

TILM3701 - Tilastotiede ja data 2022

Koonneet Henri Nyberg¹ Roope Rihtamo²

2022-09-22

¹Turun yliopisto, matematiikan ja tilastotieteen laitos, henri.nyberg@utu.fi
²Turun yliopisto, matematiikan ja tilastotieteen laitos, roope.rihtamo@utu.fi

Sisällys

Kurssin rakenne	7
Kurssimateriaali	8
1 Johdantoa ja johdattelua tilastotieteeseen	11
1.1 Tilastotiede ja kurssin idea	11
1.2 Tilastotieteen asema tutkimusyhteisön ulkopuolella	13
1.3 Kurssin luonne tilastotieteen opintojen esittelijänä	14
2 Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa	15
2.1 Mitä on tiede?	15
2.2 Tieteellinen menetelmä	20
2.3 Tilastojoen yleisestä roolista yhteiskunnassa	23
2.4 Mitä on tutkimus?	25
2.5 Tutkimuksen vaiheet ja tulosten julkaiseminen	28
3 Tilastotiede tieteenalana	31
3.1 Lisää tilastotieteen perustermejä	31
3.2 Mitä tilastotiede on ja mitä se ei ole?	33
3.3 Tilastotieteen suhde lähitieteisiin	38
3.4 Tilastotieteen osa-alueet	42
3.5 Tilastotieteen kritiikkiä	46
3.6 Tilastotieteen sovelluskohteita ja “rajatieteitä”	52

4 Sattuma ja satunnaisuus tilastotieteessä	55
4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä	56
4.2 Satunnaisuus ja todennäköisyydet	59
4.3 Tilastolliset mallit, jakaumat ja parametrit	62
4.4 Odotusarvo ja varianssi	64
4.5 Joitain jakaumia	65
4.6 Sattuman rooli tieteenteossa: Vale-emäviale-tilasto?	71
5 Tilastolliset aineistot, niiden kerääminen ja mittaaminen	73
5.1 Kertausta: Data eli aineisto	74
5.2 Otannan idea	78
5.3 Mittaaminen ja mitta-asteikot	81
5.4 Kontrolloidut kokeet ja suorat havainnot	86
5.5 Otantamenetelmät	89
5.6 Otantaesimerkkejä	97
5.7 Otannan haasteita vielä kootusti	98
6 Otokset ja otosjakaumat: tilastollisen päätelyn näkökulma	101
6.1 Satunnaisotos, yhteisjakauma ja tilastollinen malli	101
6.2 Otosjakauma: Estimaattori ja estimaatti	104
6.3 Otoskeskiarvo ja otosvarianssi (estimaattoreina)	107
6.4 Suhteellisen frekvenssin otosjakauma	110
6.5 Muita tunnuslukuja	112
6.6 Luottamusvälit	113
6.7 Otoskoko	119
7 Tilastollinen riippuvuus ja korrelaatio	127
7.1 Muuttujien väliset riippuvuudet	127
7.2 Kahden muuttujan havaintoaineiston kuvaaminen	129
7.3 Tunnusluvut	131
7.4 Satunnaismuuttujien kovarianssi ja korrelaatio	133

SISÄLLYS	5
8 Regressioanalyysi	141
8.1 Johdatus regressioanalyysin ideaan	141
8.2 Yhden selittäjän lineaarinen regressiomalli	141
8.3 Muita regressiomalleja	141
9 Tilastotieteen rooli uuden tiedon tuottamisessa	143
9.1 Tilastollisen tutkimuksen yhteisiä elementtejä	143
9.2 Tutkimusprosessi	143
10 Aineisto- ja tutkimustyyppit ja koeasetelmat	145
10.1 Tutkimustyyppit	145
10.2 Tutkimusstrategiat	145
10.3 Eriisia aineistoja ja aineistolähteitä	145
11 Tilastollisesta ennustamisesta	147
11.1 Tilastollinen selittäminen vs. ennustaminen	147
11.2 Tilastolliseen ennustamiseen liittyviä huomioita	147
12 Tilastotieteen kehityksen nykytrendejä	149

Kurssin rakenne

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

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

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

Kurssimateriaali

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

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

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

- Kurssin sisältöä:

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

Materiaalin seassa on eritylty väärinkoodatuin tietolaatikoin erinäisiä tärkeitä tilastotieteellisiä konsepteja ja termejä sekä esimerkkejä tilastotieteen sovelliuksesta. Näistä ensin mainitut löytyvät Deltan violetista laatikoista ja jälkimmäiset Statistikkan oransseista.¹ Alla esimerkkilaatikot.

Konsepti tai termi

Konseptin tai termin löyhä määritelmä.

Esimerkki

Aihetta koskeva esimerkki.

¹Toim. Huom. värit eivät täysin alkuperäisten värien kanssa yhteneväisiä.

Luku 1

Johdantoa ja johdattelua tilastotieteeseen

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

Holopainen ja Pulkkinen (2008)

1.1 Tilastotiede ja kurssin idea

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

12 LUKU 1. JOHDANTOA JA JOHDAATELUA TILASTOTIETEESEEN

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

1.2. TILASTOTIETEEN ASEMA TUTKIMUSYHTEISÖN ULKOPUOLELLA13

- * **Otanta** (ja mittaaminen): miten tutkimusaineisto kerätään niin, että se täyttää aineiston ennakkoehdot? Erilaisissa tutkimuksissa käytetään erilaisia aineistoja kuten:
 - Survey- eli haastatteluaineistot: aineisto kerätään haastattelemalla tutkimuskohteita
 - Rekisteriaineistot: aineisto on kerätty valmiiksi rekisteriin ja sitä käytetään tutkimukseen
 - Aikasarja-aineistot tai pitkittäisaineistot: useita mahdollisesti korreloituneita havaintoja samoista tutkimuskohteista
 - Ynnä muita, ks. [10](#)
- **Aineiston kuvaaminen:** minkälaisista aineistoa on kerätty ja vastaako se ennakkoehtoja?
- **Aineiston analyysin** lähtökohtia
 - Mitä tilastollista mallia/malleja käytetään?
 - Mitä tarkoitetaan mallien tuntemattomien parametrien arvojen estimoinnilla?
 - Tilastollinen päättely (estimointitulosten pohjalta)
- **Johtopäätelmien** tekeminen tilastollisen päättelyn pohjalta: saatiinko tutkimusongelmaan vastaus ja kuinka luotettava saatu vastaus on?

1.2 Tilastotieteen asema tutkimusyhteisön ulkopuolella

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

1.3 Kurssin luonne tilastotieteen opintojen esittelijänä

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

Luku 2

Tieteellinen tieto, tilastot ja arkitieto yhteiskunnassa

Tässä luvussa tarkastellaan tieteen ja tieteellisen tutkimusprosessin luonnetta erityisesti uuden **tutkitun** tiedon tuottamisen näkökulmasta. **Tiedelukutaidon** merkitys on kasvanut nyky-yhteiskunnassa, kun tiedejulkaisujen saavutettavuus ja tunnettiuus on lisääntynyt mm. tieteen popularisoinnin ja median laajemman tiede-uutisoinnin vuoksi. Tiedon, erityisesti tieteellisen tiedon, rooli korostuu yhä enemmän myös kaikilla elämän osa-alueilla: terveysteknologia (esim. sykemittarit tai Oura-sormus) perustuu lääke- ja terveystieteellisiin läpimurtoihin, talouspoliittisia päätöksiä edeltää entistä suurempi määrä asiantuntijoiden taloustiedeperusteista analyysia ja jopa peruskouluopetus on murroksessa kasvatustieteen saavutusten myötä.

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

2.1 Mitä on tiede?

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

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

16LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

ja hankitaan **tietoa**.

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

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

Tieteellinen teoria

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

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

18LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

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

Hypoteesi

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

toihin, joita ei voida selittää olemassa olevilla tieteellisillä teorioilla tyydyttävästi.

- Uuden tieteellisen tiedon tuottaminen ja jo tuotetun tiedon ymmärtäminen vaatii **tieteellisen ajattelutavan** omaksumista, jonka **perustana on lähes aina tilastollinen päätely**.
 - Tieteelliselle ajattelulle ja tiedon tuottamiselle on tunnusomaista, että se pohtii ja kehittelee **paradigmojaan** eli oman toimintansa perusteita.

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

- Esimerkkinä toimii taloustieteen nk. “**uskottavuusvallankumous**”, jossa tilastollisten menetelmien myötä taloustieteellisen tutkimuksen painopiste tuntuu siirtyneen vahemmin empiirisen kausaalitutkimuksen puolelle.

- Paradigmat siis ohjaavat uuden tieteellisen tiedon tuottamista asettamalla tutkimukselle yhtenevät raamat, jotka ohjaavat sitä, miten tutkimuskysymksiä asetetaan ja miten niihin etsitään vastauksia sekä myös sitä, miten saatuja tuloksia tulkitaan.
 - Tieteellinen tieto perustuu siis eri tutkimusalojen tiedeyhteisöjen paradigmoihin ja täten siihen, minkälaisista tutkimusta, ja mistä ilmiöistä, kannattaa tehdä.
 - Paradigmojen ei pidä ajatella olevan kaavoihin kangistuneita 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 estävät osaltaan uusien löytöjen syntymistä: liian vahasti alan paradigmoiden 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!
 - Tieteelliseen ajattelutapaan kuuluu olennaisesti juuri tiedon kumuloitumisen ymmärtäminen: yksittäinen vahva tulos on vasta alku ja

20LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

vahvistettu tieto jostain ilmiöstä, yhteydestä tai vaikutuksesta syntyy monien mittausten ja tutkimusten jatkumona.

- Tietoa ei siis voida johtaa siitä, miltä asiat näyttävät, kuten on tyypillistä “arkiajattelussa”.
 - * Tiede kehittää teorioita kriittisesti ja määritetietoisesti rationaalisen ajattelun keinoin.
 - * Teorioita ja niihin liitettäviä hypoteeseja testataan tieteellisin menetelmin ja näin saadaan uutta tietoa tutkittavasta ilmiöstä.
- Tiivistetysti voidaan sanoa että tiede on kumulatiivinen tutkimusprosessi, jossa hankitaan uutta tietoa ja samalla vahvistetaan vanhaa, mutta epävarmaa tietoa tieteellisin menetelmin.
 - * Tieteellisten menetelmien käyttöä ohjaa tutkimusalakohtaiset paradigmat, jotka ovat suuntaviivoja ja viiteistöjä siitä, minkälainen tutkimus tuottaa uskottavia tuloksia.

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

Kuva 2.1: Arkitieto ja tieteellinen tieto

2.2 Tieteellinen menetelmä

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

Tieteellinen menetelmä: Tieteellinen menetelmä on kyllakin tieteen alalla vallitseva, ajan myötä kehittynyt ja nykyisten paradigmoiden mukainen menettelytapa, jolla uutta tietoa tuotetaan ja vanhaa, mutta epävarmaa tietoa vahvistetaan. Se ei ole selkeä työvaiheiden luettelo tai menetelmähakemisto, vaan yleisesti hyväksytty ja hyväksi todettu tapa pyrkii totuuteen erilaisten tutkimusongelmien ratkaisussa. Hyvälle tieteelliselle menetelmälle voidaan lukea seuraavia kriteerejä.

- **Objektiivisuus ja loogisuus**

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

- **Kriittisyys**

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

- **Autonomisuus**

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

- **Edistyyvyys**

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

- **Toistettavuus ja yleistettävyys**

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

22LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- Tarkastellaan lähemmin erästä määritelmää erilaisille toistettavuuden lajeille. Esittemme tässä Hamermeshin (2007)² esittämän erilaisten replikointien jaottelun:
 - **Puhdas replikointi:** toinen tutkija, käyttäen täysin samaa tutkimusaineistoa ja samaa tilastollista menetelmää kuin alkuperäisessä tutkimuksessa, saa täsmälleen samat tutkimustulokset.
 - **Tilastollinen replikointi:** toinen tutkija, käyttäen eri tutkimusaineistoa (otosta), joka on kuitenkin poimittu samasta populaatiosta (ks. Luku 5), mutta samaa menetelmää, saa vastaavanlaisia tuloksia, jotka vahvistavat alkuperäisen tutkimuksen perustulokset.
 - **Tieteellinen replikointi:** toinen tutkija, käyttäen samoja asioita mittaavaa tutkimusaineistoa, joka on kuitenkin kerätty eri populaatiosta, ja käyttäen samankaltaista, mutta ei identtistä menetelmää, saa vastaavanlaisia tuloksia, jotka vahvistavat alkuperäisen tutkimuksen perustulokset.
- Teorioiden sisältämiä väitteitä voidaan muotoilla tieteellisiksi malleiksi, joihin voidaan liittää hypoteeseja, joita testataan tieteellisin menetelmin käyttäen ilmiö(i)stä mitattua havaintoaineistoa.
 - Tieteelliset mallit ovat yksinkertaistuksia reaalimaailmasta ja ne kuvaavat tutkimuksen aihetta jostain näkökulmasta tarkasteltavana systeeminä.
 - Mallit hyödyntävät matemaattista esitystapaa, sillä se tarjoaa formaalin ja objektiivisen tutkimusaiheen kuvaukseen sekä mahdollistaa siihen liittyvän loogisen päättelyn havaitun, empiirisen aineiston pohjalta.
 - Tilastolliset mallit ovat käytännössä tieteellisten mallien formaaleja matemaattisia esityksiä, jotka lisäksi mahdollistavat mallia koskevan tilastollisen päättelyn esimerkiksi hypoteesien ja niiden testaamisen avulla. Päättely perustuu tilastotieteen teoriaan, joka mahdollistaa päättelyn epävarman ja satunnaisen aineiston tapauksissa.
 - Hypoteesien asettamisen voidaan ajatella tutkittavaa ilmiötä koskevaksi ennusteiksi, joita verrataan havaittuun aineistoon. Mikäli havaittu aineisto ei sovi testattavaan teoriaan tai siihen liittyviin hypoteeseihin, voidaan (hieman yksinkertaistaen) teoriaa kehittää paremaksi. Tämä vuoropuhelu vie tiedettä eteenpäin ja tuottaa lisää tutkittua tietoa ympäröivästä maailmasta.

²Hamermesh, D. S. (2007). Replication in economics. *Canadian Journal of Economics/-Revue canadienne d'économique* 40 (3), 715–733.

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

2.3 Tilastojen yleisestä roolista yhteiskunnassa

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

Esimerkki (väite): Naiset puhuvat enemmän kuin miehet.

- Lähtökohta väitteen (hypoteesin) tutkimiseen:
 - Uskomus on väärä kunnes toisin todistetaan.
 - Lähdetään liikkeelle olettamuksesta, että miehet ja naiset puhuvat yhtä paljon.
 - Olettamuksen tueksi tai kumoamiseksi täytyy kerätä todistusaineistoa.
 - Jotta tutkimukseen saataisiin täysin varma vastaus, kaikki miesten ja naisten puheet ihmiskunnan olemassa olon ajalta pitäisi pystyä laskemaan = mahdotonta.
- Mitä siis tehdä?
 - Täytyy tyytyä tutkimaan osajoukkoja miehistä ja naisista (otos), mihin tarvitaan **otantamenetelmiä** (käsitellään tarkemmin myöhemmin luvussa 5).
 - Arvotaan satunnaisesti tutkimushenkilötä miesten ja naisten joukosta ja mitataan kuinka paljon he puhuvat.
 - Satunnaisuus tärkeää, sillä jos valikoitaisiin tarkoituksella puheliaita tai vähäsanaisia tutkimushenkilötä, tulokset väärityisivät.
- Jokaiseen mittaukseen liittyy virhe.
 - Täysin satunnainenkaan otos ei edusta täydellisesti koko väestöä. Joukkoon saattaa valikoitua puhtaasti sattumaltakin poikkeuksellisen puheliaita tai harvasanaisia naisia tai miehiä.
 - Millaisia sekoittavia tekijöitä tulee mieleen? Mitkä seikat voisivat vaikuttaa tutkittavaan asiaan?
 - Otoskoolla, eli sillä kuinka monta tutkimushenkilöä tutkitaan, on keskeinen rooli tutkimuksen luotettavuudelle. Mitä suurempi otos, sitä pienemmäksi sattuman osuus käy ja vastavasti mitä pienempi otos, sitä suurempi on yksittäisten sattumienvaikutus.
 - * Tilastolliset mallit turvautuvat todennäköisyyskuviin erottaakseen sattuman vaikutuksen: kun aineisto on kerätty, halutaan tietää kuinka todennäköistä on, että uskomus pitää paikkaansa.
- 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!

Are Women Really More Talkative Than Men?

Matthias R. Mehl^{1,*}, Simine Vazire², Nairán Ramírez-Esparza³, Richard B. Slatcher³, James W. Pennebaker³

+ Author Affiliations

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

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

Abstract

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

Kuva 2.2: Are women really more talkative than men?

2.4 Mitä on tutkimus?

- Tiede tavoittelee tietoa, mutta mistä?
- Jokaisen tutkimuksen lähtökohtana on (tai ainakin pitäisi useimmissa olla) tiedollisen uteliaisuuden, käytännön tarpeiden tai teorian kehittämispyrkimyksen herättämä ongelma, johon tutkimuksen avulla etsitään vastausta. Tutkimus yrittää käsittää sekä tulkitun ilmiön, että sen tajunnassa synnyttämät spontaanit mielikuvat tai arkipäivän tiedot.
- Tutkimus siis pyrkii löytämään täysin uutta tietoa, varmentamaan (mahd. aiempien tutkimusten myötä) syntyneitä vallitsevia mutta epävarmoja käsityksiä sekä tarkistamaan vakiintuneen tiedon paikkansapitävyyttä.
- Valtaosa tieteestä asemoituu erityisesti kahden viimeisen kohdan alaisuuteen vaikka tieteen popularisoinnissa (mm. median toimesta) usein keskitytäänkin uusiin tiedemaailmaa ja joskus "käytännön"

26LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

elämää järisyttäviin löydöksiin, jotka tosin voivat usein olla hyvin epävarmoja!

* Lisää tieteen popularisoinnista jaksossa [4.6](#).

- Millaisia kysymyksiä **tutkimuksessa** asetetaan (voidaan asettaa)?
 - **Kuvaus:** Kuinka suuri on yli 65-vuotiaiden osuus Suomen väestöstä?
 - **Riippuvuuden kuvaus:** Ovatko paljon mainostavat yritykset kannattavampia kuin vähän mainostavat?
 - Kuvattujen ilmiöiden **selittäminen ja ymmärtäminen.** Miksi vanhempien sosioekonominen asema vaikuttaa ekonomien työhönsijoitumiseen? Tämän tutkimuskysymyksen tapauksessa pyrkimys on lähtinä selittää (ymmärtää) ilmiötä.
 - **Ennustaminen:** Jos kansantulon kasvu pienenee x%, työttömyyden ennustetaan kasvavan y tuhannella.
 - Kohdetta kuvaavien käsitteiden ja teorioiden rakentaminen, teorioiden ansioiden ja puutteiden arviointi.
- Myöhemmin materiaalissa (luvussa [11](#)) keskustellaan vielä tarkemmin miten tilastotieteessä ilmiön ymmärtäminen (selittäminen) ja ennustaminen eroavat toisistaan.
- **Tutkimuksen rajat?** Onko niitä?
 - Tutkimus antaa aina vajavaisen kuvan tutkimuskohteesta.
 - * Kehittynytkin tieteellinen teoria tai malli on aina reaalimaailman yksinkertaistus: tutkimus on aina alisteinen käytetylle menetelmälle ja sen oletuksille!
 - Ymmärtämiseen tarvittava havaintomaailman hahmotus (saattaa) tuottaa ideologisesti ja historiallisesti sitoutuneita yksinkertaistavia sekä luonteeltaan usein hyvin teoreettisia abstraktioita.
 - * Alakohtainen substanssitetous sekä sen vahvuksien ja puutteiden sekä historiallisen ja ideologisen kontekstin tiedostaminen on ensiarvoisen tärkeää kaikessa tutkimuksessa!
 - Joka tapauksessa täyneen neutraaliuteen ja objektiivisuuteen on mahdotonta päästä. Tästä huolimatta on hyvä ja tärkeää pystyä tunnistamaan tämä haaste.
 - Tutkimusta voi tehdä joistakin arvolähtökohdista, mutta sen tulisi olla näkyvää. Omien arvojen mahdollisimman selvä eksplikointi on yksi keino, jolla voi yrittää vähentää piiloarvojen vaikutusta tutkimukseen.
 - * Arvot ilmenevät esimerkiksi tutkimuksessa käytetyissä käsitteisissä, jotka harvoin ovat arvovapaita. Useimmat käsitteet voidaan

korvata toisilla, joilla on paikoin hyvin erilainen arvosisältö joskin arvottava lataus saattaa myös olla paikoin tarkoituksellista! Joka tapauksessa arvopainotteisten valintojen tunnistaminen on vaikeaa.

- * Toisaalta arvoihin sitoutuminen on väistämätöntä, sillä se on sosiaalisen olemassaolon sivutuote. Yhteiskunnan jäseninä meillä on tuskia mahdollisuksia (täydellisesti) irroittautua arvoistamme kun pyrimme esim. ammatillisii päämääriin.
- Myös päinvastainen ongelma olemassa: Tutkimusta arvioitaan siihen perustellusti tai perusteettomasti kiinnitettyjen arvonäkökohtien mukaan!

- Tutkimukseen kuuluu olennaisesti myös oman tutkimustyön kuvaaminen, ts. kertomus siitä, miten esitetyihin tuloksiin on päästy.
 - Tämän myötä tieteelliselle ajattelulle on ominaista automaattinen **itsensä korjaaminen**.
 - Tutkimuskysymys, valitut menetelmät, käytetty aineisto ja tehdyt johtopäätökset perataan auki tutkimusartikelissa/raportissa, joka sitten lähetetään **vertaisarvioitavaksi** tietelliseen julkaisuun, jossa muut alan asiantuntijat arvioivat sen ja päätävät hyväksytäänkö se julkistavaksi.
- **Vertaisarvioinnissa** yksi tai useampi, tehdystä tutkimuksesta riippumaton, saman alan tutkija lukee ja tarkastaa tehdyn tutkimusartikkelin, arvioi sitä ja suosittaa tietellisen julkaisun arvioinnista vastaavalle päätoimittajalle (editorille) kyseisen artikkelin hyväksymistä tai hylkäämistä.
 - Vertaisarvointi ei aina takaa sitä, että julkaistu tutkimus olisi virheetön ja erinomaisesti tehty, vaan myös väärää tietoa pääsee välillä vertaisarvointiprosessin läpi.
 - Tämä ei kuitenkaan poista tieteellisen prosessin luotettavuutta, sillä uusi tieto varmentuu vasta usean samaa tutkimuskysymystä tutkineen ja vastaavat tulokset saaneen tutkimukseen myötä. Toisin sanoen, tieteellisen prosessin voidaan ajatella konvergoituvan totuuteen, vaikka yksittäisiä virhearvointeja sattuisikin.
- **Tutkimuksen kieli**
 - Tutkimus edellyttää arkikieltä täsmällisempää kommunikaatiota.
 - Ongelmaan liittyvien käsitteiden huolellinen määritteleminen ja erityyli on tarpeellista.

28LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- * Käsitteiden ja eri aloilla, osin samoistaasioista käytettävien, toisistaan eroavien termien systemaattinen määrittely ja jäsen-tely selkeyttää tiedeyhteisön välistä kommunikointia.
- * Eivät korvaa empiiristä tietoa vaan vaikuttavat tiedon järjestymiseen ja sen perusteella tehtäviin päätelmiin.

Esimerkki: Luonnontieteelliset vs. yhteiskunnalliset sovellukset:

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

2.5 Tutkimuksen vaiheet ja tulosten julkaiseminen

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

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

Vastuullisen tieteen

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

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

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

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

30LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

Luku 3

Tilastotiede tieteenalana

Tässä luvussa hahmottelemme tilastotieteen piirteitä tieteenalana. Käymme läpi tilastotieteelle ominaisia piirteitä, jotka erottavat sen niin lähitieteistä, kuten matematiikasta ja tietojenkäsittelytieteestä, kuin myös sovellusaloista. Usein näkee tilastotieteen typistettävän vain työkaluksi eri sovellusalojen empiriseen tutkimukseen siitäkin huolimatta että tilastotieteellä on oma rikas teoriapohjansa sekä kiistaton asema omana tieteenalanaan.

Tieteenalan määritteleminen lyhyesti on aina hieman hankala. Tästä huolimatta seuraavassa yritymme osaltaan vastata seuraaviin kysymyksiin:

- Mitä tilastotiede on ja mitä se ei ole? Miksi tilastotiede ei ole vain soveltuu matematiikkaa tai matematiikalla höystettyä tietojenkäsittelyä?
- Mihin tilastotiedettä käytetään? Onko tilastotieteellä käyttöä ns. "akateemian" eli tutkimusyhteisön ulkopuolella?
- Minkälaisista on tyypillinen tilastotiedettä kohtaan esitetty kriitikki?

3.1 Lisää tilastotieteen perustermejä

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

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

- Esimerkiksi jos tutkitaan kuntavaaleissa äänestävien tuloja niin jokainen äänestysikäinen muodostaa oman tilastoyksikkönsä (ks. alalla) ja täten populaationa (perusjoukkona) toimii kaikki äänestysikäiset kansalaiset. Jos taas tutkitaan äänestysaktiivisuutta eri kunnissa, muodostaa jokainen kunta oman tilastoyksikkönsä ja kaikki Suomen kunnat muodostavat populaation.

Populaatio

Konkreettinen tai hypoteettinen tutkimuskohteiden joukko, joka koostuu kaikista tilastoyksiköistä

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

Tilastoyksikkö ja tilastollinen muuttuja

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

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

Havainto

Havainto muodostuu tilastoyksikön tarkasteltavien tilastollisten muuttujien havaitusta arvoista.

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

Havaintoaineisto/data

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

Tiivistettynä:

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

3.2 Mitä tilastotiede on ja mitä se ei ole?

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

¹Leo Törnqvistin, Suomen ensimmäisen tilastotieteen professorin, esittämä luonnehdinta (Vartia, 1989).

²Mellin (2005).

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

⁴(A.C. Davison)

⁵(Sund, 2003)

- Tilastotiede siis **kehittää ja soveltaa menetelmiä ja (tilastollisia) malluja**, joiden avulla reaalimaailman ilmiöistä voidaan tehdä johtopäätöksiä ilmiötä kuvaavien numeeristen tai kvantitatiivisten tietojen perusteella tilanteissa, joissa tietoihin liittyy **epävarmuutta ja satunnaisuutta**.
 - Tilastollisten menetelmien avulla pyritään löytämään reaalimaailman satunnaisia ilmiötä kuvaavista numeerisista (eli kvantitatiivisista) tiedoista **systemaattisia piirteitä** joita jalostetaan sellaiseen muotoon, että ilmiöstä voidaan tehdä päätelmiä.
 - * Vrt. signaalin ja kohinan erottaminen (ks. Silver, 2014)⁶.
 - Tilastolliset mallit perustuvat todennäköisyyslaskentaan ja niillä mallinnetaan reaalielämän ilmiöiden alla piileviä prosesseja tai mekanismeja. Näiden prosessien tuottamia tietoja (aineistoja) tiivistetään usein graafisiksi esityksiksi ja tunnusluvuiksi sekä tilastollisten mallien parametreiksi, joiden pohjalta johtopäätöksiä tehdään.
 - Tässä onnistuakseen tilastollisten menetelmien tuleekin pyrkii erottamaan **sattuma ja systemaattisuus** tarkasteltavissa ilmiöissä tai, tarkemmin, niitä kuvaavissa aineistoissa, jotta johtopäätökset olisivat luotettavia.

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

Mitä tilastotiede ei ole

- **Tilastotiede ei ole vain tilastojen tuotantoa**
 - Vaikka sana **tilasto** tuo useimmiten ensimmäisenä mieleen yhteiskuntaa ja sen toimintaa kuvaavat **numeeristen tietojen järjestelmäliset kokonaisuudet**, tilastotiede ei suinkaan ole ainoastaan tilastojen ja niiden tekemisen oppia.
 - * Tämä siitätäkin huolimatta, että niiden menetelmien konstruointi, joilla näitä tilastoja tuotetaan, jalostetaan ja analysoidaan on keskeinen osa tilastotiedettä. Tilastot ovat siis usein tilastotieteen soveltajan tutkimuskohteena ja tilastojen laadinnassa käytetään apuna tilastotieteen menetelmiä.
 - * Suomessa **Tilastokeskus** toimii virallisena tilastoviranomaisena ja tilastotuottajana. Tätä **tilastotuotannon** kokonaisuutta nimitetään ajoittain **tilastotoimeksi**. **Tilastotieteen käytöalue on paljon tätä laajempi**.

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

* Terminologiaa:

- Tilastoala = Tilastotiede + Tilastotoimi
- Tilastotiede = Teoreettinen tilastotiede + Soveltava tilastotiede
- Tilastotoimi = Tilastojen tuotanto + Tilastojen hyödyntäminen

• Tilastotieteen kannalta mikä tahansa reaalimaailman ilmiötä kuvaava **numeristen tai kvantitatiivisten tietojen järjestelmällinen kokoelma** voi muodostaa **tilastollisen aineiston** ja siten tilastollisen tutkimuksen mahdollisen kohteen.

– Esimerkiksi kaikki **empiriisen** tai **kvantitatiivisen** tutkimuksen tutkimus- tai havaintoaineistot ovat tilastotieteen kannalta tilastollisia aineistoja.

• Tilastotiede sijoittuu tieteiden kentässä matematiikan, filosofian ja tietojenkäsittelytieteen rinnalle. Tästä huolimatta se ei kuitenkaan ole yksiselitteisesti minkään näiden osa-alue.

– **Tilastotiede ei ole matematiikan osa-alue**, sillä tilastotiede lähestyy tieteellistä ongelmanratkaisua eri tavoin:

* Matematiikka on tietyllä tavalla aina eksaktia ja sen tulokset perustuvat formaaliin deduktioon ja loogisiin todistuksiin, johtuen usein ”eksaktiin” ratkaisuun tai matemaattisesti formaaliin ratkaisun loogiseen esitystapaan.

* Tilastotiede sen sijaan on aina konteksti- ja aineistopohjaista ja perustuu induktiiviseen päättelyyn. Saadut tulokset ovat aina epävarmoja - koska ne kuvailevat epävarmaa tietoa generoivia prosesseja!

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

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

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

- Vaikka nämä ja jotkin muut alat jakavat tilastotieteen kanssa useita piireitä ja ominaisuuksia, on tilastotiede kuitenkin siis perustellusti oma tieteenalansa. Tämä erottelun vaikeus jo itsessään todistaa kuinka keskeinen rooli tilastotieteellä on eri aloilla!
 - Tilastotiede ei siis kuulu yksiselitteisesti sen lähitieteiden alle, vaan muodostaa oman tieteenalan omine teorioineen ja tieteellisine premissineen. Käsittelemme myöhemmin tilastotieteen roolia matematiikan ja/tai datatieteiden (“data science”) kokonaisuudessa ja keskustelemme tarkemmin näiden erojen luonteesta.

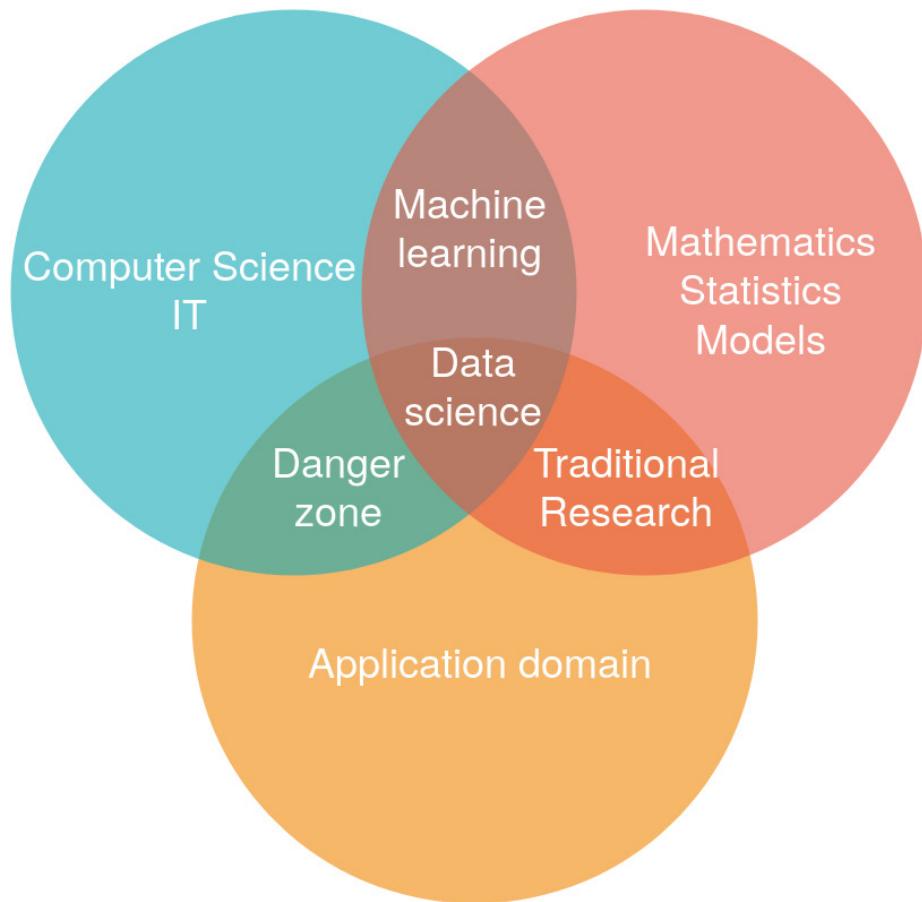
Mitä tilastotiede (ainakin) on

- **Tilastotiede yleisenä menetelmätieteenä**
 - Tieteellistä tietoa ympäröivästä maailmasta hankitaan tieteellisillä **menetelmillä/metodeilla** (Ks. tieteellisen menetelmän kriteerit luku 2)), joiden avulla tutkitaan jotain ilmiötä tai sen generoimaa kvantitatiivista mutta epävarmaa tietoa sisältävää aineistoa.
 - Tilastotieteessä kehitetyt ja kehitettävät menetelmät antavat tutki joille yhtenevät ja tiedeyhteisön hyväksymät raamatit, 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 ainakin alakohtaista. Tästä huolimatta tilastollisia menetelmiä sovelletaan aina johonkin **aineistoon!**
 - Tilastotieteen nähdäänkin usein kuuluvan ns. **menetelmätieteisiin**, joissa mm.:
 - * Kehitetään työkaluja muiden tieteiden tutkimusongelmien ratkaisuksi
 - * On myös oma sovelluksista vapaa teorianmuodostuksensa
 - Menetelmäkehityksen näkökulma tilastotieteeseen: *tilastotiede kehitää matemaattisia malleja satunnaisilmiöitä kuvaavia kvantitatiivisia tietoja generoiville prosesseille*. Koska tietoihin liittyy **epävarmuutta** tai **satunnaisuutta**, **tilastolliset mallit** perustuvat **todennäköisyyslaskentaan**.
 - * Juuri sattuman ja epävarmuuden huomioiminen tutkimusasetelmissa erottaa tilastotieteen muista menetelmätieteistä!

- Tilastollisia menetelmiä voidaan soveltaa tietojen keruun, jalostuksen ja analysoinnin jokaisessa vaiheessa. Päämäääränä on jalostaa tiedot muotoon, joka mahdollistaa tutkittavaa reaalimaailman ilmiötä koskevien joh-topäätösten tekemisen käytettyjen menetelmien pohjalta, eli ns. **tilastollisen päätelyn**.
 - Tutkimuksessa on pystyttävä valitsemaan ja käyttämään menetelmiä, jotka antavat aineistosta vastauksia haluttuihin kysymyksiin. Tämä vaatii yhtä lailla sovellusalakohtaista osaamista (ns. substanssiosaamista) kuin myös kattavaa menetelmäosaamista.
- Tilastotieteessä lähtökohtana ja ratkaisevassa asemassa on siis aina jokin satunnaisilmiön generoima **aineisto**, josta haluamme oppia tai tietää lisää, kenties voidaksemme tehdä suuria yhteiskunnallisia päätöksiä sen pohjalta!
 - Tämä aineistokeskeisyys yhtäältä erottaa tilastotieteen rajatieteistään ja toisaalta tuo sen lähemmäksi niitä ja sovellusalojaan.
 - Aineistoa analysoidaan, kuvallaan ja mallinnetaan tilastollisin menetelmin, joiden kehittäminen on keskeinen osa tilastotiedettä.
 - Pelkkä menetelmien kehittäminen kuuluu pitkälti matemaattisen/-teoreettisen tilastotieteen osa-alueelle.
 - Pelkkä ainestoon 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 harrastajista asettuvat näiden äärimmäisten luonnehdintojen välimaastoon eikä tarkkaa luokittelua ole sinänsä tarpeen tehdä ja korostaa.
 - Joka tapauksessa tilastotieteen kehityksen keskiössä ovat aina sovellusalakohtaiset ongelmat, joista useat palautuvat yleisemmälle tasolle teoreettisen tilastotieteen kehityspolkuihin.

3.3 Tilastotieteen suhde lähitieteisiin

- Kuvio 3.1⁷ tarjoaa karkean yleistyksen tietojenkäsittelytieteen (Computer Science) ja sovellusalan (Application domain) sekä tilastotieteen (Statistics) ja matematiikan (Mathematics) välisistä yhteyksistä. On selvää että tilastotieteellä on paljon päälekäisyksiä lähitieteiden kanssa ja joskus näkeekin (huolimatta edellä tehdystä huomioista) että tilastotiede nipputaan yhteen matematiikan tai tietojenkäsittelytieteen kanssa.



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

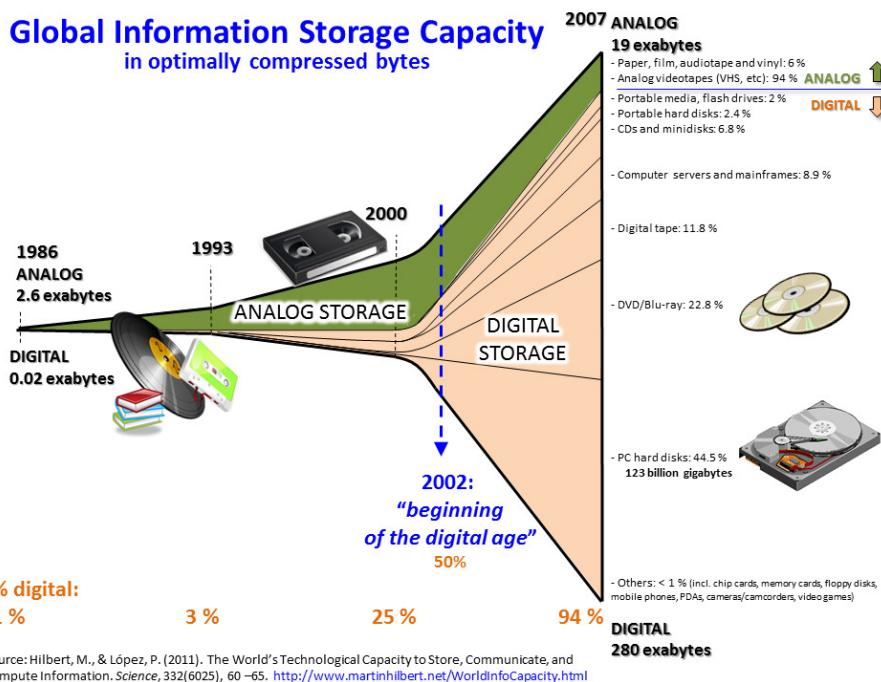
- Yritetään siis vielä hahmotella tilastotieteen suhdetta sitä lähimpänä olevaan (soveltavaan) matematiikkaan.

⁷Kuvan lähde: [Duchesnay \(2020\)](#)

- Tilastotieteessä olennaisen otantateorian (Luku 5) voisi ajatella olevan matemaattisesti määritelty teoria, jossa myös on aineiston käsite, mutta se ei tee siitä vielä varsinaisesti tilastotiedettä.
- Matematiikassa kuvataan ongelma ja esitetään se teorian muodossa, eli malli on “*parametreista havaintoihin*”.
- Tilastotieteessä ongelma on käänneinen, edetään “*havainnoista parametreihin*”, mutta ongelman matemaattinen kuvaus vaaditaan ensin.
- Tilastotiede esittää menetelmiä ja käsittää tämän käänneisen ongelman ratkaisemiseen.
 - * Karkeasti erotellen tilastotieteessä käsitteltävät ongelmat lähtevät aina havainnoista eli aineistosta ja matematiikassa suunta on teoriasta aineistoon.
 - * Voidaan siis sanoa, että tilastotieteen erottaa puhtaasta matematiikasta se, että siinä tutkitaan metodeja, jotka mahdollistavat päättelyn/tiedon hankinnan puutteellisesta tai epävarmasta tiedosta.
- Ilmiöiden kuvaamiseen ja käyttäytymisen ennakoimiseen käytetään usein **mallia**. Mallit (matemaattiset/tilastolliset mallit) voidaan jakaa **deterministisiin** ja **stokastisiin** malleihin.
 - Deterministisen mallin tapauksessa, tiettyjen alkuehtojen (alkuarvojen) vallitessa voidaan määrittää tarkaltevan 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 esim. rahanheitto tai säännostaminen.
 - Kun joitain ilmiöitä 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 muulta osin.
- Tilastotiede ei myöskään ole puhtaasti tietojenkäsittelyä, vaikka tilastotiede onkin luonteeltaan aineistopohjaista ja aineistojen sisältämää tietoa on käsittely 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) johon voidaan katsoa lukeutuvan mm.
 - Tilastotiede ja matematiikka
 - * Erityisesti tilastollinen data-analytiikka ja satunnaisen aineiston mallintaminen sekä soveltuват soveltavan matematiikan osa-alueet.
 - Tietojenkäsittely
 - * Tietoteknologian kehityksen myötä taitavien tietojenkäsittelijöiden kysyntä on kasvanut merkittävästi. Lähes jokaisella alalla kerätään entistä enemmän dataa lähes kaikesta, jonka pitäisi osata myös käsitellä sitä!
 - * Datatieteen voidaankin osaltaan katsoa syntyneen tästä elinkeinoelämän tarpeesta asiantuntijoille, jotka osaavat käsitellä suuria tietoaineistoja (dataa) sekä mallintaa niitä hyödyllisellä tavalla.
 - Sovellusala
 - * Datatiede on luonteeltaan pääosin soveltavaa ja sen alaan lukeutuvia menetelmiä sovelletaan aina johonkin tosielämän ongelmaan. Tästä syystä nk. substanssiosaaminen sovellusalalta on datatieteilijälle erityisen tärkeää ja nykypäivänä datatieteilijän rooli onkin pirstaloitunut yhä enemmän eri sovellusalojen datatieteisiin.
 - * Tästä huolimatta datatieteilijöiden käyttämät mallinnusmenetelmät ovat usein varsin samanlaisia, sillä ne pohjautuvat edelleen tilastotieteen ja matematiikan teoriapohjaan. Ilman jälkimmäisten riittävää osaamista, liikutaan datatieteen osalta vaarallisilla vesillä! (Ks. oheinen kuva ja keskustelu alla).
 - Datatieteellä ei usein nähdä olevan omaa historiallisen tieteellisen prosessin luomaa teoriapohjaa vaan sen voidaan katsoa olevan kokoelma eri

alojen tieteellisiä menetelmiä ja tuloksia, jotka voidaan yhdistää tavalla, jonka ”datavallankumous” (ks. kuva 3.2) mahdollistaa ja jotka ovat keskeisessä roolissa dataintensiivisissä sovellutuksissa.



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

- “Danger zone”
 - Kuvan 3.1 “danger zone” (Duchesnay, 2020) kuvaa tilannetta, jossa ilmiöiden/mallien tilastotieteellinen perusta unohdetaan.
 - Tilastotieteen näkökulman ohittava (laiminlyövä) soveltaja ei aina kykene suhtautumaan kriittisesti muodostuvaa ennustemallia, tai ennustetulosta, kohtaan eikä täten päädy parhaisiin mahdollisiin (tarkeimpiin) ennustetuloksiin tilanteessa, jossa jokin toinen malli kuvaisi ilmiötä annettua mallia paremmin.
 - Ko. soveltaja ottaa mallin sekä sen antaman ennustetuloksen annettuna, eikä mietti *mistä kyseinen ennustetulos johtuu*. Jotta tarkat ennustetulokset toteutuvat jatkossakin (kun uutta aineistoa, dataa, tulee saataville), on ennustajan oleellista huomioida mitkä tekijät johtivat tarkkaan ennustulokseen.

