Tilastotiede ja data

Henri Nyberg & Roope Rihtamo

Invalid Date

Table of contents

Esipuhe/kurssi-informaatiota

Kurssin (osat I ja II) rakenne

Tilastotiede ja data -kurssi koostuu kahdesta erillisestä osasta, jotka tarjoavat opiskelijalle työkaluja tieteen ja tutkimuksen ymmärtämiseen tilastotieteen näkökulmasta käsin. Lisäksi kurssi pyrkii antamaan, varsinkin alkuosaltaan, melko ei-matemaattisen johdannon tilastotieteen keskeisiin ideoihin ja perusteisiin.

- **Osa I** (2op): Ensimmäinen osa toimii johdantokurssina tieteen ja tutkimuksen tekoon esittelemällä keskeisiä käsitteitä, käytäntöjä ja termejä, joita kvantitatiivisessa/määrällisessä tutkimuksessa tilastotieteen ja sen menetelmien osalta tarvitaan (ks. osaamistavoitteet alla).
  - Tarkoituksena on yleisellä tasolla johdatella tilastotieteen ja aineistojen (datan) maailmaan sekä pohtia myös näiden laajempaa merkitystä tieteellisen tutkimuksen hyvin keskeisinä osina.
  - Kurssin ensimmäisen osan lopuksi järjestetään tentti, joka arvioidaan hyväksytty/hylätty asteikolla.
- **Osa II** (4op): Kurssin toisessa osassa syvennytään ensimmäisen osan teemoihin tarkastelemalla erityisesti tilastollisten menetelmien roolia tutkimuksessa.
  - Tarkoituksena on luoda pohja tosielämän satunnaisilmiöiden tutkimukselle tilastotieteen ja tilastollisten menetelmien keinoin.
  - Harjoitustehtävien tavoite on totuttaa opiskelija erityisesti matematiikan ja tilastotieteen opinnoissa yleiseen viikottaisten harjoitustehtävien tekemisen käytäntöön.

  - Kurssin toisen osan lopuksi järjestetään tentti, joka arvioidaan 0-5 arviointiasteikolla.

Kurssilla vältetään kovin teknistä matemaattista esitystapaa, mutta erityisesti Osassa II, tarvittavissa määrin, tullaan kuitenkin käyttämään tilastotieteen perusopinnoissa käytettäviä 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** (https://opas.peppi.utu.fi/fi/opintojakso/TILM3553/1734?period=2022-2024) tai **TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille** (https://opas.peppi.utu.fi/fi/opintojakso/TILM3568/3385?period=2%20022-2024) sekä **TILM3555 Tilastollisen päättelyn peruskurssi** (https://opas.peppi.utu.fi/fi/opintojakso/TILM3555/1731?period=2022-2024) asiaksi. Nämä kurssit, yhdessä alkuvaiheen pakollisten matematiikan kurssien lisäksi, muodostavat siis tämän kurssin kanssa lähtökohdan tilastotieteen opinnoille (Turun yliopistossa).

Kurssimateriaali

Luennot eivät suoraan perustu yhteen kirjaan tai lähteeseen. Käytettyjä lähdemateriaaleja luetellaan alapuolella oheislukemiston myötä.

Oheislukemistoa (sopivilta osin):

- Alho, J., Arjas, E., Läärä, E., ja P. Pere (2023). Tilastotieteen sanasto. Suomen Tilastoseuran julkaisuja, no. 8 2. laitos. Suomen Tilastoseura. Helsinki.

  - Ks. myös alapuolella linkki sanaston verkkoversioon.

- Holopainen, M. ja P. Pulkkinen (2008). Tilastolliset menetelmät. Sanoma Pro Oy.

- 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).

- Pesonen, M. (2017). Kurssimateriaali kurssille Aineistonhankinta ja tutkimusasetelmat, Turun 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
- Spiegelhalter, D. (2020). The Art of Statistics. Learning from Data. Penguin Books Ltd.
- Sund, R. (2003). Tilastotiede käytännön tutkimuksessa -kurssi. Helsingin yliopisto.

Näiden lisäksi materiaalissa viitataan seuraaviin artikkeleihin ja kirjoihin täydentävänä materiaalina:

- Box, G. E. P. (1976). Science and Statistics. Journal of the American Statistical Association, 71(356), 791–799.
- Haaparanta, L., ja I. Niiniluoto (2016). Johdatus Tieteelliseen Ajatteluun. Gaudeamus.
- Hamermesh, D.S. (2007). Replication in economics. *Canadian Journal of Economics/Revue canadienne d' economique* 40 (3), 715–733.
- Friendly, M., ja D. Meyer (2015). *Discrete Data Analysis with R. Visualization and Modeling Techniques for Categorical and Count Data.* Chapman & Hall/CRC.
- Marks, H. (2020). Mastering the Market Cycle. Getting the Odds on Your Side. Nicholas Brealey Publishing (paperback).
- 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.
- Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable. Random House.
- Taleb, N. N. (2012). Antifragile: Things That Gain from Disorder. Random House.
- Vartia, Y. (1989). Tilastotieteen perusteet. Yliopistopaino, Helsinki. II painos.

Muita taustamateriaaleja

- Tilastokeskuksen tilastokoulu (https://tilastokoulu.stat.fi/verkkokoulu_v2.xql?course_id=tkoulu_tilaj&lesson_id=1&subject_id=0&page_type=sisalto)
- Tilastotieteen sanasto suomi-englanti-suomi hakukone. Juha Alho, Elja Arjas, Juha Karvanen, Lasse Leskelä, Esa Läärä ja Pekka Pere (2025). Tilastotieteen sanasto. Verkkoversio 8.6.2025. Suomen Tilastoseura. https://sanasto.tilastoseura.fi/.

Suuret kiitokset VTM Visa Kuntzelle, VTM Emil Lehdelle ja filosofian ylioppilas Roni Karjanlahdelle kommenteista ja avusta materiaalin työstämisessä. Kaikki jäljelle jääneet painovirheet ovat materiaalin kokoajien.

Osaamistavoitteet

Kurssin materiaali on koostettu em. lähteistä ja pyrkii paikoin pelkistettyyn esitysmuotoon mutta kuitenkin niin että materiaalin opiskelemalla kurssien osaamistavoitteet täyttyvät kokonaisuudessaan.

Osaamistavoitteet on listattu Turun yliopiston opinto-oppaassa matematiikan ja tilastotieteen laitoksen opintotarjonnasta kurssikuvauksien alta **Osa I** (https://opas.peppi.utu.fi/fi/opintojakso/TILM3712/102471?period=2024-2027) ja **Osa II** (https://opas.peppi.utu.fi/fi/opintojakso/TILM3713/102472?period=2024-2027) ja ne löytyvät alta vielä laajemmin.

Osa I: Opintojakson suoritettuaan opiskelija

- on saanut kokonaiskuvan tilastotieteestä ja sen perusteista
- on sisäistänyt tilastotieteen keskeisiä käsitteitä ja osaa niiden avulla tarkastella kriittisesti tieteellisiä tutkimuksia
- pystyy erottamaan edustavan otoksen ja näytteen
- tuntee tilastotieteen keskeiset perustunnusluvut

Osa II: Opintojakson suoritettuaan opiskelija on täydentänyt osassa I opittuja tilastotieteen perustietoja mm. seuraavilta osin

- osaa hahmottaa tilastotieteen roolin omana tieteenalana ja eri sovellusalueiden yhteydessä

- tunnistaa erilaiset tutkimusasetelmat ja aineistotyypit
- ymmärtää otannan ja otantateorian perusteet
- hahmottaa tilastollisen riippuvuuden, korrelaation ja yksinkertaisen lineaarisen regressiomallin idean
- ymmärtää tilastollisen ennustaminen perusajatukset

Kurssin sisältö

Kurssin osien I ja II sisältöjä on listattu opinto-oppaassa ja laajemmin alla. Tämä listaus toimii hyvänä luettelona kurssin keskeisistä teemoista.

Osa I

- Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa
- Tilastotiede tieteenalana ja sen suhde lähitieteisiin, kuten datatieteeseen (data science)
- Sattuma ja satunnaisuus tilastotieteessä
- Tilastolliset aineistot (data), niiden kerääminen ja mittaaminen
- Tilastollinen riippuvuus ja korrelaatio

Osa 2

- Tilastolliset mallit ja jakaumat
- Otantamenetelmien perusteet
- Tilastollisen päättelyn perusteita
- Regressioanalyysin perusteet
- Tilastotieteen rooli uuden tieteellisen tiedon tuottamisessa
- Aineisto- ja tutkimustyypit ja koeasetelmat tilastotieteessä
- Tilastotieteen sovellusten ja sovellusalueiden esittelyä

Merkinnät

Huomaa, että tässä materiaalissa käytetään desimaalipistettä desimaalipilkun asemasta. Tämä on tarkalleen ottaen suomenkielen sääntöjen vastaista (sen yleistymisestä huolimatta), mutta tekstissä, johon liittyy matemaattista kielenkäyttöä, tämä kv. käytäntö käyttää desimaalipistettä on selvempi.

Materiaalin seassa on eritelty väärikoodatuin tietolaatikoin erinäisiä tärkeitä tilastotieteellisiä konsepteja ja termejä sekä esimerkkejä tilastotieteen sovelluksista. Näistä ensin mainitut löytyvät Deltan violeteista laatikoista ja jälkimmäiset Statistikan oransseista.

- Toim. Huom. värit eivät täysin alkuperäisten värien kanssa yhteneväisiä.

Alla esimerkkilaatikot.

Konsepti tai termi

Konseptin tai termin löyhä määritelmä.

Esimerkki

Aihetta koskeva esimerkki.

Chapter 1

title: “Johdantoa tilastotieteeseen”

OSA I

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

(Holopainen ja Pulkkinen, 2008)

Getting the Odds on Your Side

(Marks, 2020)

Huom. 1: Tämän kurssin materiaalissa määritellään useita eri termejä ja (tilastollisia) kokonaisuuksia. Vaikka näiden (riittävän tarkkaan) oikeellisuuteen pyritään, niin monessa kohtaa tehdään yksinkertaistuksia ja vältetään

kovin matemaattista esitystapaa, mikä paikoin olisi välttämätöntä tarkan määritelmän antamisen kannalta.

Huom. 2: Tämänkaltaisessa luentomonisteessa olisi hyödyllistä ja tyylikästä numeroida yhtälöitä, kuvioita ja taulukoita niiden esiintymisjärjestyksessä ja tehdä niihin viittauksia. Tämän materiaalin pohjalta on kuitenkin luotu luentovideoita, jolloin niiden toiminnan ja pidemmän aikavälin käytettävyyden kannalta on järkevää toimia tässä yhteydessä ilman numerointia yhtälöiden ja sivunumeroiden osalta.

1.1 Tilastotiede ja kurssin idea

- Tämän tilastotieteen ensimmäisen kurssin ideana on (ainakin)
  - Esitellä ja johdatella **tilastolliseen ja tieteelliseen ajatteluun** ja sen hyödyntämiseen eri tyyppisissä tutkimusongelmissa.
  - Esitellä tilastotieteen roolia **empiirisen tutkimusaineiston keräämisessä ja analyysissä** sekä tarkastella tieteentekemisen ja tilastotieteen suhdetta.
  - Pohtia **tilastotieteen olemusta tieteenalana** ja tarkastella tilastotieteen ja **datatieteen** (data sciencen) samankaltaisuuksia 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.
  - 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, kuten **odotusarvo**, **varianssi** ja kahden (tai useamman) satunnaismuuttujan **korrelaatio** ja mahdollinen kausaalinen yhteys.
  - Satunnaismuuttujien **todennäköisyysjakaumien** perusteita ja niiden yhteyksiä mm. **normaalijakaumaan** ja muutamiin muihin keskeisiin jakaumiin.

  - **Tilastollinen malli** työkaluna satunnaismuuttujien formaalissa mallintamisessa ja päättelyssä. Tilastollisiin malleihin liittyy (usein) **parametreja** joihin tilastollinen päättely kohdistuu.
  - Tilastollisten mallien **estimoinnin** perusidea, eli miten tilastollisen mallin tuntemattomille parametreille muodostetaan arvot käytettävissä olevan aineiston pohjalta. Esimerkiksi: mitä tarkoittaa tilastollisen mallin parametrin **estimaattori**, sen **tarkentuvuus** ja **harhattomuus**.
  - Alustavia tarkasteluja tilastollisen malliin liitettävän **uskottavuuden** käsitteelle.

- Toinen kurssin keskeisistä teemoista on tarkastella **tieteellistä tutkimusprosessia** teoriassa ja käytännössä tilastotieteen näkökulmasta. Tämä sisältää mm. seuraavia aiheita, joita siis käsitellään tällä kurssilla päällisin puolin varsin yleisestä näkökulmasta katsoen ja tarkemmat yksityiskohdat jätetään tätä kurssia seuraavien tilastotieteen kurssien aihepiireiksi:

  - **Tutkimusongelman** asettaminen. Mitä halutaan tutkia?
  - Tutkimusongelman täsmentäminen 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?
    * **Otanta** (ja mittaaminen): miten tutkimusaineisto kerätään niin, että se täyttää hyvältä aineistolta vaadittavat ehdot? 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 eri ajanhetkiltä
  - **Aineiston kuvaaminen**: minkälaista aineistoa on kerätty ja vastaako se ennakkoehtoja?
  - **Aineiston tilastollisen analyysin** lähtökohtia:
    * Mitä tilastollista mallia/malleja käytetään?
    * Mitä tarkoitetaan mallien tuntemattomien parametrien arvojen estimoinnilla?

    * Tilastollinen päättelyn perusteita

  – **Johtopäätelmien** tekeminen tilastollisen päättelyn pohjalta: saatiinko tutkimusongelmaan vastaus ja kuinka luotettava saatu vastaus on?

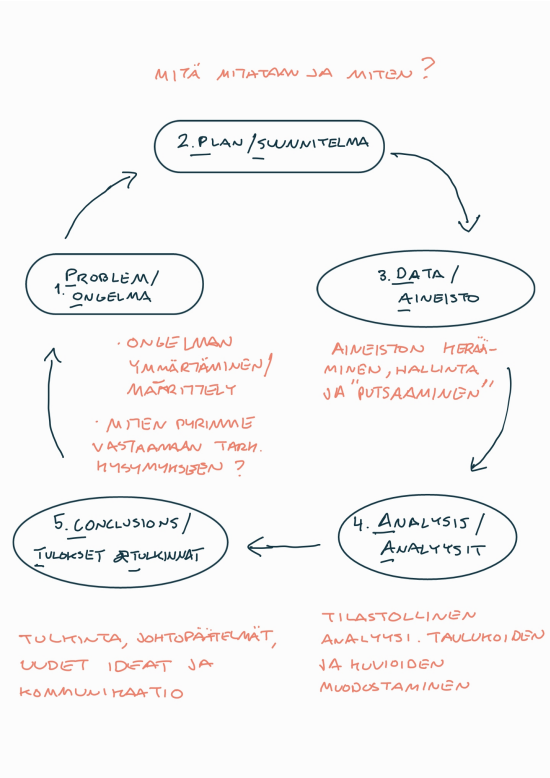
1.2 Tilastollinen lukutaito ja OSAAT-analyysisykli

Tilastotieteellä on pitkä ja menestyksekäs historia. Parhaillaan se on kuitenkin nyt muuttumassa laajojen tilastoaineistojen saatavuuden kasvaessa ja laajentuessa myös sellaisille aluille, joiden yhteydessä ei tilastollisia analyysejä ole vielä juurikaan tehty.

Tilastollinen lukutaito (datalukutaito) on keskeinen taito nykymaailmassa. Tämän jatkumona omatoiminen kyky ja taito tehdä tilastollisia analyysejä on tärkeää myös monille muille kuin tilasto- ja datatieteilijöille. Näiden huomioiden myötä myös tilastotieteen opetus on, ainakin joltain osin, muuttumassa lähinnä vain matemaattisiin menetelmiin keskittymisestä ja menetelmälistoista sellaiseen esitystapaan, mikä perustuu ongelma-ratkaisu-sykliä vastaavaan esitystapaan.

- Sukupolvien ajan opiskelijat ovat "vastaanottaneet" varsin kuivia tilastotieteen (perus)opintokursseja, jotka tyypillisesti pitivät sisällään suuren joukon erilaisia tekniikoita, ja muutamia niihin liittyviä sovelluksia, keskittyen enemmän tilastomatemaattiseen esitystapaan.

Tätä muutosta korostaa myös oheinen (englanniksi) ns. "**PPDAC-sykli**", mikä voidaan hieman vapaasti mutta myös osuvasti kääntää suomeksi "**OSAAT-sykliksi**", ja sitä käsitellään ja siihen palataan koko tämän kurssin materiaalin aikana.



Kuva 1: PPDAC-sykli (Spiegelhalter, 2020, kuva 0.3) eli suomeksi "OSAAT-sykli".

PPDAC/OSAAT-syklin (Spigelhalter, 2020, kuva 0.3) ensimmäiset kaksi vaihetta ovat (1.) ongelman määrittely ja (2.) tutkimuksen toteuttamisen suunnittelu.

- Tilastollinen tutkimus, kuten kyselytutkimus, alkaa aina kysymyksellä. Useimmiten tämä merkitsee pyrkimystä koota mahdollisimman paljon ja mahdollisimman tarkkaa tietoa. Huom: On usein houkuttelevaa ohittaa huolellisen suunnitelman tarve.
- Ts. kyse on tiedon tarpeesta (jotakin ilmiötä koskien), mikä johtaa osaltaan myös samalla ongelman määrittelyyn.

Esimerkiksi puiden lukumääriä laskeneet ihmiset kiinnittivät pikkutarkkaa huomiota tarkkoihin määritelmiin ja siten siihen miten mittaukset tehdään, sillä luotettavat johtopäätökset voidaan tehdä vain tutkimuksesta, joka on asianmukaisesti suunniteltu.

- Kohdan 2 voidaan nähdä sisältävän myös kirjallisuuteen perehtymisen sekä tutkimuksen toteutuksen tarkemman suunnitelman, kuten sopivan otantamenetelmän valinnan.
- Valitettavasti, usein kiireestä johtuen, saatujen tietojen (ts. data, kohta 3) jälkeen on suuri houkutus aloittaa tilastolliset analyysit (4.) ilman, että mietitään tarkemmin mihin (ja miten) vastauksia haetaan.
  - Otanta voidaan nähdä myös osaksi kohtaa 3.
- Pääpaino tilastotieteen (perus)kursseilla on perinteisesti ollut lopulta analyysivaiheessa (kohta 4) menetelmälistojen läpikäynnin jälkeen. Tällä kurssilla esitellään muutamia keskeisiä analyyttisiä tekniikoita, joita tilasto- ja datatieteessä voidaan käyttää, kuten aineiston visualisoinnissa ja regressioanalyysin perusteita. Nämä liittyvät tilastolliseen mittaamiseen ja laajemmin data-analyysin toteututtamiseen.
- Kohta 5: Hyvä tilastotiede perustuu perusteltujen johtopäätösten tekemiseen ja tulosten raportointiin, jotka tunnustavat saatujen empiirisen "todistusaineiston" rajoitukset ja välittävät ne selkeästi, esimerkiksi graafisten esitysten avulla.
  - Johtopäätökset herättävät yleensä lisää kysymyksiä, jolloin syklin kierto alkaa alusta.

Vaikka PPDAC/OSAAT-sykliä ei aina noudateta, eikä edes voida noudattaa, tarkasti, se korostaa tilastollisen analyysin muodollisten tekniikoiden merkitystä tilastotieteilijän tai datatieteilijän työssä. Tilastotiede on siis selvästikin paljon enemmän kuin matematiikan haara, johon liittyy "esoteerisia" kaavoja, joiden kanssa eri sukupolvet ovat usein kamppailleet osana yliopisto-opintojaan!

1.3 Tilastotieteen asema tutkimusyhteisön ulkopuolella

Tilastotiede on oppiaineena usein varsin tuntematon toisen asteen opinnoista valmistuneelle, sillä sitä ei juurikaan opeteta lukioissa (ja/tai ammattikouluissa) **pl. tilastojen ja todennäköisyyslaskennan perusteet** (joihin myös palataan tämän materiaalin aikana) huolimatta tilastotieteen keskeisestä ja kasvavasta roolista tieteenteossa.

- Tiedeyhteisön ulkopuolellakin **tilastotiedettä ja tilastotieteilijöitä tarvitaan ja arvostetaan laajalti**!

- **Tilastotiede on nostanut profiiliaan viimeisten vuosikymmenien aikana** tietoteknisen kehityksen tuotua laajat tietoaineistot ja kehittyneet laskennalliset menetelmät lähes jokaisen käyttäjän ja viime kädessä kansalaisen saataville.

- Tämä "**datavallankumous**" näkyy tilastotieteilijöiden kysynnässä työmarkkinoilla: erilaisten aineistojen laajentuessa ja määrän lisääntyessä kasvaa myös kysyntä työntekijöistä, jotka osaavat ammatitaitoisesti käsitellä (ml. koodaamistaidot), tulkita ja mallintaa tilastollisia aineistoja.

- Ei siis liene ihmekään, että erilaisten "data"-alkuisten työpaikkojen, kuten **datatieteilijä** (eng. *data scientist*) tai **data-analyytikko** (*data analyst*) määrä on kasvanut voimakkaasti jo pidempään. Kaikkia tieto- ja dataintensiivisten ammattien tekijöitä yhdistää yksi tekijä: **heidän tulee hallita ja osata tilastotiedettä!**

  - Karkeasti yksinkertaistaen, mitä paremmin ja enemmän (laajemmin), sen parempi palkka ja monipuolisemmat työtehtävät!

1.4 Kurssin luonne tilastotieteen opintojen esittelijänä

Kurssin mittaan esitellään tilastotieteen perusteiden lisäksi **miten Turun yliopistossa tilastotieteen opinnoissa syvennytään** tällä kurssilla lyhyesti esiteltäviin menetelmiin, aineistotyyppeihin ja mallinnuskokonaisuuksiin.

- Tilastotieteen opintotarjontaan voi perehtyä TY:n opinto-oppaan avulla (ks. Matematiikan ja tilastotieteen laitos, vuosien 2024–2027 opinto-opas. Tarkista ja varmista, että käytät aina uusinta opinto-oppaan versiota)

https://opas.peppi.utu.fi/fi/perustutkintokoulutus/matemaattis-luonnontieteellinen-tiedekunta/14002/13354?period=2024-2027

- Ks. lisäksi Turun yliopiston tilastotieteen keskuksen opetusta koskevat sivut:

https://www.tilastotieteenkeskus.fi/

Kaiken kaikkiaan tämä materiaali käyttää reaalimaailman esimerkkejä ja ongelmanratkaisua lähtökohtana tilastollisten menetelmien ja ideoiden käyttöönotolle. Jotkut näistä ajatuksista voivat tuntua itsestään selvältä, mutta jotkut ovat hienovaraisempia ja saattavat vaatia jonkin verran ponnistelua, vaikka kovin pitkälle meneviä matemaattisia taitoja ei tarvita.

- Perinteisiin tilastotieteen perusopintojen materiaaleihin verrattuna tämä kurssi keskittyy käsitteellisiin kysymyksiin, ei niinkään tilastomatemaattisiin ja teknisiin yksityiskohtiin, puhumattakaan pitkistä listauksista erilaisista tilastollisista menetelmistä ja testeistä.
- Kurssin ja tämän materiaalin esimerkeissä käytetään ajoittain **R/RStudio** -ohjelmointikieltä. Mitä tahansa muutakin ohjelmaa, kuten Pythonia tai Stataa, voi hyvin myös käyttää, mutta kurssimateriaali siis "tukee" nimenomaan R:ää. R-kielen alkeita koskevan peruskurssin käymistä suositellaan tämän materiaalin aikana, mutta toisaalta tämä materiaali ei kuitenkaan varsinaisesti vaadi R/RStudion käyttämistä ja hallintaa.

Chapter 2

Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa

Tiedelukutaito Tässä luvussa tarkastellaan tieteen ja tieteellisen tutkimusprosessin luonnetta erityisesti **uuden tutkitun** tiedon tuottamisen näkökulmasta.

Tieteellinen tutkimus ja asiantuntijatyö tuottavat valtavan määrän perusteltua, luotettavaa tutkimustietoa. Vastuulliseesti tuotetut tiedeartikkelit tarjoavat tietoa siitä, kuinka tutkittua tietoa tuotetaan, julkaistaan ja arvioidaan luotettavasti ja yhteisesti hyväksytyllä tavalla. Jotta tiede vaikuttaa koko yhteiskunnan hyväksi, toiminnan on oltava vastuullista tutkimuksen jokaisessa vaiheessa.

Tiedeviestintä on tiedeyhteisöjen sisäistä ja ulkoista tiedonvälitystä ja vuorovaikutusta:

- Julkisuus ja avoimuus tekevät tutkimuksesta tiedettä.
- 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.

Tiedelukutaidon merkitys on kasvanut nyky-yhteiskunnassa. Tämä on osin tiedejulkaisujen saavutettavuuden ja tunnettavuuden lisääntymisen tulosta,

mikä saattaa liittyä mm. tieteen popularisointiin ja median laajemman tiedeuutisointiin.

- Ks. tarkemmin tieteellisestä julkaisemisesta ja tutkimuksen tekemisestä Helsingin yliopiston tarjoaman Tiedelukutaidon perusteet -kurssin materiaalia seuraten, joka on julkaistu oheisena MOOC-kurssina (Massive Open Online Course):

https://tiedelukutaito.mooc.fi/

- Keskustelethan ennen kurssin käymistä (jos siis haluat suorittaa ko. kurssin) oman alasi koulutussuunnittelijan (tai vastaavan vastuuhenkilön) kanssa siitä, soveltuuko kyseinen kurssi sisällytettäväksi johonkin omaan opintokokonaisuuteesi.

Esimerkki (tiedon rooli): Tiedon, erityisesti tieteellisen tiedon, rooli korostuu yhä enemmän kaikilla elämän osa-alueilla. Näistä muutamia esimerkkejä:

- Terveysteknologia (esim. sykemittarit tai Oura-sormus) perustuu lääke- ja terveystieteellisiin läpimurtoihin.
- Talouspoliittisia päätöksiä edeltää entistä suurempi määrä asiantuntijoiden taloustiedeperusteista (ei välttämättä kuitenkaan yksimielistä) analyysiä.
- Jopa peruskouluopetus on murroksessa kasvatustieteen tutkimussaavutusten myötä.

Voidakseen ymmärtää ja arvioida kriittisesti tiedeuutisia tulee lukijan olla tietoinen tieteellisen tutkimuksen luonteesta, kuten:

- miten tutkimusartikkeleja luetaan,
- mitä niiltä voidaan odottaa, ja
- minkälaiset tulokset ovat uskottavia?

Tilastotiede näyttelee keskeistä roolia lähes kaikessa tieteellisessä tutkimuksessa ja erityisesti erilaisten tutkimuskysymysten ja niitä vastaavien hypoteesien testauksessa.

2.1 Mitä on tiede?

Aloitetaan kurssin varsinainen oppimateriaali kunnianhimoisesti tarkastelemalla mitä tiede oikeastaan on (lopulta erityisesti tilastotieteen näkökulmasta).

Annetaan tieteen määritelmälle ensin muutamia pohtivia suuntaviivoja:

- Tiede ja tieto *Tiede on **järjestelmällistä** ja **järkiperäistä uuden tiedon hankintaa**. Näin tiede määritellään toiminnaksi, jossa **tavoitellaan** ja **hankitaan tietoa**. (Lähde:* Haaparanta ja Niiniluoto (2016, s.28) (Ks. Oheislukemistoa))
- Tieteellinen tutkimus on tutkivan subjektin ja tutkimusobjektin välistä vuorovaikutusta.
- Tiede pyrkii järjestämään tiedon yksinkertaisiksi kokonaisuuksiksi 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 ilmiöitä ja niiden välisiä suhteita:

- selitetään,
- niitä koskevia käsityksiä vahvistetaan osoittamalla ne tosiksi (tai päinvastoin), sekä
- löydetään niistä uutta tietoa.

Tiede siis erottaa intuition ja "arkitiedon" oikeasta, tutkitusta tiedosta esittämällä reaalimaailmaa koskevia väitteitä ja osoittamalla ne tosiksi, tai epätosiksi, tieteellisin menetelmin.

- Tiede käsittää siis myös aiemman tutkimuksen ja se toimii kaiken tieteellisen tiedon jäsenneltynä kokonaisuutena.

Arkitieto ja tieteellinen tieto. Tieteen tekemiseen liittyvä vaatimus **uudesta tiedosta** kuitenkin sulkee tieteen ulkopuolelle toiminnot, joissa on kyse vain aikaisemmin hankittujen tietojen omaksumisesta ja järjestämisestä (Haaparanta ja Niiniluoto, 2016, s.28).

- Vrt. esimerkiksi opiskelu tai (useimmiten) myös erilaiset komitea- ja 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 (Haaparanta ja Niiniluoto, 2016, s. 29):

- **Järjestelmällisyys**: Tieteellinen tiedonhankinta on yhteiskunnallisesti organisoitu tutkimusta tekevien (ja opetusta järjestävien) instituutioiden tehtäväksi, joka kokoaa tutkimustulokset systemaattisiksi tietojärjestelmiksi niin kansallisella kuin kansainvälisellä tasolla. Näihin instituutioihin lukeutuvat mm. yliopistot, korkeakoulut ja tutkimuslaitokset ja vastaavasti tietojärjestelmiksi mm. tieteelliset julkaisut. Tiede ylittää järjestelmällisyytensä vuoksi tiedostamisen "arkitason".
- **Järkiperäisyys**: Järkiperäisyyden vaatimus asettaa rajoitteita tieteelliselle ajattelutavalle. Tiede ei siis voi nojautua esim. yksilölliseen vaistoon tai intuitioon, suostutteluun, propagandaan tai "Jumalalliseen ilmoitukseen" tai vastaavaan.

Tieteen keskiössä on todellista maailmaa koskevat **(tieteelliset) teoriat** ja niihin liitettävät **hypoteesit**.

Tieteellinen teoria

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

Esimerkkejä tieteelisistä teorioista ovat esim. Einstenin suhteellisuusteoria tai Darwinin evoluutioteoria.

- Teoria muodostuu tieteellistä menetelmää käyttämällä 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/tai **ennustaa**, sen kohteena olevaa ilmiötä tyylikkäästi sekä yksinkertaisesti. Huomiona, että tässä yhteydessä ennustamisella tarkoitetaan yleismaailmallista ennustamista (ennakointia jne.), joka saattaa poiketa tarkemmasta tilastollisesta ennustamisesta.
  - Se on luonteeltaan *induktiivinen* ja alisteinen muutoksille tai jopa hylkäämiselle empiirisen todistusaineiston ("evidenssin") osoittaessa sen olevan puutteellinen tai väärä.

  - Tieteellisen teorian tulee siis olla **empiirisesti testattavissa/koeteltavissa** ja sen tekemät ennusteet on tarvittaessa osoitettavissa vääriksi. Teoriaan liittyvät ennustukset määrittelevät sen hyödyllisyyden, sillä teoria joka ei tee testattavia ennustuksia on usein hyödytön.

Tieteelliset 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**, mitä voidaan pitää parhaana mahdollisena selityksenä kyseiselle ilmiölle ko. hetkellä (ja se voi ja usein kehittyykin jatkossa).
- 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 mahdollisesti muuttaa kokonaan.

Hypoteesi

Hypoteesi tarkoittaa esim. teorioista johdettua tai aikaisemman tutkimuksen perusteella esitettyä ennakoitua ratkaisua tai selitystä tutkittavaan ongelmaan.

- Erityisesti ja myöskin tilastotieteen näkökulmasta hypoteesi ilmaistaan teoriaa koskevana väitteenä, jonka paikkansapitävyyttä halutaan tutkia. Käytännössä hypoteesit liittyvät **tilastollisten mallien parametreihin** (tähän palataan lyhyesti myöhemmin jo tällä kurssilla).
- Hypoteeseja voidaan tilastollisesti testata ja näin saatavan empiirisen todistusaineiston perusteella voidaan hypoteesi/hypoteesit osoittaa vääriksi tai jättää voimaan.

Tiedon kumuloituminen

Edellä kuvattu tieteellisen **tiedon kumuloituminen** muokkaa teorioita vuosien saatossa täsmällisemmiksi ja paremmiksi kuvauksiksi ympäröivästä maailmasta.

- Yksittäinen (vahva) tutkimustulos on vasta alku ja vahvistettu tieto jostain ilmiöstä, yhteydestä tai vaikutuksesta syntyy monien mittausten ja tutkimusten jatkumona. Tietoa ei siis voida johtaa siitä, miltä asiat näyttävät, kuten on tyypillistä "arkiajattelussa".

- 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.

Esimerkkejä (ks. myös Haaparanta ja Niiniluoto, 2016, s. 130):

1900-luvun alussa klassinen newtonilainen fysiikka ei enää riittänyt selittämään kaikkia havaintoja, ja se täydentyi kahdella uudella teoriakokonaisuudella: kvanttimekaniikalla ja Einsteinin suhteellisuusteorialla.

Biologian alalla tapahtui vastaava murros: Darwinin Lajien synty (1859) esitti evoluutioteorian, joka myöhemmin yhdistettiin Mendelin perinnöllisyystutkimuksiin. Näin syntyi synteettinen evoluutioteoria, jota on sittemmin täydennetty mm. DNA:n rakenteen ja molekyylibiologian löydöillä.

Täyttävätkö nämä tieteellisen teorian määritelmät? Kyllä.

- Teoria on muodostunut tieteellistä menetelmää käyttäen (havainnot, kokeet, vertailut)
- Se on kumulatiivinen: uusia havaintoja ja teknologioita on lisätty ajan myötä
- Se on toistuvasti vahvistunut uusilla löydöillä (fossiilit, DNA, kokeelliset havainnot)

Tieteelliselle ajattelulle ja tiedon tuottamiselle on vastaavasti tunnusomaista, että se pohtii ja kehittelee (tutkimusalakohtaisia) **paradigmojaan** eli oman toimintansa perusteita (raameja). Paradigmat antavat suuntaviivoja ja viiteistöjä siitä, minkälainen tutkimus tuottaa uskottavia tuloksia.

Paradigma

Paradigma on tietyn alan oman tieteellisen toiminnan oppirakennelma, ajattelutapa ja peruste, joka mm. ohjaa tutkimuskysymysten asettelua, käytettäviä menetelmiä ja tulosten tulkintoja. Paradigmat elävät jatkuvassa muutoksessa tieteen kehityksen myötä.

Esimerkkeinä paradigmoista voitaneen ajatella mm. laskennallisuuden kasvamista tilastotieteen yhteydessä (laskentaintensiiviset menetelmät) sekä taloustieteen nk. “uskottavuusvallankumousta”, jossa tilastollisten menetelmien myötä taloustieteellisen tutkimuksen painopiste tuntuu siirtyneen vahvemmin ns. empiirisen kausaalitutkimuksen puolelle.

Paradigmojen ei pidä ajatella olevan kaavoihin kangistuneita ajattelu- ja menettelytapoja, 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 ajoittain estävät osaltaan uusien tieteellisten 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.

2.2 Tilastollinen päättely, populaatio ja otos

Tieteellinen ajattelutapa Tieteilijät yleensä perustavat hypoteesinsa aikaisemmin tehtyihin havaintoihin joita ei voida selittää olemassa olevilla tieteellisillä teorioilla tyydyttävästi. Tilastollinen päättely Uuden tieteellisen tiedon tuottaminen ja jo tuotetun tiedon ymmärtäminen vaatii **tieteellisen ajattelutavan** omaksumista, jonka perustana on lähes aina **tilastollinen päättely**. Tilastollisen päättelyn perusteita tarkastellaan myös tämän kurssimateriaalin myötä.

Tilastollisen päättely mahdollistaa ja sen keskeinen tavoite on tehdä perusteltuja päätelmiä (yleistyksiä) tarkasteltavasta **populaatiosta** käytettävissä olevan aineiston, usein **otoksen**, perusteella. Tilastollinen päättely sisältää mm. hypoteesien testauksen, tilastollisten mallien tuntemattomien parametrien (optimaalisten) numeeristen arvojen eli estimaattien muodostamisen ja parametrien luottamusvälien määrittämisen.

Tilastollinen päättely perustuu siis tilastotieteen matemaattiseen perustaan ja teoriaan, joka mahdollistaa päättelyn luonteeltaan epävarman ja satunnaisen aineiston tapauksissa.

- Esimerkiksi hypoteesien testaaminen osana tilastollista päättelyä on yhtäältä tieteellisten teorioiden kehittämistä ja vahvistamista ja toisaalta kritiikin keskiössä.
- Hypoteesien asettaminen voidaan ajatella tutkittavaa ilmiötä koskeviksi ennustuksiksi, joita verrataan havaittuun aineistoon. Mikäli havaittu aineisto ei ole yhteensopivaa testattavan teorian tai siihen liittyvien hypoteesien kanssa, voidaan (hieman yksinkertaistaen) teoriaa kehittää paremmaksi. Tämä vuoropuhelu vie tiedettä eteenpäin ja tuottaa lisää tutkittua tietoa ympäröivästä maailmasta.

Määritellään tässä vaiheessa (melko yleisellä tasolla) populaatio ja otos. Näihin palataan useaan kertaan vielä myöhemmin tämän kurssin aikana.

Populaatio/perusjoukko

Populaatio eli **perusjoukko**. Konkreettinen tai hypoteettinen tutkimuskohteiden joukko, joka koostuu kaikista tutkimuksen kohteena olevista tilastoyksiköistä.

Tilastoyksikkö

Tilastoyksikkö ja tilastollinen muuttuja. Populaation muodostavilta tilastoyksiköiltä (populaation alkioilta) tarkastellaan tilastollisia muuttujia, joita voidaan mitata tai havaita. Ts. tutkimuskohteita kutsutaan tilastoyksiköiksi.

Otos

Otos on populaation osajoukko, jota käytännössä tutkitaan tilastollisia menetelmiä käyttäen.

2.3 Tieteelliset ja tilastolliset menetelmät

Milloin tutkimus sitten on tieteellistä? Tiede voidaan nähdä tiedonhankintana, jossa käytetään erityistä, mahdollisesti tilanteesta (sovelluksesta) riippuvaa, **tieteellistä menetelmää** eli **metodia**.

Tieteellinen menetelmä/metodi

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 pyrkiä totuuteen erilaisten tutkimusongelmien ratkomisessa.
- Metodologinen pluralismi: Kaikkia menetelmiä voi soveltaa hyvin tai huonosti, mutta niitä voi käyttää myös luovasti väärin.

Hyvälle tieteelliselle menetelmälle voidaan lukea seuraavia kriteerejä (ks. esim. Haaparanta ja Niiniluoto, 2016, s. 38–40):

1. **Objektiivisuus ja loogisuus**

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

2. **Kriittisyys**

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

3. **Autonomisuus**

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

4. **Edistyvyys**

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

5. **Toistettavuus ja yleistettävyys**

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

Tarkastellaan lähemmin erästä määritelmää erilaisille **toistettavuuden** (**replikoinnin**) lajeille. Esittelemme tässä Hamermeshin (2007) 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, 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.

Malli Pyrittäessä jäsentämään ja ymmärtämään havaintoaineistoa, ja mahdollisesti tämän pohjalta ennusteita tehtäessä, tutkijat sanovat usein rakentavansa ja käyttävänsä tieteellisten teorioiden sijaan **malleja**. Mallin käsite palvelee monissa eri tehtävissä sekä tieteessä että sen ulkopuolella ml. konkreettiset työtehtävät monissa eri tapauksissa.

Malli on yksinkertaistettu (idealisoitu) esitys todellisuudesta. Niiden ajatuksena on tiivistää oleellista tietoa todellisuudesta (tarkasteltavasta ilmiöstä), minkä seurauksena ne ovat järjestään helpompia ymmärtää (ja tulkita) usein hyvin monimutkaiseen todellisuuteen verrattuna.

Tieteellinen malli Tieteellisten 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.
- Ajoittain tieteellisellä mallilla tarkoitetaan teoreettista mallia. Se muistuttaa (tieteellistä) teoriaa, mutta se ei kuitenkaan pyri olemaan yhtä tarkka kuvaus tarkasteltavasta ilmiöstä (ks. esim. Haaparanta ja Niiniluoto, 2016, s. 60)
- Yksi teoreettisen mallin muoto on matemaattinen tai tilastollinen (teoreettinen) malli, joka tyypillisesti koostuu joukosta yhtälöitä ja erilaisia matemaattisia merkintöjä.

Simulointi

Esimerkiksi simulaatiomallit tarjoavat mahdollisuuden tarkastella miten (teoreettinen) malli käyttäytyy käytäntöä vastaavissa tilanteissa pyrittäessä jäljittelemään mahdollisesti monimutkaisenkin kohdesysteemin toimintaa. Näitä voivat olla esim. sää tai (makro)taloutta koskevat mallit. Simulaatiomalleihin liittyy myös usein melko arkikielessäkin nähtävät viittaukset **Monte Carlo**

-menetelmiin, jotka ovat esimerkkitapauksia simulointiin perustuvista tilastollisista menetelmistä.

Vaikka malli ei ole itsessään teoria tai sellaisen ehdotus, se voi kuitenkin olla käyttökelpoinen uusien teorioiden kehittelyssä ja vanhojen laajentamisessa.

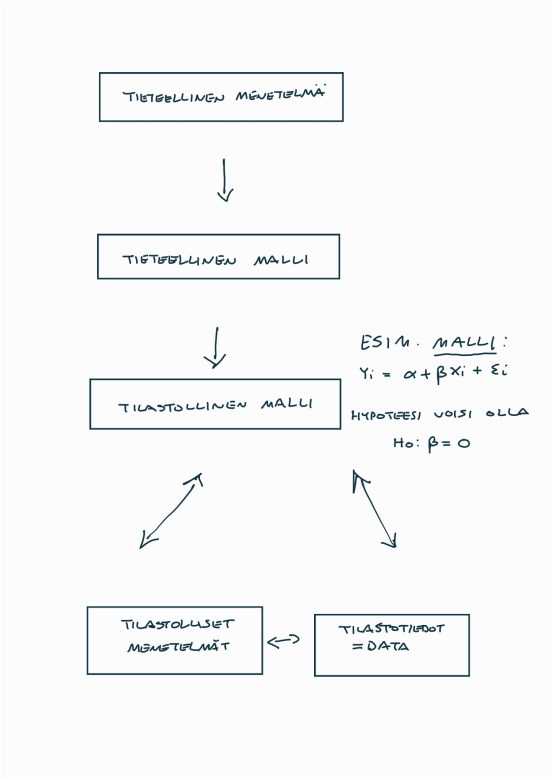
- (Tilastolliset) mallit hyödyntävät matemaattista esitystapaa, sillä se tarjoaa formaalin ja objektiivisen tutkimusaiheen kuvauksen sekä mahdollistaa siihen liittyvän loogisen päättelyn käytettävissä olevan aineiston pohjalta.

Tilastollinen malli

Tilastolliset mallit ovat käytännössä tieteellisiin malleihin kuuluvia formaaleja tilastomatemaattisia malleja ja yksinkertaistuksia todellisuudesta. Tilastollinen malli auttaa ymmärtämään ja ennustamaan mallin taustalla olevan tai olevien **satunnaismuuttujien** käyttäytymistä. Malli koostuu matemaattisista yhtälöistä, jotka kuvaavat, miten muuttujat liittyvät toisiinsa ja miten ne **jakautuvat**.

- Mallit sisältävät tuntemattomia **parametreja**, joiden arvot estimoidaan (arvioidaan) datan perusteella.

- Mallien avulla voidaan tehdä päätelmiä ja halutessamme myös ennusteita.



Kuva: Tieteellisen tiedon ja tilastojen/tilastotieteen välisiä suhteita. Malliin (esimerkiksi lineaariseen regressioon) ja sen merkintöihin palataan myöhemmin.

2.4 Tilastojen yleisestä roolista yhteiskunnassa

Tilastotiedot Ihminen ei voi (nykyaikana) toimia maailmassa järkevästi, ellei hän pysty muodostamaan oikeata kuvaa maailmasta ja sen tilasta. Yhteiskunnan kaikilla sektoreilla toiminnan seuranta, päätöksenteko ja ennakointi perustuvat oikein niitä kuvaaviin ja ajantasaisiin **(tilasto)tietoihin** ja niiden analysoinnissa käytettäviin **tilastollisiin menetelmiin**.

- (Tilasto)tietojen saatavuutta voidaan pitää jopa toimivan demokratian edellytyksenä! Tilastolliset menetelmät

Esimerkki: **Tilastotietojen (yhteiskunnallinen) vaikuttavuus**. Esimerkiksi päätöksenteko sekä julkisella että yksityisellä sektorilla (elinkeinoelämässä) perustuu 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.

Samoin esimerkiksi tuotantoprosessien ohjaus ja laadunvalvonta teollisuudessa sekä markkinatutkimus kaupan alalla perustuvat tyypillisesti tilastollisiin menetelmiin.

Epävarmuus ja satunnaisuus Koska todellisuutta kuvaaviin (tilasto)tietoihin sisältyy (lähes) aina **epävarmuutta** ja **satunnaisuutta**, tilastotiede ja tilastolliset menetelmät luovat perustan **tilastojen tuotannolle, jalostukselle ja analysoinnille** epävarmuuden vallitessa.

- Niinpä tilastojen tuotantoon, jalostukseen ja analysointiin liittyvien menetelmien kehittäminen on keskeinen osa-alue tilastotieteen tehtäväkentällä.
- Ylipäätään tilastotieteen menetelmien ymmärtämisellä on siis keskeinen rooli tietoyhteiskunnassa toimimisessa ja vaikuttamisessa.

Esimerkki (väite): Naiset puhuvat enemmän kuin miehet.

- Matthias R. Mehl ym. (Are Women Really More Talkative Than Men? *Science*, 317 (5834), 82). https://www.science.org/doi/10.1126/science.1139940
- Lähtökohta väitteen (hypoteesin, ks. yllä) tutkimiseen:
  - Uskomus on väärä kunnes toisin todistetaan. Lähdetään siis 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. Tämä on mahdotonta!
- Mitä siis tehdä?
- Täytyy tyytyä tutkimaan osajoukkoja miehistä ja naisista (otos), mihin tarvitaan **otantamenetelmiä** (käsitellään laajemmin vielä myöhemmissä luvuissa)

  - Arvotaan satunnaisesti tutkimushenkilöitä miesten ja naisten joukosta ja mitataan kuinka paljon he puhuvat.
  - Satunnaisuus on tässä tärkeää, sillä jos valikoitaisiin tarkoituksella puheliaita tai vähäsanaisia tutkimushenkilöitä, tulokset vääristyisivät.

Jokaiseen mittaukseen liittyy virhettä. Nimittäin täysin satunnainenkaan otos ei edusta täydellisesti koko väestöä. Joukkoon saattaa valikoitua puhtaasti sattumaltakin poikkeuksellisen puheliaita tai harvasanaisia naisia tai miehiä.

- Ks. aiemmin esitetyt populaation ja otoksen määrittelyt!

Otoskoolla, eli sillä kuinka monta tutkimishenkilöä tutkitaan, on myöskin keskeinen rooli tutkimuksen luotettavuudelle. Mitä suurempi otos, sitä pienemmäksi sattuman osuus käy ja vastaavasti mitä pienempi otos, sitä suurempi on yksittäisten sattumien vaikutus.

Tilastolliset mallit turvautuvat todennäköisyyksiin erottaakseen sattuman vaikutuksen: kun aineisto on kerätty, halutaan tietää kuinka todennäköistä on, että uskomus pitää paikkaansa.

Esimerkki, jatkoa. 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 puheen määrään paljon suurempi vaikutus kuin sukupuolella.
- Kiitos tilastotieteen, väärä uskomus on korvautunut tiedolla!

2.5 Mitä on tutkimus?

Tutkimus **Tiede tavoittelee tietoa, mutta mistä?** Jokaisen tutkimuksen lähtökohtana on (tai ainakin pitäisi useimmiten olla) tiedollisen uteliaisuuden, käytännön tarpeiden tai teorian kehittämispyrkimyksen herättämä **ongelma**, johon tutkimuksen avulla etsitään vastausta.

- Tutkimus yrittää käsittää sekä tutkitun 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, ja
  - 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 järisyttäviin löydöksiin, jotka tosin voivat olla hyvin epävarmoja!

Millaisia kysymyksiä tutkimuksessa asetetaan (voidaan asettaa)?

- **Kuvaus**

  Esimerkki: Kuinka suuri on yli 65-vuotiaiden osuus Suomen väestöstä?

- **Riippuvuuden kuvaus**

  Esimerkki: Ovatko paljon mainostavat yritykset kannattavampia kuin vähän mainostavat?

- Kuvattujen ilmiöiden **selittäminen** ja **ymmärtäminen**.

  Esimerkki: Miksi vanhempien sosioekonominen asema vaikuttaa ekonomien työhönsijoittumiseen? Tämän tutkimuskysymyksen tapauksessa pyrkimys on lähinnä selittää (ymmärtää) ilmiötä.

- **Ennustaminen**

  Esimerkki: Jos kansantulon kasvu pienenee x%, työttömyyden ennustetaan kasvavan y tuhannella.

Näiden kohtien yhdistelmänä ja jatkona vielä mm. tutkimuksen kohdetta kuvaavien käsitteiden ja teorioiden rakentaminen, teorioiden ansioiden ja puutteiden arviointi. Myöhemmin tässä materiaalissa keskustellaan vielä hieman tarkemmin miten tilastotieteessä ilmiön ymmärtäminen/selittäminen ja ennustaminen eroavat toisistaan.

Tutkimuksen rajat? Onko niitä?

- Tutkimus antaa aina **vajavaisen kuvan** tutkimuskohteesta.
  - Kehittynytkin tieteellinen (ml. tilastotieteellinen) teoria tai malli on aina **reaalimaailman yksinkertaistus**: tutkimus on aina alisteinen käytetylle menetelmälle ja sen oletuksille!

- Ymmärtämiseen tarvittava havaintomaailman hahmotus (saattaa) tuottaa ideologisesti ja/tai historiallisesti sitoutuneita yksinkertaistavia, mahd. hyvinkin teoreettisia, abstraktioita.
  - Alakohtainen **substanssitietous** sekä sen vahvuuksien ja puutteiden sekä historiallisen ja mahd. ideologisen kontekstin tiedostaminen on ensiarvoisen tärkeää kaikessa tutkimuksessa.

Substanssitieto

Substanssitieto tarkoittaa syvällistä ja asiantuntevaa tietoa jostakin tietystä aihealueesta tai alasta. Se viittaa siihen, että henkilöllä on:

- Laaja ja syvällinen ymmärrys tietystä aiheesta (esim. lääketiede, oikeustiede, kasvatustiede tai tekniikka).
- Kyky soveltaa tietoa käytännön tilanteissa, ongelmanratkaisussa tai päätöksenteossa.
- Ajankohtainen ja luotettava tieto, joka perustuu tutkimukseen, kokemukseen ja/tai koulutukseen.

Ajoittain ja sovelluskentästäkin riippuen tietyissä tilanteissa tutkimusta voi ja saattaa joutua tekemään joistakin **arvolähtökohdista** lähtien, mutta sen tulisi olla näkyvää. Omien arvojen mahdollisimman selvä eksplikointi on yksi keino, jolla voi yrittää vähentää erilaisten piilossa olevien arvojen vaikutusta tutkimukseen.

- Arvot ilmenevät esimerkiksi tutkimuksessa käytetyissä käsitteissä, jotka eivät aina ole arvovapaita. Useimmat käsitteet voidaan korvata toisilla, joilla on paikoin hyvin erilainen arvosisältö, joskin arvottava lataus saattaa myös olla paikoin tarkoituksellista! Joka tapauksessa arvopainotteisten valintojen tunnistaminen saattaa olla vaikeaa.
- Toisaalta arvoihin sitoutuminen on väistämätöntä, sillä se on sosiaalisen olemassaolon sivutuote. Yhteiskunnan jäseninä meillä on tuskin mahdollisuuksia (täydellisesti) irroittautua arvoistamme kun pyrimme esim. ammatillisiin päämääriin.
  - Myös päinvastainen ongelma on olemassa: Tutkimusta arvioidaan siihen perustellusti tai perusteettomasti kiinnitettyjen arvonäkökohtien mukaan.

Joka tapauksessa täyteen neutraaliuteen ja objektiivisuuteen on usein mahdotonta päästä. Tästä huolimatta on hyvä ja tärkeää pystyä tunnistamaan tämä haaste.

Esimerkki: Luonnontieteelliset vs. yhteiskunnalliset sovellutukset:

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

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 tieteelliseen julkaisuun, jossa muut alan asiantuntijat arvioivat sen ja päättävät hyväksytäänkö se julkaistavaksi. Vertaisarviointi

(Akateemisen tutkimuksen) **vertaisarvioinnissa** yksi tai useampi, tehdystä tutkimuksesta riippumaton, saman alan tutkija lukee ja tarkastaa tehdyn (vielä julkaisemattoman) tutkimusartikkelin käsikirjoituksen, arvioi sitä ja suosittaa tieteellisen julkaisun arvioinnista vastaavalle päätoimittajalle (*editorille*) kyseisen artikkelin hyväksymistä tai hylkäämistä.

- Vertaisarviointi ei aina takaa sitä, että julkaistu tutkimus olisi virheetön ja erinomaisesti tehty, vaan myös väärää tietoa ja heikosti valmisteltuja artikkelikäsikirjoituksia 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 virhearviointeja sattuisikin.

Tutkimuksen kieli

- Tutkimus edellyttää arkikieltä täsmällisempää kommunikaatiota.
- Ongelmaan liittyvien käsitteiden huolellinen määritteleminen ja erittely on tarpeellista.
  - Käsitteiden ja eri aloilla, osin samoista asioista käytettävien, toisistaan eroavien termien systemaattinen määrittely ja jäsentely selkeyttää tiedeyhteisön välistä kommunikointia.
  - Eivät korvaa empiiristä tietoa vaan vaikuttavat tiedon järjestymiseen ja sen perusteella tehtäviin päätelmiin.

Tieteen ja etiikan suhde on kahtalainen (Haaparanta ja Niiniluoto, 2016, s. 153):

- Tieteellisen tutkimuksen tuloksilla on merkitystä eettisille valinnoille.
- Toisaalta taas eettisillä kannoilla on vaikutusta tutkijan ratkaisuille tieteellisessä työssä.

Vaikkei tiede asettaisikaan eettisiä päämääriä, se tarjoaa tietoia keinoista ja niiden yhteyksistä päämääriin sekä vaihtoehtoisista menettelyistä ja niiden seurauksista.

- Ks. Suomessa toimiva Tutkimuseettinen neuvottelukunta (TENK) ja sen hyvää tieteellistä tutkimusta koskeva määrittely:

https://tenk.fi/fi

Tutkimuseettinen neuvottelukunta on opetus- ja kulttuuriministeriön asiantuntijaelin, joka edistää hyvää tieteellistä käytäntöä, ennaltaehkäisee tiedevilppiä sekä edistää tutkimusetiikkaa koskevaa keskustelua ja tiedotusta… Tieteellinen tutkimus voi olla eettisesti hyväksyttävää ja luotettavaa ja sen tulokset uskottavia vain, jos tutkimus on suoritettu hyvän tieteellisen käytännön edellyttämällä tavalla.

Chapter 3

Tilastotiede tieteenalana

Tässä luvussa hahmottelemme tilastotieteen piirteitä tieteenalana.

- Käymme läpi tilastotieteelle ominaisia piirteitä, jotka erottavat sen niin lähitieteistä, kuten matematiikasta ja tietojenkäsittelytieteestä, sekä myös eri sovellusaloista.
- Usein näkee tilastotieteen typistettävän vain työkaluksi eri sovellusalojen empiiriseen tutkimukseen. Tämä siitäkin 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?

Tlastotiedettä kohtaan esitettyä tyypillistä kritiikkiä tarkastellaan vielä tämän materiaalin (Osan II) loppupuolella. Siis sen jälkeen kun olemme ensin tutustuneet tämän kurssin myötä tarkemmin mistä tilastotieteessä on kysymys!

