Äärellisyyskorjaus:

- Äärellisyyskorjausta käytetään, jos otos poimitaan äärellisestä perusjoukosta palauttamatta ja (nyrkkisääntönä)

$$\frac{n}{N} > 0.05,$$

jossa  $n$  on edelleen otoskoko,  $N$  perusjoukon koko ja  $n < N$ .

- Jos suhde  $n/N$  on lähellä arvoa 1, tarkoittaa se, että perusjoukosta huomattava osa kuuluu otokseen.
  - Tällöin otoskeskiarvon poikkeama populaatiokeskiarvosta on luonnollisesti pienempi kuin pienemmän otoksen tilanteessa.
  - Otoskoon kasvattaminen lisää siis estimoiminnan tarkkuutta, ja juuri äärellisyyskertoimen avulla hajonta “korjataan” vastaamaan käytettyä otoskookoa.

### Otoskoko, äärellisyyskorjaus: Normaalijakauman odotusarvon estiointi

- Oletetaan, että otannan taustalla oleva perusjoukko on äärellinen (pieni).
- Tällöin luottamusvälin konstruloinnissa huomioidaan äärellisyyskorjaus (vrt. aiemmat kaavat):

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

- Tarvittava otoskoko on tällöin välivaiheiden jälkeen

$$n = \frac{1}{\frac{d^2}{z_{\alpha/2}^2 \sigma^2} + \frac{1}{N}}.$$

### Esimerkki: otoskoko (jatkoa)

- Oletetaan aiemman esimerkin tilanne kuitenkin siten, että perusjoukon koko on nyt  $N = 100$ .

## 128LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Tällöin otoskooon tulisi olla

$$n \geq \frac{1}{\frac{1}{1.96^2 \times 5} + \frac{1}{100}} \approx 16.11.$$

- Tarvittava otoskoko on siis noin 17.

### Otoskoko: Suhteellinen osuus

- Palautetaan mieleen Bernoulli-jakauman odotusarvoparametrin  $p$  luottamusvälin muodostaminen.
  - Kuten normaalijakauman odotusarvoparametrin tapauksessa, pyrimme muodostamaan mahdollisimman lyhyen luottamusvälin, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman korkea.
- Oletetaan aiempaan tapaan, että  $p$ :lle halutaan muodostaa luottamusväli, jonka toivottu pituus on  $2d$ .
- Tarvittava otoskoko saadaan kaavasta (kun perusjoukko oletetaan ääretömäksi)

$$n = \left( \frac{z_{\alpha/2} \sqrt{p(1-p)}}{d} \right)^2.$$

### Suhteellinen osuus, äärellisyyskorjaus:

- Tarvittava otoskoko saadaan äärellisyyskorjausta käytettäessä kaavasta
$$n = \frac{Np(1-p)}{\frac{(N-1)d^2}{z_{\alpha/2}^2} + p(1-p)}.$$
  - Voidaan osoittaa, että jos perusjoukko  $N$  on iso (ääretön), niin tällöin edellinen lauseke supistuu aiempaan otoskokoon osoittavaan lausekkeeseen.
- Usein otoskokoa määrättääessä suhteellisesta osuudesta ei ole olemassa arviota.
- Tällöin suhteellisen osuuden  $p$  arvoksi asetetaan useimmiten  $p = 0.5$ , jolloin suhteellisen osuuden varianssi on suurin.

**Esimerkki: Otoskoko ja suhteellinen osuus**

- Geologi haluaa arvioida kallion kultapitoisuuden ottamalla kivenäytteen  $n$  eri pisteestä. Jokaisesta näytteestä havaitaan sisältyykö siihen kultaa. Kuinka suuri otos on poimittava, jotta kultapitoisuuden estimointivirheen  $d$  arvo on korkeintaan 0.05 todennäköisyydellä 0.95?
- Tässä kullen suhteellinen osuus on tuntematon, joten  $p$ :lle asetetaan  $p = 0.5$ .
- Äärellisyyskorjaus voidaan unohtaa, sillä näytteenottopisteiden pinta-alat ovat pieniä (eli niitä on äärettömän paljon, ts. tarkasteltavaan populaatioon niitä sisältyy hyvin suuri määrä).
- Tällöin otoskoko

$$n = \frac{1.96^2 \cdot 0.5 \cdot 0.5}{0.05^2} \approx 384.16.$$

## 6.8 Luvun 6 yhteenvetö, keskeisiä termejä ja kokonaisuuksia.

- Satunnaisotos
- Yhteisjakauma
- (Satunnaisotoksen) tilastollinen malli
- Yhteisjakauma
- Riippumattomat satunnaismuuttujat
- Otosjakauma
- Tunnusluku/otossuure
- Otosjakauma
- Estimaattori ja estimaatti
- (Hyvä) estimaattorin ominaisuuksia: Harhattomuus, tyhjentävyys, tehokkuus ja tarkentuvuus
- Otoskesiarvo ja otosvarianssi estimaattoreina
- Standardipoikkeama
- Normaalijakautunut satunnaisotos
- Standardoitu satunnaismuuttuja
- Standardoidun aritmeettisen keskiarvon otosjakauma
- Suhteellisen frekvenssin otosjakauma
- Luottamusväli
- Virhemarginaali

## **130 LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA**

- Luottamustaso ja luottamuskerroimet
- Normaalijakauman odotusarvon luottamusväli
- Luottamusväli: Suhteellisen osuuden odotusarvo
- Otoskoon määräämisen perusteita

## Luku 7

# Tilastollinen riippuvuus ja korrelaatio

- Tarkastelemme tässä luvussa tilastollisia tutkimusasetelmia, joissa on muun kankaan kaksi tai useampia **muuttujia**.
- 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 käsittellään koko tilastotieteen opinto-ohjelman lävitse, kuten perusteellisesti [Lineaariset ja yleistetyt lineaariset mallit -kurssin](#) myötä.

### 7.1 Muuttujien väliset riippuvuudet

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

- 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) kasvuvaahdista Suomessa, Suomen viennin volyyymista sekä BKT:n kasvuvaahdista muissa EU-maissa ja USA:ssa? Taloustieteilijät pyrkivät yleisesti löytämään muitakin lainsäädäntöä. 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 käytettävissä olevista tuloiista ja alkoholin saatavuudesta?
- Miten todennäköisyys sairastua keuhkosyöpään riippuu tupakointin määstä 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 (täydellisesti) 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 määrittämisessä ja mahdollisesti myös ennustamisessa.
- Tilastollinen riippuvuus ja **korrelaatio**
  - Kahden muuttujan välistä (lineaarista) tilastollista riippuvuutta kutsutaan tilastotieteessä (tavallisesti) **korrelatioksi**.
  - Korrelaation eli (lineaarisen) tilastollisen riippuvuuden voimakkuita mittavia 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. selittävistä muuttujista pyritään mallintamaan regressiomalliksi kutsuttavalla tilastollisella mallilla. Käsittelemme johdantoa regressioanalyysiin 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 kuvilla ja esitellä tiivistämällä havaintoarvoihin sisältyvä informaatio sopivaan muotoon:
  - Havaintoarvojen jakaumaa kokonaisuutena voidaan kuvata sopivasti valituilla graafisilla esityksillä.
  - Havaintoarvojen jakauman karakteristisia ominaisuuksia voidaan kuvata sopivasti valituilla otostunnusluvuilla (ks. otostunnusluvut ja otosjakaumat luvussa 6).
- Koska useampiulotteisten kuvioiden kuin kaksiulotteisten muodostaminen ei ole usein kovin mielekästä, kolmen tai useaman muuttujan havaintoaineistoja havainnollistetaan tavallisesti niin, että muuttuja tarkastellaan pareittain.

- Kahden järjestys-, välimatka- tai suhdeasteikollisen muuttujan havaittujen arvojen parja havainnollistetaan tavallisesti graafisella esityksellä, jota kutsutaan hajontakuvioksi tai pistediagrammiksi (“pistekaavio” engl. scatter plot). Ks. esimerkiksi kuva 7.1.
- Usean muuttujan havaintoaineistojen karakteristisia ominaisuuksia voidaan kuvata muuttujakohtaisilla otostunnusluvuilla.
- Muuttujakohtaiset otostunnusluvut eivät kuitenkaan voi antaa informaatiota muuttujien välisistä riippuvuuksista.
- Muuttujien pareittaisia tilastollisia riippuvuuksia voidaan kuvata sopivasti valitulla korrelaation mitalla.

### Pistediagrammi (hajontakuvio)

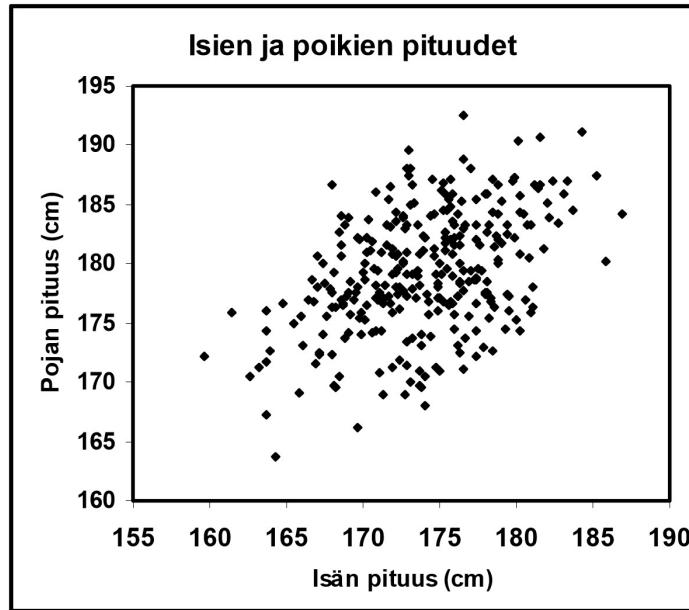
- Tarkastellaan tilannetta, jossa tutkimuksen kohteina olevista havaintoyksiköistä on mitattu kahden järjestys-, välimatka- tai suhdeasteikollisen muuttujan  $X$  ja  $Y$  arvot.
- Muuttujien  $X$  ja  $Y$  arvojen samaan havaintoysikköön liittyvien parien  $(X, Y)$  muodostamaa havaintoaineistoa voidaan kuvata graafisesti pistediagrammillä.
- 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 havaintoysikköön kaikille  $i = 1, 2, \dots, n$ . Havaintoarvojen parien  $(x_i, y_i)$  pistediagrammi saadaan esittämällä lukuparit niiden määrittelemien pisteiden tasokoordinaatistossa.

**Esimerkki: Isän ja pojien pituus**

- Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.
- Periytyykö isän pituus heidän pojilleen?
- Havaintoaineisto koostuu 300:ta isän ja heidän poikiensa pituuskysymyksistä 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 tunnusluvuilla:
  - 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 käytettäväissä 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  $i = 1, 2, \dots, n$  muodostaen havaintoyksikkökohtaisia havaintoarvojen pareja  $(x_i, y_i)$ .
- 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 pisteiden painopiste.
- Havaintoarvojen aritmeettinen keskiarvo kuvaa havaintoarvojen keskimääräistä sijaintia.
- Osoittautuu, että (aritmeettinen) keskiarvo toimii tilastollisessa mielessä hyvänä 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 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 tulomentti-) korrelatiokerrointa  $\rho_{XY}$  ja sen estimointia.
- Tällä kurssilla emme tarkastele tarkemmin tilastollisia testejä korrelatiokerioimelle  $\rho_{XY}$ , kuten:
  - Yhden otoksen testi korrelatiokerioimelle
  - Korrelatiokerioimien 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

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

satunnaismuuttujien  $X$  ja  $Y$  odotusarvot ja

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

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

- Satunnaismuuttujien  $X$  ja  $Y$  korrelaatiota

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

kutsutaan ajoittain siis **Pearsonin korrelatiokertoimeksi** (tulomomenttikorrelatiokertoimeksi).