- Eri menetelmät sopivat eri sovelluskohteisiin. Tilastotieteilijä osaa useimmiten tunnistaa eri sovelluskohteisiin sopivat menetelmät parremmin kuin tietojenkäsittelijä. Vastaavasti tehokkaan/onnistuneen ohjelmointikoodin kirjoittamisessa tilanne on usein toisinpäin.

3.4 Tilastotieteen osa-alueet

- Tilastotiede on saanut alkunsa siitä, että yhteiskunnan modernisoitussa on tarvittu yhä enemmän tietoja erilaisiin hallinnollisiin tarpeisiin. Samalla on syntynyt tarve kehittää menetelmiä joiden avulla tilastojen luotettavuutta on voitu parantaa.
 - Kehitys oli pitkään ns. ongelmasta menetelmään ja tutkimusalojen erilaisudesta johtuen myös tilastotiede on kehittynyt vastaamaan monipuolisesti erilaisiin menetelmällisiin ongelmuihin!
 - Tämä on johtanut osaltaan siihen, että tilastotiede jakautuu moniin osa-alueisiin. Osa-alueita on niin paljon, että alan huiputkaan eivät voi hallita niitä kaikkia!
- Tästä huolimatta tilastotiede voidaan karkeasti jakaa teoreettiseen ja soveltavaan osa-alueeseen, jotka toimivat alituisessa vuoropuhelussa.

Soveltava tilastotiede

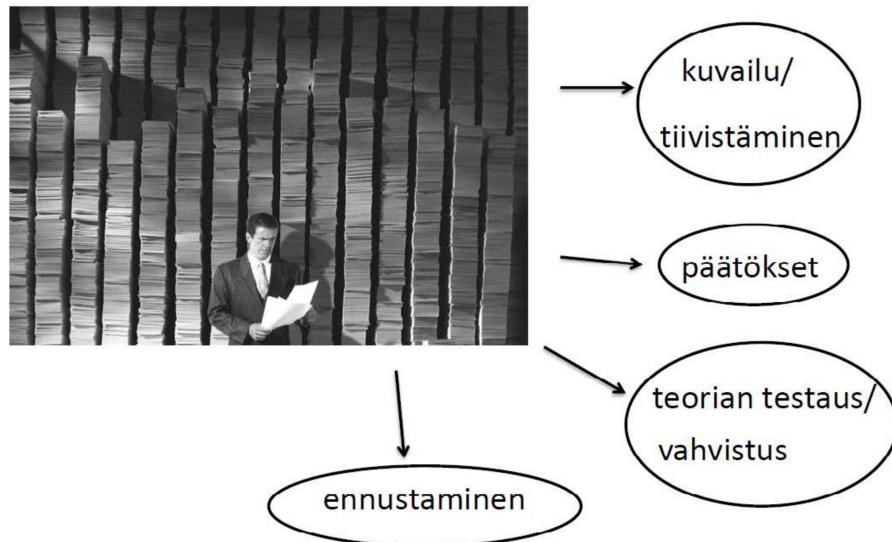
Soveltava tilastotiede

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

- Yleisesti ottaen eri tieteenoilla kohdattavat menetelmäsuuntaukset voidaan jakaa kahteen luokkaan tutkimusaineistojen tyypin perusteella:
 - **Kvantitatiivinen:** eli määrällinen tutkimus on tutkimusta, jossa tutkimusongelma on muotoiltu tarkasti etukäteen ja tutkimuskysymyksiin vastataan käyttäen tilastollisia menetelmiä pyrkien **selittämään ja ennustamaan** tutkimuksen kohteena olevaa ilmiötä.
 - * Täsmällisten ja laskennallisten tilastollisten menetelmien käytäminen numeeriseen aineistoon on kvantitatiiviselle tutkimukselle ominaisin piirre.
 - * Perustuu yleensä satunnaisotokseen (kts. luvut 4, 5 ja 6) ja tutkimusaineisto on tiivistetty numeeriseksi havaintomatriisiksi, jolle oleellinen vaatimus on sen totuudellisuus.

- * **Kritiikki:** määrällinen tutkimus on (paikoin) sokeaa tutkittavien ilmiöiden sellaiselle luonteeelle, 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:** eli laadullinen tutkimus on tutkimusta, jossa tutkimuksen kohteena olevaa ilmiötä ja sen merkitystä sekä tarkoitusta pyritään **ymmärtämään** kokonaivaltaisella tavalla.
 - * Laadullisessa tutkimuksessa annetaan usein tilaa tutkimuksen kohteena olevien ilmiöiden ja/tai ihmisten näkökulmille, vaikuttimille, kokemuksille ja tuntemuksille. Tutkimusyksikköjen otanta on täten usein harkinnanvaraista.
 - * Laadullisessa tutkimuksessa tutkimusongelma muotoutuu tutkimuksen edetessä ja sillä tyypillistä on hypoteesittomuus, eli tutkimus on tarkoitus aloittaa mahdollisimman vähin ennakkoletuksin. 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 (numeraalisia) 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 pitääsi) soveltaa kaikilla reaalimaailmaa tutkivilla tieteinaloilla, joiden tutkimusaineistot voidaan esittää kvantitatiivisessa muodossa.**
 - Tilastollisten menetelmien käyttö on siis huomattavan paljon yleisempää määrällisessä kuin laadullisessa tutkimuksessa.
 - Menetelmien soveltamisen tarkoituksena on (voi olla): **i) kuvalla ja tiivistää tietoa**, jota havaittu aineisto sisältää **ii) sovellusalan oman teorian empiirinen testaus** tai **iii) edellisten pohjalta tehtävä tilastollinen päätteily**.

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



Kuva 3.3: Soveltava tilastotiede

Teoreettinen tilastotiede

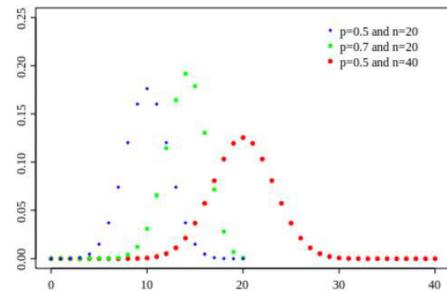
Teoreettinen tilastotiede kehittää (tilasto)matemaattisia malleja kuvaamaan satunnaisilmiötä- ja prosesseja, jotka generoivat reaalimaailman ilmiötä kuvaavia numeerisia tai kvantitatiivisia tietoja, joihin liittyy epävarmuutta ja satunnaisuutta.

- Teoreettinen tilastotiede luo pohjan tilastollisten menetelmien ymmärtämiselle, soveltamiselle ja kehittämiselle.
 - Ilman riittävää ymmärrystä tilastollisten menetelmien toimintaperiaatteista niiden soveltaja on vaarassa tehdä virhepäätelmiä! (Ks. alaluku 3.5 tilastotieteen kriiikistä)
- Mallit perustuvat todennäköisyyslaskentaan, ja niitä kutsutaan tilastollisiksi malleiksi, stokastisiksi malleiksi tai todennäköisyysmalleiksi.
 - Tilastolliset mallit perustuvat laajalti niin kutsuttuun uskottavuusfunktioon. Se on malli, joka riippuu havaintoaineiston lisäksi yhdestä tai useammasta parametrista. (ks. tarkemmin luku 6)
 - Uskottavuusfunktion arvo kertoo kuinka todennäköisenä voidaan havaittaa aineistoa pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvoilla.
 - Uskottavuuspäätelyn perusajatuksena on, että se tai ne parametriarvot, joilla uskottavuusfunktion arvo maksimoituu kuvaan aineiston generoinutta prosessia parhaiten.
 - Aineistoa koskevia hypoteeseja voidaan testata käyttäen uskottavuusfunktion maksimia vastaavaa tilastollista mallia!
 - “*Kaikki mallit ovat vääräitä, mutta jotkut ovat käytökkäisiä.*” (Box, 1976).
- Uskottavuusfunktiot perustuvat aina satunnaisilmiöiden mahdollisia arvoja kuvaaviin nk. **tiheysfunktioihin** tai **pistetodennäköisyysfunktioihin**.
 - Tiheysfunktiot kuvaavat jonkin satunnaismuuttujan (satunnaisilmiön) saamien arvojen jakaumaa.
 - Esimerkiksi kolikonheitto on satunnaisilmiö ja sillä on vain kaksi arvoa⁸ ja kolikonheittoa voidaan kuvata nk. binomijakaumalla, jossa merkitään $\text{Bin}(n, p)$ missä n on heittojen lukumäärä ja p on kruunan todennäköisyys.

⁸Kolikon kantilleen jäämistä ei tässä lasketa mahdolliseksi tapahtumaksi.

- Esimerkki: heitetään kolikkoa 40 kertaa ja saadaan kruuna 40/40 tapauksessa. Onko tämän havaintoaineiston perusteella uskottavaa, että kolikonheitto noudattaa binomijakaumaa $\text{Bin}(40, 0.5)$? Eli kuinka uskottavan voidaan pitää että kyseinen kolikko on tavallinen, painotamaton kolikko?

Tilastotiede perustuu uskottavuksiin, jotka taas perustuvat todennäköisyyteen ja tiheysfunktioihin.



Kuva 3.4: Tilastotiede ja todennäköisyys

- Todennäköisyysslaskenta luo tilastotieteelliselle epävarmuuden mallintamiselle vahvan ja uskottavan matemaattisen perustan.
 - Todennäköisyysslaskentaa opetetaan tarkemmin (tätä kurssia seuraavilla) kursseilla [TILM3553 Todennäköisyysslaskennan peruskurssi pääaineopiskelijoille](#), [TILM3568 Todennäköisyysslaskenta sivuaineopiskelijoille](#) ja [SMAT5306 Todennäköisyysslaskennan jatkokurssi](#).

3.5 Tilastotieteen kritiikkiä

- Tilastotieteen rooli tiedeyhteisössä on niin tärkeä että sitä kohtaan on ymmärrettävästi esitetty myös paljon kritiikkiä. Valtaosa kritiikistä kohdistuu joko tilastotieteen matemaattisuteen tai sitten siinä tarvittaviin oletuksiin, jotka mahdollistavat esimerkiksi hypoteesien testaamisen.

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

Pohja tilastollisten menetelmien ymmärtämiselle, soveltamiselle ja kehittämiselle

Kuva 3.5: Teoreettinen tilastotiede

- Usein kritiikki on aiheetonta ja johtuu sen esittäjän puutteellisesta tilastotieteen ymmärryksestä. Perusteettoman kritiikin esittäminen toista tieteenalaa kohtaa ei kuitenkaan ole vieraas ilmiö juuri millään alalla.
- Tässä alaluvussa käymme läpi yleisimpiä kritiikin muotoja, joita tilastotiedettä kohtaan esitetään ja pyrimme tarjoamaan vastauksia/vastineita silloin kun niitä voidaan antaa.

“Yleismaailmallinen” kritiikki

- Aloitetaan yleismaailmallisella kritiikkillä, jota tilastollista tutkimusta vastaan on esitetty:
 - Tilastotieteessä käytettävien tunnuslukujen, kuten keskiarvon, reaalimaailman vastineet ovat joskus mielivaltaisia. Esimerkiksi keskiarvo on ajoittain ongelmallinen tunnusluku, sillä lienee varsin selvää, että keskimääräistä ihmistä ei ole olemassa vaikka tilastotieteessä näitä tunnuslukuja usein lasketaankin.
 - * Esimerkiksi puhekielessä yleinen nk. “Keskiarvo-Kalle”, eli 1,8 lapsen vanhempi ja 1,5 auton omistaja on tietenkin täysin kuvitteellinen.

- * Lisäksi joskus kuulee tilastotieteilijöitä kritisoidavan lausumalla *“Jos toinen jalka on jääkylmässä vedessä ja toinen kiehuvassa vedessä, niin tilastotieteilijän mielestä ihmisellä on tällöin keskimäärin hyvä olla”*
- Korrelaatio on tunnusluku, joka kuvaa kahden muuttujan välistä riippuvutta (palaamme tähän tarkemmin luvussa 7). Se ei kuitenkaan kuvaa millään tavoin kausaalisuutta, eli sitä kumpi aiheuttaa kumman, jos kumpikaan.⁹
 - Esimerkiksi “jäätelön syönti ja hukkumiskuolemat” -tapauksessa ha-vainnollisesti todetaan jäätelönkulutuksen ja hukkumiskuolemien lu-kumääärän korreloivan keskenään, mutta taustalla vaikuttava tekijä onkin lämmmin kesä, joka vaikuttaa molempia.
- Vaikkei näiden esimerkkien oikeellisuutta ole syytä kiistää, niin tilastollisen tiedon arvioinnissa on kuitenkin syytä päästää syvemmälle.

Kritiikki matemaattisuutta kohtaan

- Ehkä merkittävin kritiikki tilastollisia menetelmiä kohtaan kohdistuu kriitikin näkökulmasta perusteettomaan, tai ainakin liian vahvaan, matemaattisuuden tuomaan itsevarmuuteen. Voidaankin siis perustellusti ky-syä, että **onko tieteellisyys = matemaattisuus?**
 - Useat tieteenalat käyttävät tutkimuksessaan edistyneitäkin tilastollisia menetelmiä siitä huolimatta, että tutkijoiden tilastomatemati-tinen pohjakoulutus ei välittämättä ole riittävällä tasolla kyseisten menetelmien kokonaisvaltaiseen ymmärtämiseen.
 - * Helppokäytöisistä tilasto-ohjelmistoista on riittävät perustaidot omaaville käyttäjille erittäin paljon hyötyä mutta koneiden ja oh-jelmien käytön opettelu ei kuitenkaan ole varsinaista tilastotie-dettä (tarvitaan enemmän tilastotieteen opintoja).
 - * Laskentatehon ja modernin tietojenkäsittelyteknologian ansiosta monimutkaisiakin tilastollisia analyysejä on kuitenkin mahdolli-s-ta tehdä vaikka tutkijalla olisi tilastotieteestä vain perustiedot, jos sitäkään.
 - * Pahimillaan tämä saattaa johtaa siihen, että analyyseja teh-däään ymmärtämättä mitä itse asiassa ollaan tekemässä.
 - Tilastollisten analyysien hyödyllisyden ja järkevyyden ehtona on kuitenkin käytettävien menetelmien, aineiston ja tutkittavan ilmiön pintaa syvemmälle ulottuva tuntemus.

⁹Tyler Vigen on kerännyt [verkkosivuilleen](#) (ks. linkki) mitä moninaisimpia esimerkkejä kahdenvälistä nk. *näennäisistä* korrelaatioista.

- * Käytettävien tilastollisten menetelmien oletukset on osattava ottaa huomioon ja toisaalta odottamattomien tulosten syyt on pysyttää jäljittämään.
 - Teknistä esitystä käyttää tutkijaa saatetaan pitää erityisen uskottavana, koska hän kykenee käyttämään vaikeita menetelmiä. Tästä huolimatta tutkimusongelma ei saisi päästää unohtumaan.
 - Tutkijan tulisikin varmistua siitä, että käytettäväät 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 välittämättä neutraaleja!
- * Kaikkia tieteentekijöitä ei voida velvoittaa opiskelemaan edistynytä abstraktia tilastotieteen teoriaa (tilastomatematiikkaa), mutta menetelmien oikeaoppinen käyttö kuitenkin vaatii riittävää ymmärrystä.

Kriitikki yksinkertaistuksia kohtaan

- Edellisiä kohtia yleisemmin tilastotiedettä on kritisoidu siitä, että se ei kykene riittävällä tasolla huomioimaan reaalimailman kompleksisuutta.
 - Merkittävässä osassa tilastollisia analyyseja lähtökohtana on usko “todellisen” maailman ja näin ollen aineistoa generoivien mekanismien olemassaoloon.
 - * Tätä saatetaan usein pitää kuitenkin kyseenalaisena: voiko “to-sielämän stokastiikasta” muka todella löytyä säännönmukaisuuksia?
 - * Tämä kysymys on kuitenkin pitkälti tieteenfilosofinen ja palautuu lopulta sovellusalaan sekä tutkimusongelmaan ja -kysymykseen: tilastollisten menetelmien toimivuutta voidaan helposti testata esimerkiksi simulaatiokokeilla.
 - Tilastotiedettä on myös kritisoidu sen “sokeudesta” sosiaaliseen vuorovaikutukseen liittyviin subjektiivisiin kokemuksiin kuten tunteisiin, kokemuksiin ja ei-numerieiisiin havaintoihin.

- * Tämä kritiikki ei kuitenkaan suoranaisesti ole tilastotieteen kriitiikkiä, vaan jälleen sovellusalakohtainen ja erityisesti tutkimuskysymyksen asettelua koskeva ongelma.
 - Tuntemuksia ja kokemuksia voidaan hyvin testata tilastollisen menetelman, mikäli tutkija osaa uskottavasti määritellä niille numeerisen mittauksen kriteeristöt!
 - Tämä on kuitenkin vaikeaa, sillä aivan kaikkea ei voida kvantifioida: kirjoitetun tekstin tai sosiaalisten merkitysten tulkinna 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 saada relevanttia tietoa, tulisi aineiston analyysin apuna joka tapauksessa aina käyttää tilastollisia menetelmiä!
 - Vaikka kvantitatiivisia aineistoja ei voi pitää objektiivisina faktoina asioiden tilasta, se ei tarkoita, etteivätkö tulokset voisi olla käytökkelpoisia.

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 menetelmäkokoelmanksi*, valla 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 kaikkissa soveltavan tutkimuksen vaiheissa, jotta tutkimusongelmaan kyetään vastaamaan eikä turhaa työtä tule tehdynksi.
 - Karkeasti luokitellen tilastotieteilijät kehittävät menetelmiä, joita soveltajat käyttävät.

- * Soveltavia tilastotieteilijöitä löytyy kuitenkin yhä kiihtyvissä määrin! Erityisesti eri rajatieteiden alueilla, kuten alaluvussa **3.6** lyhyesti esitellään.

Tilastotieteen väärinkäytö

- Tilastotiedettä on myös mahdollista käyttää väärin monin eri tavoin, joka edelleen altistaa koko tieteenalan (perusteettomalle) kritiikille!
 - Tilastoja ja tilastotiedettä käytetään paljon väärin, mutta tämä on usein tahatonta (esim. puutteellisesta koulutuksesta johtuvaa).
 - * Joskus kuitenkin näkee tarkoituksellista tilastojen vääristelyä tai tahallista tilastollisten menetelmien väärinkäytöä!
 - * Kansalaisten tiedelukutaidon ja tilastollisten menetelmien tuntemuksen merkitys on kasvanut viime vuosikymmeninä ja kasvanee jatkossa yhä, kun esimerkiksi erilaiset “vaihtoehtotieteet” ovat nousseet suositummiaksi.
 - * Tilastotieteen ymmärrys auttaa itse kutakin tunnistamaan virheellisiä tai puutteellisia tiedoja tehtyjä päätelmiä ja täten helpottaa tietoyhteiskunnassa toimimista ja kriittistä ajattelua!
- Yleisiä tilastollisten menetelmien väärinkäytötapoja ovat esimerkiksi seuraavat:
 - “**Kolmannen tyypin virhe**”: kun tilastollisia menetelmiä käytetään saadaan oikeita vastauksia, mutta väärin kysymyksiin! Esimerkiksi jos tutkija ei täysin ymmärrä minkälaisia vastauksia käytetävissä olevasta aineistosta ja valitulla menetelmällä voidaan saada, voi hän syystyä kolmannen tyypin virheeseen. Tällöin voi nimittäin käydä niin, että hän tulkitsee tilastolliset testit 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 käytettävässä edustava otos molempien ikäluokkien edustajista. Pääätät tutkia *yksisuuntaisesti* onko toisen ryhmän, ryhmän A, keskipituus *pienempi* kuin ryhmän B. Testitulos osoittaa, että voit hylätä nollahypoteesin, jonka mukaan ryhmien *keskipituus oli si sama*. Kolmannen tyypin virhe syntyy silloin, jos tosiasiallisesti testin hylkääminen johtui siitä, että ryhmän A keskipituus olikin *suurempi*

kuin ryhmän B keskipituus, mutta tästä et testin tuloksen perusteella voi tietää!

3.6 Tilastotieteen sovelluskohteita ja “rajatieteitä”

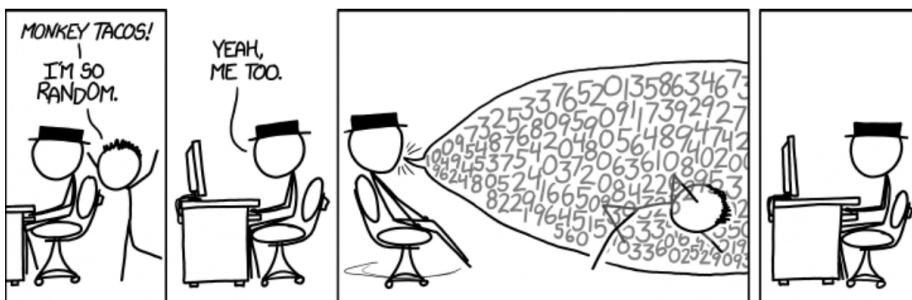
- Yleisenä menetelmätieteenä tilastotiedettä sovelletaan useilla eri tieteenoilla.
 - Jokaisella sovellusalalla on oma erillinen teoriapohjansa sekä empiriset käytänteet, joten substanssitietous on sovellettaessa erityisen tärkeää.
 - * Huolimatta vaihtelevista empiirisistä käytännöistä sovellusmenetelmän 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 yhteisvaikutukessa tilastotieteen ja lähiroteiden (viime aikoina erityisesti koneoppimisen) yhteydessä.
- Usein on pystyttävä arvioimaan ongelmanasettelun ja tulosten tarkoitukseenmukaisuutta ja pyrkiä välttymään siltä että tutkijan tieteelliset ja yhteisölliset sitoumuksivat tutkimuksen kulkun.
- Tilastotieteen pääaineopiskelun osalta substanssitietous saavutetaan usein sivuaineopintojen perusteella. Vastaavasti toisinpäin muiden aineiden pääaineopiskelijoiden kohdalla, jolloin tilastotiede voi yhtä hyvin toimia (laajalti opiskeltuna) vahvana sivuaineena.
- Jokaisella tieteenalalla, jonka tutkimusaineistot voidaan esittää numeerisessa tai kvantitatiivisessä muodossa voi soveltaa/voisi soveltaa/pitäisi soveltaa tilastollisia menetelmiä sekä tutkimusaineistoja keräättäessä 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 syntynyt nk. “rajatieteitä”:

3.6. TILASTOTIETEEN SOVELLUSKOHTEITA JA "RAJATIETEITÄ" 53

- Sovellusalojen joukossa tilastotieteen soveltaminen on muodostunut omaksi tutkimuskohteen/tieteenlajikseen (ks. linkit):
 - * Psykologia: psykometriikka,
 - * Sosiaalitieteet: sosiometria,
 - * Taloustiede: ekonometria,
 - * Kemia: kemometria,
 - * Bio- ja lääketiede: biometria,
 - * Epidemiologia,
- Soveltavan matematiikan tutkimusalojen joukossa ovat osaltaan pääallekäisiä tilastotieteen kanssa
 - * Informaatioteoria,
 - * Matemaattinen tilastotiede,
 - * Todennäköisyyslaskenta,
 - * Operaatioanalyysi
- Tietojenkäsittelytieteen alaan (osittain) lukeutuvia tutkimusalojen joukossa
 - * Laskennalliset menetelmät,
 - * Data mining,
 - * Knowledge discovery,
 - * Hahmontunnistus,
 - * Tekoäly,
 - * Koneoppiminen
- Ja paljon muita!