3.1 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 tietoaineiston keräämistä, analysointia ja tulkintaa koskeva tiede* (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* (Mellin, 2004).
- **Vale, emävale, tilasto** (Mark Twain, 1907).
  - Twain popularisoi tämän lausahduksen teoksessaan *Chapters from My Autobiography* jo vuonna 1907. Huomionarvoista toki on, että valtaosa "modernin" tilastotieteen teoriakehityksestä, jolle nykytilastotiede pohjautuu, 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**.
- *Statistics concerns what can be learned from data* (Davison, 2003. *Statistical Models. Cambridge University Press.*)
- *Maalaisjärjen tehostamista* (Sund, 2003)

Tilastotieteen määrittelyä haettaessa on viimeistään tässä vaiheessa syytä esitellä tilastotieteen keskeinen (keskeisin) konsepti eli **todennäköisyys**. Todennäköisyys on tilastotieteen perusta, joka mahdollistaa datan analysoinnin, tulkinnan ja päätöksenteon epävarmuuden vallitessa.

- Se tarjoaa matemaattisen perustan epävarmuuden ja satunnaisuuden käsittelylle.
- Todennäköisyydellä on ratkaisevan tärkeä rooli tilastollisessa päättelyssä eri muodoissaan sekä mm. ennustamisessa ja ylipäätään data-analyysissä.

Todennäköisyys

Todennäköisyys on epävarmuuden muodollinen matemaattinen ilmaus. Olkoon $P(A)$ tapahtuman A todennäköisyys. Todennäköisyyden keskeisiä sääntöjä ovat:

- **Rajat**: $0 \leq P(A) \leq 1$, jossa $P(A) = 0$, jos tapahtuma A on mahdoton, ja $P(A) = 1$, jos A on varma tapahtuma.
  - Komplementti: $P(A) = 1 - P(\text{ei A})$.

- Yhteenlaskusääntö: Jos A ja B ovat toisensa **poissulkevia** tapahtumia, eli vain toinen voi tapahtua, $P(A \text{ tai } B) = P(A) + P(B)$.
- Kertolaskusääntö: Minkä tahansa tapahtumien A ja B osalta, $P(A \text{ ja } B) = P(A|B)P(B)$, jossa $P(A|B)$ edustaa A:n todennäköisyyttä, kun B on tapahtunut.
  - A ja B ovat **riippumattomia**, jos ja vain jos $P(A|B) = P(A)$, eli B:n tapahtuminen ei vaikuta A:n todennäköisyyteen. Tässä erikoistapauksessa siis $P(A \text{ ja } B) = P(A)P(B)$.

Todennäköisyyden käsitteelle voidaan esittää erilaisia tulkintoja, joista kolme tärkeimmät ovat seuraavat (esim. Mellin (2004, s. 240)):

- Todennäköisyyden tulkinta *suhteellisena frekvenssinä*. Tapahtuman tn. voidaan samaistaa tapahtuman suhteelliseen frekvenssiin, kun tarkasteltava satunnaisilmiö esiintyy toistuvasti. Jos tapahtuman suhteellinen frekvenssi lähestyy satunnaisilmiön toistuessa jotain lukua, on tuo luku tapahtuman empiirinen todennäköisyys, mikä puolestaan liittyy tilastotieteen keskeiseen tulokseen eli ns. **suurten lukujen lakiin**.

Esimerkiksi harhattoman kolikon heittämisessä kruunien suhteellinen osuus kaikista heitoista lähestyy arvoa 1/2 eli 50% heittojen lukumäärän kasvaessa.

- Todennäköisyyden tulkinta *subjektiivisena todennäköisyytenä ja vedonlyöntisuhteena*. Ainutkertaisten tapahtumien tn. on henkilökohtainen ja se voidaan määritellä sen vedonlyöntisuhteen avulla, johon henkilö suostuu.

Esimerkki. Oletetaan, että ennen Daavidin ja Goljatin kohtaamista vedonlyöntisuhteeksi ilmoitetaan 1:9. Tämä tarkoittaa vedonlyöjien uskovan Daavidin voittoon tn:llä $1/10 = 10$ % ja Goljatin voittoon $9/10 = 90$ %. Vedonlyöntisuhde on näiden todennäköisyyksien suhde. Sivuuttamalla edelleen vedonvälittäjien osuuden (komission), voidaan näistä päätellä että meille tutumpien kertoimien kautta Goljatin voitosta saisi kertoimen 1.11 ja Daavidin voitosta kertoimen 10.

- Todennäkäisyystulkinta *klassisena todennäköisyytenä*. Tällöin satunnaisilmiö koostuu toisensa poissulkevista symmetrisistä alkeistapahtumista. Klassinen tn. saadaan määräämällä (päättelemällä) tapahtumalle suotuisten alkeistapahtumien lukumäärän suhteena kaikkiin mahdollisten alkeistapahtumien lukumäärään.

Esimerkki. Nopanheitto ja siihen kohdistuva tutkimuskysymys on esimerkki klassisesta todennäköisyydestä. Kuten esim. olkoon määrättävä todennäköisyys

tapahtumalle, että kummallakin nopalla saadaan sama silmäluku. Osoittautuu, että tämän tn. on 1/6 eli 16.7 %.

Todennäköisyyden tarkempi käsittely, kuten edellä mainittujen ominaisuuksien tarkempi tarkastelu tapahtuu myöhemmillä todennäköisyyslaskennan ja tilastotieteen kursseilla.

Tilastolliset mallit ja todennäköisyyslaskenta. Tilastollisten mallien ajatusta esiteltiin jo aiemmin lyhyesti ja tähän palataan tarkemmin vielä myöhemmin. Ne perustuvat todennäköisyyslaskentaan ja niillä mallinnetaan reaalielämän ilmiöiden alla piileviä prosesseja ja mekanismeja. Näiden prosessien tuottamia tietoja (aineistoja) tiivistetään usein graafisiksi esityksiksi ja tunnusluvuiksi sekä lopulta erityisesti tilastollisten mallien parametreiksi, joiden pohjalta johtopäätöksiä tehdään. Tässä onnistuakseen tilastollisten menetelmien tuleekin pyrkiä erottelemaan **sattuma** ja **systemaattisuus** tarkasteltavissa ilmiöissä, tai tarkemmin, niitä kuvaavissa aineistoissa, jotta johtopäätökset olisivat luotettavia.

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

Mitä tilastotiede ei ole

- **Tilastotiede ei ole vain tilastojen tuotantoa** ja/tai **oppia tilastoista ja niiden tekemisestä**
  - 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 tilastoja tuotetaan, jalostetaan ja analysoidaan on keskeinen osa tilastotiedettä. Tilastot ovat siis usein tilastotieteen soveltajan tutkimuskohteena ja tilastojen laadinnassa käytetään apuna tilastotieteen menetelmiä.
    * Suomessa erityisesti Tilastokeskus (https://stat.fi/fi) toimii virallisena tilastoviranomaisena ja tilastotuottajana. Tätä **tilastotuotannon** kokonaisuutta nimitetään ajoittain **tilastotoimeksi**. **Tilastotieteen käyttöalue on paljon tätä laajempi**.
    * Ajoittain käytettävää terminologiaa ja luokittelua: Tilastoala ja tilastotoimi
      · 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 **empiirisen** 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 tietyllä tavalla eksaktia ja sen tulokset perustuvat formaaliin deduktioon ja loogisiin todistuksiin, johtaen useimmiten "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 (tarkasteltavan otoksen perusteella)!

Tilastotiede on siis hyvä nähdä omana tieteenalanaan matemaattisesta esitystavastaan huolimatta. Eihän esimerkiksi myöskään fysiikkaa (sentään) pidetä matematiikan osa-alueena!

Tilastotiede ei ole myöskään tietojenkäsittelytieteen osa-alue, vaikkakin useiden laskennallisten menetelmien ja tehokkaan tietojenkäsittelyn rooli tilastollisissa analyyseissä on jatkuvasti kasvanut. Tietojenkäsittelytiede

- Tietojenkäsittelytieteen teoria ei rakennu tilastotieteen tavoin ajatukselle epävarmoista ja satunnaisista reaalimaailman ilmiöistä.

Vaikka matematiikka ja tietojenkäsittelytiede, 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 premisseineen. Datatiede/data science
- Käsittelemme myöhemmin tilastotieteen roolia matematiikan ja/tai **datatieteiden** (''data science'') kokonaisuudessa ja keskustelemme tarkemmin näiden välisistä eroista.

Mitä tilastotiede (ainakin) on

Tilastotiede yleisenä menetelmätieteenä. Tieteellistä tietoa ympäröivästä maailmasta hankitaan tieteellisillä **menetelmillä/metodeilla** (ks. tieteellisen menetelmän kriteerit), joiden avulla tutkitaan jotain ilmiötä tai sen generoimaa kvantitatiivista mutta epävarmaa tietoa sisältävää aineistoa.

- Tilastotieteessä kehitetyt ja kehitettävät menetelmät antavat tutkijoille yhtenevät ja tiedeyhteisön hyväksymät raamit, jotka mahdollistavat (tilastollisen) päättelyn ja päätöksenteon epävarman tiedon vallitessa. Näin voidaan uskottavasti ja luotettavasti tiivistää tietoa, jota erilaiset aineistot sisältävät, perustaa johtopäätöksiä näille tiivistyksille ja saavuttaa uusia tieteellisiä löytöjä.
- Tilastotieteen menetelmien käyttö ja soveltaminen onkin siis aina alakohtaista. Tästä huolimatta tilastollisia menetelmiä sovelletaan (aina) johonkin **aineistoon**!
- Tilastotiede nähdäänkin usein kuuluvan ns. **menetelmätieteisiin**, joissa mm kehitetään työkaluja muiden tieteiden tutkimusongelmien ratkaisuksi ja jolla on myös oma sovelluksista vapaa teorianmuodostuksensa

Summarisoidaan nyt mitä tilastotiede on:

Tilastotiede kehittää ja soveltaa tilastollisia menetelmiä ja malleja satunnaisilmiöitä kuvaaville kvantitatiivisia tietoja generoiville prosesseille, joiden avulla reaalimaailman ilmiöistä voidaan tehdä johtopäätöksiä ilmiöitä kuvaavien numeeristen tilastotietojen perusteella.

- Näin tilastotietoihin liittyy **epävarmuutta ja satunnaisuutta**.
- Tilastolliset mallit ja menetelmät perustuvat osaltaan todennäköisyyslaskentaan.

Tilastollisten menetelmien avulla pyritään löytämään reaalimaailman satunnaisia ilmiöitä kuvaavista numeerisista (eli kvantitatiivisista) tiedoista **systemaattisia piirteitä** joita jalostetaan sellaiseen muotoon, että ilmiöistä voidaan tehdä päätelmiä. Tässä voidaankin nähdä olevan kysymys **signaalin ja kohinan erottamisesta** (ks. Silver, 2014).

Juuri sattuman ja epävarmuuden huomioiminen tutkimusasetelmissa erottaa tilastotieteen muista menetelmätieteistä!

Tilastollisia menetelmiä voidaan soveltaa tietojen keruun, jalostuksen ja analysoinnin jokaisessa vaiheessa.

- Päämääränä on jalostaa tiedot muotoon, joka mahdollistaa tutkittavaa reaalimaailman ilmiötä koskevien johtopäätösten tekemisen (tilastollisessa päättelyssä) käytettävien menetelmien pohjalta.
- Tutkimuksessa on pystyttävä valitsemaan ja käyttämään menetelmiä, jotka antavat aineistosta vastauksia haluttuihin kysymyksiin. Tämä vaatii yhtä lailla sovellusalakohtaista osaamista, eli substanssiosaamista, kuin myös kattavaa menetelmäosaamista.

Aineisto (data)

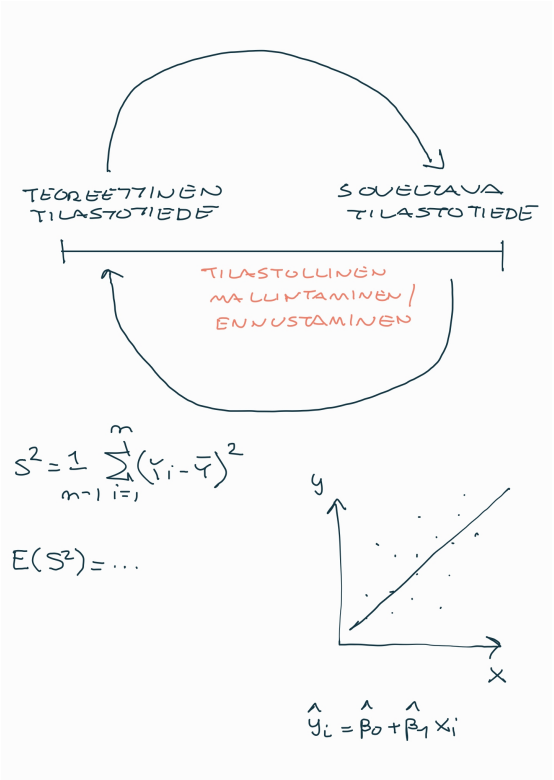
Tilastotiede, jatkoa. Tilastotieteessä lähtökohtana ja ratkaisevassa asemassa on siis aina jonkin satunnaisilmiön generoima **aineisto** (**data**), josta haluamme oppia tai tietää lisää.

Tämä aineistokeskeisyys yhtäältä erottaa tilastotieteen rajatieteistään ja toisaalta tuo sen lähemmäksi niitä ja sovellusalojaan. Aineistoa **analysoidaan**, **kuvaillaan** ja **mallinnetaan** tilastollisin menetelmin, joiden kehittäminen on keskeinen osa tilastotiedettä.

Pelkkä menetelmien kehittäminen kuuluu pitkälti **matemaattisen/teoreettisen tilastotieteen** osa-alueelle (palataan vielä myöhemmin).

- Toisaalta pelkkä aineistoon 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 "**takaisinkytkennän**" **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 tutkijoita (erityisosaajia) asettuvat näiden äärimmäisten luonnehdintojen välimaastoon eikä tarkkaa luokittelua ole sinänsä tarpeen tehdä ja korostaa.
- Joka tapauksessa tilastotieteen, ja sen osa-alueiden, kehityksen keskiössä ovat aina sovellusalakohtaiset ongelmat, joista useat palautuvat yleisemmälle tasolle teoreettisen tilastotieteen kehityspolkuihin.

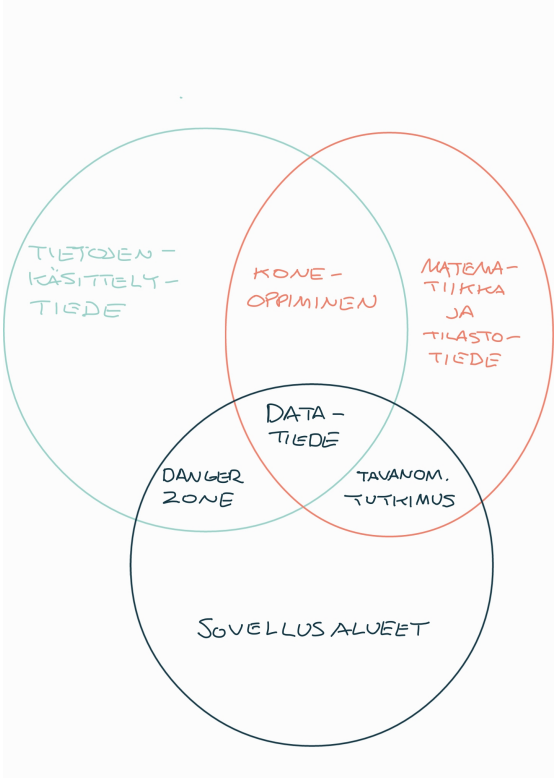


Kuva: Teoreettinen ja soveltava tilastotiede (ja niihin liittyviä joitain kaavoja ja kuvioita)

3.2 Tilastotieteen suhde lähitieteisiin

Oheinen kuvio (lähde: Duchesnay (2020): https://duchesnay.github.io/pystatsml/introduction/machine_learning.html) tarjoaa karkean yleistyksen

tietojenkäsittelytieteen (Computer Science), sovellusalan (Application domain) sekä tilastotieteen (Statistics) ja matematiikan (Mathematics) välisistä yhteyksistä. On selvää että tilastotieteellä on paljon päällekkäisyyksiä lähitieteidensä kanssa ja joskus näkeekin, huolimatta edellä tehdyistä huomioista, että tilastotiede niputetaan yhteen matematiikan tai tietojenkäsittelytieteen kanssa.



Kuva: Tilastotieteen ja rajatieteiden yhteyksiä kuvaava Venn-diagrammi (ks. Duchesnay (2020) yläpuolella).

Yritetään siis hahmotella tilastotieteen suhdetta sitä lähimpänä olevaan (soveltavaan/sovellettuun) **matematiikkaan**.

- Tilastotieteessä olennaisen otantateorian 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 käänteinen, edetään *havainnoista parametreihin*, mutta ongelman matemaattinen kuvaus vaaditaan ensin. Tilastotiede esittää menetelmiä ja käsitteitä tämän käänteisen ongelman ratkaisemiseen.

  - Karkeasti erotellen tilastotieteessä käsiteltävät ongelmat lähtevät aina havainnoista eli aineistosta ja matematiikassa suunta on teoriasta aineistoon.
  - Voidaankin siis sanoa, että tilastotieteen erottaa puhtaasta matematiikasta se, että siinä tutkitaan menetelmiä eli metodeja, jotka mahdollistavat päättelyn/tiedon hankinnan puutteellisesta tai epävarmasta tiedosta.

Matemaattiset ja tilastolliset mallit voidaan jakaa **deterministisiin** ja **stokastisiin** malleihin.

- 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 tietyt **esiintymistodennäköisyydet**.

  Esimerkkejä ovat mm. kolikonheitto tai sään ennustaminen.

- Kun jotain ilmiötä kuvataan stokastisen mallin avulla, voidaan käyttää (joudutaan käyttämään) tilastollisia menetelmiä. Vaikka käytännössä laskenta hoidetaan tietokoneohjelmien avulla, meidän tilastotieteen tutkijoina ja käyttäjinä on huolehdittava tutkimusprosessin onnistuneesta toteutuksesta muilta 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 teoriapohjansa.
  - 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)**.

Datatiede (**data science**). Tieteenala, joka keskittyy tekniikoihin, joilla saadaan muodostettua analyysejä ja käsityksiä analysoitavasta aineistosta (datasta). Tämä sisältää algoritmien rakentamisen ennustamista varten, datan analysoinnin, visualisoinnin ja mallintamisen.

Perinteinen tilastotiede on olennainen osa datatiedettä, mutta siihen luetaan kuuluvaksi myös vahva ohjelmoinnin ja datan hallinnan elementti. Datatiede hyödyntää monipuolisia menetelmiä ja työkaluja, kuten koneoppimista, tekoälyä, suurten datamäärien käsittelyä ja tilastollista analyysiä, jotta voidaan tehdä perusteltuja päätöksiä ja ennusteita.

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** mahdollistaa ja jotka ovat keskeisessä roolissa dataintensiivisissä sovellutuksissa.

- Datavallankumouksella viitataan tässä jatkuvasti kasvaneeseen laskentakapasiteettiin ja datamassojen hallintaan mitä yhä paremmat tietokoneet, ja nyt myös tekoäly eri muodoissaan, mahdollistavat eri sovellusalueilla.

Koneoppiminen

Yksi edeltävän kuvan alueista liittyy **koneoppimiseen**.

Koneoppimiseen (machine learning) luetaan menetelmiä ja algoritmeja, jotka liittyvät esim. luokittelun, ennustamisen tai klusteroinnin, saavutttamiseksi mahdollisesti monimutkaista aineistoa analysoitaessa. Koneoppiminen hyödyntää ja korostaa laskennallisia menetelmiä ja algoritmeja, joiden avulla tietokoneet voivat oppia ja tehdä päätöksiä mm. ilman tarkemman tilastollisen analyysimallin rakentamista.

Koneoppimiseen luettavia menetelmiä käytetään monenlaisiin sovelluksiin, kuten kuvantunnistukseen, luonnollisen kielen käsittelyyn, suositusjärjestelmiin ja moniin muihin datan analysointitehtäviin

Datatieteeseen voidaan katsoa lukeutuvan piirteitä mm. seuraavilta aloilta:

- Tilastotiede ja matematiikka
  - Erityisesti tilastollinen data-analytiikka ja satunnaismekanismien perusteella saatujen aineistojen mallintaminen sekä soveltuvat soveltavan/sovelletun matematiikan osa-alueet.

- Tietojenkäsittely
  - Tietoteknologian kehityksen myötä taitavien tietojenkäsitteljöiden kysyntä on kasvanut merkittävästi. Lähes jokaisella alalla kerätään entistä enemmän dataa lähes kaikesta, ja jonkun pitäisi osata myös käsitellä näitä aineistoja!
  - Datatieteen voidaankin osaltaan katsoa syntyneen tästä elinkeinoelämän tarpeesta asiantuntijoille, jotka osaavat käsitellä ja hallita suuria tietoaineistoja (dataa) sekä mallintaa niitä hyödyllisellä tavalla. Nämä piirteet korostavat tietojenkäsittelytiedettä.
- Sovellusala(t)
  - Datatiede on luonteeltaan pääosin soveltavaa ja sen alaan lukeutuvia menetelmiä sovelletaan aina johonkin tosielämän ongelmaan. Tästä syystä substanssiosaaminen tarkasteltavalta 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 (ellei olennaisesti aivan samoja), sillä ne pohjautuvat edelleen tilastotieteen ja matematiikan teoriapohjaan. Ilman jälkimmäisten riittävää osaamista, liikutaan datatieteen osalta vaarallisilla vesillä! Ks. oheisen kuvan **danger zone** ja keskustelu alla.

“Danger zone”. Oheisen Duchesnayn (2020) kuvan “danger zone” kuvaa tilannetta, jossa ilmiöiden ja toisaalta myös mallien/menetelmien, tilastotieteellinen perusta unohdetaan.

Tilastotieteen näkökulman ohittava (laiminlyövä) soveltaja ei aina kykene suhtautumaan kriittisesti muodostuvaa ennustemallia, tai ennustetulosta, kohtaan eikä täten päädy parhaisiin mahdollisiin (tarkimpiin) ennustetuloksiin tilanteessa, jossa jokin toinen malli kuvaisi ilmiötä paremmin.

- Ko. soveltaja ottaa mallin sekä sen antaman ennustetuloksen annettuna, eikä mieti *mistä kyseinen ennustetulos johtuu*. Jotta tarkat ennustetulokset toteutuvat jatkossakin (kun uutta aineistoa, dataa, tulee saataville), on ennustajan oleellista huomioida mitkä tekijät johtivat tarkkaan ennustetulokseen.
- Eri menetelmät sopivat eri sovelluskohteisiin. Tilastotieteilijä osaa useimmiten tunnistaa eri sovelluskohteisiin sopivat menetelmät paremmin kuin tietojenkäsittelijä. Vastaavasti tehokkaan/onnistuneen ohjelmointikoodin kirjoittamisessa tilanne on usein toisinpäin.
  - Ajoittain sovelluskohteet tai ennusteongelmat ovat riippuvuusrakenteiltaan niin monimutkaisia, että kovin vahvoja tulkintoja ei ole edes mahdollista muodostaa!

Hyvä toki korostaa vielä, että tilastotieteilijä voi myös erehtyä analyyseissään!

3.3 Tilastotieteen osa-alueet

Tilastotiede on saanut alkunsa siitä, että yhteiskunnan modernisoituessa 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 ongelmiin.
- Tämä on johtanut osaltaan siihen, että tilastotiede jakautuu moniin osa-alueisiin. Osa-alueita on niin paljon, että alan huiputkaan eivät voi hallita niitä kaikkia!

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 tutkimus

Kvantitatiivinen eli **määrällinen tutkimus** on tutkimusta, jossa tutkimusongelma on muotoiltu tarkasti etukäteen ja tutkimuskysymyksiin vastataan käyttäen tilastollisia menetelmiä pyrkien **mallintamaan**, **selittämään** ja/tai **ennustamaan** tutkimuksen kohteena olevaa ilmiötä.

- Täsmällisten ja laskennallisten tilastollisten menetelmien käyttäminen numeeriseen aineistoon on kvantitatiiviselle tutkimukselle ominaista.
- Perustuu yleensä satunnaisotokseen ja tutkimusaineisto on tiivistetty numeeriseksi havaintomatriisiksi, jolle oleellinen vaatimus on sen totuudellisuus.

Kritiikki: määrällinen tutkimus on (paikoin) sokea tutkittavien ilmiöiden sellaiselle luonteelle, jota ei pystytä kvantifioimaan, eli muuntamaan numeeriseen

muotoon. Näihin voidaan katsoa lukeutuvan mm. tunteet, merkitykset ja kokemukset, ellei tutkija keksi niiden numeeriselle mittaamiselle uskottavaa keinoa.

Kvalitatiivinen tutkimus

Kvalitatiivinen eli **laadullinen tutkimus** on tutkimusta, jossa tutkimuksen kohteena olevaa ilmiötä ja sen merkitystä sekä tarkoitusta pyritään **ymmärtämään** kokonaisvaltaisella tavalla.

- Laadullisessa tutkimuksessa annetaan usein tilaa tutkimuksen kohteena olevien ilmiöiden ja/tai ihmisten näkökulmille, vaikuttimille, kokemuksille ja tuntemuksille. Tutkimusyksikköjen otanta on täten usein harkinnanvaraista.
- Laadullisessa tutkimuksessa tutkimusongelma muotoutuu (voi muotoutua) tutkimuksen edetessä ja sille tyypillistä on hypoteesittomuus, eli tutkimus on tarkoitus aloittaa mahdollisimman vähin ennakko-oletuksin. Ennakko-oletuksista on kuitenkin mahdotonta täysin irtautua, joten niiden ilmi tuominen esioletuksina tai "tutkimushypoteeseina" eli arvauksina tuloksista on osa tutkimusta.

Kritiikkiä: laadullinen tutkimus ei pysty vastaamaan kysymykseen miksi, sillä ilman määrällisiä (numeerisia) aineistoja ei ilmiöiden välisiä riippuvuuksia kyetä tutkimaan:

- **Laadullisessa tutkimuksessa menetetäänkin mahdollisuus tutkia ilmiöiden todellisia syitä**.
- Laadullinen tutkimus nähdään usein vähemmän objektiivisena ja sen otosta koskevia tuloksia ei useinkaan voida yleistää koskemaan perusjoukkoa.

Yleisenä menetelmätieteenä tilastotiedettä voidaan (ja myös pitäisi) soveltaa kaikilla reaalimaailmaa 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):

- **kuvailla ja tiivistää tietoa**, jota havaittu aineisto sisältää
- sovellusalan oman **teorian empiirinen testaus** tai
- edellisten pohjalta tehtävä **tilastollinen päättely**.

Deskriptiivisellä eli kuvailevalla tilastotieteellä tarkoitetaan sellaisten menetelmien soveltamista, joiden avulla havaintoaineistosta 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 perustuvaa päätöksentekoa, jossa kvantitatiiviseen aineistoon kuuluva epävarmuus ja satunnaisuus on otettu huomioon.

- Keskeinen tilastollisen päättelyn käyttötarkoitus soveltajille on usein **teorian ja siihen liitettävien hypoteesien testaaminen**, joka voi johtaa joko teorian vahvistumiseen (*verifiointiin*) tai sen vääräksi osoittamiseen (*falsifioimiseen*).
  - On myös syytä muistaa, että yksi tutkimus ei vielä osoita teoriaa oikeaksi tai vääräksi vaan siihen tarvitaan useita tutkimuksia sekä erilaisia tutkimusasetelmia ja -menetelmiä.
  - Kuvaileva tilastotiede ja tilastollinen päättely kulkevat soveltavassa tilastollisessa tutkimuksessa käsi kädessä.

Teoreettinen tilastotiede

Teoreettinen tilastotiede kehittää (tilasto)matemaattisia malleja kuvaamaan satunnaisilmiöitä- ja prosesseja, jotka generoivat reaalimaailman ilmiöitä 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!

Tilastolliset mallit perustuvat siis todennäköisyyslaskentaan, ja niitä kutsutaan tilastollisiksi malleiksi, tai ajoittain **stokastisiksi malleiksi** tai **todennäköisyysmalleiksi**.

Kaikki mallit ovat vääriä, mutta jotkut ovat käyttökelpoisia. (Box, 1976).

Todennäköisyyslaskenta luo tilastotieteelliselle epävarmuuden mallintamiselle vahvan ja uskottavan matemaattisen perustan. Vastaavasti tilastolliset mallit perustuvat laajalti niin kutsuttuun uskottavuusfunktioon, mikä vastaavasti liittyy tn-laskentaan. Se on malli, joka riippuu havaintoaineiston lisäksi yhdestä tai useammasta parametrista.

- Uskottavuusfunktion arvo kertoo kuinka todennäköisenä havaittua aineistoa voidaan pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista

jollain parametriarvoilla. Ts. uskottavuuspäättelyn perusajatuksena on, että se tai ne parametriarvot (tehdyillä til. mallia koskevilla oletuksilla), joilla uskottavuusfunktion arvo maksimoituu kuvaa aineiston generoinutta prosessia parhaiten.

- Aineistoa koskevia hypoteeseja voidaan testata käyttäen uskottavuusfunktiota maksimia vastaavaa tilastollista mallia!

Uskottavuusfunktiot perustuvat aina satunnaisilmiöiden mahdollisia arvoja kuvaaviin nk. **tiheysfunktioihin** ja diskreettien sm:jien tapauksessa (tähän palataan myöhemmin) **pistetodennäköisyysfunktioihin**.

- Nämä funktiot kuvaavat jonkin satunnaismuuttujan (satunnaisilmiön) saamien arvojen **jakaumaa**.
- Esimerkiksi kolikonheitto on satunnaisilmiö ja sillä on vain kaksi arvoa (kolikon kantilleen jäämistä ei tässä lasketa mahdolliseksi tapahtumaksi) ja kolikonheittoa voidaan kuvata nk. binomijakaumalla, jota merkitään (Bin(n,p)), jossa (n) on heittojen lukumäärä ja (p) on kruunan todennäköisyys

Esimerkki. 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 noudataa binomijakaumaa (Bin(40, 0.5))? (Binomijakauma esitellään tarkemmin tulevissa luvuissa)
- Toisin sanoen, kuinka uskottavana voidaan pitää sitä että kyseinen kolikko on tavallinen, painottamaton kolikko?

3.4 Tilastotieteen sovellusaloja ja “rajatieteitä”

Yleisenä menetelmätieteenä tilastotiedettä sovelletaan useilla eri tieteenaloilla. Jokaisella sovellusalalla on kuitenkin oma erillinen teoriapohjansa sekä empiiriset käytänteet, joten **substanssitietous** on sovellettaessa erityisen tärkeää.

- Huolimatta vaihtelevista empiirisistä käytännöistä sovellusmenetelmän taustalla on (lähes aina) kuitenkin tilastotieteen alalla kehitetty menetelmä.

- Sovellusaloilla ongelmanratkaisussa yhdistetäänkin metodiseen osaamiseen välttämättä myös substanssitietoutta. Tämän myötä soveltavan tilastollisen tutkimuksen kenttä on laaja ja rikas.

Osa näistä sovelluskentistä on kehittynyt vahvassa yhteisvaikutuksessa tilastotieteen ja lähitieteiden, kuten kasvavissa määrin erityisesti koneoppimisen, yhteydessä. Usein on pystyttävä arvioimaan ongelmanasettelun ja tulosten tarkoituksenmukaisuutta ja pyrkiä välttymään siltä että tutkijan tieteelliset ja yhteisölliset sitoumukset heijastuisivat tutkimuksen kulkuun.

- Tilastotieteen pääaineopiskelun kannalta substanssitietous saavutetaan sivuaineopintojen perusteella. Vastaavasti toisinpäin muiden aineiden pääaineopiskelijoiden kohdalla tilastotiede 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 empiirisen tutkimuksen havaintoaineisto on tilastollisen tutkimuksen mahdollinen kohde. Esim. kokeellinen tutkimus käyttää apunaan tilastollisia menetelmiä.

Koska tilastotieteellä on sovelluksensa miltei kaikilta tieteenhaaroilla, on siis syntynyt joukko nk. **rajatieteitä**:

- Sovellusaloja, joilla tilastotieteen soveltaminen on muodostunut omaksi tutkimuskohteekseen/tieteenlajikseen (ks. linkit):
  - Psykologia: psykometriikka
  - Sosiaalitieteet: sosiometria
  - Taloustiede: ekonometria
  - Kemia: kemometria
  - Bio- ja lääketiede: biometria
  - Epidemiologia
- Soveltavan (sovelletun) matematiikan tutkimusaloja, jotka ovat osaltaan päällekkäisiä tilastotieteen kanssa
  - Informaatioteoria
  - Matemaattinen tilastotiede
  - Todennäköisyyslaskenta
  - Operaatioanalyysi
- Tietojenkäsittelytieteen alaan (osittain) lukeutuvia tutkimusaloja

- Laskennalliset menetelmät
- Data mining
- Knowledge discovery
- Hahmontunnistus
- Tekoäly
- Koneoppiminen

Ja paljon muita!

Chapter 4

Sattuma ja satunnaisuus tilastotieteessä

Tässä luvussa pohdimme sattuman ja satunnaisuuden roolia yleisesti, mutta erityisesti tilastotieteessä ja tieteessä ja millä tavalla satunnaisuus liittyy tilastoaineistojen muodostumiseen ja siten tilastotieteeseen.

- **Satunnaisuudella** tarkoitetaan yleensä säännönmukaisuuden ja (täydellisen) ennustettavuuden puuttumista, ja kenties juuri siksi sitä voidaan pitää **yhtenä maailman vaikuttavimmista ilmiöistä**.

Jokainen haluaisi tietää *mitä tuleman pitää* ja siksi sattuma tekee elämästä mielenkiintoista! Se muokkaa niin meitä itseämme kuin meitä ympäröivää maailmaa mitä merkityksellisimmin tavoin - joskus jopa vasten tahtoamme ja usein vailla täyttä ymmärtystämme!

- Ihmisen oma kokemus on kuitenkin altis kaikenlaisille virhepäätelmille, joita kutsutaan myös **kognitiivisiksi vinoumiksi**.
- Haluamme löytää systematiikkaa ja tarkoitusta kaaoksesta sekä merkityksiä ja syy-seuraussuhteita sellaisista tapahtumista, jotka kuuluvat normaalin satunnaisen vaihtelun piiriin (vrt. signaali vs. kohina). Tällaisissa tilanteissa usein tilastollinen tarkastelu paljastaakin ilmiön todellisen, alkuperäisestä kuvitelmasta poikkeavan luonteen.
- Erottaakseen systemaattisen vaihtelun satunnaisesta ja ymmärtääkseen oikeasti merkityksellisiä syy-seuraussuhteita, satunnaisuutta on välttämätöntä ymmärtää. Tämä välttämättömyys pätee erityisesti tiedeyhteisön jäseniin, jotka pyrkivät tutkimaan ympäröivän maailman satunnaisia ilmiöitä.

Tilastotiede perustuu satunnaisilmiöiden ja niiden generoimien aineistojen tutkimiseen, joten sattuman luonteen ymmärtäminen on keskeisessä roolissa niin tilastotieteen kuin muidenkin tieteiden ja lopulta maailman ymmärtämisessä.

4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä

Tilastolliset muuttujat tulkitaan satunnaisiksi, ja tilastollisen tutkimuksen tavoite onkin siis **tutkia sitä satunnaisilmiötä, joka havaitut eli toteutuneet havaintoarvot on generoinut**.

Olemme jo edellä todenneet:

- Yksi tilastotieteen olennainen tehtävä onkin kehittää **tilastollisia malleja**, joiden avulla tutkimuksen kohteena olevaa satunnaisilmiöitä voidaan kuvata, selittää ja ennustaa.
- Tilastollisen mallin satunnaisten piirteiden kuvaus perustuu **todennäköisyyslaskentaan**.

Satunnaisilmiö

Satunnaisilmiö. Reaalimaailman ilmiö on satunnaisilmiö, jos seuraavat ehdot pätevät:

- Ilmiöllä on useita erilaisia tulosvaihtoehtoja.
- Sattuma määrää mikä tulosvaihtoehdoista toteutuu, eli yksittäistä tulosta ei voida tietää etukäteen.
- Vaikka tulos vaihtelee ilmiön toistuessa satunnaisesti, käyttäytyy tulosvaihtoehtojen suhteellisten osuuksien jakauma tilastollisesti stabiilisti ilmiön toistokertojen lukumäärän kasvaessa.

Tilastollisella stabiiliudella tarkoitetaan sitä, että on mahdollista arvioida kuinka **todennäköisiä** erilaiset tapahtumat eli satunnaisilmiön tulosvaihtoehdot ovat.

- Toisin sanoen satunnaisilmiön tulosvaihtoehtoihin liittyy säännönmukaisuutta, mikä tulee esille ilmiön toistuessa.
- Tämä liittyy myös siihen, että (useimmiten) satunnaisilmiön lopputulos ei ole täysin ennustamaton. Ts. saatamme kyetä (ajoittain) ennustamaan/ennakoimaan lopputulemaa, mutta täyttä varmuutta lopputulemasta ei siis ole.

Esimerkkejä satunnaisilmiöistä:

- Tyypillinen esimerkki on uhkapelit, kuten kortti- ja noppapelit, arpajaiset, lotto tai ruletti: näitä käytetäänkin usein todennäköisyyslaskennan peruskursseilla satunnaisilmiöiden esittelyyn.

  - Huom: Osakesijoittaminen, urheiluvedonlyönti tai esim. pokeri eivät kuulu uhkapeleihin, jos niitä harjoitetaan systemaattisesti ja ammattimaisesti (ml. asianmukainen riskienhallinta).

- Lukion biologian tunneilta muistetaan, että perinnöllisyyskin on osaltaan sattumaa: se määrää 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, ihmisryhmien ja ihmisten muodostamien organisaatioiden sisäisessä ja välisessä käyttäytymisessä on useita satunnaisia elementtejä.

- Jopa hyvin deterministiseen toimintaperiaatteeseen tähtäävässä 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 perusluonteeltaan satunnaisia.

Tilastollista vaihtelua ilmentävät tilastolliset muuttujat tulkitaan **satunnaismuuttujiksi** ja havainnot (havaintoarvot) voidaan näin ollen tulkita näiden satunnaismuuttujien **realisoituneiksi arvoiksi**. Tällöin tilastollisen tutkimuksen kohteena on **nämä havainnot generoinut satunnaisilmiö**. Satunnaismuuttuja

Satunnaismuuttuja (usein lyhyesti sm., englanniksi *random variable*, ja 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 (tilastoyksiköiden tutkimusmuuttujien) realisoituneista arvoista.
- Realisoituneiden arvojen vaihtelua tilastoyksiköiden välillä kutsutaan satunnaisvaihteluksi ja tätä vaihtelua kuvataan satunnaismuuttujan todennäköisyysjakaumalla.

Satunnaismuuttuja siis kuvaa tarkasteltavan mitattavan ominaisuuden (satunnais)vaihtelua tutkimuksen kohteiden eli tilastoyksiköiden joukossa.

- Mitattavan ominaisuuden mahdolliset arvot määräävät satunnaismuuttujan luonteen. Yleisesti satunnaismuuttujat jaetaan kahteen luokkaan: **jatkuviin** ja **diskreetteihin**.
  - Tähän jakoon voidaan liittää myös jako **kvantitatiivisiin** ja **kvalitatiivisiin** satunnaismuuttujiin. Myöhemmin esiteltäviin mittaasteikkoihin liittyen kvalitatiiviset sm:jat liittyvät luokittelu- tai järjestysasteikkoon ja vastaavasti kvantitatiiviset sm:jat välimatka- ja suhdeasteikkoon.
  - Huom. ajoittain kvantitatiivisten ja kvalitatiivisten muuttujien jako tehdään niin, että myös diskreetit sm:jat luetaan kvantitatiivisiksi muuttujiksi.
- Satunnaismuuttujan **todennäköisyysjakauma** määrää erilaisten tulosvaihtoehtojen todennäköisyydet ja mahdollistaa täten tilastollisen päättelyn. Satunnaisuus eroaa mielivaltaisesta prosessista siinä, että satunnaista ilmiötä voidaan kuvata jollakin **tilastollisella lailla/mallilla** kun taas mielivaltaista prosessia ei.

Jatkuvat ja diskreetit sm:jat

Jatkuvat ja diskreetit satunnaismuuttujat

- Satunnaismuuttuja (Y) on **jatkuva**, jos se voi saada ylinumeroituvan määrän arvoja tai ts. minkä tahansa arvon joltain väliltä, kuten tyypillisesti minkä tahansa arvon joltain reaalilukuväliltä $(-\infty, \infty)$.
- Satunnaismuuttuja (Y) on **diskreetti**, jos se voi saada vain joitain mahdollisia arvoja (vain yksittäisiä, äärellisen tai numeroituvasti äärettömän määrän, arvoja). Yksinkertaisimmillaan diskreetti satunnaismuuttuja (Y) on kaksiarvoinen eli binäärinen, jolloin sen mahdollisia arvoja tyypillisesti merkitään (y=0) ja (y=1).

Todetaan lyhyesti vielä diskreetteihin sm:jiin liittyen, että mahdollisia luokkia voi olla myös enemmän kuin kaksi. Tällöin ko. *kategorinen* vastemuuttuja voi olla sellainen, että kategoriat voidaan järkevästi järjestää järjestykseen

Esimerkkejä:

- Järjestämättömät luokat: esim. henkilön kotimaa, auton väri tai sairaala, jossa lääketieteellinen operaatio tapahtuu.

- Järjestetyt luokat: esim. sotilashenkilöiden arvo.

Esimerkki: sademäärä satunnaismuuttujana. Huomista sadantaa eli sademäärää tietyllä alueella voidaan pitää (ennen huomista sadepäivää) satunnaismuuttujana (Y) ja mitattua sademäärää (sadepäivän jälkeen) täten sademäärän yhtenä realisaationa (y).

- Yleensä sademäärää kohdellaan jatkuvana muuttujana millilitroissa. Mikäli kuitenkin määritetään toteumaksi jonkin sademäärän kynnysarvon, esimerkiksi yksi milli, ylittävä sademäärä, on kyseessä diskreetti kaksiarvoinen (binäärinen) muuttuja (sademäärä on joko yli tai alle yksi milli).

- Tutkija voisi olla esimerkiksi kiinnostunut sadannan eroista kuntien tai alueiden välillä tietyssä maasssa sekä tähän kenties vaikuttavista syistä. Tätä tutkiakseen hän tarvitsisi tilastollisen aineiston tutkimus- ja taustamuuttujineen eri alueilta, joka voitaisiin kerätä esimerkiksi satunnaisotannalla (palataan myöhemmin) asettamalla sadannan mittaavia astioita satunnaisesti kyseisille alueille.

Huomioita: Ajoittain tietyn suureen/ilmiön mallinnuksessa voidaan perustellusti käyttää näkökulmasta riippuen kumpaan vaan luokkaan (diskreetit ja jatkuvat sm:jat) kuuluvaa tilastollista mallityyppiä.

Esimerkkejä. Suomen COVID19-tartuntatapauksia tutkittaessa, tartunnan saaneiden lukumäärä oli periaatteessa diskreetti satunnaismuuttuja, joka sai yksittäisen (kokonaisluku)arvon joka kuukausi, mutta käytännössä lukumäärät ovat tässä tapauksessa sen verran suuria, että niitä on perusteltua kohdalla jatkuva-arvoisena muuttujana.

Vastaavasti esimerkiksi potilaan jonotusaika päivystyksessä voi periaatteessa saada minuuttitasolla hyvinkin diskreettejä arvoja, mutta toisaalta myös minkä tahansa arvon tietyltä reaalilukuväliltä, kuten $[0, \infty)$, ts. mikä vain positiivinen arvo) aikayksikköä muutettaessa ja tällöin käytettäisiin jatkuviin sm:jiin perustuvia tilastollisia menetelmiä.

4.2 Satunnaisilmiöiden tilastollisen mallintamisen perusteita

Seuraavaksi käydään vaiheittain läpi (yleisellä tasolla) todennäköisyyden (todennäköisyyslaskennan) näkökulmasta sitä, miten reaalimaailman satunnaisilmiöitä voidaan tilastollisin menetelmin mallintaa. Kaikkiin näihin vaiheisiin syvennytään tulevissa jaksoissa vielä tarkemmin, mutta tämän osion tarkoituksena on havainnollistaa sitä, miten tilastotieteessä mallinnetaan satunnaisilmiöitä.

Jo aiemmin on todettu, että **tilastollisen tutkimuksen lähtökohta on aineisto** eli **data**. Aineiston tulee kuvata tutkittavaa satunnaisilmiötä sillä tavalla, että esitettyihin tutkimuskysymyksiin voidaan vastata ja/tai hypoteeseja testata.

- Aineiston havaittujen arvojen taustalla olevat satunnaismuuttujat määräävät käytettävän tilastollisen mallin, joka kuvaa tutkittavan ilmiön satunnaista luonnetta. Lopulta tätä tilastollista mallia voidaan käyttää tilastollisten analyysien tekemiseen, kuten ennusteiden muodostamiseen ja/tai hypoteesien testaamiseen.

Kerätyn (tai havaitun) aineiston pohjalta pyritään tekemään tutkimuskysymystä vastaavia päätelmiä tutkimuksen kohteena olevasta ja aineiston generoineesta satunnaisilmiöstä.

Tilastotieteessä tilastollisen tutkimusaineiston muodostumista voidaan pitää esimerkkinä satunnaisilmiöstä. Voimme ajatella (tässä kohtaa kurssia ja opintoja), että tilastollisen tutkimuksen kohteet on valittu, tavalla taikka toisella, **arpomalla**.

- Arvonta on mainio esimerkki satunnaisilmiöstä, sillä siihen liittyy aina ennustamattomuutta: vaikka yksittäisen arvonnan tulosta ei voi tietää etukäteen, noudattaa se kuitenkin todennäköisyyden lakeja.
- Koska arvonnan tulos vaihtelee satunnaisesti arvontakerrasta toiseen, myös tutkimuksen kohteita kuvaavat tiedot vaihtelevat satunnaisesti arvontakerrasta toiseen.

Tilastollisten aineistojen kerääminen arvontaa hyödyntäen

Satunnaisotanta: Otannalla tarkoitetaan laveasti tutkimusaineistojen keräämisen menetelmiä. Erilaisten virhelähteiden kontrolloimiseksi tutkimuksen kohteet on syytä valita arpomalla.

Esimerkiksi tutkittaessa peruskouluopetusta, kaikkien koulujen tutkiminen olisi liian työlästä ja kallista. Tällöin sovelletaan erilaisia otantamenetelmiä, jotka

varmistavat että tutkimukseen valikoituu satunnaisesti edustava otos Suomen kouluista.

Satunnaistetut kokeet: Kokeellisessa tutkimuksessa tavoitteena on vertailla erilaisten käsittelyiden vaikutuksia kokeen kohteisiin. Erilaisten virhelähteiden kontrolloimiseksi käsittelyt on syytä arpoa kohteille.

Esimerkiksi lääketutkimuksessa on tärkeää, että testattavan lääkeaineen vaikutusta tutkittaessa testattavaksi henkilöiksi ei valikoidu esimerkiksi juuri tietyn ikäisiä ihmisiä. Lääkeaine saattaa vaikuttaa eri tavoin eri ikäisiin koehenkilöihin, joten satunnaisotanta eri ikäisistä on tarpeen tutkimusta tehdessä.

Tilastollisen aineiston keräämistä, eli otantateoriaa, ja sen tavoitteita käsitellään myöhemmin tämän materiaalin kuluessa.

Satunnaisotannalla kerätyissä aineistoissa eli **otoksissa** satunnaisuus perustuu siihen, että satunnaismuuttujien toteutuvat arvot (ja niistä lasketut tunnusluvut, kuten keskiarvo, joita tarkastellaan hetken kuluttua) vaihtelevat satunnaisesti otoksesta toiseen, koska otokseen kuuluvat havaintoarvot vaihtelevat otoksesta toiseen.

- Tämä vaihtelu on kuitenkin tilastollisesti stabiilia (ks. yllä), eli eri otokset ilmentävät samaa satunnaisilmiötä. Ts. otantaa toistettaessa eri aineistojen sisältämä satunnaisvaihtelu on aineiston generoineen todennäköisyysjakauman mukaista myös otosten välillä.
- Tilastotieteen tehtävä on tuottaa **tilastollisia malleja** tälle satunnaisilmiöissä havaittavalle **tilastolliselle stabiliteetille**. **Tämä säännönmukaisuus on siis tilastollisen tutkimuksen kohde.**

4.3 Havaintoaineisto eli data

Tilastotieteellinen tutkimus tarkastelee reaalimaailman ilmiöitä. Täten tutkimuskohteena on tavallisessa elämässä tavattavia asioita, ihmisiä tai tapahtumia.

Tilastollinen tutkimus aloitetaan tutkimusaineiston keruun suunnittelulla.

- ks. “PPADC/OSAAT-sykli” luvussa 2

Edellä esitellyt satunnaismuuttujat liittyvät hyvin olennaisesti lopulta analysoitavan tutkimusaineiston muodostumiseen. Tilastotieteessä nimittäin ajatellaan, että satunnaismuuttujien reaalisaatioiden kerätty joukko muodostaa lopulta tarkasteltavan **havaintoaineiston** eli **datan**.

- Kuten jo aiemmin määrittelimmme, tutkimuskohteita kutsutaan tilastoyksiköiksi, joilta tarkasteltaan erilaisia mitattavia tilastollisia muuttujia.
- Kun tarkasteltavien tilastoyksikön tilastollisten muuttujien (numeeriset) arvot havaitaan, kutsutaan näiden arvojen joukkoa **havainnoksi**. Havainto/havainnot

Havainto muodostuu tilastoyksikön tarkasteltavien tilastollisten muuttujien havaitusta arvoista.

Lopulta havainnot muodostavat havaintoaineiston eli datan. Data

Havaintoaineisto eli data on tilastoyksiköiden tilastollisista muuttujista kerätty havaintojen joukko. See koostuu populaation (tyypillisesti sen osajoukon) tilastoyksiköiden havaituista tilastomuuttujien arvoista eli havainnoista.

Havaintoaineisto voidaan koota taulukoksi, johon listataan tilastoyksiköt riveille ja tilastomuuttujat sarakkeisiin. Jos havaintoaineisto koostuu n:stä tilastoyksiköstä, joista jokaisesta on kerätty esim. (m):stä tilastomuuttujasta havainnot, niin aineisto voidaan esittää taulukon muodossa seuraavasti:

	tilastomuuttuja 1	tilastomuuttuja 2	…	tilastomuuttuja (m)
tilastoyksikkö 1	(y_{1,1})	(y_{1,2})	…	(y_{1,m})
tilastoyksikkö 2	(y_{2,1})	(y_{2,2})	…	(y_{2,m})
( )	( )	( )		( )
tilastoyksikkö (n)	(y_{n,1})	(y_{n,2})	…	(y_{n,m})

Tässä siis rivillä (i) on (i). **tilastoyksikön** havainto ja sarakkeessa (j) on (j):tä tilastollisesta muuttujasta havaittu arvo (y_{i,j}). Ts. yhdellä rivillä on yhden tilastoyksikön tiedot kaikista tilastomuuttujista ja yksi sarake on kaikkien tilastoyksiköiden tiedot yhdestä tilastomuuttujasta.

Esimerkki. Empiirisenä esimerkkinä ylläolevasta taulukkomuotoisesta aineistosta on isän ja pojan pituuksista koostuva aineisto (tätä tarkastellaan vielä tarkemmin myöhemmin), jossa on havaintoja 1078 kappaletta (ts. $n = 1078$). Tämän aineiston havainnot merkitsevät oheista taulukkoa

Havaintopari (isä_i, poika_i)	Isän pituus (cm)	Pojan pituus (cm)
(isä_1, poika_1)	165.1	151.9

Havaintopari (isä_i, poika_i)	Isän pituus (cm)	Pojan pituus (cm)
(isä_2, poika_2)	160.8	160.5
( )	( )	( )
(isä_1078, poika_1078)	178.6	170.2

Usein, varsinkin parhaillaan kiihtyvällä vauhdilla, kerättävät havaintoaineistot ovat niin suuria, ettei edellisenkaltaisesta havaintotaulukosta voida usein suoraan tarkastelemalla nähdä aineiston pääpiirteitä. Tällöin voi olla tarpeen **luokitella aineistoa** taulukon muodostamiseksi.

- Luokittelussa on kysymys aineiston tiivistämisestä kohtuullisen kokoiseksi 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 tapana luokitella ikäjakauman kuvaamisessa 10-vuotisluokkiin (15–24, 25–34,…), vaikka periaatteessa ikä voitaisiin ilmoittaa minuutinkin tarkkuudella.
  - Luokkien lukumäärään vaikuttavat muun muassa tilastomuuttujan arvojen vaihteluväli ja havaintoaineiston laajuus.
- Luokittelussa pyritään siihen, että luokkien lukumäärä saadaan tarvittaessa luokkia yhdistämällä kohtuulliseksi ja että luokat valitaan tasavälisesti eli siten, että kahden peräkkäisen luokan alarajojen erotus on vakio.
  - Kun aineistoa luokitellaan, aineiston luettavuus paranee mutta toisaalta osa tiedoista (informaatiosta) menetetään eivätkä yksittäiset havaintoarvot ole enää tiedossa.

Kvantitatiivisen tutkimuksen aineistoksi kelpaa periaatteessa kaikki havaintoihin perustuva informaatio, joka on **mittauksen** avulla muutettavissa numeeriseen muotoon.

- Kaikki havaitut tilastolliset muuttujat eivät ole aina mielenkiintoisia. Tutkimuksen kannalta mielenkiintoisia muuttujia kutsutaan **tutkimusmuuttujiksi**, joiden lisäksi havaintoaineisto pitää mahdollisesti sisällään **taustamuuttujia**.

Esimerkiksi, jos tutkimuksella halutaan tietoa suomalaisen aikuisväestön mielipiteistä, havaintoyksikköinä ovat aikuisväestöön kuuluvat henkilöt. Jos halutaan tietoa suomalaisista kunnista, havaintoyksikköinä ovat Suomen kunnat jne.

- Ensimmäisessä tapauksessa tilastollisina muuttujina on aikuisväestön mielipiteet, joita voidaan selvittää esimerkiksi kyselytutkimuksella. Toisaalta voidaan myös kerätä taustamuuttujiksi haastatelluista muita tietoja, kuten asuinpaikka, ikä ja ammatti.

Kaikkia mielenkiintoisia muuttujia ei kuitenkaan välttämättä voida havaita, eli niille ei voida määrittää numeerista arvoa. Tällöin puhutaan nk. **latenteista muuttujista**, eli muuttujista joita ei suoraan havaita mutta joiden oletetaan vaikuttavan havaittavien muuttujien taustalla.

- Latentteja muuttujia ovat esimerkiksi elämänlaatu, onnellisuus, konservatiivisuus, yms.
- Latentteja muuttujia voidaan rakentaa tilastollisten mallien avulla käyttäen hyödyksi niihin liittyviä havaittuja muuttujia.

Myöhemmin tässä materiaalissa palataan vielä eri yhteyksissä muutamiin perustason graafisiin esitystapoihin miten tilastotaineistoja voidaan havainnollistaa. Graafiset menetelmät ovat erittäin tärkeä osa aineiston havainnollistamista. Kuvat helpottavat aineiston tulkitsemista ja toimivat usein perusteltuna lähtökohtana monimutkaisempien tilastollisten mallien käyttämiselle.

4.4 Populaation luonteesta

Populaatio on siis ryhmä, josta otos on peräisin.

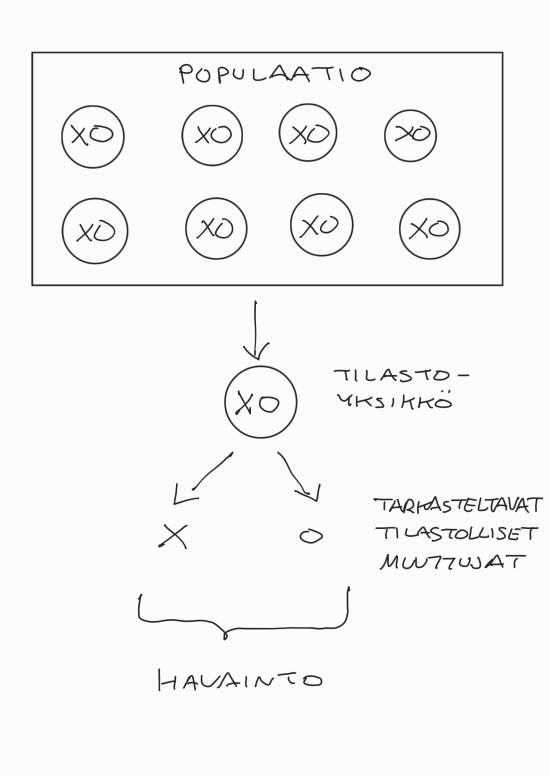
- Populaatio ja otos määriteltiin lyhyesti jo aiemmin (kertaa siis näiden määritelmät!).
- Esim. kyselytutkimuksessa populaatio voi olla kirjaimellinen populaatio, kuten tietyn alueen asukkaat, mutta mittauksia tehtäessä, tai kun kaikki mahdollinen data on käytettävissä, populaatio muuttuu matemaattiseksi idealisaatioksi, joka edustaa kaikkia mahdollisia ko. populaatioon kuuluvia havaintoja.

