

TILM3701 - Tilastotiede ja data 2022

Koonneet

Henri Nyberg¹

Roope Rihtamo²

2022-08-26

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 Tilastojen yleisestä roolista yhteiskunnassa	23
2.4 Mitä on tutkimus?	24
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	48
3.6 Tilastotieteen sovelluskohteita ja “rajatieteitä”	53

4	Sattuma ja satunnaisuus tilastotieteessä	55
4.1	Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä	55
4.2	Satunnaisuus ja todennäköisyydet	55
4.3	Tilastolliset mallit, jakaumat ja parametrit	55
4.4	Odotusarvo ja varianssi	55
4.5	Joitain jakaumia	55
4.6	Sattuman rooli tieteenteossa: Vale-emävale-tilasto?	55
5	Tilastolliset aineistot, niiden kerääminen ja mittaaminen	57
5.1	Kertausta: Data eli aineisto	57
5.2	Otannan idea	57
5.3	Mittaaminen ja mitta-asteikot	57
5.4	Kontrolloidut kokeet ja suorat havainnot	57
5.5	Otantamenetelmät	57
5.6	Otantaesimerkkejä	57
5.7	Otannan haasteita vielä kootusti	57
6	Otokset ja otosjakaumat: tilastollisen päättelyn näkökulma	59
6.1	Satunnaisotos, yhteisjakauma ja tilastollinen malli	59
6.2	Otosjakauma: Estimaattori ja estimaatti	59
6.3	Otoskeskiarvo ja otosvariassi (estimaattoreinta)	59
6.4	Suhteellisen frekvenssin otosjakauma	59
6.5	Muita tunnuslukuja	59
6.6	Luottamusvälit	59
6.7	Otoskoko	59
7	Tilastollinen riippuvuus ja korrelaatio	61
7.1	Muuttujien väliset riippuvuudet	61
7.2	Kahden muuttujan havaintoaineiston kuvaaminen	61
7.3	Tunnusluvut	61
7.4	Satunnaismuuttujien kovarianssi ja korrelaatio	61

<i>SISÄLLYS</i>	5
8 Regressioanalyysi	63
8.1 Johdatus regressioanalyysin ideaan	63
8.2 Yhden selittäjän lineaarinen regressiomalli	63
8.3 Muita regressiomalleja	63
9 Tilastotieteen rooli uuden tiedon tuottamisessa	65
9.1 Tilastollisen tutkimuksen yhteisiä elementtejä	65
9.2 Tutkimusprosessi	65
10 Aineisto- ja tutkimustyytit ja koeasetelmat	67
10.1 Tutkimustyytit	67
10.2 Tutkimusstrategiat	67
10.3 Erilaisia aineistoja ja aineistolähteitä	67
11 Tilastollisesta ennustamisesta	69
11.1 Tilastollinen selittäminen vs. ennustaminen	69
11.2 Tilastolliseen ennustamiseen liittyviä huomioita	69
12 Tilastotieteen kehityksen nykytrendejä	71

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, mahdollisuuksien mukaan, kovin teknistä matemaattista esitystapaa, mutta tarvittavissa määrin tullaan myös käyttämään tilastotieteen perusopinnoissa tarvittavia matemaattisia merkintöjä ja määritelmiä. Esim. todennäköisyyslaskennan ja tilastollisen päättelyn perusteita ei käydä vielä riittävällä matemaattisella tarkkuudella lävitse, vaan nämä tarkastelut jäävät tätä kurssia seuraavien kurssien (TILM3553 Todennäköisyyslaskennan peruskurssi tai TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille sekä TILM3555 Tilastollisen päättelyn peruskurssi) asiaksi. Nämä kurssit, yhdessä alkuvaiheen pakollisten matematiikan kurssin lisäksi, muodostavat siis tämän kurssin johdannon kanssa lähtökohdan tilastotieteen opinnoille.
- Luennot eivät suoraan perustu yhteen kirjaan tai lähteeseen. Käytettyjä lähdemateriaaleja luetellaan alapuolella oheislukemiston myötä.
- Oheislukemistoa (sopivilta osin):
 - Mellin, I. (2004). Johdatus tilastotieteeseen: Tilastotieteen johdantokurssi (1.kirja). Yliopistopaino, Helsingin yliopisto.
 - Mellin, I. (2000). Johdatus tilastotieteeseen: Tilastotieteen jatkokurssi (2.kirja). Yliopistopaino, Helsingin yliopisto.
 - Mellin, I. (2006). Tilastolliset menetelmät. Luentomoniste, Aalto yliopisto (TKK).
 - Holopainen, M. ja P. Pulkkinen (2008). Tilastolliset menetelmät. Sanoma 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 kommenteista ja avusta materiaalin työstämisessä. Kaikki jäljelle jääneet painovirheet ovat materiaalin koajien.

Kurssimateriaali

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

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

Lisäksi on avattu opinto-oppaassa ja vielä laajemmin alla. Tämä listaus toimii hyvänä luettelona kurssin keskeisistä teemoista.

- Kurssin sisältöä:

- Tilastotiede tieteenalana ja sen suhde lähitieteisiin, kuten datatieteen (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
- Osaamistavoitteet: Opintojakson suoritettuaan opiskelija

Materiaalin seassa on eritelty vääriskoodatuin tietolaatikoin erinäisiä tärkeitä tilastotieteellisiä konsepteja ja termejä sekä esimerkkejä tilastotieteen sovelluksista. Näistä ensin mainitut löytyvät Deltan violeteista laatikoista ja jälkimmäiset Statistikan oransseista.¹ Alla esimerkkilaatikat.

Konsepti tai termi

Konseptin tai termin löyhä määritelmä.

Esimerkki

Aihetta koskeva esimerkki.

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

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 tyyppisissä tutkimusongelmissa.
 - Esitellä tilastotieteen roolia **empiirisen tutkimusaineiston keräämisessä ja analyysissä** sekä tarkastella tieteentekemisen ja tilastotieteen suhdetta.
 - Pohtia **tilastotieteen olemusta tieteenalana** ja tarkastella tilastotieteen ja datatieteiden (data sciencen) samankaltaisuuksia ja eroja.
 - Pohtia **sattuman ja satunnaisuuden roolia** jokapäiväisessä elämässä ja erityisesti osana tieteellistä tutkimusprosessia.
 - Oppia tilastotieteen peruskäsitteitä ja (tilastollisen) tutkimuksenteon alkeita ja siihen liittyviä mahdollisia ongelmia esimerkiksi tilastollisten aineistojen keräämisessä.

- Oppia tilastollisten aineistojen **kuvaamisen ja käsittelyn** alkeita sekä tilasto(tieteellisen)llisen **mallintamisen ja koeasetelmien** peruskäsitteitä.
- Kurssilla käsitellään myös **tilastollisen päättelyn** peruskäsitteitä ja perusteita kuten
 - Mitä on **todennäköisyys** ja miten se tulkitaan tilastotieteessä sekä laajemmin tieteessä. Erityisesti tilastotieteen osalta keskiössä on tämän kurssin osalta **satunnaismuuttujat** sekä niihin liitettävät käsitteet
 - * **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 formaalisessa mallintamisessa ja päättelyssä. Tilastollisen malliin liittyy (usein) **parametreja** joihin tilastollinen päättely kohdistuu.
 - * Tilastollisten mallien **estimoinnin** perusidea, eli miten tilastollisen mallin parametreille muodostetaan arvot käytettävissä olevan aineiston pohjalta. Esimerkiksi: mitä tarkoittaa tilastollisen mallin parametrin **estimaattori** ja sen **harhattomuus**?
 - * Alustavia tarkasteluja tilastollisen mallin uskottavuuden käsitteelle ja **luottamusväleille** tilastollisen mallin estimoiduille parametreille.
- Toinen kurssin keskeisistä teemoista on tarkastella tieteellistä tutkimusprosessia teoriassa ja käytännössä. Tämä sisältää mm. seuraavia aiheita (joita siis käsitellään tällä kurssilla päällisin puolin varsin yleisestä näkökulmasta katsoen ja tarkemmat yksityiskohdat jätetään tätä kurssia seuraavien tilastotieteen kurssien aihepiireiksi):
 - **Tutkimusongelman** asettaminen: mitä halutaan tutkia?
 - Tutkimusongelman täsmentäminen ja **tutkimusstrategian** laatiminen: millä keinoin asetettuun tutkimusongelmaan voidaan vastata?
 - **Tutkimusaineiston** (tai vain lyhyemmin **aineiston** eli **datan**) kerääminen
 - * **Aineiston ennakkoehdot**: mitkä ehdot tulee täyttyä, jotta asetettuun tutkimusongelmaan voidaan vastata?