Luku 4

Sattuma ja satunnaisuus tilastotieteessä



Kuva 4.1: Hauska kuva satunnaisuudesta.

Tässä luvussa pohdimme sattuman ja satunnaisuuden roolia tilastotieteessä ja tieteessä ylipäättäään. Satunnaisuudella tarkoitetaan yleensä säännönmukaisuuden puuttumista ja ennustamattomuutta ja kenties juuri siksi sitä voidaan pitää yhtenä maailman vaikuttavimmista ilmiöistä. Jokainen haluaisi tietää mitä tulenan pitää ja siksi sattuma tekee elämästä mielenkiintoista: se vaikuttaa ja muokkaa niin meitä itseämme kuin ympäröivää maailmaa mitä merkityksellisimmin tavoin - joskus jopa vasten tahtoamme ja usein vailla 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 normaalivaihtelun piiriin. Tällaisissa tilanteissa usein tilastollinen tarkastelu paljastaakin ilmiön todellisen, alkuperäisestä kuvitelmasesta poikkeavan luonteen. Erotaakseen systemaattinen vaihtelu satunnaisesta ja

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

4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä

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

Satunnaisilmiö

Reaalimaailman ilmiö on satunnaisilmiö, jos seuraavat ehdot pätevät:

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

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

4.1. SATUNNAISILMIÖT JA SATUNNAISMUUTTUJAT TILASTOTIETEESSÄ 57

Esimerkkejä satunnaisilmiöistä

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

Satunnaismuuttujat

- Tilastollista vaihtelua ilmentävät tilastolliset muuttujat tulkitaan **satunnaismuuttujiksi** ja havainnot (havaintoarvot) voidaan näin ollen tulkitä näiden satunnaismuuttujien realisoituneiksi arvoiksi. Tällöin tilastollisen tutkimuksen kohteena on nämä havainnot generoinut *satunnaisilmiö*.
 - Satunnaismuuttuja siis kuvaa tarkasteltavan mitattavan ominaisuuden (satunnais)vaihtelua tutkimuksen kohteiden, eli tilastoysiköiden joukossa.
 - Mitattavan ominaisuuden mahdolliset arvot määrääävät satunnaismuuttujan luonteen. Yleisesti satunnaismuuttujat jaetaan kahteen luokkaan: **jatkuviin** ja **diskreetteihin**.
 - Satunnaismuuttujan **todennäköisyysjakama**, määräää erilaisten

tulosvaihtoehtojen todennäköisyyden ja mahdollistaa täten tilastollisen analyysin ja päätelyn.

- * Satunnaisuus eroaa mielivaltaisesta prosessista siinä, että satunnaista ilmiötä voidaan kuvata jollakin **tilastollisella lailla** kun taas mielivaltaista prosessia ei.

Satunnaismuuttuja

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

- Satunnaismuuttujan Y realisoituvaa arvoa y kutsutaan realisaatioksi tai toteumaksi.
- Tilastollinen aineisto muodostuu useiden satunnaismuuttujien (tilastoyksiköiden tutkimusmuuttujien) realisoituneista arvoista.
- Realisoituneiden arvojen vaihetusta tilastoyksiköiden välillä kutsutaan satunnaisvaihteluksi.

Jatkuvat ja diskreetit satunnaismuuttujat

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

Esimerkki: satunnaismuuttuja

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

- Muuttujat voidaan luokitella myös **kvalitatiivisiin** ja **kvantitatiivisiin** muuttujien.
 - Kvalitatiivisiin muuttuihin liittyy luokittelut- tai järjestysasteikko
 - Kvantitatiivisiin muuttuihin välimatka- ja suhdeasteikko.
- Tilastolliset menetelmät perustuvat todennäköisyyslaskennan¹ tuloksiin ja tarjoavat keinon hallita satunnaisuuden aiheuttamaa epävarmuutta sekä tavan erottaa systemaattinen ja satunnainen vaihtelu, eli signaali ja kohina, toisistaan.
- Tilastollisen aineiston **tilastollisella mallilla** tarkoitetaan täten niiden satunnaismuuttujien todennäköisyysjakaumaa, jonka ajatellaan generoivien havainnot.
 - Yksinkertaisimillaan esimerkiksi yksinkertaiseen satunnaisotantaan takaisinpanolla perustuva satunnaismalli (palaamme tähän otantaa käsitleväässä luvussa 5).
 - Satunnaisuus perustuu siihen, että satunnaismuuttujien toteutuvat arvot (ja niistä lasketut tunnusluvut kuten keskiarvo) vaihtelevat satunnaisesti otoksesta toiseen.
- Todennäköisyyslaskennan ja tilastotieteen tehtävä on tuottaa **matematisia ja tilastollisia malleja** satunnaisilmiöissä havaittavalle tilastolliselle stabiliteetille.

4.2 Satunnaisuus ja todennäköisydet

- Tilastotieteessä **tutkimusaineiston keräämistä** voidaan pitää hyvänä esimerkinä satunnaisilmiöstä.
 - Voimme ajatella, että tilastollisen tutkimuksen kohteet on aina valittu arpomalla.
 - Arvonta on mainio esimerkki satunnaisilmiöstä, sillä siihen liittyy aina ennustamattomuutta: vaikka yksittäisen arvonnan tulosta ei voi tietää etukäteen, noudattaa se kuitenkin todennäköisyyden lakeja.
 - Koska arvonnan tulos vaihtelee satunnaisesti arvontakerrasta toiseen, myös tutkimuksen kohteita kuvaavat tiedot vaihtelevat satunnaisesti arvontakerrasta toiseen.
 - Tutkimuksen kohteita kuvaavien tietojen käyttäytymisessä havaitaan kuitenkin arvontaa toistettaessa juuri sitä säännönmukaisuutta, jota kutsutaan tilastolliseksi stabiliteetiksi. **Tämä säännönmukaisuus on tilastollisen tutkimuksen kohde.**

¹Todennäköisyyslaskentaa käsitellään väilläisesti tulevissa luvuissa mutta varsinaisesti tarkeimmin 2. periodin kurssilla [TILM3553 Todennäköisyyslaskennan peruskurssi](#) ja (erityisesti sivuaineopiskelijoille) [TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille](#).

- Esimerkkejä tilastollisten aineistojen keräämisen menetelmistä, jotka perustuvat arvontaan:
 - **Satunnaistetut kokeet:** Kokeellisessa tutkimuksessa tavoitteena on vertailla erilaisten käsitellyiden vaikutuksia kokeen kohteisiin. Erilaisten virhelähteiden kontrolloimiseksi käsitteily on syytä arpoa kohdeille.
 - **Satunnaisotanta:** Otannalla² tarkoitetaan laveasti tutkimusaineistojen keräämisen menetelmiä. Erilaisten virhelähteiden kontrolloimiseksi tutkimuksen kohteet on syytä valita arpomalla. (Ks. Luku 5)
- Kerätyn (tai havaitun) aineiston pohjalta tehdään päätelmiä sen generoineesta satunnaisilmiöstä esimerkiksi testaamalla erilaisia siihen liittyviä hypoteeseja.
 - Tilastotiede voidaan jakaa kahteen merkittävään paradigmaan sen mukaan, miten tilastolliseen päättelyyn, ml. hypoteesihin ja niiden testaamiseen, suhtaudutaan. Näitä ovat **klassinen eli frekventistinen tilastotiede** sekä **Bayesilainen tilastotiede**. Tarkastellaan seuraavaksi minkälaisia eroja ja yhtäläisyyskiä näiden koulukuntien välillä on.

Frekventistinen tilastotiede

- Klassisessa eli frekventistisessä tilastotieteessä ajatellaan että hypoteesien testaaminen tulee perustua yksinomaan havaittuun aineistoon ja siihen liitettävään tilastolliseen malliin.
- Nimi “frekventistinen” juontuu siitä, että tilastollisen mallin perustana oleva todennäköisyysjakama määrittää satunnaismuuttujan mahdollisten arvojen todennäköisyydekseen niiden suhteellisen osuuden äärettömästä määrästä realisaatioita, ts. niiden suhteellisen frekvenssin.
- Klassisessa tilastotieteessä havaittuun aineistoon *sovitetaan* tilastollinen malli, joka vastaa saattua aineistoa parhaiten.
 - Tämä tilastollinen malli voidaan (useimmiten) perustaa nk. **uskottavuusfunktioon**, joka on *aineiston* sekä yhden tai useaman *parametrin* funktio ja joka saavuttaa suurimman arvonsa nk. “suurimman uskottavuuden pisteesä”.
 - Uskottavuusfunktio kertoo kuinka todennäköisenä havaittua aineistoa voidaan pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvolla.

² Erityisesti erilaisten otantamenetelmien yhteydessä, joita tarkastellaan tarkemmin luvussa 5.

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

Bayesilainen tilastotiede

- Bayesilainen tilastotiede on tilastotieteen toinen suuri paradigma ja on saanut nimensä englantilaiselta harrastelijamatemaatikko ja presbyteripappi **Thomas Bayesilta**, jota pidetään Bayesilaisen tilastotieteen isänä.
- Bayesilainen tilastotiede ulottaa todennäköisyyskäsityksen, eli tajauksia, myös aineistoa koskevien hypoteesien puolelle: kuinka todennäköisenä joitain hypoteesia voidaan pitää jo ennen tutkimusaineiston keräämistä?
 - Myös Bayesilaisessa tilastotieteessä hyödynnetään uskottavuusfunktioita, mutta hypoteesien testaus ei perustu niinkään frekventistiseen ajatukseen todennäköisyysistä suhteellisina osuuksina äärettömässä sarjassa.
 - Bayesilaiset perustavat sen sijaan hypoteesien testaamisen tutkimuskysymystä koskevien ennakkokäsitysten päivittämiselle sen jälkeen, kun aineisto on havaittu.
 - Nämä ennakkokäsitykset voidaan kuvata todennäköisyysjakaumana, priorijakaumana, jota päivitetään ns. posteriorijakaumaksi kun aineisto havaitaan. Nämä päättely perustuu priorijakauman ja aineiston uskottavuusfunktion väliselle kompromissille!
- Ajatusta ennakkokäsityksistä todennäköisyksinä käytetään niin Bayesilaisen tilastotieteen kritiikkinä kuin puolustuksena.
 - Lopulta olemme kaikki Bayesilaisia: jokaisella on sisäisiä ennakkokäsityksiä, myös tutkijoilla! Nämä ennakkokäsitykset voivat perustua esimerkiksi aiempaan tutkittuun tietoon, mutta myös uskomuksiin.

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

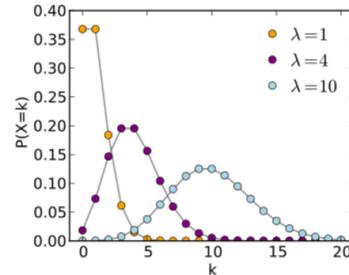
4.3 Tilastolliset mallit, jakaumat ja parametrit

- Tilastolliset mallit perustuvat satunnaismuuttujan mahdollisten tulosvaihtoehtojen todennäköisyyksiä kuvaavalle **todennäköisyysjakaumale**, joka määräät millä todennäköisyydellä satunnaismuuttuja saa erilaisia arvoja.
 - Kuten aiemmin todettiin, satunnaismuuttujat jaetaan kahteen luokkaan: diskreetteihin ja jatkuviin.
- Toisaalta ajoittain tietyn suureen/ilmiön mallinnuksessa voidaan perustusti käyttää molempien luokkiin kuuluvien satunnaismuuttuja- ja tilastollisen mallityypin vaihtoehtoja.
 - Esimerkki: Esimerkiksi COVID19-tartuntatapausten lukumäärä Suomessa on periaatteessa diskreetti satunnaismuuttuja, joka saa yksittäisen (kokonaisluku)arvon joka kuukausi, mutta käytännössä lukumääät ovat tässä tapauksessa sen verran suuria, että niitä mallinneitaan jatkuva-arvoisena muuttujana.
 - Vastaavasti esimerkiksi potilaan jonotusaika päivystyksessä voi periaatteessa saada minkä tahansa arvon tietyltä reaalilukuväliltä ($[0, \infty)$, ts. mikä vain positiivinen arvo) ja tällöin käytettäisiin jatkuviin sm:jiin perustuvia tilastollisia menetelmiä.
- Satunnaismuuttujan mahdolliset arvot määräväät myös mahdollisen todennäköisyysjakauman ja täten myös käytettävän tilastollisen mallin.
 - **Diskreetin satunnaismuuttujan** jakauma voidaan usein esittää taulukkomuodossa. Eri arvojen todennäköisyydet muodostavat kyseisen satunnaismuuttujan todennäköisyysjakauman (**pistetodennäköisyysfunktion**), jota voidaan havainnollistaa esimerkiksi pylväsdiagrammilla.
 - Jatkuvan satunnaismuuttujan Y arvot muodostavat jonkin reaaliakselin välin, joka sisältää äärettömän määren lukuja. Tämän vuoksi jatkuvan satunnaismuuttujan jakauman esittäminen taulukon kautta ei ole luonteva, vaan jakauma esitetään yleensä satunnaismuuttujan **tiheysfunktion** avulla.
 - * Pistetodennäköisyys- ja tiheysfunktioit siis määräväät satunnaismuuttujan mahdollisille arvoille todennäköisyydet väliltä $[0, 1]$ ja näin voidaan arvioida havaitun aineiston uskottavuutta ja testata siihen liitettäviä hypoteeseja suhteessa estimoituun suurimman uskottavuuden estimaattiin.

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

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

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



Kuva 4.2: Esimerkki: Poisson-jakauman sovelluskohteita ja sen pistetodennäköisyysfunktio eri parametrin λ arvoilla.

Parametrien estimointi ja niiden testaus

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

- Perusjoukko parhaiten kuvaavan (eli “aineiston generoineen”) parametrin arvo pyritään siis estimoimaan aineiston perusteella.
- Parametrien estimoinnin lisäksi usein **testataan** parametreja koskevia oletuksia (eli hypoteeseja).
- Estimoointi ja testaus ovat tilastolliseen tutkimukseen liittyvän **tilastollisen päättelyn** keskeisiä välineitä, joiden avulla tutkittavasta ilmiöstä pyritään tekemään johtopäätöksiä siitä kerätyn havaintoaineiston perusteella.
 - Estimoitujen parametrien testaus voi vastata esimerkiksi seuraavaksi laisiin kysymyksiin:
 - * Onko suomalaisten miesten keskipituus 180 cm?
 - * Vaikuttaako yliopistokoulutus tulevaisuuden ansioihin?
 - * Auttaako tietty lääkeaine jonkin sairauden hoidossa?
 - * Voiko osakemarkkinoiden tuottoja ennustaa?
- Parametrien testaus on osa tilastollista päättelyä, johon palataan tarkemmin luvussa [6](#)

4.4 Odotusarvo ja varianssi

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

Odotusarvo

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

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

Esimerkki: Odotusarvo

Perinteikäs esimerkki odotusarvosta on tavallisen kuusitahoinen 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

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

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

Varianssi

Satunnaismuuttujan Y hajontaa voidaan mitata varianssilla

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

tai sen neliöjuuren eli **keskihajonnan** avulla

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

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

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

4.5 Joitain jakaumia

Tarkastellaan seuraavassa muutamia keskeisiä tilastollisia jakaumia. Esittelemme ensin keskeisintä jatkuvien satunnaismuuttujien jakaumaa, normaalijakau-

maa, ennen muutamien diskreettien satunnaismuuttujien jakaumia.

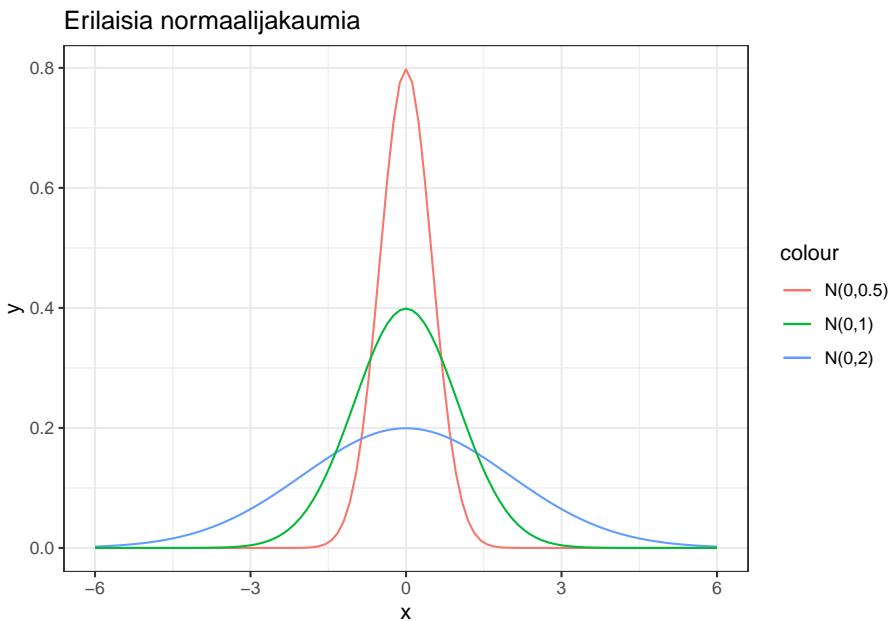
4.5.1 Normaalijakauma

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

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

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

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



Kuva 4.3: Normaalijakaumien muotoja eri parametriarvoilla.

Esimerkki: Miesten pituus

- Tutkitaan miesten pituutta hyvin määritellyssä joukossa, kuten varusmiespalvelusta tietynä vuonna suorittavien joukossa.
 - Pituus on ominaisuus, jonka voidaan nähdä määrätyvän monista perintö- ja ympäristötekijöistä. Pituutta voidaan siis pitää satunnaismuuttujana.
 - Oletetaan, että pituus noudattaa normaalijakaumaa. Näin ollessa Y on valitun miehen pituus ja $Y \sim N(\mu, \sigma^2)$.
- Tuntemattomien parametrien μ ja σ^2 tulkinta:
 - Odotusarvo $\mu = E(Y)$ on satunnaisesti valitun miehen pituuden odotettavissa oleva arvo.
 - Varianssi $\sigma^2 = \text{Var}(Y) = E[(Y - \mu)^2]$ kuvailee valitun miehen pituuden odotusarvostaan määrätyyn poikkeaman (keskijonnan) neliön odotettavissa olevaa arvoa (kuvaten ts. pituksien jakauman keskityneisyyttä/hajaantuneisuutta pituksien odotusarvon ympärillä).

4.5.2 Bernoulli-, binomi- ja Poisson-jakauma

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

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

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

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

- **Binomijakauma**

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

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

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

Esimerkki: Miesten lukumäärä Saksin osavaltion perheissä 1876–1885^a

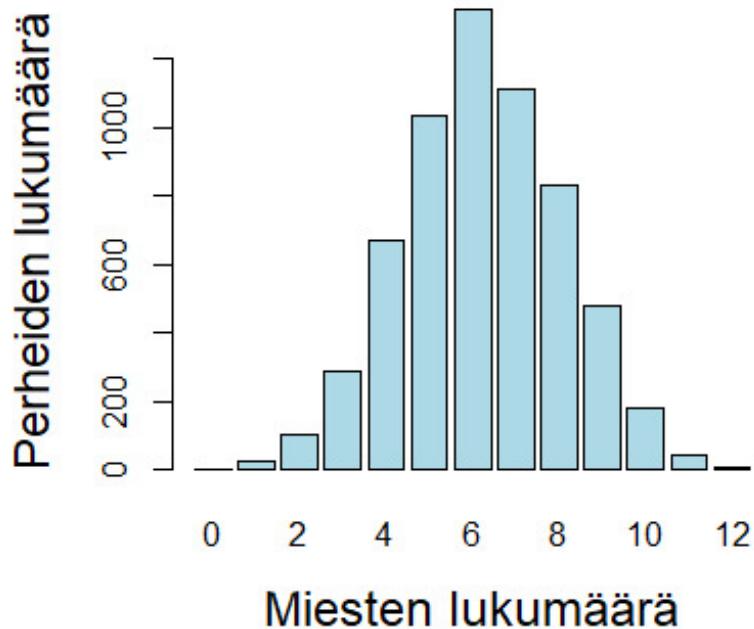
Vuosien 1876–1885 aikana Saksin osavaltiossa rekisteröitiin yli neljä miljoonaa syntynyttä lasta. Tällöin vanhempien tuli ilmoittaa lapsen suku puoli (mies tai nainen) heidän syntymätodistuksensa. Myöhemmässä tutkimuksessa tutkittiin tarkemmin 6115 perhettä, joissa asui 12 lasta ja tarkemmin miesten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

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

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

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

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



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

Poisson-jakauma

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

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

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

Esimerkki: Poisson-jakauma

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

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

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

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

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

Esimerkki (jatkuu): Poisson-jakauma

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

lastollisen päättelyn ja erityisesti estimointiteorian piiriin. Vastaavasti vieraajoukkueen maalimäärität: $Y_i^A \sim P(\lambda_A)$.

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

	Kotijoukkue (home)	Vieraajoukkue (away)	Yht.
Keskiarvo	1.486	1.063	2.550
Varianssi	1.316	1.172	2.618

Tuloksista voidaan siis päätellä, että kotijoukkueen (odottavissa oleva) maalimäärä on vieraajoukkuetta korkeampi (osoittaen kotiedun merkitystä jalkapallossa). Lisäksi edellä todetun Poisson-jakauman teoreettisten ominaisuuksien mukaisesti keskimäärität maalimäärität ovat lähes länniiden variansseja, mikä osoittaa osaltaan (tässä yksinkertaistetussa tilanteessa), että Poisson-jakauman perustuva jakaumaoletus on kelvollinen.