- Pearsonin korrelatiokerroin  $\rho_{XY}$  mittaa satunnaismuuttujien  $X$  ja  $Y$  lineaarisen riippuvuuden voimakkuutta. Ts. sm:jien välistä (lineaarista) yhteyttä.
- Pearsonin korrelatiokerointa voidaan estimaoida Pearsonin **otoskorrelatiokertoimella**.

### 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  $s_y$ .

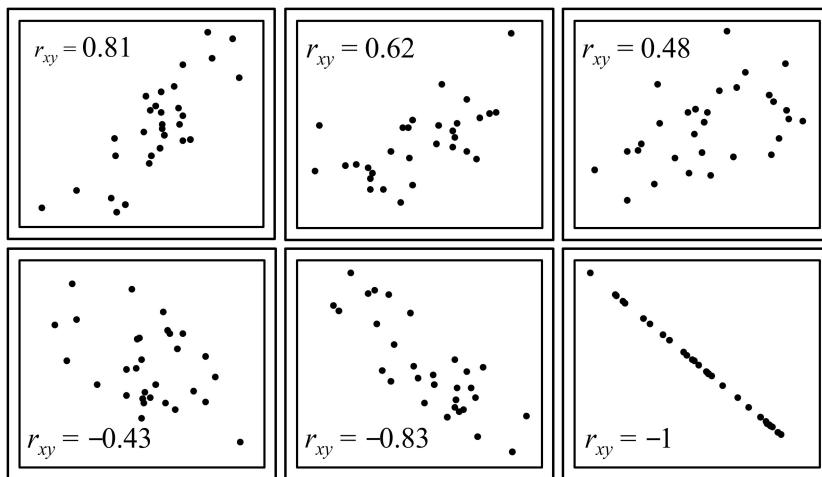
Ts. havaintoarvojen pareista  $(x_i, y_i), i = 1, 2, \dots, n$ , laskettu Pearsonin otoskorrelatiokerroin on siis

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

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

- Otoskorrelatiokertoimen estimaattori voidaan johtaa sekä momenttimenetelmällä että suurimman uskottavuuden menetelmällä, jotka ovat tyyppilisiä estimointimenetelmiä tilastotieteessä ja tarkemmin tilastollisessa päättelyssä.
- Otoskovarianssi:
  - Huomaa, että  $x$ - ja  $y$ -havaintoarvojen otoskovariansit 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  $\alpha, \beta \neq 0$
  - 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.

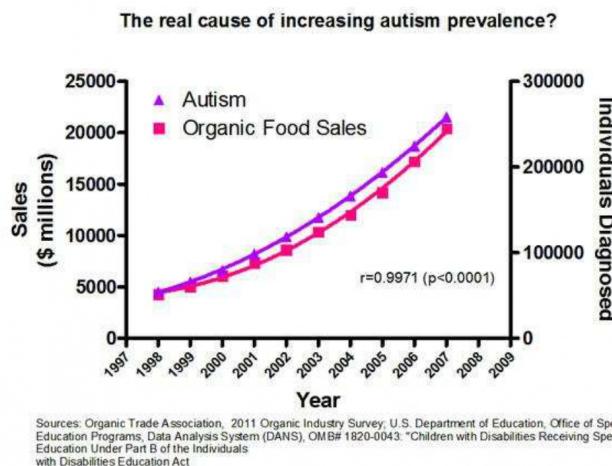
- Vaikka  $r_{xy} = 0$ ,  $x$ - ja  $y$ -havaintoarvojen välillä saattaa silti olla jopa eksakti epälineaarinen riippuvuus.
- **Havainnollistus:** Alapuolella esitettävät kuviot (Kuva 7.2) havainnollistavat kahden muuttujan havaittujen arvojen ( $n = 30$ ) pistediagrammin ilmeen ja korrelaation välistä yhteyttä.



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

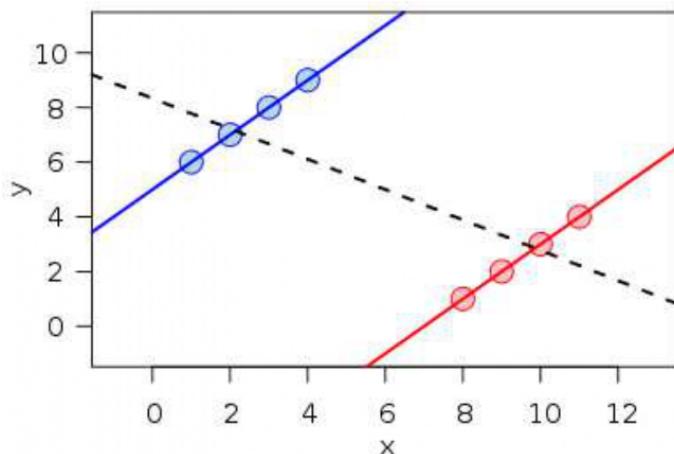
- Ks. seuraavasta [linkistä](#) lisää havainnollistuksia.
  - *Guess the correlation* pelissä pääset arvioimaan esitettävän pisteparven korrelaation voimakkuutta erilaisissa simuloiduissa tilanteissa: <http://guessthecorrelation.com/>
- **Kausaalisuus**
  - Muuttujan  $x$  arvojen muutos vaikuttaa muuttujan  $y$  arvoihin (syvaikutussuhde), jos seuraavat kolme ehtoa täytyvät:
    - \* muuttujan  $x$  muutos esiintyy ajallisesti ennen  $y$ :n muutosta.
    - \* muuttujissa  $x$  ja  $y$  tapahtuvien muutosten välillä on riippuvuutta.
    - \* muuttujassa  $y$  tapahtunutta muutosta ei voida selittää millään muilla tekijöillä.
  - 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 kausaalisuude kuvaa: **kahden muuttujan yhteisvaihtelu ei riitä todisteeksi siitä, että kyseessä olevien muuttujien välillä on kausaalista yhteyttä**
- Yhteisvaihtelu voi johtua myös kolmannen muuttujan vaikutuksesta molempien muuttujaan tai virheellisestä otannasta, vaikka muuttujat olisivatkin perusjoukossa toisistaan riippumattomia



Kuva 7.3: Esimerkkejä: luomuruoka syypää lisääntyneisiin autismitapauksiin?

- Simpsonin paradoksi
  - Simpsonin paradoksi syntyy, kun kahden muuttujan välinen korrelaatio muuttuu pääväistaiseksi, otettaessa huomioon jokin kolmas muuttuja, joka korreloii molempien muuttujien kanssa



Kuva 7.4: Simpsonin paradoksi

**Esimerkki: Berkeleyn sukupuolisyrjintää**

Yksi tunnetuimmista esimerkeistä Simpsonin paradoksista on Berkeleyn yliopiston sukupuolisyrjintätapaus. Yliopisto haastettiin oikeuteen vuonna 1973 sukupuolisyrjinnästä. Väitettiin, että yliopistoon olisi miesten helpompi päästää kuin naisten, sillä yhteensä 8442:sta mieshakijasta 44 % hyväksyttiin kun samat luvut olivat naisilla 4321 ja 35 %. Mieshakijoista pääsi siis 9 prosenttiyksikköä enemmän sisälle kuin naisista.

- Tarkasteltaessa erikseen eri tiedekuntia huomataan, että itseasissa useammassa tiedekunnissa naisia on päässyt sisälle isompi osuus hakijoista. Aineisto kuudesta isoimmasta tiedekunnasta on listattu alla olevaan taulukkoon.

Miehet			Naiset	
Tiedekunta	Hakijat	Hyväksyttyt %	Hakijat	Hyväksyttyt %
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 tiivistään korrelatiokertoimen tulkintavirheitä aiheuttavat useimmiten seuraavat seikat:
  - Riippuvuudesta ei välttämättä seuraa syy-seuraussuhdetta.
  - Kolmas muuttuja eli kahden muuttujan välinen yhteys selittyy yhteenestää syystä (esimerkiksi lämpimästä kesästä).
  - 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.

## 7.5 Luvun 7 yhteenveton, keskeisiä termejä ja kokonaisuuksia.

- Eksakti ja tilastollinen riippuvuus
- Korrelaatio
- Kovarianssi
- Pearsoninkin korrelatiokerroin ja otoskorrelatiokerroin
- Pistediagrammi
- Kausalisuus vs. korrelaatio



## Luku 8

# Regressioanalyysi

Tilastollinen riippuvuus ja korrelaatio -jakson laajennuksena pyrimme tässä luvussa vastaamaan seuraavaan kysymykseen: *Miten jonkin selittävän muuttujan tilastollista riippuvuutta joistakin toisista, selittäväksi muuttujiksi kutsutuista muuttujista voidaan mallintaa?* Muuttujien välisen riippuvuuden, eli erilaisten tosielämän asioiden ja ilmiöiden välisen yhteyksien analysointi on tavallisesti keskeinen kysymys tieteellisessä tutkimuksessa. Regressioanalyysi 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 kriteeriston 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: **selittävä muuttuja**, jonka tilastollista vaihtelua pyritään selittämään **selittävän muuttujan**, tai **selittävien muuttujien**, vaihtelulla.
  - Toisin sanoen, pyritään erottamaan se selittävän 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 se loppuu, 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**.
- **Lineaarinen regressioanalyysi** siis (teknisesti) rajoittuu muuttujien *lineaarisesti* riippuvuuksien kuvaamiseen. Kuitenkin, laajemmin asiaa pohdittaessa, lineaaristen regressiomallien suuri käyttökelpoisuus muuttujien välisen riippuvuusien tilastollisessa analyysissa perustuu (ainakin) seuraaviin seikkoihin:
  - Lineaarilla regressiomallilla voidaan usein vähintään kohtuullisella (riittävällä) tarkkuudella approksimoida epälineaarisiakin muuttujien väliä riippuvuksia!
  - 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 tilastotieteiden 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.
  - Luokittelua regressiomallin funktionaalisen muodon mukaan:
    - \* Lineaariset regressiomallit
    - \* Epälineaariset regressiomallit
  - Luokittelua regressiomallin yhtälöiden lukumäärän mukaan:
    - \* Yhden yhtälön regressiomallit
    - \* Moniyhtälömallit

Tällä kurssilla käsitellään vain **lineaarisia yhden yhtälön regressiomalleja**. Kuitenkin luvussa 8.3 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 vaihtelun 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 tulosmuuttujaksi*
  - Vastaavasti selittävää muuttuja kutsutaan paikoin *selittäjäksi, riippumattomaksi muuttujaksi tai ennustavaksi muuttujaksi*.
- Tässä luvussa tarkastellaan lyhyesti ja tiivistetysti seuraavia yhden selittävän muuttujan lineaarisen 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 vaihtelua halutaan selittää selittävän muuttujan eli selittäjän  $x$  havaittujen arvojen vaihtelun avulla. Tulkitaan selittävä muuttuja tässä kohtaa kiinteäksi eli sen arvot oletetaan tunnetuksi.<sup>1</sup>
- Tehdään siis seuraavat oletukset:
  - i) Selitettävä muuttuja  $Y$  on suhdeasteikollinen satunnaismuuttuja.
  - ii) Selittävä muuttuja  $x$  on kiinteä eli ei-satunnainen muuttuja.
- 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  $i = 1, 2, \dots, n$ .