Populaatio voidaan ajatella yksilöiden joukkona, mutta myös todennäköisyysjakaumana satunnaisille havainnoille, joka on otettu kyseisestä populaatiosta.

- Tämä tarkoittaa, että populaatio ei ole vain fyysinen joukko yksilöitä, vaan se edustaa myös kaikkia mahdollisia havaintoja, jotka voidaan poimia kyseisestä joukosta.
- Populaatioita voidaan tiivistää esim. odotusarvon ja varianssin kautta, jotka kuvaavat populaation ominaisuuksia samalla tavalla kuin otoskeskiarvo ja otosvarianssi kuvaavat otoksen ominaisuuksia.

Aineisto on siis lopulta arvonnan tulos eli mitkä perusjoukon (populaation) tilastoyksiköt tulevat valituksi otokseen, ja siten myös tutkimuksen kohteita kuvaavat tiedot vaihtelevat satunnaisesti arvontakerrasta toiseen.

Usein data ei käytännössä synny kirjaimellisesta populaatiosta otoksena. Kun meillä on kaikki olemassa oleva data, voimme kuvitella sen olevan peräisin sellaisesta populaatiosta tapahtumia, jotka olisivat voineet tapahtua, mutta eivät tapahtuneet. Tämä tarkoittaa, että vaikka meillä olisi kaikki mahdollinen data, voimme silti käyttää tilastollisia menetelmiä ja malleja analysoidaksemme sitä ikään kuin se olisi otos suuremmasta, hypoteettisesta populaatiosta.



Kuva: Populaatiosta havaintoon, joka koostuu kahdesta tilastomuuttujasta.

Populaation muodostavilta tilastoyksiköiltä havaitaan/mitataan/tarkastellaan tutkimuksen kannalta niiden kiinnostavia ominaisuuksia eli **tilastollisia muuttujia** ja lopulta niiden arvoja.

- Mielenkiinnon kohteena olevia tilastollisia muuttujia kutsutaan **tutkimusmuuttujiksi** (kuten tulot ja kuntien äänestysprosentti) ja niiden lisäksi voidaan kerätä lisätietoa eli **taustamuuttujia** (näitä voisivat olla esimerkiksi asuinpaikka ja kunnan väkiluku).

Esimerkki: vaalitutkimukset. Politiikan tutkimuksen alalla yksi mielenkiintoinen tutkimuskohde on tutkia kuntavaaleissa äänestävien ihmisten tuloja.

- Tällöin jokainen äänioikeuttaan käyttävä muodostaa oman tilastoyksikkönsä. Vastaavasti populaationa (perusjoukkona) toimii kaikki äänestysikäiset kansalaiset, jotka äänioikeuttaan käyttävät.
- Toinen tutkimuskysymys voisi käsitellä kuntien välistä äänestysaktiivisuutta.Tällöin jokainen kunta muodostaa oman tilastoyksikkönsä ja vastaavasti 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.
  - Pohdi, miksi pelkästään äänioikeuttaan käyttävien tutkiminen saattaisi olla tutkimuksen tulosten luotettavuuden kannalta ongelmallista?

- Tilastoyksiköiden tilastollisilla muuttujilla on tietty mahdollisten arvojen joukko, ja näillä arvoilla on jokin **jakauma** populaatiossa. Palaamme myöhemmin tilastotieteen keskeisiin jakaumiin ja niiden esittelyyn tarkemmin, mutta määritellään seuraavassa kuitenkin populaatiojakauma.

Populaatiojakaumalla tarkoitetaan potentiaalisten havaintojen jakaumaa koko populaatiossa. Se viittaa myös geneerisen satunnaismuuttujan todennäköisyysjakautumaan, joka kuvaa, kuinka todennäköistä on saada tietty arvo satunnaismuuttujan realisaationa.

Esimerkiksi edelliseen esimerkkiin viitaten äänestysikäisten tulot voivat määritelmästä riippuen saada minkä tahansa positiivisen arvon mutta kunnan äänestysprosentti on luonnollisesti rajattu nollan ja sadan prosentin väliin.

4.5 Todennäköisyysjakauma

4.5.1 Yleistä taustaa todennäköisyysjakaumille

Tilastolliset menetelmät perustuvat todennäköisyyslaskennan tuloksiin ja tarjoavat keinon hallita satunnaisuuden aiheuttamaa epävarmuutta.

Todennäköisyysjakauma

Tilastolliset mallit perustuvat satunnaismuuttujan mahdollisten tulosvaihtoehtojen todennäköisyyksiä kuvaavalle **todennäköisyysjakaumalle**, joka määrää millä todennäköisyydellä satunnaismuuttuja saa erilaisia arvoja.

Kertymäfunktio

Todennäköisyysjakauma on yleisnimitys matemaattiselle ilmaisulle, joka kuvaa satunnaismuuttujan Y mahdollisuutta saada arvon y. Tämä jakauma auttaa ymmärtämään satunnaismuuttujan käyttäytymistä ja tekee mahdolliseksi tilastollisten analyysien, kuten esim. ennusteiden tekemisen.

Satunnaismuuttujalla Y on **kertymäfunktio**, joka määritellään kaavalla

$$F_Y(y) \stackrel{merk.}{=} F(y) = P(Y \leq y)$$

eli kyseessä on todennäköisyys, että Y saa arvon, joka on enintään y. Ts. kertymäfunktion arvo on nollan ja yhden välillä (eli ts. $[0, 1]$): $\lim_{y \to -\infty} F(y) = 0$ ja $\lim_{y \to \infty} F(y) = 1$, ja kuvaa paljonko todennäköisyysmassaa on kertynyt vasemmalta pisteeseen y saakka.

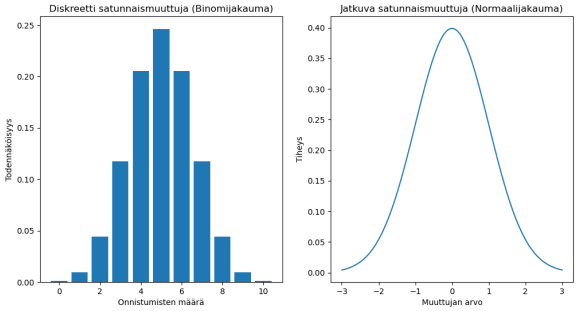
Pistetodennäköisyysfunktio **Diskreetin satunnaismuuttujan** todennäköisyysjakauma voidaan usein esittää taulukkomuodossa. Eri arvojen todennäköisyydet muodostavat kyseisen satunnaismuuttujan todennäköisyysjakauman, **pistetodennäköisyysfunktion** (ptnf:n), jota voidaan havainnollistaa esimerkiksi pylväsdiagrammilla

- Esimerkkejä diskreeteistä tn-jakaumista ovat mm. Bernoulli-jakauma, binomijakauma ja Poisson-jakauma. Näitä käsitellään tarkemmin Osassa II.

Esimerkiksi allaesiteltävä puoluiden kannatusosuuksia koskeva esimerkkitapaus on esimerkki pistetodennäköisyysfunktiosta, jonka taustalla voidaan ajatella olevan satunnaismuuttuja Y, joka kuvaa yksittäisen satunnaisen äänestäjän kannattamaa puoluetta. Ennen hänen vastaustaan voidaan ajatella, että on olemassa tietyt kyselyhetken aikaiset todennäköisyydet, että ko. henkilö kannattaa tiettyä puoluetta.

Tiheysfunktio Vastaavasti **jatkuvan satunnaismuuttujan** arvot muodostavat jonkin reaaliakselin välin, joka sisältää äärettömän määrän lukuja. Tämän

vuoksi jatkuvan satunnaismuuttujan todennäköisyysjakauman esittäminen pistetodennäköisyysfunktion kautta ei ole luontevaa, vaan jakauma esitetään satunnaismuuttujan **tiheysfunktion** avulla.



Kuva: Esimerkkitapaukset pistetodennäköisyysfunktiosta (binomijakauma) ja tiheysfunktiosta (normaalijakauma).

Havaitulle aineistolle perustettava tilastollinen malli perustuu juuri sm:jan todennäköisyysjakaumaan, joka riippuu yhdestä tai useammasta (arvoltaan yleisesti tuntemattomasta) **parametrista** ja kuvaa kyseisen muuttujan säännönmukaista satunnaisvaihtelua.

Parametrit määrittävät todennäköisyysjakauman (ja/tai tilastollisen mallin) käyttäytymisen. Parametrit estimoidaan (=arvioidaan) käytettävissä olevan aineiston perusteella.

Esimerkiksi myöhemmin (Osassa II) esiteltävässä yhden selittäjän lineaarisessa regressiomalli sisältää kaksi parametria (ja varianssiparametrin), jotka määrittävät regressiosuoran yhtälön (ns. vakiotermi ja kulmakerrointa mittaavaa parametri).

Parametrien estimointi on keskeinen osa tilastollista analyysiä, kuten erinäisten päätelmiä/tulkintojen tekemistä ja ennusteiden muodostamista.

Tilastollinen malli riippuu siis tehdystä todennäköisyysjakaumaoletuksesta sekä havaitusta aineistosta eli mielenkiinnon kohteena olevan satunnaismuuttujan realisaatioista. **Tavoitteena on pyrkiä arvioimaan sitä populaatiotason todennäköisyysjakaumaa, joka on tutkimusaineiston generoinut**.

- Tarkemmin sanottuna tavoitteena on arvioida populaatiotason todennäköisyysjakauman parametreja, jotka määrittävät jakauman muodon ja siten eri tulosvaihtoehtojen todennäköisyydet!

Esimerkki diskreettiä aineistoa koskevasta jakaumasta: puolueiden kannatusmittaus

Poliittisten puolueiden kannatustutkimuksia mittaavia galluppeja tehdään erityisesti vaalien alla usean tilastollisia kyselytutkimuksia tekevän tahon toimesta.

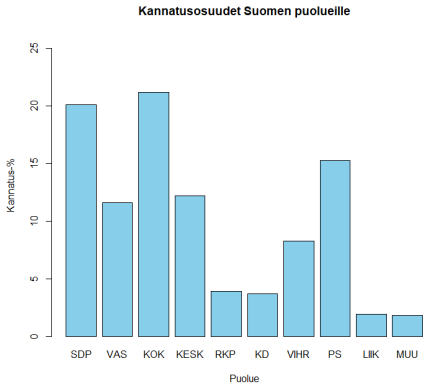
Puoluevalintaa (tai valitsematta jättämistä) voidaan pitää äänestysikäisten keskuudessa satunnaismuuttujana, jolla on äärellinen määrä tulosvaihtoehtoja: puolueiden määrän lisäksi vastaaja voi ilmoittaa ettei tiedä ketä aikoo äänestää tai että ei aio äänestää ollenkaan.

- Täten puolueiden kannatus on diskreetti muuttuja ja kannatustutkimuksessa kohdepopulaatio on äänestysikäiset (täysi-ikäiset) Suomen kansalaiset.
- Osoittautuu, että tässä yksinkertaistetussa tapauksessa paras arvio (nk. suurimman uskottavuuden estimaatti, josta lyhyesti Osassa II) todennäköisyydelle, että satunnaisesti valittu äänestysikäinen kansalainen äänestää tiettyä puoluetta on kyseisen puolueen saama osuus kaikista kyselyyn vastanneista, kunhan otos on edustava. Otoksen edustavuus on erittäin tärkeä yksityiskohta ja sitä tarkastellaan hetken päästä vielä tarkemmin

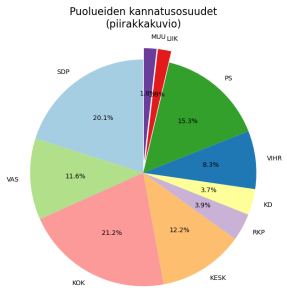
Oheisessa taulukossa on esitelty tuloksia Taloustutkimuksen Ylelle suorittamasta puolueiden eduskuntavaalikannatusta mittaavasta tutkimuksesta, jossa haastateltiin 2481 ihmistä puhelinhaastatteluina ja internetpaneelina aikavälillä 7.6.2024–2.7.2024.

- Haastatelluista 1910 kertoi puoluekantansa ja tutkimuksen virhemarginaalin kerrottiin olevan suurimmillaan 1,9 prosenttiyksikköä suuntaansa (virhemarginaalista myös lisää Osassa II).
- Tässä tapauksessa mielenkiinnon kohteena olevat tilastollisen mallin parametrit, jotka kuvaavat kannatusosuuksia koko valtakunnan tasolla ovat juurikin edellä mainitut äänestystodennäköisyydet! Saadut todennäköisyydet (estimaatit) on esitetty alla olevassa taulukossa ja kuvaajissa.

Puolue	SDP	VAS	KOK	KESK	RKP	KD	VIHR	PS	LIIK	MUU
Kannatus-%	20.1	11.6	21.2	12.2	3.9	3.7	8.3	15.3	1.9	1.8



Kuva: Puolueiden kannatukset pylväsdiagrammeina.



Kuva: Puoluiden kannatukset piirakkakuviona.

Pylväsdiagrammi

Edellä kuvioissa esitellään samalla kaksi tilastotieteelle tyypillistä graafista esitysmuotoa eli **pylväsdiagrammi** ja **piirakkakuvio**. Pylväsdiagrammi (bar chart) on kuvio, jossa kategoriset tiedot (edellä puolueet) esitetään suorakulmaisina pylväinä. Pylvään korkeus tai pituus kuvaa havaintojen määrää tai arvoa eli tässä esimerkissä kannatusosuuksia. Piirakkakuvio Vastaavasti piirakkakuvio (pie chart) merkitsee ympyrää, joka on jaettu sektoreihin, jotka

kuvaavat eri kategorioiden prosentuaalisia osuuksia kokonaisuudesta. Molemmat esitystavat ovat hyödyllisiä – mutta ne palvelevat hieman eri tarkoituksia. Pylväsdiagrammi korostaa vertailua, piirakkakuvio osuuksia. Tilastotieteessä on tärkeää valita esitystapa, joka tukee tutkimuskysymystä ja auttaa tulkitsemaan aineistoa mahdollisimman selkeästi.

4.5.2 Odotusarvo ja varianssi

Satunnaismuuttujan todennäköisyysjakauman keskeisiä piirteitä voidaan tiivistää **odotusarvon**, **varianssin** ja **keskihajonnan** avulla. Näille on olemassa vastaavat otosvastineet, joita käytetään jakauman ominaisuuksien arvioimiseen otosaineiston perusteella.

Odotusarvo

Odotusarvo. Satunnaismuuttujan (Y) odotusarvo ($\mathsf{E}(Y)$) kuvaa satunnaismuuttujan odotettavissa olevaa arvoa.

- Merkinnän ($\mathsf{E}(Y)$) käyttö juontaa juurensa englannin kielen sanoihin expectation (“odotus”) ja **expected value** (“odotusarvo”).
- Odotusarvo kuvaa siis jakauman painopistettä.
- Odotusarvo on satunnaiskokeen tulosvaihtoehtojen **painotettu keskiarvo**, jossa kunkin tuloksen painona on vastaavan tapauksen todennäköisyys.

Esimerkki: Nopanheiton odotusarvo. Perinteinen esimerkki odotusarvosta on tavallisen kuusitahoisen nopan silmäluvun odotusarvo. Nopanheitto on diskreetti satunnaisilmiö 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 ($P(Y = y) = \frac{1}{6}$).

Nopanheiton silmäluvun odotusarvo määritetään siis painotettuna keskiarvona:

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

- Sivuhuomiona edelliseen liittyen edellä lukujen y_i painotetun summan yhteydessä käytetään **summamerkintää**. Yleisesti tällä tarkoitetaan sitä, että (esimerkinomaisesti) havaintojen (lukujen) $y_1, \ldots, y_n$ summaa $y_1 + \cdots + y_n$ voidaan merkitä lyhyesti $\sum_{i=1}^{n} y_i$. Summamerkinnästä näkyy mistä indeksointi summassa alkaa (havaintoyksiköstä 1) ja mihin se päättyy (havaintoyksikköön n). Summamerkintä luetaan tässä tapauksessa

siis "summa y_i, jossa i käy 1:stä n:ään". Ajoittain indeksoinnit jätetään myös merkitsemättä, jos sillä ei ole tarkastelussa merkitystä tai se on asiayhteydestä muuten selvä.

Diskreettien sm:jien (kuten yllä olevassa esimerkissä) sijaan jatkuvien satunnaismuuttujien tapauksessa odotusarvon $\mathbb{E}(Y)$ määritteleminen merkitsee integroinnin käyttämistä:

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y)\, dy, \tag{4.1}$$

jossa $f_Y(y)$ on satunnaismuuttujan Y tiheysfunktio. Tarkemmat yksityiskohdat jatkuvien sm:jien odotusarvoista jäävät todennäköisyyslaskennan kursseille.

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

Varianssi ja keskihajonta

Satunnaismuuttujan (Y) hajontaa voidaan mitata **varianssilla**

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

tai sen neliöjuuren eli **keskihajonnan** avulla

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

- Merkintöjen ($\mathsf{Var}(Y)$) ja ($\mathrm{D}(Y)$) taustalla on englannin kielen sanat **variance** (varianssi) ja **deviation** ("poikkeama" tai "hajonta").

- Ts. mitä lähempänä nollaa (nollaa ei kuitenkaan saavuteta!) keskihajonta ja varianssi ovat, sitä todennäköisempää on, että satunnaismuuttujan arvo on lähellä odotusarvoa.

Odotusarvon ja varianssin (keskihajonnan) tavanomaiset **estimaattorit**, eli konkreettiseen numeeriseen aineistoon, otokseen, liittyvät ja siitä laskettavat vastineet, ovat otoskeskiarvo ja otosvarianssi (otoshajonta), joihin tutustuaan pian myöhemmin. Tähän liittyen määritellään tässä vaiheessa tunnusluku

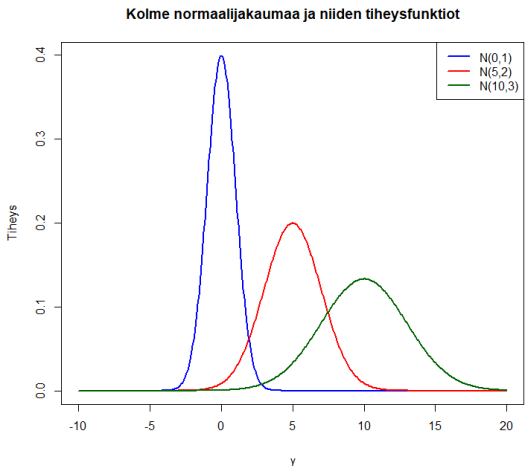
4.5.3 Todennäköisyysjakauma: Esimerkkinä normaalijakauma

Käydään seuraavaksi läpi tilastollisen mallin parametrien merkitystä **normaalijakauman** tapauksessa (ja ylipäätään esitellään samalla normaalijakaumaa). Normaalijakauma on yksi keskeisimpiä ja tärkeimpiä todennäköisyysjakaumia tilastotieteessä ja sen avulla voidaan helposti kuvata jakauman parametrien vaikutusta jakauman muotoon. Normaalijakauma

Normaalijakauma. Satunnaismuuttujan Y noudattaessa normaalijakaumaa merkitään yleisesti $Y \sim \mathsf{N}(\mu, \sigma^2)$, jossa μ on jakauman odotusarvo ja σ^2 sen varianssi. Normaalijakauman **tiheysfunktio**

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

määrittelee parven normaalijakaumia kun parametreille ( ) (odotusarvo) ja (^2) (varianssi) annetaan erilaisia arvoja. Alla olevassa kuvassa on kuvattu erilaisia normaalijakauman tiheysfunktion muotoja eri parametriarvoilla (ks. oikea yläkulma).



Kuva: Normaalijakauman tiheysfunktion muotoja eri parametriarvoilla.

Tiheysfunktio kuvaa siis satunnaismuuttujan eri tulosvaihtoehtojen (vaaka-akselin numeeriset arvot) todennäköisyyksiä.

Tarkasteltaessa ylläolevia normaalistijakautuneita sm:jia, esimerkiksi satunnaismuuttujaa, joka on normaalisti jakautunut odotusarvoltaan (= 5) ja varianssiltaan ($\sigma 2 = 2$) merkitään ($Y \sim \mathsf{N}(5, 2)$) (ts. kuvion punainen tf.). Tämän sm:jan

realisaatiot ovat keskittyneet odotusarvon ympärille niin, että suuret poikkeamat odotusarvosta kumpaankaan suuntaan ovat vähemmän todennäköisiä kuin pienemmät poikkeamat.

- Toisin sanoen, odotusarvon määritelmän mukaisesti, normaalijakauman parametreista odotusarvo kuvaa sitä mikä on satunnaisilmiön generoimien realisaatioiden odotettavissa oleva arvo. Täten odotusarvo ( ) kuvaa normaalijakauman tapauksessa sen "sijaintia" (huomioi, että tämä on erityisesti normaalijakauman ominaisuus).

- Vastaavasti varianssi kuvaa sitä, kuinka paljon realisaatiot keskimäärin poikkeavat odotusarvosta. Suurempi varianssi "leventää" normaalijakauman tiheysfunktiota, jolloin suuremmat poikkeamat ovat suhteellisesti todennäköisempiä.

- Ajoittain on tulkinnan kannalta helpompaa hahmottaa satunnaismuuttujan hajontaa varianssin neliöjuurella, eli keskihajonnalla.

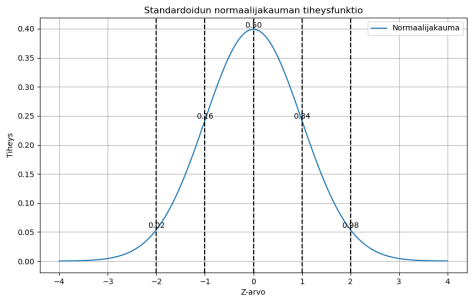
Standardoitu normaalijakauma Erityisesti ns. **standardoidun normaalijakauman** tapauksessa odotusarvo on 0 ja varianssi 1 eli $\mathsf{N}(0,1)$. Se voidaan määritellä normaalisti jakautuneen sm. $Y \sim \mathsf{N}(\mu, \sigma^2)$ perusteella standardoimalla

$$Z = \frac{Y - \mu}{\sigma},$$

jossa siis $Z \sim \mathsf{N}(0,1)$. Tätä jakaumaa tarvitaan ja käsitellään myöhemmin. Seuraavassa kuvassa esitetään standardoidun normaalijakauman tiheysfunktion kuvio ja siihen liittyvien eri standardoitujen **z-arvojen** mukaisia **prosenttipisteitä**. z-arvo

- Standardoidun normaalijakauman kertymäfunktion arvo tietyssä pisteessä (ts, tietyn z-arvon kohdalla allaolevassa kuvassa) on sen määritelmän mukaisesti kertynyt todennäköisyys (tiheys), joka jää tästä pisteestä katsoen vasemmalle puolelle jakaumaa. Ts. Prosenttipiste p. prosenttipiste on se havaintoarvo, joka jakaa havainnot kahteen osaan niin, että p% havaintoarvoista on ko. havaintoarvoja pienempiä tai yhtäsuuria kuin se.

- Esimerkiksi z-arvoon $z = -2$ liittyvä kertymäfunktion arvo on n. 0.02 (eli n. 2%) eli ts. tähän pisteeseen mennessä kertynyt todennäköisyysmassa vasemmalta katsoen on tämä n. 2%. Vastaavasti arvoon $z = 2$ liittyen kertymäfunktion arvo on n. 0.98 (eli n. 98%. Tarkalleen ottaen pisteessä $z = 1.96$ kertymä on 97,5%). Koska standardinormaalijakauman tiheysfunktio on symmetrinen odotusarvon 0 ympärillä, niin kohta $z = 0$ erottaa todennäköisyysmassaa 50% verran sen molemmille puolille.



Kuva: Standardoitu normaalijakauma ja sen tiheysfunktio. Kuvaan on korostettu muutamia z-arvojen mukaisia kertymäfunktion arvoja.

4.6 Parametrien estimointi

Tilastollinen malli, joka kuvaa satunnaisilmiötä, voidaan usein (ja tällä kurssilla) ajatella perustuvan johonkin parametriseen todennäköisyysjakaumaan. Tilastollisen analyysin tavoitteena on selvittää, mikä prosessi on tuottanut havaittavan aineiston. Kuten aiemmin todettiin, todennäköisyysjakaumat riippuvat **parametreista**, ja erityisesti klassisessa tilastotieteessä tehdään oletus, että aineiston taustalla vaikuttavat vakioarvoiset mutta tuntemattomat parametriarvot.

- Toisin sanoen tilastollinen aineisto koostuu usean tilastoyksikön havaituista tutkimusmuuttujan arvoista (realisaatioista), ja tavoitteena on aineiston avulla arvioida sitä todennäköisyysjakaumaa, joka on generoinut nämä havainnot.

Lähtökohtana, **tiettyjen yksinkertaistavien ja vahvojen oletusten nojalla**, pidetään usein sitä, että kaikki tutkimusmuuttujan realisaatiot ovat peräisin samasta todennäköisyysjakaumasta. Tälle oletukselle voidaan perustaa havaintoaineiston **yhteisjakauma** (ks. Osa II), joka kuvaa koko aineiston todennäköisyysjakaumaa. Parametriarvot ovat kuitenkin edelleen tuntemattomia.

- Huom. Yhteisjakauma voidaan määritellä myös ilman oletusta, että kaikki havainnot noudattavat samaa jakaumaa.

Tilastotieteen keskeinen tehtävä on **estimoida** eli arvioida, hyvin perustelluin menetelmin, mitkä nämä tuntemattomat parametriarvot ovat. Parametrit kuvaavat satunnaisilmiötä, joka on lopulta tutkimuksen varsinainen mielenkiinnon kohde.

Nk. **suurimman uskottavuuden estimoinnissa** aineiston generoiman (oletetun) todennäköisyysjakauman parametriarvot **estimoidaan** käytettävän otoksen/aineiston avulla siten, että aineistoa kuvaavan tilastollisen mallin **uskottavuusfunktio** maksimoituu. Tämä merkitsee käytännössä valitun todennäköisyysjakauman sovittamista havaintoaineistoon mahdollisimman hyvin. Näin muodostuvia parametrien arvoja kutsutaan **estimaateiksi** ja ne kuvaavat käsillä olevan aineiston ja tehtyjen oletusten jälkeen parhaiten perusjoukkoa.

Uskottavuusfunktio. Tilastollisen mallin pohjana oleva todennäköisyysjakauma määrittää yhdessä kaikkien havaintojen kanssa uskottavuusfunktion, jota käytetään tilastollisen mallin tuntemattomien parametriarvojen estimoimiseen suurimman uskottavuuden (SU) menetelmällä.

Tähän SU-estimointimenetelmään palataan hieman tarkemmin vielä Osassa II ja erityisesti useaan kertaan myöhemmin tilastotieteen perus-, aine- ja myös syventävissä opinnoissa.

4.7 Hypoteesien testaaminen

Parametrien estimointi ja niihin liittyvä **hypoteesien testaus** ovat tilastolliseen tutkimukseen liittyvän tilastollisen päättelyn keskeisiä välineitä, joiden avulla pyritään tekemään johtopäätöksiä tutkittavasta ilmiöstä havaintoaineiston perusteella.

Hypoteesien testauksessa testataan tilastollisen mallin parametreihin liittyviä väittämiä vertaamalla niitä havaintoaineistosta saatuihin estimaatteihin.

Esimerkki. Voidaan esimerkiksi testata hypoteesia, jonka mukaan aineistoa generoivan prosessin odotusarvo olisi 0 eli $\mu = 0$.

Hypoteesien testaamisella voidaan siis testata, onko jokin asetettu hypoteesi uskottava nyt havaitun aineiston ja siihen perustuvien estimointitulosten valossa.

Esimerkkejä. Parametreihin kohdistuva tilastollinen testaus voi (tarkemmin vielä määritellyissä ja johdetuissa tilastollisissa malleissa) vastata esimerkiksi seuraavanlaisiin tutkimuskysymyksiin:

- Onko suomalaisten miesten keskipituus 180 cm?
- Vaikuttaako yliopistokoulutus tulevaisuuden ansioihin?
- Auttaako tietty lääkeaine jonkin sairauden hoidossa?

- Voiko osakemarkkinoiden tuottoja ennustaa jonkin (selittävän/ennustavan) tekijän avulla?

Nollahypoteesi ja vastahypoteesi

Nollahypoteesi on tilastollisessa testauksessa käytettävä lähtöoletus, joka kuvaa odotettavissa olevaa tilannetta. Tyypillisesti (vaikkakaan ei aina) nollahypoteesi esittää, että kahden ilmiön välillä "ei ole yhteyttä" tai että esimerkiksi tietty hoito ei ole tehokas. Tilastollisen testauksen tarkoituksena on arvioida, voidaanko nollahypoteesi hylätä havaintoaineistosta saadun tilastollisesti merkitsevän löydöksen perusteella.

- Nollahypoteesin paikkansapitävyyttä arvioidaan usein **p-arvon** avulla. Pieni p-arvo viittaa siihen, että havaittu ilmiö olisi epätodennäköinen, jos nollahypoteesi olisi tosi. Tällöin nollahypoteesi voidaan asettaa kyseenalaiseksi ja mahdollisesti hylätä.
- On tärkeää huomata, että nollahypoteesia ei voida todistaa "oikeaksi" — voidaan ainoastaan todeta, että aineisto ei anna riittävästi näyttöä sen hylkäämiseksi, jos p-arvo on esimerkiksi suurempi kuin 0.1 (eli yli 10 %).

Vastahypoteesi sisältää usein mielenkiinnon kohteena olevan tapahtuman, kuten "on eroa" tai "on vaikutusta".

Testisuure ja p-arvo

(Seuraava pitkälle Mellinin (2004, s. 328-329) mukaan)

Tilastollisen testin toteuttaminen perustuu erityiseen **testisuureeseen**, joka pyrkii keräämään yhteen "todisteet" tehtyjä oletuksia vastaan. Testisuure mittaa, kuinka hyvin havaittu aineisto sopii yhteen nollahypoteesin mukaisen tilanteen kanssa.

- Mikäli aineistosta laskettu testisuureen arvo poikkeaa niin paljon siitä arvosta, joka saataisiin tilanteessa, jossa nollahypoteesi pitää paikkansa, voidaan nollahypoteesi hylätä ja vaihtoehtoinen hypoteesi ottaa tilalle.

p-arvo (eli **havaittu merkitsevyystaso**) on yleisellä tasolla (**ilman tarkempaa tilastomatemaattista määritelmää**) todennäköisyys saada testisuureen arvo, joka poikkeaa nollahypoteesin mukaisesta arvosta yhtä paljon tai enemmän kuin havaittu arvo, **olettaen että nollahypoteesi pätee**.

- Pieni p-arvo (esimerkiksi $p < 0.05$ kun merkitsevyystaso α on 0.05 (5 %), eli $p < \alpha$) viittaa siihen, että havaittu testisuureen arvo olisi epätodennäköinen, jos nollahypoteesi olisi voimassa. Tällöin on perusteltua asettaa nollahypoteesi kyseenalaiseksi.

- On tärkeää huomata, että **p-arvo ei ole todennäköisyys sille, että nollahypoteesi on tosi**. Tämä on yleinen väärinkäsitys.

- p-arvo riippuu valitusta tilastollisesta testistä ja käytetystä testisuureesta. Näitä yksityiskohtia ei käsitellä tarkemmin vielä tällä kurssilla.

Käytännössä johtopäätelmiä varten valitaan **ennalta** (ennen tilastollisen testin suorittamista) tilastollinen merkitsevyystaso α, joka määrittää kuinka pieni p-arvo vaaditaan nollahypoteesin hylkäämiseksi. Tyypillisiä valintoja ovat $\alpha = 0.1, \alpha = 0.05$ tai $\alpha = 0.01$ vastaten 10 %, 5 % ja 1 % merkitsevyystasoja.

Havaittu vaikutus katsotaan **tilastollisesti merkitseväksi**, kun tilastolliseen testiin liitettävä p-arvo on pienempi kuin α. Tämä tarkoittaa, että havaittu tulos olisi epätodennäköinen, jos nollahypoteesi ja kaikki muut mallinnusoletukset pitäisivät paikkansa. Tällöin nollahypoteesi voidaan perustellusti **hylätä**.

- Todetaan hyvin yleisellä tasolla, että tiedeyhteisöllä on usein taipumus jättää julkaisematta tutkimustuloksia, joissa nollahypoteesi jää voimaan. Yleensä tämä tilanne syntyy, kun lopputulos ei eroa jo aikaisemmin otaksutusta. (Toki ajoittain tilanne on myös toisinpäin eli "toivotaan" nollahypoteesin hylkäämistä).

Tilastolliseen testaamiseen palataan vielä tarkemmin tämän materiaalin Osassa II.

Esimerkki: Testaus uskottavuuspohjaisessa tilastollisessa päättelyssä. Palataan aiemmin esitettyyn kuviteltuun esimerkkiin tutkijasta, joka haluaa tutkia sademääriä.

- Sanotaan, että tutkija on kiinnostunut tutkimaan väitettä, jonka mukaan Turussa sataa keskimäärin yli 800 milliä vuodessa.

- Oletetaan lisäksi, että tutkijalla on käytettävänä satunnaisotantaan perustuva arvio Turun sademääristä vuosilta 1980–2023 ja että vuosittaiset sademäärät ovat riippumattomia ja samoin jakautuneet.

- Nyt tutkija voi suorittaa väitteen tilastotieteellisen tutkimuksen muodostamalla tilastollisen mallin, jonka tavoite on kuvata Turun keskimääräistä vuosittaista sademäärää.

- Asetettua hypoteesia voidaan siis testata suhteessa estimointituloksiin, jotka kuvaavat vuosittaisen sademäärän odotettavissa olevaa arvoa.

- Tutkija voi lopulta muodostaa tilastollisen ennustemallin, joka huomioi sademäärissä mahdollisesti esiintyvän pitkän aikavälin muutoksen ja tehdä mallin pohjalta ennusteen tulevan vuoden sademäärästä. Tilastollista ennustamista käsitellään Osassa II.

Chapter 5

Tilastolliset aineistot, niiden kerääminen ja mittaaminen

Edellä käsiteltiin tilastotieteen suhtautumista satunnaisilmiöihin. Tässä luvussa tarkastelemme lähemmin miten reaalimaailman satunnaisilmiöistä kerätään tietoa ja miten niitä voidaan mitata. Ts. miten käytännössä, tässä kohtaa yleisellä tasolla, havaintoaineistot muodostuvat ja miten niiden analysointia lähestytään.

Tilastotiede rakentuu ajatukselle ilmiöiden tutkimisesta rajallisen ja epävarman tiedon vallitessa. Käytännössä tämä tarkoittaa sitä, että tutkimuksen kohteena olevat rajalliset aineistot sisältävät niin systemaattista kuin satunnaisuudesta johtuvaa vaihtelua. Tilastollisten menetelmien avulla **pyrimme erottamaan systemaattisen vaihtelun satunnaisesta**, eli **erottamaan signaalin kohinasta**, tekemällä tilastollista päättelyä aineiston generoimasta mekanismista. Lyhyesti tämä tarkoittaa aineiston systemaattisen vaihtelun tilastollista mallintamista, pitäen näin sisällään mm. tarkasteltavan tilastollisen mallin tuntemattomien parametrien estimoinnin otoksen pohjalta, joka kattaa vain (pienen) osajoukon koko populaation (perusjoukon) tilastoyksiköistä.

Voidaksemme tehdä uskottavaa päättelyä "havainnoista parametreihin", tulee käytettävän aineiston, usein **otoksen**, olla riittävän **edustava**. Tämän luvun keskeisin oppi onkin, että miten **otanta** tulisi suorittaa, jotta havaintoaineisto olisi **edustava otos** populaatiosta. Vaikka aineiston hankinta vaatii yleensä runsaasti käytännön työtä, kannattaa se suunnitella ja tehdä huolellisesti, sillä huonosti toteutetun otannan vuoksi tutkimusongelman kannalta keskeisiä johtopäätöksiä ei voida (useimmiten) tehdä myöhemmässä vaiheessa jo muodostunutta aineistoa analysoitaessa.

5.1 Kokonaistutkimus ja otantatutkimus

Lähtökohtaisesti 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.

Esimerkkejä. Jos tutkitaan Suomen kuntia, niin kokonaistutkimuksessa tutkitaan kaikki kunnat. Kunnista on useimmissa tilanteissa mahdollista kerätä mielenkiinnon kohteena olevia aineistoja eli tilastollisten muuttujien arvoja.

Toisaalta, jos tutkitaan jonkin lääkeaineen vaikutuksia ihmisiin, niin kokonaistutkimuksessa tutkittaisiin jokainen ihminen erikseen. Selvää on, että tällainen kokonaistutkimus olisi liian vaikeaa ja järjetöntäkin käytännössä toteuttaa.

Otantatutkimus

Otantatutkimus. Otantatutkimuksessa tutkimus kohdistetaan johonkin (populaation/perusjoukon) osajoukkoon, joka poimitaan sopivaa **otantamenetelmää** käyttäen (ks. Osa II) ja populaatiota/perusjoukkoa koskevat johtopäätelmät tehdään tähän otokseen perustuen.

- Otantatutkimus on usein luonnollinen valinta, sillä koko populaation tutkiminen ei useinkaan ole mahdollista tai kannattavaa.
- Perusjoukosta otokseen poimittuja alkioita kutsutaan **otosyksiköiksi** ja niiden muodostama osajoukko, eli **otos**, on se osa perusjoukkoa, joka tutkitaan tutkimusaineiston keräämisen jälkeen.

Esimerkkejä. Esimerkiksi aseiden patruunoita valmistava tehtailija ei voi tutkia toimivatko kaikki ammukset! Myöskään valaisimien valmistaja tuskin tekee kokonaistutkimuksia valmistamiensa tuotteiden kestoajan selvittämiseksi.

Lääketutkimusta tehdäänkin poikkeuksetta otantatutkimuksena (ja kontrolloituina kokeina), 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.

Otantatutkimuksessa keskitytään siis perusjoukkoa edustavan mutta pienemmän, mieluusti satunnaisesti valitun otoksen tutkimiseen. Otantatutkimus on halvempi kuin kokonaistutkimus ja tulokset saadaan nopeammin!

- Otantatutkimuksissa tiedot kerätään useimmiten haastattelemalla, kirjallisella/sähköisellä kyselyllä tai suoraan tietorekistereistä. Tiedonkeruun toteuttaminen (eri sovelluksissa) määrää osaltaan käytettävän otantamenetelmän.

- Teoriassa äärelliseen perusjoukkoon kohdistuvat kokonaistutkimukset voidaan aina tulkita otantatutkimuksiksi! Tällöin siis perusjoukko tulkitaan otokseksi hypoteettisesta äärettömästä perusjoukosta!

  Esimerkki. Esimerkiksi Galilein tekemät painovoiman vaikutusta kappaleiden putoamisaikaan liittyneet mittaukset. Koetuloksia (mittauksia) voidaan pitää otoksena äärettömästä mahdollisten koetulosten joukosta.

Ehkä hieman yllättäen, otantatutkimuksen tulokset voivat olla myös luotettavampia kuin kokonaistutkimuksen. Otantatutkimuksessa voidaan panostaa enemmän huolelliseen ja tarkkaan mittaamiseen sekä valitun otoksen tavoittamiseen. Kokonaistutkimuksessa vastauskato ja tarkasteltavan populaation valintavirhe ovat mahdollisia siinä missä otantatutkimuksessakin.

Otantateoria onkin yksi tilastotieteen keskeisimpiä oppeja ja tarjoaa teoreettisen kehikon useiden empiiristen tutkimusten tulosten yleistämiseen. Otannan tarkempaa onnistunutta toteuttamista, ml. yksityiskohtia eri otantamenetelmistä, tarkastellaan Osassa II.

5.2 Mittaaminen

Kumpaa tahansa tutkimusotetta (kokonais- tai otantatutkimus) noudatettaessa tietojen keräämisessä on olennaisena osana kohteiden ominaisuuksien **mittaaminen**. Tilastotieteellinen tutkimus perustuu aina mitattaviin satunnaisilmiöihin. Tavoitteena on mittaamalla liittää jokin luku ilmiötä kuvaavaan ominaisuuteen, jota todennäköisyyslaskennan näkökulmasta katsoen mallinnettaisiin satunnaismuuttujan kautta.

- Mittaaminen vaatii aina mittauksen kohteen, hyvin määritellyn mitattavan ominaisuuden ja **mittarin**, joka liittää mielekkäät lukuarvot mitattavaan ominaisuuteen.

- Erilaiset 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

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 laajemminkin tilastotieteessä.

Mittarin 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 yhtenä keskeisimmistä hyvistä ominaisuuksista joita tilastollisten mallien parametrien estimaattoreilta voidaan vaatia. Tähän palataan Osassa II.

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

Huonosti toistettava mittari antaa tilastoyksiköiden samankaltaisille ominaisuuksille hyvin erilaisia arvoja riippuen otoksesta. **Mittausten reliabiliteettia/luotettavuutta** arvioidessa voidaan pohtia esimerkiksi seuraavia kysymyksiä:

- Kuinka hyvin mittaustulokset ovat toistettavissa? Kuinka paljon niissä on ei-sattumanvaraisuutta?
- Mittausten validiteetti: kuinka hyvin pystyttiin mittaamaan sitä, mitä 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. Esimerkkinä tästä on mm. ns. **luottamusväli** eli väli, joka vaihtelee otoksesta toiseen ja rakennetaan siten, että se sisältää mielenkiinnon kohteena olevan parametrin arvon tietyllä varmuudella, kun otantakoetta toistetaan!

- Luottamusväliä käytetään siis osaltaan määrittämään saatavan estimaatin luotettavuutta.

Luotettavuudella voidaan tarkoittaa myös tutkimuksen **objektiivisuutta**.

- **Objektiivinen totuus**: tutkimustulokset ovat samat riippumatta siitä kuka pätevä tutkija tutkimuksen on tehnyt. Tulosten tulisi olla luotettavia, mutta luotettavatkin tulokset voivat olla vikaantuneita siinä mielessä, että ne tarkastelevat asiaa vain yhdestä näkökannalta!

Esimerkiksi. Tarkastellaan C-vitamiinin vaikutusta 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, primääri-kasvaimen sijaintipaikan ja histologisen kasvaintyypin suhteen.
- Seuranta-aika: aika hetkestä, jolloin todettiin tavanomaisten hoitojen olevan tehottomia, kuolinhetkeen saakka.
- Tulos: C-vitamiinia saaneet käsittelyryhmän potilaat elivät 4 kertaa kauemmin (p-arvo < 0.0001)
- Ristiriitaista evidenssiä saatiin tutkimuksessa, jossa on vastaava tutkimusongelma, mutta toteutettu satunnaistettuna kokeena (Moertel et al.~1985). Satunnaistettiin potilaat, joilla pitkälle edennyt paksunsuolen tai peräsuolen syöpä, C-vitamiinia saavien ja lumelääkettä saavien ryhmiin.
- Tulos: kontrolliryhmän potilaat elivät keskimäärin hieman pidempään, mutta ero ei tilastollisesti merkitsevä.

Mistä näiden kahden tutkimuksen erot johtuivat?

- “Huonolla tuurilla” kaltaistetut verrokit erosivat käsittelyryhmän potilaista joillakin merkittävillä tavoilla, joita ei oltu mitattu! Miten kvantifioida “huonoa tuuria”?

- Tilastolliset menetelmät ja tässä tapauksessa p-arvo 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!

5.3 Mitta-asteikot

Mitta-asteikot

Mitta-asteikot. Kuten satunnaismuuttujia koskeneessa luvussa opittiin, satunnaisilmiöillä on erilaisia tulosvaihtoehtoja, jotka osaltaan määrittävät myös satunnaismuuttujien todennäköisyysjakaumia. Ilmiön luonteesta riippuen voidaan näille tulosvaihtoehdoille käyttää erilaisia **mitta-asteikkoja**:

- Laatueroasteikko (luokitteluasteikko, nominaaliasteikko)
- Järjestysasteikko (ordinaaliasteikko)
- Välimatka-asteikko (intervalliasteikko)
- Suhdeasteikko

Luokitteluasteikko

Laatueroasteikko/luokitteluasteikko (nominaaliasteikko): Muuttujan mittaustaso on tällöin sellainen, että sen arvot voidaan luokittaa toisistaan eroaviin luokkiin. Ts. mihin luokkaan kohde kuuluu mitattavan ominaisuuden perusteella?

- Havainnot (tilastoyksiköt) luokitellaan ennaltamääriteltyihin luokkiin ja yksittäinen havainto vain yhteen luokkaan. Luokkien järjestyksellä ei ole merkitystä.
- Kukin tilastoyksikkö kuuluu vain yhteen luokkaan. Tällöin kahdesta tilastoyksiköstä/havainnosta voidaan päätellä vain kuuluvatko ne saamaan luokkaan vai eivät.
- Emme pysty määrittelemään empiirisesti mielekästä järjestystä havaintoarvojen välillä.

Esimerkkejä: Sukupuoli, veriryhmä tai kotikunta.

On syytä huomauttaa, että vaikka mitattava ilmiö ei olisikaan numeerinen, se voidaan aina "koodata" eli muuntaa numeeriseksi. Esimerkiksi perinteisen kaksiarvoisen mies-nainen -muuttujan tapauksessa voidaan käyttää tunnuksia 0 ja 1.

Järjestysasteikko

Järjestysasteikko (ordinaaliasteikko): Tällöin muuttujan arvot voidaan luokittelun lisäksi asettaa (empiirisesti) mielekkääseen järjestykseen. Tällöin siis mittauksen kohteella on "enemmän mitattavaa ominaisuutta" kuin jollakin toisella kohteella.

- Tilastoyksiköt luokitellaan ennalta määrättyihin luokkiin, joilla on yksikäsitteinen järjestys.

Esimerkkejä: Sotilasarvo, sosiaaliryhmä, kilpailun tulos tai sairauksien tarttuvuus.

Välimatka-asteikko

Välimatka-asteikko (intervalliasteikko): Luokittamisen ja järjestyksen asettamisen lisäksi havaintoarvojen välimatkalla on empiirisesti mielekäs tulkinta. Ts. intervalliasteikon tasoisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi).

- Välimatka-asteikolla pystytään mittaamaan yksittäisten luokkien tai havaintoarvojen ero. Kuinka paljon kahden mittauksen kohteen ominaisuudet eroavat toisistaan?
- Intervalliasteikon tasoisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi). Muuttujan voidaan siis ajatella (useimmiten) saavan minkä tahansa reaalilukuarvon.
- Absoluuttista nollapistettä ei kuitenkaan ole vs. suhdeasteikko. Mittarin nollapiste on "keinotekoinen" ja siten vapaasti valittavissa. Samoin voidaan valita käytettävä mittayksikkö vapaasti sekä yhteen- ja vähennyslasku ovat sallittuja. Oleellista on se, että havaintojen välisellä välimatkalla on aina empiirisesti mielekäs tulkinta.

Esimerkkejä: Lämpötilan mittaaminen celcius-asteina. Pystymme numeroarvoina ilmoittamaan onko tänään lämpimämpi, yhtä lämmin vai kylmempi sää kuin eilen ja kuinka monta astetta muutos on. Sen sijaan, jos ilman lämpötila on 0 (celciusastetta), ei se tarkoita, että lämpöä ei ole.

Suhdeasteikko

Suhdeasteikko: Intervalliasteikon ominaisuuksien lisäksi on määriteltynä yksikäsitteinen mittalukujen absoluuttinen nollapiste, minkä myötä havaintoarvojen suuruuksien keskinäiset suhteet voidaan määritellä.

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

Toisaalta toisena esimerkkinä kunnan veroäyri tai henkilön pituus: Absoluuttinen nollapiste on 0.

- Nollapisteen ollessa absoluuttinen, se "pysyy paikallaan" ja mittalukujen suhteet pysyvät samoina.

Tavallisesti eroon väli- ja suhdeasteikon välillä ei tarvitse kiinnittää (liikaa) huomiota. Sen sijaan tärkeää ja useimmiten myös helppoa tehdä ero nominaali-, ordinaari ja väli-tai suhdeasteikollisten muuttujien välillä.

Mitattavien ominaisuuksien luonne voidaan siis selvästikin jakaa kahteen luokkaan:

- **Kvalitatiiviset** ominaisuudet ja siten kvalitatiiviset/laadulliset muuttujat
  - Tutkimuksen kohteet voidaan luokitella kvalitatiivisesti eli laadullisesti toisistaan eroaviin kategorioihin eli luokkiin.
  - Kvalitatiivista muuttujaa kutsutaan usein myös kategoriseksi muuttujaksi.
  - Tällöin muuttujien arvot kuvaavat vain tilastoyksiköiden laadullisia piirteitä.
- **Kvantitatiiviset ominaisuudet**
  - Mittaukseen liittyy tällöin suoraviivaisesti lukuja (ei siis sellaisiksi koodattavia ominaisuuksia)

Luokittelu- ja järjestysasteikkoa kutsutaan ajoittain myös **kvalitatiivisiksi asteikoiksi**. Vastavasti välimatka- ja suhdeasteikkoa kutsutaan myös **kvantitatiivisiksi asteikoiksi**, koska tällöin mittaluvut kuvaavat jonkin ominaisuuden määrää.

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 käytettävissä olevia analyysimenetelmiä.

Esimerkki. Esimerkiksi keskiarvon laskeminen on eräs tilastollinen operaatio, ja se ei ole sallittu kvalitatiivisille muuttujille.

5.4 Kontrolloidut kokeet ja suorat havainnot

Tarkastellaan seuraavaksi kahta tyyppiesimerkkiä tilastollisen tutkimusaineiston keräämisen asetelmista:

Kontrolloidut kokeet

Kontrolloidut kokeet ovat sellaisia kokeellisia tutkimusasetelmia, joissa tutkimuksen kohteet altistetaan suunnitelmallisesti erilaisiin koeolosuhteisiin, jotta voidaan selvittää miten kohteet reagoivat muutoksiin.

Suorat havainnot

Suoria havaintoja kerätessä koeolosuhteita 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 tällöin mahdollista tarkastella tutkittavaa asiaa koeolosuhteissa eli ikäänkuin "eristyksissä" (muiden tekijöiden suhteen). Kontrolloidut kokeet eivät kuitenkaan ole aina mahdollisia, jolloin on käytettävä suoria havaintoja.

- Tällöin tutkimuskohdetta ei suunnitelmallisesti altisteta koeolosuhteille (''käsittelyille'') vaan muuttuvien olosuhteiden vaikutuksia tilastoyksikköihin seurataan passiivisesti.
- Toisin sanoen tutkimuksen kohteena olevat tilastoyksiköt eivät välttämättä edes tiedä osallistuvansa tutkimukseen.

Lisäksi usein tehdään hoito- ja käsittelyvastetta koskevia vertailuja erilaisissa olosuhteissa, mitkä osaltaan vaikuttavat tulosten uskottavuuteen, sillä tutkittaviin tilastoyksiköihin voi vaikuttaa olosuhteiden muutosten lisäksi muut ulkopuoliset tekijät.

- Näiden **selittävien** ja **sekoittavien tekijöiden** vaikutusten kontrollointi on suoria havaintoja tehtäessä vaativa tehtävä.

- Mikäli ulkopuolisia tekijöitä ei havaita ja/tai pystytä mittaamaan, tai muuten jostain syystä olla lisätty ja käytetty käytettävässä tilastollisessa mallissa, voi kyseeseen tulla ns. **puuttuvien selittäjien harha**.
  - Tämä tarkoittaa sitä, että syystä tai toisesta jotain keskeistä tekijää ei ole huomioitu tilastollisessa analyysissa.
  - Esimerkiksi havaittuihin tuloksiin vaikuttaa jokin havaitsematon tekijä, jonka vaikutusta ei kyetä kvantifioimaan puutteellisten havaintoarvojen vuoksi.

Kausaalisuus Suoria havaintoja tehtäessä ei voida (usein) selvittää vasteen ja olosuhteiden **kausaalista yhteyttä**. Suorilla havainnoilla voidaan lähinnä saada selville onko vasteella ja olosuhteilla jokin yhteys (korrelaatio).

Tilastolliset **kausaalisuhteet** ovat muuttujien välisiä syy-seuraus -suhteita, joita pyritään todentamaan tilastollisin menetelmin empiirisissä tutkimuksissa.

- **On syytä korostaa**, että kahden muuttujan välinen riippuvuus, kuten (lineaarinen) korrelaatio, **ei ole** suoraan tulkittavissa syy-seuraus -suhteeksi, sillä riippuvuus voi aiheutua esim. puuttuvasta kolmannesta muuttujasta.
- Kausaaliyhteyden toteaminen perustuu sekä teoriaan että muuttujien tilastolliseen käsittelyyn.
- Kausaalisuuden suuntaa on usein vaikeaa todentaa, koska kausaalisuuden matemaattiset kriteerit ovat hyvin tiukkoja.

Suorien havaintojen keräämiseen liittyy olennaisesti joitain riskejä ja toisaalta rajoituksia. Riskit liittyvät käytännössä otoksen harhaisuuteen, kuten erityisesti ns. valikoitumisharhaan. Tämä tilastoyksiköiden **valikoituminen** otokseen aiheuttaa harhaa, sillä otokseen valikoituva osajoukko saattaa ylikorostaa perusjoukon joitain ominaisuuksia. Näin tapahtuu esimerkiksi, jos havaintoja tehtäessä suositaan systemaattisesti joitakin tulosvaihtoehtoja. Tämä suosiminen voi olla tahallista tai tahatonta. Valikoituminen

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

- Esimerkiksi verrattaessa sydän- ja verisuonitautipotilaiden hoitotoimenpiteitä potilaat eivät mahdollisesti ole valikoituneet yhtä todennäköisesti pallolaajennukseen, ohitusleikkaukseen tai lääkehoitoryhmään, sillä taudin vakavuus saattaa jo määritellä mikä hoitotoimenpide valitaan.

- Valikoituminen on iso ongelma seurantatutkimuksissa, sillä harhaisten 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

Satunnaistaminen on 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.

Palataksemme vielä kontrolloituihin kokeisiin, niissä satunnaistaminen jakaa yksilöt riippumatta yksilön omista ilmiöön vaikuttavista muuttujista joko **käsittely- tai kontrolliryhmään** (eng. treatment ja control groups).

- Se takaa, ettei valikoitumista jonkin käsittelyä edeltävän ominaisuuden mukaan esiinny.
- Tämä tarkoittaa **altisteen** (käsittely, *treatment*) antamista (täysin) satunnaisesti kokeeseen valituille yksilöille, riippumatta näiden taustamuuttujien arvoista.
- Nämä yksilöt sinänsä voivat olla satunnaisotos jostain populaatiosta (tai ainakin niiden toivotaan olevan), mutta satunnaistaminen tarkoittaa siis käsittelyn kohdentamista koeyksilöille, ei satunnaisotantaa sinänsä.
- Esimerkiksi tutkittavat voidaan satunnaistaa lääkehoito- ja placeboryhmiin, jotta mahdolliset erot tutkittavien iässä, sukupuolessa ja muissa taustamuuttujissa eivät aiheuta systemaattista harhaa, kun tutkitaan lääkehoidon vaikutusta.

Satunnaistaminen (osaltaan) mahdollistaa tilastollisen päättelyn, jonka avulla otoksesta saatuja tietoja voidaan hyödyntää tehtäessä päätelmiä koko perusjoukosta. Johtopäätelmien pätevyys riippuu mm. siitä, kuinka hyvin otanta

on suoritettu. Tämän vuoksi on tärkeää ymmärtää otannan perusperiaatteet ja erilaisten otantamenetelmien luonne.