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

4.6 Sattuman rooli tieteenteossa: Vale-emävale-tilasto?

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

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

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

- Kiinnostuksen kohteena on siis hyvin harvoin vain jokin yksittäinen tunnusluku, kuten keskiarvo, varianssi tai korrelaatio (palaamme näihin myöhemmin luvussa 6).
- Tieteen popularisointi on yksi tutkijoiden ja yliopistojen tiedeyhteisön tärkeimmistä yhteiskunnallisista tehtävistä, mutta valitettavan usein se typistyy yksittäisen viimeisimmän tutkimustuloksen esitellyksi.
- Yliopistoyhteisössä kuitenkin luonnollisesti luotamme kumuloituneeseen tutkittuun tietoon ja tiedämme, että **yksittäinen tutkimus on vasta hyvä alku**.
 - Ihmistieteitä, kuten ilmeisesti erityisesti psykologiaa sekä osin myös muiden ohella lääke- ja taloustiedettä, on viimeisen vuosikymmenen ajan puhuttanut paljon niin sanottu **replikaatiokriisi**, sillä useaa arvostettuaan tutkimusta ei ole saatu **toistettua eli replikoitua**.
 - On ymmärrettävä, että replikaatiokriisi, varsinkin jos se on (ala-kohdaisesti) laajalle levinnyttä, murentaa kansalaisten luottamusta tieteellisiin tuloksiin.
 - Toistettavuus on yksi tutkimuksen peruskriteereistä, joka erottaa tieteellisen tiedon muista tietolähteistä, jotka sen puuttuminen herättää ymmärrettävästi huolta tieteellisen prosessin toimivuudesta.
 - Replikaatiokriisiin voi kuitenkin myös tulkita toisin: ilman kriittisyyttä omia (ja muiden) tuloksia kohtaan, ei mitään kriisiä olisikaan, joen silkka sen olemassaolo on osoitus tieteellisen prosessin toimivuudesta.
- Kun tuntee ja tunnistaa sattuman voiman ja ymmärtää kaikki mahdolliset satunnaisuuden lähteet, jotka altistavat tutkimusprosessin virheille, tulee samalla ymmärtääneeksi että eri tavoin koeteltu, useassa tutkimuksessa kumuloitunut tieto tulisi olla kaiken tieteen popularisoinnin keskiössä yksittäisten, mahdollisesti uusien ja yllättävien tutkimustulosten sijaan.
 - Tähän mennessä olemme jo oppineet, että tälle on myös vahvat tilastolliset perustelut: satunnaisen tiedon maailmassa mikään ei ole täysin varmaa, ei edes kaikkein edistyneimpien tilastomenetelmien avulla!

Luku 5

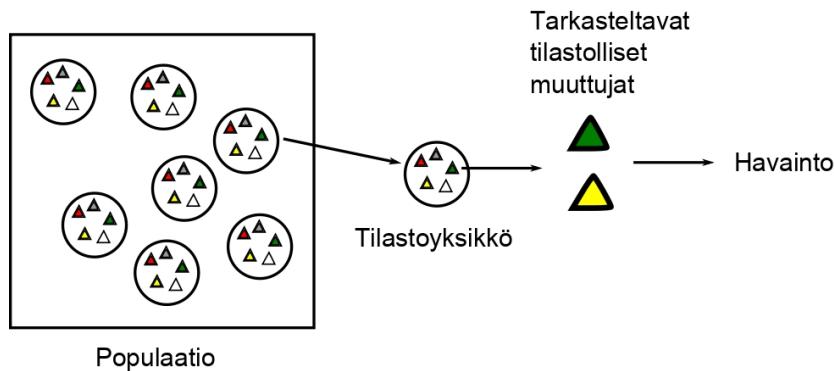
Tilastolliset aineistot, niiden kerääminen ja mittaaminen

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

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

5.1 Kertausta: Data eli aineisto

- Tilastollinen tutkimus aloitetaan tutkimusaineiston keruun suunnitellulla.
- Kertauksen vuoksi: tilastollinen tutkimusaineisto (havaintoaineisto) koostuu tilastoyksiköiden populaatiosta havaituista tilastomuuttujien arvoista.



Kuva 5.1: Populaatiosta havaintoon.

- Havaintoaineisto voidaan koota taulukoksi, johon listataan tilastoyksiköt riveille ja tilastomuuttujat sarakkeisiin. Jos havaintoaineisto koostuu n tilastoyksiköstä, joista jokaisesta on kerätty esim. m :stä tilastomuuttujasta havainnot, niin aineisto voidaan kirjoittaa taulukon muotoon:

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

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

- Usein (varsinkin parhaillaan kiihyväällä vauhdilla) kerättävät havaintoaineistot ovat niin suuria, ettei edellisenkaltaisesta havaintotaulukosta voida usein suoraan tarkastelemalla nähdä aineiston pääpiirteitä.

- Tällöin voi olla tarpeen luokitella aineistoa taulukon muodostamiseksi.
 - Luokittelussa on kysymys aineiston tiivistämisestä kohtuullisen koikiseksi ja havainnollisempaan muotoon. Luokittelussa tilastomuuttujan arvot sijoitetaan eri luokkiin siten, että yhden tilastomuuttujan arvo voi kuulua vain yhteen luokkaan. Luokka ilmoitetaan yleensä luokkavälinä, kuten reaalilukuvälinä. Esimerkiksi henkilön ikä on tapan luokitella ikäjakauaman kuvaamisessa 10-vuotislukkiin (15-24, 25-34, ...), vaikka periaatteessa ikä voitaisiin ilmoittaa minuutinkin tarkkuudella.
 - Luokkien lukumääärään vaikuttavat muun muassa tilastomuuttujan arvojen vaihteluväli ja havaintoaineiston laajuus. Luokittelussa pyritään siihen, että luokkien lukumäärä saadaan tarvittaessa luokkia yhdistämällä kohtuulliseksi ja että luokat valitaan tasavälisesti eli siten, että kahden peräkkäisen luokan alarajojen erotus on vakio. Kun aineistoa luokitellaan, aineiston luettavuus paranee mutta toisaalta osa tiedoista menetetään eivätkä yksittäiset havaintoarvot ole enää tiedossa.
 - Emme vielä tällä kurssilla käsittele tilastografiikan esittämistä tarkemmin. Muun muassa tilastollisen päättelyn peruskurssi (TILM3555) vastaa näihin kysymyksiin tarkemmin. Graafiset menetelmät ovat joka tapauksessa erittäin tärkeä osa aineiston havainnollistamista. Kuvat helpottavat aineiston tulkitsemista ja toimivat usein perusteltuna lähtökohtana monimutkaisempien tilastollisten mallien (ja algoritmien) sovittamiselle.
-
- Kvantitatiivisen tutkimuksen aineistoksi kelpaa periaatteessa kaikki havaintoihin perustuva informaatio, joka on **mittauksen** avulla muutettavissa numeeriseen muotoon.
 - Havaintoyksiköiden tilastollisten muuttujien numeerisia arvoja kutsutaan **havaintoarvoiksi** tai **havainnoiksi**.
 - Kaikki havaitut tilastolliset muuttujat eivät ole aina mielenkiintoisia. Tutkimuksen kannalta mielenkiintoisia muuttuja kutsutaan **tutkimusmuuttujiksi**, joiden lisäksi havaintoaineisto pitää mahdollisesti sisällään **taustamuuttuja**.
 - * Esimerkiksi, jos tutkimuksella halutaan tietoa suomalaisen aikuisväestön mielipiteistä, havaintoyksikköinä ovat aikuisväestöön kuuluvat henkilöt. Jos halutaan tietoa suomalaisista kunnista, havaintoyksikköinä ovat Suomen kunnat jne.

76LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

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

Kokonaistutkimus

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

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

Otantatutkimus

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

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

- Otantatutkimuksessa keskitytään siis perusjoukko edustavan pienemään, mieluusti satunnaisesti valitun otoksen tutkimiseen.
 - Otantatutkimuksissa tiedot kerätään useimmiten haastattelemella, kirjallisella/sähköisellä kyselyllä tai suoraan tietorekistereistä. Tiedonkeruun toteuttaminen (eri sovelluksissa) määrää osaltaan käytetään otantamenetelmän.
 - Teoriassa äärelliseen perusjoukkoon kohdistuvat kokonaistutkimukset voidaan aina tulkita otantatutkimuksiksi (perusjoukko tulkitaan otokseksi hypoteettisesta äärettömästä perusjoukosta)!
 - * Esimerkiksi Galilein tekemät painovoiman vaikutusta kappaleiden putoamisaikaan liittyneet mittaukset. Koetuloksia (mittauksia) voidaan pitää otoksena äärettömästä mahdollisten koetulosten joukosta. Tällöin ainoa mahdollisuus ilmiön tutkimiseen on käyttää otantaa.
- Otantatutkimuksen tulokset voivat olla luotettavampia kuin kokonaistutkimuksen.

78LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

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

5.2 Otannan idea

- Otantatutkimuksen (karkeat) suunnittelua- ja työvaiheet ovat seuraavat:
 1. Tavoitteiden asettaminen
 2. Perusjoukon (populaation) asettaminen
 3. Kehikko
 4. Kerättävän informaation sisältö (mitä tietoa todella tarvitaan, mitä voidaan jättää pois, suunnitellaan kysymykset ja mahdollinen kyselylomake)
 5. Otoskoon määrittäminen
 6. Suoritetaan otoksen poiminta, tietojen keräys ja tarkastus
 7. Aineiston taulukointi ja analysointi
 8. Raportin laatiminen
- Otantatutkimuksessa ajatuksena on siis poimia **edustava otos** siitä populaatiosta (perusjoukosta), joka on mielenkiinnon kohtena eli jota halutaan tutkia ja josta halutaan tietoja.
 - **Tavoiteperusjoukko** on joukko, johon otannan myötä saatavat tutkimustulokset halutaan yleistää. Toisin sanoen, se mistä haluamme tietoja määräää populaation.
 - **Kohdeperusjoukko** on joukko, jota koskevia tietoja halutaan keräää.
 - * Esimerkiksi äänestysikäiset Suomen kansalaiset.
 - * Usein tavoiteperusjoukko = kohdeperusjoukko.
 - * Tavoiteperusjoukko voi joskus olla laajempi (esim. "ihmiset" vs. "suomalaiset").
- Tutkimuksessa (edustavaan) otokseen poimitut tilastoysiköt, näiden tilastolliset muuttujat ja niiden arvot muodostavat **otosaineiston** eli siis tutkimus- tai havaintoaineiston (**datan**).
 - Tutkimuskysymykseen vastatakseen tutkija valitsee sopivan tilastollisen mallin ja estimoii sen parametrit tähän otokseen perustuen.
 - Perusoletuksena on otoksen ja valitun tilastollisten mallin pohjalta suoritettavan tilastollisen päätelyn **yleistävyys koko populaatioon**.

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

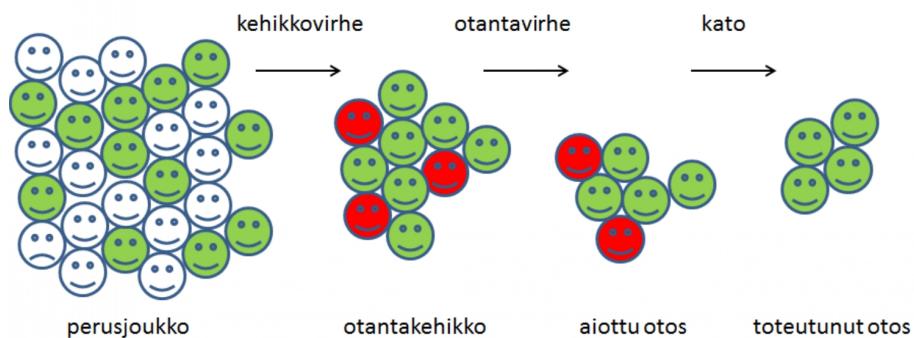
Edustavuus

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

- Keskeistä tutkimuksen ja sen edustavuuden kannalta on, että tutkija osaa kerätä sisällöllisesti ja määrällisesti **sopivan kokoinen** aineiston.
- Tietyn otoksen edustavuutta arviodessa voi käyttää apuna seuraavia kysymyksiä:
 - Miksi päädyttiin tämän kokoiseen otokseen?
 - * **Otoskoko** vaikuttaa siihen miten hyvin otoksesta tehdyt johdotöökset voidaan yleistää koskemaan koko perusjoukkoa, ts. kuinka luotettavia ne ovat. Tämä johtuu siitä että yksittäisten otosyksiköiden ominaisuudet saattavat vaihdella suuresti ja kasvattamalla otoskokoa perusjoukon systemaattiset piirteet tulevat otoskoon kasvaessa yhä paremmin esille. Kun otoskoko vastaa populaation kokoa, on kyseessä tietenkin kokonaistutkimus, joka kertoo kaiken perusjoukosta. Otoskoon valintaan ja määräämiseen palataan myöhemmin luvussa [6](#).
 - Käytettiinkö apuna tilastotieteellisesti vankkaa suunnittelua otoskoon määrittämiseksi ja/tai miten pyrittiin varmistamaan tutkimuksen kannalta tärkeisiin analyysiryhmiin kuuluvien riittävä määrä aiheistossa?
 - Harkittiinko muita otantamenetelmiä ja miksi päädyttiin juuri käytössä olleeseen menetelmään?
 - Edustavuuteen vaikuttaa keskeisesti se, millä tavoin otanta pystytään suorittamaan, ts. mihin kohdeperusjoukkoon otanta kohdistetaan.
 - **Kehikkoperusjoukko** on rekisterin, luetteloon tms. peittämä osa kohdeperusjoukkoa. Kyseessä on siis se osa kohdeperusjoukkoa, josta otanta ylipäänsä pystytään suorittamaan eli **otantakehikko**.
 - **Otantakehikon alipeitto** esiintyy, kun otantakehikosta puuttuu osa kohdeperusjoukon alkioista (esim. tutkimus suoritetaan puhelin-haastattelulla, mutta osa aiottuun otokseen kuuluvista haastateltavista ei omista puhelinta). Vastaavasti **otantakehikon ylipeittoa** esiintyy, kun otantakehikkoon kuuluu kohdeperusjoukkoon kuulumattomia alkioita.
 - * Nämä ovat nk. **kehikkovirheetä**. Lisäksi esimerkiksi kyselytutkimuksissa tai rekisteriaineistoissa saattaa esiintyä **katoa**, eli osa vastauksista jää uupumaan tai niitää ei jostain syystä mitata.

80LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

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

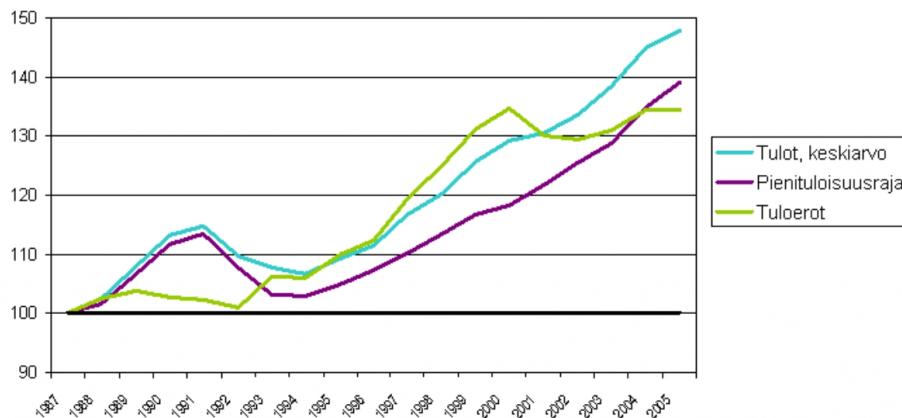


Kuva 5.2: Otannan idea.

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

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

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



Kuva 5.3: Tuloerot.

5.3 Mittaaminen ja mitta-asteikot

Mittaaminen

- Tilastotieteellinen tutkimus perustuu aina mitattaviin satunnaisilmiöihin: tavoitteena on mittamalla liittää jokin luku ilmiötä kuvavaan ominaisuuteen, ts. mitata kyseisen satunnaismuuttujan havaittua arvoa.
- Kumpaa tahansa tutkimusotetta (kokonais- tai otantatutkimus) noudattaessa tietojen keräämisessä on olennaisena osana kohteiden ominaisuuksien **mittaaminen**.
 - Mittaaminen vaatii aina mittauksen kohteen, hyvin määritellyn mitattavan ominaisuuden ja **mittarin**, joka liittää mielekkäät lukuarvot mitattavaan ominaisuuteen.
 - Eriaiset mittarit heijastavat ilmiön ominaisuuksia eri tavoin ja eri tarkkuudella
 - * Esimerkiksi, jos tutkitaan opiskelijoiden pituuden kehitystä, niin mitataan pituutta eri aikoina. Pituudet voidaan mitata senttimetreissä, metreissä, kilometreissä tai vaikkapa tuumissa.
 - * Mittari on hyvä, jos sen antama mittaus on
 - (i) **validi** eli mittaus esittää oikein mitattavaa ominaisuutta (senttimetri mittaa pituutta, gramma ei) ja
 - (ii) **luotettava** eli mittaus on **harhaton** ja **toistettavissa**.
 - * Määritellään nämä termit vielä erikseen, sillä ne ovat keskeisiä tilastotieteessä.

82LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

Harhattomuus

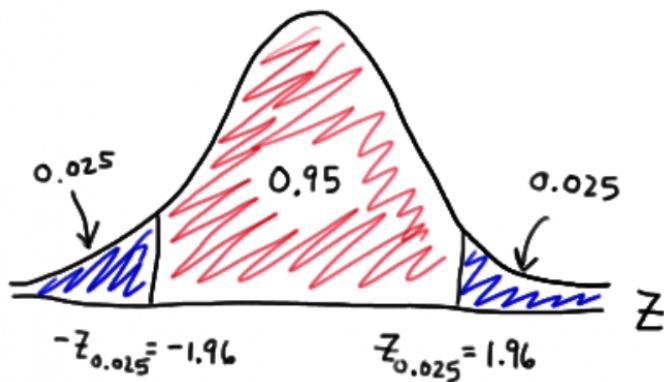
Mittari on harhaton, jos se ei systemaattisesti ali- tai yliarvioi mitattavan ominaisuuden määrää.

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

Toistettavuus

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

- Huonosti toistettava mittari antaa tilastoysiköiden samankaltaisille ominaisuuksille hyvin erilaisia arvoja riippuen otoksesta.
- **Mittausten reliabiliteettiä/luotettavuutta** arvioidessa voidaan pohjata esimerkiksi seuraavia kysymyksiä:
 - Kuinka hyvin mittaustulokset ovat toistettavissa, kuinka paljon niissä on ei-sattumanvaraisuutta?
 - Mittausten validiteetti: kuinka hyvin pystytteiin mittamaan 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. Esimerkinä **luottamusväli**:
 - Väli, joka vaihtelee otoksesta toiseen ja joka usein sisältää mielenkiinnon kohteenan olevan parametrin, kun otantakoetta toistetaan!
 - Luottamusväliä käytetään määrittämään estimaatin luotettavuutta.
 - Väliestimointia tarkastellaan tarkemmin luvussa 6.



Kuva 5.4: Normaalijakaumaan perustuva 95% luottamusväli.

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

Esimerkki: C-vitamiinin vaikutus syövän hoidossa

- Annettiin C-vitamiinia 100:lle terminaalivaiheen syöpäpotilaalle ja seurattiin kuolleisuutta (Cameron and Pauling, 1976).
 - Pyrittiin luomaan tärkeiden ominaisuuksien suhteen samanlaisia verrokkiryhmiä ja valittiin kutakin potilasta kohden 10 verrokkia, jotka olivat samanlaisia iän, sukupuolen, primäärikasvaimen sijaintipaikan ja histologisen kasvaintyyppin suhteita.
 - Seuranta-aika: aika hetkestä, jolloin todettiin tavanomaisten hoitojen olevan tehotonta, kuolinhetkeen saakka.
 - Tulos: C-vitamiinia saaneet käsitellyryhmän potilaat elivät 4 kertaa kauemmin ($p < 0.0001$).
- Ristiriitaista evidenssiä saatatiin tutkimuksessa, jossa vastaava tutkimusongelma, mutta toteutettu satunnaistettuna kokeena (Moertel et al. 1985).

84LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Satunnaistettiin potilaat, joilla pitkälle edennyt paksunsuolen tai peräsuolen syöpää, C-vitamiinia saavien ja lumelääketää saavien ryhmiin.
- Tulos: kontrolliryhmän potilaat elivät keskimäärin hieman pidempään, mutta ero ei tilastollisesti merkitsevä.
- Mistä kahden tutkimuksen erot johtuivat?
 - Huonolla tuurilla kaltaistetut verrokit erosivat käsittelyryhmän potilaista joillakin merkittävillä tavoilla, joita ei oltu mitattu! Miten kvantifioida “huonoa tuuria”?
 - Tilastolliset menetelmät tekevät juuri tämän: “Mikä on todennäköisyys, että havaittu tulos (tai sitä enemmän nollahypoteesista poikkeava tulos) olisi syntynyt vain sattumalta?”
 - Ilman satunnaistamista tuota kenties merkittävää ei-mitattua eroa ei pystytä varmuudella kontrolloimaan.
 - Todellisuudessa ero johtui siitä, että ensin mainitun tutkimuksen kontrollit valittiin jo kuolleista syöpäpotilaista, eikä heihin liittynyt enää mitään satunnaisuutta!

Mitta-asteikot

- Kuten satunnaismuuttujia koskeneessa luvussa 4 opittiin, satunnaismiljöillä on erilaisia tulosvaihtoehtoja, jotka kantavat satunnaismuuttujien todennäköisyysjakaumia.
 - On syytä huomauttaa, että vaikka mitattava ilmiö ei olisikaan numerinen, se voidaan aina “koodata” eli muuntaa numeeriseksi. Esimerkiksi perinteinen kaksiarvoinen mies-nainen -muuttujan tapauksessa voidaan käyttää tunnuksia 0 ja 1.
- Ilmiön luonteesta riippuen voidaan näille tulosvaihtoehdolle käyttää erilaisia **mitta-asteikkoja**.
 - **Laatueroasteikko/luokitteluateikko** (nominaaliasteikko): Muuttujan mittautaso on tällöin sellainen, että sen arvot voidaan luokittaa toisistaan eroaviin luokkiin. Ts. mihin luokkaan kohde kuuluu mitattavan ominaisuuden perusteella?
 - * Tilastoysiköt luokitellaan ennaltamääriteltyihin luokkiin. Luokkien järjestyksellä ei ole merkitystä.