---

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

- Matemaattisemmin tämä tarkoittaa sitä, että tällöin havaintoarvot muodostavat pisteitä 2-ulotteisessa  $(x_i, Y_i)$  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 selitettävä muuttujan  $Y$  satunnainen ja havaittu arvo havaintoyksikölle  $i$ .
  - $x_i$  selittävä muuttujan eli selittäjän  $x$  ei-satunnainen ja havaittu arvo havaintoyksikölle  $i$ .
  - $\varepsilon_i$  on virhetermi (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 regressiokertoimeksi. 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 päde esitettävässä muodossa, jos mallissa ei ole vakiota (vakioselittääjää), joka yleensä automaattisesti 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: Regressiokertoimet  $\beta_0$  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 yhtiseksi oletettua varianssia kutsutaan ajoittain jäännösvarianssiksi.
  - iii) Virhetermit 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 modifioituna), kun siirrytään tarkastelemaan selittävän muuttujan ehdollista odotusarvoa selittäjien suhteen.
  - Voidaan (ajoittain) olettaa, että selittävä muuttuja ja selittäjät noudattavat yhdessä **multinormaalijakaumaa** eli aiemmin esitellyn yksilotteisen normaalijakauman moniulotteista laajennusta.
  
  
  
- Regressioanalyysille voidaan esittää kaksi asialoogisesti varsin erilaista lähtökohtaa, joilla on kuitenkin myös monia yhtymäkohtia:
  - i) Ongelmat determinististen mallien sovittamisessa havaintoihin: Havainnoille esitetty malli ei sovi täsmällisesti kaikkiin havaintoihin. Tämä onkin osaltaan tilastollisen mallinnuksen yksi ominaispiirteistä: Täydellistä sopivuutta aineiston kanssa ei käytännössä koskaan saavuteta.
  - ii) Tavoitteena on moniulotteisen todennäköisyysjakauman regressiofunktion parametrien estimointi.
  - Vaikka moniulotteisten todennäköisyysjakaumien regressiofunktiot ovat yleisesti epälineaarisia, lineaariset regressiomallit muodostavat tärkeän ja paljon sovelletun mallihuokan.
- 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  $y_i$ ,  $i = 1, 2, \dots, n$  käyttäen.
  - 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:

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

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

- Nämä ns. **normaaliyhtälöt** johtavat lopulta pienen sieventämisen jälkeen regressiokertoimien  $\beta_0$  ja  $\beta_1$  pienimmän neliösumman (PNS-) estimaattoreihin (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 (matemaatisesti 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  $y$  saavat **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. Vidaan 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 lineaariselle malliin liittyviä sovitteita ja residuaaleja.
  - 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 selittäväälle muuttujalle antama arvo havaintopisteessä  $x_i$ . Vastaavasti residuaali on selittäväinen muuttujan havaitun arvon  $y_i$  ja sovitteen  $\hat{y}_i$  eli estimoidun regressiosuoran yhtälön selittäväille muuttujalle havaintopisteessä  $x_i$  antaman arvon erotus.
  - Estimoitu regressiomalli selittää selittävän muuttujan havaittujen arvojen vaihtelua sitä paremmin mitä lähempänä estimoidun mallin sovitteet  $\hat{y}_i$  ovat selittävän muuttujan havaittuja arvoja  $y_i$ .
  - Yhtäpitävästi edellisen kanssa: Estimoitu regressiomalli selittää selittävän muuttujan havaittujen arvojen  $y_i$  vaihtelua sitä paremmin mitä pienempiä ovat estimoidun mallin residuaalit  $\hat{\varepsilon}_i$ .

- Liitteen 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 aineistoon. 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 **ylios-vittamiseen** usean selittäjän lineaarisia malleja käsiteltäessä.

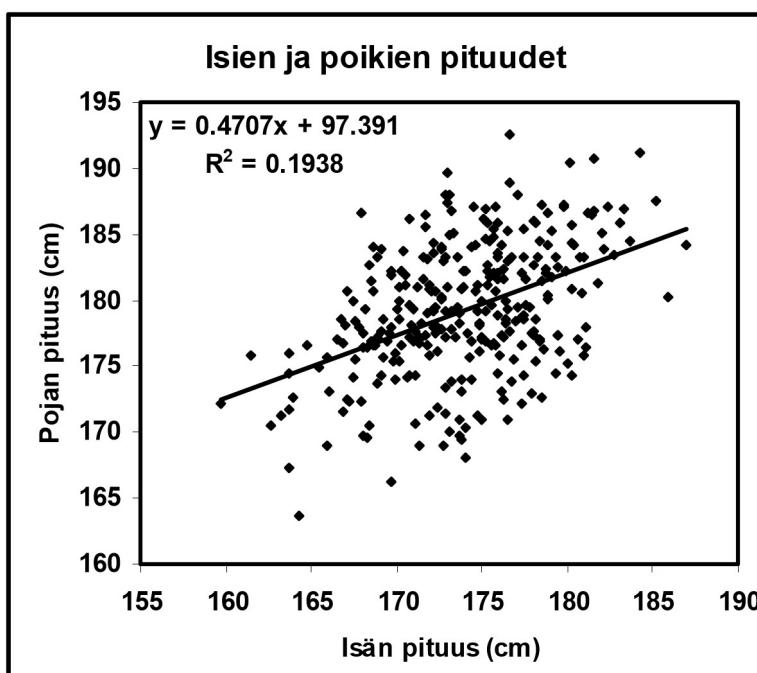
**Esimerkki: isän ja poikien pituus, tarkemmin**

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 **binäärisen** ( $Y_i = 0$  tai  $Y_i = 1$ ) tai **lukumäärä**  $Y_i \in \{0, 1, 2, 3, \dots\}$
  - Mikäli vaste on binäärisen, 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 malliluokka on ns. **Poisson-regressiomalli**. Tässä yhteydessä oletetaan siis, ettei sm.  $Y$  noudattaa Poisson-jakaumaa ja regressiomalli rakennetaan tämän oletuksen ympärille.
- **Vastemuuttajan roolin/luonteen selvittäminen on hyvin keskeistä tilastollista mallia rakennettaessa.** Tässä pätee samat eroavaisuudet mitkä tulevat tutuiksi todennäköisyyslaskennan kursseilla kun käsitellään diskreettien ja jatkuva-arvoisten satunnaismuuttujien jakaumia ja näihin liittyviä yksityiskohtia.
- Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään useilla myöhemmillä tilastotieteen kursseilla.
  - Erityisesti aineopintojen kurssien **TILM3561 Tilastollinen päättely I** ja **TILM3562 Tilastollinen päättely II** jälkeisellä **TILM3588 Lineaariset ja yleistetyt lineaariset mallit -kurssilla**, jossa 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.

### 8.4 Luvun 8 yhteenvetö, keskeisiä termejä ja konaisuuksia.

- Lineaarinen regressioanalyysi

*8.4. LUVUN 8 YHTEENVETO, KESKEISIÄ TERMEJÄ JA KOKONAISSUUKSIA.*155

- Yhden selittäjän lineaarinen regressiomalli
- Selitettävä muuttuja
- Selittävä muuttuja/selittävät muuttujat
- Virhetermi
- Pienimmän neliösumman (PNS) menetelmä
- PNS-estimaattorit ja PNS-estimaatit
- Sovitteet ja residuaalit
- Selitysaste



# 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 tieteeseen, 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 yleisemällä 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 miinkään 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 käsiteltyjä asioita ja yleisiä elementtejä, jotka jokaisen tieteentekijän tulee hallita.

### 9.1 Tilastollisen tutkimuksen yhteisiä elementtejä

#### 1. Satunnaisvaihtelu

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

## 158LUKU 9. TILASTOTIETEEN ROOLI UUDEN TIEDON TUOTTAMISESSA

satunnaisvaihtelua ilmentävää aineistoa on (tai tulisi olla) tilastotieteellistä.

- Tilastollisen tutkimuksen tavoitteena on (useimmiten) pyrkiä erottamaan satunnaisilmiön systemaattinen ja satunnainen vaihtelu. Tämä vaatii substanssiosaamisen lisäksi menetelmäosaamista sekä hyvää tilastotieteellistä intuiota.
- Satunnaisvaihtelun “välttämättömyys” satunnaisilmiöiden tutkimuksessa on tiedostettava ja ymmärrettävä. 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 olioista, ilmiöistä ja toisaalta myös rajoituksista.
  - Lisäksi tutkimusongelman holistinen käsitteily 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 pakkinoit aseettomia: mitkä ympäröivistä ilmiöistä 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 korkealotteisesta aineistosta, eli aineistosta jossa jokaiselta tutkimusyksiköltä mitataan jopa miljoonia eri muuttujia, kuten DNA-tutkimuksessa genomitietoa.<sup>1</sup>

#### • Hahmottamisen vaiheet:

<sup>1</sup>Tilastotieteessä näitä menetelmiä kutsutaan monimuuttujamenetelmiksi ja niitä käsitellään tarkemmin erikoiskursseilla [TILM3704 Monimuuttujamenetelmät](#) sekä [TILM3611 Monimuuttujamenetelmien jatkokurssi](#)

- “Todellisen” järjestelmän operationalisointi kvantitatiiviseksi kuvaukseen 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 väriä, mutta jotkut ovat käyttökelpoisia.
  - Tilastollinen malli on “vain” kuvaus aineiston sisältämästä vaiheesta: se ei käytännössä 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 on 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>2</sup>
- Tutkimusta ei yleensä ole mahdollista jakaa täysin selvästi erillisin ja ajallisesti 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 kulussa.
  - Päätelmien tieteellisyden (periaatteellinen) tarkistusmahdollisuus, ja nykyään yhä useammin jo toistettavuus, on tärkeää.
- Usein saattaa kuitenkin olla järkevää jäsentää tutkimuksessa kohdattavia tehtäviä ja vaiheita sekä niiden välistä suhteita osana tutkimusprosessia.
  - Tutkimuksen lähtökohtana on jokin ongelma, johon tutkimuksen avulla etsitään vastausta.
  - 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 asemoiдаan 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 päätelmien tekemiseen.
- Havaittava tieto
  - Yleensä ajatellaan, että todellisuudesta saadaan tietoa tavalla taikka toisella havaintoja tekemällä.
  - Havaittava tieto ei mitenkään pysty kattamaan kaikkea tutkimuskohdeeseen liittyvää ja toisaalta ymmärtämiseen tarvittava havaintomaailman hahmotus tuottaa ideologisesti ja historiallisesti sitoutuneita yksinkertaistavia sekä luonteeltaan usein hyvin teoreettisia abstraktioita.
- Operationalisointi: Siirrytään teoriasta empiriaan

<sup>2</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 systeemattisuusvaatimusten 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 siirrytään 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 tulkinta edellyttää, että niitä on tulittava siinä kontekstissa, josta ne ovat peräisin.
  - \* On esim. mahdollista, että esitetty kysymys ei välttämättä vastaa tutkimuskohteteen ominaisuuksia.
- Aineisto eli data
  - 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 käsitetty 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.