Chapter 6

Otannan idea

6.1 Otannan perusteet

Otantatutkimuksen (karkeat) suunnittelu- ja työvaiheet ovat mm. seuraavat:

1. Tavoitteiden asettaminen
2. Perusjoukon (populaation) asettaminen
3. Otantakehikko
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 määrää populaation.
- **Kohdeperusjoukko** on joukko, jota koskevia tietoja halutaan kerätä.

Esimerkiksi äänestysikäiset Suomen kansalaiset voivat muodostaa kohdeperusjoukon.

Usein tavoiteperusjoukko = kohdeperusjoukko. Tavoiteperusjoukko voi joskus olla laajempi (esim. “ihmiset” vs. “suomalaiset”).

Tutkimuksessa (edustavaan) otokseen poimitut tilastoyksiköt, näiden tilastolliset muuttujat ja niiden arvot muodostavat siis **otosaineiston** eli (otannalla saadun) **datan**.

- Tutkimuskysymykseen vastatakseen tutkija valitsee sopivan tilastollisen mallin ja estimoi sen parametrit poimittuun otokseen perustuen.
- Perusoletuksena ja tavoitteena on otoksen ja valitun tilastollisten mallin pohjalta suoritettavan tilastollisen päättelyn **yleistettävyys koko populaatioon**.
- Otos tuotetaan tilanteeseen sopivaa **otantamenetelmää** hyödyntäen pyrkien varmistamaan otoksen edustavuus (ts. perusjoukko pienoiskoossa).

Edustava otos

Edustavuus (lopulta tämän myötä muodostuva **edustava otos**). Tutkimukseen valitut yksiköt edustavat koko populaatiota. Ts. tutkimukseen valittu osajoukko kuvaa perusjoukon ominaisuuksia kattavasti.

Keskeistä tutkimuksen ja sen edustavuuden kannalta on, että tutkija osaa kerätä sisällöllisesti ja määrällisesti **sopivan kokoisen** aineiston. Tietyn otoksen edustavuutta arvioidessa voidaan käyttää apuna seuraavia kysymyksiä. Ts. **miksi päädyttiin tämän kokoiseen otokseen**?

Otoskoko **Otoskoko** vaikuttaa siihen miten hyvin otoksesta tehdyt johtopäätökset 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 tarkempaan valintaan ja määräämiseen palataan myöhemmin.
- Käytettiinkö apuna tilastotieteellisesti vankkaa suunnittelua otoskoon määrittämiseksi, vai/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. Kaiken kaikkiaan **edustavan otoksen** avulla on **mahdollista tehdä perusjoukkoa koskevaa tilastollista päättelyä**, sillä otos kuvaa perusjoukon ominaisuuksia riittävän hyvin. **Tämä on yksi tilastotieteen keskeisimpiä oppeja** mutta myös kriittisen tiedelukutaidon ja arkijärjen kannalta tärkeää.

6.2 Näytteistä ja otannan haasteista

Otantekehikko

Kehikkoperusjoukko on rekisterin tai luettelon tms. peittämä osa kohdeperusjoukkoa. Kyseessä on siis se osa kohdeperusjoukkoa, josta otanta ylipäänsä pystytään suorittamaan eli **otantakehikko**.

Ali- ja ylipeitto

Otantakehikon ali- ja ylipeitto. 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.

Vajaapeittävyys. Otantakehikon yli- ja alipeitto ovat nk. **kehikkovirheitä**. Lisäksi esimerkiksi kyselytutkimuksissa tai rekisteriaineistoissa saattaa esiintyä **katoa** (**vastauskatoa**, ks. alla), eli osa vastauksista jää uupumaan tai niitä ei jostain syystä mitata.

- Vajaapeittävyys voi esiintyä myös silloin kun populaation alkioista ei ole välttämättä täydellistä luetteloa.

Vastauskato ja vastausharha

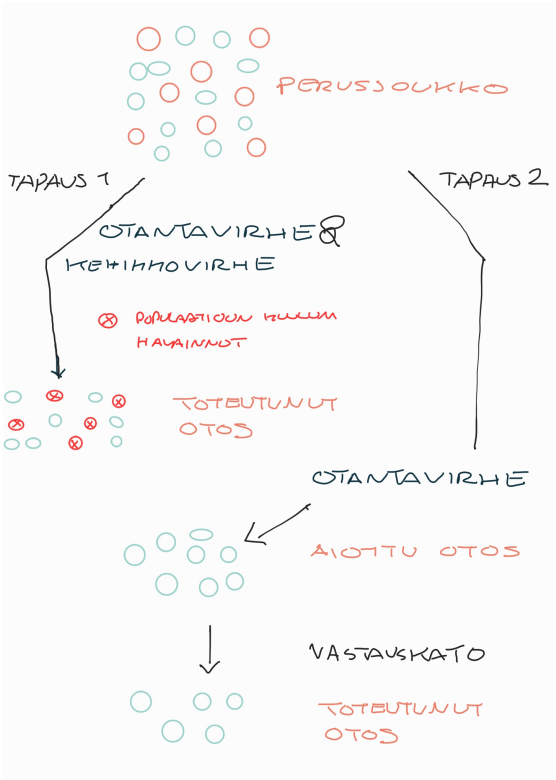
Vastauskato ja vastausharha.

- **Vastauskatoa** syntyy, kun tutkimuksen kohteita ei tavoiteta tai he kieltäytyvät vastaamatta. Kadon vuoksi lopullinen otoskoko saattaa jopa karsiutua pois tai jokin osajoukko on aliedustettuna.
- **Vastausharhaa** voi esiintyä, kun kysymykset voivat olla huonosti muotoiltuja tai vastaajat voivat antaa vääriä tietoja.

Otantavirhe

Otantavirhe on vastaavasti satunnaisuudesta johtuvaa tilastollisten muuttujien vaihtelua otoksesta toiseen, minkä tuloksena mm. otoksen perusteella lasketun tarkasteltavan estimaatin ja sen populaatiovastineen välillä on eroavaisuutta.

- Otantavirhe onkin ainoa virhelaji, jonka suuruutta voidaan tilastollisin menetelmin arvioida.
- Otantavirheen myötä on myös mahdollista, että tietyssä otoksessa havaittava tilastollinen systemaattisuus tai ylipäätään "signaalilta" (vs. kohina) vaikuttava informaatio voi johtua pelkästään satunnaisesta vaihtelusta.



Kuva: Otannan virhelajeja ja vaiheita kahdessa eri tapauksessa.

Esimerkki:Kotitalouksien tulot, tuloerot ja pienituloisuusrajan kehitys.

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

Otos vai näyte?

Joskus kohdeperusjoukko on sellainen, ettei siitä voi kerätä edustavaa otosta. Tämä voi johtua esimerkiksi siitä, että ei tiedetä keitä kohdeperusjoukkoon kuuluu eikä täten myöskään sen kokoa. Tällöin ei tietenkään voida myöskään suorittaa satunnaisotantaa.

Näyte

Poimintaharha ja näyte. Poimintaharhaa tapahtuu kun otos ei edusta populaatiota. Otos ei siis ole edustava otos. Jos tutkimukseen poimitaan ne tiedettyyn perusjoukkoon kuuluvat tilastoyksiköt, jotka sattuvat olemaan "saatavilla" tutkimuksen tekemishetkellä, tai ovat jopa itse valinneet itsensä otokseen, niin kyseessä on **näyte**. Näyte ei kata ilmiön koko vaihtelua edustavan satunnaisotoksen tapaan.

Koska **näyte ei ole edustava otos**, ei sen perusteella voida tehdä perusjoukkoa koskevia yleistyksiä luotettavasti.

Edustavan otoksen vaatimus on yksi tilastollisen tutkimuksen tavoitteista, joten näytteiden käyttö ei ole suositeltavaa!

Esimerkkejä näytteistä.

- Jos television ajankohtaisohjelma pyytää katsojia twiittaamaan mielipiteensä ajankohtaisesta asiasta, kyseessä on itse valikoituva näyte (osallistujat valitsevat itse itsensä).
- Esimerkiksi perinteiset katukyselyt eivät ole myöskään edusta erityisen hyvää otantatapaa, sillä kadulla liikkujat eivät välttämättä kovin hyvin edusta tutkittavaa perusjoukkoa, ellei perusjoukkona ole kyseisellä kadulla kyseiseen aikaan liikkuvat ihmiset.

- Vastaavasti surullisenkuuluisat X:ssä (ent. Twitter) toteutetut, tietyn henkilön omalle seuraajakunnalleen osoittamat kyselyt **eivät muodosta edustavaa otosta** edes X:n käyttäjistä, saati sitten koko kohdeperusjoukosta! (Ellei sitten kohdeperusjoukko ole juurikin kyseisen käyttäjän seuraajat, joskin näin kyseisiä kyselytuloksia harvemmin tulkitaan.)

Esimerkki: **Rikollisuuden tutkiminen harkinnanvaraisen näytteen avulla**. Kriminologian tutkija haluaa selvittää mitkä tekijät selittävät kesämökkimurtoja tehtailevien rikollisten motivaatiota rikoksilleen.

- Mikään rekisteri Suomessa ei kata kaikkia kesämökkimurtoja tehneitä ihmisiä, sillä osa näistä rikollisista ei jää koskaan kiinni.
- Täten tutkija ei voi muodostaa satunnaisotantaa kohdeperusjoukosta, vaan hän joutuu tyytymään esimerkiksi oikeuden pöytäkirjoista selviäviin kesämökkimurroista kiinnijääneisiin tilastoyksiköihin, joista hän valikoi tutkimukseensa sopivat yksilöt.
- Tällöin kyseessä on **harkinnanvarainen näyte**, sillä ei voida taata näytteeseen valikoitujen tilastoyksiköiden olevan edustava otos kohdeperusjoukosta!
- Harkinnanvarainen näyte voi kuvata kohdeperusjoukkoa hyvin, mutta on kuitenkin alisteinen tutkijan suorittamalle harkinnalle, joka voi olla tiedostamatta tai tiedostaen jollain tavalla vääristynyttä.

Chapter 7

Perustunnusluvuista

Perehdytään seuraavaksi muutamiin keskeisiin **tilastollisiin tunnuslukuihin** ja sitä myöden **estimaattoreihin** käyttäen tilastotieteessä tarvittavaa matemaattista notaatiota.

Tunnusluku

Tunnusluku (engl. statistic) tarkoittaa aineistosta laskettua lukuarvoa, joka kuvaa otoksen ominaisuuksia.

- Esimerkkejä tunnusluvuista ovat otoskeskiarvo, otosvarianssi ja mediaani.

Kun tarkastellaan satunnaismuuttujia, tunnusluku ei ole enää kiinteä luku vaan satunnaismuuttuja. Tämä johtuu siitä, että otosaineisto koostuu satunnaisista havainnoista, ja siten myös siitä laskettu tunnusluku vaihtelee otoksesta toiseen.

7.1 Perustunnuslukuja

Välimatka- tai suhdeasteikollisen muuttujan havaintoarvojen jakaumaa voidaan karakterisoida mm. seuraavilla perustunnusluvuilla, joita tyypillisesti raportoidaan tilastollisten analyysien yhteydessä.

- Havaintoarvojen keskimääräistä sijaintia kuvataan aritmeettisilla keskiarvoilla.
- Havaintoarvojen hajaantuneisuutta tai keskittyneisyyttä kuvataan otosvarianssilla tai otoskeskihajonnalla.

- Kahden muuttujan havaintoarvojen parien (lineaarista) riippuvuutta kuvataan otoskovarianssilla ja otoskorrelaatiokertoimella (ks. seuraava luku).

Ts. oletetaan seuraavassa, että meillä on käytettävissä välimatka- tai suhdeasteikollisen muuttujan havaittuja arvoja eli **otoskooltaan** n:n havainnon suuruinen otos havaintoja $(y_1, y_2, \dots, y_n)$. Käsitellään seuraavassa otoskeskiarvoa ja otosvarianssia.

Otoskeskiarvo

Havaintoarvojen (y_1, y_2,..., y_n) aritmeettinen **otoskeskiarvo** on

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(y_1 + \cdots + y_n).$$

Havaintoarvojen aritmeettinen keskiarvo kuvaa havaintoarvojen keskimääräistä arvoa. Osoittautuu, että (aritmeettinen) keskiarvo toimii tilastollisessa mielessä hyvänä estimaattorina satunnaismuuttujan (Y) odotusarvolle $\mathsf{E}(Y)$.

Esimerkki. Tarkastellaan tämän luvun läpi seuraavaa pientä tilastoaineistoa (otosta), jossa havaintoarvot y_i ovat 45, 67, 23, 89, 45, 56, 78, 34, 45, 67, 90, 12. Ts. $y_1 = 45, y_2 = 67$, ja lopulta $y_{12} = 12$ eli selvästikin otoskoko $n = 12$.

Otoskeskiarvo on tällöin

$$\bar{y} = \frac{1}{12}\sum_{i=1}^{12} y_i = \frac{1}{12}(45{+}67{+}23{+}89{+}45{+}56{+}78{+}34{+}45{+}67{+}90{+}12) = 54.25.$$

Otosvarianssi

Otosvarianssi: Havaintoarvojen $(y_1, y_2, \dots, y_n)$ otosvarianssi on muotoa

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

jossa $(\bar{y})$ on y-havaintoarvojen aritmeettinen otoskeskiarvo (määriteltiin yläpuolella).

Esimerkki, **jatkoa**. Jatketaan ylläolevan aineiston parissa. Otosvarianssi saadaan siis seuraavasti

$$s_y^2 = \frac{1}{11}\sum_{i=1}^{12}(y_i - 54.25)^2 = \frac{1}{11}\Big((45 - 54.25)^2 + \cdots + (12 - 54.25)^2\Big) = 618.75.$$

Vastaavalla tavalla mitä edellä voidaan määritellä esimerkiksi x havaintoarvojen otoskeskiarvo $\bar{x}$ ja otosvarianssi s_x^2. Havaintoarvojen varianssi mittaa havaintoarvojen hajaantuneisuutta tai keskittyneisyyttä havaintoarvojen aritmeettisen keskiarvon suhteen.

Otoskeskihajonta

Otoskeskihajonta: Havaintoarvojen (y_1, y_2,..., y_n) otoskeskihajonta

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

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

Esimerkki, **jatkoa**. Otoskeskihajonta saadaan siis otosvarianssista suoraviivaisesti

$$s_y = \sqrt{s_y^2} = \sqrt{\frac{1}{11}\sum_{i=1}^{12}(y_i - 54.25)^2} = \sqrt{\frac{1}{11}\Big((45-54.25)^2 + \cdots + (12-54.25)^2\Big)} \approx 24.87.$$

Jälleen vastaavalla tavalla voidaan määritellä x-havaintoarvojen otoskeskihajonta s_x. Havaintoarvojen keskihajonta (ja varianssi) mittaa siis havaintoarvojen hajaantuneisuutta tai keskittyneisyyttä havaintoarvojen aritmeettisen keskiarvon suhteen.

7.2 Otoskeskiarvo ja otosvarianssi estimaattoreina

Tilastollisessa tutkimuksessa on tavoitteena estimoida aineiston generoineen todennäköisyysjakauman tuntemattomat parametrit käyttäen havaintoaineistoa. Edellä esiteltiin jo havaintoaineistolle laskettavia tyypillisiä tunnuslukuja.

Miten luodaan tilastomatemaattinen yhteys edellisen luvun aineistoon ja sille laskettuihin konkreettisiin numeerisiin tunnuslukujen arvoihin? Seuraavassa muodostetaan tilastotieteelle keskeistä yhteyttä parametrien ja niitä arvioivien, "estimoivien", **estimaattorien** välille.

Tässä luvussa keskitytään lyhyesti vain otoskeskiarvoon ja otosvarianssiin normaalijakauman parametrien estimaattoreina ja siihen miten tämä näkökulma liittyy ja toisaalta eroaa edellä esitellyistä tietylle yksittäiselle otokselle laskettavista tunnusluvuista.

Merkitään satunnaismuuttujaa yhä Y (isolla kirjaimella) ja satunnaismuuttujan realisaatiota pienellä kirjaimella y.

- Otoskokoa, eli otokseen osallistuvien tilastoyksiköiden määrää, merkitään edelleen n:llä ja tilastoyksiköitä indeksoidaan alaindeksillä (i=1,...,n).

- Otoksen poimimisen jälkeen satunnaismuuttujat (Y_1, …, Y_n) saavat havaituiksi arvoikseen havaintoarvot (y_1, …, y_n) (ts. (Y_1=y_1, …, Y_n = y_n)). Näin jo edellä analysoitu yksittäinen havaintoaineisto on siis yksi **satunnaisotos** siitä jakaumasta, jota satunnaismuuttuja Y noudattaa.
- Otosvaihtelu Tämä tarkoittaa siis sitä, että kun otos poimitaan useita kertoja, niin saamme käytännössä aina erilaisen havaitun/realisoituneen satunnaisotoksen (y_1, …, y_n). Tätä vaihtelua kutsutaan myös **otosvaihteluksi**.

Satunnaisotos normaalijakaumasta. Olkoot (Y_1, …, Y_n) riippumattomia ja normaalisti jakautuneita satunnaismuuttujia, joille pätee ($Y_i \sim \mathsf{N}(\mu, \sigma^2)$), jossa ($\mu$) ja ($\sigma^2$) ovat normaalijakauman muodon määräävät parametrit. Parametrien (μ) ja (σ^2), eli normaalijakauman odotusarvon ja varianssin (ks. odotusarvon ja varianssin määritelmät), arvoja ei tunneta ja tavoitteena onkin päätellä, **estimoida**, niiden arvot käytettävissä olevaa aineistoa käyttäen.

Otoskeskiarvo. **Satunnaismuuttujien** (huom!) (Y_1, …, Y_n) **otoskeskiarvo** on

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

Yksittäisen otoksen otoskeskiarvo on tällöin sm:jien realisaatioiden aritmeettinen keskiarvo (ks. edellä ja esimerkki konkreettisesta otoksesta)

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

Otoskeskiarvo $\bar{Y}$ on siis satunnaismuuttuja, jonka saama arvo $\bar{y}$ vaihtelee satunnaisesti otoksesta toiseen johtuen satunnaisotannasta.

Voidaan osoittaa, että kun satunnaismuuttujat ovat samoin jakautuneet (noudattavat samaa jakaumaa) odotusarvoltaan ( ), on otoskeskiarvo jakauman odotusarvon harhaton estimaattori. Ts. pätee tulos

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

Täten aineiston otoskeskiarvo kuvaa aineiston perusjoukon tilastollisen mallin odotusarvoa. Ts. **otoskeskiarvo on estimaattori**, joka estimoi **harhattomasti** aineistoa generoivan todennäköisyysjakauman odotusarvoa!

- Huomioi, että satunnaisuudesta johtuen yksittäinen numeerinen otoskeskiarvo voi poiketa paljonkin jakauman odotusarvosta. Tästä huolimatta otoskeskiarvo on kuitenkin (tietyssä mielessä) paras mahdollinen arvio todellisen jakauman odotusarvosta.

Populaatiovarianssi on satunnaismuuttujan kautta ajateltuna on $\sigma^2 = \mathsf{E}[(Y-\mu)^2]$.

Populaatiovarianssia voidaan harhattomasti estimoida käyttäen **otosvarianssia** (satunnaismuuttujille)

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

Tämän yksi realisaatioista perustuu yksittäiselle aineistolle laskettuun otosvarianssiin (merkitään s^2 tai s_y^2), ks. edellä olevat kaavat.

Ts. **tämän (ala)luvun keskeinen huomio on**, että otoskeskiarvo ($\bar{Y}$) ja otosvarianssi (S^2) ovat siis satunnaismuuttujien kautta asiaa ajateltaessa satunnaismuuttujia, joiden saamat konkreettiset numeeriset arvot ($\bar{y}$ ja s^2) vaihtelevat satunnaisesti otoksesta toiseen, ja toimivat saatavilla olevasta aineistosta laskettuna kohdepopulaation jakauman parametrien estimaattoreina.

Estimaattoreihin, niiden ominaisuuksiin ja tarkempaan perusteoriaan perehdytään tarkemmin Osassa II.

7.3 Muita tunnuslukuja

Tilastollisia analyysejä tehtäessä johtopäätösten ja objektiivisten tulkintojen tueksi tarvitaan tunnuslukuja. (Otos)keskiarvoa ja (otos)varianssia tunnuslukuina 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 (kuten edellä nähtiin) keskiarvo ja varianssi tai keskihajonta (välimatka- ja suhdeasteikon havaintojen tapauksessa).

- Todetaan tässä kohtaa havaintojen hajaantuneisuutta mittaavista tunnusluvuista, että erityisesti järjestysasteikollisten muuttujien tapauksessa hajaantuneisuutta voidaan perustaa ns. **järjestystunnuslukuihin**. Ts. havaintoarvot voidaan järjestää suuruusjärjestykseen pienimmästä suurempaan. Ts. k. järjestystunnusluku on k. havaintoarvo suuruusjärjestyksessä. Jo aiemmin esitelty **prosenttipiste** liittyy kiinteästi suuruusjärjestykseen järjestettyyn aineistoon.

Esitellään seuraavassa vielä lyhyesti muutamia muita tunnuslukuja.

Moodi

Moodi eli tyyppiarvo on havaintoaineiston yleisin muuttujan arvo tai se on luokka, jolla on suurin frekvenssi.

Esimerkki, **jatkoa**. Jatketaan ylläolevan 12 havainnon esimerkkiaineiston käsittelyä. Ko. aineiston moodi on 45.

Mediaani

Mediaani on järjestetyn havaintoaineiston keskimmäinen arvo. Mediaani siis jakaa järjestetyn havaintoaineiston kahteen osaan siten, että puolet arvoista on mediaania pienempiä ja puolet arvoltaan mediaania suurempia.

- Jos havaintoarvoja on parillinen määrä, niin tällöin esitetään jompikumpi keskimmäisistä arvoista tai joskus niiden keskiarvo.

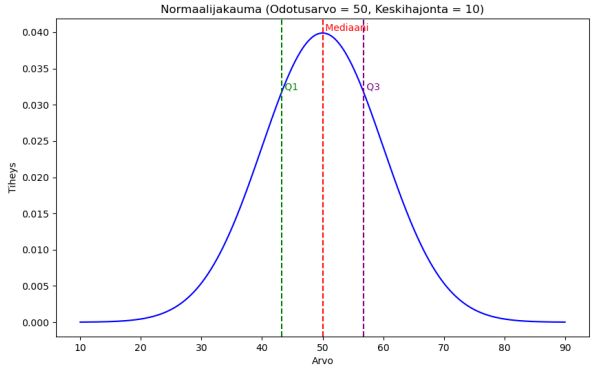
Esimerkki, **jatkoa**. Tarkasteltavan esimerkkiaineiston mediaania etsittäessä järjestetään havainnot suuruusjärjestykseen: 12, 23, 34, 45, 45, 45, 56, 67, 67, 78, 89, 90. Tällöin mediaanina voidaan pitää havaintoarvojen 45 ja 56 keskiarvoa eli 50.5.

Luokitteluasteikolla mitattaville muuttujille ei ole olemassa luontevia sijaintilukuja keskilukujen yhteydessä pl. moodi.

Fraktiilit ja desiilit

Järjestysasteikolla mitatuille muuttujille voidaan mediaanin lisäksi määrittää **fraktiileja**: pp %:n fraktiili jakaa tilastoaineiston kahteen osaan siten, että kyseistä fraktiilia pienempiä havaintoarvoja on pp %.

- Eniten käytettyjä fraktiileja ovat **kvartiilit**. **Alakvartiili** (Q_1) on 25 %:n fraktiili, ja **yläkvartiili** (Q_3) on 75 % fraktiili. Ts. alakvartiili on mediaania pienempien havaintoarvojen mediaani ja yläkvartiili on mediaania suurempien havaintoarvojen mediaani.
- Tietyistä fraktiileista käytetään nimitystä **desiili**. Ensimmäinen desiili on 10 % fraktiili ja esim. yhdeksäs desiili on 90 % fraktiili.



Kuva: Mediaani, ala- ja yläkvartiili esimerkkitilanteessa (kyseessä ei ole sama aineisto mitä tarkastellaan alapuolella!).

Esimerkki, **jatkoa**. Tarkasteltavan esimerkkiaineiston (ks. yllä) alakvartiili (Q1) on 42.25 ja yläkvartiili (Q3) 69.75.

Alakvartiili (Q1, 25 %):

$$Q_1^{\text{pos}} = 1 + (n-1) \cdot 0.25 = 1 + 11 \cdot 0.25 = 3.75$$

$$\begin{aligned} Q_1 &= y^{(3)} + 0.75 \cdot (y^{(4)} - y^{(3)}) \\ &= 34 + 0.75 \cdot (45 - 34) \\ &= 34 + 8.25 \\ &= 42.25, \end{aligned}$$

jossa $y^{(k)}$ on k:nnes järjestetty tarkasteltavan otoksen havaintoarvo (sen järjestetyssä jonossa).

Yläkvartiili (Q3, 75 %):

$$Q_3^{\text{pos}} = 1 + (n-1) \cdot 0.75 = 1 + 11 \cdot 0.75 = 9.25$$

$$\begin{aligned} Q_3 &= y^{(9)} + 0.25 \cdot (y^{(10)} - y^{(9)}) \\ &= 67 + 0.25 \cdot (78 - 67) \\ &= 67 + 2.75 \\ &= 69.75. \end{aligned}$$

Huom. ala-ja yläkvartiilin muodostuksessa voidaan käyttää muita/erilaisia valintoja mitä edellä ja siten lopputulos voi olla myös (hieman) erilainen.

Hajontalukuja: Varianssin/keskihajonnan lisäksi, jos muuttuja on mitattu vähintään järjestysasteikolla, sille voidaan määrittää vaihteluväli ja kvartiiliväli.

Vaihteluväli

Vaihteluväli kuvaa aineiston kokonaispeittoa ja siinä ilmoitetaan aineiston pienin havainto ja suurin havainto. Ts. vaihteluväli=(pienin havainto, suurin havainto). Voidaan myös laskea suurimman ja pienimmän havainnon välinen erotus.

Esimerkki, **jatkoa**. Tarkasteltavan esimerkkiaineiston tapauksessa vaihteluväli on selvästikin (12,90). Näiden välinen erotus 90-12=78.

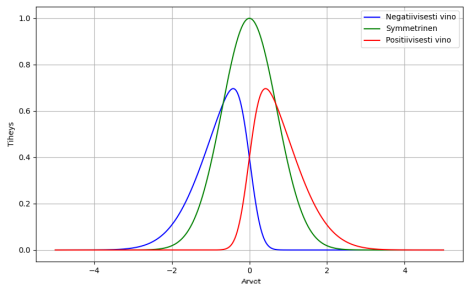
Vinous ja kurtoosi (ja ylihuipukkuus)

Satunnaisotoksille voidaan laskea myös niiden jakauman muotoa kuvaavat tunnusluvut **vinous** ja **kurtoosi** (ks. Alho ym., 2023), jota kutsutaan ajoittain myös **huipukkuudeksi**. Vinous ja kurtoosi/huipukkuus voidaan määrittää välimatka- ja suhdeasteikon muuttujille.

Vinous ja **kurtoosi**. Tarkastellaan satunnaismuuttujaa Y, jolla on odotusarvo $\mu = \mathsf{E}(Y)$ sekä keskusmomentit $\mu_k = \mathsf{E}[(Y-\mu)^k], k = 2, 3, 4$, keskihajonnan ollessa $\sigma = \sqrt{\mu_2}$ (ks. keskihajonnan määritelmä).

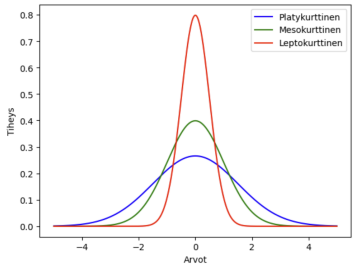
Standardoitua 3. momenttia μ_3/σ^3 kutsutaan (jakauman) **vinoudeksi** (eng. **skewness**).

- Normaalijakauma on symmetrinen jakauma ja sen vinous on siten 0.
- Vinoutunut jakauma (skewed distribution) kuvaa tilannetta, jossa aineiston jakauma ei ole symmetrinen, vaan se kallistuu voimakkaasti jompaan kumpaan suuntaan.



Vastaavasti standardoitua 4. momenttia $\gamma = \mu_4/\sigma_4$ kutsutaan **kurtoosiksi** (**huipukkuudeksi**, eng. **kurtosis**).

- Ajoittain kurtoosi määritellään vielä $\kappa = \gamma - 3$, josta käytetään englanniksi termiä **excess kurtosis** (suoraviivaisena käännöksenä **ylihuipukkuus**). Jos Y noudattaa normaalijakaumaa, niin tällöin $\gamma = 3$ ja siten $\kappa = 0$.
- Jakaumille joille $\kappa < 0$, $\kappa = 0$ ja $\kappa > 0$ käytetään englanninkielisiä termejä *platykurtic*, *mesokurtic* ja *leptokurtic*. Näiden suomenkieliset käännökset ovat (ks. Alho ym., 2023) *platykurttinen*, *mesokurttinen* ja *leptokurttinen*.



Vinous ja kurtoosi mittaavat siis kumpikin omalla tavallaan jakauman poikkeamaa normaalijakaumasta. Alho ym. (2023) ehdottavat termin kurtoosi käyttöä kun viitataan pelkästään standardoituun 4. momenttiin γ tai sen muunnokseen κ ja/tai ao. suureiden arvoihin. Synonyymeiksi soveltuvat *peakedness* ja *huipukkuus* silloin, kun puhutaan jakauman muodosta yleisellä tasolla ao. ominaisuuden tarkastelun perustuessa asianmukaisesti määriteltyyn järjestysrelaatioon ja sitä noudattavaan huipukkuusmittaan.

Vinoutuneet jakaumat ovat yleisiä esimerkiksi tulo- ja myyntidatassa, joissa pieni osa havaintoja voi vaikuttaa voimakkaasti kokonaisuuteen. Tällöin perinteiset tunnusluvut, kuten keskiarvo ja keskihajonta, voivat antaa harhaanjohtavan kuvan, koska ne ovat herkkiä poikkeaville arvoille. Siksi on tärkeää käyttää analyysimenetelmiä ja tunnuslukuja, jotka huomioivat vinoutuneisuuden – kuten mediaani ja kvartiiliväli. Tämänkaltainen datan monipuolinen tarkastelu auttaa paljastamaan aineiston mahdollisia yksityiskohtia, jotka muuten jäisivät piiloon tavanomaisia perusmenetelmiä käytettäessä.

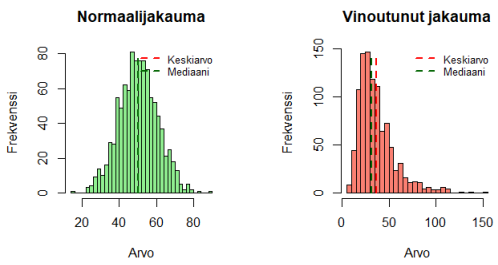
Esimerkki. Alla olevassa kuvassa on kuvattuna oikealla satunnaisotos jakaumasta, joka on oikealle vino ja huipukas.

- Vertailun vuoksi vasemmalla on myös normaalijakaumaa noudattava aineisto, jossa otoskeskiarvo ja mediaani ovat samalla kohdalla.

- Huomataan, että oikealle vinossa jakaumassa suuremmat realisaatiot ovat suhteellisesti todennäköisempiä kuin pienemmät. Kuvatun aineiston vinous on (1.4), joka tarkoittaa oikealle vinoa jakaumaa. Vastaavasti negatiiviset arvot vinoudelle kuvaavat vasemmalle vinoa jakaumaa.

- Vastaavasti huomataan että aineiston havainnot ovat verrattain keskittyneitä, eli että jakauma josta aineisto on peräisin on *huipukas/kurtoosinen*. Kuvatun aineiston kurtoosi saa yli 6:n olevan arvon, eli se on selvästi huipukkaampi kuin normaalijakauma (kurtoosi $\gamma = 3$). Tällöin on siis kysymys ylihuipukkuudesta (excess kurtosis) on yli kolmen.

- Vinoa ja huipukasta jakaumaa ilmentävät havaintoaineistot ovat tyypillisiä tutkimusta tehdessä ja niiden käsittely, mallintaminen ja niihin sovellettavat tilastollisen päättelyn menetelmät vaativat jälleen pidemmälle meneviä tilastotieteen opintoja!



Kuva: Esimerkkikuva symmetrisestä ja vinosta jakaumasta.

Vinoutuneisiin jakaumia (normaalijakaumaan verrattuna) ja ylipäätään jakaumien "**häntätodennäköisyyksien**" tärkeyttä korostaa myös Nassim Nicholas Taleb teoksissaan The Black Swan (2007) ja Antifragile (2012). Normaalijakaumaan perustuva ajattelu voi johtaa harhaan erityisesti epävarmoissa ja monimutkaisissa järjestelmissä. Mustat joutsenet Hänen mukaansa niin sanotut **mustat joutsenet** – harvinaiset, yllättävät mutta vaikutuksiltaan suuret tapahtumat – eivät ole poikkeuksia, vaan usein juuri ne, jotka muovaavat historiaa ja taloutta ratkaisevasti.

Taleb kritisoi riskimallinnusta, joka sivuuttaa paksuhäntäisten jakaumien merkityksen: ääripäät eivät ole vain tilastollisia poikkeuksia, vaan voivat hallita koko ilmiön käyttäytymistä. Hänen viestinsä on, että meidän tulisi suhtautua epävarmuuteen vakavasti ja rakentaa järjestelmiä, jotka eivät vain kestä sokkeja – vaan jopa hyötyvät niistä (antifragiliteetti).

Chapter 8

Tilastollinen riippuvuus, korrelaatio ja kausaalisuus

Tarkastelemme tässä luvussa tilastollisia tutkimusasetelmia, joissa on mukana **kaksi tai useampia muuttujia**. Pyrimme tässä luvussa vastaamaan (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 korrelaatiolla?
- Mikä on korrelaation ja riippuvuuden suhde?
- Miten korrelaatiota ja sen voimakkuutta voidaan estimoida?

Käsittelemme myöhemmin Osassa II **regressioanalyysia yhden selittäjän lineaarisen regressiomallin** tapauksessa. Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään koko tilastotieteen opinto-ohjelman lävitse, kuten perusteellisesti lineaarisia ja yleistettyjä lineaarisia malleja koskevilla kursseilla.

8.1 Muuttujien välisistä riippuvuuksista

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

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

Esimerkkejä riippuvuustarkasteluista:

- Miten työttömyysaste Suomessa (% työvoimasta) riippuu BKT:n (bruttokansantuotteen) kasvuvauhdista Suomessa, Suomen viennin volyymista sekä BKT:n kasvuvauhdista muissa EU-maissa ja USA:ssa? Taloustieteilijät pyrkivät yleisesti löytämään muitakin lainalaisuuksia.
- Millainen on riskin ja tuoton välinen suhde osakesijoittamisessa? Oletettavasti hajauttaminen pienentää riskiä ja/tai alhainen korkotaso suosii sijoittamista pörssiin.
- Miten alkoholin kulutus (l per capita vuodessa) riippuu alkoholijuomien hintatasosta, ihmisten käytettävissä olevista tuloista ja alkoholin saatavuudesta?
- Miten todennäköisyys sairastua keuhkosyöpään riippuu tupakoinnin määrästä ja kestosta?
- Miten vehnän hehtaarisato (t/ha) riippuu kesän keskilämpötilasta ja sademäärästä sekä maan muokkauksesta, lannoituksesta ja tuholaisten torjunnasta?
- Miten betonin lujuus (kg/cm2) riippuu sen kuivumisajasta?
- Miten kemiallisen aineen saanto (%) riippuu valmistusprosessissa käytettävästä lämpötilasta?

Tilastollinen riippuvuus

Eksakti vs. **tilastollinen riippuvuus**. Tarkastelemme tässä esityksessä yksinkertaisuuden vuoksi pääasiassa kahden muuttujan välistä riippuvuutta:

(i) Muuttujien välinen riippuvuus on **eksaktia**, jos toisen arvot voidaan mallintaa/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 mallintamisessa ja mahdollisesti myös ennustamisessa.

Korrelaatio

Tilastollinen riippuvuus ja korrelaatio. Kahden muuttujan välistä (lineaarista) tilastollista riippuvuutta kutsutaan tilastotieteessä **korrelaatioksi**.

- Korrelaation voimakkuutta mittaavia tilastollisia tunnuslukuja kutsutaan korrelaatiokertoimiksi.
- Korrelaatiot muodostavat perustan muuttujien välisten riippuvuuksien ymmärtämiselle.

Vaikka korrelaatiot muodostavat perustan muuttujien välisten riippuvuuksien ymmärtämiselle, riippuvuuksia halutaan tavallisesti analysoida vielä paljon tarkemmin ja formaalimmin valittavan tilastollisen mallin kautta. Tämä tapahtuu perinteisesti mm. regressioanalyysiä hyödyntäen.

- Viime vuosina erityisesti koneoppimisen ja tilastollisen oppimisen myötä myös muita kehittyneitä vaihtoehtoja on käsitelty kiihtyvällä tahdilla.

Regressioanalyysi

Regressioanalyysi on tilastollinen menetelmä, jossa jonkin ns. selitettävän muuttujan tilastollista riippuvuutta joistakin toisista ns. selittävistä muuttujista pyritään mallintamaan regressiomalliksi kutsuttavalla tilastollisella mallilla.

- Käsittelemme regressioanalyysiä koskevaa johdantoa myöhemmin Osassa II.

8.2 Kahden muuttujan havaintoaineiston kuvaaminen

Kuten yhden muuttujan havaintoaineistojen tapauksessa, lähtökohdan kahden tai useamman muuttujan havaintoaineistojen kuvaamiselle muodostaa tutustuminen havaintoarvojen jakaumaan.

- Havaintoarvojen jakaumaa kokonaisuutena voidaan kuvata sopivasti valituilla graafisilla esityksillä.

- Havaintoarvojen jakauman karakteristisia ominaisuuksia voidaan kuvata sopivasti valituilla otostunnusluvuilla (ks. aiempi ja myös myöhemmät luvut koskien otostunnuslukuja ja otosjakaumia).

Koska useampiulotteisten kuvioiden kuin kaksiulotteisten muodostaminen ei ole usein kovin mielekästä, kolmen tai useamman muuttujan havaintoaineistoja havainnollistetaan tavallisesti niin, että muuttujia tarkastellaan pareittain.

- Kahden järjestys-, välimatka- tai suhdeasteikoillisen muuttujan havaittujen arvojen pareja havainnollistetaan tavallisesti graafisella esityksellä, jota kutsutaan hajontakuvioksi tai **pistediagrammiksi** ("pistekaavio", engl. **scatter plot**).

- Ks. esimerkkikuva pistediagrammista allaolevasta isien ja poikien pituuksia koskevasta esimerkistä.

Pistediagrammi

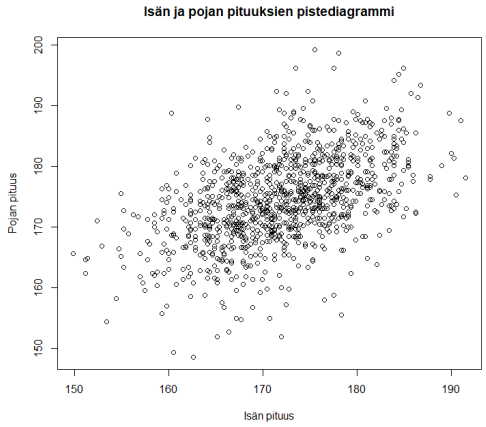
Pistediagrammi (*scatter plot*). Olkoot (X) ja (Y) järjestys-, välimatka- tai suhdeasteikollisia satunnaismuuttujia, joiden n kappaletta havaittuja arvoja ovat (x_1, x_2, ..., x_n) ja (y_1, y_2, ..., y_n). Oletetaan lisäksi, että havaintoarvot (x_i) ja (y_i) liittyvät samaan havaintoyksikköön kaikille (i = 1, 2, ..., n) eli tarkastellaan pareja (x_i, y_i), $i = 1, ..., n$.

- Havaintoarvojen parien (lukuparien) ((x_i, y_i)) pistediagrammi saadaan esittämällä lukuparit niiden määrittelemien pisteiden tasokoordinaatistossa.

- Pistediagrammi sopii erityisesti kahden muuttujan välisen riippuvuuden havainnollistamiseen. Se on keskeinen työväline korrelaatio- ja regressioanalyysissa.

Esimerkki: Isän ja pojan pituus. Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan. Periytyykö isän pituus heidän pojilleen?

Jo aiemmin tilastoaineiston esimerkkikohdassa viitattiin seuraavaan Karl Pearsonin klassikkoesimerkkiin 1078 havainnon aineistosta, joka sisältää isän ja heidän poikiensa pituuksia. Ts. aineisto koostuu lukuparista ((x_i , y_i)„ i = 1, 2, ..., 1078), jossa (x_i) = isän (i) pituus ja (y_i) = isän (i) pojan pituus. Yksittäinen piste (kuvassa "ympyrä") vastaa yhtä ko. lukuparien havaintoa. (Lähde: Kaggle: https://www.kaggle.com/datasets/abhilash04/fathersandsonheight.)



Kuva: Isän ja pojan 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 pitempiä poikia kuin lyhyillä isillä.

Tällaisten tilastollisten riippuvuuksien analysoimista lineaaristen regressiomallien avulla tarkastellaan myöhemmin. Ts. miten pojan pituutta voidaan mallintaa isän pituutta hyödyntäen.

Usean muuttujan havaintoaineistojen karakteristisia ominaisuuksia voidaan kuvata muuttujakohtaisilla otostunnusluvuilla. Muuttujakohtaiset otostunnusluvut eivät kuitenkaan voi antaa informaatiota muuttujien välisistä riippuvuuksista. Muuttujien parittaisia tilastollisia riippuvuuksia voidaan (usein ja osin) kuvata sopivasti valitulla korrelaation mitalla, mitä tarkastellaankin seuraavaksi.

8.3 Satunnaismuuttujien kovarianssi ja korrelaatio

Tarkastellaan välimatka- tai suhdeasteikollisten satunnaismuuttujien (X) ja (Y) (Pearsonin tulomomentti-) korrelaatiokerrointa (_{XY}) ja sen estimointia.

- Tällä kurssilla **emme tarkastele** tarkemmin mm. seuraavia tilastollisia testejä korrelaatiokertoimelle (_{XY}), kuten:
  - Yhden otoksen testi korrelaatiokertoimelle
  - Korrelaatiokertoimien vertailutestiä
  - Korreloimattomuuden testaamista
- Todetaan myös, että lisätietoja ja tarkempia yksityiskohtia moniulotteisista satunnaismuuttujista ja jakaumista tarkastellaan todennäköisyyslaskennan kursseilla.

Kovarianssi ja korrelaatio

Satunnaismuuttujien kovarianssi ja korrelaatio. Olkoon $((X, Y))$ satunnaismuuttujien (X) ja (Y) muodostama pari. Lisäksi

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

ovat ko. satunnaismuuttujien (X) ja (Y) odotusarvot ja

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

satunnaismuuttujien (X) ja (Y) varianssit.

Määritellään satunnaismuuttujien (X) ja (Y) välinen **kovarianssi** (σ_{XY}) kaavalla

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

Sm:jien (X) ja (Y) **korrelaatio** (ρ_{XY}) on vastaavasti

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

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

Satunnaismuuttujien (X) ja (Y) korrelaatiota

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

kutsutaan ajoittain siis **Pearsonin korrelaatiokertoimeksi** (tulomomenttikorrelaatiokertoimeksi).

- Pearsonin korrelaatiokerroin (ρ_{XY}) **mittaa satunnaismuuttujien** (X) **ja** (Y) **lineaarisen riippuvuuden voimakkuutta**. Ts. sm:jien välistä (lineaarista) yhteyttä.

- Pearsonin korrelaatiokerrointa voidaan estimoida otoksen pohjalta Pearsonin **otoskorrelaatiokertoimella**.

Otoskorrelaatio

Pearsonin otoskorrelaatiokerroin. Havaintoarvojen pareista (x_i, y_i), $i = 1, ..., n$, 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 otoskeskiarvot.

Otoskovarianssin (s_{xy}) avulla voidaan määritellä (x)- ja (y)-havaintoarvojen lineaarisen tilastollisen riippuvuuden voimakkuuden mittari eli **Pearsonin otoskorrelaatiokerroin** (r_{xy}), mikä saadaan otoskovarianssista (s_{xy}) normeerausoperaatiolla, jossa otoskovarianssi (s_{xy}) jaetaan (x)- ja (y)-havaintoarvojen keskihajonnoilla (s_x) ja (s_y):

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

Todetaan tässä kohtaa vain lyhyesti, että otoskorrelaatiokertoimen estimaattori voidaan johtaa sekä momenttimenetelmällä että suurimman uskottavuuden menetelmällä, jotka ovat tyypillisiä estimointimenetelmiä tilastotieteessä ja tarkemmin tilastollisessa päättelyssä.

Otoskovarianssin ominaisuuksia:

- Huomaa, että (x)- ja (y)-havaintoarvojen otoskovarianssit niiden itsensä kanssa ovat niiden variansseja.
- Otoskovarianssi (s_{xy}) mittaa (x)- ja (y)-havaintoarvojen yhteisvaihtelua niiden aritmeettisten keskiarvojen ympärillä.
- Otoskovarianssilla on taipumus saada positiivisia (negatiivisia) arvoja, jos havaintopisteiden muodostama "pistepilvi" ("pisteparvi)" näyttää nousevalta (laskevalta) oikealle mentäessä; ks. pistediagrammin ilmeen ja Pearsonin otoskorrelaatiokertoimen yhteys, jota käsitellään seuraavaksi.

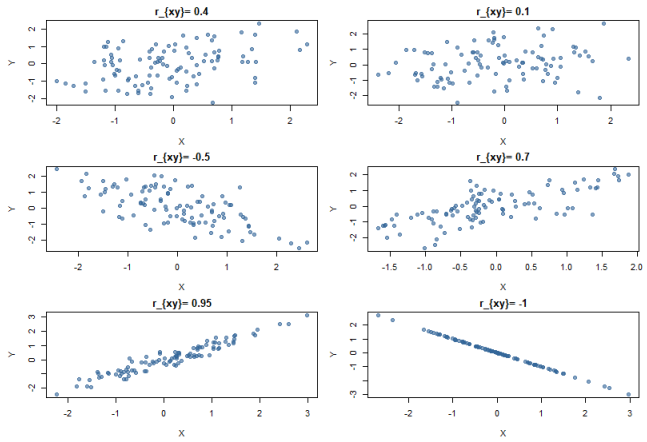
Pearsonin otoskorrelaatiokertoimella (r_{xy}) on seuraavat ominaisuudet:

i) $(-1 \leq r\_xy \leq 1)$
ii) $(r_{xy} = \pm 1)$, jos ja vain jos $(y_i = \alpha + \beta x_i)$, jossa (α) ja (β) ovat reaalisia vakiota ja $(\alpha, \beta \neq 0)$.
iii) Korrelaatiokertoimella (r_{xy}) ja kovarianssilla (s_{xy}) on aina sama etumerkki

Näiden ominaisuuksien (i)-(iii) perusteella voidaan tehdä seuraavia huomioita ja tulkintoja:

- Havaintoarvojen pareista $((x_i, y_i), i = 1, 2, \ldots, n,)$ laskettu Pearsonin otoskorrelaatiokerroin (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.
- Karkeasti ajatellen voidaan todeta, että jos $|r_{xy}| > 0.8$, niin korrelaatio on voimakasta, välillä 0.6-0.8 huomattavaa, ja tätä ennen kohtalaista tai lähes merkityksetöntä kun $r_{xy} \approx 0$.
- Vaikka $(r_{xy} = 0)$, (x)- ja (y)-havaintoarvojen välillä saattaa silti olla jopa eksakti **epälineaarinen riippuvuus**.

Esimerkki: Alapuolella esitettävät (simuloidut) kuviot havainnollistavat kahden muuttujan havaittujen arvojen $((n = 100))$ pistediagrammin ilmeen ja korrelaation välistä yhteyttä.



Kuva: Havainnollistuksia Pearsonin otoskorrelaatiokertoimen arvosta ja erilaisista xy-pisteparvista.

Ks. seuraavasta linkistä lisää havainnollistuksia: *Guess the correlation* pelissä pääset arvioimaan esitettävän pisteparven korrelaation voimakkuutta erilaisissa simuloiduissa tilanteissa: http://guessthecorrelation.com/

Kausaalisuuden ehdot

Kausaalisuus. Muuttujan (x) arvojen muutos vaikuttaa muuttujan (y) arvoihin (syy-vaikutussuhde), jos seuraavat kolme aika yleiselle tasolle tuotua ehtoa täyttyvät:

- Muuttujan (x) muutos esiintyy ajallisesti ennen (y):n muutosta.
- Muuttujissa (x) ja (y) tapahtuvien muutosten välillä on tilastollista riippuvuutta (korrelaatiota tai muuta riippuvuutta).
- Ei sekoittumista. Muuttujassa (y) tapahtunutta muutosta ei voida selittää millään muilla tekijöillä.

Kausaalisuhteita on vaikea todentaa tilastollisesti, mutta hyvin suunnitellut satunnaistetut kokeet ovat usemmiten paras käytettävissä oleva viitekehys tämänkaltaisen tutkimusasetelman saavuttamiseksi, mikäli se on ylipäätään ko. sovelluksen kohdalla mahdollista.

- Satunnaistetut kokeet vähentävät harhan ja sekoittavien tekijöiden vaikutusta, mikä tekee niistä luotettavimman tavan osoittaa kausaalisuhteita.

Käytännössä kausaalisuhteita selvitettäessä on tunnettava etukäteen ilmiötä koskevat aiemmat teoriat ja tutkimukset tarkasti, jotta voidaan ottaa huomioon ilmiöön vaikuttavat tekijät.

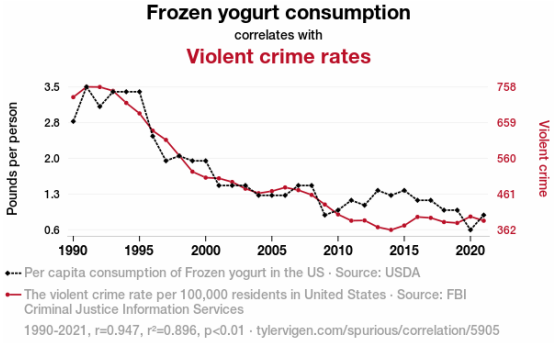
- Todellisuus (ja mahdollinen kausaalisuussuhde) on usein monimutkaisempi, kuin mitä tällainen yksinkertainen (kahden muuttujan välinen) analyysi kykenee kuvaamaan:

Voidaan siis todeta, että **kahden muuttujan yhteisvaihtelu ei riitä todisteeksi siitä, että kyseessä olevien muuttujien välillä on kausaalista yhteyttä**.

Tilastollisia menetelmiä, kuten regressioanalyysiä ja siinä yhteydessä erilaisten selittävien muuttujien myötä tapahtuvaa niiden vaikutuksen kontrollointia voidaan käyttää ko. tekijöiden huomioon ottamiseksi, mutta aina tarvitaan harkintaa sen suhteen, kuinka luottavaisesti kausaalisuudesta voidaan väittää jotain. Tämä johtuu (osaltaan) siitä, että tilastolliset mallit perustuvat oletuksiin, jotka eivät aina pidä paikkaansa todellisessa maailmassa.

Esimerkki. Alla olevassa kuvassa on esimerkki todella vahvasta yhteisvaihtelusta jäädykejogurtin (eng. frozen yogurt) kulutuksen ja väkivaltarikosten välillä Yhdysvalloissa. Kyseisten havaintojen korrelaatioksi vuosien 1990-2021 välillä saadaan ($r_{xy} = 0.947$), joka on erittäin suurta, kun huomioidaan että kyseiset muuttujat eivät tunnu liittyvän toisiinsa mitenkään.

Kyseessä onkin nk. “näennäinen korrelaatio”, eli tilanne, jossa kaksi muuttujaa korreloi vahvasti vailla mitään syytä. Näin voi tapahtua täysin sattumalta! Voit etsiä lisää esimerkkejä Tyler Vigenin tarjoamalta Spurious Correlations -verkkosivulta!



Kuva: Esimerkki nk. näennäisestä korrelaatiosta (eng. spurious correlation). Lähde: Tyler Vigen.

Yhteisvaihtelu voi johtua myös kolmannen muuttujan vaikutuksesta molempiin muuttujiin tai virheellisestä otannasta, vaikka muuttujat olisivatkin perusjoukossa toisistaan riippumattomia.

- Klassinen esimerkki tällaisesta "puuttuvan muuttujan harhasta" on hukkumiskuolemien ja jäätelön kulutuksen näennäinen yhteys, jos tarkastellaan vain niiden välistä korrelaatiota.
- Tosiasiassa molempia selittää lämpimät kelit, jolloin ihmiset uivat enemmän, mutta myös syövät enemmän jäätelöä!

Esimerkki: Simpsonin paradoksi: U.C. Berkeleyn sukupuolisyrjintä. Simpsonin paradoksilla tarkoitetaan tilannetta, jossa kahden muuttujan välinen korrelaatio muuttuu päinvastaiseksi otettaessa huomioon jokin kolmas muuttuja, joka korreloi molempien muuttujien kanssa.

- Yksi tunnetuimmista esimerkeistä Simpsonin paradoksista on Berkeleyn yliopiston sukupuolisyrjintätapaus.
  - Yliopisto haastettiin oikeuteen vuonna 1973 sukupuolisyrjinnästä.
  - Väitettiin, että yliopistoon olisi miesten helpompi päästä 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 huomattiin, että itse asiassa useammassa tiedekunnassa naisia on päässyt sisälle isompi osuus hakijoista.
  - Tämä johtui siitä, että naiset hakivat opiskelemaan aloja, joille sisäänpääsystä käytiin kovempaa kilpailua. Toisin sanoen, naisten hakemukset keskittyivät aloille, joilla oli vähemmän aloituspaikkoja.

– Aineisto kuudesta isoimmasta tiedekunnasta on listattu alla olevaan taulukkoon.

Tiedekunta	Hakijat (Miehet)	Hyväksytyt % (Miehet)	Hakijat (Naiset)	Hyväksytyt % (Naiset
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 siis tiivistäen **korrelaatiokertoimeen liittyviä tulkintavirheitä** aiheuttavat useimmiten seuraavat seikat:

- Riippuvuudesta ei välttämättä seuraa syy-seuraussuhdetta.
- Kolmas muuttuja eli kahden muuttujan välinen yhteys selittyy yhteisestä syystä (esimerkiksi lämpimästä kesästä).
- Muuttujien välinen yhteys ei ole lineaarinen (on epälineaarinen).
- Poikkeavien havaintojen vaikutus.

Puutteita: Korrelaatiokertoimella on kaksi puutetta:

- Se mittaa vain lineaarista riippuvuutta.
- Se ei ole (tilastollinen) malli, jonka avulla nähtäisiin, miten toinen muuttuja vaikuttaa toiseen muuttujaan.

::::::::

Chapter 9

Osan I yhteenvetoa signaalin ja kohinan näkökulmasta

Palataan vielä Osan I lopuksi yhteenvetoon tähän mennessä opitusta. Kuten todettua, tilastollisessa tutkimuksessa mielenkiinnon kohteena on satunnaisilmiöiden tutkiminen ja erityisesti **systemaattisen ja satunnaisen vaihtelun** eli **signaalin ja kohinan erottaminen toisistaan**. Tämä saattaa näyttäytyä selvimmin (tähän mennessä opitun pohjalta) edellä käsitellyssä kahden muuttujan (lineaarista) riippuvutta koskevassa kohdassa. Ylipäätään siis **muuttujien välisten riippuvuuksien tutkiminen** on keskiössä kun pyrimme löytämään signaalin tai signaaleja ja erottamaan ne puhtaasta kohinasta.

Signaali ja kohina

Signaali ja kohina (tilastotieteessä) perustuu ajatukselle, että havaittu aineisto (= data) koostuu kahdesta komponentista: (deterministisestä) signaalista, joka on se, mistä olemme erityisesti kiinnostuneita, ja satunnaisesta kohinasta, joka muodostaa satunnaisen jäännösvirheen. Tilastollisen analyysin (ml. tilastollisen päättelyn) haasteena on tunnistaa nämä kaksi oikein ja **olla erehtymättä luulemaan kohinaa signaaliksi**.