- * Kukin tilastoyksikkö kuuluu vain yhteen luokkaan. Tällöin kahdesta tilastoyksiköstä/havainnosta voidaan päätellä vain kuuluvatko ne saamaan luokkaan vai eivät.
- * Emme pysty määrittelemään empiirisesti mielekästä järjestystä havaintoarvojen väillä.
- * Esimerkkejä: Sukupuoli, veriryhmä tai kotikunta.
- **Järjestysasteikko** (ordinaaliasteikko): Tällöin muuttujan arvot voidaan luokitteluun lisäksi asettaa empiirisesti mielekkääseen järjestykseen. Tällöin siis mittauksen kohteella on “enemmän mitattavaa ominaisuutta” kuin jollakin toisella kohteella
 - * Tilastoyksiköt luokitellaan ennalta määritettyihin luokkiin, joilla on yksikäsitteinen järjestys.
 - * Esimerkkejä: Sotilasarvo, sosiaaliryhmä, kilpailun tulos tai sairauksien tarttuvuuus.
- **Välimatka-asteikko** (intervalliasteikko): Luokittamisen ja järjestyskseen asettamisen lisäksi havaintoarvojen välimatkalla on empiirisesti mielekäs tulkinta. Ts. intervalliasteikon tasaisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi).
 - * Välimatka-asteikolla pystytään mittamaan yksittäisten luokkien tai havaintoarvojen ero. Esimerkiksi: Lämpötilan mittäminen esim. celcius-asteina. Pystymme numeroarvoina ilmoittamaan onko tänään lämpimämpi, yhtä lämmin vai kylmempi sää kuin eilen ja kuinka monta astetta muutos on.
 - * Kuinka paljon kahden mittauksen koteen ominaisuudet eroavat toisistaan.
 - * Intervalliasteikon tasaisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi). Mittarin nollapiste on kuitenkin ”keinotekoinen” ja siten vapaasti valittavissa. Samoin voidaan valita käytettävä mittayksikkö vapaasti. Oleellista on vain se, että havaintojen välisellä välimatkalla on aina empiirisesti mielekäs tulkinta.
 - * Yhteen- ja vähennyslasku ovat sallittuja.
- **Suhdeasteikko:** Jos intervalliasteikon ominaisuuksien lisäksi on määriteltyä yksikäsitteinen mittalukujen absoluuttinen nollapiste.
 - * Esimerkiksi kuuden euron hintainen tuote on kaksi kertaa niin kallis kuin kolmen euron tuote.
 - * Kunnan veroäyri tai henkilön pituus: Absoluuttinen nollapiste on 0.
 - * Nollapisteen ollessa absoluuttinen, se ”pysyy paikallaan” ja mittalukujen suhteet pysyvät samoina.
- Mitta-asteikot voidaan jakaa kahteen luokkaan: **Luokittelu- ja järjestysasteikkoja kutsutaan kvalitatiivisiksi asteikkoiksi.** Tällöin muuttujien arvot kuvaavat vain tilastoyksiköiden laadullisia piirteitä.

86LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Vastavasti **välimatka- ja suhdeasteikko kutsutaan kvantitatiiviseksi asteikoiksi**, koska tällöin mittaluvut kuvaavat jonkin ominaisuuden määräätä.
- Tilastollisen analyysin kannalta mitta-asteikkojen merkitys on siinä, ettei tilastollisten (matemaattisten) operaatioiden sallittavuus määräytyy muuttujan mitta-asteikon mukaan. Mitä ”korkeampi” mitta-asteikko, sitä enemmän on käytettäväissä olevia analyysimenetelmiä. Esimerkiksi keskiarvon laskeminen on eräs tilastollinen operaatio, ja se ei ole sallittu kvantitatiivisille muuttujille.

Aineistotyypejä

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

5.4 Kontrolloidut kokeet ja suorat havainnot

- Tilastollinen tutkimusaineisto voidaan kerätä:
 - **Kontrolloidulla kokeilla**, joissa tutkimuksen kohteet altistetaan suunnitelmallisesti erilaisiin koeolosuhteisiin selvittääkseen miten kohteet reagoivat muutoksiin.
 - **Suoria havaintoja** tehtäessä 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 mahdollista tarkastella tutkittavaa asiaa koeolosuhteissa ”eristyksissä”.
- Kontrolloidut kokeet eivät kuitenkaan ole aina mahdollisia, jolloin on käytettävä suoria havaintoja.
 - Tällöin tutkimuskohdetta ei suunnitelmallisesti altisteta koeolosuhdeille (”käsittelyille”) vaan muuttuvien olosuhteiden vaikutuksia tilastoyksikköihin seurataan passiivisesti.

- Toisin sanoen tutkimuksen kohteena olevat tilastoyksiköt eivät välttämättä edes tiedä osallistuvansa tutkimukseen.
- Lisäksi usein tehdään hoito/käsittelyvastetta koskevia vertailuja erilaisissa olosuhteissa, joka osaltaan vaikuttaa tulosten uskottavuuteen, sillä tutkitavien tilastoyksiköihin voi vaikuttaa olosuhteiden muutosten lisäksi muut ulkopuoliset tekijät.
 - Näiden **selittävien ja sekoittavien tekijöiden** vaikutusten kontrollointi on suoria havaintoja tehtäessä vaativa tehtävä.
 - Mikäli ulkopuolisia tekijöitä ei havaita ja/tai pystytä mittamaan, tai muuten jostain syystä olla lisätty ja käytetty käytettävässä tilastollisessa mallissa, voi kyseeseen tulla ns. **puuttuvien selittäjien harha**, joka tarkoittaa sitä että havaittuun tuloksiin vaikuttaa joakin havaitsematon tekijä, mutta jonka vaikutusta ei kyetä kvantifioimaan puutteellisten havaintoarvojen vuoksi.
- Suoria havaintoja tehtäessä ei voida (usein) selvittää vasteen ja olosuhteiden **kausaalista** yhteyttä. Suorilla havainnoilla voidaan lähiinä saada selville onko vasteella ja olosuilla jokin yhteys (korrelatio) (ks. luku 7).
- Suorien havaintojen keräämiseen liittyy ollenaisesti joitain riskejä ja toisaalta rajoituksia. Riskit liittyvät käytännössä otoksen harhaisuuteen (erit. valikoitumisharha).
 - Esimerkiksi jos havaintoja tehtäessä suositaan systemaattisesti joitakin tulosvaihtoehtoja. Tämä suosiminen voi olla tahallista tai tahaonta.
 - Tämä tilastoyksiköiden **valikoituminen** otokseen aiheuttaa harhaa, sillä otokseen valikoituvia osajoukko saattaa ylikorostaa perusjoukon joitain ominaisuuksia.

Valikoituminen

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

- Esimerkiksi verrattaessa sydän- ja verisuonitautipilaiden hoito-toimenpiteitä potilaat eivät mahdollisesti ole valikoituneet yhtä todennäköisesti pallolaajennukseen, ohitusleikkaukseen tai lääkehointiryhmään, sillä taudin vakavuus saattaa jo määritellä mikä hiototoimenpide valitaan.
- Valikoituminen on iso ongelma seurantatutkimuksissa, sillä harhaisien 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 määräät mitkä perusjoukon alkioista tulevat poimituksi otokseen (tutkimuksen kohteiksi)!

Satunnaistaminen

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

- Satunnaistaminen takaa sen, että mahdolliset sekoittavat tekijät ovat jakaantuneet tasaisesti tutkittavassa joukossa. Tällöin sekoittavat tekijät eivät aiheuta harhaa otokseen ja tutkimuksen tulokset voidaan yleistää koko populaatioon.
- Satunnaistaminen poistaa otannasta valikoitumisharhan, sillä otokseen poiminta suoritetaan riippumatta tilastoyksiköiden ominaisuuksista. Satunnaistaminen on ainoa puolueeton tapa poimia otos (ei suosi mitään perusjoukon osaa)!
- Satunnaistaminen (osaltaan) mahdollistaa **tilastollisen päättelyn**, jolla avulla otoksesta saatuja tietoja voidaan hyödyntää tehtäessä päätelmiä koko perusjoukosta.
 - Tilastollisen päättelyn avulla voidaan muodostaa esimerkiksi jakaukien ja tilastollisten mallien tuntemattomille parametreille arviot (piste-estimaatit) ja arvioida niiden epävarmuutta (keskivirheet ja luottamusväli) sekä testata tarkasteltavaan ilmiöön liittyviä hypoteeseja (ks. luku 6).
- Johtopäätelmien pätevyys riippuu mm. siitä, kuinka hyvin otanta on suoritettu. Tämän vuoksi on tärkeää ymmärtää otannan perusperiaatteet ja erilaisten otantamenetelmien luonne.
- Kontrolloiduissa kokeissa satunnaistaminen jakaa yksilöt **riippumatta yksilön omista ilmiöön vaikuttavista muuttujista joko käsitellyt tai kontrolliryhmään** (eng. treatment ja control).
 - Se takaa, ettei valikoitumista jonkin käsitellytä edeltävän ominaisuden mukaan esiinny.

- Tämä tarkoittaa **altisteen** (käsittely / “treatment”) antamista (täyssin) satunnaisesti kokeeseen valituille yksilöille, riippumatta näiden taustamuuttujien arvoista.
- Nämä yksilöt sinänsä voivat olla satunnaisotos jostain populaatiosta (tai ainakin niiden toivotaan olevan), mutta satunnaistaminen tarkoittaa siis käsittelyn kohdentamista koeyksilöille, ei satunnaisotantaa sinänsä.
- Esimerkiksi tutkittavat voidaan satunnaistaa lääkehoito- ja placebo-ryhmiin, jotta mahdolliset erot tutkittavien iässä, sukupuolessa ja muissa taustamuuttujissa eivät aiheuta systemaattista harhaa, kun tutkitaan lääkehoidon vaikutusta.

5.5 Otantamenetelmät

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

Sisältymistodennäköisyys

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

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

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

90LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

jossa siis N on perusjoukon alkioiden lukumäärä. Toisin sanoen, kaikilla perusjoukon alkioilla on oma nollaa suurempi todennäköisyys (voi olla 1), 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 **analyysipainojojen** muodostamisessa sekä uudelleenpainotuksessa (vastauskadon korjaus).
- Tässä luvussa käsitellään erilaisia perinteisiä otantamenetelmiä sekä siitä, minkälaisista 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 poimin-takehikkoon (perusjoukkoon) kuuluvat alkiot järjestetään jonoon ja siitä poimitaan otokseen joka k. alkio.
 - **Oositettu otanta**: perusjoukko (populaatio) jaetaan ominaisuuksiltaan yhtenäisiin eli homogenisiin **ositteisiin**, joista jokaisesta poimitaan erillinen otos.
 - **Ryväsatanta** tai joskus myös **moniasteinen otanta**: Hyödynnetään perusjoukossa esiintyvää kerroksellisuutta, eli hierarkkisuutta otannassa.

5.5.1 Yksinkertainen satunnaisotanta

- **Yksinkertaisessa satunnaisotannassa** (YSO) jokaisella tilastoyksiköllä (perusjoukon alkiolla) on nollasta poikkeava todennäköisyys tulla valituksi otokseen.
 - Otannan satunnaisuus tulee siis siitä, että jokainen tilastoyksikkö poimitaan otokseen *satunnaisesti!* (Ks. luku 4)
 - YSOa pidetään otannan perusmuotonä, jossa jokaisella perusjoukon alkiolla on lähtökohtaisesti yhtä suuri todennäköisyys tulla valituksi otokseen.
 - * Yksinkertainen satunnaisotanta on periaatteeltaan intuitiivinen ja helppo ymmärtää. Lisäksi se on tietyissä tilanteissa usein helppo toteuttaa.
 - Tällöin on selvää että myös jokaisella perusjoukon samankokoisella osajoukolla on sama todennäköisyys tulla valituksi.

- Toisin sanoen, todennäköisyys tulla poimituksi ei riipu tilastoyksikön ominaisuuksista tai siitä minkälaisia ominaisuuksia jo poimituilla otosyksiköillä on.
- Satunnaisotanta siis selvästi korjaa valikoitumisharhaa (ks. aiempi luku 5.4) satunnaistamalla otokseen valikoitumisen täysin! YSO voi daankin aina tulkita arvonnaksi. Käytännön työssä arvonta onkin oiva satunnaistamisen keino.

- **YSO:n toteuttaminen**

- Käytännössä yksinkertainen satunnaisotanta etenee vaiheittain:
 - * Tutkimuksen alussa tutkijalla tulisi olla käytettäväänään (ts. tulisi koostaa) lista kaikista perusjoukon havaintoyksiköistä (alkioista). Tämä muodostaa tutkimuksen **otantakehikon**.
 - * Tämän jälkeen jokaiseen perusjoukon alkioon voidaan liittää numeriset tunnukset.
 - * Sitten valitaan haluttu otoksen koko. Otoskoon määrittäminen on keskeinen osa koesuunnittelua, ks. luku 6.6
 - * Otantakehikosta arvotaan perusjoukon alkiot otokseen yksi kerrallaan.
 - * Käytännössä arvonta voidaan toteuttaa satunnaislukuja generoimalla (tuottamalla) niin että jokaisen otantakehikon alkion sisältymistodennäköisyyss 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 tarkoituksesta on poimia n :n alkion kokoinen otos (huom. $n < N$). Olkoon i yksittäisen alkion indeksiluku (ts. jokainen alkio on numeroitu esimerkiksi tavalla $i = 1, \dots, N$).

YSO:n poiminta palauttaen

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

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

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

92LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

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

YSO:n poiminta palauttamatta

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

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

- Tässä A_i on jo poimittujen alkioiden lukumäärä ennen kyseistä **otosite- raatiota**: ensimmäisen poiminnan kohdalla $A_i = 0$, toisen kohdalla $A_i = 1$ ja niin edespäin.
 - Ilman takaisinpanoa populaatiosta voidaan poimia $\binom{N}{n}$ erilaista otosta.²
 - Otantaan palauttamatta liittyviä todennäköisyyksiä hallitaan **hy- pergeometrisen jakauman** avulla, joka johtaa (melko) yksinkertaiseen **tilastolliseen malliin** YSO:a käytettäessä.

Esimerkki: Yksinkertaisen satunnaisotannan poimintastrate- giat

- Esimerkki: Poimitaan palloja kulhosta satunnaisesti.
 - Jos yksittäinen pallo (alkio) voi tulla poimituksi useammin kuin kerran, eli pallo palautetaan kulhoon sen poiminnan jäl-

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

keen, on kyseessä yksinkertainen satunnaisotanta takaisinpanolla.

- Vastaavasti jos pallo voi tulla valituksi vain kerran, eli pallo poistetaan kulhosta sen poiminnan jälkeen, on kyseessä otanta ilman takaisinpanoa.

Otoskoon vaikutus YSO:n

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

YSO: Potentiaaliset ongelmat

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

5.5.2 Systemaattinen otanta

- Systemaattisessa, eli tasavälistessä, otannassa poimintakehikkoon (perusjoukkoon) kuuluvat alkiot järjestetään jonoon ja siitä poimitaan otokseen joka k . alkio.

94LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- 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 omaisuudesta, jota tutkitaan.
- Systemaattisen otannan suhteen potentiaaliseksi ongelmaksi muotoutuu havaintoysikkölistan mahdollinen säädöllinen jaksollisuus, jota se ei havaitse ja jolloin satunnaisotanta toimisi (kenties) paremmin.
 - * Ongelmaa syntyy esimerkiksi silloin, jos tiedot perusjoukosta koostuvat heteropariskunnista ja poimintaintervalli on parilleen luku. Tällöin seurausena voi olla, että otokseen saattaisi valikoitua ainoastaan joko miehiä tai naisia.
- Myös systemaattisessa otannassa tarvitaan siis lista tai rekisteri kaikista perusjoukon havaintoysiköistä ja sitä soveltaankin tavallisesti YSO:n sijasta silloin, kun perusjoukon alkioista on käytettäväissä tietorekisteri, luettelo tai havaintoja kerätään ajassa tai tilassa.
 - Esimerkiksi mielipidekyselyn kohteet poimitaan (voitiin poimia) puhelinluettelosta (tai vastaavasta rekisteristä) valitsemalla haastateltavaksi jokaiselta aukeamalta ensimmäisenä esiintyvää henkilö tai joitain tuotetta valmistavan tehtaan laaduvalvonnassa valitsemalla laatuvarvointiin joka sadas tuote, joka hihnalta valmistuu. Muita esimerkkejä ovat esim. liikenne-, jäsenrekisteri- tai kassajonossa seisivien otantayksiköiden poiminta otokseen.

5.5.3 Ositettu otanta

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

- Esimerkiksi jos tavoitteena on tutkia jonkin maan erilaisten ja usein hyvin eri kokoisten kieliryhmien taloudellista asemaa. Kaikista ryhmistä tulisi saada edustava otos.
- Tällöin maan koko populaatioon kohdistettu yksinkertainen satunnaisotanta ei olisi järkevä, sillä otoskoon pitäisi olla (todennäköisesti) hyvin suuri, että jokaisesta kieliryhmästä saataisiin poimittua edustava otos.
- Ositetun otannan avulla otos voitaisiin kerätä niin, että jokaisesta ryhmästä (ositteesta) poimitaan osaotos yksinkertaisella satunnaisotannalla tai systemaattisella otannalla ja nämä osaotokset yhdistetään yhdeksi otokseksi.
- Osittu 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 ongelmaa kuin yksinkertaiseen ja systemaattiseen satunnaisotantaan.
 - Otokseen valikoituneet vastaajat voivat olla mm. levittätyneinä suulle maantieteelliselle alueelle. Näin ollen otannan suorittaminen vaatii suuria kustannuksia.
 - Onko (järkevä) osittaminen ylipäätään mahdollista toteuttaa tarkasteltavassa sovelluskohteessa?

5.5.4 Ryvästonta

- Ryvästonta soveltuu tilanteisiin, joissa perusjoukko on “ryvästeistä” eli se voidaan jakaa luonnollisiin ryhmiin eli rypäisiin (eng. *clusters*).
- Rypäät indikoivat aineiston luontaisista 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 onkin, että toisin kuin ositetussa otannassa, ryvästannassa rypäiden oletetaan olevan toistensa kanssa riittävän samankaltaisia, että jokaista rypästä ei tarvitse erikseen tutkia.
 - Tämä onkin yksi ryvästannan tärkeimpää motivointeja, sillä sitä usein perustellaan kustannustehokkuudella: sen sijaan että poimitaan satunnaisia koululaisia mahdollisesti suuresta määrästä kouluja, voidaan poimia satunnaisia rypäitä (kouluja), joista tutkimusyksiköt eli koululaiset poimitaan.

96LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERAÄMINEN JA MITTAAMINEN

- Lisäksi koulun sisällä koululuokat muodostavat alirypäitä, joista voidaan edelleen poimia satunnaisotos, jotta päästään tutkimaan perusjoukon alkioita eli koululaisia esim. haastattelututkimuksen muodossa.
- Tavoitteena on vähentää tietojen keruun aiheuttamia kustannuksia samalla varmistaen, että otos on kuitenkin mahdollisimman edustava!
- Ryvästannan voi suorittaa **yksi-** tai **kaksivaiheisena** (**yksiasteinen/kaksiasteinen ryvästanta**).
 - **Kaksivaiheisessa ryvästannassa**
 - * **Ensimmäisessä vaiheessa** poimitaan joukko rypäitä kaikkien rypäiden joukosta, eli vain osa rypäistä on mukana lopullisessa otoksessa.
 - * **Toisessa vaiheessa** poimitaan ensimmäisessä vaiheessa poimittuista rypäistä alkiotason otokset.
 - **Yksivaiheisessa ryvästannassa** toisessa vaiheessa valitaan kaikki ensimmäisen vaiheen otosrypäiden alkiot, jolloin toisen vaiheen otanta typistyy ensimmäisen vaiheen rypäiden alkioiden kokonais-tutkimukseksi.
 - Poiminnan eri vaiheissa voidaan soveltaa yksinkertaista satunnaiso-tantaa tai systemaattista otantaa.
- Ryvästantaa käytetään usein suuria haastattelututkimuksia tehtäessä. Erityisesti, ryvästantaa voidaan hyödyntää myös silloin, kun tutkijalla ei ole käytettävissään kattavaa listaa kaikista havaintoyksiköistä, mutta näiden muodostamat rypät on määritettävissä.
- Ryvästannan heikkoutena pidetään sitä, ettei aina ole helppoa muodostaa rypäitä, jotka ovat toistensa kaltaisia. Tulosten tarkkuus myös riippuu moninpaikoin siitä, kuinka hyvin rypäisiin jako onnistuu.

Esimerkkejä ryvästannasta

- Esimerkki 1:
 - Poimitaan oppilaitoksen opiskelijoista otos arpomalla ensin otos luokkahuoneista (=rypäistä).
 - Arvottuissa luokkahuoneissa käydään sitten suorittamassa ky-sely.
 - * Esim. Oppilaitoksen opiskelijoista voidaan poimia otos arpomalla ensin otos luokkahuoneista, jolloin luokkahuo-neet ovat nk. rypäitä.
 - * Mahdollisia ongelmia? Miten huomoida päivä- ja iltao-piskelijat? Tämän voisi toteuttaa arpomalla otos luokka-

huoneista päiväsaikaan ja toinen otos ilta-aikaan. Tässä yhdistetään ryvästontaan ositettu otanta, jolla taataan päivä- ja iltaopiskelijoiden edustus.