## 162 LUKU 9. TILASTOTIETEEN ROOLI UUDEN TIEDON TUOTTAMISESSA

- Analyysien tulokset on tulkittava eli käännettävä ne takaisin empiirian kieleltä teorian kielelle. Tavoitteena on siis substanssitietouteen perustuen tuottaa uutta tietoa siten, että se lisää myös substanssitietoutta
- Tulkinnan voi ajatella olevan operationalisoinnin käänteistapahtuma: Tutkimuksen läpiviennin sekä tulkinnan kannalta onnistunut operationalisointi ovat loppujen lopaksi yksi ja sama asia.
- Raportointi
  - Parhaimmillaan tutkimusraportti on vakuuttava, ja periaatteessa (ja toivottavasti) toiston mahdollistava, kuvaus tutkimusprosessin kaikista vaiheista, 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 yleistettävyyttä ja analyysin arvioitavuutta ja uskottavuutta tulisi pohtia raportissa. Tutkimuksen kuluessa tehdyt valinnat tulisi perustella tiedostaen muukaan myös omat arvopainoitteiset valinnat (ja ehkä oletuksetkin).

**Esimerkki tilastollisesta kyselytutkimuksesta** (lisää esimerkkejä myöhemmin luvussa 10).

- Päätöksentekijät ja tiedotusvälineet kartoittavat säännöllisin välein suomalaisten mielipiteet erilaisista yhteiskuntaa koskevista kysymyksistä.
  - Esimerkkejä:
    - \* Miten suomalaiset suhtautuvat NATO-jäsenyyteen?
    - \* Miten suomalaiset suhtautuvat ydinvoiman lisärakentamiseen (osana vihreää siirtymää)?
    - \* Mitkä ovat poliittisten puolueiden kannatussuudet?
- 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 kohteeksi poimittujen suomalaisten joukko muodostaa edustavan pienoiskuvan Suomen kansasta (huom. aiemmin käsitellyn onnistuneen otannan idea ja vaatimukset).

- \* 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 arvonnoissa 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.

**Esimerkki: laadunvalvonta (Mellin, 2006)**

Tehdas valmistaa korkealuokkaisia sulkimia kameroihin. Tehdas pyrkii siihen, että yli 90 % sulkimista kestää vähintään 100 000 kameran laukaisua. - Sulkimien laadunvalvonta on toteutettu seuraavalla tavalla: - (i) Tuotantolinjalta poimitaan arpomalla joukko sulkimia rasituskokeeseen. - (ii) Rasituskokeessa määräätään vähintään 100 000 laukaisua kestävien sulkimien suhteellinen osuus. - Kokeen tavoitteena on tehdä kokeen tulosten perusteella yleisiä johtopäätöksiä 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 perusjoukossa. - Rasituskokeen kohteiden poiminta arpomalla on ainoa menetelmä, joka mahdollistaa edustavan pienoiskuvan saamisen. - Rasituskokeen kohteiden poiminta kaikkien valmistettujen sulkimien muodostamasta perusjoustosta arpomalla merkitsee satunnaisotannan soveltamista ja tutkimuksen kohteeksi poimittu perusjoukon osa toimii muodostettavana (satunnais)otokseksena. - Arvonnan käytöö 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 pystytä arvioimaan, ellei kokeen kohteiden poiminnassa ole käytetty satunnaisotantaa. - Kokeen suunnittelussa, toteutuksessa ja tulosten analysoinnissa sovelletaan mm. seuraavia tilastollisia menetelmiä: koesuunnittelu, otanta, estimointi ja testaus.

### **9.3 Luvun 9 yhteenvetö, keskeisiä termejä ja konkaisuuksia.**

- Tässä luvussa vedetään yhteen paljon mm. otantaan liittyviä seikkoja ja laajennetaan niiden merkitystä tilastotieteellisen tutkimuksen osana.

*9.3. LUVUN 9 YHTEENVETO, KESKEISIÄ TERMEJÄ JA KOKONAISSUUKSIA.*165

- Satunnaisvaihtelun merkitys
- Ilmiön ja ongelman hahmottaminen ja muotoilu tilastolliseksi tutkimusasetelmaksi
- Tilastollisen mallin muodostaminen, siihen perustuva tilastollinen päättely ja synteesi ilmiön ymmärtämiseen liittyen

166 LUKU 9. TILASTOTIETEEN ROOLI UUDEN TIEDON TUOTTAMISESSA

## Luku 10

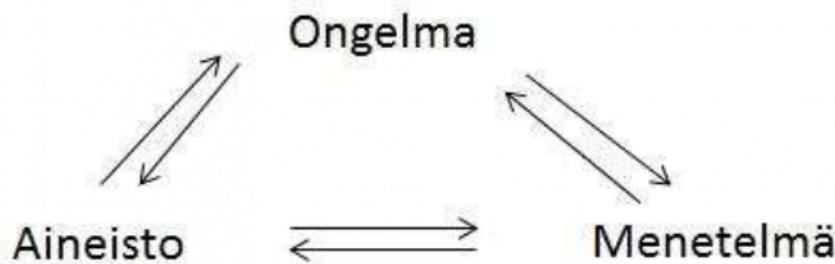
# Aineisto- ja tutkimustyyppit ja koeasetelmat

Tässä luvussa käsitellään erilaisia tapoja toteuttaa tilastollista tutkimusta. Emppiirisen tutkimuksen lähtökohtana on aina tutkimusongelma, joka sisältää kysymyksien tai kysymyksiä, joihin tutkimuksella haetaan vastauksia. Tilastotieteen näkökulmasta tutkimusongelman keskiössä on kuitenkin **aineisto**, **data**, ja se miten käytettävissä olevasta aineistosta saadaan vastauksia tutkimuskysymyksiin. Tarkastelemme tässä luvussa myös tutkimuksenteon käytäntöä käsitteellä erilaisia aineistotyyppejä. Käymme läpi eri alojen ja tutkimusongelmien käytännön tutkimustyyppisissä 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 määritellään 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 esimerkiksi seuraaviin kysymyksiin:

- Miten tilastotiede liittyy tiedon keruuseen?
- Miten aineisto generoituu?
- Millaisiin kysymyksiin saadaan kussakin asetelmassa vastauksia?
- Tarkemmat asetelmiin ja analyyseihin liittyvät yksityiskohdat käsitellään [TILM3579 Kokeiden suunnittelu 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 tutki-mukset voidaan karkeasti jakaa neljään eri luokkaan: **kuvaileva**, **vertaileva**, **kokeellinen** ja **havainnoiva** tutkimus.

- **Kuvaileva tutkimus**

- Tarkoituksesta on kuvata jonkin ilmiön, tilanteen tai tapahtuman luonnetta, yleisyyttä, historiallista kehitystä tai muita tunnuspiirteitä mahdollisimman todennäköisesti ja tarkasti.
- Keskeistä **tiemon lisääminen** ja pyrkimys vastata kysymyksiin **mitä, millainen tai miten**.
  - \* Yleisesti ottaen kuvaileva tilastollinen tutkimus perustuu aineiston laskettuihin tunnusluvuille, jotka kuvaavat aineiston ominaisuuksia. Esimerkkinä toimivat kesiarvon lisäksi sen kaltaiset keskimääräistä havaintoja mittavaat suuret kuten mediaani ja moodi tai vaihtelua kuvaavat eri muuttujien vaihteluvälit ja keskihajonnat (ks. luku @ref{luku6}).
- Saadakseen luotettavia tunnuslukuja, tulee otoksen olla edustava ja havaintojen luotettavia ja päteviä eli saatujen mittausten pitää kuvata kohteena olevaa ilmiötä ilman virheitä.
- Kuvailevassa tutkimuksessa ei tutkita muuttujien väisiä yhteyksiä tai riippuvuuksia eikä täten yleensä tehdä jakoa selittäviin ja selittäviin muuttuijiin vaan muuttujat ovat asetelmallisesti samantasoisia.
  - \* Vastaavasti kuvailevassa tutkimuksessa ei välttämättä testata hypoteeseja, ei tehdä ennusteita, ei anneta selityksiä tai pohdita

seurausia: 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 väliä 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, seurausille.
  - \* Syy-seuraussuhteita ei kuitenkaan vertailevassa tutkimuksessa pohdita, ts. vertaileva tutkimus ei ole suoranaisesti kiinnostunut kohtena olevien ilmiöiden/ryhmiä vertailussa löydettyjen erojen syistä vaan mielenkiinnon kohtena on kys. erot itsessään.
- Vertailevaa tutkimusta tehdessä on tarpeen pohtia:
  - \* Miksi jotakin tutkimuskohdetta vertaillaan eli mitä tutkimuskohdeesta 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 laboratoriassa tai sairaalan koehuoneissa, jotta mittaukset ja käsittelytekijät on tutkimuksen tekijän puolesta kontrolloitu ja täten halutunlaisia.

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

- \* Tutkimusasetelman kontrollointi vähentää mittauksiin ja käsitteleytöihin 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 lainsäädännölliisiin ongelmiin 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 yleistettävyyttä: liian pitkälle kontrolloidut ja pelkistetyt koeolosuhteet eivät toimi kaikkien tutkimuskysymysten kannalta yleistettävyyden 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?

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

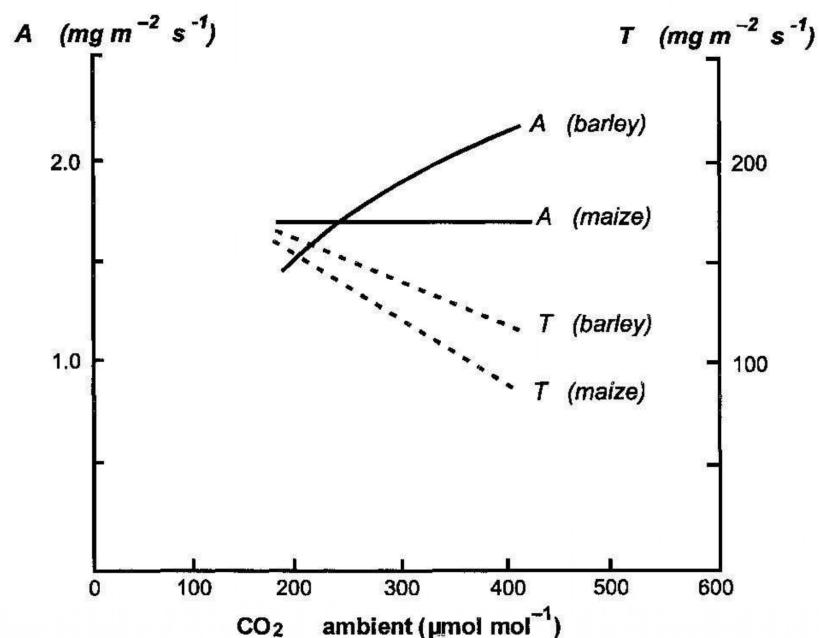


Figure 3: Measured canopy responses to ambient  $\text{CO}_2$  of photosynthesis (A) and transpiration (T) in barley (C3) and maize (C4) (from: Goudriaan and Unsworth, 1990).