On syytä vielä korostaa, että kun dataa analysoidaan signaalit tulevat aina kohinan kanssa. Tämän vuoksi niiden erottaminen toisistaan tekee aiheesta mielenkiintoisen! Havaittu data sisältää sekä todellista informaatiota (signaalia) että satunnaista vaihtelua (kohinaa).

- Väistämätön satunnaisvaihtelusta johtuva vaihtelu (kun kyse on satunnaisilmiöistä!) vaikeuttaa tätä tunnistamisoperaatiota.

- Tästä vaihtelusta johtuen rakennamme tilastollisia malleja, jotka ovat hyödyllisiä abstraktioita tarkasteltavalle ilmiölle. Mallit auttavat kuvaamaan ja ymmärtämään datan vaihtelua ja epävarmuutta tarjoamalla matemaattisen kehyksen, jonka avulla voidaan tehdä ennusteita ja arvioida, kuinka hyvin mallit sopivat havaittuun aineistoon.

Esimerkki. Elämänviisautta pohdittavaksi signaalin kehittymisestä suhteessa kohinaan (ts. miten signaali ja kohina näyttäytyvät tässä?):

"*I'm greater believer in luck, and I find the harder I work the more I have of it*"

(Thomas Jefferson, Yhdysvaltojen 3. presidentti 1801–1809)

Soveltavassa tilastotieteessä kiinnostuksen kohteena on siis hyvin harvoin vain jokin yksittäinen tunnusluku, kuten keskiarvo, varianssi tai korrelaatio. Kysymys on lähes aina useiden tunnuslukujen ja viime kädessä formaalin tilastollisen mallin täsmentämisen kautta pyrkimyksestä mahdollisimman hyvään signaalin erottamiseen kohinasta.

- Tähän tavoitteeseen pyrittäessä olemme nyt nähneet edellä ensimmäisiä ja alustavia askeleita, mutta matkaa on vielä runsaasti jäljellä tämän vaiheen jälkeen!

Tieteen popularisointi on yksi tutkijoiden ja yliopistojen tiedeyhteisön tärkeimmistä yhteiskunnallisista tehtävistä, mutta valitettavan usein se typistyy yksittäisen viimeisimmän tutkimustuloksen esittelyksi. Kuinka usein kyseessä on lopulta lähinnä kohinasta signaalin sijaan?

- Yliopistoyhteisössä kuitenkin, luonnollisesti, luotamme kumuloituneeseen tutkittuun tietoon ja tiedämme, kuten luvuissa tähän asti olemme käyneet lävitse, että **yksittäinen tutkimus on vasta hyvä alku**.
  - Ihmistieteitä, kuten ilmeisesti erityisesti psykologiaa sekä osin myös mm. lääke- ja taloustiedettä, on viimeisen vuosikymmenen ajan puhuttanut paljon ns. **replikaatiokriisi**, sillä useaa arvostettuakaan tutkimusta ei ole saatu **toistettua eli replikoitua**. On ymmärrettävää, että replikaatiokriisi, varsinkin jos se on (alakohtaisesti) laajalle levinnyttä, murentaa kansalaisten luottamusta tieteellisiin tuloksiin.
  - Toistettavuus on yksi tutkimuksen peruskriteereistä, joka erottaa tieteellisen tiedon muista tietolähteistä, joten 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, joten silkka sen olemassaolo on osoitus tieteellisen prosessin toimivuudesta.

- Kun tuntee ja tunnistaa sattuman voiman ja ymmärtää mahdolliset satunnaisuuden lähteet, jotka altistavat tutkimusprosessin virheille, tulee samalla ymmärtäneeksi että eri tavoin koeteltu, useassa tutkimuksessa kumuloitunut tieto tulisi olla kaiken tieteen popularisoinnin keskiössä yksittäisten, mahdollisesti uusien ja yllättävien tutkimustulosten sijaan.

Tähän mennessä olemme jo oppineet, että tälle (tutkimustulosten vaihtelulle ja kehittymiselle) on myös vahvat tilastolliset perustelut: **satunnaisen tiedon maailmassa mikään ei ole täysin varmaa**, ei edes kaikkein edistyneimpien tilastomenetelmien avulla!

- Erityisesti nykypäivänä ei-tieteellinen tieto ja tarkoituksellinen **disinformaatio**, joita perustellaan heppoisin havainnoin, leviävät internetissä kulovalkean tavoin.

- Voidaan sanoa, että on tiedeyhteisön ja tutkijoiden moraalinen vastuu taistella näitä uskomuksia vastaan **popularisoimalla tiedettä**.

Tieteen (liiankin vahvat) popularisointipyrkimykset saattavat kuitenkin ajoittain jopa pahentaa ongelmaa, sillä tutkijoilta voi unohtua **satunnaisuuden voima**. Eli se, mitä olemme osaltaan hahmotelleet edellä, kuten millä tavalla tilastoaineistot muodostuvat ja mitä mahdollisia ongelmia näihin vaiheisiin voi liittyä.

Chapter 10

Otannan tilastollisia perusteita

OSA II

Tässä luvussa jatketaan otannan tarkastelua. Erityisesti tarkastellaan miten satunnaisotos kohdepopulaatiosta voidaan käytännössä kerätä. **Otantamenetelmiksi** kutsutaan niitä menetelmiä, joilla kohdepopulaatiosta **edustava otos** kerätään.

Satunnaisuuden hyödyntäminen näyttelee keskeistä roolia otoksen edustavuuden varmentamisessa. Kohdepopulaation koko, maantieteellinen jakauma, ikärakenne ja muut yksityiskohdat vaikuttavat kuitenkin keskeisellä tavalla siihen, miten otanta tulee käytännössä suorittaa.

10.1 Otantamenetelmät

Tässä jaksossa tarkastellaan erilaisia **otantamenetelmiä**. Näiden menetelmien tarkoitus on suorittaa otosaineiston (tutkimusaineiston) kerääminen niin, että se huomioi aiemmin esitellyt hyvän otannan kriteerit, ts. että sen tuottama otos on edustava ja luotettava. Näin ollen otos kuvaa koko perusjoukkoa.

- Otantamenetelmän, joskus myös **otanta-asetelman**, valinta on tietenkin vahvasti sovellusalakohtainen: käytettävät aineistot ja täten otantamenetelmät määräytyvät 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 **sisältymistodennäköisyydeksi**.

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

Käytännössä otoksen poiminta suoritetaan niin, että n:n alkion otos (n on tässä siis edelleen otoskoko) poimitaan jollakin satunnaisotannan menetelmällä N:n alkion perusjoukosta (N on siis edelleen perusjoukon koon merkintä).

- Perusjoukon yksittäinen alkio (tilastoyksikkö) k tulee poimituksi n:n alkion otokseen (tutkimusaineistoon) tunnetulla **sisältymistodennäköisyydellä** π_k,

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

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

- Sisältymistodennäköisyys voi olla sama kaikille perusjoukon alkioille tai vaihdella perusjoukon eri osajoukkojen (alkioryhmien) välillä. Tämä tulee huomioida otantamenetelmän valinnassa, jotta saadun otoksen edustavuus ei vaarannu.

- Sisältymistodennäköisyyttä voidaan käyttää monimutkaisemmassa otantateoriassa **asetelma-** ja **analyysipainojen** muodostamisessa sekä uudelleenpainotuksessa (vastauskadon korjaus).

Tässä luvussa esitellään seuraavaksi erilaisia perinteisiä otantamenetelmiä (niiden paikoin yksinkertaistettuja kuvauksia ja toteutuksia) sekä sitä, minkälaisten perusjoukkojen tilanteissa mikäkin otantamenetelmä on sopivin.

- **Yksinkertainen satunnaisotanta** (YSO): perinteisin otantamenetelmä, jossa jokaisella tietyn kokoisella otoksella sama mahdollisuus tulla valituksi.
- **Systemaattinen otanta** (SYS): eli tasavälisessä otannassa poimintakehikkoon (perusjoukkoon) kuuluvat alkiot järjestetään jonoon ja siitä poimitaan otokseen joka k. alkio.

- **Ositettu otanta**: perusjoukko (populaatio) jaetaan ominaisuuksiltaan yhtenäisiin eli homogeenisiin **ositteisiin**, joista jokaisesta poimitaan erillinen otos.
- **Ryväsotanta** tai joskus myös **moniasteinen otanta**: hyödynnetään perusjoukossa esiintyvää kerroksellisuutta eli hierarkkisuutta otannassa.

10.2 Yksinkertainen satunnaisotanta

Yksinkertainen satunnaisotanta

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

YSO:ssa otannan satunnaisuus tulee siis siitä, että jokainen tilastoyksikkö poimitaan otokseen *satunnaisesti*!

- YSOaa pidetään otannan perusmuotona, jossa jokaisella perusjoukon alkiolla on lähtökohtaisesti yhtä suuri todennäköisyys tulla valituksi otokseen.
- YSO on periaatteiltaan 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 satunnaistamalla otokseen valikoitumisen täysin! YSO voidaankin aina tulkita arvonnaksi. Käytännön työssä arvonta onkin oiva satunnaistamisen keino.

YSO:n toteuttaminen. Käytännössä yksinkertainen satunnaisotanta etenee vaiheittain:

- Tutkimuksen alussa tutkijalla tulisi olla käytettävänään (ts. tulisi koostaa) **lista** kaikista perusjoukon havaintoyksiköistä **(alkioista)**. Tämä muodostaa tutkimuksen **otantakehikon**.
- Tämän jälkeen jokaiseen perusjoukon alkioon voidaan liittää numeeriset tunnukset.
- Sitten valitaan haluttu otoksen koko. Otoskoon määrittäminen on keskeinen osa koesuunnittelua.

- Otantakehikosta arvotaan perusjoukon alkiot otokseen yksi kerrallaan.
- Käytännössä arvonta voidaan toteuttaa satunnaislukuja generoimalla (tuottamalla) niin että jokaisen otantakehikon alkion sisältymistodennäköisyys on yhtä suuri.[1]

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 tarkoituksena on poimia n:n alkion kokoinen otos (huom. $n < N$). Olkoon (i) yksittäisen alkion indeksiluku (ts. jokainen alkio on numeroitu esimerkiksi tavalla $i = 1, \ldots, N$).

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 alkiota otokseen on jo poimittu.
- Alkion k **poimintatodennäköisyys** poimintakierroksella/"otositeraatiolle" j ($j = 1, \ldots, n$, jossa n on tavoiteltava otoskoko) on tällöin selvästi

$$\frac{1}{N} \quad \forall\, k = 1, \ldots, N,$$

  eli sama todennäköisyys kaikille alkioille $k = 1, \ldots, N$.
- Otantaan palauttaen liittyviä todennäköisyyksiä hallitaan **binomijakauman** avulla, joka johtaa yksinkertaiseen **tilastolliseen malliin** YSO:a käytettäessä.
- Poiminta palauttaen (otanta takaisinpanolla) on toisaalta usein varsin epärealistinen otantamenetelmä eri tutkimustilanteissa ja sovelluksissa.

Esimerkiksi lienee mahdotonta testata samaa lääkettä useaan otteeseen samaan aikaan yhdellä koehenkilöllä.

YSO:n poiminta palauttamatta. Kun poiminta suoritetaan **palauttamatta**, poimittua alkiota ei palauteta perusjoukkoon poiminnan jälkeen eikä se täten voi tulla poimituksi otokseen kuin kerran.

[1] Satunnaislukujen generointia käsitellään ja opetellaan mm. kursseilla TILM3517 R-kielen alkeet ja TILM3705 Johdatus laskennalliseen tilastotieteeseen.

- 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 k **poimintatodennäköisyys** on tällöin vastaavasti (niin kauan kun alkio ei ole jo poimittu otokseen) poimintakierroksella j ($j = 1, \ldots, n$)

$$\frac{1}{N - A_k},$$

ja kun k:s alkio tulee poimituksi, niin tämän jälkeen poimintatodennäköisyys on 0. Tässä A_k on jo poimittujen alkioiden lukumäärä ennen kyseistä poimintakierrosta/otositeraatiota: ensimmäisen poiminnan kohdalla $A_k = 0$, toisen kohdalla $A_k = 1$ ja niin edespäin. Ts. $A_k = j - 1$.

- Ilman takaisinpanoa populaatiosta voidaan poimia $\binom{N}{n}$ erilaista otosta.
- Kun otosyksiköiden järjestyksellä ei ole merkitystä. $\binom{N}{n}$ on ns. binomikerroin, joka saadaan kaavasta $\frac{N!}{n!(N-n)!}$, jossa $N! = N \cdot (N-1) \cdot (N-2) \cdots 1$ on N:n kertoma.
- Otantaan palauttamatta liittyviä todennäköisyyksiä hallitaan **hypergeometrisen jakauman** avulla, joka johtaa (melko) yksinkertaiseen **tilastolliseen malliin** YSO:a käytettäessä.

Esimerkki yksinkertaisen satunnaisotannan poimintastrategioista. 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 1:tä) menetelmät poikkeavat olennaisesti.
- Toisaalta, jos perusjoukko on ääretön, niin menetelmillä ei ole käytännössä eroa
  - Toisin sanoen, kun $(N \longrightarrow \infty)$, niin $(\frac{n}{N} \longrightarrow 0)$ eli todennäköisyys että sama alkio poimittaisiin otokseen useammin kuin kerran lähestyy nollaa otoskoon lähestyessä ääretöntä.

- Monesti onkin (teoreettiselta) kannalta järkevää olettaa että otos poimitaan äärettömästä perusjoukosta vaikka perusjoukko tosiasiallisesti olisikin äärellinen, mutta siis riittävän "iso".
- Tällöin voidaan olettaa käytettävän otantaa takaisinpanolla, sillä siinä käytettävät tilastolliset mallit ovat yksinkertaisempia kuin otannassa ilman takaisinpanoa ja tämä helpottaa tilastollisessa päättelyssä käytettäviä kaavoja.

YSO:n potentiaaliset ongelmat

- Monissa tapauksissa ei kuitenkaan ole helppoa saada listaa kaikista perusjoukon havaintoyksiköistä (jolloin menetelmän käyttö on periaatteessa mahdotonta).
- Kyselytutkimuksissa perusjoukko on usein suuri ja laajalle alueelle hajaantunut. Henkilökohtaisten, kasvotusten toteutettavien, haastattelujen tekeminen vaatisi suuria resursseja. Haastattelijat joutuisivat esim. matkustamaan ympäri Suomea satunnaisotokseen valikoituneiden henkilöiden asuinpaikkojen mukaan. Tällaisissa tutkimustilanteissa voidaankin käyttää muunlaisia otantamenetelmiä.

10.3 Systemaattinen otanta

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 järjestetystä 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 ominaisuudesta, jota tutkitaan.

Mahdolliset ongelmat: Systemaattisen otannan suhteen potentiaaliseksi ongelmaksi muotoutuu havaintoyksikkölistan mahdollinen säännöllinen jaksollisuus, jota systemaattinen poiminta ei havaitse ja jolloin satunnaisotanta toimisi (kenties) paremmin.

- Ongelmia syntyy esimerkiksi silloin, jos tiedot perusjoukosta koostuvat heteropariskunnista ja poimintaintervalli on parillinen luku. Tällöin seurauksena voi olla, että otokseen saattaisi valikoitua ainoastaan joko miehiä tai naisia.
- Myös systemaattisessa otannassa tarvitaan siis lista tai rekisteri kaikista perusjoukon havaintoyksiköistä ja sitä sovelletaankin tavallisesti YSO:n sijasta silloin, kun perusjoukon alkioista on käytettävissä tietorekisteri, luettelo tai havaintoja kerätään ajassa tai tilassa.

Esimerkkejä

- Esimerkiksi mielipidekyselyn kohteet poimitaan (voitiin poimia) puhelinluettelosta (tai vastaavasta rekisteristä) valitsemalla haastateltavaksi jokaiselta aukeamalta ensimmäisenä esiintyvä henkilö tai jotain tuotetta valmistavan tehtaan laaduvalvonnassa valitsemalla laatuarviointiin joka sadas tuote, joka hihnalta valmistuu.
- Muita esimerkkejä ovat esim. liikenne-, jäsenrekisteri- tai kassajonossa seisovien otantayksiköiden poiminta otokseen.

10.4 Ositettu otanta

Ositettu otanta

Ositettu otanta on sopiva menetelmä tilanteisiin, joissa perusjoukko koostuu jonkin ominaisuuden suhteen homogeenisista ryhmistä eli ts. alkioryhmistä (osista). Ositettu otanta pyrkii varmistamaan, että tutkittava otos on edustava kaikkien (tutkimuksen kannalta) olennaisten ryhmien osalta.

Esimerkkinä voitaisiin 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 järkevää, sillä otoskoon pitäisi olla (todennäköisesti) hyvin suuri, että jokaisesta kieliryhmästä saataisiin poimittua edustava otos.

Ositetun otannan avulla otos voitaisiin kerätä niin, että jokaisesta ryhmästä (ositteesta) poimitaan osaotos yksinkertaisella satunnaisotannalla tai systemaattisella otannalla ja nämä osaotokset yhdistetään yhdeksi otokseksi.

Ositettu otanta voi (oikein toteutettuna ja sopivassa asetelmassa) tuottaa paljon tarkempaa tietoa kuin yksinkertainen satunnaisotanta samaa otoskokoa käytettäessä! Voidaan esimerkiksi käyttää tietoa siitä, että otosyksiköt ovat joka ositteessa keskenään samankaltaisia.

- Ositetun otannan käyttöön suurissa kyselytutkimuksissa liittyy samoja ongelma kuin yksinkertaiseen ja systemaattiseen satunnaisotantaan.
- Otokseen valikoituneet vastaajat voivat olla mm. levittäytyneinä suurelle maantieteelliselle alueelle. Näin ollen otannan suorittaminen vaatii suuria kustannuksia.
- Onko (järkevä) osittaminen ylipäätään mahdollista toteuttaa tarkasteltavassa sovelluksessa?

10.5 Ryväsotanta

Ryväsotanta

Ryväsotanta (klusteriotanta) 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 rypäitä, joiden alkiot ovat työntekijöitä ja koululuokat muodostavat koulun sisällä omia luonnollisia rypäitään ja opiskelijat ovat alkioita näissä rypäissä.

Huomionarvoista on, että toisin kuin ositetussa otannassa, ryväsotannassa rypäiden oletetaan olevan toistensa kanssa riittävän samankaltaisia, että jokaista rypästä ei tarvitse erikseen tutkia.

- Tämä onkin yksi ryväsotannan tärkeimpiä etuja, sillä sitä usein perustellaan kustannustehokkuudella. Tavoitteena on vähentää tietojen keruun aiheuttamia kustannuksia samalla varmistaen, että otos on kuitenkin mahdollisimman edustava!

Esimerkki. Sen sijaan että poimitaan satunnaisia koululaisia mahdollisesti suuresta määrästä kouluja, voidaan poimia satunnaisia rypäitä (kouluja), joista tutkimusyksiköt eli koululaiset poimitaan.

Lisäksi koulun sisällä koululuokat muodostavat alirypäitä, joista voidaan edelleen poimia satunnaisotos, jotta päästään tutkimaan perusjoukon alkioita eli koululaisia esim. haastattelututkimuksen muodossa.

Ryväsotannan voi suorittaa yksi- tai kaksivaiheisena (ts. kyseessä on yksiasteinen/kaksiasteinen ryväsotanta).

Kaksivaiheinen ryväsotanta

- **Ensimmäisessä vaiheessa** poimitaan joukko rypäitä kaikkien rypäiden joukosta, eli vain osa rypäistä on mukana lopullisessa otoksessa.
- **Toisessa vaiheessa** poimitaan ensimmäisessä vaiheessa poimituista rypäistä alkiotason otokset.

Yksivaiheisessa ryväsotannassa toisessa vaiheessa valitaan kaikki ensimmäisen vaiheen otosrypäiden alkiot, jolloin toisen vaiheen otanta typistyy ensimmäisen vaiheen rypäiden alkioiden kokonaistutkimukseksi.

- Poiminnan eri vaiheissa voidaan soveltaa yksinkertaista satunnaisotantaa tai systemaattista otantaa.

Ryväsotantaa käytetään usein suuria haastattelututkimuksia tehtäessä. Ryväsotantaa voidaan erityisesti hyödyntää myös silloin, kun tutkijalla ei ole käytettävissään kattavaa listaa kaikista havaintoyksiköistä, mutta näiden muodostamat rypäät ovat määritettävissä.

Ryväsotannan heikkoutena pidetään sitä, ettei aina ole helppoa muodostaa rypäitä, jotka ovat toistensa kaltaisia. Tulosten tarkkuus myös riippuu monin paikoin siitä, kuinka hyvin rypäisiin jako onnistuu.

Esimerkkejä ryväsotannasta.

Esimerkki 1: Poimitaan oppilaitoksen opiskelijoista otos arpomalla ensin otos luokkahuoneista (=rypäistä). Arvotuissa luokkahuoneissa käydään sitten suorittamassa kysely.

- Esim. Oppilaitoksen opiskelijoista voidaan poimia otos arpomalla ensin otos luokkahuoneista, jolloin luokkahuoneet ovat rypäitä.

- Mahdollisia ongelmia? Miten huomoida päivä- ja iltaopiskelijat? Tämän voisi toteuttaa arpomalla otos luokkahuoneista päiväsaikaan ja toinen otos ilta-aikaan. Tässä yhdistetään ryväsotantaan ositettu otanta, jolla taataan päivä- ja iltaopiskelijoiden edustus.

Esimerkki 2: Tutkittaessa tänä vuonna peruskoulun aloittavia voidaan ensin poimia otos kouluista, jolloin koulut ovat rypäitä. Tämän jälkeen arvotaan kustakin otokseen tulleesta koulusta tietty määrä tutkimuksen kohderyhmään kuuluvia oppilaita.

10.6 Esimerkkejä otantatutkimuksista

Esimerkki: Työllisyys ja työttömyys, Tilastokeskuksen työvoimatutkimus (toukokuussa 2025)

https://stat.fi/tilasto/dokumentaatio/tyti

Työvoimatutkimus antaa tuoreen ja kattavan kuvan työvoimasta ja työmarkkinoiden muutoksista. Julkisuudessa seurataan kuukausittain erityisesti työllisyyden ja työttömyyden muutoksia edellisen vuoden vastaavasta kuukaudesta. Erityisesti kausitasoitettuja aikasarjoja ja trendiaikasarjoja käytetään seurattaessa pitkän aikavälin kehitystä ja suhdannevaihtelua. (Aikasarja-aineistoja esitellään vielä tarkemmin myöhemmin)

- Työvoimatutkimus on otostutkimus, jonka avulla tilastoidaan 15–89-vuotiaan väestön työmarkkinoille osallistumista, työllisyyttä, työttömyyttä ja työaikaa kuukausittain, neljännesvuosittain ja vuosittain.
- Tutkimuksen tietosisältö perustuu EU:n asetukseen, ja tutkimuksen otokseen kuuluu joka kuukausi noin 12 500 henkilöä.
- Työvoimatilastoja käytetään työvoimapoliittisten ennusteiden ja suunnitelmien laadinnassa, toimien seurannassa ja päätöksenteon tukena.
  - Työmarkkina-aseman perusluokittelussa väestö jaetaan työllisiin, työttömiin ja työvoiman ulkopuolisiin.
  - Työlliset ja työttömät muodostavat työvoiman.
- Tilasto perustuu kuukausittain verkkokyselylomakkeella ja haastattelemalla kerättävään otosaineistoon.
- Työvoimatutkimuksen **perusjoukon** muodostavat Suomessa vakinaisesti asuvat 15–89-vuotiaat henkilöt.

  - Perusjoukkoon kuuluvat myös tilapäisesti ulkomailla (alle vuoden) oleskelevat sekä Suomen väestötietojärjestelmään rekisteröidyt ulkomaalaiset, joiden oleskelu Suomessa kestää vähintään vuoden.

- Työvoimatutkimuksen otos poimitaan **ositetulla systemaattisella otannalla** väestön keskusrekisteriin perustuvasta Tilastokeskuksen väestötietokannasta kahdesti vuodessa.

  - Ositteiden muodostamisessa käytetään NUTS1-jakoa ja ikätietoa. Ositteet ovat: Manner-Suomi (15–74-vuotiaat), Ahvenanmaan maakunta (15–74-vuotiaat) ja 75–89-vuotiaiden ikäryhmä.

- Tutkimus on paneelitutkimus, jossa samaa henkilöä haastatellaan viisi kertaa. Haastattelut tehdään kolmen kuukauden välein, paitsi neljäs haastattelu, joka tehdään kuuden kuukauden kuluttua kolmannesta haastattelusta.

Esimerkki: Terveys 2000

https://thl.fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/terveys-2000-2011/terveys-2000-tiiviisti

Terveys 2000 -tutkimuksen tavoite oli tuottaa ajankohtainen kattava kuva työikäisen ja iäkkään väestön terveydestä ja toimintakyvystä selvittämällä tärkeimpien terveysongelmien yleisyyttä ja syitä sekä niihin liittyvän hoidon, kuntoutuksen ja avun tarvetta.

- Tutkimus koskee (koski) 18 vuotta täyttänyttä Suomen aikuisväestöä (perusjoukko), josta valitaan valtakunnallisesti edustava 10 000 henkilön otos.
- Poimittiin kaksivaiheinen ryväsotos terveyskeskuspiireistä.
  - Ositus perustui yliopistosairaaloiden vastuualuiden väestömäärään suhteutettuun kiintiöintiin.
  - Suurimmat 15 terveyskeskuspiiriä poimittiin otokseen ja lopuista 65:stä piiristä poimittiin loppuotos kussakin ositteessa systemaattisella (PPS) otannalla (sisältymistodennäköisyys suhteessa alkion kokoon).

Chapter 11

Satunnaisotokset: Tilastollisen päättelyn näkökulma

Tarkastellaan seuraavaksi otoksia ja otosjakaumia hieman "tilastollisemmin" mitä aiempien otantaa koskevien lukujen yhteydessä. Tilastollinen päättely on keskeinen osa tilastotiedettä, sillä se mahdollistaa päätelmien yleistämisen otoksesta populaatioon/perusjoukkoon.

- Tämä ja seuraava luku toimivat esimerkkeinä formaaliin matemaattiseen esitykseen perustuvan **tilastollisen päättelyn perusteista** (otannan ja otantajakaumien näkökulmasta), jonka yleinen iso idea on yleisesti tehdä luotettavia johtopäätöksiä perusjoukosta otoksen perusteella.
- Tällä kurssilla käydään läpi (vain) tarvittavia yksityiskohtia sekä rakennetaan pohjia todennäköisyyslaskennan ja tilastollisen päättelyn peruskurssille.

11.1 Satunnaisotos, yhteisjakauma ja tilastollinen malli

Aiemmista luvuista muistamme, että tilastollisen tutkimuksen kohteena ovat satunnaisilmiöt, joita kuvataan satunnaismuuttujia käyttäen. Satunnaismuuttujilla on todennäköisyysjakaumat, joita tilastotieteessä kuvataan diskreettien sm:jien tapauksessa **pistetodennäköisyysfunktion** ja jatkuvien sm:jien tapauksessa **tiheysfunktion** avulla.

- Merkitään satunnaismuuttujaa edelleen isolla kirjaimella, Y, ja satunnaismuuttujan realisaatiota pienellä kirjaimella y. Otoskokoa, eli otokseen osallistuvien tilastoyksiköiden määrää merkitään n:llä ja tilastoyksiköitä indeksöidään alaindeksillä $i = 1, \ldots, n$.
- Otoksen poimimisen jälkeen satunnaismuuttujat $Y_1, \ldots, Y_n$ saavat havaituiksi arvoikseen havaintoarvot $y_1, \ldots, y_n$ (ts. $Y_1 = y_1, \ldots, Y_n = y_n$).
- Näin havaintoaineisto on siis **satunnaisotos**, joka voidaan määritellä tarkemmin seuraavasti.

Satunnaisotos. Olkoot $Y_1, \ldots, Y_n$ riippumattomia ja samoinjakautuneita satunnaismuuttujia, joiden tiheysfunktiota (tf., tai pistetodennäköisyysfunktiota (ptnf)) merkitään $f(y, \theta)$:llä, jossa y:n on yksittäisen sm:n Y reaalisaatio ja θ on jokin jakauman muodon määräävä parametri (tai parametrit).

Parametrin θ arvoa ei yleensä tunneta ja tavoitteena onkin päätellä, **estimoida**, sen arvo käytettävissä olevasta aineistosta.

Satunnaisotoksen tilastollinen malli

- Havaintoarvot $y_1, \ldots, y_n$ ovat kiinteitä lukuja, mutta ne vaihtelevat satunnaisesti otoksesta toiseen. Satunnaisotannassa **satunnaisuus liittyy siis havaintoarvojen vaihteluun satunnaisesti otoksesta toiseen**.
  - Satunnaisuus ei siis liity otannan tuloksena saatuihin havaintoarvoihin, vaan otoksen poimintaan. Toisin sanoen, yksittäisen tilastoyksikön havaintoarvo (esim. pituus) on kiinteä luku. Satunnaisotannassa satunnaisuus kumpuaa siis siitä, että tilastoyksikkö tulee valituksi otokseen sattumanvaraisesti.

Yhteisjakauma - Satunnaismuuttujien $Y_1, \ldots, Y_n$ **yhteisjakauma** muodostaa **tilastollisen mallin** havaintoarvojen satunnaiselle vaihtelulle eri otoksissa.

- Yhteisjakaumaa merkitään $f(y_1, \ldots, y_n; \theta)$, jossa havaitut arvot y_i ovat kiinteitä ja parametri θ on tuntematon.
- Koska tällä kurssilla satunnaismuuttujat $Y_1, \ldots, Y_n$ oletetaan **riippumattomiksi toisiinsa nähden**, niiden yhteisjakauma on tulomuotoa

$$f(y_1, \ldots, y_n; \theta) = f(y_1; \theta) \cdots f(y_n; \theta).$$

Tilastollisen päättelyn ensimmäinen tavoite ja tehtävä on pyrkiä havaitun aineiston $y_1, \ldots, y_n$ avulla estimoimaan parametrin θ arvo niin, että havaintojen yhteisjakauma kuvaa aineistoa parhaalla mahdollisella tavalla.

- Tässä $f(y_1, \ldots, y_n; \theta)$ on siis (tilastomatemaattisessa mielessä) tilastollinen malli. Tilastollisen mallin monimutkaisuus ilmenee sen parametrien määrästä eli ts. kuinka monta parametria sisältyy θ:aan. Mitä enemmän parametreja (erit. suhteessa havaintojen määrään), sitä monimutkaisempi malli.

Parsimoonisuusperiaate - Useimmiten ja perinteisesti ajatellaan, että on käytettävä niin yksinkertaisia menetelmiä kuin mahdollista, mutta ei yhtään yksinkertaisempia. Tämä on ns. **parsimoonisuusperiaate** eli **vähäparametrisuus-** tai **säästeliäisyysperiaate**.

Harhan ja varianssin kompromissi - Vähäparametrisuusperiaatteen voidaan nähdä perustuvan ns. Occamin partaveitsen -periaatteeseen, jonka mukaan *"ilmiöitä selittävien 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 tärkeää tilastollisen ennustamisen ja viime vuosikymmeninä yleistyneen tilastollisen (kone)oppimisen sovellutuksissa. Tähän palataan vielä myöhemmin ennustamista koskevassa luvussa.

- Mallin muodon määrää tutkijan tekemä aineistoa koskeva jakaumaoletus satunnaismuuttujille Y_i, $i = 1, ..., n$, mikä voi paikoin olla hyvinkin monimutkainen.

Oletetaan, että $Y_1, \ldots, Y_n$ ovat aiempien oletusten pätiessä riippumattomia sm:jia ja että ne muodostavat satunnaisotoksen jakaumasta, jonka odotusarvo on μ ja varianssi on σ^2.

- Ts. oletamme

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

- Tässä tapauksessa mielenkiinnon kohteena olevat parametrit ovat siis μ ja σ^2 eli $\theta = (\mu \quad \sigma^2)$. Tässä θ on siis vektori, joka koostuu kahdesta parametrista.
- Tutkimuskysymyksestä johdetut hypoteesit voisivat koskea esimerkiksi näitä parametreja, tai tarkemmin niiden oikeita arvoja, jotka ovat siis lähtökohtaisesti tuntemattomia.
- Tilastollisten mallien tehtävänä on siis estimoida nämä todennäköisyysjakaumien parametrit havaitun aineiston perusteella, joten keskeinen tilastollinen kysymys on, **että miten estimointi suoritetaan luotettavasti**? Estimointi voidaan perustaa (tiettyjen oletusten pätiessä) nk.

uskottavuusfunktion käyttöön ja **suurimman uskottavuuden (SU) menetelmään**, mitä sivuttiin hyvin lyhyesti Osassa I ja tarkastellaan tarkemmin myöhemmissä tilastotieteen opinnoissa.

Esimerkki: Satunnaisotos normaalijakaumasta. Normaalijakautuneiden satunnaismuuttujien satunnaisotokselle $Y_1, \ldots, Y_n$ pätee $Y_i \sim \mathrm{N}(\mu, \sigma^2), , , i = 1, \ldots, n)$.

- Huom. tässä pätee myös oletus että sm:t $Y_1, \ldots, Y_n$ ovat riippumattomia ja samoin jakautuneita. Toisinaan käytetään lyhenteitä **iid** tai *i.i.d*, jotka tulevat englannin kielen ilmaisusta "independent and identically distributed".
  - Jos tätä oletusta vahvistetaan vielä normaalisuusoletuksella, kuten yläpuolella, niin käytettäisiin lyhennettä **nid** eli *normally and independently distributed*.
  - Merkintä soveltuu käytettäväksi muidenkin jakaumien tapauksessa.
- Esimerkiksi R-ohjelmassa voidaan generoida 10 havainnon ($n = 10$) satunnaisotos standardinormaalijakaumasta, eli kun $\mu = 0$ ja $\sigma^2 = 1$ (ts. $Y_i \sim \mathrm{N}(0, 1),\ i = 1, \ldots, 10$) komennolla rnorm(10).

11.2 Tilastollisia jakaumia

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

11.2.1 Normaalijakauma

Normaalijakauma

Jos satunnaismuuttuja Y noudattaa **normaalijakaumaa** odotusarvolla $\mathsf{E}(Y) = \mu$ ja varianssilla $\mathsf{Var}(Y) = \sigma^2$, niin tällöin merkitään $Y \sim \mathrm{N}(\mu, \sigma^2)$.

- Y:n tiheysfunktio on normaalijakauman tapauksessa muotoa (ks. myös Osan I normaalijakauman tiheysfunktioiden kuvaajia eri parametriarvojen tapauksessa)

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

  jossa e viittaa potenssia laskettaessa Neperin lukuun $e \approx 2.71828$.

- Ylläoleva tiheysfunktio määrittelee parven normaalijakaumia kun parametreille (vakioille) μ ja σ^2 annetaan erilaisia arvoja. Nämä kaksi parametria määräävät normaalijakauman tarkemman muodon.

Esimerkki: Miesten pituus. Tutkitaan miesten pituutta hyvin määritellyssä joukossa, kuten varusmiespalvelusta tiettynä vuonna suorittavien joukossa (vrt. jo aiemmin esitelty Pearsonin klassinen isien ja poikien pituusesimerkki)

- Pituus on ominaisuus, jonka voidaan nähdä määräytyvän monista perintö- ja ympäristötekijöistä. Pituutta voidaan siis pitää satunnaismuuttujana.
- Oletetaan, että pituus noudattaa normaalijakaumaa. Näin ollen sm. Y on valitun miehen pituus ja $Y \sim \mathrm{N}(\mu, \sigma^2)$.
- Tuntemattomien parametrien μ ja σ^2 tulkinta:
  - Odotusarvo $\mu = \mathsf{E}(Y)$ on satunnaisesti valitun miehen pituuden odotettavissa oleva arvo.
  - Varianssi $\sigma^2 = \mathsf{Var}(Y) = \mathsf{E}\left[(Y-\mu)^2\right]$ kuvaa valitun miehen pituuden odotusarvostaan määrätyn poikkeaman (keskihajonnan) neliön odotettavissa olevaa arvoa (kuvaten ts. pituuksien jakauman keskittyneisyyttä/hajaantuneisuutta pituuksien odotusarvon ympärillä).

11.2.2 Bernoulli- ja binomijakauma

Bernoulli-jakauma ja sen laajennuksena toistokoetyyppisissä tilanteissa käytettävä **binomijakauma** ovat diskreettejä jakaumia, joiden avulla voidaan tiivistää onnistumiskertojen lukumääriä ja tapausten prosentuaalisia osuuksia, joissa onnistunut tapahtuma sattui. Kyseessä on siis diskreettien satunnaismuuttujien realisaatiot ja niiden suhteellinen osuus ko. otoksessa. Bernoulli-jakauma

Bernoulli-jakauma (tai Bernoullin jakauma) on (diskreetti) 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 $\mathrm{P}(Y = 1) = p$ ja vastaavasti vastatodennäköisyys $\mathrm{P}(Y = 0) = 1 - p$.
- Bernoulli-jakaumaa merkitään $Y \sim B(p)$, jossa siis $0 < p < 1$.

- Bernoulli-jakauman pistetodennäköisyysfunktio on muotoa

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

  jossa (y) on sm:n (Y) realisaatio (havaittu arvo) ja parametri (p) on tuntematon, jota voidaan estimoida otoksen avulla, kuten myöhemmin tullaan näkemään.

- Bernoulli-jakautuneen sm:jan odotusarvo $\mathsf{E}(Y) = p$ ja varianssi $\mathsf{Var}(Y) = p(1-p)$.

Todetaan tässä kohtaa lyhyesti seuraavasta eli Bernoulli-jakauma (ja binomijakauma sen laajennuksena) liittyy **logistiseen regressioon** eli regressiomalliin, jossa vastemuuttuja on binäärinen. Käytännössä malli voidaan nähdä mallina logaritmiselle vetosuhteelle onnistumisen ja epäonnistumisen todennäköisyyksien välillä.

- Käsitteet **odds** (**vetokerroin**) ja laajemmissa tarkasteluissa **odds ratio** (**ristisuhde** tai **vetosuhde**) liittyvät tähän.

- Esimerkiksi, koska sadasta pekonia syömättömästä henkilöstä 6 saa suolistosyövän ja 94 ei saa, odds on tässä tapauksessa 6/94, jota joskus kutsutaan myös 6-94.

- Kertoimia käytetään yleisesti Isossa-Britanniassa vedonlyönnissä, mutta niitä käytetään myös laajasti mittasuhteiden tilastollisessa mallintamisessa. Tämä tarkoittaa, että lääketieteellisessä tutkimuksessa hoitotoimenpiteiden tai käyttäytymisen vaikutuksia ilmaistaan usein kertoimien suhteena.

Binomijakauma

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

- Binomijakaumalla kyetään vastaamaan mm. kysymykseen millä todennäköisyydellä n:n kokoisessa otoksessa tapahtuu k onnistumista.

Esimerkki: Miesten lukumäärä Saksin osavaltion perheissä 1876–1885. 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.

- Ks. tarkemmin esimerkki 3.2 Friedlyn ja Meyerin kirjassa (s. 67-68, 2015).

Miesten lukumäärä Saksin osavaltiossa 12:n lapsen perheissä:

	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

Tässä tilantessa mielenkiinnon kohteena saattaisi olla hypoteesi, jonka mukaan pojan (miehen) syntymätodennäköisyys (P(mies) $= p$ on $p = 0.5$.

11.2.3 Poisson-jakauma

Poisson-jakauma

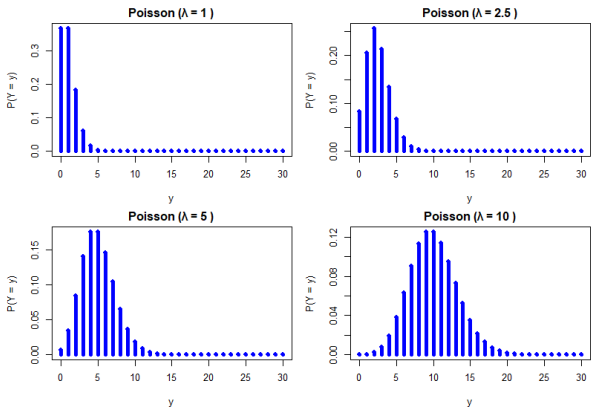
Poisson-jakauma (Poissonin jakauma). Jos satunnaismuuttuja (Y) on Poisson-jakautunut, merkitään $Y \sim P(\lambda)$, jossa parametri $\lambda > 0$ on Poisson-jakauman parametri.

- Poisson-jakaumaa voidaan käyttää tilanteissa, joissa sm. (Y) on jokin lukumäärä ja sen pistetodennäköisyysfunktio on muotoa

$$\mathrm{P}(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \qquad y = 0, 1, 2, ...$$

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

- Alla kuvassa on kuvattu Poisson-jakauman pistetodennäköisyysfunktion muotoja parametrin (λ) eri arvoilla.



Kuva: Poisson-jakauman pistetodennäköisyysfunktiota mallin parametrin (lambda) eri arvoilla.

Esimerkki: Poisson-jakauma ja urheilutilastotiede. 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 joukkuetta (aiemmin Valioliigan perustamiskauden 1992–1993 jälkeen 22 joukkuetta) 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, jonka esittämään kysymykseen (hypoteesiin) vastaus on tietenkin ilmeinen!

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

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.

Seuraavassa ristitaulukossa (ristiintaulukossa) raportoidaan 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+.

	Vierasjoukkueen maalien lkm.					
Kotij. maalien lkm.	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

Olettamalla, että koti- ja vierasjoukkueen todennäköisyys tehdä maali ottelun aikana on vakio ja riippumattomia toisistaan (vahva yksinkertaistava oletus), niin tällöin koti- ja vierasjoukkueen ottelun aikana tekemien maalien lukumäärää (ilman edellä käytettyä maalimäärien "katkaisua" neljään) voidaan melko hyvin approksimoida oletuksella, että nämä lukumäärät ovat Poisson-jakautuneita. Ts. $Y_i^H \sim P(\lambda_H)$ on sm., joka kuvaa i:n ottelun kotijoukkueen

tekemien maalien lukumäärää ja intensiteettiparametrin λ_H arvon määrittäminen kuuluu tilastollisen päättelyn ja erityisesti estimointiteorian piiriin. Vastaavasti vierasjoukkueen maalimäärät: $Y_i^A \sim P(\lambda_A)$.

Osoittautuu, että parametreille λ_H ja λ_A saatavat estimaatit tarkateltavassa aineistossa ovat $\lambda_H = 1.49$ ja $\lambda_A = 1.06$v ja ne vastaavat tässä yksinkertaistetussa tilanteessa koti- ja vierasjoukkueen keskimääräisiä maalimääriä:

	Kotijoukkue (home)	Vierasjoukkue (away)	Yht.
Otoskeskiarvo	1.486	1.063	2.550
Otosvarianssi	1.316	1.172	

Tuloksista voidaan siis päätellä, että kotijoukkueen (odotettavissa oleva) maalimäärä on vierasjoukkuetta korkeampi (osoittaen kotiedun merkitystä jalkapallossa). Lisäksi edellä todetun Poisson-jakauman teoreettisten ominaisuuksien mukaisesti keskimääräiset maalimäärät ovat lähellä niiden otosvariansseja, mikä osoittaa osaltaan (tässä yksinkertaistetussa tilanteessa), että Poisson-jakaumaan perustuva jakaumaoletus on vähintään kohtuullisen kelvollinen.

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

11.3 Tunnusluvut ja parametrien estimaattorit

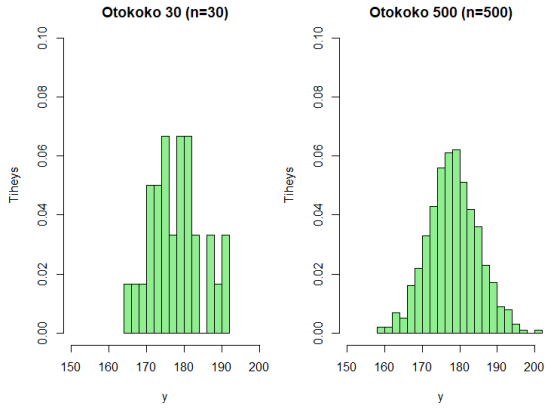
Jo aiemmin tämän materiaalin puitteissa (Osassa I) totesimme, että erityisesti klassisessa tilastotieteessä tilastollinen päättely pohjautuu aineiston tilastollisen mallin kuvaamalle **tilastolliselle stabiliteetille**, joka ilmenee ajatuksena aineiston keruun toistamisesta. Tilastollinen stabiliteetti

- Oletetaan, että tarkasteltavan aineiston on tuottanut satunnaisotanta tai satunnaiskoe, joka noudattaa tilastollista mallia $f(y_1, \ldots, y_n; \theta)$ (ks. edellä tehty päättely tässä suhteessa).
- Toistetaan aineiston keruu samoissa olosuhteissa yhä uudelleen ja uudelleen.
  - Saatava aineisto (numeeriset arvot) $(y_1, \ldots, y_n)$ vaihtelevat näin ollen täsmennetyn tilastollisen mallin jakauman kuvaamalla tavalla.

Esimerkki: Normaalisti jakautunut aineisto. Tilastollisella stabiliteetillä tarkoitetaan sitä, että saman tilastollisen mallin ja todennäköisyysjakauman generoimat aineistot ovat ominaisuuksiltaan samankaltaisia.

- Esimerkiksi normaalijakautuneen aineiston tapaus: Oletetaan, että sm:jat $Y_1, ..., Y_n$ noudattavat samaa normaalijakaumaa $\mathsf{N}(\mu, \sigma^2)$, eli niillä on sama odotusarvo ja varianssi eri havainnoille $i = 1, ..., n$. Tämä tarkoittaa sitä että eri otosten $y_1, \ldots, y_n$ havainnot vaihtelevat saman odotusarvon ympärillä ja keskimäärin samalla tavalla (varianssilla). Edelleen, tilastollinen stabiliteetti voidaan tässä yhteydessä ilmaista myös *suurten lukujen lain (SLL)* kautta. Yksi SLL:n muoto sanoo, että tässä tapauksessa havaintojen aritmeettinen keskiarvo lähestyy havaintoarvojen odotusarvoa, kun havaintojen lukumäärä n kasvaa.

- Allaolevassa kuvassa on esimerkkinä kaksi simulointia. Ensimmäisessä on arvottu 30 realisaatioita ($n = 30$) normaalijakaumasta, jonka odotusarvo $\mu = 178$ ja varianssi $\sigma^2 = 49$ (eli keskihajonta on 7). Vastaavasti toisessa kuvassa simuloidaan samasta jakaumasta, mutta nyt otoskoko on 500 havaintoa ($n = 500$).



Kuva: Kaksi otoskooltaan 30 ja 500 havaintoa sisältävää otosta normaalijakaumasta, jonka odotusarvo on 178 ja varianssi 49.

Näissä kahdessa kuvassa havainnot on esitetty histogrammeina, joissa vihreät palkit kuvaavat kyseiseen havaintoarvojen väliin kuuluvien havaintojen osuutta. Kuvista selviää otoskoon vaikutus. Suuremman otoskoon tapauksessa jakauma näyttää vielä enemmän normaalijakaumalta, mitä melko pienen otoskoon tapauksessa. Kuvaajista huomataan lisäksi kuinka valtaosa havainnoista keskittyy odotusarvon ympärille ja erityisesti oikealla olevassa kuviossa arvojen vaihtelu on hyvin symmetristä odotusarvon ympärillä.

Histogrammi Huomiona edellisiin satunnaisotosten kuvaajiin liittyen, ne ovat siis ns. **histogrammeja**. Histogrammi on yleinen tapa esittää tämänkaltaista aineistoa. Tässä tapauksessa jatkuvan muuttujan havaintoarvot on luokiteltu niin, että jokainen havainto kuuluu yhteen tasaväliseen luokkaan ja yhden luokan yleisyyttä kuvaa yksi pylväs. Pylväiden "luokittelut" tehdään graafisen esityksen aikaansaamiseksi. Pylvään korkeus määräytyy luokan sisältämien arvojen lukumäärän eli frekvenssin mukaan.

Tunnusluku

Satunnaisotoksesta voidaan laskea erilaisia **tunnuslukuja/otossuureita**, joita merkitään $T(Y)$:llä, ts. ne ovat aineiston funktioita

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

Tunnusluvut ovat siis satunnaismuuttujien funktioina myös satunnaismuuttujia!

Otoksen poimimisen jälkeen, havaintoarvoja käyttäen, voidaan laskea **tunnuslukujen havaitut arvot** (jolloin ne ovat siis ei-satunnaisia).

- Ts. havaitussa aineistossa (realisaatio pisteessä) $(y_1, \dots, y_n)$ pätee

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

  jossa pieni-t $(t(y))$ korostaa tunnusluvun numeerista arvoa vs. iso-T $(T(Y))$ yläpuolella.

Esimerkkinä tunnusluvusta on keskiarvo $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

- Otoskeskiarvo on havaittujen arvojen keskiarvo, kun se lasketaan kerätystä aineistosta, ts. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Estimaattori ja estimaatti Laajemmin vielä edelliseen määrittelyyn liittyen:

Jos tunnuslukua $T(Y)$ käytetään tilastollisen mallin parametrin, tai samoin myös usean parametrin tapauksessa parametrin (ajoittain parametrivektorin kun parametreja on useita) (θ) estimointiin, niin tätä sanotaan tällöin parametrin **estimaattoriksi**, jota voidaan merkitä esim. $\hat{\theta}(Y) = g(Y_1, \dots, Y_n)$. Estimaattorin otoskohtaisia arvoja kutsutaan **estimaateiksi** ja merkitään esim. $\hat{\theta}(y)$.

Oletetaan, että estimaattorilla on nk. todellinen arvo, $g(\theta)$, joka vastaa parametrin arvoa perusjoukon tasolla ja jota pyritään aineistoa käyttäen estimoimaan. Toivottavaa olisi, että estimaatit $\hat{\theta}(y) = g(y_1, \dots, y_n)$ osuisivat mahdollisimman lähelle tunnusluvun todellista arvoa $g(\theta)$. Ts. satunnaismuuttujan eli tässä tapauksessa estimaattorin $\widehat{\theta}(Y) = g(Y_1, \dots, Y_n)$ jakauman tulisi keskittyä mahdollisimman tiiviisti $g(\theta)$:n ympärille.

Uskottavuusfunktio **Uskottavuusfunktio**. Erityisesti klassisessa tilastotieteessä tilastollisen mallin parametrien estimointi perustetaan usein nk. **suurimman uskottavuuden menetelmään**.

- Koska mielenkiinnon kohteena on tilastollisen mallin eli tehdyn jakaumaoletuksen alaisen yhteisjakauman parametrit, perustuu suurimman uskottavuuden estimointi näiden parametrien estimaattoreihin. Tavoitteena on löytää sellaiset parametriarvot, jotka ovat havaitun aineiston kannalta uskottavimmat tilastollisen mallin parametrien arvot.

Käytännössä suurimman uskottavuuden estimointi perustuu valitun tilastollisen mallin määrittelevään tiheysfunktioon tai pistetodennäköisyysfunktioon, kun aineisto on havaittu.

- Tarkemmin, nk. **uskottavuusfunktiota** merkitään $L(\theta) = L(\theta; y_1, \ldots, y_n) = f(y_1, \ldots, y_n; \theta)$, jossa kirjain L tulee englannin kielen sanasta *likelihood function*. Huomaa yhteys yhteisjakaumaan tehdyillä oletuksilla. Uskottavuusfunktiossa aineisto on havaittu eli se tulkitaan kiinteäksi ja parametri (θ) on tuntematon muuttuja, jonka arvoa pyritään estimoimaan. Ajoittain uskottavuusfunktiosta jätetään lopulta parametrista θ riippumattomat komponentit pois.
- Koska tehty jakaumaoletus määrää yhteisjakauman muodon, voidaan suurimman uskottavuuden estimointi perustaa siitä johdettuihin todennäköisyyksiin.
- Lopulta valitaan sellaiset parametriarvot, jotka ovat aineiston valossa kaikkein *uskottavimmat* ja kutsutaan niitä **suurimman uskottavuuden estimaateiksi**.

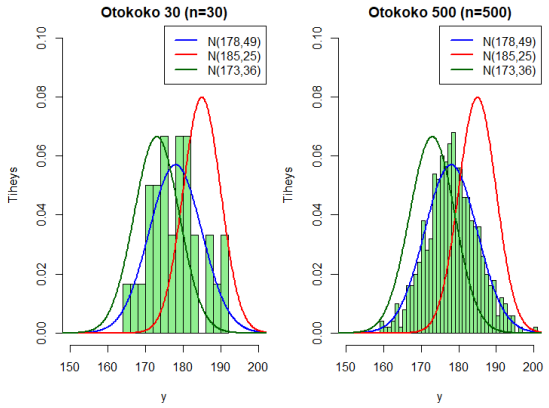
Suurimman uskottavuuden estimointiin ja siihen pohjaavaan päättelyyn syvennytään tarkemmin tilastollisen päättelyn perus- ja aineopintokursseilla.

Esimerkki: Normaalisti jakautuneen aineiston parametrien estimointi. Normaalijakauma on odotusarvon suhteen symmetrinen jakauma, joten odotusarvon estimointi voidaan perustaa havaitun aineiston "sijaintiin".

- Toisin sanoen, normaalisti jakautuneen aineiston ja sm:jan Y odotusarvon $\mathsf{E}(Y) = \mu$ estimaattorina toimii havaintoarvojen otoskeskiarvo $\bar{Y}$.
- Vastaavasti normaalisti jakautuneen aineiston varianssi kuvaa havaintoarvojen keskittyneisyyttä/hajaantuneisuutta odotusarvon ympärillä.

Uskottavuuspohjaisen päättelyn toimintaperiaatetta voidaan kuvata jälleen käyttäen normaalijakautunutta aineistoa esimerkkinä. Jatketaan siis edellä aloitettua esimerkkiä. Seuraavissa kuvioissa toistetaan aiemmat simuloitujen havaintojen histogrammit. Samaan kuvaan on nyt molemmissa tapauksessa piirretty myös kolmen normaalijakauman tiheysfunktiot.

- Mikä kolmesta vaihtoehdosta näyttäisi sopivan parhaiten kuvaamaan aineistoa? Ts. mikä näistä olisi mielestäsi *uskottavin* kandidaatti havaitulle aineistolle?
- Huomioi, että käytännössä todellisen aineiston tapauksessa parametrien arvoja ei tiedetä, vaan ne pitää päätellä eli estimoida aineistosta! Tässä tapauksessa pienemmän aineiston otoskeskiarvo 177.7 ja otosvarianssi 47.2. Vastaavasti isomman aineiston tapauksessa otoskeskiarvo on 177.7 ja otosvarianssi 45.4.
- Luonnollisesti aineistoa generoineen normaalijakauman parametrien ollessa $\mu = 178$ ja $\sigma^2 = 49$, niin sopivin on sininen käyrä eli jakauman $\mathsf{N}(178, 49)$ tiheysfunktio. Muiden jakaumien "paikat" eli odotusarvot ovat väärässä kohdassa. Havaintojen lukumäärän kasvaessa tämä käy selvemmäksi, mikä vastaa täsmälleen tilastotieteen keskeisiä ideoita otoskoon vaikutuksesta.



Kuva: Kolmen normaalijakauman tiheysfunktiot suhteessa simuloituihin aineistoihin.

Hyvän estimaattorin ominaisuudet. Merkitään seuraavassa parametrin θ estimaattoria $\widehat{\theta}$:lla. Siltä voidaan toivoa seuraavia ominaisuuksia:

- **Harhattomuus**: Estimaattorin odotettavissa oleva arvo yhtyy tuntemattoman parametrin θ todelliseen arvoon eli $\mathsf{E}(\widehat{\theta}) = \theta$.
  - 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 otoskohtaiset arvot parametrille jakautuvat otantaa toistettaessa (symmetrisesti) parametrin todellisen arvon ympärille.