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

5.6 Otantaesimerkkejä

Esimerkki: Työllisyys ja työttömyys, Tilastokeskuksen työvoimatutkimus

- Työvoimatutkimus on otostutkimus, jonka avulla tilastoidaan 15–74-vuotiaan väestön työmarkkinoille osallistumista, työllisyyttä, työttömyyttä ja työaikaa (yhden viikon aikana) kuukausittain, neljännesvuosittain ja vuosittain.
 - Työvoimatilastoja käytetään työvoimapoliittisten ennusteiden ja suunnitelmien laadinnassa, toimien seurannassa ja päätöksenteon tukena.
 - Työmarkkina-aseman perusluokittelussa väestö jaetaan työllisiin, työttömiin ja työvoiman ulkopuolisiin.
 - * Työlliset ja työttömät muodostavat työvoiman.
 - Työvoimatutkimuksen **perusjoukon** muodostavat Suomessa vakinaisesti asuvat 15–74-vuotiaat henkilöt.
 - Työvoimatutkimuksen otos poimitaan **ositetulla satunnaisotannalla** väestön keskusrekisteriin perustuvasta Tilastokeskuksen väestötietokannasta kahdesti vuodessa.
- Ositetun satunnaisotoksen poiminta:
 - Tutkimus on paneelitutkimus, jossa samaa henkilöä haastatellaan viisi kertaa.
 - Joka kuukauden otokseen kuuluu noin 12 000 henkilöä, keskimäärin noin joka 300. henkilö perusjoukosta.
 - Yhden tutkimuskuukauden otos koostuu viidestä rotaatioryhmästä, jotka ovat tulleet tutkimukseen mukaan eri aikoina. Otos vaihtuu asteittain siten, että kolmena peräkkäisenä kuukautena vastaamisvuorossa ovat eri henkilöt.

- Julkisuudessa seurataan useimmiten kuukausittain työllisyyden ja työttömyyden muutoksia edellisen vuoden vastaavasta kuukaudesta. Vaihtoehtoisesti voidaan käyttää kausitasoitettuja lukuja, jolloin tilannetta voidaan verrata edelliseen kuukauteen.

Esimerkki: Terveys 2000

- Terveys 2000 -tutkimuksen tavoite oli tuottaa ajankohtainen kattava kuva työikäisen ja iäkkäään väestön terveydestä ja toimintakyvystä selvittämällä tärkeimpien terveysongelmien yleisyyttä ja sitä 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 ryväatosos terveyskeskuspiireistä.
 - Ositus perustui yliopistosairaaloiden vastuualueiden väestömäärään suhteutettuun kiintiöintiin.
 - Suurimmat 15 terveyskeskuspiiriä poimittiin otokseen ja lopuista 65:stä piiristä poimittiin loppuotos kussakin ositteessa systemaattisella (PPS) otannalla (sisältymistodennäköisyys suhteessa alkion kokoon).

5.7 Otannan haasteita vielä kootusti

- **Poimintaharha:** Otos ei edusta populaatiota. Vaarana varsinkin silloin, kun otokseen tulleet populaation alkiot ovat valikoituneet tai ovat itse valinneet itsensä otokseen. Vastaavasti toisinaan otoksen peitto ei ole hyvä eli tällöin otanta ei kata koko perusjoukkoa tai se kattaa perusjoukon ja vähän muutakin.
 - Jos television ajankohtaisohjelma pyytää katsojia twiittaamaan mielipiteensä ajankohtaisesta asiasta, kyseessä on itse valikoituvia näyte (osallistujat valitsevat itse itsensä).
- Jos poimitaan tutkimukseen ne perusjoukon alkiot, jotka ovat tutkimuksen

tekemishetkellä ‘saatavilla’, niin kyseessä on **näyte**. Näyte ei siis kata ilmiön koko vaiotelua edustavan satunnaisotoksen tapaan.

– Esimerkiksi perinteiset katukyselyt eivät ole hyvä otantatapa, sillä kadulla liikkujat eivät välttämättä kovin hyvin edusta tutkittavaa perusjoukkoa, ellei perusjoukkona ole kyseisellä kadulla kyseiseen aikaan liikkuvat ihmiset.

- **Vajaapeittävyys:** Populaation alkioista ei ole välttämättä täydellistä luetteloa
- **Vastauskato:** Tutkimuksen kohteita ei tavoiteta tai he kieltyyväät vastaamatta. Kadon vuoksi lopullinen otoskoko saattaa jopa karsiautua pois tai jokin osajoukko on alioidustettuna.
- **Vastausharha:** Kysymykset voivat olla huonosti muotoiltuja tai vastajat voivat antaa väärää tietoja.

100LUKU 5. TILASTOLLISET AINEISTOT, NIIDEN KERÄÄMINEN JA MITTAAMINEN

Luku 6

Otokset ja otosjakaumat: tilastollisen päätelyn näkökulma

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

6.1 Satunnaisotos, yhteisjakauma ja tilastollinen malli

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

102LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

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

Satunnaisotos

Olkoot Y_1, \dots, Y_n riippumattomia ja samoinjakautuneita satunnaismuuttuja, joiden tiheysfunktiota (tf., tai pistetodennäköisyysfunktiota (ptnf)) merkitään $f(y, \theta)$:llä, jossa y :n on yksittäisen sm:jan Y reaalisaatio ja θ on jokin jakauman muodon määrävä parametri (tai parametrit). Parametrin θ arvoa ei yleensä tunneta ja tavoitteena onkin päättää, **estimoida**, sen arvo lopulta käytettävässä olevasta aineistosta.

Satunnaisotoksen tilastollinen malli

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

6.1. SATUNNAISOTOS, YHTEISJAKAUMA JA TILASTOLLINEN MALLI103

tekijöiden määrän tulee olla mahdollisimman vähäinen, ts. tilastotieteessä menetelmien (mallien) tulee olla mahdollisimman yksinkertaisia, mutta silti riittäviä.

- Tämä periaate ja sen suhde ns. **varianssin ja harhan väliseen kompromissiin** on erityisen tärkeä erityisesti tilastollisen ennustamisen ja viime vuosikymmeninä yleistyneen tilastollisen (kone)oppimisen sovellutuksissa (ks. tarkemmin alaluku 3.3 ja luku 6).
- Oletetaan, että Y_1, \dots, Y_n ovat aiempien oletusten pätissä riippumattomia sm:jia ja että ne muodostavat satunnaisotoksen jakaumasta, jonka odotusarvo on μ ja varianssi on σ^2 .
 - Ts. oletamme

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

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

Esimerkki: satunnaisotos normaalijakaumasta

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

- Merkintä $\perp\!\!\!\perp$ tarkoittaa, että sm:t Y_1, \dots, Y_n ovat riippumattomia ja samoin jakautuneita (toisinaan myös lyhyesti *iid* tai *i.i.d.*, joka tulee englannin kielen ilmaisusta “independent and identically distributed”). Merkintä soveltuu käytettäväksi muidenkin jakaumien tapauksessa.
- Esimerkiksi R-ohjelmassa voidaan generoida 10 havainnon ($n = 10$) satunnaisotos standardoidusta normaalijakaumasta (ts. $Y_i \sim N(0, 1), i = 1, \dots, 10$) komennolla `rnorm(10)`.

Esimerkki: miesten pituus

- Kerätään havaintoja miesten pituuksista yksinkertaisella satunnaisotannalla (takaisinpalauttaen) n kappaletta.
- Tällöin havaintoarvoja Y_1, \dots, Y_n voidaan pitää riippumattomina

satunnaismuuttujina, joista jokainen noudattaa tehdyn jakaumaoletuksen mukaan normaalijakaumaa $N(\mu, \sigma^2)$.

- Estimoinnin tehtäväänä on muodostaa parhaat mahdolliset arviot parametreille μ ja σ^2 , ja mahdolisesti testata esimerkiksi odotusarvollle μ asetettua hypoteesia.

6.2 Otosjakauma: Estimaattori ja estimaatti

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

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

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

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

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

- Esimerkiksi keskiarvo on havaittujen arvojen keskiarvo, kun se lasketaan kerätystä aineistosta.
- Jos tunnuslukua T käytetään tilastollisen mallin parametrin (parametri) θ estimointiin, niin tästä sanotaan tällöin parametrin **estimaattoriksi**.
 - Estimaattorin otoskohtaisia arvoja, kuten yllä t , kutsutaan **estimaatteiksi**.
 - Toivottavaa olisi, että estimaatit $t = g(y_1, \dots, y_n)$ osuisivat mahdollisimman lähelle tunnusluvun todellista arvoa $g(\theta)$. Ts. satunnaismuuttujan eli tässä tapauksessa estimaattorin $T = g(Y_1, \dots, Y_n)$ jakauman tulisi keskityä mahdollisimman tiiviisti $g(\theta)$:n ympärille.
- Koska tunnusluku/estimaattori T on satunnaismuuttuja, sillä on todennäköisyysjakauma, jota kutsutaan tunnusluvun T **otosjakaumaksi**.
 - Otosjakauma muodostaa (tilastollisen mallin) todennäköisyysmallin tunnusluvun T arvojen satunnaisvaihtelulle otoksesta toiseen.
 - 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 yllä jo todettiin, havaintoaineesta käyttäen.

Estimaattorin ominaisuudet

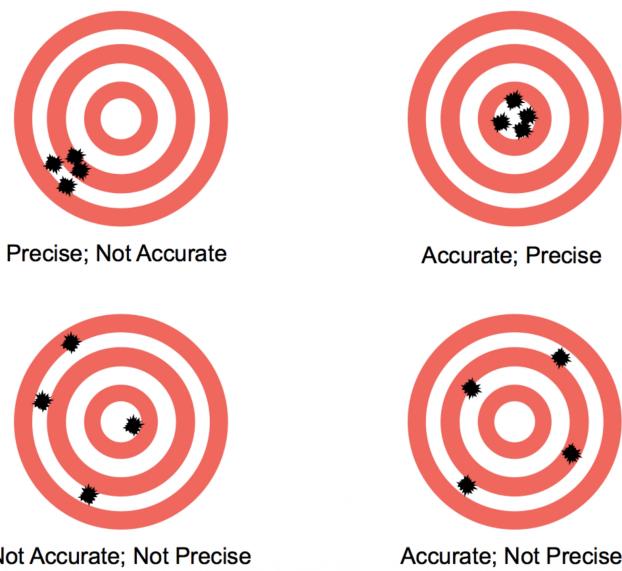
- Merkitään seuraavassa parametrin θ estimaattoria $\hat{\theta}$:lla ja siltä voidaan toivoa seuraavia ominaisuuksia:

Harhattomuus

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

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

106LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA



Kuva 6.1: Harhaton estimaattori

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 $\text{Var}(\hat{\theta}^{(1)}) \leq \text{Var}(\hat{\theta}^{(2)})$.

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

6.3 Otoskeskiarvo ja otosvarianssi (estimaattoreina)

Otoskeskiarvo

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

108LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

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

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

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

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

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

- Täten otoskeskiarvo kuvaaa aineiston perusjoukon tilastollisen mallin odotusarvoa.

Aritmeettisen keskiarvon ominaisuuksia

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

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

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

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

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

Otosvarianssi

- Aineiston sisältämää vaihtelua kuvataan **otosvarianssilla**

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

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

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

jota otosvarianssi harhattomasti estimoii.

- Huomioi, että **otosvarianssi** on eri asia kuin **otoskeskiarvon varianssi**.
- Otoskeskiarvo \bar{Y} ja otosvarianssi S^2 ovat siis satunnaismuuttuja, joiden saamat arvot vaihtelevat satunnaisesti otoksesta toiseen.

Normaalijakautunut otos

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

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

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

Standardoidun aritmeettisen keskiarvon otosjakauma

- Tarkastellaan **standardoitua** satunnaismuuttujaa

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

- Tällöin Z :n odotusarvo $E(Z) = 0$ ja varianssi $\text{Var}(Z) = 1$.
 - Jos $Y_i \sim N(\mu, \sigma^2), i = 1, \dots, n$, niin tällöin Z noudattaa standardoitua normaalijakaumaa:
- $$Z \sim N(0, 1).$$
- Jälleen voidaan osoittaa, että tämä tulos pätee asymptoottisesti (suurissa otoksissa) myös ilman yllä tehtyä normaalisuusoletusta.

6.4 Suhteellisen frekvenssin otosjakauma

Frekvenssi ja suhteellinen frekvenssi

- Oletetaan, että tapahtuman A todennäköisyys on

$$P(A) = p,$$

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

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

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

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

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

Frekvenssin otosjakauma

- Frekvensillä f on odotusarvo

$$E(f) = np,$$

ja varianssi

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

- Frekvenssi f noudattaa binomijakaumaa parametrein n ja p :

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

Suhteellinen frekvenssi: Odotusarvo ja varianssi

- Suhteellisen frekvenssin \hat{p} odotusarvo

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

ja varianssi

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

- Suhteellisen frekvenssin \hat{p} standardipoikkeamaa

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

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

Suhteellisen frekvenssin otosjakauma

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

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

- Aritmeettisen keskiarvon tapaan standardoitut sm.

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

noudattaa suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa.

EU-kansanäänestys

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

kannatusosuuden jakauma olisi

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

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

- Nän ollen haluttu todennäköisyys (ts. saada sellainen satunnaismuuttujan $Z \sim N(0, 1)$ arvo että suhteellinen osuus on pienempi kuin 0.5)

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

6.5 Muita tunnuslukuja

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

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

- **Moodi:** Moodi eli tyyppiarvo on havaintoaineiston yleisin muuttujan arvo tai se on luokka, jolla on suurin frekvenssi.
- **Mediaani:** Mediaani on järjestetyn havaintoaineiston keskimmäinen arvo (jos havaintoarvoja on pariton määrä, parillisessa tapauksessa esitetään jompikumpi keskimmäisistä arvoista). Mediaani siis jakaa järjestetyn havaintoaineiston kahteen osaan siten, että puolet arvoista on mediaania pienempiä ja puolet arvoltaan mediaania suurempia.
 - Luokittelustaikolla mitattaville muuttujille ei ole olemassa luontevia sijaintilukuja keskilukujen yhteydessä pl. moodi.
- Järjestysasteikolla mitatuille muuttujille voidaan mediaanin lisäksi määrittää **fraktiileja**: pp%:n fraktiili jakaa tilastoaineiston kahteen osaan siten, että kyseistä fraktiilia pienempiä havaintoarvoja on pp%.
 - Eniten käytettyjä fraktiileja ovat **kvartiilit**. **Alakvartiili** Q_1 on 25 %:n fraktiili, ja **yläkvartiili** Q_3 on 75 % fraktiili.

- Tietystä fraktiileista käytetään nimitystä **desili**. Ensimmäinen desili D_1 on 10 % fraktiili ja esim. yhdeksäs fraktiili D_9 on 90 % fraktiili.
- Hajontalukuja: Varianssin/keskijonnan lisäksi, jos muuttuja on mitattu vähintään järjestysasteikolla, sille voidaan määrittää vaihteluväli ja kvartiiliväli. **Vaihteluväli** kuvaa aineiston kokonaispeittoa ja siinä ilmoitetaan aineiston pienin havainto ja suurin havainto. Ts. vaihteluväli=(pienin havainto, suurin havainto). **Kvartiiliväli** = (Q_1, Q_3) .
- Muita tunnuslukuja: Tilastollisen päätöksenteon yhteydessä käytettäviä tunnuslukuja ovat **vinous** ja **huipukkuus**. Vinous ja huipukkuus voidaan määrittää välimatka- ja suhdeasteikon muuttujille. Vinous ja huipukkuus mittaaavat kumpikin omalla tavallaan jakauman poikkeamaa normaalijakaumasta. Normaalijakauman vinous on 0 ja huipukkuus on 3.

6.6 Luottamusvälit

- Satunnaisesti saadusta aineistosta laskettujen tunnuslukujen luotettavuus on tilastollisen mallin parametrien estimoinnissa keskeinen tilastollinen kysymys.
 - Otoksen poimintaan liittyvän satunnaisvaihtelon vuoksi emme voi varmuudella tietää onko saatu otokseen perustuva parametriestimaatti ”lähellä” vai ”kaukana” sen todellisesta arvosta.
 - Täten tarvitaan jokin tapa, jolla saadun parametrestimaatin luotettavuutta voidaan arvioida.

Luottamusväli

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

- Otoskoko on luottamusvälejä koskevissa tarkasteluissa keskeinen ja luottamusväleihin palataankin otoskoon käsittelyn yhteydessä.
- Valittua luottamustasoa merkitään usein $1 - \alpha$:lla, jossa **merkitsevyystaso (riskitaso)** α on esimerkiksi $\alpha = 0.05$.
- Tulkinta: Jos **otantaa** jakaumasta $f(y; \theta)$ toistetaan, niin keskimäärin $100 \times (1 - \alpha)\%$ otoksista kontstruloiduista luottamusväleistä peittää parametrin θ todellisen arvon.

114 LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- Oletetaan, että olemme tehneet johtopäätöksen, että konstruloitu luottamusväli peittää parametrin θ tuntemattoman todellisen arvon.
 - Tällöin otantaa toistettaessa luottamusvälin konstruktiosista seuraa, että tehty johtopäätös on oikea keskimäärin $100 \times (1 - \alpha)\%$ tapauksista.
 - Vastaavasti taas $100 \times \alpha\%$ ei peitä parametrin todellista arvoa.
- Luottamusväli on kenties tunnetumpi kansankieliseltä nimitykseltään **virhemarginaali**, joka on itseasiassa luottamusvälin puolikas: todellinen parametriarvo kuuluu saadun estimaatin ja virhemarginaalien sisään jäävälle osuudelle.
 - Normaalisti mm. otoskoon kasvu pienentää virhemarginaalia.
 - Kuten jatkossa tullaan havaitsemaan, virhemarginaalin suuruuteen vaikuttavat otosasetelma, otoskoko, luottamustaso ja tutkittavan tilastollisen tunnusluvun jakauma.
- Luottamusväleissä ei kuitenkaan varsinaisesti ole kyse “virheestä” vaan saadun/muodostetun tiedon tarkkuudesta.
 - Luottamusvälit, eli virhemarginaalit, siis (yleisesti) riippuvat valitavasta luottamustasosta $1 - \alpha$ ja näin ollen samasta aineistosta on saatavissa useita virhemarginaaleja.* Täten on tarkalleen ottaen virheellistä sanoa, että “tutkimuksen virhemarginaali on 3,5 puoleen tai toiseen”.* Oikeammin olisi sanoa esimerkiksi “tutkimuksessa saadun kananatuksen virhemarginaali on 3,5 puoleen tai toiseen 95 % luottamustasolla.”* Virhemarginaali kasvaa, kun aineistoa lohkotaan: jos tuhannen hengen otoksesta esitetään tietoja, jotka kuvaavat erikseen miesten ja naisten ominaisuuksia, sukupuolittain lasketut ovat estimaatit epävarmempia kuin koko otoksesta esitetyt.
 - Vastaavasti on virheellistä sanoa että tutkimuksella olisi virhemarginaali, sillä virhemarginaali liittyy aina vain tutkimuksen antamiin numeroisiin arvoihin.
 - Aitoja virhelähteitä ovat mm. otantatutkimukseen liittyvien kysymysten muotoilu, käsitteiden monitulkintaisuus, vastaajien valikointuminen ja vastauskato.

Normaalijakauman odotusarvon luottamusväli

- Käsittelemme seuraavassa (normaalijakauman) odotusarvon μ luottamusvälejä ja jatkossa oletetaan (ellei toisin mainita), että taustalla oleva populaatio, N , on “iso” (ääretön).

- Näin ollen ns. äärellisyyskorjausta ei käytetä (yksinkertaisuuden vuoksi).
- Tarkastellaan satunnaisotosta normaalijakaumasta $Y_1, \dots, Y_n \perp\!\!\!\perp, Y_i \sim N(\mu, \sigma^2), i = 1, \dots, n.$
- Tarkastellaan normaalijakauman odotusarvon μ luottamusvälin määräämistä otannan avulla olettaen että jakauman varianssi σ^2 on tunnettu.
 - Muistetaan että normaalijakauman odotusarvoparametrin $E(Y_i) = \mu$ **harhaton estimaattori** on aritmeettinen keskiarvo

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

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

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

jossa standardoitu satunnaismuuttuja

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

(ks. aiemmat alaluvut tässä luvussa 6) noudattaa $N(0, 1)$ -jakaumaa.

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

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

- Joka voidaan kirjoittaa uudelleen muodossa

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

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

116 LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

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

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

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

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

- Luottamusvälin pituus

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

- **Virhemarginaali** on luottamusvälin pituuden puolikas eli

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

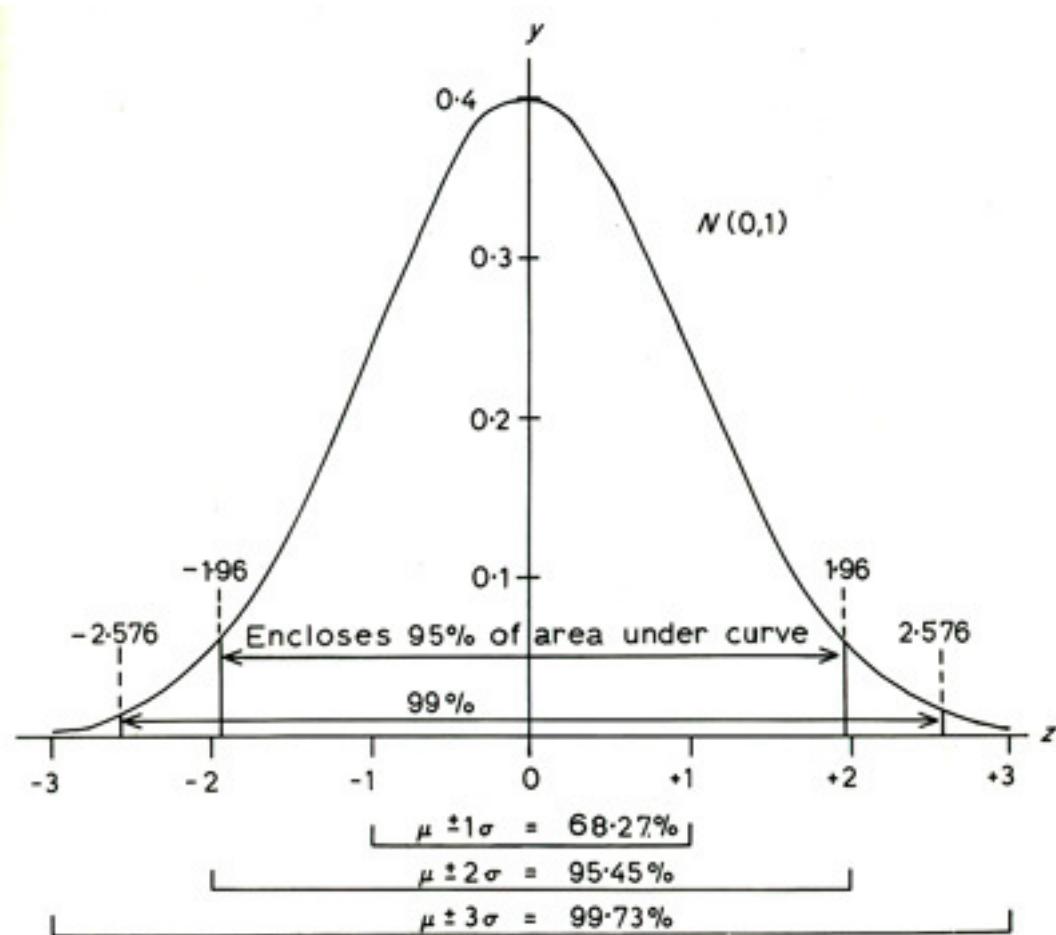
- Edellä tiettyyn otokseen liittyvä luottamusväli perustetaan realisoituneeseen otoskeskiarvoon $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.
- Olisi toivottavaa pystyä konstruoimaan parametrille μ mahdollisimman lyhyt luottamusväli, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman korkea. 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 lyhtenäminen pienentää luottamustasoa, jolloin tieto parametrin μ todellisesta arvosta tulee epävarmemmaksi.