Kuva 10.2: Hiilidioksidipitoisuuden kasvun vaikutus satomääriin.

### 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:
    - \* 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 mahdollistaa samankaltaisten ryhmien saamisen.
    - \* Kokeen kohteiden jakamista erilaisen käsitteilyn kohteiksi joutuvien ryhmiin arpomalla kutsutaan siis **satunnais-tamiseksi**.
  - Arvonnан käyttö ryhmiin jaossa merkitsee sitä, että koetulokset ovat satunnaisia seuraavassa mielessä: Jos arvontaa toistetaisiin, kokeesta saataisiin (suurella todennäköisyydellä) erilaiset ryhmäjaot.
- 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 mahdollistaa

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-seuraussuhdeita tarkastellaan tilanteissa, joissa tutkijalla ei ole välttämättä mitään kontrollia (tai syytä sille) tutkimusyksikköihin tai heihin vaikuttaviin muutujiin (käsitteleytekijöihin).
  - \* Esimerkiksi tutkimusasetelmat, joissa tutkimuksen kohteena olevia yksiköitä (esim. ihmiset, kunnat, valtiot) ei voida satunnaisestaan kuuluvaksi osaksi joukkoa, joka altistetaan jollekin käsittelylle.
- Tällöin tutkijan on tyydyttävä havainnoimaan sitä mitä tapahtuu luonnonstaan 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 käytäytymisestä.
  - \* 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 kielellisiä 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 emotionaaliseksi tutkittavaan ryhmään tai tilanteeseen.

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

- Kokeellinen asetelma: Poimitaan satunnaisesti  $n$  kappaletta raskaana olevia naisia ja heistä  $n_1$  kappaletta satunnaistetaan käsitellyryhmään (raskauden keskeytys) ja  $n_2$  kontrolliryhmään. Kaikki naiset käyvät muutaman seuraavan vuoden ajan syöpäseulonnoissa.
- Kokeelliseen asetelma ei selvästikään 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 tilastoyksikötä: **poikkileikkaus-** ja **pitkittäistutkimuksiin**. Tilastoyksikköjen erilainen ryhmittely tuottaa erilaisia aineistotyyppejä, jotka voidaan jaotella karkeasti kolmeen eri tyyppiin

- 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 hyödyntäävät eri aineistotyyppejä sen mukaan, miten ne sopivat tutkimuskysymykseen ja valittuun menetelmään. Tarkastellaan seuraavaksi mitä em. kaksi tutkimustrategiaa tarkoittavat, miten ne eroavat ja minkälaisia tutkimustyyppejä-, asetelmia- ja aineistoja kumpaankin kuuluu.

### 10.2.1 Poikittaistutkimus eli poikkileikkaustutkimus

- Poikittaistutkimukseksi kutsutaan tutkimusstrategiaa, jossa tarkoituksesta on tutkia kohdetta tai ilmiötä laaja-alaisesti tietynä ajankohtana käytetään poikkileikkausaineistoja.
  - Voidaan tarkastella useita ryhmiä, joissa on esimerkiksi eri-ikäisiä henkilöitä ja ryhmistä saattua tietoa vertaillaan toisiinsa.
  - Voidaan käyttää kuvailemaan riskisuhteita (odds ratio) tai kuvailemaan tiettyyn populaation osaan kohdistuvaa ilmiötä tai riskiä (esimerkiksi sydän- ja verisuonitaudit).
    - \* Esimerkiksi tutkitaessa sydän- ja verisuonitauteja binäärисellä vastemuuttujalla käytetään aineistoa, joka koostuu eri ikäisistä ja kuntoisista ihmisiästä voidaan arvioida iän ja muiden muuttujien vaikutuksia sydän- ja verisuonitauteihin sairastumisen riskitekijöinä.
  - Poikittaistutkimuksessa ei saada tietoa tilastoysiön mielenkiinnon kohteena olevien muuttujien arvojen muutoksesta yli ajan mutta tutkimuksessa voidaan kuitenkin kerätä tietoa menneisyyteen liittyen.
  - Eri ikäryhmiä vertailtaessa ongelmana on myös niin sanottu kohorttivaikutus: tietynä aikana syntyneiden, eli tietyn kohortin, elinolosuhheet saattavat olla täysin erilaiset kuin jonakin toisena aikana syntyneiden, minkä vuoksi ikäryhmien väliset erot saattavat johtua esimerkiksi yhteiskunnallisia olosuhteista.
  - Poikittaistutkimukseen osallistutaan vain yhden kerran, jolloin tietoa saadaan kerralla paljon.
    - \* Tämä on kuitenkin usein työlästä ja suuren poikkileikkausainesiton kerääminen voi olla kallista.
    - \* Poikittaistutkimuksessa hyödynnetäänkin usein rutiinitoimenpiteinä kerättyjä aineistoja (esimerkiksi tietyn ikävuoden terveys-tarkastuksista)

- \* Nämä voidaan selvittää korrelaatioita ilmiöiden välillä (esimerkiksi alkoholin käyttö ja maksakirroosi) ja siten luoda hypoteeseja tarkemmille jatkokutkimuksille
- \* Tällöin on kuitenkin taas vaara sekoittavista tekijöistä, jos aineisto 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 were 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 poikkileikkausdimension lisäksi myös aikasarjadimensioita.
  - Yleinen tutkimuskysymys pitkittäistutkimuksessa on jonkin **käsitelyn vaikutuksen arvointi**. Tällaisia ovat esimerkiksi lääkeaineetutkimus, poliittisten päätösten arvointi tai markkinointitutkimus.
- \* Pitkittäistutkimuksessa voidaan siis tarkastella **muutosta** mutta on tärkeää muistaa, että pitkittäistutkimuksen eri mittauskerrat eivät ole toisistaan **riippumattomia** ja tämä tulee ottaa tilastollisessa mallissa huomioon!

- Pitkittäästutkimuksen 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 vääristäävät tilastoysiköiden ominaisuuksista erilliset ympäristön haittamuuttujat ovat kaikille samat.
  - \* Pitkittäästutkimuksen pitkän keston vuoksi tutkittavien määrään 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 pitkittäästutkimuksesta

**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.
  - Prevalensi kuvailee sairauden tai haitan omaavien henkilöiden määärää tarkasteltavasta väestöstä tietynä ajankohtana.
  - Usein mitataan vallitsevuustiheyttä eli sairaiden lukumääärää tietynä ajanhetkenä / väkiluku samana ajankohtana.
- Pitkittäistutkimussa mitataan sairauksien ilmaantuvuutta eli insidenssiä.
  - Tutkimuksessa seurataan väestössä ilmaantuvien uusien sairaustabuosten lukumääärää tietyn ajanjakson aikana.
  - Useimmiten mitataan ilmaantuvuustiheyttä, joka ilmoittaa uusien sairastabuosten määrästä henkilöaikaa kohden.
  - Henkilöaika muodostuu tarkasteltavan henkilöryhmän yhteenlasketusta seuranta-aikasta ennen sairastumista, esimerkiksi 100 henkilövuotta muodostuu seurattaessa 100 henkilöä vuoden ajan tai 10 henkilöä 10 vuoden ajan.

**10.2.3 Kohorttitutkimus**

- Kohorttitutkimus on altistelähtoinen 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 alkuaankin riittävän suuri.
    - \* Valitun kohortin tilastoyksiköt 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 mielenkiinnon 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 kuntakokeiluun 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 työllisyden paraneminen)
- Kohorttitutkimus voi vastata kysymykseen: “Mitkä ilmiöt johtuvat tästä altisteesta?”

#### **Esimerkki kohorttitutkimuksesta**

Toisen maailmansodan aikana räjäytettiin Japanissa kakso atomipommia. Tämän traagisen tapahtuman jälkeen tutkijat alkoivat selvittää, mitä terveysvaikutuksia ionisoiva säteily aiheuttaisi altistuneille. Tutkimuksessa seurattiin altistuneiden ja altistumattomien sairastumista vuodesta 1945 vuoteen 1970. Tutkimuksen mukaan ionisoiva säteily aiheutti etupäässä monenlaisia kasvaimia; mm. keuhkosyöpää, rintasyöpää ja kilpirauhasen syöpää.

#### **10.2.4 Tapaus-verrokkitutkimus**

- On **retrospektiivinen havainnoiva** pitkittäistutkimusmenetelmä, jossa tutkimukseen valitaan esimerkiksi tutkittavaan sairauteen (tai muulle altisteeelle/käsittelylle altistuneita) potilaita (**tapaaukset**) ja lisäksi henkilötä, jotka eivät ole sairastuneita tähän sairauteen (**verrokkit**) (tai altistuneet altisteeelle/käsittelylle).
  - Tavoitteena on tutkia miten tutkimusyksiköt reagoivat altistuttuaan jollekin altisteeelle tai käsittelylle. Soveltuu erityisesti harvinaisten ilmiöiden aiheuttajien selvittämiseen.
    - \* Esimerkkinä altistuminen Covid-19 virukselle: retrospektiivisesti (jälkikäteen) voidaan tarkastella viruksen kantajan kanssa samassa tilassa olleita (virukselle altistuneita (virus on altiste)) kysiseissä 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).



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

- \* Toisena esimerkkinä voitaisiin jälleen pitää em. työllisyyden 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 työllisyyden kehitys altistumisen jälkeen.
- Käsittelyn tai altistuksen seuraauksia, esimerkiksi sairauden, syitä etiitään 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.
  - \* Verrokkien valinta kuitenkin kriittinen, sillä valitsemalla verrokkit/kontrollitapaukset väärin mikään tilastollinen testi tai menetelmä ei korjaa tai kvantifioi tästä virhettä!
  - \* Esimerkki verrokkiryhmän epäkelvosta valinnasta on huonosti mitattu aiempi altistuminen ja/tai jos jokin tutkimuksen kannalta keskeinen taustamuutuja 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-verrokitutkimuksesta**

Länsi-Saksassa tuotiin 50-luvun lopulla markkinoille talidomidi-niminen uni- ja rauhoittava lääke. Varsin pian markkinoille tulon jälkeen tietyntyyppisten synnynnäisten epämuodostumien määrä alkoi lisääntyä rajusti. Talidomidin ja lasten raajojen muodostumishäiriöiden yhteys paljastettiin tapaus-verrokitutkimussilla. Tutkimuksissa selvitettiin sekä sairaiden lasten (tapaukset) että terveiden lasten (verrokkien) äitien altistuminen talidomidille raskauden kriittisten viikkojen aikana. Melkein kaikki sairaiden lasten äidit olivat saaneet talidomidia ensimmäisten raskausviikkojen aikana (talidomidin oli myös havaittu helpottavan odottavien

äitienv raskauspahoinvoingtia). Talidomidi poistettiin markkinoilta ja epämudostumatapausten määrä putosi jyrkästi.