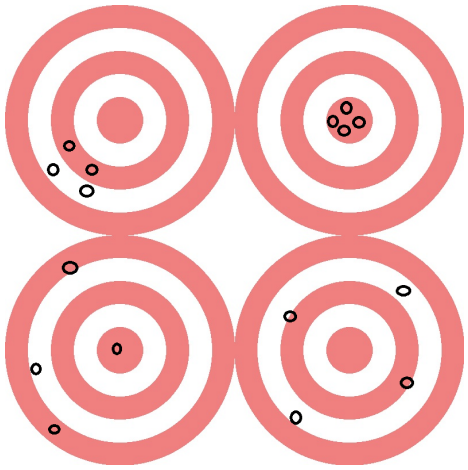
- **Tyhjentävyys**. Tyhjentävä estimaattori käyttää kaiken otokseen sisältyvän parametria θ koskevan informaation.

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

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

Voidaan osoittaa (yksityiskohdat sivuutetaan tällä kurssilla), että esimerkiksi yksinkertaisen satunnaisotoksen tapauksessa tavanomaisilla binomi- ja normaalijakauman parametrien estimaattoreilla on kaikki edellä mainitut hyvät ominaisuudet. Näin ei ole yleisesti monimutkaisemmissa tilastollisisssa malleissa.

Seuraavassa kuvassa havainnollistetaan vielä erilaisten estimaattorien ominaisuuksia. Vasemmalla ylhäällä oleva estimaattori on täsmällinen (precise), mutta ei tarkka (accurate), toisin kuin oikealla ylhäällä olevan estimaattori, joka on myös tarkka ja siten suositeltavin. Vastaavasti vasemmalla alhaalla oleva estimaattori ei ole tarkka eikä täsmällinen, kun taas oikealla alhaalla estimaattori on (keskimäärin) tarkka mutta ei täsmällinen.



Kuva: Havainnollistuksia estimaattoreiden ominaisuuksista.

Todetaan vielä tiivistetysti, että estimaattoreiden kehittäminen erilaisten tilastollisten mallien tapauksessa kuuluu teoreettisen tilastotieteen alaan.

Seuraavaksi perehdytään tarkemmin kahteen kenties useimmiten tarkasteltavaan tunnuslukuun: otoskeskiarvoon ja otosvarianssiin.

11.4 Tarkemmin otoskeskiarvosta ja otosvarianssista estimaattoreina

Oletetaan, kuten aiemminkin, että $Y_1, \ldots, Y_n$ ovat riippumattomia sm:jia ja että ne muodostavat satunnaisotoksen jakaumasta, jonka odotusarvo on μ, ts. $\mathsf{E}(Y_i) = \mu$, ja varianssi on σ^2, ts. $\mathrm{Var}(Y_i) = \sigma^2$.

Havaintojen (satunnaismuuttujien) $Y_1, \ldots, Y_n$ **otoskeskiarvo** on

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

- Kuten edellä todettiin, yksittäisen jo "realisoituneen" otoksen otoskeskiarvo on tällöin sm:jien realisaatioiden aritmeettinen keskiarvo

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

  Vastaavasti otoskeskiarvo $\bar{Y}$ on satunnaismuuttuja, jonka saama arvo $\bar{y}$ vaihtelee satunnaisesti otoksesta toiseen satunnaisotannasta johtuen.

- Kun satunnaismuuttujien odotusarvo on μ, on otoskeskiarvo jakauman odotusarvon harhaton estimaattori, ts. voidaan osoittaa, että pätee

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

  Täten otoskeskiarvo kuvaa aineiston perusjoukon tilastollisen mallin odotusarvoa.

Otosvarianssi. Aineiston sisältämää vaihtelua voidaan kuvata **otosvarianssilla**

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

- Sm:jien vaihtelua perusjoukon tasolla kuvataan **populaatiovarianssilla**

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

  ja voidaan osoittaa, että otosvarianssi estimoi tätä harhattomasti.

- Otokseen $y_1, ..., y_n$ perustuva (havaittava) otosvarianssi (vrt. otoskeskiarvon käyttäytyminen)

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

Yhteenvetona siis, otoskeskiarvo $\bar{Y}$ ja otosvarianssi S^2 ovat satunnaismuuttujia, joiden saamat arvot (estimaatit) $\bar{y}$ ja s_y^2 vaihtelevat satunnaisesti otoksesta toiseen, ja näillä on estimaattoreina (useita) suotuisia ominaisuuksia

- Vrt. edellisen jakson estimaattorin neljä keskeistä ominaisuutta, joita ei kuitenkaan vielä juuri tämän enempää käsitellä tällä kurssilla.

Esimerkki: Normaalisti jakautuneen aineiston parametrien estimointi (jatkoa). Tarkastellaan normaalisti jakautuneen aineiston parametriestimaatteja edellä aloitettuun esimerkkiin liittyen. Edellä jo raportoitiin otoskokoihin 30 ja 500 havaintoa perustuneiden otosten otoskeskiarvot ja otosvarianssi:

- Pienemmän aineiston otoskeskiarvo 177.7 ja otosvarianssi 47.2. Vastaavasti isomman aineiston tapauksessa otoskeskiarvo 177.7 ja otosvarianssi 45.4

Generoidaan seuraavaksi suurempi satunnaisotos samasta aineistosta valitsemalla otoskooksi $n = 2000$. Nyt estimaateiksi saadaan $\bar{y} = 178.2$ ja $s^2 = 49.1$. Ts. (aiempiin estimaatteihin nähden) otoskeskiarvo on edelleen yhä tarkempi arvio oikeasta parametriarvosta otosvarianssikin ollessa nyt lähempänä varianssin oikeaa arvoa. Otoskokoa kasvattamalla estimaatit lähestyvät siis oikeita arvoja. Tämä liittyy (yleisellä tasolla tässä kohtaa) ko. estimaattorien **tarkentuvuusominaisuuteen** eli otoskoon kasvaessa tarkentuvat estimaattorit (ja niiden myötä saatavat estimaatit) ovat yhä lähempänä estimoitavien parametrien todellisia arvoja.

Chapter 12

Otosjakaumat ja epävarmuuden arvioiminen

Tässä luvussa pohditaan tarkemmin sitä, **miten tilastollisessa tutkimuksessa saatujen tulosten epävarmuutta voidaan arvioida**.

- Erilaisten tutkimustulosten yhteydessä törmätään usein esim. ns. **virhemarginaalin** käsitteeseen. Virhemarginaalilla tarkoitetaan kansankielessä jonkin tunnusluvun, kuten esimerkiksi keskiarvon, arvojen sellaista vaihteluväliä, joka johtuu satunnaisuudesta. Käytännössä se tarkoittaa sitä, että saatu tulos, estimaatti, on ilmiön käyttäytymiseen (otantaa toistettaessa) nähden epävarma.
- Mistä tämä epävarmuus tutkimustuloksissa kumpuaa ja miten siihen voidaan tutkimusta tai koeasetelmaa suunnitellessa vaikuttaa?

Tilastotieteen yksi keskeisimpiä tavoitteita on pyrkiä tuottamaan tarkkaa ja tutkittua tietoa. Kun tutkimuskohteena on epävarma ja satunnainen ympäröivä maailma, onkin tilastollisten menetelmien luotettavuuden kannalta keskeistä pyrkiä arvioimaan tätä epävarmuutta, jotta saatuihin tuloksiin voidaan luottaa ja ymmärtää samalla tuloksiin liittyvän epävarmuuden suuruutta. Tässä luvussa käsitellään tähän epävarmuuteen liittyviä tilastotieteen perusteita.

12.1 Otosjakauma

Hieman kertauksena jo edellisistä luvuista, mutta nyt täydennettynä otosjakaumaa koskevalla määrittelyllä voidaan todeta seuraavaa.

Otosjakauma

Tunnusluvun/estimaattorin otosjakauma. Tunnusluku (T) (kuten esim. otoskeskiarvo) on satunnaismuuttujien $Y_1, \ldots, Y_n$ funktiona myös satunnaismuuttuja. Tämä tarkoittaa että tunnusluvun (T) arvot vaihtelevat otoksesta toiseen (tämä on ns. otosvaihtelua) jonkin todennäköisyysjakauman mukaisesti. Tätä kutsutaan **tunnusluvun (T) otosjakaumaksi** (otantajakaumaksi) ja sen avulla saatujen tunnuslukujen (estimaattien) luotettavuutta voidaan arvioida.

Tähän määrittelyyn liittyen on syytä vielä todeta, että todellisissa tutkimustilanteissa otantaa ei yleensä toisteta, jolloin otoksen poiminnassa käytetty otantamenetelmä (kuten arvonta) on ainutkertainen tapahtuma. Usein otantaa ei voida toistaa edes periaattessa. Tämä tarkoittaa, mikä voi tuntua yllättävältä, että otostunnuslukujen teoreettinen otosjakauma voidaan silti saada selville ja hyödyntää siten tilastollisessa päättelyssä ja lopulta aineiston käytännön analysoinnissa.

- Ts. todennäköisyyteen perustuvassa otannassa otantavirhettä (eli, kuten aiemmin todettiin, otannasta aiheutava satunnaisuus tunnusluvun arvossa) voidaan arvioida yhden otoksen perusteella, koska tunnetaan tunnusluvun otantajakauma.

Edellisiin lukuihin viitaten, tässäkin tapauksessa otosjakaumat 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 olemme jo aiemmin nähneet, käytännön havaintoaineistoa käyttäen.

Otosjakaumien teoria muodostaa perustan parametrien estimoinnille sekä estimaattorien ominaisuuksille sekä parametreja koskevien hypoteesien testaamiselle. Tähän liittyvä tilastollinen päättely tähän asti opittuna:

- Tilastollisessa tutkimuksessa pyritään tekemään päätelmiä aineiston generoineen satunnaisilmiön luonteesta. Ts. pyrkimyksenä on tehdä yleistyksiä otoksesta perusjoukkoon.
- Satunnaisilmiötä ja sen generoinutta aineistoa kuvataan todennäköisyysjakaumalla, jonka muodon määrää kyseisen jakauman parametrit.
- Parametrit ovat tilastollisessa päättelyssä mielenkiinnon kohteena ja niitä pyritään estimoimaan havaintoaineistosta otostunnusluvuilla/estimaattoreilla.

Aritmeettisen keskiarvon ominaisuuksia. Edellisestä luvusta muistamme, että aritmeettinen keskiarvo on eräs aineistosta yleisesti laskettu tunnusluku.

Tarkastellaan seuraavaksi aritmeettisen keskiarvon otosjakauman ominaisuuksia.

Edellä tehtyjen (sm:jia Y_i koskevien) oletusten pätiessä voidaan osoittaa (perustellaan myöhemmillä kursseilla) että aritmeettisella keskiarvolla $\bar{Y}$ on seuraava odotusarvo ja varianssi:

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

Aritmeettisen keskiarvon $\bar{Y}$ **standardipoikkeama** on tällöin

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

Standardipoikkeamaa kutsutaan myös **keskiarvon keskivirheeksi** ja se kuvaa otoskeskiarvon otosvaihtelua odotusarvon μ ympärillä.

Standardipoikkeama ja keskivirhe

Huomioi, että **otoskeskiarvon varianssi** on eri asia kuin **otosvarianssi**. Otoskeskiarvon varianssi kuvaa sitä, miten otoskeskiarvon toteumat vaihtelevat otantaa toistettaessa, eli otoksesta toiseen. Näin ollen otoskeskiarvon varianssia voidaan käyttää saadun otoskeskiarvon toteuman luotettavuuden arviointiin.

Huomataan myös, että aritmeettisen keskiarvon otosjakauma keskittyy yhä voimakkaammin havaintojen yhteisen odotusarvon μ ympärille, kun otoskoko n kasvaa.

- Ts. otoskoon n kasvaessa $\mathsf{Var}(\bar{Y}) = \frac{\sigma^2}{n}$ pienenee (koska nimittäjä suurenee suhteessa osoittajaan). Näin ollen aritmeettinen keskiarvo tuottaa tarkempia estimaatteja odotusarvosta silloin, kun otoskoko on suuri.

Otoskeskiarvo ja normaalijakautunut otos. Muodostakoot sm:jat $Y_1, \ldots, Y_n$ satunnaisotoksen normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$. Tällöin voidaan osoittaa, että havaintojen $Y_1, \ldots, Y_n$ keskiarvo $\bar{Y}$ noudattaa normaalijakaumaa odotusarvolla μ ja varianssilla σ^2/n (ks. yllä) ja pätee jakaumatulos

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

Itse asiassa ns. **asymptoottiseen teoriaan** vedoten eli suurten otosten tapauksessa (ts. n on "iso") voidaan osoittaa, tämä jakaumatulos pätee myös ilman sm:jien Y_i normaalisuusoletusta.

- Tämä tulos perustuu yhteen tilastotieteen kaikkein keskeisimmistä tuloksista eli **keskeisen raja-arvolauseeseen (KRL)**.

- KRL:n olennainen ja intuitiivinen tulkinta merkitsee satunnaismuuttujien otoskeskiarvon taipumusta noudattaa normaalijakaumaa riippumatta satunnaismuuttujan taustalla olevan otosjakauman muodosta. Tällä on merkitystä monessa yhteydessä, joista mm. alempana tarkasteltava luottamusvälin muodostaminen on yksi esimerkki. Ylipäätään tämä hyvin tärkeä tulos on merkittävä peruste normaalijakauman keskeiselle asemalle tilastotieteessä.
- KRL:n tarkempi tarkastelu vaatii jälleen selvästi enemmän käytyjä tilastotieteen (ja matematiikan) opintoja.

Esimerkki: Aritmeettisen keskiarvon otosjakauma normaalijakautuneen aineiston tapauksessa

Tarkastellaan seuraavaksi simuloidun esimerkin avulla aritmeettisen keskiarvon ominaisuuksia. Erityisesti klassisessa tilastotieteessä tilastollinen päättely, esimerkiksi luottamusvälien konstrulointi, perustuu ajatukselle toistetusta aineistonkeruusta. Tätä on helppoa havainnollistaa **simuloimalla** käyttäen moderneja tietoteknisiä ratkaisuja, joiden avulla voidaan generoida halutulla tavalla jakautuneiden satunnaismuuttujien realisaatioita.

- Aritmeettisen otoskeskiarvon tapauksessa tämä tarkoittaa sitä, miten realisoituneet otoskeskiarvot vaihtelevat satunnaisesti, kun otantaa samasta kohdepopulaatiosta toistetaan.
- Yllä esitetyn tuloksen nojalla otoskoko näyttelee keskeistä roolia otoskeskiarvojen vaihtelun suhteen.

Olkoon $Y_1^{(j)}, \ldots, Y_n(j)$ (j):s otos suuresta (äärettömästä) populaatiosta, jossa $j = 1, \ldots, J$ indeksoi eri otoksia. Oletetaan lisäksi, että

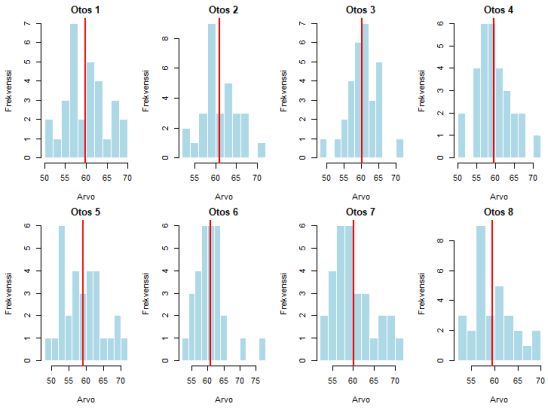
$$Y_i^{(j)} \sim \mathsf{N}(60, 25)$$

kaikille $j = 1, \ldots, J$ ja $i = 1, \ldots, n$ ja sm:jat (tilastoyksiköt) ovat keskenään riippumattomia kaikissa otoksissa. Toisin sanoen, kerätään (J) otosta kohdepopulaatiosta, joista jokaisessa tilastoyksiköt ovat samoin ja riippumattomasti jakautuneita.

Tiedetään edellä esitellyn pohjalta, että normaalisti jakautuneen (ja laajemminkin ilman jakaumaoletusta) populaation tapauksessa otoskeskiarvo on odotusarvon harhaton estimaattori (ts. että $\mathsf{E}(\bar{Y}) = \mu$) ja että sen otosjakauma on muotoa $\bar{Y} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$. Täten selvästikin **otoskoolla** on suuri merkitys sille kuinka keskittynyt aritmeettisen keskiarvon otosjakauma on. Suurella otoskoolla tulisi teoriassa siis saada parempia arvioita aineiston generoineen jakauman odotusarvosta!

Alla havainnollistetaan yllä kuvatun *toistetun aineistonkeruun* ajatusta aritmeettisen keskiarvon otosjakauman taustalla. Jakaumasta $\mathsf{N}(60, 25)$

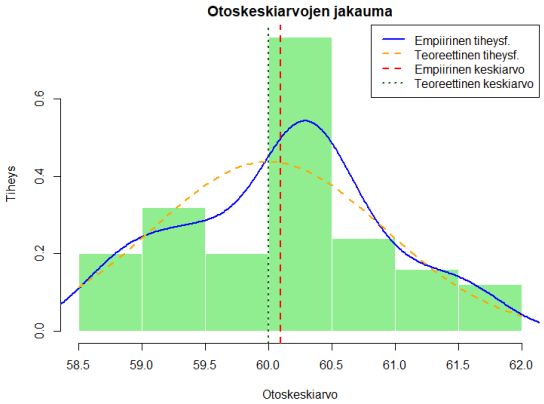
generoidaan (generoitiin) $J = 50$ otosta, joista jokaisen otoskoko on $n = 30$. Seuraavassa kuviossa on havainnollistuksen vuoksi esitetty ensimmäisen kahdeksan otoksen (yhteensä $J = 50$ otoksesta) histogrammit sekä näiden aritmeettiset keskiarvot (punaiset pisteet).



Kuva: Simulointikokeen kahdeksan ensimmäisen otoksen histogrammit ja niiden otoskeskiarvot (punaiset viivat).

Seuraavassa kuvassa on kuvattu kaikkien (J=50) otoksen otoskeskiarvojen histogrammi. Lisäksi on laskettu otoskeskiarvojen keskiarvo, jonka pitäisi harhattomuuden vuoksi olla lähellä populaatiotason odotusarvoa $\mu = 60$. Lisäksi on laskettu **otoskeskiarvojen varianssi**, jonka pitäisi saada arvo $\frac{\sigma^2}{n} = \frac{25}{30} \approx 0.83$. Estimaatteja ja teoreettisia normaalijakauman parametreja vastaavat tiheysfunktiot on piirretty kuvaan.

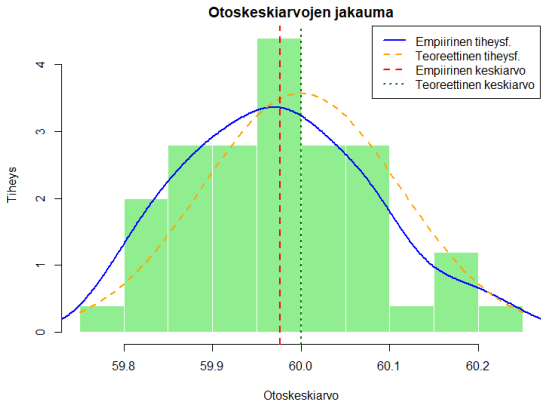
- Osoittautuu, että tässä simuloinnissa otoskeskiarvojen keskiarvo on 60.10 ja otoskeskiarvojen varianssi 0.629.



Kuva: Simulointikokeen 50 toiston otoskeskiarvojen jakauma ja normaalijakauman teoreettinen ja nyt saavutettu empiirinen tiheysfunktio. “Teoreettisella keskiarvolla” viitataan otoskeskiarvojen odotettavissa olevaan keskiarvoon, joka populaatiotasolla tulisi vastata odotusarvoa μ (tässä tapauksessa $\mu = 60$).

Mitä tapahtuu kun yksittäisen otoksen otoskokoa n kasvatetaan? Teorian perusteella otoskeskiarvojen otosjakauman varianssi pitäisi pienentyä! Alla olevassa kuviossa havainnollistetaan tätä. Jälleen generoidaan/generoitiin $J = 50$ otosta populaatiojakaumasta $N(60, 25)$, mutta tällä kertaa otoskokoa kasvatettiin tasolle $n = 2000$. Kuvassa on kuvattu kaikkien $J = 50$ otoksen keskiarvojen histogrammi sekä otoskeskiarvojen teoreettista otosjakaumaa ja vastaavaa estimoitua otosjakaumaa vastaavat tiheysfunktiot. Edelleen otoskeskiarvojen pitäisi harhattomuuden vuoksi olla lähellä arvoa $\mu = 60$, mutta tällä kertaa niiden varianssin pitäisi olla pienempi eli $\frac{\sigma^2}{n} = \frac{25}{2000} = 0.0125$.

- Otoskeskiarvojen keskiarvoksi saatiin 60.02 ja varianssiksi 0.0138. Havaitaan, että nyt siis estimoitu otoskeskiarvojen varianssi on pienempi kuin edellisessä ($n = 30$) esimerkissä ja lisäksi olemme jälleen lähempänä todellista oikeaa odotusarvoa! Huomaa, että kuviossa myös vaaka-akseli muuttuu eli se kapenee merkittävästi edelliseen kuvioon ja tilanteeseen verrattuna.



Kuva: Simulointikokeen otoskeskiarvojen jakauma kun otoskoko on isompi eli 2000 havaintoa ($n = 2000$).

Standardoidun aritmeettisen keskiarvon otosjakauma. Tarkastellaan **standardoitua** satunnaismuuttujaa

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

Tällöin (Z):n odotusarvo $\mathsf{E}(Z) = 0$ ja varianssi $\mathsf{Var}(Z) = 1$.

Jos lisäksi oletetaan $Y_i \sim \mathrm{N}(\mu, \sigma^2), i = 1, \ldots, n$, niin tällöin (Z) noudattaa standardoitua normaalijakaumaa:

$$Z \sim \mathrm{N}(0, 1).$$

Tällä tuloksella on tärkeä merkitys mm. kun tarkastellaan luottamusvälien muodostamista

- Jälleen voidaan osoittaa, että Z:taa koskeva jakaumatulos pätee asymptoottisesti (suurissa otoksissa) myös ilman sm:jien Y_i normaalisuusoletusta.

12.2 Suhteellisen frekvenssin otosjakauma

Frekvenssi ja suhteellinen frekvenssi. Oletetaan, että tapahtuman (A) todennäköisyys on

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

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

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

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

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

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

Ts. tässä pätee sama logiikka kuin yläpuolella otoskeskiarvon kohdalla ja normaalistijakautuneen satunnaismuuttujien tapauksessa. Näin ollen myös alapuolella käsiteltävissä (suhteellisen) frekvenssin otosjakauman käyttäymisessä on paljon samaa mitä aritmeettisen keskiarvon tapauksessa, mutta nyt toki merkinnät ja lopputulos koskee eri tilannetta.

Frekvenssin otosjakauma. Frekvenssillä f on odotusarvo

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

ja varianssi

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

(Luvussa aiemmin tehtyjen oletusten ollessa voimassa) frekvenssi f noudattaa binomijakaumaa parametrein n ja p:

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

Suhteellinen frekvenssi: Odotusarvo ja varianssi. Suhteellisen frekvenssin $\hat{p}$ odotusarvo

$$\mathsf{E}(\hat{p}) = \mathsf{E}\Big(\frac{f}{n}\Big) = p,$$

ja varianssi

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

Suhteellisen frekvenssin $\hat{p}$ standardipoikkeamaa

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

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

Suhteellisen frekvenssin otosjakauma. Koska $\mathsf{E}(\hat{p}) = p$ ja $\mathsf{Var}(\hat{p}) = \frac{pq}{n}$, niin suhteellisen frekvenssin otosjakauma keskittyy yhä voimakkaammin tapahtuman A todennäköisyyden $\mathrm{P}(A) = p$ ympärille, kun otoskoko n kasvaa.

- Tämän näkee erityisesti nollaa lähestyvästä suhteellisen frekvenssin varianssin lausekkeesta, kun havaintojen lukumäärä n kasvaa.

Jälleen suurten otosten tapauksessa voidaan myös osoittaa, että suhteellinen frekvenssi noudattaa em. oletusten pätiessä normaalijakaumaa:

$$\hat{p} \sim \mathrm{N}\Big(p, \frac{pq}{n}\Big).$$

Aritmeettisen keskiarvon tapaan standardoidulle satunnaismuuttujalle (ks. yllä) pätee

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

eli tämä standardoitu sm. noudattaa suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa.

Suomen EU-kansanäänestys (Muokattu esimerkki kirjasta Mellin (2004, s. 307)). Oletetaan, että juuri ennen Suomen EU-kansanäänestyksessä vuonna 1994 jäsenyyttä kannattaneiden suhteellinen osuus eräässä gallupissa oli 0.54 (54 %).

- Lopulta kansanäänestyksessä kyllä-äänteen (kannattajien) osuus oli 56.9 %.

Mikä olisi ollut tällöin todennäköisyys gallupiin perustuen, siis juuri ennen äänestystä, 200 havainnon otoksessa kyllä-osuus olisi ollut alle 50 %?

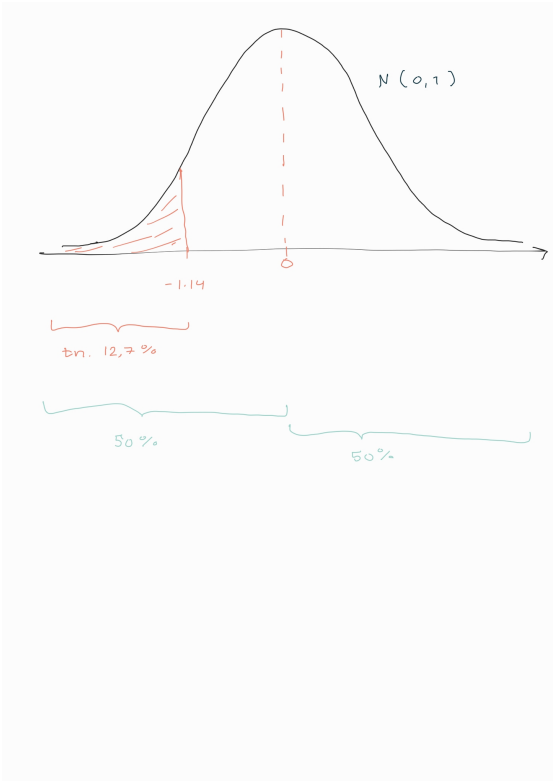
Suhteellisen frekvenssin otosjakauman perusteella kyllä-kannatusosuuden jakauma olisi

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

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

Näin ollen haluttu todennäköisyys (ts. saada sellainen satunnaismuuttujan $Z \sim \mathrm{N}(0, 1)$ arvo että suhteellinen osuus on pienempi kuin 0.5)

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



Kuva: Standardinormaalijakauman tiheysfunktion alle jäävä värjätty osuus kuvaa todennäköisyyttä, jolla saadaan sellainen realisaatio että EU:n kannatuksen suhteellinen osuus on pienempi kuin 50 %.

12.3 Luottamusvälit

Tilastotieteen yleisen idean mukaisesti satunnaisesti saadusta aineistosta laskettujen tunnuslukujen luotettavuus on tilastollisen mallin parametrien estimoinnissa keskeinen tilastollinen kysymys.

- Otoksen poimintaan liittyvän satunnaisvaihtelun vuoksi **emme voi varmuudella tietää** onko saatu otokseen perustuva parametriestimaatti "lähellä" vai "kaukana" sen todellisesta arvosta.
- Täten tarvitaan jokin tapa, jolla saadun parametriestimaatin luotettavuutta voidaan arvioida.

Luottamusväli

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

- Otoskoko on luottamusvälejä koskevissa tarkasteluissa keskeinen ja luottamusväleihin palataankin otoskoon käsittelyn yhteydessä.

Merkitsevyystaso ja luottamustaso

- Valittua **luottamustasoa** merkitään usein $1-\alpha$:lla, jossa **merkitsevyystaso** (**riskitaso**) α on esimerkiksi $\alpha = 0.05$ eli 5 % (vrt. hypoteesien testaamisen esittely aiemmin).
- **Tulkinta**: Jos **otantaa** jakaumasta $f(y;\theta)$ **toistetaan**, niin keskimäärin $100 \times (1-\alpha)\%$ otoksista konstruloiduista luottamusväleistä peittää parametrin θ todellisen arvon.

Virhemarginaali

Luottamusväli on kenties tunnetumpi kansankieliseltä nimitykseltään **virhemarginaali**, joka on itse asiassa luottamusvälin puolikas.

Kuten jatkossa tullaan havaitsemaan, virhemarginaalin suuruuteen vaikuttavat otosasetelma, otoskoko, luottamustaso ja tutkittavan tilastollisen tunnusluvun jakauma. Normaalisti mm. otoskoon kasvu pienentää virhemarginaalia.

Luottamusväleissä ei varsinaisesti ole kyse "virheestä" vaan saadun/muodostetun tiedon tarkkuudesta.

Luottamusvälit, eli virhemarginaalit, siis (yleisesti) riippuvat valittavasta luottamustasosta $1-\alpha$ ja näin ollen samasta aineistosta on saatavissa useita virhemarginaaleja. Täten on tarkalleen ottaen virheellistä sanoa, että "tutkimuksen virhemarginaali on 3,5% puoleen tai toiseen". Oikeammin olisi sanoa esimerkiksi "tutkimuksessa saadun kannatuksen virhemarginaali on 3,5 % puoleen tai toiseen 95 % luottamustasolla".

- 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 valikoituminen ja vastauskato

Normaalijakauman odotusarvon luottamusväli. Käsittelemme seuraavassa (normaalijakauman) odotusarvon μ luottamusväliä. Ellei toisin mainita, oletetaan että taustalla oleva populaatio, N, on "iso" (ääretön). Näin ollen ns. äärellisyyskorjausta ei käytetä (yksinkertaisuuden vuoksi).

Tarkastellaan siis satunnaisotosta normaalijakaumasta

$$Y_1, \ldots, Y_n \perp\!\!\!\perp,\ Y_i \sim \mathrm{N}(\mu, \sigma^2),\ i = 1, \ldots, n,$$

ja erityisesti normaalijakauman odotusarvon μ luottamusvälin määräämistä otannan avulla olettaen että jakauman varianssi σ^2 on tunnettu. Lisäksi muistetaan että normaalijakauman odotusarvoparametrin $\mathsf{E}(Y_i) = \mu$ harhaton estimaattori on aritmeettinen keskiarvo

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

Valitaan **luottamustasoksi** $1-\alpha$, jossa α on siis valittu merkitsevyystaso. Yleinen valinta ihmistieteissä on $\alpha = 0.05$ tai $\alpha = 0.1$ vastaten 95 % ja 90 % prosentin luottamustasoa. Luonnontieteissä ja lääketieteellisissä sovelluksissa α on usein pienempi, kuten $\alpha = 0.01$.

Määrätään **luottamuskertoimet** $-z_{\alpha/2}$ ja $z_{\alpha/2}$ (luottamusväli on ns. kaksisuuntainen), joille pätee

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

jossa standardoitu satunnaismuuttuja (ks. aiemmat määritelmät edellä)

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

ja $\mathrm{P}(\cdot)$:llä merkitään todennäköisyyttä, joka tässä tapauksessa liittyy normaalijakaumaan, ja $z_{\alpha/2}$ on jakaumafunktion arvo pisteessä $\alpha/2$.

- Ks. aiemmat alaluvut tässä luvussa osoittaen, että (tehdyillä oletuksilla) Z noudattaa $\mathrm{N}(0,1)$-jakaumaa.

Etsitään siis odotusarvoparametrille μ sellainen arvo, jolla oheinen epäyhtälö pätee ja siten päädytään luottamusväliin. Nyt ylläoleva epäyhtälöketju voidaan kirjoittaa muodossa

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

joka voidaan kirjoittaa uudelleen muodossa

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

kertomalla nimittäjällä puolittain ja vähentämällä sm:jien keskiarvo molemmin puolin. Normaalijakauman odotusarvon $(1-\alpha)\times 100\%$ luottamusväli on siis

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

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

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

Luottamusvälin pituus on

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

Näin ollen (tässä tilanteessa) virhemarginaali on luottamusvälin pituuden puolikas eli

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

- Huomaa, että edellä tiettyyn otokseen liittyvä luottamusväli perustetaan tietysti realisoituneeseen otoskeskiarvoon $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Olisi toivottavaa pystyä konstruoimaan parametrille μ 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 μ todellisesta arvosta tulee epätarkemmaksi.
  - Luottamusvälin lyhentäminen pienentää luottamustasoa, jolloin tieto parametrin μ todellisesta arvosta tulee epävarmemmaksi.

Normaalijakauman odotusarvon luottamusväli kun varianssi σ^2 on tuntematon. Normaalijakauman odotusarvon $(1-\alpha)\times 100\%$ luottamusväli on nyt

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

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

- (Studentin) t-jakauma muistuttaa silmämääräisesti normaalijakaumaa, mutta se on **paksuhäntäisempi**. Vapausasteluvun kasvaesssa t-jakauma lähestyy normaalijakaumaa.
- Suurissa otoksissa (n iso) luottamuskertoimet voidaan poimia (approksimatiivisesti) myös normaalijakaumasta eli korvata edellä kertoimet $t_{\alpha/2}$ aiemmin käytetyillä kertoimilla $z_{\alpha/2}$.
- Normaalijakauman odotusarvon luottamusväli (σ^2 tuntematon), t-jakauma eri vapausastein df

t-jakauma

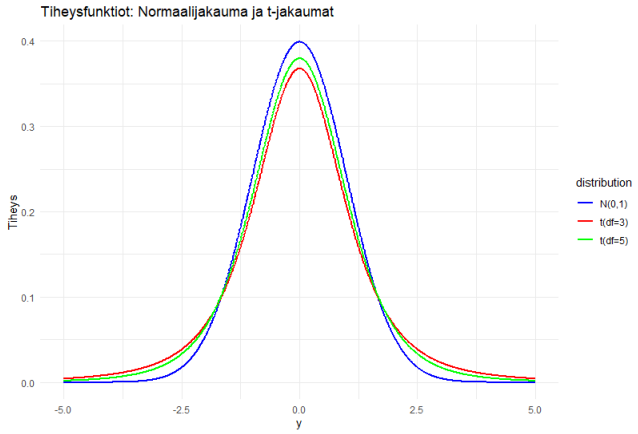


Figure 12.1: .

Kuva: Standardoidun normaalijakauman ja t-jakauman tiheysfunktioita t-jakauman kahteen eri vapauasteeseen liittyen. Vapausasteluvun kasvaessa t-jakauma lähestyy normaalijakaumaa.

Luottamusväli: Suhteellisen osuuden odotusarvo. Käsittelemme seuraavassa siis suhteellisen osuuden p luottamusvälejä. Tässä tilanteessa tarkastellaan siis 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 $p = \mathsf{E}(Y_i)$ harhaton estimaattori on tapahtuman A suhteellinen otosfrekvenssi

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

Bernoulli-jakauman (vrt. binomijakauma) ominaisuuksien perusteella $\mathsf{E}(Y_i) = p$ ja $\mathsf{Var}(Y_i) = pq$, jossa $q = 1 - p$.

Näin ollen voimme normaalijakauman odotusarvoparametrin luottamusvälin konstruloinnin tapaan määritellä satunnaismuuttujan Z:

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

joka noudattaa (suurissa otoksissa) $\mathrm{N}(0, 1)$-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 määrätään ylläpuolella nähtyyn tapaan:

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

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

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

ja luottamusväli voidaan kirjoittaa

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

ja luottamusvälin pituus on (ja tämän puolikas siis virhemarginaali)

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

12.4 Bootstrap ja Monte Carlo-menetelmät

Tilastotieteessä käytännön sovelluksissa turvaudutaan usein laskennallisiin menetelmiin, kun analyyttiset ratkaisut ovat hankalia tai mahdottomia. Tällaisia menetelmiä käytetään esimerkiksi juuri vaihtoehtona edellä esiteltyjen

luottamusvälien muodostamiselle tai testisuureiden jakaumien arvioinnissa. Yksi keskeinen ja laajasti käytetty menetelmä on **bootstrap**, eli **uusio-otanta**. Emme käsittele tätä tällä kurssilla tarkemmin, mutta ohessa kuitenkin joitain keskeisiä yleisiä huomioita.

Bootstrap

Bootstrap ("**uusio-otanta**") on menetelmä, jolla luodaan mm. luottamusvälejä uudelleennäytteistämällä havaittua dataa sen sijaan, että oletettaisiin todennäköisyysmalli taustalla olevalle satunnaismuuttujalle.

Bootstrap perustuu ajatukseen, että havaittu otos sisältää riittävästi tietoa perusjoukosta, jotta siitä voidaan simuloida uusiotanta-aineistoja. Käytännössä tämä tarkoittaa sitä, että alkuperäisestä aineistosta poimitaan satunnaisotoksia takaisinpanolla (eli sama havainto voi esiintyä useita kertoja yhdessä bootstrap-otoksessa), ja näistä uusista otoksista lasketaan kiinnostavia tilastollisia tunnuslukuja, kuten keskiarvoja, mediaaneja tai regressiomallien tapauksessa parametriestimaatteja ja niiden keskivirheitä. Näin saadaan empiirinen arvio ko. tunnuslukujen jakaumista.

Bootstrap ei ole yksi yksittäinen menetelmä, vaan pikemminkin yleistermi useille eri uusio-otantatekniikoille. Näitä ovat esimerkiksi (tämä ei ole suinkaan täydellinen lista):

- perinteinen (ei-parametrinen) bootstrap,
- parametrinen bootstrap, jossa oletetaan jokin jakauma ja simuloidaan siitä,
- block bootstrap, jota käytetään aikasarjoissa säilyttämään riippuvuusrakenne.

Käytettävä tarkempi menetelmä tulee aina valita aineiston rakenteen ja tutkimuskysymyksen mukaan, ja sen valinta on syytä perustella huolellisesti.

Monte Carlo-menetelmät

Bootstrap ja Monte Carlo -menetelmät. Bootstrap on esimerkki laajemmasta laskennallisten menetelmien perheestä, jota kutsutaan usein nimellä Monte Carlo -menetelmät. Näiden menetelmien ytimessä on satunnaissimulointi: toistetaan suuri määrä satunnaisia kokeita tai otoksia ja tarkastellaan tulosten jakaumia ja eri tunnuslukuja.

Monte Carlo -menetelmät tarjoavat tilastotieteelle ja data-analyysille monia etuja:

- Joustavuus: Ne eivät (välttämättä) vaadi tiukkoja oletuksia jakaumista tai mallin muodosta.

- Sovellettavuus monimutkaisissa tilanteissa: Esimerkiksi silloin, kun analysoidaan epälineaarisia malleja, pieniä otoksia tai epäsymmetrisiä jakaumia.
- Visualisointi ja ymmärrettävyys: Simuloidut jakaumat voidaan helposti visualisoida, mikä helpottaa tulosten tulkintaa ja viestintää.
- Monte Carlo -menetelmät ovat keskeisiä myös bayesilaisessa tilastotieteessä, jossa posteriorijakaumia arvioidaan usein simuloimalla (ks. esim. ns. Markov Chain Monte Carlo (MCMC) menetelmät).

12.5 Otoskoko

Tilastollisen päättelyn keskeinen tavoite on siis yleistää otoksen pohjalta tehty päättely koskemaan koko perusjoukkoa. Seuraavaksi käymme läpi seikkoja, jotka tulee ottaa huomioon **otoskokoa** (eli merkinnöissä n) miettiessä.

Kun on päätetty, millainen tutkimusaineisto halutaan kerätä, on päätettävä, kuinka suuri otoksen on oltava, jotta se edustaa tutkittavaa perusjoukkoa kattavasti.

- Liian pieni otos, eli pieni määrä otokseen poimittuja tilastoyksiköitä, voi **sattumalta** poiketa paljonkin perusjoukosta. Tämä ns. **otantavirhe** on sitä suurempi mitä pienempää otosta käytetään. Liian pienen otoksen vuoksi muuten hyvin toteuttu tutkimus- ja otanta-asetelma saattaa epäonnistua vastaamaan tutkimuksen mielenkiinnon kohteena olevaan kysymykseen.
- Vastaavasti 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 määrätä etukäteen.

Perusjoukon rooli otoskoon määrittämisessä. Ensiksi tulee 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 siististi otoskoon kasvaessa, kuten edellä nähtiin.
- Kuuluuko tutkimukseesi esimerkiksi otoksen sisällä olevien ryhmien keskiarvojen vertailua tai muita otoksen osajoukkojen tunnuslukujen vertailua? Jos kuuluu, niin otoskoko tulee valita pienimmän ryhmäkoon 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 (arvioida) tulosten halutun tarkkuuden avulla. Tarkastellaan kuitenkin ensin minkälaiset kysymykset liittyvät otoskoon pohdintaan tulosten tarkkuuden osalta.

- Kuinka varma sinun on oltava, että tulokset ovat täsmällisiä? Tämä on virhemarginaalisi!

Esimerkiksi vaalikannatuksen arvioimiseen 2 % virhemarginaalilla riittää huomattavasti pienempi otoskoko kuin 0.2 % virhemarginaalilla toimittaessa. Politiikan tutkija voisikin kasvattaa otoskokoa vaalien lähestyessä, mikäli mielii saada tarkempia tuloksia.

- Kuinka varma haluat olla, että otos edustaa joukkoa oikein? Tämä on luottamustaso, josta keskusteltiin jo luottamusvälin määräämisen yhteydessä!

Odotetun vastauskadon vaikutus otoskokoon. **Kuinka suuri vastauskato tulee mahdollisesti olemaan?**

- Yleensä osa kyselytutkimukseen valituista jättää vastaamatta eli syntyy **vastauskatoa**. Kato vinouttaa otosta, jos vastaamatta jättäneet ovat mielipiteiltään erilaisia kuin vastanneet.

- Otoskoon kasvattaminen ei paranna kadon aiheuttamaa vinoutumista.

Esimerkki: Jos Alkon myymälän asiakastutkimus suoritetaan ovensuukyselynä 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 asiakkailla!

Esimerkki: Otoskoko normaalijakauman odotusarvon estimoinnissa. Palautetaan mieleen normaalijakauman $\mathrm{N}(\mu, \sigma^2)$ odotusarvon luottamusvälin määrääminen (kun varianssi σ^2 oletetaan tunnetuksi). Luottamusväliksi saatiin edellä johdetuksi

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

j a luottamusvälin symmetrisyydestä johtuen luottamusvälin pituus

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

Oletetaan, että normaalijakauman odotusarvoparametrille μ halutaan konstruoida luottamusväli, jonka toivottu pituus on $2d$ (eli luottamusvälin pituus $= 2d$. Huomio siis luottamusvälin symmetrisyys). Luottamusvälin pituuden lausekkeesta (kun se asetetaan yhtäsuureksi $2d$:n kanssa) saadaan yhtälöä järjestelemällä lauseke

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

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

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

- Yksinkertaisuuden vuoksi pitäydytään tässä tapauksessa luottamuskertoimissa $z_{\alpha/2}$ vaikka varianssi σ^2 olisikin tuntematon.

Esimerkki: **Käytännön otoskoon määrääminen**. Oletetaan, että haluamme määrätä otoskoon niin, että otoskeskiarvo poikkeaa populaatiokeskiarvosta 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

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

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

jossa n on edelleen otoskoko, N perusjoukon koko.

- 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 estimoinnin tarkkuutta, ja juuri äärellisyyskertoimen avulla hajonta "korjataan" vastaamaan käytettyä otoskokoa.

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

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

Tarvittava otoskoko on tällöin (välivaiheiden jälkeen)

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

Esimerkki: otoskoko (jatkoa). Oletetaan aiemman esimerkin tilanne kuitenkin siten, että perusjoukon koko on nyt $N = 100$. 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. Palauta mieleen yläpuolelta Bernoullijakauman 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 äärettömäksi)

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

Esimerkki: Otoskoko ja suhteellinen osuus. Geologi haluaa arvioida kallion kultapitoisuuden ottamalla kivinäytteen n:tä 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 kullan suhteellinen osuus on tuntematon, joten p:lle asetetaan arvio $p = 0.5$.
- Äärellisyyskorjaus voidaan unohtaa, sillä näytteenottopisteiden pinta-alat ovat pieniä (eli niitä on äärettömän paljon, ts. tarkasteltavaan populaatioon niitä sisältyy hyvin suuri määrä).

Tällöin otoskoko

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

Suhteellinen osuus, äärellisyyskorjaus. Tarvittava otoskoko saadaan äärellisyyskorjausta käytettäessä kaavasta

$$n = \frac{Np(1-p)}{\frac{(N-1)d^2}{z^2_{\alpha/2}} + p(1-p)}.$$

- Voidaan osoittaa, että jos perusjoukko N on iso (ääretön), niin tällöin tämä edellinen lauseke supistuu aiempaan otoskokoa osoittavaan lausekkeeseen.
- Usein otoskokoa määrättäessä suhteellisesta osuudesta ei ole olemassa arviota. Tällöin suhteellisen osuuden p arvoksi asetetaan useimmiten $p = 0.5$, jolloin suhteellisen osuuden varianssi on suurin.

Chapter 13

Regressioanalyysi

Tilastollinen riippuvuus ja korrelaatio -jakson laajennuksena pyrimme tässä luvussa vastaamaan seuraavaan kysymykseen:

Miten jonkin selitettävän muuttujan tilastollista riippuvuutta joistakin toisista selittäviksi muuttujiksi kutsutuista muuttujista voidaan mallintaa?

Muuttujien välisten riippuvuuksien, eli erilaisten tosielämän ilmiöiden välisten 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 säännönmukaisuutta ja muuttujien välillä näyttäisi olevan järkevä (asialooginen) yhteys, niin päästään "malliajatteluun". Ts. pyritään rakentamaan tilastollista mallia ko. aineistolle, mikä valitun kriteeristön perusteella parhaiten kuvaa analysoitavaa aineistoa..

13.1 Johdatus regressioanalyysin ideaan

Regressioanalyysi pyrkii havaintoaineiston perusteella mallintamaan tilastoyksikköjen tilastollisten muuttujien välistä riippuvuutta.

Selitettävä ja selittävät muuttujat

- Regressiomallissa tilastollisia muuttujia on kahdenlaisia: **selitettävä muuttuja**, jonka tilastollista vaihtelua pyritään selittämään **selittävän muuttujan**, tai **selittävien muuttujien**, avulla.
- Toisin sanoen, pyritään erottamaan se selitettävän muuttujan arvojen vaihtelu, joka voidaan selittää selittävän muuttujan (selittävien muuttujien)

arvojen vaihtelulla siitä vaihtelusta, joka on täysin satunnaista. Jälleen on siis kysymys signaalista ja kohinasta!

Esimerkiksi voitaisiin tutkia selittääkö vaaleissa puolueiden/ehdokkaiden vaalimainontabudjetti heidän äänimääriään, ja jos selittää, niin kuinka suuren osan äänimääristä?

Jos **tilastollisesti merkitsevä osa** selitettävän muuttujan havaittujen arvojen vaihtelusta voidaan selittää selittävien muuttujien arvojen vaihtelun avulla, sanomme, että selitettävä muuttuja **riippuu tilastollisesti** merkitsevästi selittäjinä käytetyistä muuttujista.

Yleisemmin regressioanalyysi pyrkii vastaamaan seuraaviin kysymyksiin koskien tilastollisten muuttujien välistä riippuvuutta:

- Muuttujien välisten **riippuvuuksien kuvaaminen**. Millainen on riippuvuuden muoto (kuten lineaarinen vai epälineaarinen)? Kuinka voimakasta riippuvuus on?
- Muuttujien välisten **riippuvuuksien selittäminen**. Tilastollisen riippuvuuden luonteen kuvaaminen ja tiivistäminen.
- Selitettävän muuttujan käyttäytymisen **ennustaminen**.

Lineaarinen regressioanalyysi (teknisesti) rajoittuu muuttujien *lineaaristen* riippuvuuksien kuvaamiseen. Kuitenkin, laajemmin asiaa pohdittaessa, lineaaristen regressiomallien suuri käyttökelpoisuus muuttujien välisten riippuvuuksien tilastollisessa analyysissa perustuu (ainakin) seuraaviin seikkoihin:

- Lineaarisella regressiomallilla voidaan ajoittain kohtuullisella tarkkuudella approksimoida (siis jossain määrin, malli on toki virheellinen) epälineaarisiakin muuttujien välisiä riippuvuuksia.
- Muuttujien välinen epälineaarinen riippuvuus voidaan usein myös linearisoida käyttäen sopivia muunnoksia alkuperäisiin muuttujiin.
- Epälineaariset regressiomallit muodostavat oman tilastollisten (regressio)mallien luokkansa (joita ei käsitellä tarkemmin vielä tällä kurssilla, mutta kylläkin myöhemmissä tilastotieteen opinnoissa).

Regressiomalleja käytetään apuvälineinä monilla tilastotieteen osa-alueilla. Esimerkkejä regressiomallien käyttökohteista tilastotieteessä:

- Varianssianalyysi
- Koesuunnittelu
- Biometria/biostatistiikka
- Aikasarja-analyysi ja ennustaminen
- Ekonometria

Regressioanalyysissa sovellettavat tilastolliset mallit voidaan luokitella usealla eri periaatteella.

- Luokittelu regressiomallin funktionaalisen muodon mukaan:
  - Lineaariset regressiomallit
  - Epälineaariset regressiomallit
- Luokittelu regressiomallin yhtälöiden lukumäärän mukaan:
  - Yhden yhtälön regressiomallit
  - Moniyhtälömallit

Yhden selittäjän lin. malli Tällä kurssilla käsitellään seuraavaksi käytännössä vain **yhden selittäjän lineaarista regressiomallia**. Myöhemmin esitellään kuitenkin lyhyesti minkälaisia laajennuksia tälle regressioanalyysin perustilanteelle tyypillisesti hyödynnetään tilastollisissa analyyseissä.

13.2 Yhden selittäjän lineaarinen regressiomalli

Yhden selittäjän lineaarinen regressiomalli pyrkii selittämään selitettävän muuttujan havaittujen arvojen vaihtelua yhden selittävän muuttujan havaittujen arvojen vaihtelun avulla. Se on siis yksinkertaisin esimerkki yhden selittäjän lineaarisista regressiomalleista, sillä se sisältää vain yhden selittävän muuttujan useamman sijaan.

- Selitettävää muuttujaa kutsutaan usein myös *vastemuuttujaksi, vasteeksi, riippuvaksi muuttujaksi tai tulosmuuttujaksi*
- Vastaavasti selittävää muuttujaa kutsutaan ajoittain myös *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 tavanomaiset oletukset?
- Miten mallin parametrit estimoidaan?
- Miten mallin parametreja koskevia hypoteeseja testataan?
- Miten mallin hyvyyttä 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.

- Kyseinen muuttuja voidaan myös tulkita satunnaismuuttujana eikä seuraavat tarkastelut muutu ratkaisevasti tämän seurauksena.

Tehdään siis seuraavat oletukset:

i) Selitettävä muuttuja Y on suhdeasteikollinen satunnaismuuttuja.
ii) Selittävä muuttuja x on kiinteä eli ei-satunnainen muuttuja.

Olkoot $y_1, y_2, \ldots, y_n$ selitettävän muuttujan Y ja $x_1, x_2, \ldots, x_n$ selittävän muuttujan x havaittuja arvoja. Oletetaan lisäksi, että havaintoarvot x_i ja y_i liittyvät samaan havaintoyksikköön kaikille $i = 1, 2, \ldots, n$, eli tarkastellaan havaintopareja (x_i, y_i).

- Matemaattisemmin tämä tarkoittaa sitä, että tällöin havaintoarvot muodostavat pisteitä 2-ulotteisessa avaruudessa (x_i, y_i).

Yhden selittäjän lineaarinen regressiomalli. Oletetaan, että havaintoarvojen y_i ja x_i välillä on **lineaarinen tilastollinen riippuvuus**, joka voidaan ilmaista yhtälöllä

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

Tämä yhtälö määrittelee yhden selittäjän lineaarisen regressiomallin, jossa

- y_i on selitettävän satunnaismuuttujan Y_i havaittu arvo havaintoyksikölle i.
- x_i selittävän muuttujan eli selittäjän x ei-satunnainen ja havaittu arvo havaintoyksikölle i.
- ε_i on **virhetermi** ja sen satunnainen ja ei-havaittu arvo havaintoyksikölle (i). Se vastaa siitä osuudesta sm:jan Y_i satunnaisvaihtelusta mitä mallin systeemaattinen osa $\beta_0 + \beta_1 x_i$ ei kykene mallintamaan.

Yhden selittäjän lineaarisen regressiomallin regressiokertoimista:

- β_0 on ei-satunnainen ja tuntematon vakio, ja sitä kutsutaan **vakioselittäjän** regressiokertoimeksi. Nimitys johtuu siitä, että kerrointa β_0 vastaa keinotekoinen selittäjä, joka saa kaikille havaintoyksiköille $i = 1, 2, \ldots, n$ vakioarvon 1. Tämä lisätään malliin olennaisesti aina (ellei ole jotain todella erityisiä syitä toimia toisin).
  - Huomautus: Jatkossa esitettävät kaavat eivät välttämättä päde esitettävässä muodossa, jos mallissa ei ole vakiota (vakioselittäjää), joka yleensä automaattisesti lisätään aina mukaan malliin.
- β_1 on ei-satunnainen ja tuntematon vakio, ja siis selittäjään x liittyvä regressiokerroin, jota kutsutaan ajoittain myös kulmakertoimeksi (se määrittelee regressiosuoran kulmakertoimen).

– Huomautus: Huomaa, että regressiokertoimet β_0 ja β_1 on oletettu samoiksi kaikille havaintoyksiköille i.

Regressiokerroin

Regressiokertoimella tarkoitetaan tilastollisen mallin (lopulta aineiston perusteella estimoidun) parametrin arvoa, joka ilmaisee selittävän muuttujan ja vastemuuttujan välisen suhteen voimakkuuden regressiomallissa. Kertoimen tulkinta vaihtelee sen mukaan, onko vastemuuttuja jatkuva muuttuja (kuten yllä lineaarisessa regressiomallissa), binäärinen tai ajoittain myös osuus (logistinen regressio) tai esimerkiksi lukumäärä (Poisson-regressio).

Regressiokerroin kuvaa, kuinka paljon ja mihin suuntaan selittävän muuttujan muutos vaikuttaa vastemuuttujaan.

- Esimerkiksi lineaarisessa regressiossa (kuten edellä) lopulta estimoitu regressiokerroin $\hat{\beta}_1$ kertoo, kuinka paljon vastemuuttujan arvo muuttuu, kun selittävä muuttuja x kasvaa yhdellä yksiköllä, pitäen muut muuttujat vakiona (=ennallaan). Yhden selittäjän tapauksessa tämä toteutuu tietysti automaattisesti.
- Logistisessa regressiossa (binääriselle vastemuuttujalle) kerroin kuvaa todennäköisyyden muutosta, kun taas Poisson-regressiossa se kuvaa odotettujen tapahtumien lukumäärän muutosta.

Keskustellaan hetken aikaa **virhetermin ε_i roolista**. Havaintoyksikkökohtaisista (huom riippuvat indeksistä i!) virhetermeistä ε_i tehdään oletuksia, jotka liittyvät tilastollisen mallin rakentamiseen. Ns. standardioletukset ovat seuraavat:

i) $\mathsf{E}(\varepsilon_i) = 0$, $i = 1, 2, \ldots, n$.
ii) Virhetermeillä on vakiovarianssi eli ne ovat homoskedastisia: $\mathsf{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$. Virhetermien ε_i tässä yhteiseksi oletettua varianssia kutsutaan ajoittain jäännösvarianssiksi.
iii) Virhetermit ovat korreloimattomia: $\mathsf{Cov}(\varepsilon_i, \varepsilon_l) = 0$, $i \neq l$, eli niiden välinen kovarianssi on siis 0.
iv) Lisäksi ajoittain tehdään normaalisuusoletus eli että virhetermit ovat normaalisti jakautuneita: $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $i = 1, 2, \ldots, n$.

- Huomautus: Oletus (iv) sisältää oletukset (i) ja (ii) (tämä liittyy normaalijakauman ominaisuuksiin)
- Oletus (i) on keskeisin. Sen sijaan kaikki oletukset (ii)–(iv) eivät ole (yhdessä) välttämättömiä ja melko usein eri tilastotieteen osa-alueilla ja sovelluksissa virhetermistä tehdään vähemmän tai lievempiä oletuksia mitä (ii)–(iv) olettavat.