- * **Otanta** (ja mittaaminen): miten tutkimusaineisto kerätään niin, että se täyttää aineiston ennakkoehdot? Erilaisissa tutkimuksissa käytetään erilaisia aineistoja kuten:
 - Survey- eli haastatteluaineistot: aineisto kerätään haastattelulla 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:** minkälaista 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 ammattikouluissa huolimatta sen keskeisestä ja kasvavasta roolista tieteenteossa.
- Tiedeyhteisön ulkopuolellakin **tilastotiedettä ja tilastotieteilijöitä arvostetaan laajalti**.
- **Tilastotiede onkin nostanut profiliaan viimeisten vuosikymmenien aikana** tietoteknisen kehityksen tuotua laajat tietoaaineistot ja kehittyneet laskennalliset menetelmät lähes jokaisen kansalaisen saataville.
- Tämä "*datavallankumous*" näkyy tilastotieteilijöiden kysynnässä työmarkkinoilla: erilaisten aineistojen määrän lisääntyessä kasvaa myös kysyntä työntekijöistä, jotka osaavat ammatitaitoisesti käsitellä, tulkita ja mallintaa tilastollisia aineistoja.
- Ei siis liene ihmeäkään, että erilaisten "data"-alkuisten työpaikkojen, kuten **datatieteilijä** (eng. **data scientist**) tai **data-analyytikko** (**data analyst**) määrä on kasvanut voimakkaasti jo pidempään. Kaikkia tieto- ja datainensiivisten 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 TY:ssä tilastotieteen opinnoissa syvennyttään** tällä kurssilla esiteltäviin menetelmiin, aineistotyyppeihin ja mallinnuskokonaisuuksiin. Tilastotieteen opintotarjontaan voi perehtyä Turun yliopiston 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 tunnettuus on lisääntynyt mm. tieteen popularisoinnin ja median laajemman tiedeuutisoinnin 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 tiedeuutisia 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 tutkimuskysymysten 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 järkipäristä uuden tiedon hankintaa.*¹ Tiede (voidaan) siis ymmärtää toiminnaksi, jossa tavoitellaan ja hankitaan **tietoa**.

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

16LUKU 2. TIETEELLINEN TIETO, TILASTOT JA ARKITIETO YHTEISKUNNASSA

- Tieteellinen tutkimus on tutkivan subjektin ja tutkimusobjektin välistä vuorovaikutusta.
 - Tiede pyrkii järjestämään tiedon yksinkertaisiksi kokonaisuuksiksi ja pyrkii löytämään säännönmukaisuuksia.
- Tiede on siis tiedon hankintaa, jonka kohteena on meitä ympäröivä todellinen maailma sen ilmiöineen ja tapahtumineen.
 - Tiedon hankinnalla tarkoitetaan kumulatiivista prosessia, jossa ympäröivän maailman ilmiöitä ja niiden välisiä suhteita
 - i) selitetään,
 - ii) niitä koskevia käsityksiä vahvistetaan osoittamalla ne tosiksi sekä
 - iii) löydetään niistä uutta tietoa.
 - Tiede siis erottaa intuition ja “arkitiedon” oikeasta, tutkitusta tiedosta esittämällä reaali maailmaa koskevia väitteitä ja osoittamalla ne todeksi tieteellisin menetelmin.
 - Tiede käsittää myös aiemman tutkimuksen ja se toimii kaiken tieteellisen tiedon jäsennehtynä 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 tiedettävä sen tuottaessa uutta tietoa.
 - Tieteelle voidaan asettaa (ainakin) seuraavat kaksi sitä määrittelevää ominaisuutta.
 - **Järjestelmällisyys:** tieteellinen tiedonhankinta on yhteiskunnallisesti organisoitu tutkimusta tekevien (ja opetusta järjestävien) instituutioiden tehtäväksi, joka kokoaa tutkimustulokset systemaattisiksi tietojärjestelmiksi niin kansallisella kuin kansainvälisellä tasolla.
 - * Näihin instituutioihin 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 liitettävät **hypoteesit**.

Tieteellinen teoria

Tieteelliset teoriat ovat hyvin perusteltuja kuvauksia ja selityksiä siitä, miten ympäröivä maailmamme toimii tai esimerkiksi siitä miten eri ilmiöt ovat yhteyksissä toisiinsa. Ne ovat luotetuin, täsmällisin ja kattavin tieteellisen tiedon muoto. Teorian vahvuus riippuu siitä, kuinka laajoja ja erilaisia reaalimaailman 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 alistainen muutoksille 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 hyödyllisyyden, 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 “verifoidaan”) 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.
 - Tämä tieteellisen tiedon kumuloituminen muokkaa teorioita vuosien saatossa täsmällisemmiksi ja paremmiksi kuvauksiksi

ympäröivästä maailmasta.

- * On kuitenkin syytä huomauttaa että tieteellisetkään teorialat 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.

Tieteilijät yleensä perustavat hypoteesinsa aikaisemmin tehtyihin havaintoihin, joita ei voida selittää olemassa olevilla tieteellisillä teorioilla tyydyttävästi.

Hypoteesi

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

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

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

- Esimerkkinä toimii taloustieteen nk. “uskottavuusvallankumous”, jossa tilastollisten menetelmien myötä taloustieteellisen tutkimuksen painopiste tuntuu siirtyneen vahvemmin empiirisen kausaali-tutkimuksen puolelle.
- Paradigmat siis ohjaavat uuden tieteellisen tiedon tuottamista asettamalla tutkimukselle yhtenevät raamit, jotka ohjaavat sitä, miten tutkimuskysymyksiä asetetaan ja miten niihin etsitään vastauksia sekä myös sitä, miten saatuja tuloksia tulkitaan.
 - Tieteellinen tieto perustuu siis eri tutkimusalojen tiedeyhteisöjen paradigmoihin ja täten siihen, minkälaista 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 syntyä: liian vahvasti alan paradigmojen kanssa ristiriidassa oleva tulos saattaa jäädä julkaisematta, mikäli tutkija ei pidä sitä lainkaan mahdollisena suhteessa vallitseviin paradigmoihin.
 - Tieteelliseen ajattelutapaan kuuluu olennaisesti juuri tiedon kumuloitumisen ymmärtäminen: yksittäinen vahva tulos on vasta alku ja vahvistettu tieto jostain ilmiöstä, yhteydestä tai vaikutuksesta syntyy monien mittausten ja tutkimusten jatkumona.

- Tietoa ei siis voida johtaa siitä, miltä asiat näyttävät, kuten on tyyppillistä “arkiajattelussa”.
 - * Tiede kehittää teorioita kriittisesti ja määrätietoisesti 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

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

Tieteellinen tieto

- ▶ perustuu tietoiseen opiskeluun, analyysiin ja yleistämiseen (otantateoria)
- ▶ muodostaa hierarkkisen 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) riippuvaa, tieteellistä **menetelmää** eli **metodia**.

Tieteellinen menetelmä: Tieteellinen menetelmä on kullakin tieteen alalla vallitseva, ajan myötä kehittynyt ja nykyisten paradigmojen mukainen menettelytapa, jolla uutta tietoa tuotetaan ja vanhaa, mutta epä-

varmaa tietoa vahvistetaan. Se ei ole selkeä työvaiheiden luettelo tai menetelmähakemisto, vaan yleisesti hyväksytty ja hyväksi todettu tapa pyrkiä totuuteen erilaisten tutkimusongelmien ratkomisessa. Hyvälle tieteelliselle menetelmälle voidaan lukea seuraavia kriteerejä.