Normaalijakauman odotusarvon luottamusväli (σ^2 tuntematon)

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

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

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



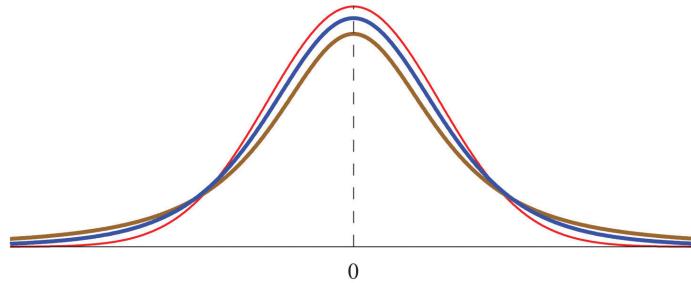
Kuva 6.2: Standardoitu normaalijakauma: Virhemarginaaleja

- (Studentin) t -jakauma muistuttaa silmämäääräisesti normaalijakamaa, mutta se on paksuhäntäisempi. Vapausasteluvun kasvaessa t -jakauma lähestyy normaalijakaumaa.
- Suurissa otoksissa (n iso) luottamuskerroimet voidaan poimia (approksimatiivisesti) myös normaalijakaumasta eli korvata edellä kerroimet $t_{\alpha/2}$ aiemmin käytettyillä kertoimilla $z_{\alpha/2}$.
- Normaalijakauman odotusarvon luottamusväli (σ^2 tuntematon), t -jakauma eri vapausastein df

Standard normal

t -distribution with $df = 5$

t -distribution with $df = 2$



Kuva 6.3: t -jakauman (ja standardoidun normaalijakauman) tiheysfunktioita

Luottamusväli: Suhteellisen osuuden odotusarvo

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

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

- Bernoulli-jakauman (vrt. binomijakauma) ominaisuuksien perusteella $E(Y_i) = p$ ja $\text{Var}(Y_i) = pq$, jossa $q = 1 - p$.

- Nämä ollen voimme normaalijakauman odotusarvoparametrin luottamusvälin konstruoinnin tapaan määritellä satunnaismuuttujan Z :

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

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

- Suhteellisen frekvenssin hajonnan estimaattori on siis

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

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

- Luottamuskertoimet määritetään aiempaan tapaan:

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

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

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

- Luottamusväli voidaan kirjoittaa

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

ja luottamusvälin pituus on

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

6.7 Otoskoko

- Yksi tilastollisen päätöslauselman keskeisiä tavoitteita on yleistää otoksen pohjalta tehty päätöslauselmaan koko perusjoukko. Seuraavaksi käymme läpi seikkoja, jotka tulee ottaa huomioon otoskokoa miettiessä.
- Kun on päätetty, millainen tutkimusaineisto halutaan kerätä, on päätetään, kuinka suuri otoksen on oltava, jotta se edustaa tutkittavaa joukkoa kattavasti.
 - Liian pieni **otoskoko**, eli pieni määrä otokseen poimittuja tilastoyksiköitä, voi **sattumalta** poiketa paljonkin perusjoukosta.

120LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

- * Tämä niin kutsuttu **otantavirhe** on sitä suurempi mitä pienemppää otosta käytetään.
- * Liian pienen otoksen vuoksi muuten hyvin toteuttu tutkimusja otanta-asetelma saattaa epäonnistua vastaamaan tutkimuksen mielenkiinnon kohteena olevaan kysymykseen.
- Todella suuren otoksen koostaminen voi olla **työlästä, kallista** tai joskus jopa täysin mahdotonta esimerkiksi siksi että käytettävissä olevat tutkimusyksiköt eivät ole käytettävissä ajallisten rajoitteiden vuoksi (kuten harvinaisten tautien kantajat).
 - * Toisaalta perusjoukon systemaattiset piirteet tulevat otoskoon kasvaessa paremmin esille, vaikka yksittäisten otosyksiköiden tilastolliset muuttujat saattavat vaihdella suuresti.
- **Otoskoko** siis vaikuttaa keskeisesti siihen, miten hyvin otoksesta tehdyt johtopäätökset voidaan yleistää koskemaan koko perusjoukkoa!
- Optimaalinen, tai ainakin tutkimusongelmaan vastaamisen kannalta vähintään riittävä arvio, otoskoosta voidaan usein määräätä etukäteen.

Perusjoukon rooli otoskoon määrittämisessä

- Ensiksi tulee kuitenkin pohtia käsillä olevaa tutkimusongelmaa esimerkiksi kysymällä: **Millainen on perusjoukkosi?**
 - Onko tutkittavan muuttujan arvoissa paljon vaihetlua? Jos on, niin tämä täytyy huomioida kasvattamalla otoskokoa.
 - * Esimerkiksi otosten keskiarvot alkavat käyttäytyä riittävän siistiä vasta noin otoskoosta $n = 30$ alkaen.
 - * Kyseinen otoskoko ei kuitenkaan ole missään tapauksessa yleinen ja pätevä peukalosääntö otoskoon koolle, vaan se tulee aina päättää tutkimusongelmakohtaisesti.
 - Kuuluuko tutkimukseesi esimerkiksi otoksen sisällä olevien ryhmien keskiarvojen vertailua tai muita otoksen 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 otoskokoaa voidaan approksimoida tulosten halutun tarkkuuden avulla.

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

Odotetun vastauskadon vaikutus otoskokoon

- Kuinka suuri vastauskato tulee mahdollisesti olemaan?
 - Yleensä osa kyselytutkimukseen valituista jättää vastaamatta. Tätä kutsutaan kadoksi. Kato vinouttaa otosta, jos vastaamatta jättäneet ovat mielipiteiltään erilaisia kuin vastanneet.
 - Otoskoon kasvattaminen ei paranna kadon aiheuttamaa vinoutumista.
- Esimerkki: Jos Alkon myymälän asiakastutkimus suoritetaan 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 asiakkaille.

Esimerkki: otoskoko normaalijakauman odotusarvon estimoinnissa

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

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

122 LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

ja 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 $2d$ (huom. symmetrisyys).
- Luottamusvälin lausekkeesta saadaan täten järjestelemällä

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

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

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

- Pitäädytään (yksinkertaisuuden vuoksi) luottamuskertoimissa $z_{\alpha/2}$ vaikka varianssi σ^2 olisikin tuntematon.

Esimerkki: otoskoko

- Oletetaan, että haluamme määräätä 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:

- Äärellisyyskorjausta käytetään, jos otos poimitaan äärellisestä perusjoukosta.

kosta palauttamatonta ja (nyrkkisääntöä)

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

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

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

Otoskoko, äärellisyyskorjaus: Normaalijakauman odotusarvon estiointi

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

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

- Tarvittava otoskoko on tällöin välivaiheiden jälkeen

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

Esimerkki: otoskoko (jatkoa)

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

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

- Tarvittava otoskoko on siis noin 17.

Otoskoko: Suhteellinen osuus

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

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

Suhteellinen osuus, äärellisyyskorjaus:

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

Esimerkki: Otoskoko ja suhteellinen osuus

- Geologi haluaa arvioida kallion kultapitoisuuden ottamalla kivinäytteen n eri pisteestä. Jokaisesta näytteestä havaitaan sisältyykö siihen kultaa. Kuinka suuri otos on poimittava, jotta kultapi-

toisuuden estimointivirheen d arvo on korkeintaan 0.05 todennäköisyydellä 0.95?

- Tässä kullan suhteellinen osuus on tuntematon, joten p :lle asetetaan $p = 0.5$.
- Äärellisyyskorjaus voidaan unohtaa, sillä näytteenottopisteiden pinta-alat ovat pieniä (eli niitä on äärettömän paljon, ts. tarkasteltavaan populaatioon niitää sisältyy hyvin suuri määrä).
- Tällöin otoskoko

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

126 LUKU 6. OTOKSET JA OTOSJAKAUMAT: TILASTOLLISEN PÄÄTTELYN NÄKÖKULMA

Luku 7

Tilastollinen riippuvuus ja korrelaatio

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

7.1 Muuttujien väliset riippuvuudet

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

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

Esimerkkejä riippuvuustarkasteluista

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

• Eksakti vs. tilastollinen riippuvuus

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

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

7.2 Kahden muuttujan havaintoaineiston kuvaaminen

- Kuten yhden muuttujan havaintoaineistojen tapauksessa, lähtökohdan kahden tai useaman muuttujan havaintoaineistojen kuvaamiselle muodostaa tutustuminen havaintoarvojen jakaumaan.
- Havaintoarvojen jakaumaa voidaan kuvilla ja esitellä tiivistämällä havaintoarvoihin sisältyvä informaatio sopivan muotoon:
 - Havaintoarvojen jakaumaa kokonaisuutena voidaan kuvata sopivasti valituilla graafisilla esityksillä.
 - Havaintoarvojen jakauman karakteristisia ominaisuuksia voidaan kuvata sopivasti valituilla otostunnusluvuilla (ks. otostunnusluvut ja otosjakaumat luvussa 6).
- Koska useampi- kuin kaksiulotteisten kuvioiden tekeminen ei ole usein kovin mielekästä, kolmen tai useaman muuttujan havaintoaineistoja havainnollistetaan tavallisesti niin, että muuttuja tarkastellaan pareittain.
- Kahden järjestys-, välimatka- tai suhdeasteikkoillisen muuttujan havaittujen arvojen paraja havainnollistetaan tavallisesti graafisella esityksellä, jota kutsutaan hajontakuvioksi tai pistediagrammiksi (“pistekaavio” engl. scatter plot). Ks. esimerkiksi kuva 7.1.

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

Pistediagrammi (hajontakuvio)

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

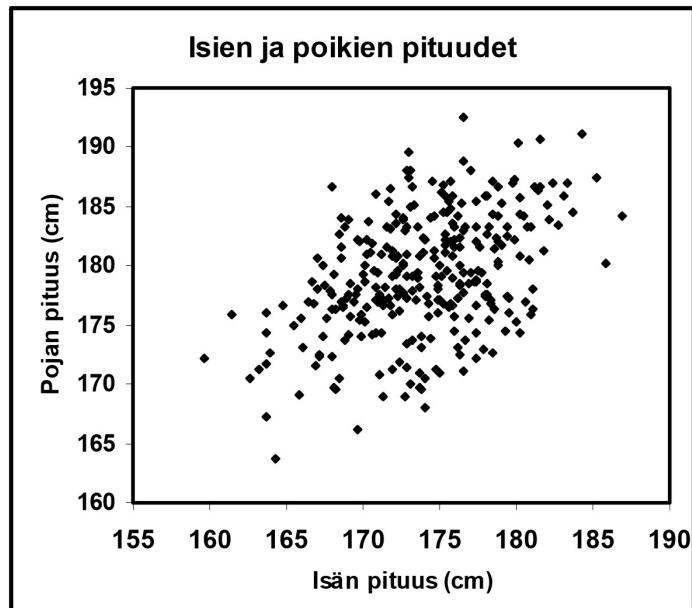
Pistediagrammi

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

Esimerkki: Isän ja pojantulokset

- Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.
- Periytyykö isän tulokset heidän pojilleen?
- Havaintoaineisto koostuu 300:ta isän ja heidän poikiensa tuluksienvien muodostamasta lukuparista (x_i, y_i) , $i = 1, 2, \dots, 300$, jossa $x_i = \text{isän } i$ tulokset ja $y_i = \text{isän } i$ pojantulokset.

- 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 pitempia poikia kuin lyhyillä isillä.
- Tällaisten tilastollisten riippuvuuksien analysoimista lineaaristen regressiomallien avulla tarkastellaan myöhemmin luvussa 8 Yksinkertainen lineaarinen regressiomalli.



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

7.3 Tunnusluvut

- Kahden välimatka- tai suhdeasteikollisen muuttujan havaintoarvojen parien muodostamaa jakaumaa voidaan karakterisoida seuraavilla tunnusluvuilla:
 - Havaintoarvojen keskimääräistä sijaintia kuvataan aritmeettisilla keskiarvoilla.
 - Havaintoarvojen hajaantuneisuutta tai keskityneisyyttä kuvataan keskihajonnoilla tai (otos-) variansseilla.

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

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

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

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

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

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

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

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

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

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

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

7.4 Satunnaismuuttujien kovarianssi ja korrelaatio

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

Satunnaismuuttujien kovarianssi ja korrelaatio

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

Olkoot

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

satunnaismuuttujien X ja Y odotusarvot ja

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

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

satunnaismuuttujien X ja Y varianssit.

Määritellään satunnaismuuttujien X ja Y kovarianssi σ_{XY} kaavalla

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

Määritellään satunnaismuuttujien X ja Y korrelaatio ρ_{XY} kaavalla

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

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

- Satunnaismuuttujien X ja Y korrelaatiota

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

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

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

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

voidaan estimaoida Pearsonin **otoskorrelatiokertoimella** (ks. alla).

- Estimaattori r_{XY} voidaan johtaa sekä momenttimenetelmällä että suurimman uskottavuuden menetelmällä, jotka ovat tyypillisiä estimointimenetelmiä tilastotieteessä ja tarkemmin tilastollisessa päätyssä.

Pearsonin otoskorrelatiokerroin

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

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

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

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

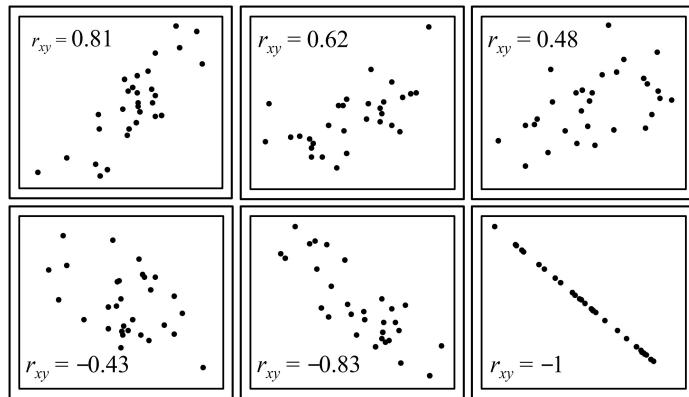
Ts. havaintoarvojen pareista $(x_i, y_i), i = 1, 2, \dots, n$ laskettu Pearsonin otoskorrelatiokerroin on siis

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

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

- Otoskovarianssi:

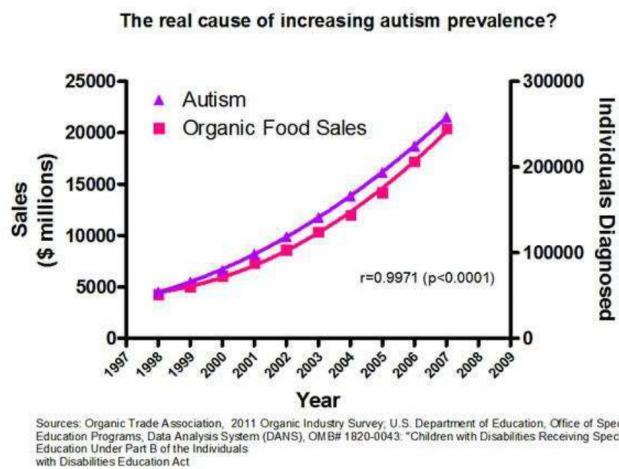
- Huomaa, että x - ja y -havaintoarvojen otoskovarianssit niiden itsensä kanssa ovat niiden variansseja.
- Otoskovarianssi s_{xy} mittaa x - ja y -havaintoarvojen yhteisvaihtelua niiden aritmeettisten keskiarvojen ympärillä.
- Otoskovarianssilla on taipumus saada positiivisia (negatiivisia) arvoja, jos havaintopisteiden muodostama ”pistepilvi (pisteparvi)” näyttää nousevalta (laskevalta) oikealle mentäessä; ks. pistediagrammin ilmeen ja Pearsonin otoskorrelatiokertoimen yhteys, jota käsitellään seuraavaksi.
- Pearsonin otoskorrelatiokertoimella r_{xy} on seuraavat ominaisuudet:
 - i) $-1 \leq r_{xy} \leq 1$
 - ii) $r_{xy} = \pm 1$, jos ja vain jos $y_i = \alpha\beta x_i$, jossa α ja β ovat reaalisia vakiota ja $\beta \neq 0$
 - iii) Korrelatiokertoimella r_{xy} ja kovarianssilla s_{xy} on aina sama etumerkki
- Pearsonin otoskorrelatiokerroin r_{xy} : Tulkinta/tulkintoja:
 - Havaintoarvojen pareista $(x_i, y_i), i = 1, 2, \dots, n$ laskettu Pearsonin otoskorrelatiokerroin r_{xy} mittaa x - ja y -havaintoarvojen lineaarisen tilastollisen riippuvuuden voimakkuutta.
 - Jos $r_{xy} = \pm 1$, niin x - ja y -havaintoarvojen välillä on eksakti eli funktioalinen lineaarinen riippuvuus, mikä merkitsee sitä, että kaikki havaintopisteet (x_i, y_i) asettuvat samalle suoralle.
 - Jos $r_{xy} = 0$, niin x - ja y -havaintoarvojen välillä ei voi olla eksaktia lineaarista riippuvuutta.
 - Vaikka $r_{xy} = 0$, x - ja y -havaintoarvojen välillä saattaa silti olla jopa eksakti epälineaarinen riippuvuus.
- **Havainnollistus:** Alapuolella esitettävät kuviot havainnollistavat kahden muuttujan havaittujen arvojen ($n = 30$) pistediagrammin ilmeen ja korrelaation välistä yhteyttä.
 - Toinen havainnollistus: Ks. seuraavasta [linkistä](#) lisää havainnollistuksia.
 - *Guess the correlation* pelissä pääset arvioimaan esitettävän pisteparen korrelaation voimakkuutta erilaisissa simuloiduissa tilanteissa: <http://guessthecorrelation.com/>
 - **Kausaalisuus**
 - Muuttujan x arvojen muutos vaikuttaa muuttujan y arvoihin (syvaikutussuhde), jos seuraavat kolme ehtoa täytyvät:
 - * muuttujan x muutos esiintyy ajallisesti ennen y :n muutosta
 - * muuttujissa x ja y tapahtuvien muuttujien välillä on riippuvuutta



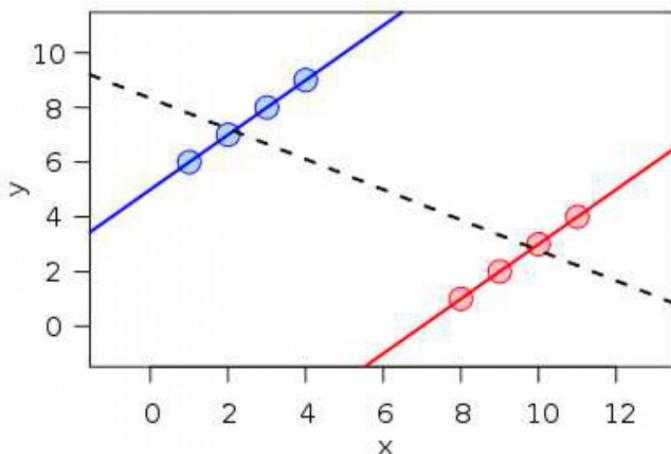
Kuva 7.2: Havainnollistuksia Pearsonin otoskorrelaatiokertoimen arvosta ja erilaisista xy -pisteparvista. Lähde: Mellin (2006).

* muuttujassa y tapahtunutta muutosta ei voida selittää millään muilla tekjöillä

- Kausaalisuhteita selvitettäessä on tunnettava etukäteen ilmiötä koskevat aiemmat teoriat ja tutkimukset tarkasti, jotta voidaan ottaa huomioon ilmiöön vaikuttavat tekijät
- Todellisuus on usein monimutkaisempi, kuin mitä kausaalisuude kuvaaa: **kahden muuttujan yhteisvaihtelu ei riitä todisteeksi siitä, että kyseessä olevien muuttujien välillä on kausaalista yhteyttä**
- Yhteisvaihtelu voi johtua myös kolmannen muuttujan vaikutuksesta molempien muuttujien tai virheellisestä otannasta, vaikka muuttujat olisivatkin perusjoukossa toisistaan riippumattomia
- Simpsonin paradoksi
 - Simpsonin paradoksi syntyy, kun kahden muuttujan välinen korrelaatio muuttuu päinvastaiseksi, otettaessa huomioon jokin kolmas muuttuja, joka korreloii molempien muuttujien kanssa



Kuva 7.3: Esimerkkejä: luomuruoka syypää lisääntyneisiin autismitapauksiin?



Kuva 7.4: Simpsonin paradoksi

Esimerkki: Berkeleyn sukupuolisyrjintä

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

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

Tiedekunta	Miehet		Naiset	
	Hakijat	Hyväksytyt %	Hakijat	Hyväksytyt %
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

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

Luku 8

Regressioanalyysi

- 8.1 Johdatus regressioanalyysin ideaan
- 8.2 Yhden selittäjän lineaarinen regressiomalli
- 8.3 Muita regressiomalleja

Luku 9

Tilastotieteen rooli uuden tiedon tuottamisessa

9.1 Tilastollisen tutkimuksen yhteisiä elementtejä

9.2 Tutkimusprosessi

144 LUKU 9. TILASTOTIETEEN ROOLI UUDEN TIEDON TUOTTAMISESSA

Luku 10

Aineisto- ja tutkimustyyppit ja koeasetelmat

10.1 Tutkimustyyppit

10.2 Tutkimusstrategiat

10.3 Erilaisia aineistoja ja aineistolähteitä

Luku 11

Tilastollisesta ennustamisesta

- 11.1 Tilastollinen selittäminen vs. ennustaminen**
- 11.2 Tilastolliseen ennustamiseen liittyviä huo-
mioita**

Luku 12

Tilastotieteen kehityksen nykytrendejä