Regressioanalyysille voidaan esittää kaksi asialoogisesti varsin erilaista lähtökohtaa, joilla on kuitenkin myös monia yhtymäkohtia:

i) Ongelmat deterministisen mallien (ks. aiempi keskustelu deterministisen ja stokastisten mallien välillä) 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 tavanomaisen lineaarisen mallin avulla.
ii) Tavoitteena on tarkemmin ottaen moniulotteisen todennäköisyysjakauman regressiofunktion parametrien estimointi. Vaikka moniulotteisten todennäköisyysjakaumien regressiofunktiot ovat yleisesti epälineaarisia, lineaariset regressiomallit muodostavat tärkeän ja paljon sovelletun malliluokan.

Parametrien estimointi

Koska regressiokertoimet β_0 ja β_1 sekä jäännösvarianssi σ^2 ovat (tavallisesti) tuntemattomia, niiden arvot on **estimoitava** muuttujien x ja Y havaittuja arvoja x_i ja y_i, $i = 1, 2, \ldots, n$, käyttäen.

Lineaaristen regressiomallien parametrien estimointiin käytetään tavallisesti **pienimmän neliösumman (PNS) menetelmää**. Tämän estimointimenetelmän tarkemmat yksityiskohdat ovat myöhempien tilastotieteen kurssien asioita. Seuraavassa kuitenkin muutamia yksityiskohtia mihin PNS-menetelmä perustuu yhden selittäjän tapauksessa.

PNS-menetelmässä edellä esitellyn yhden selittäjän lineaarisen regressiomallin regressiokertoimien β_0 ja β_1 estimaattorit määrätään minimoimalla virhetermien ε_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 β_0 ja β_1 suhteen.

- Tämä minimointi tapahtuu tavanomaiseen tapaan derivoimalla funktio $S(\beta_0, \beta_1)$ kertoimien β_0 ja β_1 suhteen ja merkitsemällä derivaatat nolliksi:

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

Nämä ns. **normaaliyhtälöt** johtavat lopulta pienen sieventämisen jälkeen regressiokertoimien β_0 ja β_1 pienimmän neliösumman (PNS-) estimaattoreihin ja lopulta käytännössä analysoitavasta aineistosta laskettaviin PNS-estimaatteihin

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \text{ja} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}, \end{aligned}$$

josta nähdään yhteys aiemmin Osassa I esiteltyihin x:n ja y:n otoskeskiarvoihin, otoskeskihajontoihin sekä otoskovarianssiin ja otoskorrelaatiokertoimeen x:n ja y:n välillä.

PNS-estimaatit $\hat{\beta}_0$ ja $\hat{\beta}_1$ määrittelevät suoran (matemaattisesti katsoen) kaksiulotteisessa avaruudessa:

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

jossa

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

Sovitteet - Tämän estimoidun suoran tuottamat arvot $\hat{y}_i$ ovat käytännössä eri havainnoille y_i saatavat **sovitteet** $(i = 1, ..., n)$ valittuun lineaariseen malliin perustuen. Näiden perusteella voidaan osaltaan tehdä estimoidusta mallista tulkintoja ja tutkia myös mallin sopivuutta aineistoon.

Estimoidun mallin tulkintoja

Tarkastellaan estimoidun (yhden selittäjän) lineaarisen mallin seurauksia. Sijoitetaan regressiokertoimien β_0 ja β_1 PNS-estimaattoreiden lausekkeet estimoidun regressiosuoran lausekkeeseen. Tällöin estimoidun regressiosuoran yhtälö voidaan kirjoittaa muodossa:

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

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

- i) Jos $r_{xy} > 0$, suora on nouseva.
- ii) Jos $r_{xy} < 0$, suora on laskeva.
- iii) Jos $r_{xy} = 0$, suora on vaakasuorassa.
- iv) Suora jyrkkenee (loivenee), jos

- korrelaation itseisarvo $|r_{xy}|$ kasvaa (pienenee)
- keskihajonta s_y kasvaa (pienenee)
- keskihajonta s_x pienenee (kasvaa)

Residuaalit Tarkastellaan vielä estimoituun lineaariseen malliin liittyviä **residuaaleja**, jotka saadaan havaintojen ja sovitteiden erotuksena

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

- Sovite on estimoidun regressiosuoran yhtälön selitettävälle muuttujalle antama arvo havaintopisteessä x_i. Vastaavasti residuaali on selitettävän muuttujan havaitun arvon y_i ja sovitteen $\hat{y}_i$ eli estimoidun regressiosuoran yhtälön selitettävälle muuttujalle havaintopisteessä x_i antaman arvon erotus.
- Estimoitu regressiomalli selittää selitettävän muuttujan havaittujen arvojen vaihtelua sitä paremmin mitä lähempänä estimoidun mallin sovitteet $\hat{y}_i$ ovat selitettävän muuttujan havaittuja arvoja y_i. Yhtäpitävästi estimoitu regressiomalli selittää selitettävän muuttujan havaittujen arvojen y_i vaihtelua sitä paremmin mitä lähempänä nollaa estimoidun mallin residuaalit $\hat{\varepsilon}_i$ ovat.

Selittäjän til. merkitsevyys **Selittäjän tilastollinen merkitsevyys**. Yhden selittäjän lineaarisessa mallissa erityisen mielenkiinnon kohteena on (ks. esim. ylläolevat tulkinnat) testata nollahypoteesin

$$H_0 : \beta_1 = 0$$

paikkansa pitävyyttä. Tätä hypoteesia voidaan testata ns. **t-testillä**, mikä perustuu mm. R:n tai RStudion tuottamiin t-arvoihin ("t values"), jotka saadaan kertoimen β_1 estimaatin $\hat{\beta}_1$ ja sen keskivirheen (eli estimoidun hajonnan) osamääränä. Jos saatava t-testisuureen arvo on itseisarvoltaan verrattaen suuri, erityisesti itseisarvoltaan suurempi kuin 2, niin ko. tilastollisen testin p-arvo on pienempi kuin 5 %, ja selittäjä x on täten tilastollisesti merkitsevä selittävä muuttuja 5 % merkitsevyystasolla.

Selitysaste Liittyen vielä estimoidun mallin sopivuuden tarkasteluun, estimoidun regressiomallin hyvyyttä mitataan (tavanomaisesti) mm. **selitysasteella** (R^2).

- Selitysasteen määritelmä perustuu ns. varianssianalyysihajotelmaan, jossa selitettävän muuttujan havaittujen arvojen vaihtelua kuvaava neliösumma on jaettu kahdeksi neliösummaksi, joista toinen kuvaa 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 selitettävän muuttujan havaittujen arvojen vaihtelua.

- Vastaavasti arvo $R^2 = 1$ tarkoittaa, että malli sopii täydellisesti aineistoon. Ts. selitysaste mittaa lineaarisen regressiomallin selittämää osuutta selitettävän muuttujan havaittujen arvojen kokonaisvaihtelusta.

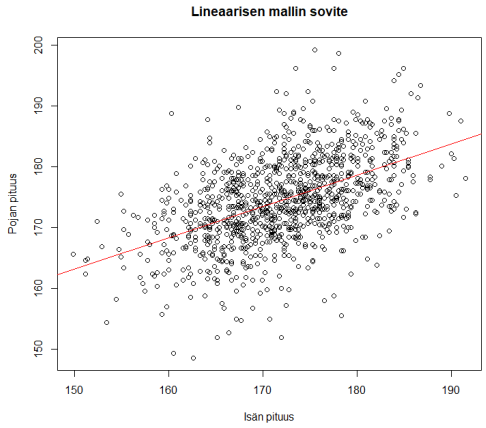
- Korkea selitysasteen arvo on siis sinänsä usein toivottava lopputulos lineaarisen mallin käytön yhteydessä. Tämän liian mekaaninen tavoittelu johtaa kuitenkin ajoittain muihin ongelmiin, kuten **ylisovittamiseen** usean selittäjän lineaarisia malleja käsiteltäessä.

Esimerkki: isän ja pojan pituus, jatkoa. Jatketaan isän (x) ja heidän poikiensa (y) pituutta koskevan aineiston tarkastelua (ks. Osa I). Periytyykö isän pituus heidän pojilleen? Käytännössä jo aiemmin tarkastelimme Pearsonin klassista havaintoaineistoa isän ja heidän poikiensa pituuksien muodostamista lukupareista.

Estimoidun regressiosuoran yhtälö on (ks. myös alla oleva kuva)

$$\hat{y}_i = 86.09 + 0.514 x_i, \quad i = 1, 2, ..., 1078.$$

Suoran kulmakertoimen $\hat{\beta}_1 = 0.514$ tulkinta on siis, että jos isä A on 1 cm pitempi kuin isä B, isä A:n poika on keskimäärin 0.514 cm pitempi kuin isä B:n poika.



Kuva: Havainnot ja pojan pituuden sovitteet lineaarisesta regressiomallista (punainen suora).

Ohessa vielä tarkemmin R:n (RStudion) lm-funktion tuottama regressiotulostus:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 86.08771    4.65555   18.49   <2e-16 ***
Father       0.51401    0.02706   19.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.193 on 1076 degrees of freedom
Multiple R-squared:  0.2512,    Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Edellä keskusteluihin seikkoihin liittyen nähdään mm. seuraavaa:

- Selitysaste (“Multiple R-squared”) on 0.251 eli poikien pituuden lineaarisesta vaihtelusta kyetään selittämään n. 25 % isien pituuden avulla.
- Testattaessa isän pituuteen liittyvää regressiokerrointa, eli nollahypoteesia $H_0 : \beta_1 = 0$, niin saamme t-suhteeksi (t-testisuureeksi) n. 19.00 ja p-arvoksi 0.000 (2e-16 on R:n tapa antaa hyvin pieni luku, olennaisesti siis nolla tässä tapauksessa). Näin ollen voimme kaikilla tyypillisillä tilastollisilla merkitsevyystasoilla α, kuten $\alpha = 0.05$, hylätä tämän nollahypoteesin eli isän pituudella on tilastollisesti merkitsevää informaatiota pojan pituutta koskien.
- Huomaa kuitenkin, mitä myös mm. selitysaste kuvaa, että pituuksien välinen suhde ei ole täysin deterministinen vaan aika runsaasti jää kuitenkin vielä satunnaista vaihtelua jäljelle.

Ennusteet **Ennusteen muodostaminen**. Mainitaan vielä tässä kohtaa ennusteen muodostamisesta. Edellä mallin sovittaminen tarkoittaa, että käytämme kaikkia havaintoja/havaintopareja (x_i, y_i), $i = 1, ...n$, mallin sovittamiseen (“opettamiseen”). Tämän tuloksena saadaan edellä mainitut sovitteet ja esimerkiksi edeltävän kuvan lineaarisen suoran tässä yhden selittäjän tapauksessa.

Sen sijaan mm. koneoppimisen suosion kasvusta johtuen ennusteilla tarkoitetaan yhä useammin tilastotieteellisessä kirjallisuudessa uusille havainnoille ja/tai tietylle selittävän muuttujan arvolle x perustuvaa y:n ennustetta. Ts. jos muuttuja x saa arvon $x = x^e$, jossa x^e on jokin x:lle asetettava arvo ennusteen pohjaksi, niin ennuste on tällöin $\hat{y}^e = \hat{\beta}_0 + \hat{\beta}_1 x^e$.

Ennusteiden muodostusta ja laajempaa roolia tilastollisissa analyyseissä ja tietysti erityisesti ennustamisessa tarkastellaan vielä tulevassa luvussa erikseen, kun tehdään myös mm. ero opetus- ja testiaineiston välille.

13.3 Muita regressiomalleja

Yksinkertaista lineaarista regressiomallia voidaan laajentaa monin tavoin ja monenlaisiin erilaisiin tilanteisiin. Käydään seuraavassa hyvin lyhyesti ja johdantomaisesti läpi muutamia mahdollisia ja tyypillisiä tilanteita.

Usean selittäjän lin. malli **Usean selittäjän lineaarinen regressiomalli**: yhden selittäjän sijaan käytetään useita selittäviä muuttujia.

Esimerkki. Jos haluamme mallintaa opiskelijan arvosanaa (olisi tässä siis vastemuuttuja Y_i), voimme ottaa huomioon paitsi opiskelutunnit, mutta myös unen määrän, stressitason ja aiemmat suoritukset. Malli näyttäisi tässä tapauksessa tältä (vrt. yhden selittäjän tapaus):

$$\text{arvosana}_i = \beta_0 + \beta_1 \times \text{opiskelutunnit}_i + \beta_2 \times \text{uni}_i + \beta_3 \times \text{stressi}_i + \beta_4 \times \text{aiemmat arvosanat}_i + \varepsilon_i.$$

Epälineaarinen regressiomalli Lineaarisen mallin sijaan malli voi olla myös **epälineaarinen**.

Esimerkki. Logistinen kasvukäyrämalli on esimerkki epälineaarisesta mallista. Se voidaan esittää seuraavassa muodossa:

$$Y_i = \frac{a}{1 + \exp(-k(t_i - t_0))} + \varepsilon_i,$$

jossa

- Y_i on opiskelijan arvosana (vaste),
- t_i on opiskelijan käyttämä opiskeluaika (esimerkiksi tuntia viikossa),
- a on vasteen maksimiarvo (esimerkiksi 10),
- k on kasvun jyrkkyyttä säätelevä parametri,
- t_0 on kohta, jossa kasvu on nopeinta,
- ε_i on satunnaisvirhe, joka kuvaa lineaarisen mallin tapaan mallin ulkopuolista vaihtelua.

Tämä malli on epälineaarinen muotonsa eli parametrien suhteen. Se kuvaa tilannetta, jossa vaste kasvaa nopeasti alussa, mutta lähestyy ylärajaa a asymptoottisesti. Malli soveltuu hyvin tilanteisiin, joissa vasteen kasvu hidastuu resurssien lisääntyessä, kuten oppimisen tai esimerkiksi esim. biologisen kasvun yhteydessä.

Erityisen tärkeitä yhden selittäjän lineaarisen mallin laajennuksia ilmenee kun **vastemuuttuja on muuta muotoa** mitä edellä oletetaan lineaarisissa regressiomalleissa, joissa käytännössä oletetaan että vaste on suhdeasteíkollinen muuttuja, kuten jokin reaaliluku. Vaste voi hyvin olla myös **diskreettiarvoinen**, kuten **binäärinen** $Y_i = 0$ tai $Y_i = 1$ tai **lukumäärä** $Y_i \in 0, 1, 2, 3, ...$

Logistinen regressio Mikäli vaste on binäärinen, niin tällöin tyypillinen tarkasteltava ja täsmennettävä tilastollinen malli on **logistinen regressiomalli** (tunnetaan myös **logistisena regressiona** tai **logit-mallina**).

Esimerkki logistisesta regressiosta. Kuvitellaan, että haluamme ennustaa, onnistuuko työnhakija saamaan työpaikan. Meillä on tietoa hakijan koulutuksesta, työkokemuksesta ja siitä, onko hän osallistunut työhaastatteluun. Vastemuuttuja Y_i on binäärinen: Työnhakija i sai työpaikan (eli $y_i = 1$) tai ei saanut ($y_i = 0$). Logistinen regressiomalli voisi tällöin olla

$$\text{logit}(p_i) = \log\Big(\frac{p_i}{1-p_i}\Big) = \beta_0 + \beta_1 \times \text{koulutus}_i + \beta_2 \times \text{kokemus}_i + \beta_3 \times \text{haastattelu}_i, \quad i = 1, ..., n,$$

jossa $\text{logit}(p_i)$ tarkoittaa siis logaritmista suhdetta työpaikan saamistodennäköisyyden p_i ja sen vastatodennäköisyyden $1 - p_i$ välillä eli sen todennäköisyyden kesken, että hakija ei saa työpaikkaa. Mallin avulla voimme arvioida, kuinka todennäköistä on, että hakija saa työpaikan, kun tiedämme hänen taustatekijänsä eli mallin sovitteena saadaan todennäköisyyksiä $\hat{p}_i$, jotka ovat tunnetusti nollan ja yhden välillä. Tämä on erittäin hyödyllistä esimerkiksi rekrytointiprosessien kehittämisessä tai koulutuksen vaikuttavuuden arvioinnissa.

- Huomaa eroavaisuudet lineaariseen malliin! Ts. mitä tapahtuu malliyhtälön vasemmalla puolella eikä myöskään mitään erillistä virhetermiä ilmene malliyhtälön oikealla puolella.
- Kaikki tarkemmat yksityiskohdat ovat tämän kurssin ulkopuolella.

Poisson-regressio Jos vaste on lukumäärä, niin tällöin yksi mahdollinen malliluokka on ns. **Poisson-regressiomalli**. Tässä yhteydessä oletetaan siis, että sm. Y_i noudattaa Poisson-jakaumaa ja regressiomalli rakennetaan tämän oletuksen ympärille.

Esimerkkejä. Yksi mahdollinen esimerkki voisi olla tutkia tarkempaa Poisson-regressiomallia aiemmin käsiteltyyn jalkapallo-ottelun kotijoukkueen ja vierasjoukkueen maalimääriä koskevalle aineistolle.

Toinen esimerkki saattaisi olla esimerkiksi seuraaava. Kuvitellaan, että haluamme mallintaa, kuinka monta asiakaspalvelupyyntöä yritys saa päivässä. Meillä on tietoa viikonpäivästä (niiden vaikutuksesta asiakaspalvelupyyntöihin), mainoskampanjoista ja verkkosivun kävijämääristä. Vaste on lukumäärä: kuinka monta pyyntöä tulee. Poisson-regressiomalli olisi tällöin muotoa

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{viikonpäivä}_i + \beta_2 \times \text{kampanja}_i + \beta_3 \times \text{kävijät}_i, \quad i = 1, ..., n,$$

jossa λ_i on odotettavissa oleva pyyntöjen lukumäärä. Malli kertoo meille, miten eri tekijät vaikuttavat asiakaspalvelun kuormitukseen – ja auttaa esimerkiksi resurssien suunnittelussa. Mallin sovitteina saadaan lukumääräsovitteita havainnoille y_i.

Kuten nämäkin kaksi diskreettien vastemuuttujien mallityyppiä osoittavat ja korostavat, **vastemuuttujan roolin/luonteen selvittäminen on hyvin**

keskeistä tilastollista mallia rakennettaessa. Tässä siis pätee samat eroavaisuudet mitkä tulevat tutuiksi todennäköisyyslaskennan kursseilla kun käsitellään diskreettien ja jatkuva-arvoisten satunnaismuuttujien jakaumia ja niihin liittyviä yksityiskohtia.

Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään useilla myöhemmillä tilastotieteen kursseilla (Turun yliopistossa) tilastotieteen aineopinnoissa, kuten erityisesti tilastollista päättelyä sekä lineaarisia ja yleistettyjä lineaarisia malleja koskevilla kursseilla.

Chapter 14

Tilastollisesta ennustamisesta

Kuten olemme jo tähän menneessä nähneet, tilastollinen analyysi ja sen erottamattomana osana tilastollinen päättely on keskeinen vaihe tieteellistä tutkimusta. Vielä ennen tilastollisen selittämisen ja ennustamisen välisiä eroja koskevia pohdintoja muistutetaan minkälaisia tilastollisia analyysitilanteita ja menetelmiä olemme tähän mennessä sivunneet:

- Aineistojen perustunnuslukujen laskeminen ja kuvaileminen (ts. kuvaileva tilastotiede).
  - Yksinkertaisin tilastollisen päättelyn muoto on hyödyntää aineistoa kuvailevia tunnuslukuja, kuten keskiarvoja ja hajontalukuja (kuten niitä on esitelty edellä). Niistä voidaan kuitenkin tehdä vain melko rajoittuneita päätelmiä. Kuvaileva tilastotiede toimii näissä yhteyksissä lähinnä alustavien analyysien ja aineistoon tutustumisen kautta.
- Muuttujien välisten (mahdollisten) yhteyksien tutkiminen
  - Varsinkin havainnoivassa tutkimuksessa selvitetään, miten selittävät muuttujat ovat yhteydessä selitettävään vastemuuttujaan. Näissä yhteyksissä voidaan käyttää esim. lineaarista tai logistista regressiota (ja niiden monenmoisia laajennuksia) tai esim. aikasarja-analyysissä aikasarjoja analysoitaessa (aikasarjoista lyhyesti vielä myöhemmin). Näiden pohjalta voidaan arvioida muuttujien yhteyksiä ja riippuvuussuhteita, ja myöskin muodostaa ennusteita, kuten tässä luvussa tullaan tarkastelemaan.
- Hypoteesien testaaminen

  - Hypoteesien testaamisella voidaan osaltaan arvioida kuinka uskottavaa on, että nyt havaittu käyttäytyminen tai ilmiö toistuu jatkossa ja mahdollisesti johtopäätelmät myös yleistyvät muihin tilanteisiin.

- Ryhmitteleminen, kuten dimension pienentäminen.

  - Tätä ei ole käsitelty eikä käsitellä juurikaan vielä tällä kurssilla. Ajoittain tilastotieteellä pyritään löytämään aineistosta erilaisia ryhmiä ja luokkarakenteita. Tässä yhteydessä laajoja aineistoja saatetaan pystyä "puristamaan" pienemmiksi ja siten mahdollisesti helpommin analysoitaviksi ja tulkitattaviksi.

Nämä tavoitteet korostavat valtaosin aineiston mallintamis- ja selittämistavoitteita. 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 tehdään tähän eroavaisuuteen liittyviä keskeisiä huomioita.

Ts. aineiston mallintaminen on tärkeää ja monesti erityisenä tavoitteena, mutta usein haluamme myös muodostaa ennusteita tulevaa kehitystä koskien. Tämä tarkoittaa siis uusien havaintojen arvojen ennustamista.

14.1 Tilastollinen selittäminen vs. ennustaminen

(Tilastollinen) **selittäminen** tarkoittaa esim. kahden muuttujan välisen yhteyden tutkimista

- Vrt. siis tämän kurssin yksinkertainen tilanne lineaarisen regressiomallin yhteydessä, jota voidaan suoraviivaisesti ja usein pitääkin laajentaa useiden selittävien (ennustavien) muuttujien samanaikaiseen käyttämiseen.

Selitysmalli

Esimerkki. Tutkijaa saattaa kiinnostaa esimerkiksi tupakoinnin vaikutus sepelvaltimotautikuolleisuuteen tai ylipainon vaikutus leikkauksen jälkeisiin infektioihin. Tällöin pyrkimyksenä on rakentaa "**selitysmalli**", jossa on perustellut syy-seuraussuhteet selittävästä (selittävistä) muuttujista selitettävään muuttujaan.

(Tilastollinen) **ennustaminen** vastaavasti tarkoittaa, että tietyillä selittävän tai selittävien (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 rakennettaessa varsinaisilla selityssuhteilla ei välttämättä ole merkitystä. Tärkeintä on mallin ennustekyky, ei niinkään esim. yksittäisen regressiokertoimen arvo ja siihen liittyvät tarkemmat tulkinnat ja mahdollinen tilastollinen testaaminen. Tilastollisesti merkitsevä regressiokerroin ei tarkoita, että muuttujalla olisi välttämättä todellista ennustekykyä.
- Ennustekyky on siis tutkittava erikseen. Esimerkiksi lineaarisen mallin perinteiset tunnusluvut, kuten selitysaste, eivät vielä kerro mallin todellisesta ennustekyvystä paljoakaan, koska ennusteet koskevat havaintoja, joita ei käytetä mallin estimoinnissa (opettamisessa). Tästä huolimatta melko usein ennustemallin rakentaminen perustetaan pitkälle samoihin tilastollisen päättelyn ja estimointiteorian lähtökohtiin mitä olemme jo sivunneet tällä kurssilla.
- Hyvin usein tutkimuksissa raportoidaan, että tietty muuttuja "ennustaa" (*predict*) toista. Usein kuitenkin taustalla on tällöinkin usean muuttujan selitysmalli, jonka regressiokertoimien tilastollista merkitsevyyttä on tulkittu siihen tapaan mitä edellä kuvattiin. 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.

14.2 Tilastolliseen ennustamiseen liittyviä huomioita

Esimerkki. Kovin usein toistuvaa pohdiskelua ja jälkiviisautta jälkikäteen.

"Olisihan se pitänyt tietää/arvata!"

Vai olisiko sittenkään!? Tähän pohdiskeluun palataan alapuolella.

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

Esimerkkejä. Kun valitsemme reitin työmatkalle, päätämmekö menemmekö toisille treffeille tai säästämme huonompia aikoja varten, teemme ennusteen tulevaisuuden kehityksestä ja siitä, miten mahdollisesti suunnitelmamme vaikuttavat suotuisan tuloksen todennäköisyyteen.

- Lisää pohdintaa ja esimerkkejä, kuten ylläolevat esimerkit, löytyvät Silverin kirjasta (2014) (ks. Oheislukemistoa).

Arkiset ongelmat eivät aina vaadi ankaraa ajattelua ja pohdiskelua erilaisten vaihtoehtojen välillä niihin käytettävissä olevan ajan ollessa rajallinen. Tästä huolimatta teet ennusteita tiedostaen ja useimmiten tiedostamatta monta kertaa päivässä!

Ennustevirhe

Ennustevirhe. Ennustetta $\widehat{y}^e$ verrataan toteutuneeseen arvoon tai kehitykseen y. Näiden erotuksena muodostuu ennustevirhe $y - \widehat{y}^e$.

Riippuen ennustettavasta kohteesta (kuten riippuen tilastollisen mallin vastemuuttujan luonteesta eli onko se esimerkiksi jatkuva vai diskreetti sm.), ennustevirhe näyttäytyy eri muodoissa.

Lähtökohtana on (luonnollisesti) minimoida ennustevirheet. Käytännössä useinmiten mm. vastemuuttujan luonteen perusteella valitaan sopiva ennustevirheitä summarisoiva tunnusluku, kuten keskineliöennustevirhe (jatkuvat vastemuuttujat) tai luokitteluvasteiden tapauksessa väärin ennustettujen luokitteluiden suhteellinen osuus.

Ajoittain ennustetarkkuutta on helpompi ja toisaalta sitten vaikeampi tarkkailla.

- Esim. taloustieteessä on paljon helpompi arvioida työttömyyttä koskevaa ennustetta kuin esimerkiksi ennustetta (jopa väitettä) velkaelvytyksen tehokkuudesta.
- Toisaalta valtio-opissa voidaan arvioida vaalitulosta koskevia ennusteita suoraviivaisesti vaalien jälkeen, mutta saattaa kulua vuosikymmeniä nähdä miten poliittisten instituutioiden ennusteisiin perustuvat ennakoidut muutokset vaikuttavat poliittisten päätösten tuloksiin.

Esimerkki: Finanssikriisi 2008. Silverin kirjan (2014) luvun 1 pohdintaa ennustevirheestä vuosien 2007–2009 finanssikriisiin liittyen.

- Tuona aikana tapahtui 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 laajempaakin asuntokuplaa eli ts. perusteettoman korkeita asuntojen hintoja), työttömyys kasvoi räjähdysmäisesti jne.

Pohditaan finanssikriisin ennustettavuutta **hieman ennen kriisin alkamista** eli tilanteessa kun **ennusteita muodostettiin vielä näkemättä laajalti yllätyksenä tullutta kriisiä**. Lopulta tapahtuneiden ennustevirheiden **yhteisiä ja tyypillisiä piirteitä** myös muita sovelluksia ajatellen:

1. Asunnonomistajat ja sijoittajat ajattelivat, että nousevat hinnat viittasivat siihen, että asuntojen hinnat jatkaisivat nousuaan (todellisuudessa historia viittasi siihen, että pankki- ja finanssikriisien yhteydessä niillä on taipumus laskea).
2. Luottoluokistuslaitokset (samoin kuin Lehman Brothersin kaltaiset pankit) eivät ymmärtäneet, miten riskialttiita asuntovakuudelliset arvopaperit olivat. Ongelma ei varsinaisesti ollut siinä, että luokituslaitokset eivät nähneet asuntokuplaa. Sen sijaan niiden ennustemallit olivat täynnä huonoja oletuksia ja väärää "itseluottamusta" mahdollisten asuntojen hintojen romahduksen riskeistä.
3. Laajasti ei ennakoitu, miten asuntokriisi laukaisee globaalin rahoituskriisin. Se johtui suurelta osin liiallisesta velkaantumisesta markkinoilla, jossa lyötiin erinäisten instrumenttien myötä kasvavissa määrin vetoa yhdysvaltalaisten halukkuuden puolesta sijoittaa uuteen kotiin.
4. Rahoituskriisin välittömissä jälkimainingeissa ei osattu ennustaa, miten laajoja taloudellisia ongelmia se aiheuttaa. Finanssikriisit tyypillisesti tuottavat erittäin syviä ja pitkäkestoisia taloudellisia taantuma- ja lamajaksoja.

Edellä olevan esimerkin tilanteessa, ja laajemminkin vastaavissa tilanteissa ja sovelluksissa, joihin liittyy epäonnistuneita ennusteita, epäonnistumisissa on (ainakin yksi) tyypillinen **yhteinen piirre**. Ts. kussakin tapauksessa ennustajat (ammattilaiset tai tavanomaiset kansalaiset) **jättävät/jättivät keskeisen asiayhteyteen liittyvän tekijän huomiotta**.

- Olennainen kysymys kuuluu, olisiko tämä tekijä ollut ennakoitavissa ennen ennusteiden kohteena olevan tapahtuman tapahtumista? Ja jos se olisi ollut tiedossa, olisiko se voinut olla osa tilastollista (ennuste)mallia?

Esimerkki (jatkoa): Jatketaan finanssikriisiä koskevan esimerkin käsittelyä. **Mitä tekijöitä ei osattu nähdä** ennen finanssikriisin puhkeamista (edellisen esimerkkikohdan numeroituihin kohtiin viitattuna):

1. Asunnonomistajien luottamus asuntojen hintoihin johtui ehkä siitä, että lähimenneisyydessä Yhdysvalloissa asuntojen hinnat eivät olleet laskeneet merkittävästi. *Kuitenkaan* koskaan aikaisemmin Yhdysvaltojen asuntojen hinnat eivät olleet nousseet niin laajalla alueella kuin romahdusta edeltävällä kaudella.
2. Pankkien luottamus luottoluokituslaitosten (kuten Moody's ja S&P) 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.
3. (Taloustietelijöiden) luottamus rahoitusjärjestelmän kykyyn kestää asuntokriisi syntyi ehkä siitä, että aikaisemmin asuntojen hintojen heilahtelulla ei yleensä ollut suuria vaikutuksia rahoitusjärjestelmässä. *Kuitenkaan* rahoitusjärjestelmä ei luultavasti koskaan aikaisemmin ole ollut yhtä vekkaantunut eikä vedonlyöntiä asuntojen hinnoista ollut tehty vastaavassa mittaluokassa.
4. Poliittisten päättäjien luottamus siihen, että talous toipuu nopeasti rahoituskriiseistä syntyi ehkä viime vuosikymmenten taantumista saaduista kokemuksista. Useampia niitä oli seurannut nopea "V-muotoinen" toipuminen, kuten nyt myös myöhemmin mm. koronapandemian aikaan. *Kuitenkaan* nämä taantumat eivät olleet liittyneet rahoituskriiseihin ja rahoituskriisit ovat (yleensä) erilaisia.

Jokaista edellistä kohtaa yhdistää ennustamiseen hyvin keskeisesti liittyvä seikka: Pyrittiin ennustamaan tilannetta/ilmiötä, joka oli kuitenkin ns. **otoksen ulkopuolella** (engl. **out-of-sample**) eikä siis vastaavasta tilanteesta ollut aiempaa kokemusta (=dataa). Kun ennustaminen epäonnistuu merkittävällä tavalla, tämä sama seikka jättää yleensä runsaasti sormenjälkiä rikospaikalle.

- Miten tämä huomio näyttäytyy siis oheisen esimerkin tapauksessa?

Esimerkki (jatkoa):

- 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 asuntomarkkina-aineistoa.
- Ongelmana oli, että 1980-luvulta 2000-luvun alkuvuosiin saakka asuntojen hinnat olivat aina vakaat tai nousevat Yhdysvalloissa. Tässä tilanteessa oletus, että asunnonomistajien asuntolainat eivät 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 ulkopuolinen tapahtuma** ja tässä tilanteessa luottoluokituslaitosten mallit olivat lähtökohtaisesti huonoja lainojen hoitamatta jättämisen riskiä arvioitaessa.

Finanssikriisiä 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**", **opetusaineistolla**, joka ei lopulta ollut relevantti juuri myöhemmin tapahtunutta kriisivaihetta ajatellen!

- Onkin tärkeää ymmärtää, että ''todellisessa'' ennustetilanteessa joudumme käyttämään aiempaa aineistoa mallien ja algoritmien rakentamiseen.
- Näin ollen tilastollisten mallien/menetelmien ennustekykyä arvioitaessa onkin mentävä otoksen ulkopuolelle, koska ''otoksen sisällä'' voimme opettaa kyseisiä malleja (ääritilanteessa) niin, että ne ovat periaatteessa äärettömän tarkkoja. Ne eivät kuitenkaan takaa missään mielessä hyvää ennustekykyä tulevia tapahtumia ennustettaessa.

Opetus- ja testiaineisto

Opetus- ja testiaineisto. Mallien ja algoritmien opettaminen ns. **opetusaineistolla** ja ennustekyvyn arviointi **ennusteotoksen** avulla pitää erottaa toisistaan.

- Seuraavassa oletetaan, että opetusaineisto=otos, jota esim. regressiomallin tapauksessa käytettiin mallin sovittamiseen.
- Opetusaineistolla (x_i, y_i), $i = 1, ..., n$ siis opetetaan (estimoidaan/optimoidaan) käytettävän ennustemallia/algoritmia ja sen parametreja.
- Lopulta muodostetaan ennusteita opetusaineiston ulkopuoliselle testiaineistolle.
  - Merkitään testiaineistoa (x_j^e, y_j^e), $j = 1, ..., n^e$. Muodostuvia ennusteita merkitään $\widehat{y}_j^e$, $j = 1, ..., n^e$. Tässä ei oteta tarkempaa kantaa minkälaista mallia/algoritmia käytetään näiden muodostamiseen.
  - Testiaineistoa ei siis käytetä mallin/algoritmin opettamiseen, vaan ainoastaan ennusteiden hyvyyden arviointiin.
- Voidaan muodostaa ennustekykyä kuvaavia tunnuslukuja, kuten

  - Jatkuvien sm:jien tapauksessa keskineliöennustevirhe (MSFE): $\frac{1}{n^e}\sum_{j=1}^{n^e}(y_j^e - \hat{y}_j^e)^2$. Tässä tapauksessa MSFE:ssä on myös ennustevirheiden neliöiden summa jaettu testiaineiston havaintojen lukumäärällä (näin ei aina toimita).
  - Diskreettien, kuten binääristen ja siten luokittelutilanteeseen muodostettavien ennusteiden kohdalla voidaan käyttää esimerkiksi oikein tai väärin ennustettujen (luokiteltujen) havaintojen suhteellista osuutta testiaineistossa.

- Tähän tapaan voidaan vertailla kahden, tai usemman, ennustemallin paremmuutta todellisessa ennustetilanteessa. Esimerkiksi keskineliöennustevirheen tapauksessa suositaan mallia, joka tuottaa testiaineiston perusteella pienimmän keskineliöennustevirheen.

Esimerkki, jatkoa. Palataan vielä **isien ja poikien pituuksia** koskevaan esimerkkiin. Tehdään pieni ennustekokeilu, jossa tavoitteena on havainnollistaa tämän luvun keskeisiä huomioita.

Jaetaan tarkasteltava aineisto opetusaineistoon ja testiaineistoon

- Otetaan ensimmäiset 700 havaintoa opetusaineistoon ja estimoidaan lin. malli uudestaan näillä opetusaineiston havainnoille. Ts. nyt vain näitä havaintoja käytetään mallin opettamiseen.
- Loput havainnot muodostavat testiaineiston. Opetusaineistolla opetetun mallin (ml. estimoidut parametrit) perusteella tehdään ennusteet $\hat{y}_j^e, j = 1, ..., n^e$, jossa $n^e = 378$.

Osoittautuu, että lineaarisen mallin (jossa siis isän pituudella nyt ennustetaan testiaineiston havaintoja poikien pituudesta) tapauksessa keskineliöennustevirhe on MSFE = 39.59.

Otetaan vertailumalliksi, selvästikin liian yksinkertainen malli, jossa poikien pituusennuste olisi suoraaan opetusaineiston poikien keskipituus. Ts. lineaarinen malli supistuu niin, että isien pituutta koskeva ennustava muuttuja jätetään mallista pois. Tällöin MSFE=51.49.

- Näemme siis kuinka lineaarisella mallilla (ml. isien pituus ennustavana muuttujana) saadaan selvästi pienempi testiaineistolle saatava keskineliöennustevirhe. On siis hyödyllistä käyttää isän pituutta ennustavana tekijänä myös ennustamisen näkökulmasta. Tämä vahvistaa osaltaan, että isien pituuden sisältämä ennusteinformaatio poikien pituudesta sisältää todellista *signaalia*, eikä tämä suhde ole siis *kohinaa*.

Pienenä lisähuomiona vielä edelliseen eli ajoittain opetus- ja testiaineiston lisäksi määritellään vielä erillinen validointiaineisto mallin/algoritmin validointia varten ennen varsinaista testiaineiston käsittelyä.

Otoksen sisäinen sovittaminen

Otoksen sisäiseen sovittamiseen (engl. **in-sample** tai **training sample** estimation tai "prediction") liittyy ennustamisen näkökulmasta katsoen ns. **ylisovittamisen** vaara. On siis mahdollista, että yritämme puristaa lähes puhtaasta kohinasta signaaleja, jotka eivät missään mielessä tule olemaan valideja otoksen ulkopuolisessa ennustamistilanteessa.

Ylisovittaminen (ylisovitettu tilastollinen malli) merkitsee tilastollisen mallin rakentamista siten, että se oppii opetusaineiston liian(kin) hyvin (malli on liian mukautettu opetusaineistoon), jolloin sen ennustamiskyky (yleistämiskyky) uusia havaintoja testiotoksessa ennustettaessa alkaa heikentyä. Tämä tarkoittaa, että malli oppii liikaa opetusaineiston sellaisia yksityiskohtia, jotka ovat kohinaa, eikä malli yleisty hyvin uusille, mallin sovittamisen aikaan tuntemattomille testiaineiston havainnoille.

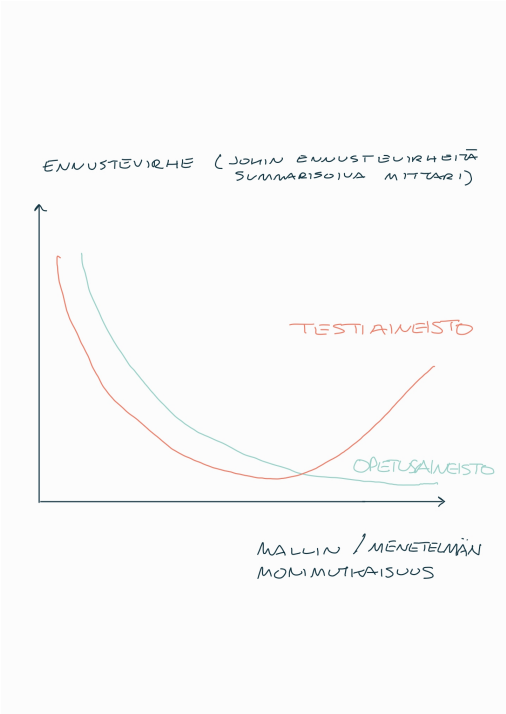
Ylisovittaminen voi siis johtaa, ja usein johtaakin, huonoon suorituskykyyn todellisissa ennustustehtävissä ja siten verrattaen suuriin ennustevirheisiin.

Vastaavasti **alisovitettu malli** ei kuvaa aineistoa tyydyttävällä tavalla. On odotettavissa, että jos malli ei sovitu edes opetusaineiston tapauksessa hyvin, niin sen todellinen ennustekyky on tyypillisesti myös heikko.

Harha-varianssi-kompromissi

Harhan ja varianssin välinen kompromissi (*bias-variance trade-off*). Kun sovitetaan tilastollista mallia/algoritmia ennustamista varten, mallin/algoritmin lisääntyvä monimutkaisuus johtaa lopulta malliin, jossa on vähemmän harhaa siinä mielessä, että sillä on suurempi potentiaali mukautua taustalla olevan prosessin yksityiskohtiin. Toisaalta samanaikaisesti syntyy enemmän varianssia, mikä perustuu siihen että pienetkin muutokset opetusaineistossa muuttavat mallin/algoritmin parametreja ja siten lopulta saatavia sovitteita (opetusaineisto) ja/tai ennusteita (testiaineisto). Näiden kahden tavoitteen välillä tulee siis tasapainotella, jotta vältetään ylisovittaminen.

Harhan ja varianssin kompromissi on **fundamentaalinen tulos** tilastotieteen, mutta erityisesti myös koneoppimiseen ja tekoälyyn pohjautuvien uusien menetelmien käyttökelpoisuudelle erilaisissa sovelluksissa.



Kuva: Harhan ja varianssin kompromissi ennustevirheen ja mallin monimutkaisuuden kesken.

Jotta asiat eivät olisi liian yksinkertaisia, aivan viime vuosien aikana uusimmassa tutkimuskirjallisuudessa on jo haastettu oheisen kuvan ja harhan ja varianssin välisen kompromissin paikkaansapitävyyttä. Tämä perustuu siihen mitä tapahtuu, jos siirtyisimme vielä enemmän kuvassa oikealle. Ts. ns. **yliparametrisoituihin malleihin ja menetelmiin**. On voitu osoittaa, että tietyissä tilanteissa testiaineiston ennustevirhe lähtee uudelleen laskuun!

- Tämän yksityiskohdan tutkimus kuitenkin jatkuu edelleen ja sen tarkempi käsittely on jälleen selvästi tämän kurssin ulkopuolella.

Esimerkki, jatkoa. Jatketaan vielä **pituus-esimerkkiä** ja siihen edellä kohdistettua ennustekokeilua. Otetaan tarkasteluun mukaan myös yksi usein (tilastollisen) koneoppimisen lukuun luettava menetelmä/algoritmi eli ns. tukivektoriregressio (SVR, support vector regression).

- Emme käsittele tätä menetelmää/algoritmia tällä kurssilla sen tarkemmin. Keskeistä on vain todeta, että kyseessä on periaatteessa lineaariseen regressioon verrattuna selvästi monimutkaisempi menetelmä.

Vertaillaan ensin lineaarisen mallin ja SVR:n kykyä mallintaa 700 havainnon opetusaineiston havaintoja. Tarkastellaan menetelmien perusteella saatavien residuaalien (havainnot-sovitteet) neliöiden summaa jaettuna 700:lla, eli MSFE:tä vastaavaa tunnuslukua opetusaineistolle. Tällöin lineaarisen mallin sovitteiden perusteella luku on 37.76 ja SVR:n tapauksessa 37.64.

- Ts. SVR tuottaa tarkempia sovitteita opetusaineistolle, mikä ei ole suinkaan yllättävää. Päinvastoin, tämä on täysin odotettua, kun puhutaan mallin/algoritmin kyvystä "sovittua" opetusaineistoon. Tämä näkyy ylläolevassa kuvassa laskevana opetusaineiston ennustevirhettä kuvaavana käyränä menetelmän monimutkaistuessa.

Siirrytään nyt tarkastelemaan opetusaineiston perusteella muodostettujen lineaarisen mallin ja SVR:n tuottamia ennusteita testiaineistolle. Saadaan lineaariselle mallille MSFE=39.59 ja SVR:n tapauksessa MSFE= 40.21!

- Ts. ennustekykyä vertailtaessa monimutkaisempi, ja opetusaineistolle paremmin sopiva, SVR tuottaa heikompia ennusteita mitä lineaarinen malli. Tässä siis toteutuu edellisen kuvion tilanne, jossa välttämättä monimutkaisempi malli/algoritmi, kuten tässä tapauksessa SVR, ei tuota parempia ennusteita!

On syytä korostaa, että **tämä on vain yksi esimerkki** eikä suinkaan tarkoita, etteikö koneoppimiseen luettavilla menetelmillä olisi paljon annettavaa ennustamiselle monien sovellusten yhteydessä.

Jälleen kerran näistä tämän luvun teemoista keskustellaan tarkemmin myöhemmin tilastotieteen aine- ja syventävien opintojen erikoiskursseilla.

Chapter 15

Tilastotieteen rooli uuden tiedon tuottamisessa

Tilastotieteen yhteiskunnallisesta roolista alustettiin keskustelua jo aiemmissa luvuissa. Tilastotieteen keskeinen yhteiskunnallinen rooli liittyy elimellisesti juuri uuden tieteellisen tiedon tuottamiseen

- Tilastotiede liittyy olennaisesti melkein kaikkeen tieteentekemiseen, joten ei liene yllätys että tilastotiede on jossain määrin tuttua kaikille tieteentekijöille.
- Tilastotiede tarjoaa pohjan uuden tiedon tuottamiselle, mutta toisaalta voitaisiin myös ajatella teoreettisen tilastotieteen ja siellä luotujen menetelmien ylipäätään mahdollistaneen uskottavan tieteenteon. Tässä luvussa emme kuitenkaan takerru tähän ''muna vai kana''-ongelmaan, vaan tarkastelemme yleisemmällä tasolla tilastotieteen roolia tieteenteossa.

Ensiksi tarkastelemme tilastollisia menetelmiä hyödyntävistä ongelmanasetteluista löydettäviä yhteisiä elementtejä. Nämä elementit ovat niin yleisiä että niitä voidaan tarkastella ja kuvata ilman yhteyttä mihinkään yksittäiseen ongelmaan.

Tämän jälkeen tarkastelemme, yleisellä tasolla, tilastollisia menetelmiä hyödyntävän tieteellisen tutkimusprosessin eri vaiheita yleisesti.

- On kuitenkin mahdotonta koostaa yleisiä ''tee se näin''-listoja tilastollisen tutkimuksen toteuttamiseksi, joten tarkastelemme kurssilla käsiteltyjä aihekokonaisuuksia ja niitä yleisiä periaatteita, joita jokaisen tieteentekijän tulee hallita.

15.1 Tilastollisen tutkimuksen tyypillisiä elementtejä

1. **Satunnaisvaihtelu**. Satunnaisilmiöiden generoima havaintoaineisto on aina tilastollisen tutkimuksen tutkimuskohde. Täten kaikki tieteellinen tutkimus, joka koskee satunnaisvaihtelua ilmentävää aineistoa on (tai tulisi olla) tilastotieteellistä.

- Tilastollisen tutkimuksen tavoitteena on (useimmiten) pyrkiä erottamaan satunnaisilmiön systemaattinen ja satunnainen vaihtelu eli signaali ja kohina. Tämä vaatii **substanssiosaamisen** lisäksi **menetelmäosaamista** sekä hyvää **tilastotieteellistä intuitiota**.
- 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 pahimmillaan asettaa kyseisen aiheen tutkimuksen vuosiksi väärille raiteille!

Substanssitietous

Substanssitietous tarkoittaa syvällistä ja asiantuntevaa tietoa jostakin tietystä aihealueesta tai alasta. Se viittaa siihen, että henkilöllä on:

- Laaja ja syvällinen ymmärrys tietystä aiheesta (esim. lääketiede, oikeustiede, kasvatustiede, tekniikka).
- Kyky soveltaa tietoa käytännön tilanteissa, ongelmanratkaisussa tai päätöksenteossa.
- Ajankohtainen ja luotettava tieto, joka perustuu tutkimukseen, kokemukseen tai koulutukseen.

Substanssitietous eroaa esimerkiksi yleissivistyksestä siinä, että se on spesifiä ja syvällistä, kun taas yleissivistys on laajempaa ja pinnallisempaa tietoa monista eri aiheista.

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 siis kuvaus järjestelmään liittyvistä olioista, ilmiöistä ja toisaalta myös rajoituksista.

- Lisäksi tutkimusongelman holistinen käsittely on tilastollisen tutkimuksen kannalta tärkeää: ilmiöön liittyvien tärkeiden ominaisuuksien unohtuminen tarkastelusta saattaa johtaa esimerkiksi *puuttuvan muuttujan harhaan*!

Tilastolliset menetelmät auttavat tutkijaa vastaamaan kysymyksiin siitä, mitkä tilastolliset muuttujat ovat tutkimuskysymyksen kannalta oleellisia.

- Varsinkin nykypäivänä kun datan määrä kasvaa alati kiihtyvällä tahdilla, olemme ihmiskuntana informaatiotulvan edessä paikoin aseettomia: mitkä ympäröivistä ilmiöistä liittyvät toisiinsa ja miten?

- Erityisesti teoreettisen tilastotieteen kentällä on viimeisten vuosikymmenien aikana kehitetty lukuisia edistyksellisiä menetelmiä nk. **dimension pienennyksen** alalla. Näiden menetelmien tavoitteena on löytää rakenteita tai olennaisia piirteitä aineistosta, jossa jokaiselta tutkimusyksiköltä mitataan suuri määrä muuttujia – esimerkiksi genomitutkimuksessa, jossa yksilöiltä voidaan mitata satojatuhansia geneettisiä markkereita. Tilastotieteessä nämä menetelmät menevät valtaosin **monimuuttujamenetelmien** alle ja niitä käsitellään tarkemmin monimuuttujamenetelmiä koskevilla erikoiskursseilla

- **Hahmottamisen vaiheet**

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

3. **Tilastollisen mallin muodostaminen ja siihen perustuva päättely**.

Muistetaan tunnetun tilastotieteen professorin George Boxin kuvaava sitaatti: "Kaikki mallit ovat vääriä, mutta jotkut ovat käyttökelpoisia". Tilastollinen malli on tosiaan "vain" kuvaus aineiston sisältämästä vaihtelusta: 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 eritteleminen voi olla vaikeaa, 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 substanssitietous ei aina riitä haluttuun käyttöön. Yhdistämällä tilastotieteen keinoja sekä substanssitietoutta saadaan ongelma ratkaistua vakuuttavalla ja perustellulla tavalla.

- Tilastollisen (soveltavan) tutkimuksen tavoitteena on tuottaa substanssitietoon perustuen ja tilastotieteen menetelmiä hyödyntäen uutta tietoa: lopputulos on menetelmä- ja substanssiosaamisen **synteesi**, joka tuottaa uutta substanssitietoutta (sekä joskus myös uusia ongelmia teoreettisen tilastotieteen menetelmäkehitykselle).
- **Jokaisen tutkijan tulisi siis 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?

15.2 Tutkimusprosessi

Soveltavassa tilastotieteessä tutkimusongelman asettelulla on erityisen tärkeä rooli. Tämä alaluku nojaa soveltavilta osin erityisesti Sundin (2003) kurssimateriaaliin.

- Yksi soveltavan tilastotieteen osa-alue on juuri kokeiden suunnittelu ja analysointi. Kokeen suunnittelussa, toteutuksessa ja tulosten analysoinnissa sovelletaan mm. seuraavia tilastollisia menetelmiä: koesuunnittelu, otanta, estimointi ja testaus.

Tutkimusta ei yleensä ole mahdollista jakaa täysin selvästi erillisiin ja ajallisesti tosiaan seuraaviin vaiheisiin.

- Tutkimusprosessin vaiheet toistuvat vuorotellen ja limittäin, sillä tutkimuksen aikana tehdyt havainnot muokkaavat tutkimuksen kulkua.
- Tutkimuksen tekeminen vaikuttaa lopulta saataviin johtopäätelmiin. Aineiston ja itse ilmiön tuntemus kasvaa tutkimuksen kuluessa.
- Päätelmien tieteellisyyden (periaatteellinen) tarkistusmahdollisuus, ja nykyään yhä useammin jo toistettavuus, on tärkeää.

Usein saattaa olla järkevää jäsentää tutkimuksessa kohdattavia tehtäviä ja vaiheita sekä niiden välisiä suhteita osana tutkimusprosessia. Palataan siis tässä kohtaa aiemmin tämän materiaalin alkupuolella käsiteltyyn **PPDAC/OSAAT-syklin** eri vaiheisiin. Siis nyt kun olemme ehtineet tähän mennessä käsitellä monia tilastotieteellisen tutkimuksen perusasioita ja vaiheita.

Vaihe 1: Ongelma

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 asemoidaan ja tulkitaan suhteessa vallitseviin olosuhteisiin ja tieteelliseen keskusteluun.

Vaihe 2: Suunnitelma

Mitä mitataan ja miten? Tämä mm. seuraavien vaiheiden pohdintaa.

Määritelmät:

- Ilmiöitä ei voida tutkia sellaisenaan, vaan vain niiden ilmentymien kautta ja käsitteiden avulla
- Tutkimus edellyttää arkikieltä täsmällisempää kommunikaatiota (kuten olemme käyneet lävitse), 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 tutkimuskohteeseen liittyvää ja toisaalta ymmärtämiseen tarvittava havaintomaailman hahmotus tuottaa ideologisesti ja historiallisesti sitoutuneita yksinkertaistavia sekä luonteeltaan usein hyvin teoreettisia abstraktioita.

"Operationalisointi": Teoriasta empiriaan siirtyminen. Havainnoiminen ja mittaaminen joudutaan suhteuttamaan valittuun käsitejärjestelmään.

- Tilastollista mittaamista käsittelimme jo aiemmin, kuten erilaisia mittaasteikoita. Numeerisen mittauksen onnistumiseksi käsitteen muotoilu on kiinnitettävä mittariksi. Numeeristen mittaustenkin tulkinta edellyttää, että niitä on tulkittava siinä kontekstissa, josta ne ovat peräisin. On esim. mahdollista, että esitetty kysymys ei välttämättä vastaa tutkimuskohteen ominaisuuksia.
- Usein joudutaankin tekemään kompromisseja mittauksen eksaktisuus- ja systemaattisuusvaatimusten ja arkielen monimerkityksellisyyden välillä. On siis *operationalisoitava* tutkimusasetelma sellaiseksi, että tutkittavasta ilmiöstä pystytään tuottamaan ongelmaratkaisun kannalta tarkoituksenmukaista tietoa.
- Näkökulman kiinnittäminen: Operationalisoinnin avulla siirrytään teorian tasolta empirian tasolle ja samalla tulee määritellyksi näkökulma, josta ongelmaa tarkastellaan. Käsitteet ja niiden yhteyksistä esitettävät näkemykset voivat ajoittain vaihtua tutkimuksen kuluessa.

Vaihe 3: **Aineisto eli data**

- Aineisto edustaa tutkimuksessa empiiristä maailmaa ja se muodostetaan ongelmanasettelun perusteella.
- Tarvitaan systemaattinen aineisto, jonka avulla on mahdollista vastata tutkimuskysymyksiin.
- Aineiston tuottamiseen liittyy useita valintoja, jotka implisiittisesti määräävät myös mahdolliset analyysimenetelmät. Näitä olemme käsitelleet useissa tämän materiaalin luvuissa.

- Aineiston esikäsittely: Aineisto ei ole keräämiseen jälkeen yleensä koskaan suoraan käytettävissä vaan vaatii erinäistä käsittelyä. Esikäsittelyssä aikaisemmin tehtyjen valintojen myötä aineistossa esiintyvät ilmentyvät sovitetaan vastaamaan ongelmankäsittelyä.

Vaihe 4: **Analyysi**

- Analyysivaiheessa sopivasti käsitelty aineisto ja ongelma 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.
- Analyysien tulokset on useimmiten lopulta 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 lopuksi yksi ja sama asia.

Vaihe 5: **Tulokset, tulkinnat ja johtopäätelmät**

- Parhaimmillaan tulokset tiivistävä tutkimusraportti on vakuuttava, ja periaatteessa (ja toivottavasti) toiston mahdollistava, kuvaus tutkimusprosessin kaikista vaiheista, jolloin tutkija voi itse arvioida kuinka uskottavina tuloksia voidaan pitää.
- 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 mukaan myös mahdolliset omat arvopainoitteiset valinnat ja oletukset.

Esimerkki: Tilastolliset kyselytutkimukset. Mielipiteitä selvitetään kyselytutkimuksilla, joiden kohteeksi poimitaan tyypillisesti esim. noin 1000-2000 suomalaista. Kyselytutkimuksen tavoitteena on tehdä kyselyn tulosten

perusteella johtopäätöksiä mielipiteiden jakautumisesta kaikkien suomalaisten joukossa. 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 kannatusosuudet?

Miten esim. 1000-2000 suomalaiseen kohdistetun kyselyn tulokset voidaan yleistää koskemaan kaikkia suomalaisia?