- **Objektiivisuus ja loogisuus**

- Tutkimuskohteen ominaisuudet ovat tutkijan mielipiteistä riippumattomia.
- Tieteellinen tieto tutkimuskohteesta syntyy tutkijan ja tutkimuskohteen vuorovaikutuksen 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 lisäoletusten 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 hyväksyttävää vedota siihen, että väitteen totuus olisi toivottavaa tai epätoivottavaa esimerkiksi poliittisista, uskonnollisista tai moraalisisista syistä.

- **Edistytvyys**

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

- **Toistettavuus ja yleistettävyyys**

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

- Tarkastellaan lähemmin erästä määritelmää erilaisille toistettavuuden la-

jeille. Esittelemme 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 reaali maailmasta ja ne kuvaavat tutkimuksen aihetta jostain näkökulmasta tarkasteltavana systeeminä.
 - Mallit hyödyntävät matemaattista esitystapaa, sillä se tarjoaa formaalin ja objektiivisen tutkimusaiheen kuvauksen sekä mahdollistaa siihen liittyvän loogisen päättelyn havaitun, empiirisen aineiston pohjalta.
 - Tilastolliset mallit ovat käytännössä tieteellisten mallien formaaleja matemaattisia esityksiä, jotka lisäksi mahdollistavat mallia koskevan tilastollisen päättelyn esimerkiksi hypoteesien ja niiden testaamisen avulla. Päättely perustuu tilastotieteen teoriaan, joka mahdollistaa päättelyn epävarman ja satunnaisen aineiston tapauksissa.
 - Hypoteesien testaamisen voidaan ajatella tutkittavaa ilmiötä koskeviksi ennusteiksi, joita verrataan havaittuun aineistoon. Mikäli havaittu aineisto ei sovi testattavaan teoriaan tai siihen liittyviin hypoteeseihin, voidaan teoriaa kehittää paremmaksi. Tämä vuoropuhelu vie tiedettä eteenpäin ja tuottaa lisää tutkittua tietoa ympäröivästä maailmasta.
 - Hypoteesien testaaminen on yhtäältä tieteellisten teorioiden kehittämisen ja vahvistamisen ja toisaalta kritiikin keskiössä.

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

- 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. Nykyaikana 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 kuvaaviin **(tilasto)tietoihin** ja niiden analysoinnissa käytettäviin **tilastollisiin menetelmiin**.
 - Oikein todellisuutta kuvaavat, ajantasaiset (tilasto)tiedot ovat välttämättömiä modernin yhteiskunnan toiminnalle.
 - Esimerkiksi päätöksenteko sekä julkisella että yksityisellä sektorilla (elinkeinoelämässä) perustuu pitkälti yhteiskuntaa ja elinkeinoelämää kuvaaviin (tilasto)tietoihin ja tilastollisten menetelmien tuottamiin tuloksiin sekä niiden perusteella tehtäviin päätöksiin.
 - * Esimerkkejä ovat tietyt konkreettiset (talous)poliittiset toimenpiteet (talous)tilastojen perusteella. Lisäksi tuotantoprosessien ohjaus ja laadunvalvonta teollisuudessa sekä markkinatutkimus kaupan alalla perustuvat tilastollisiin menetelmiin.
 - (Tilasto)tietojen saatavuutta voidaan pitää jopa toimivan demokration edellytyksenä.
- Koska todellisuutta kuvaaviin (tilasto)tietoihin sisältyy (lähes) aina epävarmuutta ja satunnaisuutta, tilastotiede ja tilastolliset menetelmät luovat perustan tilastojen tuotannolle, jalostukselle ja analysoinnille.
 - Niinpä tilastojen tuotannon, jalostuksen ja analysoinnin menetelmien kehittäminen on keskeinen osa tilastotieteen tehtäväkenttää.
 - Samoin tilastotieteen menetelmien ymmärtämisellä on keskeinen rooli tietoyhteiskunnassa toimimisessa 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 tutkimushenkilöitä miesten ja naisten joukosta ja mitataan kuinka paljon he puhuvat.
 - Satunnaisuus tärkeää, sillä jos valikoitaisiin tarkoituksella puheliaita tai vähäsanaisia tutkimushenkilöitä, tulokset vääristyisivät.
- Jokaiseen mittaukseen liittyy virhe.
 - Täysin satunnainenkaan otos ei edusta täydellisesti koko väestöä. Joukkoon saattaa valikoitua puhtaasti sattumaltakin poikkeuksellisen puheliaita tai harvasanaisia naisia tai miehiä.
 - Millaisia sekoittavia tekijöitä tulee mieleen? Mitkä seikat voisivat vaikuttaa tutkittavaan asiaan?
 - Tosin mitä suurempi otos, sitä pienemmäksi sattuman osuus käy ja joudutaan turvautumaan todennäköisyyksiin: 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!

2.4 Mitä on tutkimus?

- Tiede tavoittelee tietoa, mutta mistä?

Are Women Really More Talkative Than Men?

Matthias R. Mehl^{1,2}, 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?

- Jokaisen tutkimuksen lähtökohtana on (tai ainakin pitäisi useimmiten olla) tiedollisen uteliaisuuden, käytännön tarpeiden tai teorian kehittämisyrittämyksen 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 paikansäilyvyyttä.
 - Valtaosa tieteestä asemoituu erityisesti kahden viimeisen kohdan alaisuuteen vaikka tieteen popularisoinnissa (mm. median toimesta) usein keskitytäänkin uusiin tiedemaailmaa järjestyttäviin löydöksiin, jotka tosin voivat usein olla hyvin epävarmoja!
 - * Lisää tieteen popularisoinnista ja jaksossa 4.6.
- Millaisia kysymyksiä **tutkimuksessa** asetetaan (voidaan asettaa)?
 - **Kuvaus:** Kuinka suuri on yli 65-vuotiaiden osuus Suomen väestöstä?
 - **Riippuvuuden kuvaus:** Ovatko paljon mainostavat yritykset kannattavampia kuin vähän mainostavat?
 - Kuvattujen ilmiöiden **selittäminen** ja **ymmärtäminen**. Miksi vanhempien sosioekonominen asema vaikuttaa ekonomien työhönsijoitumiseen? Tämän tutkimuskysymyksen tapauksessa pyrkimys on lä-

hinnä 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.
 - * Kehittynekin tieteellinen teoria tai malli on aina reaali maailman 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 vahvuuksien ja puutteiden sekä historiallisen ja ideologisen kontekstin tiedostaminen on ensiarvoisen tärkeää kaikessa tutkimuksessa!
- Joka tapauksessa täyteen 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 käsitteissä, jotka harvoin ovat arvovapaita. Useimmat käsitteet voidaan korvata toisilla, joilla on paikoin hyvin erilainen arvo sisältö 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 jäsenenä meillä on tuskin mahdollisuuksia (täydellisesti) irroittautua arvoistamamme kun pyrimme esim. ammatillisiin päämääriin.
 - Myös päinvastainen ongelma olemassa: Tutkimusta arvioidaan siihen perustellusti tai perusteettomasti kiinnitettyjen arvonäkökohtien mukaan!

- Tutkimukseen kuuluu olennaisesti myös oman tutkimustyön kuvaaminen, ts. kertomus siitä, miten esitettyihin tuloksiin on päästy.
 - Tämän myötä tieteelliselle ajattelulle on ominaista automaattinen **itsensä korjaaminen**.
 - Tutkimuskysymys, valitut menetelmät, käytetty aineisto ja tehdyt johtopäätökset perataan auki tutkimusartikkelissa/raportissa, joka sitten lähetetään **vertaisarvioitavaksi** tietelliseen julkaisuun, jossa muut alan asiantuntijat arvioivat sen ja päättävät hyväksytäänkö se julkaistavaksi.
- **Vertaisarvioinnissa** yksi tai useampi, tehdystä tutkimuksesta riippumaton, saman alan tutkija lukee ja tarkastaa tehdyn tutkimusartikkelin, arvioi sitä ja suosittaa tietellisen julkaisun arvioinnista vastaavalle päätoimittajalle (editorille) kyseisen artikkelin hyväksymistä tai hylkäämistä.
 - Vertaisarviointi ei aina takaa sitä, että julkaistu tutkimus olisi virheetön ja erinomaisesti tehty, vaan myös väärää tietoa pääsee välillä vertaisarviointiprosessin läpi.
 - Tämä ei kuitenkaan poista tieteellisen prosessin luotettavuutta, sillä uusi tieto varmentuu vasta usean samaa tutkimuskysymystä tutkineen ja vastaavat tulokset saaneen tutkimuksen myötä. Toisin sanoen, tieteellisen prosessin voidaan ajatella konvergoituvan totuuteen, vaikka yksittäisiä virhearviointeja sattuisikin.
- **Tutkimuksen kieli**
 - Tutkimus edellyttää arkikieltä täsmällisempää kommunikaatiota.
 - Ongelmaan liittyvien käsitteiden huolellinen määrittäminen ja erittely on tarpeellista.
 - * Käsitteiden ja eri aloilla, osin samoista asioista käytettävien, toisistaan eroavien termien systemaattinen määrittely ja jäsentely selkeyttää tiedeyhteisön välistä kommunikointia.
 - * Eivät korvaa empiiristä tietoa vaan vaikuttavat tiedon järjesty miseen ja sen perusteella tehtäviin päätelmiin.