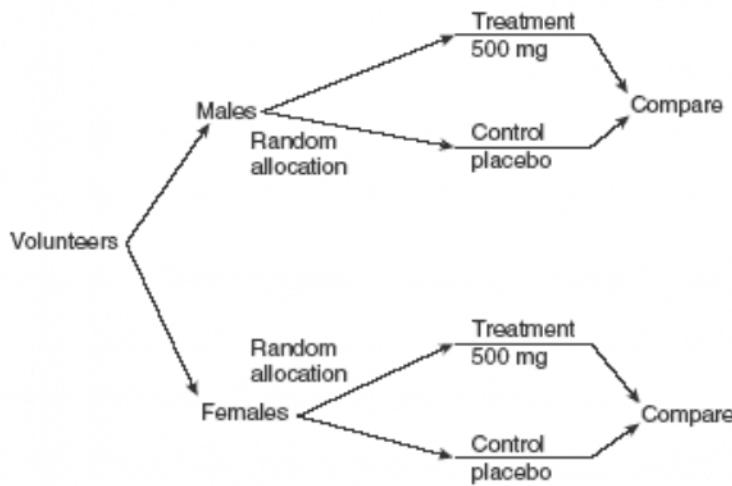
Kuva 10.6: Esimerkki tapaus-verrokkitutkimuksesta

- **Lohkot/osittaminen**

- Lohkomisella / osittamisella tarkoitetaan tutkimusyksiköiden järjestämistä ryhmiin niin, että ryhmät ovat mahdollisten sekoittavien tekijöiden suhteen mahdollisimman samankaltaiset
- Lohkotekijä on yleensä muuttuja, jonka aiheuttama vaihtelu vasteenseen ei ole tutkijan päämielenkiinnon kohtena
- Lohkotekijän kontrollointi johtaa usein tarkkuudeltaan parempiin tuloksiin
- “Block what you can, randomize what you cannot.”

- **Sokkouttaminen**

- Tutkimushoitojen sokkoutuksen tavoitteena on vähentää sekä potilaan että tutkimushenkilökunnan mahdollisten ennakkokäsitysten vaikutusta tutkimuksen tuloksiin.



Kuva 10.7: Esimerkki ositetusta koeasetelmosta

- **Yksöissokkotutkimuksessa** (single-blind) tutkittava ei tiedä, mihin hoitoryhmään hän kuuluu.
- **Kaksoissokkotutkimuksessa** (double-blind) ei tutkittava eikä tutkija ja muukaan tutkimushenkilökunta saa tutkimuksen aikana tietää mihin hoitoryhmään tutkittava kuuluu.
- **Kolmoissokkotutkimuksessa** tutkimuksessa edes havaintojen analysoija (esim. tilastotutkija) ei tiedä, miten kokeellisen tutkimuksen käsitteily on koodattu.
- **Avoimessa tutkimuksessa** (open) sekä tutkittava että tutkimushenkilökunta tietävät, mihin hoitoryhmään tutkittava kuuluu. Myös avoimessa tutkimuksessa tulisi yleensä käyttää satunnaistamista.

## 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 eduiksi voidaan lukea mm. seuraavat seikat
  - Aineiston muodostaminen/kerääminen on verrattain helppoa ja Suomessa on paljon korkealaatuisia rekisteriaineistoja. Tätä edesauttaa tietotekniikan nopea kehittyminen, joka on mahdollistanut erittäin suurten aineistojen rutuininomaisen keräämisen.
  - Ei tarvetta erikseen tuottaa tutkimusaineistoa: vältetään mahdollisuksi kallis aineiston keräysvaihe.
  - Suomalainen henkilötunnusjärjestelmä mahdollistaa tietojen tehokkaan käytön ja laadukkaan tutkimuksen.
- Rekisteriaineistojen ongelmina ja haittoina voidaan pitää mm. seuraavia
  - Mikäli tutkimuksessa lähtökohtaisesti käytetään rekisteriaineistoja, määäräävät ne välillisesti myös mahdolliset tutkimuskohteet: rekisteriaineistot kerätään eri tarkoitusta varten eivätkä ne täten välttämättä sisällä kaikkea haluttua informaatiota.
    - \* Tutkimuksen ongelmalähtöisyyss saattaa unohtua helpommin, kun tutkimusongelman aihiota asetellaan sopimaan rekisteriaineistojen tarjoamiin mahdollisuuksiin.
    - \* Rekisteriaineistoilla on myös omat rajansa: tutkimuskysymysten kannalta väärin mitattua muuttujaa ei useinkaan voida millään tavalla muuntaa täydellisesti haluttuun muotoon.
  - Rekisteriaineisto pitää usein esikäsitellä sopivan muotoon laadullista tutkimusta muistuttavalla tavalla.
  - Rekisteriaineistojen analyysistä ja niihin soveltuviista tilastollisista menetelmistä on vähän lähesti metodologisia oppikirjoja ja/tai esimerkkitutkimuksia.
  - ”Ulkopuolisille” tutkijoille aineistojen käyttö saattaa olla hankalaa mm. korkeiden pääsykustannusten (rekisterien ylläpitäjien, viranomaisten ja tutkimuslaitosten ulkopuolella), tietosuojakysymysten tai teknisten hankaluuksien takia.
    - \* Rekisteriaineiston käyttö vaatii tutkimussuunnitelman ja tutkimussuunnitelman perusteella myönnetyn käyttöluvan rekisterin ylläpitäjältä.
  - Tietotekniikan kehitymisen 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 substanssitetoutta.
    - \* Eri rekisterejä ei aina saadakaan linkattua tehokkaasti yhteen esimerkiksi jos ne ovat mitanneet mielenkiinnon kohteina olevia muuttuja 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 otantavirheeseen perustuvan tilastollisen päättelyn 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 kerättävänkin aineistoja.
  - Tutkimusongelman pitäisi aina olla keskeinen lähtökohta myös rekisteriaineiston käytössä.
  - Itse kerättessä aineisto on mahdollista rääältöiden tuottaa vastaanmaan 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 rekisterinpitäjän ja rekisteröitän 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 siinä, 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.

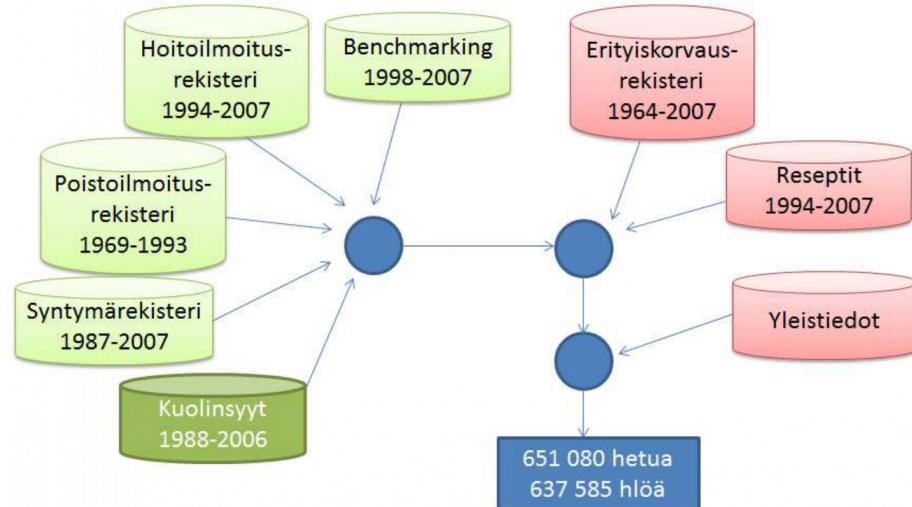
- Rekisteriaineistot tietosuojalain takaaman yksityisyysdien suojan näkökulmasta
  - Suomessa rekisteriaineistot pyritään luovuttamaan käyttöön vain tunnistettomina yksityisyysdien suojan säilyttämiseksi.
  - Tieteellinen tutkimustarkoitus luokitellaan kuitenkin poikkeustapauksaksi tietosuojalaissa ja 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 tietosuojaa koskevat säännökset ovat niin selkeitä ja kattavia, että yleisössä ei synny epäilyksiä tietojen väärinkäytön mahdollisuksista.
  - Rekisteritietojen käyttöön on aina haettava lupaa. Erityisesti eri rekistereitä yhdistettäessä on hankittava myös tietosuojaavaltuutetun lausunto suunnitteilla olevan tutkimuksen laillisuudesta ja käyttöehdoista.
- 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.
  - 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.
  - Minkälaisia tietorakenteita aineistossa käytetään ja miten se vaikuttaa eri muuttujien tallentamiseen tietojärjestelmään?
    - \* Tutkimuskysymyksen kannalta on voi olla merkitystä esimerkiksi sillä, onko rekisterinpitää kerännyt henkilöiden ikätietoa vuoden vai kymmenen vuoden tarkkuudella.

**Esimerkki: diabeteksen ja sen lisäsairauksien esiintyvyyden ja ilmaantuvuuden rekisteriperusteinen mittaaminen**

- Vaihe 1: Diabeteskohtortin identifointi (Tilastokeskus: kuolinsyyt,

THL: diagnoosit, Kela: erityiskorvaukset, reseptit).

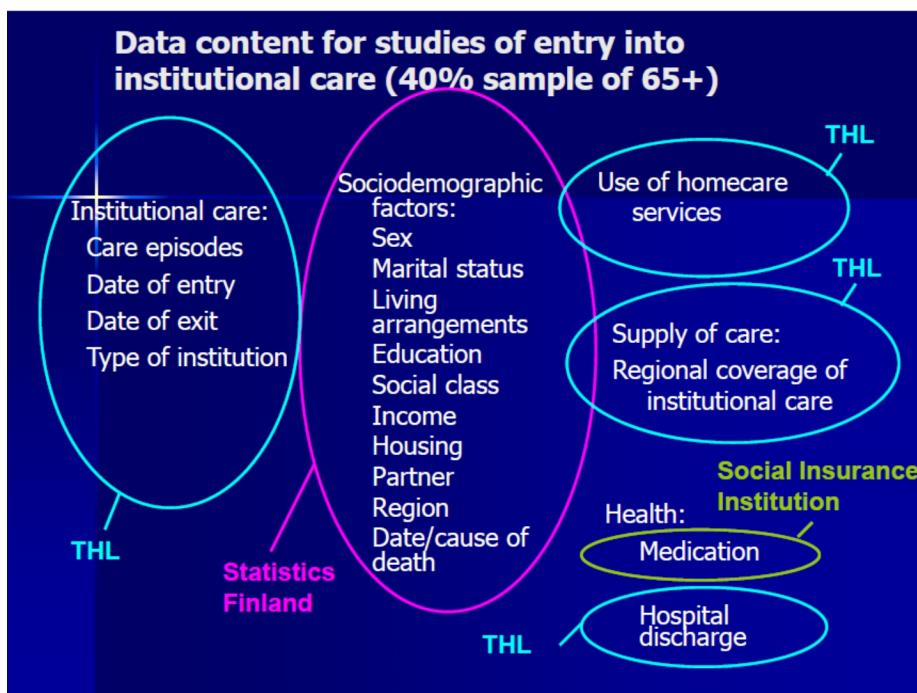
- Vaihe 2: Seurantatiedot (syöpärekisteri, sairaspäivärahat, eläkkerekisteri...).
- Ongelma: kuinka monta henkilöä diabeteskohortista on kuollut seuranta-aikana?
  - Kuolema on vakaa käsite, johon ei liity mittausvirhettä tai subjektiivisuutta.
  - Katsotaan aineistosta "yhden rivin säännöllä" kuinka monelle löytyy tieto kuolemasta.
- Kysymys: Kuinka moni diabeteskohorttiin kuuluvista sairastaa tyypin 1 diabetesta?
  - Tyypin 1 diabetes johtuu insuliinia tuottavien beetasolujen tuhoutumisesta autoimmuuniprosessin seurauksena.
  - Tyypin 1 diabeetikko tarvitsee jatkuvasti insuliinia, mutta ei hyödy haiman omaa insuliinineritystä tehostavista lääkeistä.
  - Rakennetaan algoritmi, jolla identifioidaan tyypin 1 diabeetikot lääkeostojen luokkien ja säännöllisyyden perusteella.