- Kyselyn tulokset voidaan yleistää, jos kyselyn kohteiksi poimittujen suomalaisten joukko muodostaa **edustavan pienoiskuvan** Suomen kansasta. Ts. onnistuneen otannan idea ja erityisesti **edustavan otoksen vaatimukset** tulee täyttyä. Otannan avulla on siis mahdollista päästä tähän tavoitteeseen.
- 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.
- Yleistyksen luotettavuutta ei pystytä arvioimaan, ellei kokeen kohteiden poiminnassa ole käytetty satunnaisotantaa.
- Kyselyn kohteiden poimintaa kaikkien suomalaisten muodostamasta perusjoukosta arpomalla voidaan nähdä satunnaisotantana ja tutkimuksen kohteeksi 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 kerran (ainakin jonkin verran) erilaiset tulokset, koska eri arvonnoissa kyselyyn poimittaisiin (suurella todennäköisyydellä) eri henkilöt.

Chapter 16

Aineisto- ja tutkimustyypit ja koeasetelmat

Tässä luvussa käsitellään erilaisia tapoja toteuttaa tilastollista tutkimusta. Empiirisen tutkimuksen lähtökohtana on aina tutkimusongelma, joka sisältää kysymyksen tai kysymyksiä, joihin tutkimuksella haetaan vastauksia. Tilastotieteen näkökulmasta tutkimusongelman keskiössä on **aineisto**, **data**, ja se miten käytettävissä olevasta aineistosta saadaan vastauksia tutkimuskysymyksiin.

Tarkastelemme tässä luvussa myös tutkimuksenteon käytäntöä käsittelemällä erilaisia aineistotyyppejä.

- Käymme läpi eri alojen ja tutkimusongelmien käytännön tutkimustyössä kohdattavia aineistoja ja erittelemme pidemmälle eri tutkimuskysymysten käytännön haasteita aineistojen osalta sekä sitä, minkälaisia ongelmia erilaisiin tutkimusykysymyksiin käytännössä liittyy ja miten eri tutkimusasetelmat pyrkivät niitä ratkomaan.

Aineistotarpeen ja sen analysoinnin lähtökohdat määrää tarkempi tutkimusongelma.

- Tutkimus voi olla esimerkiksi kuvailevaa, vertailevaa, selittävää tai kokeellista ja aineistolle sekä menetelmille asetetaan kussakin tapauksessa erilaiset vaatimukset ja odotukset.

Erilaisiin tutkimuskysymyksiin ja niihin vastausta etsiviin koeasetelmiin liittyvien esimerkkien avulla pyrimme löytämään vastauksia esimerkiksi seuraaviin kysymyksiin:

- Miten tilastotiede liittyy tiedon keruuseen?

- Miten aineisto generoituu?
- Millaisiin kysymyksiin saadaan kussakin asetelmassa vastauksia?

Tässä luvussa käsiteltävät asiat kuuluvat tilastotieteelle ominaisesti kvantitatiivisen tutkimussuuntauksen alaisuuteen. Luvussa esiteltävät karkeat tutkimustyyppien ja aineistojen jaot ovat vain yksi jaottelutapa ja todennäköisesti poikkeaa eri oppikirjoissa ja lähteissä esitetyistä.

Tässä luvussa tarkastellaan myös erilaisia tutkimusasetelmia ja niiden mukaisia aineistotyyppejä. Erilaiset tutkimusasetelmat johtavat erilaisiin aineistotyyppeihin, jotka voidaan jaotella karkeasti kolmeen eri tyyppiin:

- Poikkileikkausaineisto: havaintoaineisto kattaa yhden ajankohdan ja mahdollisesti useita tilastollisia muuttujia
- Aikasarja-aineisto: havaintoaineisto kattaa vain yhden tilastollisen muuttujan mitattuna useana ajanhetkenä
- Paneeli- ja pitkittäisaineisto: havaintoaineisto kattaa mahdollisesti useita tilastollisia muuttujia mitattuna useana ajanhetkenä

Eri tutkimusstrategiat hyödyntävät eri aineistotyyppejä sen mukaan, miten ne sopivat tutkimuskysymykseen ja valittuun menetelmään. Tarkastellaan seuraavassa siis miten tutkimusstrategiat eroavat ja minkälaisia tutkimustyyppejä-, asetelmia- ja aineistoja näihin kuuluu.

16.1 Tutkimustyypit

Tarkastellaan ensin erilaisia tutkimustyyppejä yleisellä tasolla. Erilaiset tutkimukset voidaan karkeasti jakaa (ainakin) neljään eri luokkaan:

- kuvaileva,
- vertaileva,
- kokeellinen ja
- havainnoiva tutkimus.

Kuvaileva tutkimus

Kuvaileva tutkimus

- Tarkoituksena on kuvata jonkin ilmiön, tilanteen tai tapahtuman luonnetta, yleisyyttä, historiallista kehitystä tai muita tunnuspiirteitä mahdollisimman todenmukaisesti ja tarkasti.
- Keskeistä on **uuden tiedon lisääminen** ja pyrkimys vastata kysymyksiin **mitä**, **millainen** tai **miten**.
  - Yleisesti ottaen kuvaileva tilastollinen tutkimus perustuu aineistosta lasketuille tunnusluvuille, jotka kuvaavat aineiston ominaisuuksia. Esimerkkeinä toimivat keskiarvon lisäksi sen kaltaiset keskimääräistä havaintoa mittaavat suureet, kuten mediaani ja moodi, tai vaihtelua kuvaavat eri muuttujien vaihteluvälit ja keskihajonnat.
  - Saadakseen luotettavia tunnuslukuja, tulee otoksen olla edustava ja havaintojen luotettavia ja päteviä eli saatujen mittausten pitää kuvata kohteena olevaa ilmiötä ilman virheitä.
- Kuvailevassa tutkimuksessa ei tutkita muuttujien välisiä yhteyksiä tai riippuvuksia eikä täten yleensä tehdä jakoa selittäviin ja selitettäviin muuttujiin vaan muuttujat ovat asetelmallisesti samantasoisia.
  - Kuvailevassa tutkimuksessa ei välttämättä testata hypoteeseja, ei tehdä ennusteita, ei anneta selityksiä tai pohdita seurauksia: 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

- Vertaileva tutkimus voidaan jakaa kahteen luokkaan
  1. Ryhmäeroja selittävään tutkimukseen
  2. Korrelaatiotutkimukseen
- **Ryhmäeroja selittävässä tutkimuksessa** pyritään selvittämään, mitkä tekijät liittyvät tutkittaviin ilmiöihin, jotka aiheuttavat ryhmissä ilmeneviä eroja.
- **Korrelaatiotutkimuksissa** pyritään löytämään ilmiöiden välisiä yhteyksiä tutkimalla kohdejoukkoa kokonaisuutena, jolloin mitattavien muuttujien joukkoon otetaan selittäviä muuttujia.
  - Selittäviä muuttujia hyödynnetään molemmissa luokissa. Niiden avulla pyritään löytämään yhteyksiä verrattavien kohteiden välillä ja niiden voidaan ajatella olevan myös mahdollisia syitä selitettäville muuttujille.
- Syy-seuraussuhteita ei kuitenkaan vertailevassa tutkimuksessa pohdita, ts. vertaileva tutkimus ei ole suoranaisesti kiinnostunut kohteena olevien

ilmiöiden/ryhmien vertailussa löydettyjen erojen syistä vaan mielenkiinnon kohteena on kys. erot itsessään.
- Vertailevaa tutkimusta tehdessä on tarpeen pohtia:
  - Miksi jotakin tutkimuskohdetta vertaillaan eli mitä tutkimuskohteesta halutaan nimenomaan saada selville.
  - Mitkä ja minkälaisia tilastoyksiköitä vertailuun kannattaa ottaa mukaan, jotta tutkimuksen tavoitteet saavutetaan.
  - Tyypillistä se, että kontrolli on puutteellista ja ns. väliin tulevia muuttujia ei voida aina eliminoida.
  - Tutkimuksessa on hyväksyttävä myös muuttujiin liittyvä luonnollinen vaihtelu.

Kokeellinen tutkimus

Kokeellinen tutkimus

- Kokeellisessa tutkimuksessa tarkastellaan syy-seuraussuhteita olosuhteissa, joissa tutkija pystyy kontrolloimaan tutkimusyksiköihin vaikuttavia tekijöitä, eli niin sanottuja käsittelytekijöitä (tai interventioita). Tavoitteena on varmistaa, että riippuvaan muuttujaan vaikuttaa vain tutkittava käsittelytekijä, jolloin muiden muuttujien vaikutus voidaan sulkea pois.
  - **Kvasikokeellisessa tutkimuksessa** kontrolli ei ole yhtä tiukka: esimerkiksi koehenkilöiden satunnaistaminen ryhmiin ei välttämättä ole mahdollista. Tällöin tutkimusasetelma muistuttaa kokeellista tutkimusta, mutta altistuu suuremmalle riskille, että muut tekijät vaikuttavat havaittuihin tuloksiin.
- Tavallisesti kokeellisella tutkimuksella viitataan sellaiseen tutkimukseen, jossa aineiston on kerätty valvotussa ja kontrolloidussa ympäristössä, kuten laboratoriossa tai sairaalan koehuoneissa, jotta mittaukset ja käsittelytekijät on tutkimuksen tekijän puolesta kontrolloitu ja täten halutunlaisia.
  - Tutkimusasetelman kontrollointi vähentää mittauksiin ja käsittelytekijöihin liittyvien virhelähteiden mahdollisuuksia ja täten jättää vähemmän sijaa epäilyksille.
  - Lisäksi tutkimuksen toistettavuus ja objektiivisuus paranevat, kun koejärjestelyt tehdään tarkasti ja huolellisesti.
- Kokeellisilla tutkimuksilla on mahdollista päästä **kausaalipäätelmiin** muuttujien välisistä syy-seuraussuhteista.
- Kokeelliset tutkimukset tuottavat yleensä nopeammin riittävään näyttöön perustuvaa evidenssiä kuin havainnoivat tutkimukset.

Kokeellinen tutkimusasetelma ei kuitenkaan ole mahdollinen kaikissa tilanteissa!

Esimerkkejä ja pohdintaa. 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 lainsäädännöllisiin ongelmiin tai niitä ei ole toteutettu riittävän hyvin, jotta asetelma riittäisi kokeelliseen analyysiin.
- Vaihtoehtoisia koeasetelmia on kuitenkin luotu ja tutkimuksia tehty näiden pohjalta.

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. Ts. joiden ilmiöiden siirtäminen laboratorio-olosuhteisiin muuttaa niiden käyttäytymistä.

- Ihmiset käyttäytyvät eri tavalla laboratorio-olosuhteissa kuin normaalissa ympäristössä!

Esimerkki: Lääketieteelliset kokeet. Erään tappavan taudin hoitoon on kehitetty uusi lääke, jonka toivotaan 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 käsittely.

  - Tällöin mahdolliset erot parantuneiden suhteellisissa osuuksissa on oltava seurausta erilaisista käsittelyistä.
  - Kokeen kohteiden jakaminen ryhmiin arpomalla on ainoa menetelmä, joka mahdollistaa samankaltaisten ryhmien saamisen.
  - Kokeen kohteiden jakamista erilaisen käsittelyn kohteiksi joutuviin ryhmiin arpomalla kutsutaan **satunnaistamiseksi**.
  - Arvonnan käyttö ryhmiin jaossa merkitsee sitä, että koetulokset ovat satunnaisia seuraavassa mielessä: Jos arvontaa toistettaisiin, 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: koesuunnittelu, estimointi ja testaus.

Havainnoiva tutkimus

Havainnoiva tutkimus

Kuten edellä mainittiin, kokeellisia tutkimusasetelmia ei useinkaan ole mahdollista järjestää. Tällaisia kysymyksiä voidaan kuitenkin tutkia havainnoivassa tutkimuksessa, jossa syy-seuraussuhteita tarkastellaan tilanteissa, joissa tutkijalla ei ole välttämättä mitään kontrollia (tai syytä sille) tutkimusyksikköihin tai heihin vaikuttaviin muuttujiin (käsittelytekijöihin).

Esimerkiksi tutkimusasetelmat, joissa tutkimuksen kohteena olevia yksiköitä (esim. ihmiset, kunnat, valtiot) ei voida satunnaistaa kuuluvaksi osaksi joukkoa, joka altistetaan jollekin käsittelylle. Niitä saatetaan voida kuitenkin seurata esim. peräkkäisinä ajankohtina, jolloin kyseessä saattaisi olla pitkittäistutkimus.

Tällöin tutkijan on tyydyttävä havainnoimaan sitä mitä tapahtuu luonnostaan tietyssä (mahdollisesti satunnaisesti poimitussa) tutkimusjoukossa tietyssä tilanteessa.

- Havainnoivan tutkimuksen aineistoa voidaan analysoida samoin menetelmin kuin kokeellisen tutkimuksenkin, mutta mitattujen tekijöiden vaikutusta ei voida erottaa kokonaisuudesta samalla tarkkuudella kuin kokeellisessa tutkimuksessa.
- Havainnoivan tutkimuksen tilastollinen teoria muodostuu periaatteista ja menetelmistä, joiden avulla aineiston tuottaman evidenssin painoarvoa voidaan arvioida mahdollisimman "puhtaasti".
- **Havainnoivan tutkimuksen edut**
  - Saadaan välitöntä ja suoraa tietoa yksilöiden, ryhmien ja organisaatioiden toiminnasta ja käyttäytymisestä.
  - Tutkija voi havainnoida tutkittavia luonnollisessa ympäristössä.
  - Sopii sekä määrällisen 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 tai kun halutaan saada selville sellaista tietoa, jota tutkittavat eivät halua suoraan kertoa tutkijalle.
- **Havainnoivan tutkimuksen haitat**
  - Tutkija saattaa häiritä tilannetta tai muuttaa sen kulkua.
  - Tutkija saattaa sitoutua emotionaalisesti tutkittavaan ryhmään tai tilanteeseen.
  - Todellisten kausaalisuhteiden selvittäminen (jos ylipäätään olemassa) voi olla vaikeaa

Esimerkki: raskauden keskeytyksen ja rintasyövän välinen kausaaliyhteys.

- Kokeellinen asetelma: Poimitaan satunnaisesti (n) kappaletta raskaana olevia naisia ja heistä (n_1) kappaletta satunnaistetaan käsittelyryhmään (raskauden keskeytys) ja (n_2) kontrolliryhmään. Kaikki naiset käyvät muutaman seuraavan vuoden ajan syöpäseulonnoissa.
- Kokeellinen 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 lääkityksen aloittamisen jälkeen)

16.2 Poikittaistutkimus eli poikkileikkaustutkimus

Poikittaistutkimus

Poikittaistutkimukseksi kutsutaan tutkimusstrategiaa, jossa tarkoituksena on tutkia kohdetta tai ilmiötä laaja-alaisesti tiettynä ajankohtana käyttäen poikkileikkausaineistoja.

Esimerkkejä. Voidaan tarkastella useita ryhmiä, joissa on esimerkiksi eri-ikäisiä henkilöitä ja ryhmistä saatua tietoa vertaillaan toisiinsa (esim. sydän- ja verisuonitaudeista).

Toisena esimerkkinä, kun tutkitaan sydän- ja verisuonitauteja, niin saatettaisiin käyttää aineistoa, joka koostuu eri ikäisistä ja kuntoisista ihmisistä. Tällöin voidaan arvioida esim. iän ja muiden muuttujien vaikutuksia sydän- ja verisuonitauteihin sairastumisen riskitekijöinä.

Poikittaistutkimuksessa **ei saada tietoa** tilastoyksikön 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: tiettynä aikana syntyneiden, eli tietyn kohortin, elinolosuhteet saattavat olla täysin erilaiset kuin jonakin toisena aikana syntyneiden, minkä vuoksi ikäryhmien väliset erot saattavat johtua esimerkiksi yhteiskunnallisista olosuhteista.
- Poikittaistutkimukseen osallistutaan vain yhden kerran, jolloin tietoa saadaan kerralla paljon.

  - Tämä on kuitenkin usein työlästä ja suuren poikkileikkausaineston kerääminen voi olla kallista.
  - Poikittaistutkimuksessa hyödynnetäänkin usein rutiinitoimenpiteinä kerättyjä aineistoja (esimerkiksi tietyn ikävuoden terveystarkastuksista)
  - Näin voidaan selvittää korrelaatioita ilmiöiden välillä (esimerkiksi alkoholin käyttö ja maksakirroosi) ja siten luoda hypoteeseja tarkemmille jatkotutkimuksille. Tällöin on kuitenkin taas vaara sekoittavista tekijöistä, jos aineistoa ei ole kerätty varta vasten tätä tarkoitusta varten

16.3 Pitkittäistutkimus

Pitkittäistutkimus

Pitkittäistutkimuksessa seurataan usein samoja tilastoyksiköitä ''yli ajan'', eli mittauspisteitä on useita ja mahdollisesti pitkältä aikaväliltä. Ts. hyödynnetään poikkileikkausdimension lisäksi myös aikasarjadimensiota. (dimensio =''ulottuvuus'')

Yleinen tutkimuskysymys pitkittäistutkimuksessa on jonkin **käsittelyn vaikutuksen arviointi**.

- Tällaisia ovat esimerkiksi lääkeainetutkimus, poliittisten päätösten arviointi tai markkinointitutkimus.

Pitkittäistutkimuksessa voidaan siis tarkastella **muutosta**, mutta on tärkeä muistaa, että pitkittäistutkimuksen eri mittauskerrat eivät ole toisistaan **riippumattomia** ja tämä tulee ottaa tilastollisessa mallissa huomioon!

- Tutkittavan ryhmän henkilöt ovat eläneet saman historiallisen ajan sekä käyneet läpi samat yhteiskunnalliset muutokset, jolloin muutoksen tutkiminen on luotettavaa, sillä tutkimusta vääristävät tilastoyksiköiden ominaisuuksista erilliset ympäristön haittamuuttujat ovat kaikille samat.
- Pitkittäistutkimuksen pitkän keston vuoksi tutkittavien määrää kuitenkin yleensä vähenee ja tutkimuksen valmistumisessa kestää kauan, jopa vuosikymmeniä.

Esimerkki: poikittais- ja pitkittäistutkimus epidemiologiassa. Epäkokeelliset epidemiologiset tutkimukset voivat olla joko poikittaistutkimuksia tai pitkittäistutkimuksia.

- Poikittaistutkimus on tiettyyn ajankohtaan rajoittuva tutkimus, jossa mitataan sairauksien vallitsevuutta eli prevalenssia. Prevalenssi kuvaa siis sairauden tai haitan omaavien henkilöiden määrää tarkasteltavasta väestöstä tiettynä ajankohtana.
- Usein mitataan vallitsevuustiheyttä eli sairaiden lukumäärää tiettynä ajanhetkenä suhteessa koko väkilukuun samana ajankohtana.

Pitkittäistutkimuksessa mitataan sairauksien ilmaantuvuutta eli insidenssiä.

- Tutkimuksessa seurataan väestössä ilmaantuvien uusien sairaustapausten lukumäärää tietyn ajanjakson aikana.
- Useimmiten mitataan ilmaantuvuustiheyttä, joka ilmoittaa uusien sairastapausten määrän henkilöaikaa kohden. Henkilöaika muodostuu tarkasteltavan henkilöryhmän yhteenlasketusta seuranta-ajasta ennen sairastumista, esimerkiksi 100 henkilövuotta muodostuu seurattaessa 100 henkilöä vuoden ajan tai 10 henkilöä 10 vuoden ajan.

Aikasarja-analyysi (yksittäisten tai ajoittain useiden aikasarjojen välisten riippuvuuksien) tutkiminen on tavallaan erikoistapaus tästä tutkimustyypistä.

- Aikasarja-aineistoja esitellään vielä erikseen myöhemmin.

16.4 Kohorttitutkimus

Kohorttitutkimus

Kohorttitutkimus on altistelähtöinen pitkittäistutkimuksen muoto, jossa seurataan tiettyä väestöryhmää (kohorttia) ajan yli. Tutkimuksen tavoitteena on selvittää, miten jokin altiste (esim. lääke, ympäristötekijä tai politiikkatoimenpide) vaikuttaa tutkittavaan ilmiöön, kuten sairastuvuuteen tai työllisyyteen.

- **Kohortti** tarkoittaa suljettua väestöryhmää, joka on valittu jonkin yhteisen ominaisuuden perusteella, kuten syntymävuosi (syntymäkohortti), työpaikka tai asuinalue.
- Kohortti voi olla kiinteä (samat henkilöt seurannassa koko ajan) tai avoin (uusia henkilöitä voi liittyä mukaan).

Suuri kohortti voidaan jakaa alakohortteihin, esimerkiksi altistuneisiin ja altistumattomiin ryhmiin. Tutkimuksessa voidaan vertailla altistuneiden ja altistumattomien ryhmien välillä ilmeneviä eroja mielenkiinnon kohteena olevassa muuttujassa.

Esimerkiksi voidaan tutkia rokotteen vaikutusta sairastuvuuteen, tai kuntakokeilun vaikutusta työllisyyteen eri kunnissa. Tällöin insidenssiä (uusien tapausten ilmaantuvuutta) mitataan kummassakin ryhmässä ja vertaillaan.

Kohorttitukimus 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 muuttujia**.
- Tutkimukset saattavat olla hyvin pitkäkestoisia, jos tutkitaan ilmiötä, joka ilmenee vasta pitkä ajan kuluttua altistuksesta (kuten sairaus tai työllisyyden paraneminen)
- Kohorttitutkimus voi vastata kysymykseen: “Mitkä ilmiöt johtuvat tästä altisteesta?”

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

16.5 Tapaus-verrokkitutkimus

Tapaus-verrokkitutkimus

Tapaus-verrokkitutkimus on **retrospektiivinen ja havainnoiva** tutkimusasetelma, jossa tutkitaan altisteiden ja sairauksien välistä yhteyttä vertailemalla kahta ryhmää:

- Tapaukset: henkilöt, joilla on tutkittava sairaus tai ilmiö.
- Verrokit: mahdollisimman samankaltaiset henkilöt, joilla sairautta ei ole.

Molemmilta ryhmiltä selvitetään jälkikäteen, ovatko he altistuneet tutkittuun tekijään. Näin voidaan arvioida, onko altistuminen yleisempää tapausten kuin verrokkien joukossa.

Tapaus-verrokkitutkimus soveltuu erityisesti harvinaisten sairauksien tai ilmiöiden syiden selvittämiseen, koska se ei vaadi suurta väestöpohjaa eikä pitkää seuranta-aikaa. Toisin kuin poikittaistutkimus, joka kuvaa väestön tilaa tiettynä ajankohtana, tapaus-verrokkitutkimus keskittyy tiettyyn populaation osaan ja pyrkii selvittämään mahdollisia syy-seuraussuhteita altisteiden ja sairauksien välillä.

Esimerkkejä:

- Esimerkkinä altistuminen Covid-19 virukselle: retrospektiivisesti (jälkikäteen) voidaan tarkastella viruksen kantajan kanssa samassa tilassa olleita (virukselle altistuneita (virus on altiste)). Kyseisessä tilassa olleet olisivat tapauksia ja hypoteettinen toinen tila ilman virusta toimisi verrokkina (ts. ei yhtäkään tartuntatapausta). Mielenkiinnon kohteena olisi tarkastella kuinka monta henkilöä sai tartunnan (ja minkälaiset olosuhteet olivat).
- Käsittelyn tai altistuksen seurauksia, esimerkiksi sairauden, syitä etsitään vertaamalla tapausten ja verrokkien aikaisempaa altistumista erityisesti mielenkiinnon kohteena oleville altisteille.

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 verrokkikuntiin, joissa kys. politiikkamuutosta ei toteutettaisi. Mielenkiinnon kohteena olisi työllisyyden kehitys altistumisen jälkeen.

Tapaus-verrokkitutkimus eroaa kohorttitutkimuksesta siten että siinä voidaan tutkia **yhtä tilastollista muuttujaa** (kuten sairastumista), mutta **useita altisteita**: mistä altisteesta sairaus on seuraus, ts. mikä on taudinaiheuttaja?

- Altistumishistoriaa voidaan selvitettää mm. mittauksilla, malleilla tai kyselylomakkeilla.

Esimerkiksi tapauksien ja verrokkien altistumiseroista saadaan epäsuora arvio altistuneiden riskistä sairastua kyseiseen sairauteen suhteessa altistumattomien riskiin.

Tapaus-verrokkitutkimukset ovat yleensä suhteellisen yksinkertaisia ja halpoja toteuttaa niiden retrospektiivisestä luonteesta johtuen: tutkimuskysymys määrittelee aineistotarpeen, jonka jälkeen se tarvitsee vain kerätä.

- Verrokkien valinta on kuitenkin kriittinen, sillä valitsemalla verrokit/kontrollitapaukset väärin mikään tilastollinen testi tai menetelmä ei korjaa tai kvantifioi tätä virhettä!

  - Esimerkki verrokkiryhmän epäkelvosta valinnasta on huonosti mitattu aiempi altistuminen ja/tai jos jokin tutkimuksen kannalta keskeinen taustamuuttuja sivuutetaan: mitä jos tauti tai sen vakavuus riippuukin sairastuneen muusta terveydentilasta?

Esimerkki tapaus-verrokkitutkimuksesta. Länsi-Saksassa tuotiin 50-luvun lopulla markkinoille talidomidi-niminen uni- ja rauhoittava lääke. Varsin pian markkinoille tulon jälkeen tietyntyyppisten synnynnäisten epämuodostumien määrä alkoi lisääntyä rajusti.

Talidomidin ja lasten raajojen muodostumishäiriöiden yhteys paljastettiin tapaus-verrokkitutkimuksilla. 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 äitien raskauspahoinvointia). Talidomidi poistettiin markkinoilta ja epämuodostumatapausten määrä putosi jyrkästi.

16.6 Erilaisia aineistoja ja aineistolähteitä

Käydään seuraavaksi läpi erilaisia aineistotyyppejä, joita käytännön tutkimuksissa Suomessa (ja maailmalla) usein käytetään. Emme käsittele tässä erikseen itse otannalla koottavia aineistoja tai otannan järjestämistä, vaan lähinnä aineistotyyppien tuunnuspiirteitä.

16.6.1 Rekisteriaineistot

Rekisteriperusteinen tutkimus hyödyntää aineistoinaan valmiiksi kerättyjä **tietokantoihin** tallennettuja aineistoja, joita kutsutaan **rekisteriaineistoiksi**. Rekisteriaineistot

- Yleensä **hallinnollisia tarpeita** varten kerättyjä tietoja.
- Rekisteriaineistojen eduiksi voidaan lukea mm. seuraavia seikkoja:
  - Aineiston muodostaminen/kerääminen on verrattain helppoa ja Suomessa on paljon korkealaatuisia rekisteriaineistoja. Tätä edesauttaa tietotekniikan nopea kehittyminen, joka on mahdollistanut erittäin suurten aineistojen rutiininomaisen keräämisen.
  - Ei tarvetta erikseen tuottaa tutkimusaineistoa: vältetään mahdollisesti 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, määräävät ne välillisesti myös mahdolliset tutkimuskohteet: rekisteriaineistot kerätään eri tarkoitusta varten eivätkä ne täten välttämättä sisällä kaikkea haluttua informaatiota.
    * Tutkimuksen ongelmalähtöisyys 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 myös esikäsitellä myöhempään analyysiin sopivaan muotoon.
  - "Ulkopuolisille" tutkijoille aineistojen käyttö saattaa olla hankalaa mm. korkeiden pääsykustannusten (rekisterien ylläpitäjien, viranomaisten ja tutkimuslaitosten ulkopuolella), tietosuojakysymysten tai teknisten hankaluuksien takia.
    * Rekisteriaineiston käyttö vaatii usein tarkemman tutkimussuunnitelman ja sen perusteella myönnetyn käyttöluvan rekisterin ylläpitäjältä.

- Tietotekniikan kehittymisen vuoksi kasvaneet rekisteriaineistot tekevät käyttökelpoisen tiedon esiin seulomisesta haastavaa. Tämä näyttäytyy esimerkiksi eri rekistereiden tietojen linkittämisessä yhteen, jolla saattaa olla tutkimuksen kannalta ratkaiseva merkitys ja saattaa edelleen korostaa substanssitietoutta.

  - Eri rekisterejä ei aina saadakaan linkattua tehokkaasti yhteen. Näin esimerkiksi, jos ne ovat mitanneet mielenkiinnon kohteina olevia muuttujia eri tilastoyksikön tasolla (vrt. kunnan vs kaupunginosan työllisyys)

Erilaisia rekisterejä Suomessa:

- Verorekisterit (Verohallinnon rekisterit)
- Kuolemansyyrekisterit
- Eläkerekisterit
- Väestölaskennat (väestörekisteri)
- Syöpärekisteri
- Lääkeostorekisteri
- Sosiaali- ja terveydenhuollon rekisterit

- Kelan etuusrekisterit
- Osoiterekisterit
- Etukorttirekisterit
- Opintosuoritusrekisteri

Rekisteriaineiston käyttämisen **tilastollisia haasteita**:

- Rekisteriaineistot ovat usein kokonaisaineistoja, joten otantavirheeseen perustuvan tilastollisen päättelyn oletukset eivät välttämättä päde. Lisäksi isoissa aineistoissa käytännössä merkityksettömistäkin eroista tulee herkästi tilastollisesti merkitseviä!
- Rekisteriaineistoja saadaan "valmiina" ja niiden kokonaistutkimukseen soveltuvasta luonteesta huolimatta niitä on arvioitava samojen periaatteiden mukaisesti kuin itse kerättäviäkin aineistoja.
- Tutkimusongelman pitäisi aina olla keskeinen lähtökohta myös rekisteriaineiston käytössä.
  - Itse kerätessä aineisto on mahdollista räätälöiden tuottaa vastaamaan juuri tutkimuskysymykseen kun taas rekisteriaineisto on "toisen käden" aineistoa ja ohjaa täten tutkimusta niin käsitteiden määrittelystä kuin tutkimuskysymysten asettelusta lähtien.

Tietosuojalaki ja tieteellinen tutkimus Tietosuojalaki

Tietosuojalain ja EU:n yleisen tietosuoja-asetuksen mukaan henkilötietoja saa käsitellä vain, jos käsittelylle on laillinen peruste. Tieteellisessä tutkimuksessa tämä peruste voi olla esimerkiksi yleinen etu. Henkilötietojen käsittelyssä on keskeistä arvioida, voidaanko aineiston tiedot tosiasiallisesti yhdistää tunnistettavissa olevaan henkilöön. Jos näin on, kyseessä on henkilötieto, ja tietosuojasääntelyä sovelletaan.

Tietosuojalainsäädännön tarkoituksena on ohjata suunnitelmallisuuteen ja huolellisuuteen henkilötietojen käsittelyssä. Sääntelyn tavoitteena on:

- estää tarpeeton henkilötietojen kerääminen ja säilyttäminen,
- suojata rekisteröityjen yksityisyyttä ja oikeuksia,
- varmistaa, että henkilötiedot ja niiden käsittely ovat suojattuja koko käsittelyn elinkaaren ajan.

Rekisteriaineistot ja yksityisyydensuoja tutkimuksessa

- Suomessa rekisteriaineistot pyritään luovuttamaan tutkimuskäyttöön ensisijaisesti tunnisteettomina, jotta yksityisyydensuoja säilyy.
- Tieteelliselle tutkimukselle on säädetty erityisiä poikkeuksia, jotka mahdollistavat henkilötietojen käsittelyn ilman suostumusta, kunhan käsittely on tarpeellista ja asianmukaisesti suojattu.
- Tutkimuksessa tulee ensisijaisesti käyttää anonymisoituja tai pseudonymisoituja tietoja, ellei tutkimuksen toteuttaminen muutoin ole mahdollista. Tällöin on noudatettava tietojen minimoinnin, tietoturvan ja osoitusvelvollisuuden periaatteita.

Luvat ja rekisteröidyn oikeudet

- Henkilötietojen käsittely tutkimuksessa voi edellyttää lupaa rekisterinpitäjältä tai lupaviranomaiselta, erityisesti jos:
  - yhdistetään useita rekistereitä,
  - käsitellään arkaluonteisia tietoja (esim. terveystiedot),
  - tai jos tietoja ei voida anonymisoida.

Joissain tapauksissa voidaan tarvita myös tietosuojavaltuutetun lausunto henkilötietojen käsittelyn lainmukaisuudesta. Rekisteröidyllä on oikeus saada riittävästi tietoa tietojen käyttötarkoituksesta ja käsittelystä, ellei tutkimuksen luonne tai anonymisointi tätä estä.

Luottamus ja läpinäkyvyys: Tietosuojasääntelyn selkeys ja kattavuus ovat myös tutkimuksen hyväksyttävyyden ja luottamuksen kannalta keskeisiä. Kun tietosuojaa koskevat säännöt ovat ymmärrettäviä ja niitä noudatetaan johdonmukaisesti, vähenee yleisön huoli tietojen mahdollisesta väärinkäytöstä.

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 tulkinnat 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 rekisterinpitäjä kerännyt henkilöiden ikätietoa vuoden vai kymmenen vuoden tarkkuudella.

Esimerkki: Diabeteksen ja sen lisäsairauksien esiintyvyyden ja ilmaantuvuuden rekisteriperusteinen mittaaminen

- Vaihe 1: Diabeteskohortin identifiointi (Tilastokeskus: kuolinsyyt, THL: diagnoosit, Kela: erityiskorvaukset, reseptit).
- Vaihe 2: Seurantatiedot (syöpärekisteri, sairaspäivärahat, eläkerekisteri…). Ongelma: kuinka monta henkilöä diabeteskohortista on kuollut seuranta-aikana?
  - Kuolema on vakaa käsite, johon ei liity mittausvirhettä tai subjektiivisuutta.
  - Katsotaan aineistosta kuinka monelle löytyy tieto kuolemasta.
- Kysymys: Kuinka moni diabeteskohorttiin kuuluvista sairastaa tyypin 1 diabetesta?
  - Rakennetaan malli/algoritmi, jolla identifioidaan tyypin 1 diabeetikot lääkeostojen luokkien ja säännöllisyyden perusteella.

Esimerkki: Rekisteritutkimus pitkäaikaisen laitoshoivan käytöstä. Miten sosiaaliset tekijät, kuten sosioekonominen asema ja perherakenne, vaikuttavat laitoshoivan käyttöön?

- Kolme erityistä tutkimusintressiä
  - Laitoshoivaan siirtymisen riskit
  - Laitoksissa vietetty aika
  - Laitoshoivan käyttö elämän loppupäässä
- Aiemmat tutkimukset samasta aiheesta
  - Perustuvat potilasaineistoihin
  - Eivät sisällä laitostumis- ja poistumistietoja samassa aineistossa
  - Kärsivät vastauskadosta
  - Kärsivät seurantakadosta ja seurannan puutteellisuudesta
  - Perustuvat pieniin aineistoihin
  - Eivät mahdollista perhevaikutusten tutkimista

16.6.2 Aikasarjat ja paneeliaineistot

Aikasarja-aineisto

Aikasarja. Aikasarjaksi kutsutaan havaintojen jonoa, jossa aika määrää jostain tilastollisesta muuttujasta tehtyjen havaintojen järjestyksen.

- Aikasarjoissa peräkkäiset havaintoarvot ovat myös tyypillisesti korreloituneita, eli autokorreloituneita, keskenään.

Aikasarjahavainnot ovat tavallisesti peräkkäisiä siten, että mittaukset on tehty tasaisin aikavälein. Väliaikojen tasaisuus ei kuitenkaan ole välttämätöntä ja monissa tutkimusasetelmissa kohdeaikasarjasta voidaan poimia havaintoja jatkuvasti tai mielivaltaisen pienin aikavälein.

- Yksittäinen aikasarja on siis pitkittäisaineiston erikoistapaus, jossa tarkastellaan vain yhtä aikasarjaa. Pitkittäisaineistoon nähden toistot eivät välttämättä ole suunniteltuja, vaan niitä havaitaan jatkuvasti ajassa.

Esimerkkejä. Esimerkkejä aikasarja-aineistoista

- Vuotuinen bruttokansantuote Suomessa
- Suomalaisten lukumäärä kunkin vuoden lopussa
- Vuorokautinen sademäärä Helsingin Kaisaniemessä
- Osakkeiden hintasarjat

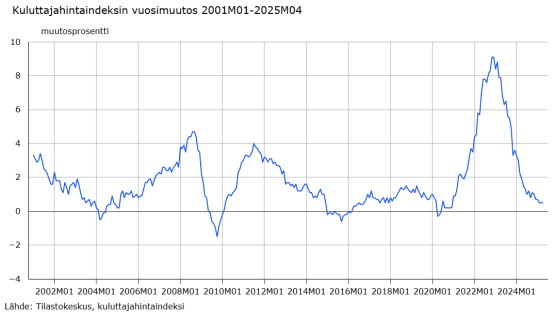
Jotkut aikasarjat ovat suunnitelmallisesti muodostettu tiettyjen sopimusten ja paikoin menttelyjen avulla muista aikasarjoista. Tällaisia tilastollisia suureita kutsutaan **indekseiksi** ja ne sisältävät tiivistettyä tietoa yhteiskunnasta, kuten esimerkiksi inflaation mittarina käytetty kuluttajahintaindeksi. Indeksit

Esimerkki: Kuluttajahintaindeksi (toukokuussa 2025).

https://stat.fi/tilasto/khi

- Kuluttajahintaindeksi kuvaa kotitalouksien Suomessa ostamien tavaroiden ja palveluiden hintakehitystä. Sitä käytetään yleisenä inflaation eli yleisen hintatason nousun mittarina.
- Kuluttajahintaindekseihin kuuluu neljä indeksiä, jotka ovat kuluttajahintaindeksi, elinkustannusindeksi, yhdenmukaistettu kuluttajahintaindeksi ja kiinteäveroinen yhdenmukaistettu kuluttajahintaindeksi. Indeksit mittaavat kuluttajien tavaroista ja palveluista maksamien hintojen kehitystä.

- Perusjoukon muodostavat (Suomen) kotitaloudet ja niiden yksityiset kulutusmenot sekä kulutusmenoihin kuuluvien hyödykkeiden hintojen seuranta.
  - Kotitalouksien yksityisillä kulutusmenoilla tarkoitetaan sitä osaa kulutusmenoista, joka aiheutui Suomen talousalueella vertailuajanjaksolla.
  - Kuluttajahintaindeksin, yhdenmukaistetun kuluttajahintaindeksin sekä yhdenmukaistettu kuluttajahintaindeksi kiintein veroin. Nykyinen perusvuosi on 2015.
- Tilastokeskuksen haastattelijat keräävät indeksiä varten hintatietoja eri hyödykkeistä ympäri maata.



Kuva: Kuluttujahintaindeksin vuosimuutos (tammikuu 2001–Huhtikuu 2025). Lähde: Tilastokeskus.

Aikasarjojen tilastollinen analyysi perustuu siihen, että sarja tulkitaan (taustalla olevan) **stokastisen prosessin** eli satunnaisprosessin realisaatioksi

- Jos aikasarjan generoinut prosessi saadaan selville, voidaan tietoja prosessista käyttää aikasarjan käyttäytymisen kuvaamiseen ja selittämiseen sekä aikasarjan tulevan käyttäytymisen ennustamiseen.
- 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-analyysia opiskellaan tarkemmin useilla Turun yliopiston erikoiskursseilla.
- Aikasarjoja analysoimalla voidaan selvittää esimerkiksi
  - Onko aikasarjassa **trendejä** eli aikasarjan tason systemaattisia muutoksia ajan kuluessa?

  - Onko aikasarjassa **syklistä vaihtelua** kuten **suhdanne**- ja/tai **kausivaihtelua**?

Paneeliaineistot **Paneeli- eli pitkittäisaineistot**. Paneeliaineistolla tarkoitetaan aineistoa, jossa tilastoyksiköistä on useita havaintoja ja aika määrää havaintojen järjestyksen (kuten aikasarjoissa) ja lisäksi jokaisena ajanhetkenä mitataan useampi kuin yksi tilastollinen muuttuja (kuten poikkileikkausaineistossa).

- Paneeliaineisto on terminä käytetympi yhteiskuntatieteissä kun taas pitkittäisaineisto esimerkiksi lääketieteessä.
- Pitkäisaineiston ideaa tarkasteltiin jo edellä.
- 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.
  - Paneeliaineisto mahdollistaa vastaamisen kysymykseen miksi? Yleisesti ottaen paneeliaineistoja käytetäänkin erityisesti ns. kausaalipäättelyyn tähtäävissä malleissa.

16.6.3 Haastattelu- tai kyselytutkimus

Kyselytutkimus

Survey-tutkimus (mielipidetiedustelu) 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 surveytutkimuksessa pyritään yleistämään tulokset otoksesta koko perusjoukkoon.

Kyselytutkimukset (surveyt) muodostavat kokonaan oman tutkimustapansa, joka mahdollistaa hyvin erilaisen informaation keräämisen kuin tavallisesti kvantitatiivissa aineistoissa. Sen voidaan katsoa koostuvan mm. seuraavista vaiheista:

- Kohdepopulaation määrittely ja otannan suunnittelu
- Kyselylomakkeen rakentaminen ja testaaminen
- Kyselymetodin määrittely (puhelinhaastattelu, elektroninen kysely, tai jotain muuta…)
- Mahdollisten haastattelijoiden koulutus
- Aineiston keräys
- Aineiston yhteneväisyyden tarkistaminen (muuttujat tallennettu oikein jne.)
- Tulosten adjustointi mahdollisten identifioitujen virhelähteiden mukaan

Esimerkkejä ovat mm. poliittiset mielipidekyselyt, markkinointitutkimukset, alkoholinkulutustottumukset ja terveyspalveluiden tyytyväisyyskyselyt, joihin voi kaikkiin uskoa liittyvän vastausharhaa eri syistä.

- Esimerkiksi vastausharhaa arkaluontoisiin kysymyksiin voidaan vähentää 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 altisteinen tehdylle kyselylle ja täten sen käyttökelpoisuus perustuu hyvin pitkälti etukäteissuunnitteluun, kunnolliseen toteutukseen ja kyselylomakkeen oikeaoppiseen rakentamiseen.

- Lisäksi aineiston käyttökelpoisuus riippuu myös vastaajaotoksen poiminnasta (edustavuudesta) ja siitä, kuinka totuudenmukaista informaatiota vastaajat ovat kyselyssä antaneet. Tässäkin hyvä etukäteissuunnitelu on keskeistä.
- Tilastollisten menetelmien avulla pyritään arvioimaan otoksen, kyselyn suunnittelun ja kerätyn vastaajaotoksen sisältämää tai aiheuttamaa harhaa.

Chapter 17

Tilastotieteen koulukunnat: Frekventistisyys vs. Bayesiläisyys

Tilastotiede voidaan karkeasti jakaa kahteen merkittävään paradigmaan sen mukaan, miten **tilastolliseen päättelyyn**, ml. hypoteeseihin ja niiden testaamiseen, suhtaudutaan. Näitä ovat **klassinen eli frekventistinen tilastotiede** sekä **Bayesilainen tilastotiede**.

Tarkastellaan seuraavaksi minkälaisia eroja ja yhtäläisyyksiä näiden koulukuntien välillä on.

Frekventistinen tilastotiede **Frekventistinen tilastotiede**

- Klassisessa eli frekventistisessä tilastotieteessä ajatellaan että hypoteesien testaaminen tulee perustua yksinomaan havaittuun aineistoon ja siihen liitettävään tilastolliseen malliin.
    - Tämän materiaalin tilastollista päättelyä koskevat kohdat ja esitellyt lähestymistavat ja menetelmät lukeutuvat klassisen tilastotieteen koulukuntaan.
- Nimi ''frekventistinen'' juontuu siitä, että tilastollisen mallin perustana oleva todennäköisyysjakauma määrittää satunnaismuuttujan mahdollisten arvojen todennäköisyydeksi niiden suhteellisen osuuden äärettömästä määrästä realisaatioita, ts. niiden suhteellisen frekvenssin.
- Klassisessa tilastotieteessä havaittuun aineistoon sovitetaan tilastollinen malli, joka vastaa saatua aineistoa parhaiten.

  - Tämä tilastollinen malli voidaan (useimmiten) perustaa uskottavuusfunktioon, joka on siis aineiston sekä yhden tai useamman parametrin funktio ja joka saavuttaa suurimman arvonsa suurimman uskottavuuden pisteessä. Uskottavuusfunktio kertoo kuinka todennäköisenä havaittua aineistoa voidaan pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvolla. Täten ne parametriarvot, joilla uskottavuusfunktion arvo maksimoituu, *kuvaavat aineiston generoimaa prosessia parhaiten*, annettuna mallissa tehdyt oletukset, kuten jakaumaoletus/oletukset.

- Tutkimuskysymyksen mukaisia 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.

  - Kaikki päättely on toki alisteista tehdyille oletuksille koskien käytettävää tilastollista mallia.

Bayesiläinen tilastotiede **Bayesilainen tilastotiede**

- Bayesilainen tilastotiede on tilastotieteen toinen suuri paradigma ja on saanut nimensä englantilaiselta harrastelijamatemaatikko ja presbyteeripappi Thomas Bayesilta, jota pidetään bayesilaisen tilastotieteen isänä.

- Bayesilainen tilastotiede ulottaa todennäköisyyskäsityksen, eli todennäköisyysjakauman, myös aineistoa koskevien hypoteesien puolelle: kuinka todennäköisenä jotain hypoteesia voidaan pitää jo ennen tutkimusaineiston keräämistä?

  - Myös bayesilaisessa tilastotieteessä hyödynnetään uskottavuusfunktiota, mutta hypoteesien testaus ei perustu niinkään frekventistiseen ajatukseen todennäköisyyksistä 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 todennäköisyyksinä käytetään niin bayesilaisen tilastotieteen kritiikkinä kuin puolustuksena.

- Lopulta olemme kaikki bayesilaisia: jokaisella on sisäisiä ennakkokäsityksiään, myös tutkijoilla! Nämä ennakkokäsitykset voivat perustua esimerkiksi aiempaan tutkittuun tietoon, mutta myös uskomuksiin.
- Prioritiedon hyödyntäminen tilastollisessa tutkimuksessa on usein perusteltua, mutta prioritiedon käyttämistä voidaan pitää myös subjektiivisena ja täten ongelmallisena tulosten luotettavuuden kannalta.

Chapter 18

Tilastotieteeseen kohdistunutta kritiikkiä

- Tilastotieteen rooli tiedeyhteisössä on niin tärkeä että sitä kohtaan on ymmärrettävästi esitetty myös melko paljon kritiikkiä. Valtaosa kritiikistä kohdistuu joko tilastotieteen matemaattisuuteen tai sitten 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 vieras 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.
- On syytä korostaa, että seuraavassa käsiteltävät näkökulmat ovat monin paikoin sellaisia, joiden ymmärtämiseksi vaaditaan laajempia tilastotieteen opintoja mitä tällä kurssilla ehditään käsitellä. Tilastotieteen opintojen edetessä seuraavat, kriittisetkin, näkökulmat tulevat yhä paremmin ymmärrettäväksi.

“Yleismaailmallinen” kritiikki.

- 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.

**Esimerkkejä: Mahdoton keskiarvo: Puhekielessä kuulee paikoin tilastotiedettä kritisoitavan nk. "Keskiarvo-Kimmon" avulla, eli kuvitteellisella ihmisellä, joka on 1,8 lapsen vanhempi ja 1,5 auton omistaja.

Lisäksi joskus kuulee tilastotieteilijöitä kritisoitavan lausumalla "Jos toinen jalka on jääkylmässä vedessä ja toinen kiehuvassa vedessä, niin tilastotieteilijän mielestä ihmisellä on tällöin keskimäärin hyvä olla"

- **Korrelaatio** on **tunnusluku**, joka kuvaa kahden muuttujan välistä riippuvuutta. Se ei kuitenkaan kuvaa millään tavoin **kausaalisuutta**, eli sitä kumpi aiheuttaa kumman, jos kumpikaan. Vrt. aiemmin materiaalissa käydyt keskustelut ja esimerkit kahdenvälisistä nk. "näennäisistä" korrelaatioista.

Esimerkiksi "jäätelön syönti ja hukkumiskuolemat' ' -tapauksessa havainnollisesti todetaan jäätelönkulutuksen ja hukkumiskuolemien lukumäärän korreloivan keskenään, mutta taustalla vaikuttava tekijä onkin lämmin kesä, joka vaikuttaa molempiin.

Kritiikki matemaattisuutta kohtaan

- Ehkä merkittävin kritiikki tilastollisia menetelmiä kohtaan kohdistuu kritiikin näkökulmasta perusteettomaan, tai ainakin liian vahvaan, matemaattisuuden tuomaan itsevarmuuteen. Voidaankin siis perustellusti kysyä, että **onko tieteellisyys = matemaattisuus**?
  - Useat tieteenalat käyttävät tutkimuksessaan edistyneitäkin tilastollisia menetelmiä siitä huolimatta, että tutkijoiden tilastomatemaattinen pohjakoulutus ei välttämättä ole riittävällä tasolla kyseisten menetelmien kokonaisvaltaiseen ymmärtämiseen.
    * Helppokäyttöisistä 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).
    * Laskentatehon ja modernin tietojenkäsittelyteknologian, ja nyt laajemmin myös kasvavissa määrin tekoälyn, ansiosta monimutkaisiakin tilastollisia analyysejä on kuitenkin mahdollista tehdä vaikka tutkijalla olisi tilastotieteestä vain perustiedot, jos sitäkään.

        * Pahimmillaan 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 pystyttävä jäljittämään.

    – Teknistä esitystä käyttävää tutkijaa saatetaan pitää erityisen uskottavana, koska hän kykenee käyttämään vaikeita menetelmiä. Tästä huolimatta tutkimusongelma ei saisi päästä 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 matemaattisuudestaan huolimatta ole välttämättä neutraaleja!

- Kaikkia tieteentekijöitä ei voida velvoittaa opiskelemaan edistynyttä tilastotieteen teoriaa (tilastomatematiikkaa), mutta menetelmien oikeaoppinen käyttö kuitenkin vaatii riittävää ymmärrystä.

Kritiikki yksinkertaistuksia kohtaan

- Edellisiä kohtia yleisemmin tilastotiedettä on kritisoitu 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 “tosielämän 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 kritisoitu sen "sokeudesta" sosiaaliseen vuorovaikutukseen liittyviin subjektiivisiin kokemuksiin kuten tunteisiin, kokemuksiin ja ei-numeerisiin havaintoihin.

  - Tämä kritiikki ei kuitenkaan suoranaisesti ole tilastotieteen kritiikkiä, vaan jälleen sovellusalakohtainen ja erityisesti tutkimuskysymyksen asettelua koskeva ongelma.

    * Tuntemuksia ja kokemuksia voidaan hyvin testata tilastollisin menetelmin, 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 kvantitatiivisilla mittauksilla saamaan 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 käyttökelpoisia.

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 oikeanlaisen menetelmän voi vastata mihin tahansa tutkimuskysymykseen!

  - Tilastolliset menetelmät tulee ymmärtää ja niitä tulee soveltaa kaikissa soveltavan tutkimuksen vaiheissa (vrt. **OSAAT-sykli**), jotta tutkimusongelmaan kyetään vastaamaan eikä turhaa työtä tule tehdyksi.
  - Karkeasti luokitellen tilastotieteilijät kehittävät menetelmiä, joita soveltajat käyttävät.
    * Soveltavia tilastotieteilijöitä löytyy kuitenkin yhä kiihtyvissä määrin! Erityisesti eri rajatieteiden alueilla, kuten biometria, ekonometria jne. (ks. aiempien lukujen tarkemmat rajatieteiden esittelyt).

Tilastotieteen väärinkäyttö

- 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 johtuvaa).
  - Joskus kuitenkin näkee tarkoituksellista tilastojen vääristelyä tai tahallista tilastollisten menetelmien väärinkäyttöä! Uutisjutussa "Jäätelö lisää hukkumiskuolemia – vai miten se menikään?
  - **Kansalaisten tiedelukutaidon** ja tilastollisten menetelmien tuntemuksen merkitys on kasvanut viime vuosikymmeninä ja kasvanee jatkossa yhä, kun esimerkiksi erilaiset "vaihtoehtotieteet'' ovat nousseet suosituimmiksi.
  - Tilastotieteen ymmärrys auttaa itse kutakin tunnistamaan virheellisiä tai puutteellisin tiedoin tehtyjä päätelmiä ja täten helpottaa tietoyhteiskunnassa toimimista ja kriittistä ajattelua!
- Yleisiä tilastollisten menetelmien väärinkäyttötapoja ovat esimerkiksi seuraavat:
  - **"Kolmannen tyypin virhe''**: kun tilastollisia menetelmiä käyttämällä saadaan oikeita vastauksia, mutta vääriin kysymyksiin! Esimerkiksi, jos tutkija ei täysin ymmärrä minkälaisia vastauksia käytettävissä olevasta aineistosta ja valitulla menetelmällä voidaan saada, voi hän syyllistyä kolmannen tyypin virheeseen. Tällöin voi nimittäin käydä niin, että hän tulkitsee tilastollisten menetelmien, kuten tilastollisten testien tulokset, täysin oikein, mutta luulee 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 käytettävissä edustava otos molempien ikäluokkien edustajista.

- Päätät tutkia *yksisuuntaisesti* onko toisen ryhmän, ryhmän A, keskipituus *pienempi* kuin ryhmän B.
  - Testitulos osoittaa, että voit hylätä nollahypoteesin, jonka mukaan 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 tätä et testin tuloksen perusteella voi tietää!

Chapter 19

Tilastotieteen kehityksen nykytrendejä

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

- **Aineistot kasvavat ja monimutkaistuvat**. Näin ollen tilastotietedettä ja tilastollisten menetelmien kehitystyötä tullaan tarvitsemaan (yhä enemmän) 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) valikoivasti ja subjektiivisesti. Luulemme haluavamme lisää informaatiota, kun todellisuudessa haluamme tietoa (ts. signaaleja kun vastaavasti kasvava määrä kohinaa yrittää vaikeuttaa tätä signaalin erottamista kohinasta).

- **Laskennallisuus kasvaa**. Tietokoneiden laskentakapasiteetti nousee vaikkakin suhteellinen kasvu ei ehkä olekaan enää niin suurta mitä muutamat viime vuosikymmenet.

- Osin laskennallisuuteen liittyvä **koneoppimisen** kehittyminen ja sen ''rajatieteiden'' yhteyteen integroituneet käytännöt ja menetelmäkehitystyö tulee jatkumaan.

  - 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 kyettävät arvioimaan ovatko tulokset uskottavia ja vapaita ilmeisistä hankaluuksista.
  - Näiltä osin yksi keskeinen kehityssuunta on jo vahvasti yleistyneet laajat tekoälymallit, kuten kielimallit. ChatGPT:tä, ja vastaavia tekoälysovelluksia, käytetäänkin jo laajalti (vuonna 2025) mm. tilastollisen ohjelmoinnin apuvälineenä. Nopeasta kehityksestä huolimatta ChatGPT:n antamat vastaukset ovat kuitenkin paikoin epäluotettavia, joten sen käytössä korostaa edelleen vahvan tilastotieteellisen osaamisen lisäksi substanssiosaaminen.
- **Poikkitieteellisyys tulennee entisestäänkin vahvistumaan**. Ts. substanttitietouden ja tilastotieteen yhdistäminen ja sen tärkeys ei tule ainakaan vähenemään. Sivuaineopintoja kannattaa siis käydä ja ottaa opinto-ohjelmaan!
  - Tämän lisäksi kokonaisvaltaisesti tilastotieteen ytimen osaajien osaamista tulla kysymään jatkossakin.

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

Part I

Liitteet

Liite A: Kreikkalaiset aakkoset

Iso kirjain	Pieni kirjain	Nimi	Latin. vastine
A	α	alpha	a
B	β	beta	b
Γ	γ	gamma	g
Δ	δ	delta	d
E	ε	epsilon	e
Z	ζ	zeta	z
H	η	eta	e
Θ	θ	theta	th
I	ι	iota	i
K	κ	kappa	k
Λ	λ	lambda	l
M	μ	mu	m
N	ν	nu	n
Ξ	ξ	xi	x
O	o	omicron	o
Π	π	pi	p
P	ρ	rho	r
Σ	σ	sigma	s
T	τ	tau	t
Υ	υ	upsilon	y / u
Φ	φ	phi	f / ph
X	χ	chi	kh
Ψ	ψ	psi	ps
Ω	ω	omega	O