Esimerkki: Luonnontieteelliset vs. yhteiskunnalliset sovellutukset:

- Luonnontieteiden lainalaisuuksia: Monet luonnontieteelliset ilmiöt ovat luonteeltaan varsin pysyviä.
 - Voidaan tehdä luotettavasti laajojakin yleistyksiä.
 - Selityksiä voidaan empiirisesti testata.

- Luotettavia matemaattisia esityksiä voidaan kehittää.
- Yhteiskuntatieteissä (yhteiskuntatieteiden historiallisuuden myötä) erinäisiä lainalaisuuksia ja tyypillisiä piirteitä:
 - Usein tutkitaan **yhteiskunnallisia ilmiöitä**, jotka eivät suurelta osin ole toistettavissa.
 - Vaihtelevat huomattavasti ajan myötä (aiemmin voimassa olleet lainalaisuudet eivät välttämättä ole enää voimassa ja päinvastoin), mikä vaikeuttaa tilastollista analyysiä.
 - Yhteiskunnallisten ilmiöiden mittaaminen?
 - * Yhteiskunnan rakenne ja toiminta on ehdollinen siinä käytettävän merkitysjärjestelmän suhteen. Kysymys **mittaamisesta** on asetettava suhteessa tähän käsitejärjestelmään. Joudutaan tekemään erilaisia kompromisseja eksaktisuus- ja systemaattisuusvaatimusten sekä arkikie- len monimerkityksellisyyden välillä.

2.5 Tutkimuksen vaiheet ja tulosten julkaiseminen

Tieteellinen tutkimus ja asiantuntijatyö tuottavat valtavan määrän perusteltua, luotettavaa tutkimustietoa. Ks. tarkemmin tieteellisestä julkaisemisesta linkin tapauksessa erityisesti yhteiskuntatieteiden alalla, mutta peruseriaatteen 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 hyväksytyllä 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 opintokokonaisuuteesi.

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

Luku 3

Tilastotiede tieteenalana

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

Tieteenalan määrittäminen lyhyesti on aina hieman hankalaa. Tästä huolimatta seuraavassa yritämme osaltaan vastata seuraaviin kysymyksiin:

- Mitä tilastotiede on ja mitä se ei ole? Miksi tilastotiede ei ole vain sovellettua matematiikkaa tai matematiikalla höystettyä tietojenkäsittelyä?
- Mihin tilastotiedettä käytetään? Onko tilastotieteellä käyttöä ns. “akatemian” eli tutkimusyhteisön ulkopuolella?
- Minkälaista on tyypillinen tilastotiedettä kohtaan esitetty kritiikki?

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 reaali maailman ilmiöitä. Täten tutkimuskohteena on tavallisessa elämässä tavattavia asioita, ihmisiä tai tapahtumia. Tutkimuskohteita kutsutaan tilastoyksiköiksi ja niiden joukkoa kutsutaan populaatioksi (perusjoukoksi).

- Esimerkiksi jos tutkitaan kuntavaaleissa äänestävien tuloja niin jokainen äänestysikäinen muodostaa oman tilastoyksikkönsä (ks. alla) 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 muuttujia**.
 - Edellisissä esimerkeissä nämä olisivat esim. äänestäjien tulot ja kuntien äänestysprosentti.
 - Mielenkiinnon kohteena olevia tilastollisia muuttujia kutsutaan **tutkimusmuuttujiksi** (tulot ja kuntien äänestysprosentti) ja niiden lisäksi voidaan kerätä ylimääräistä tietoa eli **taustamuuttujia** (näitä voisi olla esimerkiksi asuinpaikka ja kunnan väkiluku).
 - Tilastoyksiköiden tilastollisilla muuttujilla on tietty mahdollisten arvojen joukko, ja näillä arvoilla on jokin **jakauma** populaatiossa.
 - * Esimerkiksi tulot voivat määritelmästä riippuen saada minkä tahansa positiivisen arvon mutta äänestysprosentti on luonnollisesti rajattu nollan ja sadan prosentin väliin.

Tilastoyksikkö ja tilastollinen muuttuja

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

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

Havainto

Havainto muodostuu tilastoyksikön tarkasteltavien tilastollisten muuttujien havaitusta arvoista.

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

Havaintoaineisto/data

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

Tiivistettynä:

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

3.2 Mitä tilastotiede on ja mitä se ei ole?

- Aloitetaan tarkastelemalla erinäisiä **tilastotieteen “karakterisointeja”** eri tahojen ja tutkijoiden toimesta:
 - ***Tilastotiede on tietotuotannon teknologiaa**, jonka avulla voidaan suorittaa kvantitatiivisten tietojen joukkotuotantoa ja havaintoihin perustuvia tieteellisiä ja käytännöllisiä päätöksiä. Tilastotiede on siis yksikköjen muodostamaan joukkoon liittyvän numeerisen tietoineiston keräämistä, analysointia ja tulkintaa koskeva tiede*¹.
 - ***Tilastotiede on yleinen menetelmätiede**, jota sovelletaan, jos reaali maailman 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 reaali maailman 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 merkitykselliset työnsä vasta 1920- ja 30-lukujen aikana, joten tällä lentävällä lausahduksella ei ole mitään tekemistä nykyisten tilastollisten menetelmien kanssa.

⁴(A.C. Davison)

⁵(Sund, 2003)