Kuva 10.8: Vaihe 1: Diabeteskohortin identifointi

**Esimerkki: rekisteritutkimus pitkääikaisen laitoshoivan käytöstä**

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



Kuva 10.9: Vaihe 1: Diabeteskohortin identifiointi

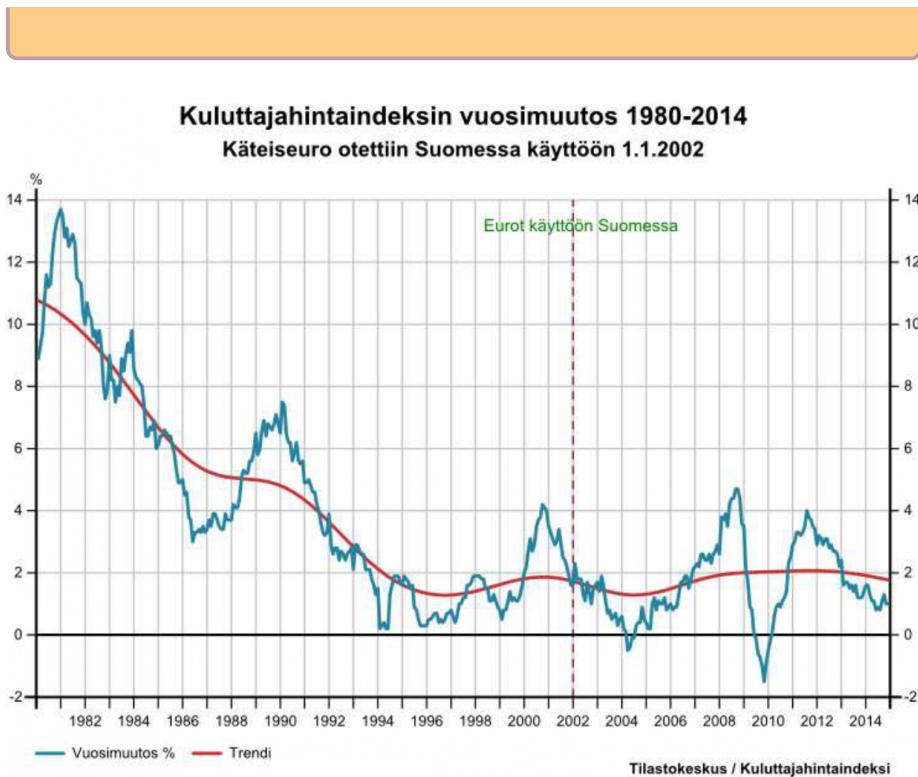
### 10.3.2 Aikasarjat ja paneeliaineistot

- **Aikasarjat**

- Aikasarjaksi kutsutaan havaintojen jonoa, jossa aika määräät jostain tilastollisesta muuttujasta tehtyjen havaintojen järjestyksien.
- Havainnot ovat tavallisesti peräkkäisiä, ja mittaukset on tehty tasaisin aikavälein, mutta väliaikojen tasaisuus ei kuitenkaan ole välttämätöntä ja monissa tutkimusasetelmissa kohde aikasarjasta voidaan poimia havaintoja jatkuvasti tai mielivaltaisen pienin aikavälein.
- Yksittäinen aikasarja on pitkittäisaineiston erikoistapaus, jossa tarjollaan vain yhtä aikasarjaa. Pitkittäisaineistoon nähdyn 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. Tällaisia tilastollisia suureita kutsutaan **indekseiksi** ja ne sisältävät tiivistettyä tietoa yhteiskunnasta, kuten esimerkiksi inflaation mittarina käytetty kuluttajahintaindeksi.

#### Esimerkki: kuluttajahintaindeksit

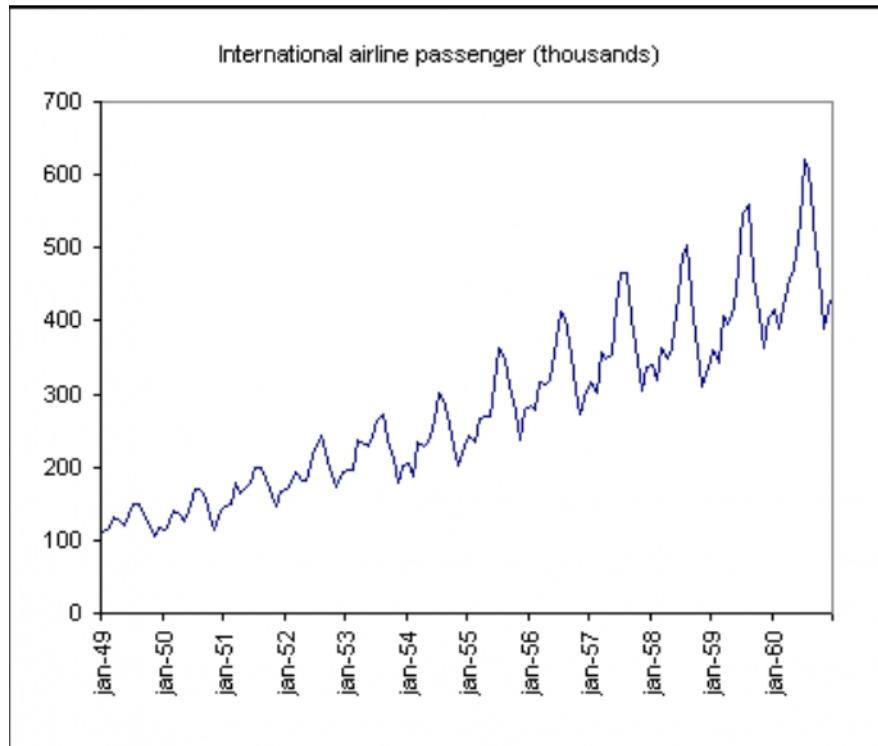
- Hintaindeksi on normalisoitu keskiarvo tarkasteltavan tuote- tai palvelukorin hinnoista, joka lasketaan säännöllisin välajoin tavoitteena helpottaa hintojen muutosten seurantaa eri ajankohtien tai alueiden välillä. Laajimmat indeksit mittavaat hintojen kehitystä koko talouden tasolla ja niitä hyödynnetään monin tavoin talouspolitiikassa.
  - Tilastokeskuksen haastattelijat keräävät indeksiä varten kaiken kaikkiaan noin 50 000 hintatietoa lähes 500 hyödykkeestä noin 2 700 (näin joitain vuosia sitten) liikkeestä aina kuukauden puolivälissä, lisäksi noin 1 000 hintatietoa kerätään keskitetysti.
  - Tilastokeskuksen laskema kuluttajahintojen vuosimuutos oli marraskuussa 2014 oli n. 1,0 prosentti
  - Marraskuussa kuluttajahintoja nostivat viime vuodesta eniten vuokrankorotukset, tupakkatuotteiden ja alkoholiuomien vähittäishintojen ja ravintola- ja kahvilopalvelujen sekä kerrostalojen kunnossapitolvelujen kallistuminen
  - Kuluttajahintojen nousua vuoden takaisesta hillitsi marraskuussa eniten elintarvikkeiden, poltonesteiden ja viihdeelektronikaan halpeneminen



Kuva 10.10: Esimerkki: kuluttajahintaindeksi

- **Aikasarjojen tilastollinen analyysi** perustuu siihen, että sarja tulkitaan jonkin **stokastisen prosessin** eli satunnaisprosessin realisaatioiksi
  - Jos aikasarjan generoinut prosessi saadaan selville, voidaan tietoja prosessista käyttää aikasarjan käyttäytymisen kuvaamiseen ja selittämiseen sekä aikasarjan tulevan käyttäytymisen ennustamiseen.
  - Aikasarja-aineistot ilmentävät ilmiöstä riippuen ns. **autokorrelaatiota**, eli ajallisesti toisiaan lähellä olevat havainnot ovat korreloituneempia kuin ajallisesti kaukana toisistaan olevat.
    - \* Aikasarja-analyysi on yksi tilastotieteen osa-alue, jolla on rikas ja pitkälle kehittynyt teoriapohja.
    - \* Aikasarja-analyssia 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?**
- Ks. esimerkiksi kuva [10.11](#).



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

#### • Paneeli- eli pitkittäisaineisto

- Paneeliaineistolla tarkoitetaan aineistoa, jossa tilastoyksiköistä on useita havaintoja ja aika määrä havaintojen järjestyksen (kuten aikasarjoissa) ja lisäksi jokaisena ajanhetkenä mitataan useampi kuin yksi tilastollinen muuttuja (kuten poikkileikkausaineistossa).
  - \* Paneeliaineisto on terminä käytettympä yhteiskuntatieteissä kun taas pitkittäisaineisto esimerkiksi lääketieteessä.
  - \* Havaintoyksiköt voivat olla esimerkiksi yrityksiä, ihmisiä, kuntia tai kouluja. Ns. “täydellisessä” paneeliaineistossa kaikista havaintoyksiköistä on havaittu kaikki muuttujat kaikkina ajanhetkinä.

”Kiertävä” paneeli on vastaavasti sellainen, jossa osa havaintoyksiköistä vaihtuu ajan kuluessa.

- Tyypillisesti havaintoja kerätään tasaisin väliajoin, kuten kuukausittain tai vuosittain, ja yksittäisen ajanhetken havainto on poikkileikkausaineisto ja kustakin havaintoyksiköstä on oma usean muuttujan aikasarjansa.
- Paneelaineisto mahdollistaa vastaamisen kysymykseen miksi? Yleisesti ottaen paneelaineistoja käytetäänkin erityisesti ns. kausaalipäättelyyn tähänvissä 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.
  - Havaintoyksiköt 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 rakentaminen ja testaaminen
  - Kyselymetodin määrittely (puhelinhästattelu, elektroninen kysely...)
  - Mahdollisten haastattelijoiden koulutus
  - Aineiston keräys
  - Aineiston yhteneväisyyden tarkistaminen (muuttujat tallennettu oikein jne.)
  - Tulosten adjustointi mahdollisten identifioitujen virhelähteiden muukaan
- Survey-tutkimusta käytetään erityisesti asenteiden, mielipiteiden ja käytäytymisen tutkimiseen.
  - Esimerkkejä ovat mm. poliittiset mielipidekyselyt, markkinointitutkimukset, alkoholinkulutustottumuksia ja terveyspalveluiden tytyväisyyksikyselyt, joihin voi kaikkiin uskova liittyvän vastausharhaa eri syistä.