- Tilastotiede siis **kehittää** ja **soveltaa menetelmiä** ja (tilastollisia) **mal-
leja**, joiden avulla reaalimaailman ilmiöistä voidaan tehdä johtopäätöksiä
ilmiöitä kuvaavien numeeristen tai kvantitatiivisten tietojen perusteella
tilanteissa, joissa tietoihin liittyy **epävarmuutta ja satunnaisuutta**.
 - Tilastollisten menetelmien avulla pyritään löytämään reaalimaailman
satunnaisia ilmiöitä kuvaavista numeerisista (eli kvantitatiivisista)
tiedoista **systemaattisia piirteitä** joita jalostetaan sellaiseen muo-
toon, että ilmiöistä 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 meka-
nismeja. Näiden prosessien tuottamia tietoja (aineistoja) tiivistetään
usein graafisiksi esityksiksi ja tunnusluvuiksi sekä tilastollisten
mallien parametreiksi, joiden pohjalta johtopäätöksiä tehdään.
 - Tässä onnistuakseen tilastollisten menetelmien tuleekin pyrkiä erot-
telemaan **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 useimmille ensimmäisenä mieleen yhteiskun-
taa ja sen toimintaa kuvaavat **numeeristen tietojen järjestelmäl-
liset kokoelmat**, tilastotiede ei suinkaan ole ainoastaan tilastojen
ja niiden tekemisen oppia.
 - * Tämä siitäkkin huolimatta, että niiden menetelmien konstruoin-
ti, joilla näitä tilastoja tuotetaan, jalostetaan ja analysoidaan on
keskeinen osa tilastotiedettä. Tilastot ovat siis usein tilastotie-
teen soveltajan tutkimuskohteena ja tilastojen laadinnassa käy-
tetään apuna tilastotieteen menetelmiä.
 - * Suomessa Tilastokeskus toimii virallisena tilastoviranomaisena
ja tilastotuottajana. Tätä **tilastotuotannon** kokonaisuutta ni-
mitetään ajoittain **tilastotoimeksi**. **Tilastotieteen käyttö-
alue 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 reaali maailman ilmiötä kuvaava **numeeristen tai kvantitatiivisten tietojen järjestelmällinen kokoelma** voi muodostaa **tilastollisen aineiston** ja siten tilastollisen tutkimuksen mahdollisen kohteen.
 - Esimerkiksi kaikki **empiirisen** tai **kvantitatiivisen** tutkimuksen tutkimus- tai havaintoaineistot ovat tilastotieteen kannalta tilastollisia aineistoja.
- Tilastotiede sijoittuu tieteiden kentässä matematiikan, filosofian ja tietojenkäsittelytieteen rinnalle. Tästä huolimatta se ei kuitenkaan ole yksiselitteisesti minkään näiden osa-alue.
 - **Tilastotiede ei ole matematiikan osa-alue**, sillä tilastotiede lähestyy tieteellistä ongelmanratkaisua eri tavoin:
 - * Matematiikka on tietyllä tavalla aina eksaktia ja sen tulokset perustuvat formaaliin deduktioon ja loogisiin todistuksiin, johtaen usein “eksaktiin” ratkaisuun tai matemaattisesti formaaliin ratkaisun loogiseen esitystapaan. - Tilastotiede sen sijaan on aina konteksti- ja aineistopohjaista ja perustuu induktiiviseen päättelyyn. Saadut tulokset ovat aina epävarmoja - koska ne kuvailevat epävarmaa tietoa generoivia prosesseja!
 - Tilastotiede on siis hyvä nähdä omana tieteenalanaan matemaattisesta esitystavastaan huolimatta. Eihän esimerkiksi myöskään fysiikkaa (sentään) pidetä matematiikan osa-alueena!
 - **Tilastotiede ei ole myöskään tietojenkäsittelytieteen osa-alue**, vaikkakin useiden laskennallisten menetelmien ja tehokkaan tietojenkäsittelyn rooli tilastollisissa analyyseissä on jatkuvasti kasvanut.
 - * Tietojenkäsittelytieteen teoria ei rakennu tilastotieteen tavoin ajatukselle epävarmoista ja satunnaisista reaali maailman ilmiöistä.

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

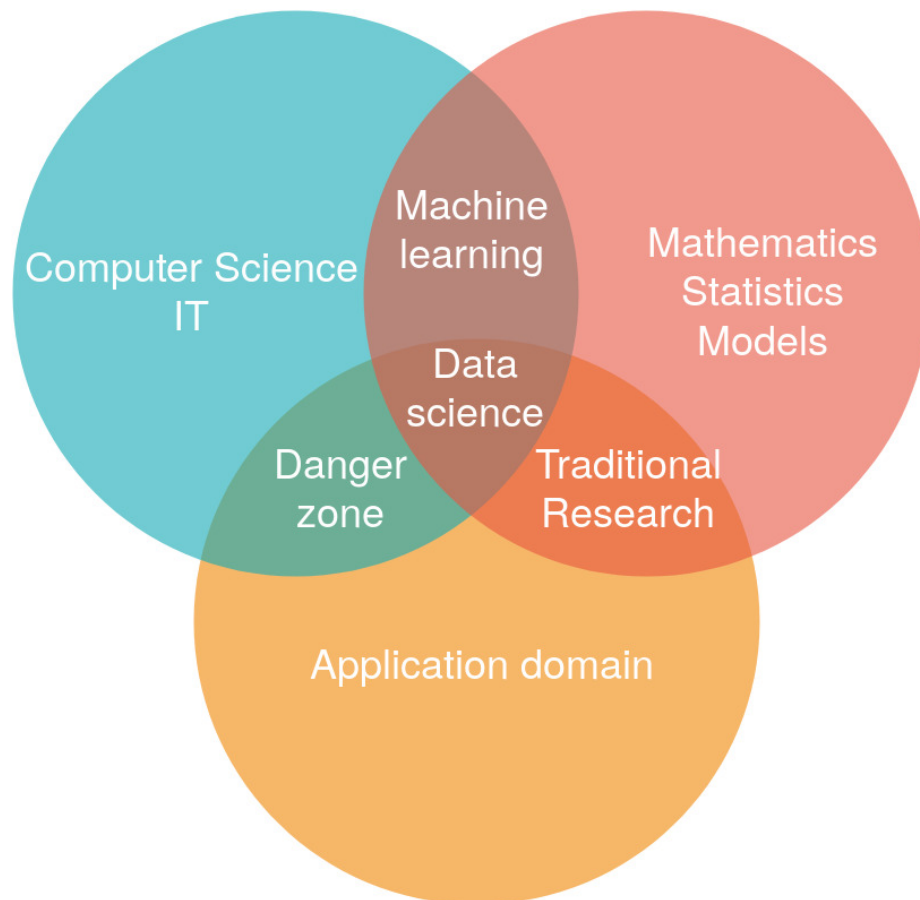
Mitä tilastotiede (ainakin) on

- 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 tutkijoille yhtenevät ja tiedeyhteisön hyväksymät raamit, jotka mahdollistavat (tilastollisen) päättelyn ja päätöksenteon epävarman tiedon vallitessa. Näin voidaan uskottavasti ja luotettavasti tiivistää tietoa, jota erilaiset aineistot sisältävät, perustaa johtopäätöksiä näille tiivistyksille ja saavuttaa uusia tieteellisiä löytöjä.
 - * Tilastotieteen menetelmien käyttö ja soveltaminen onkin siis aina alakohtaista. Tästä huolimatta tilastollisia menetelmiä sovelletaan aina johonkin **aineistoon!**
 - Tilastotieteen nähdäänkin usein kuuluvan ns. **menetelmätieteisiin**, joissa mm.:
 - * Kehitetään työkaluja muiden tieteiden tutkimusongelmien ratkaisuksi
 - * On myös oma sovelluksista vapaa teorianmuodostuksensa
 - Menetelmäkehityksen näkökulma tilastotieteeseen: *tilastotiede kehittää matemaattisia **malleja** satunnaisilmiöitä kuvaavia kvantitatiivisia tietoja generoiville prosesseille*. Koska tietoihin liittyy **epävarmuutta** tai **satunnaisuutta**, **tilastolliset mallit** perustuvat **todennäköisyyslaskentaan**.
 - * Juuri sattuman ja epävarmuuden huomioiminen tutkimusasetelmissä erottaa tilastotieteen muista menetelmätieteistä!

- Tilastollisia menetelmiä voidaan soveltaa tietojen keruun, jalostuksen ja analysoinnin jokaisessa vaiheessa. Päämääränä on jalostaa tiedot muotoon, joka mahdollistaa tutkittavaa reaalimaailman ilmiötä koskevien johtopäätösten tekemisen käytettyjen menetelmien pohjalta, eli ns. **tilastollisen päättelyn**.
 - Tutkimuksessa on pystyttävä valitsemaan ja käyttämään menetelmiä, jotka antavat aineistosta vastauksia haluttuihin kysymyksiin. Tämä vaatii yhtä lailla sovellusalaakohtaista osaamista (ns. substansiosaamista) kuin myös kattavaa menetelmäosaamista.
- Tilastotieteessä lähtökohtana ja ratkaisevassa asemassa on siis aina jonkin satunnaishetken 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, kuvaillaan ja mallinnetaan tilastollisin menetelmin, joiden kehittäminen on keskeinen osa tilastotiedettä.
 - Pelkkä menetelmien kehittäminen kuuluu pitkälti matemaattisen/teoreettisen tilastotieteen osa-alueelle.
 - Pelkkä aineistoon keskittyminen ja (mekaaninen) analysointi voi sen sijaan olla joissain tilanteissa pitkälti tietojenkäsittelyä.
 - **Tilastollinen “mallintaminen”** löytyykin näiden välistä ja se sisältää eri alojen sovelluksista kumpuavan tarpeen uusien menetelmien kehittämiseen.
 - * Tämä vuoropuhelu muodostaa tilastotieteelle luonnollisen “takaisinkytkennän” teoreettisen ja soveltavan puolen välillä: uudet teoreettiset menetelmät vastaavat soveltavan tilastotieteen ongelmiin mutta herättävät aina uusia kysymyksiä, jotka palautuvat taas teoreettisen tilastotieteilijän pöydälle!
 - Luonnollisesti valtaosa tilastotieteilijöistä ja lähitieteiden harrastajista asettuvat näiden äärimmäisten luonnehdintojen välimaastoon eikä tarkkaa luokittelua ole sinänsä tarpeen tehdä ja korostaa.
 - Joka tapauksessa tilastotieteen kehityksen keskiössä ovat aina sovellusalaakohtaiset ongelmat, joista useat palautuvat yleisemmälle tasolle teoreettisen tilastotieteen kehityspolkuihin.

3.3 Tilastotieteen suhde lähitieteisiin

- Kuvio 3.1 tarjoaa karkean yleistyksen tietojenkäsittelytieteen (Computer Science) ja sovellusalan (Application domain) sekä tilastotieteen (Statistics) ja matematiikan (Mathematics) välisistä yhteyksistä. On selvää että tilastotieteellä on paljon päällekkäisyyksiä lähitieteidensä kanssa ja joskus näkeekin (huolimatta edellä tehdyistä huomioista) että tilastotiede niputetaan yhteen matematiikan tai tietojenkäsittelytieteen kanssa.



Kuva 3.1: Tilastotieteen ja rajatieteiden yhteyksiä kuvaava Venn-diagrammi

- Yritetään siis vielä hahmotella tilastotieteen suhdetta sitä lähimpänä olevaan (soveltavaan) matematiikkaan.
 - Tilastotieteessä olennaisen otantateorian (Luku 5) voisi ajatella olevan matemaattisesti määritelty teoria, jossa myös on aineiston käsite,

mutta se ei tee siitä vielä varsinaisesti tilastotiedettä.

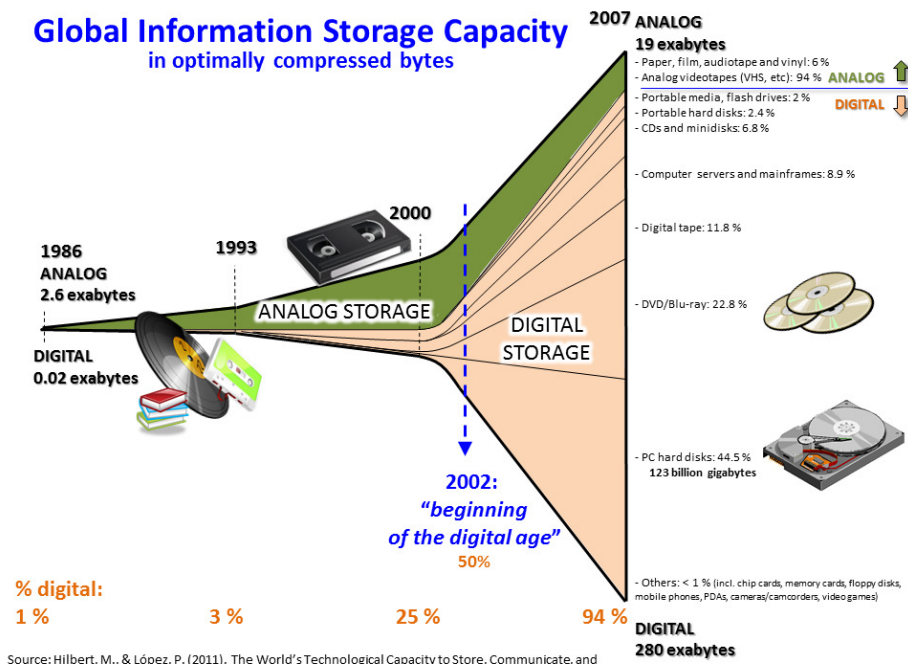
- Matematiikassa kuvataan ongelma ja esitetään se teorian muodossa, eli malli on *“parametreista havaintoihin”*.
- Tilastotieteessä ongelma on käänteinen, edetään *“havainnoista parametreihin”*, mutta ongelman matemaattinen kuvaus vaaditaan ensin.
- Tilastotiede esittää menetelmiä ja käsitteitä tämän käänteisen ongelman ratkaisemiseen.
 - * Karkeasti erotellen tilastotieteessä käsiteltävät ongelmat lähtevät aina havainnoista eli aineistosta ja matematiikassa suunta on teoriasta aineistoon.
 - * Voidaankin siis sanoa, että tilastotieteen erottaa puhtaasta matematiikasta se, että siinä tutkitaan 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 tarkalteen 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 sään ennustaminen.
 - Kun jotain ilmiötä kuvataan stokastisen mallin avulla, voidaan käyttää (joudutaan käyttämään) tilastollisia menetelmiä. Vaikka käytännössä laskenta hoidetaan tietokoneohjelmien avulla, meidän tilastotieteen tutkijoina ja käyttäjinä on huolehdittava tutkimusprosessin onnistuneesta toteutuksesta muilta osin.
- Tilastotiede ei myöskään ole puhtaasti tietojenkäsittelyä, vaikka tilastotiede onkin luonteeltaan aineistopohjaista ja aineistojen sisältämää tietoa on käsitelty osin samoin kuin tietojenkäsittelyssä siitä asti kun se on ollut mahdollista (tietokoneen keksimisen myötä).
 - Tilastotieteen ja tietojenkäsittelytieteen ero on lähitieteistä selvin: tilastotieteellä on *“mekaanisesta”* tai teoreettisesta tietojenkäsittelystä selkeästi erillinen ja oma teoriapohjansa.
 - * Siinä missä tilastotieteen teoria perustuu aineiston stokastiselle mallintamiselle, tietojenkäsittely on enemmänkin algoritmista

ajattelua, missä aineistolla on ratkaisevalla tavalla erilainen rooli.

- Lisäksi suomen kielessä tietojenkäsittely ymmärretään laajemmassa mielessä ohjelmoitavissa olevaksi automatisoimiseksi, jota tilastotiede ei perusolemukseltaan suinkaan ole.

- Tarkastellaan seuraavaksi tilastotieteen suhdetta viime vuosien aikana paljon suosiota keränneeseen datatieteeseen (data science) johon voidaan katsoa lukeutuvan mm.
 - Tilastotiede ja matematiikka
 - * Erityisesti tilastollinen data-analytiikka ja satunnaisen aineiston mallintaminen sekä soveltuvat soveltavan matematiikan osa-alueet.
 - Tietojenkäsittely
 - * Tietoteknologian kehityksen myötä taitavien tietojenkäsittelijöiden kysyntä on kasvanut merkittävästi. Lähes jokaisella alalla kerätään entistä enemmän dataa lähes kaikesta, jonkun pitäisi osata myös käsitellä sitä!
 - * Datatieteen voidaankin osaltaan katsoa syntyneen tästä elinkeinoelämän tarpeesta asiantuntijoille, jotka osaavat käsitellä suuria tietoaaineistoja (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. alta).

- Datatieteellä ei usein nähdä olevan omaa historiallisen tieteellisen prosessin luomaa teoriapohjaa vaan sen voidaan katsoa olevan kokoelma eri alojen tieteellisiä tuloksia, jotka voidaan yhdistää tavalla, jonka “datavalankumous” (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 (tarkimpiin) ennustetuloksiin tilanteessa, jossa jokin toinen malli kuvaaisi ilmiötä annettua mallia paremmin.
 - Ko. soveltaja ottaa mallin sekä sen antaman ennustetuloksen annettuna, eikä mieti *mistä kyseinen ennustetulos johtuu*. Jotta tarkat ennustetulokset toteutuvat jatkossakin (kun uutta aineistoa, dataa, tulee saataville), on ennustajan oleellista huomioida mitkä tekijät johtivat tarkkaan ennustulokseen.
 - Eri menetelmät sopivat eri sovelluskohteisiin. Tilastotieteilijä osaa useimmiten tunnistaa eri sovelluskohteisiin sopivat menetelmät paremmin kuin tietojenkäsittelijä. Vastaavasti tehokkaan/onnistuneen ohjelmointikoodin kirjoittamisessa tilanne on usein toisinpäin.