---

<sup>2</sup>Kausaalimalleja opetetaan kurssilla [TILM3529 Kausaalipäättely havainnoivissa tutkimuksissa](#)

- Esimerkiksi vastausharhaa arkaluontoisiin kysymyksiin voidaan vähtää avoimilla kysymyksillä tai alustamalla kysymystä johdannolla, jossa suvaitaan / ymmärretään kaikenlaiset vastaukset.
- “Randomized response”: vastaukseen lisätään jonkin todennäköisyysmallin mukaisesti harhaa todellisen vastauksen salaamiseksi.
- Kerätty aineisto on siten altisteinen tehdylle kyselylle ja täten sen käytökkelpoisuus perustuu hyvin pitkälti etukäteissuunnitteluihin, kunnolliseen toteutukseen ja kyselylomakkeen oikeaoppiseen rakentamiseen.
  - Lisäksi aineiston käytökkelpoisuus riippuu myös vastaajaotoksen pojminnasta (edustavuudesta) ja siitä, kuinka totuudenmukaista informaatiota vastaajat ovat kyselyssä antaneet. Tässäkin hyvä etukäteissuunnitelu on keskeistä.
  - Tilastollisten menetelmien avulla pyritään arvioivaan otoksen, kyselyn suunnittelun ja kerätyn vastaajaotoksen sisältämää (tai aiheuttaa) harhaa.

#### **10.4 Luvun 10 yhteenveto, keskeisiä termejä ja kokonaisuuksia.**

- Kuvaileva tutkimus
- Vertaileva tutkimus
- Kokeellinen tutkimus
- Havainnoiva tutkimus
- Poikittais/poikkileikkaustutkimus
- Pitkittäistutkimus (paneeli- ja pitkittäisaineistot)
- Kohorttitutkimus
- Tapaus-verrokkitutkimus ml. lohkot/osittaminen ja sokkouttaminen
- Rekisteriaineisto
- Aikasarja-aineisto
- Survey eli kyselytutkimus

## Luku 11

# Tilastollisesta ennustamisesta

Kuten olemme jo tähän menneessä näheet, tilastollinen analyysi ja sen erottamattomana osana tilastollinen päättely on keskeinen vaihe tieteellistä tutkimusta. Vielä ennen tilastollisen selittämisen ja ennustamisen väliä eroja koskevia pohdintoja muistutetaan jo aiemmin käsitellystä **kuvailevasta tilastotieteestä**. Tämä voidaan nähdä vielä (ainakin) kolmantena yleisenä tilastotieteen tavoitteena mallintamisen/selittämisen ja ennustamisen lisäksi. Yksinkertaisin tilastollisen päättelyn muoto on hyödyntää aineistoaa kuvailevia tunnuslukuja, kuten keskiarvoja ja keskihajontalukuja. Niistä voidaan kuitenkin tehdä vain melko rajoittuneita pääteliä. Varsinkin havainnoivassa tutkimuksessa sen selittämiseksi, miten selittävät muuttujat ovat yhteydessä selittävään vaste-muuttujaan, käytetään esim. lineaarista tai logistista regressiota (ja niiden monenmoria 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 tarkemmat 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 infektiointiin.

- Tällöin **pyrkimyksenä** on rakentaa “**selitysmalli**”, jossa on perustellut syy-seuraussuhheet selittävästä (selittävistä) muuttujista selittävään muuttujaan.
- (Tilastollinen) **ennustaminen** tarkoittaa, että tietyillä selittävän tai seittävien (tai “ennustavien”) muuttujien yhdistelmillä voidaan ennustaa ennustettavan muuttujan arvoa.
  - Ts. siis ennustettavana muuttujana toimii tilastollisen mallin näkökulmasta katsoen vastemuuttujan arvo, jota pyritään ennustemallin avulla ennustamaan.
  - Ennustemalleja tutkittaessa varsinaisilla selityssuhteilla ei välittämätä ole merkitystä. Tärkeintä on mallin ennustekyky, ei niinkään esim. yksittäisen regressiokertoimen arvo ja siihen liittyvät tarkemmat tulkinnot. Tilastollisesti merkitsevä regressiokerroin ei tarkoita, että muuttujalla olisi välittämättä 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 päätölyn ja estimointiteorian lähtökohtiin mitä olemme jo sivunneet täällä 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**) kasvatetulta hurjasti suosioitaan viime vuosina ennustaminen on levinyt, ja leviää jatkossakin, voimakkaasti myös hyvin keskeiseksi osaksi muutakin tilastollista analyysiä.

## 11.2 Tilastolliseen ennustamiseen liittyviä huo-mioita

- Ennustamista on kaikkialla! Sen rooli on paljon keskeisempi osa meidän kaikkien arkea mitä ensiajatukselta saattaa tulla mieleen.

## 11.2. TILASTOLLISEEN ENNUSTAMISEEN LIITTYVIÄ HUOMIOITA 197

- Ennustaminen on elämässämme korvaamatonta.<sup>1</sup>
  - Kun valitsemme reitin työmatkalle, päättämmekö menemmekö toisille treffeille tai säätämme 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. vastemuuttajan luonteen perusteella valitaan sopiva ennustevirheitä summarisoiva tunnusluku, kuten keskineliöennustevirhe (jatkuvat vastemuuttujat) tai luokitteluvasteiden tapauksessa väärin (tai oikein) ennustettujen luokitteluiden suhteellinen osuuus.
  - 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:** Silverin 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 laskivat 40 prosenttia (osoittaen osaltaan vallinnutta laajempaan **“asuntokuplaa”** (perustettoman korkeita asuntojen hintoja), työttömyys kasvoi räjähdyväisesti jne.

**Ennustevirheen yhteisiä ja tyypillisiä piirteitä** (tässä tapauksessa), jotka laajentuvat moniin muihinkin tilanteisiin ja sovelluksiin:

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

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 taimumus laskea (näissä olosuhteissa).
2. Luottoluokituslaitokset (samoin kuin Lehman Brothersin kaltaiset pankit) eivät ymmärtäneet, miten riskialtiita asuntovakuudelliset arvopaperit olivat. Ongelma ei varsinaisesti ollut siinä, että luottoluokituslaitokset eivät nähtäneet asuntokuplaa. Sen sijaan niiden ennustemallit olivat täynnä huonoja oletuksia ja väärää “itseluottamusta” mahdollisten asuntojen hintojen romahduksen riskeistä.
3. Laajasti ei ennakoitu, miten asuntokriisi laukaisee globaalim rahoituskriisiin. Se johtui suuresta osin liiallisesta velkaantumisesta markkinoilla, jossa lyötiin erinäisten instrumenttien myötä vetaa yhdysvaltalaisen halukkuuden puolesta sijoittaa uuteen kotiin.
4. Rahoituskriisiin välittömässä jälkimainingeissa ei osattu ennustaa, miten laajoja taloudellisia ongelmia se aiheuttaa. Rahoituskriisit tyyppillisesti 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 siiä, 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 asuntovakuudellisia 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 siiä, että aikaisemmin asuntojen hintojen heilahtelulla ei yleensä ollut suuria vaikuttuksia rahoitusjärjestelmässä. **Kuitenkaan** rahoitusjärjestelmä ei luultavasti koskaan aikaisemmin ole ollut yhtä vekkaantunut eikä vedonlyöntiä asuntojen hinnoista ollut tehty vastaanvassa 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. koro-

napandemian 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 näyttäätyy 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 ulottuvaa Yhdysvaltain asuntomarkkinaaineistoa.
- Ongelmana oli, että 1980-luvulta 2000-luvun alkuvuosiin saakka asuntojen hinnat olivat aina vakaat tai nousevat Yhdysvalloissa. Tässä tilanteessa oletus, että asumnonomistajien asuntolainat eivät juurikaan 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 ulkopuolisiä 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 rakenntamiseen. Nämä ollen näiden ennustekykyä arvioitaessa onkin mentävä otoksen ulkopuolelle, koska “otoksen sisällä” voimme opettaa kyseisiä malleja (ääritilanteessa) niin, että ne ovat periaatteessa ääretömän 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) lähes yksinomaan ennustesovelluksiin. Tällöin mallien ja algoritmien 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!

### 11.3 Luvun 11 yhteenvetö, keskeisiä termejä ja kokonaisuuksia.

- Tilastollinen ennustaminen
- Selitysmalli vs. ennustemalli
- Ennuste ja ennustevirhe
- Opetusaineisto vs. ennusteotos
- In-sample/training sample vs. out-of-sample (forecasting)

## Luku 12

# Tilastotieteen kehityksen nykytrendejä

### 12.1 Modernin tilastotieteen suuntaviivoja

Seuraavassa vielä joitain tilastotiedettä parhaillaan koskevia kehitystrendejä, joita olemme myös osaltaan sivunneet tämän kurssimateriaalin myötä:

- **Aineistot kasvavat ja monimutkaistuvat.** Näin ollen tilastotiedettä 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. signaaleja kun vastaanvastii kasvava määrä kohinaa yrittää vaikeuttaa tätä signaalin erottamista kohinasta).
- **Laskennallisuus** kasvaa. Tietokoneiden laskentakapasiteetti nousee vaikka kokin suhteellinen kasvu ei ehkä olekaan enää niin suurta mitä muutamat viime vuosikymmenet.
- Osin laskennallisuuteen liittyvä **koneoppimisen** kehittyminen ja sen “rajoitiateiden” yhteyteen integroituneet käytännöt ja menetelmäkehitystyö tulee jatkumaan.
  - Toisaalta “perinteiselle” tilastotieteelle on edelleen myöskin vahva oma roolinsa monilla eri tieteenaloilla.

- **Analyysien automatisointi:** Tilastolliset ohjelmat alkanevat jatkossa tulkita 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.
  - Näiltä osin yksi keskeinen kehityssuunta on jo vahvasti yleistyneet laajat teköälymallit kuten OpenAI:n toteuttama [ChatGPT](#).
    - \* ChatGPT:tä käytetäänkin jo laajalti mm. tilastollisen ohjelmoinnin apuvälineenä. Nopeasta kehityksestä huolimatta ChatGPT:n antamat vastaukset ovat kuitenkin paikoin epäluotettavia tai suorastaan väärin, joten sen käytössä korostaa edelleen vahvan tilastotieteellisen osaamisen lisäksi substanssiosaaminen.
- **Poikkitieteellisyys** tulee entisestäänkin vahvistumaan. Ts. substanttitietouden ja tilastotieteen yhdistäminen ja sen tärkeys ei tule ainakaan vähemään. Sivuaineopintoja kannattaa siis käydä ja ottaa opinto-ohjelmaan.
  - Tämän lisäksi kokonaisvaltaisesti tilastotieteen ytimen osaajien osaamista tulla kysymään jatkossakin.
- Oheisten huomioiden ohella lopuksi on syttää 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 kyyjä ottaa vastaan monia jo varsin monimutkaisia tilastollisia menetelmiä!

## 12.2 Luvun 12 yhteenveto, keskeisiä termejä ja kokonaisuksia.

- Aineistot kasvavat ja monimutkaistuvat
- Informaatio ei ole enää niukkaa
- Laskennallisuus kasvaa
- Koneoppimisen vahvistuminen ja kehittyminen
- Analyysien automatisoituminen
- Poikkitieteellisyys