3.4 Tilastotieteen osa-alueet

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

Soveltava tilastotiede

Soveltava tilastotiede

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

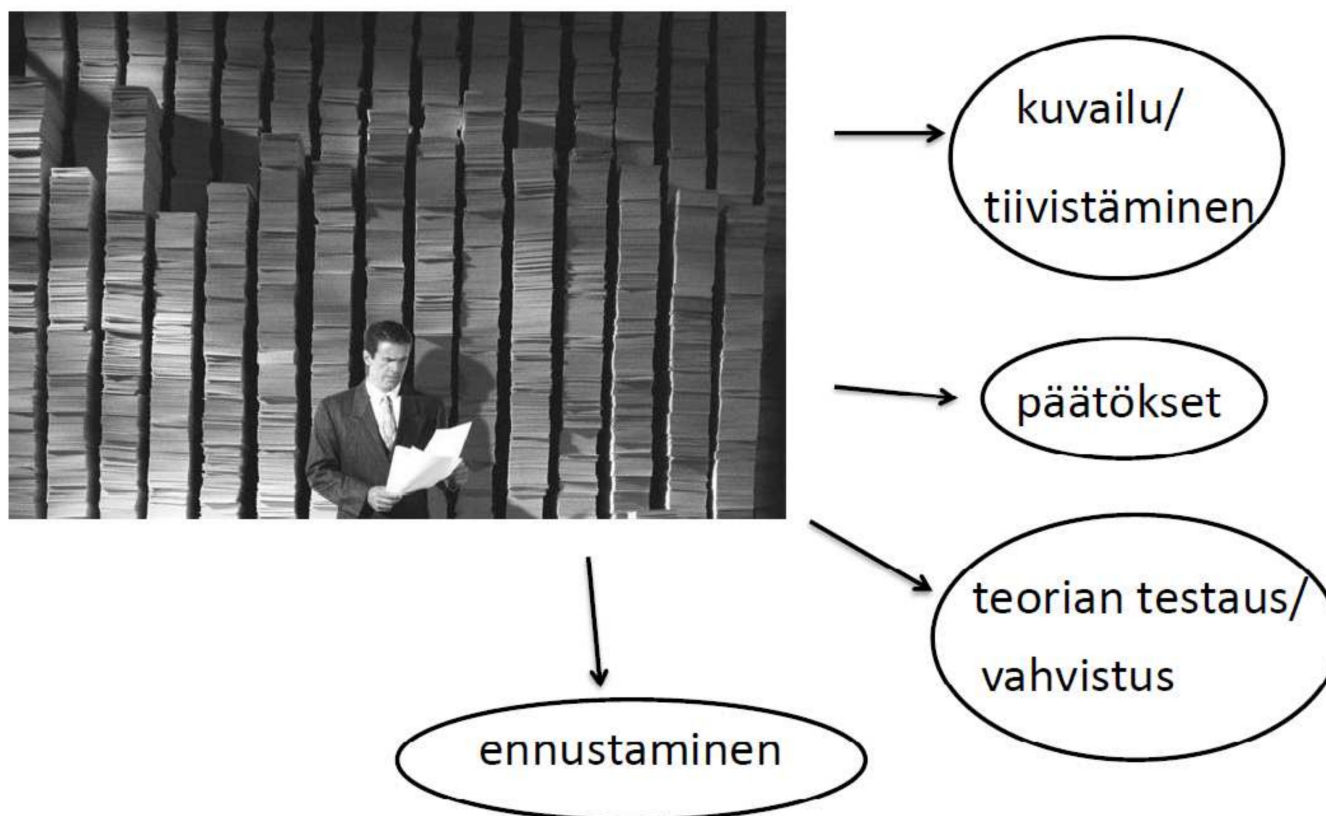
- Yleisesti ottaen eri tieteenaloilla kohdattavat menetelmäsuuntaukset voidaan jakaa kahteen luokkaan tutkimusaineistojen tyypin perusteella:
 - **Kvantitatiivinen:** eli määrällinen tutkimus on tutkimusta, jossa tutkimusongelma on muotoiltu tarkasti etukäteen ja tutkimuskysymyksiin vastataan käyttäen tilastollisia menetelmiä pyrkien **selittämään ja ennustamaan** tutkimuksen kohteena olevaa ilmiötä.
 - * Täsmällisten ja laskennallisten tilastollisten menetelmien käyttäminen numeeriseen aineistoon on kvantitatiiviselle tutkimukselle ominaisin piirre.
 - * Perustuu yleensä satunnaisotokseen (kts. luvut 4, 5 ja 6) ja tutkimusaineisto on tiivistetty numeeriseksi havaintomatriisiksi, jolle oleellinen vaatimus on sen totuudellisuus.
 - * **Kritiikki:** määrällinen tutkimus on (paikoin) sokea tutkittavien ilmiöiden sellaiselle luonteelle, jota ei pystytä kvantifioimaan, eli muuntamaan numeeriseen muotoon. Näihin voidaan katsoa lukeutuvan mm. tunteet, merkitykset ja kokemukset, ellei tutkija keksi niiden numeeriselle mittaamiselle uskottavaa keinoa.
 - **Kvalitatiivinen:** eli laadullinen tutkimus on tutkimusta, jossa tutkimuksen kohteena olevaa ilmiötä ja sen merkitystä sekä tarkoitusta pyritään **ymmärtämään** kokonaisvaltaisella tavalla.
 - * Laadullisessa tutkimuksessa annetaan usein tilaa tutkimuksen kohteena olevien ilmiöiden ja/tai ihmisten näkökulmille, vaikuttimille, kokemuksille ja tuntemuksille. Tutkimusyksiköjen otanta on täten usein harkinnanvaraista.
 - * Laadullisessa tutkimuksessa tutkimusongelma muotoutuu tutkimuksen edetessä ja sille tyypillistä on hypoteesittomuus, eli tutkimus on tarkoitus aloittaa mahdollisimman vähin ennakkooletuksin. Ennakkooletuksista on kuitenkin mahdotonta täysin irtautua, joten niiden ilmi tuominen esioletuksina tai ”tutkimushypoteeseina” eli arvauksina tuloksista on osa tutkimusta.
 - * Kritiikkiä: laadullinen tutkimus ei pysty vastaamaan kysymyseen 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 pitäisi) soveltaa kaikilla reaalimaailmaa tutkivilla tieteenaloilla, joiden tutkimusaineistot voidaan esittää kvantitatiivisessa muodossa.
 - Tilastollisten menetelmien käyttö on siis huomattavan paljon yleisempää määrällisessä kuin laadullisessa tutkimuksessa.
- Menetelmien soveltamisen tarkoituksena on (voi olla): **i) kuvailla ja tiivistää tietoa**, jota havaittu aineisto sisältää **ii) sovellusalan oman teorian empiirinen testaus** tai **iii) edellisten pohjalta tehtävä tilastollinen päättely**.
 - **Deskriptiivisellä eli kuvailevalla tilastotieteellä** tarkoitetaan sellaisten menetelmien soveltamista, joiden avulla havaintoaineistosta voidaan esimerkiksi laskea tunnuslukuja, kuvata havaintomuuttujien jakaumia ja visualisoida aineiston generoimaa ilmiötä tai siitä johdettuja tunnuslukuja.
 - **Tilastollinen päättely** on sen sijaan aineiston tarkasteluun/kuvailuun sekä mallintamiseen perustuvaa päätöksentekoa, jossa kvantitatiiviseen aineistoon kuuluva epävarmuus ja satunnaisuus on otettu huomioon.
 - * Keskeinen tilastollisen päättelyn käyttötarkoitus soveltajille on usein **teorian ja siihen liitettävien hypoteesien testaaminen**, joka voi johtaa joko teorian vahvistumiseen (*verifiointiin*) tai sen vääräksi osoittamiseen (*falsifioimiseen*) (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.

Teoreettinen tilastotiede

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

- Teoreettinen tilastotiede luo pohjan tilastollisten menetelmien ymmärtämiselle, soveltamiselle ja kehittämiselle.

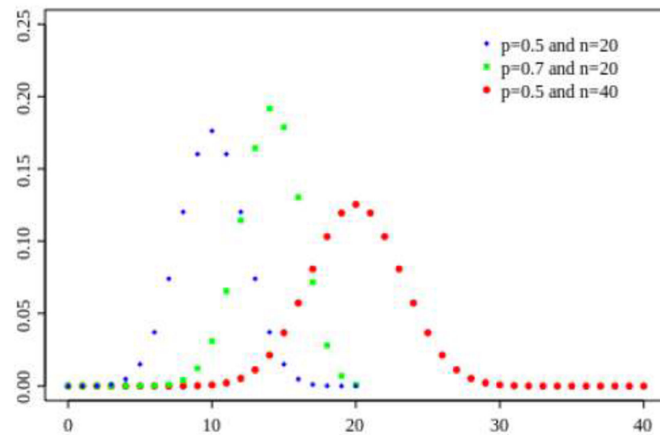


Kuva 3.3: Soveltava tilastotiede

- Ilman riittävää ymmärrystä tilastollisten menetelmien toimintaperiaatteista niiden soveltaja on vaarassa tehdä virhepäätelmiä! (Ks. alaluku 3.5 tilastotieteen kritiikistä)
- 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 parametrasta. (ks. tarkemmin luku 6)
 - Uskottavuusfunktion arvo kertoo kuinka todennäköisenä voidaan havaittua aineistoa pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvoilla.
 - Uskottavuuspäätelyn perusajatuksena on, että se tai ne parametriarvot, joilla uskottavuusfunktion arvo maksimoituu kuvaa aineiston generoinutta prosessia parhaiten.
 - Aineistoa koskevia hypoteeseja voidaan testata käyttäen uskottavuusfunktion maksimia vastaavaa tilastollista mallia!
 - *“Kaikki mallit ovat väärinä, mutta jotkut ovat käyttökelpoisia.”* (Box, 1976).
- Uskottavuusfunktiot perustuvat aina satunnaisilmiöiden mahdollisia arvoja kuvaaviin nk. **tiheysfunktioihin** tai todennäköisyysfunktioihin.
 - Tiheysfunktiot kuvaavat jonkin satunnaismuuttujan (satunnaisilmiön) saamien arvojen jakaumaa.
 - Esimerkiksi kolikonheitto on satunnaisilmiö ja sillä on vain kaksi arvoa⁷ ja kolikonheittoa voidaan kuvata nk. binomijakaumalla, merkitään $\text{Bin}(n, p)$ missä n on heittojen lukumäärä ja p on kruunan todennäköisyys.
 - Esimerkki: heitetään kolikkoa 40 kertaa ja saadaan kruuna 40/40 tapauksessa. Onko tämän havaintoaineiston perusteella uskottavaa, että kolikonheitto noudattaa binomijakaumaa $\text{Bin}(40, 0.5)$? Eli kuinka uskottavan voidaan pitää että kyseinen kolikko on tavallinen, painotamaton kolikko?
- Todennäköisyyslaskenta luo tilastotieteelliselle epävarmuuden mallintamiselle vahvan ja uskottavan matemaattisen perustan.

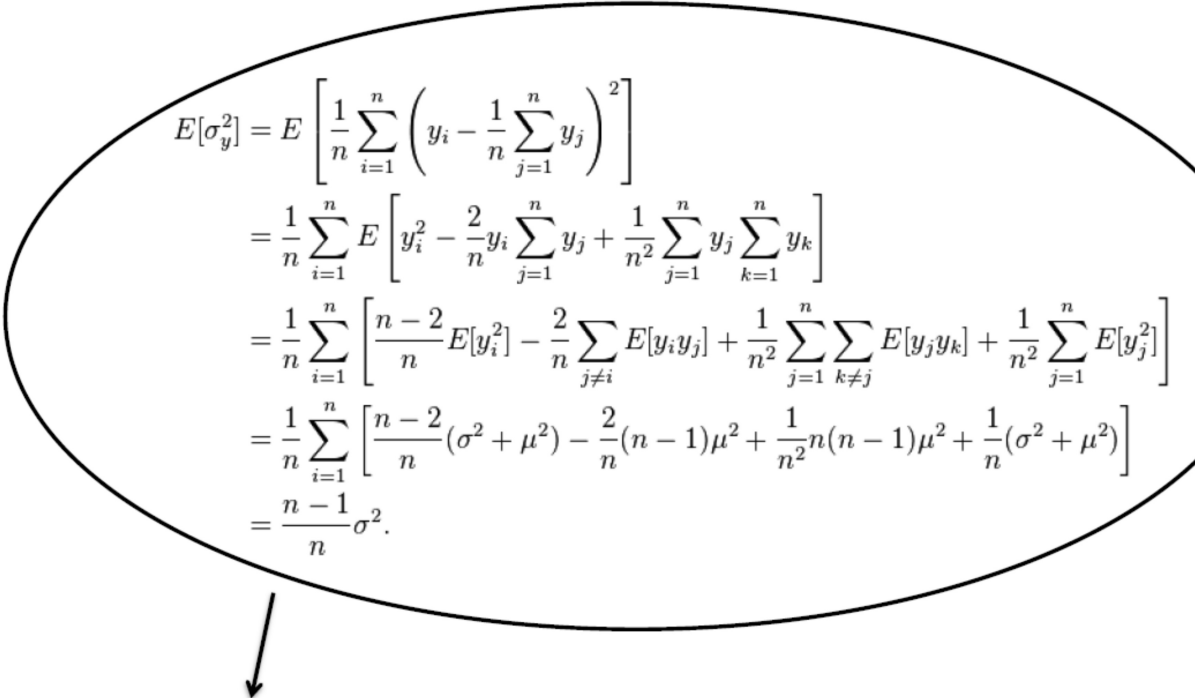
⁷Kolikon kantilleen jäämistä ei tässä lasketa mahdolliseksi tapahtumaksi.

Tilastotiede perustuu uskottavuuksiin, jotka taas perustuvat todennäköisyyteen ja tiheysfunktioihin.



Kuva 3.4: Tilastotiede ja todennäköisyys

- Todennäköisyyslaskentaa opetetaan tarkemmin (tätä kurssia seuraavilla) kursseilla TILM3553 Todennäköisyyslaskennan peruskurssi pääaineopiskelijoille, TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille ja SMAT5306 Todennäköisyyslaskennan jatkokurssi.



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

Pohja tilastollisten menetelmien ymmärtämiselle, soveltamiseksi ja kehittämiselle

Kuva 3.5: Teoreettinen tilastotiede

3.5 Tilastotieteen kritiikkiä

- Tilastotieteen rooli tiedeyhteisössä on niin tärkeä että sitä kohtaan on ymmärrettävästi esitetty myös paljon kritiikkiä. Valtaosa kritiikistä kohdistuu joko tilastotieteen matemaattisuuteen tai sitten siinä tarvittaviin oletuksiin, jotka mahdollistavat esimerkiksi hypoteesien testaamisen.
 - Usein kritiikki on aiheetonta ja johtuu sen esittäjän puutteellisesta tilastotieteen ymmärryksestä. Perusteettoman kritiikin esittäminen toista tieteenalaa kohtaa ei kuitenkaan ole vieras ilmiö juuri millään alalla.

- Tässä alaluvussa käymme läpi yleisimpiä kritiikin muotoja, joita tilastotiedettä kohtaan esitetään ja pyrimme tarjoamaan vastauksia/vastineita silloin kun niitä voidaan antaa.

“Yleismaailmallinen” kritiikki

- Aloitetaan yleismaailmallisella kritiikillä, jota tilastollista tutkimusta vastaan on esitetty:
 - Tilastotieteessä käytettävien tunnuslukujen, kuten keskiarvon, reaaliaikailman 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 lasketaan.
 - * Esimerkiksi puhekielessä yleinen nk. “Keskiarvo-Kalle”, eli 1,8 lapsen vanhempi ja 1,5 auton omistaja on tietenkin täysin kuvitteellinen.
 - * Lisäksi joskus kuulee tilastotieteilijöitä kritisoitavan lausumalla *“Jos toinen jalka on jääkylmässä vedessä ja toinen kiehuu vedessä, niin tilastotieteilijän mielestä ihmisellä on tällöin keskimäärin hyvä olla”*
- Korrelaatio on tunnusluku, joka kuvaa kahden muuttujan välistä riippuvuutta (palaamme tähän tarkemmin luvussa 7). Se ei kuitenkaan kuvaa millään tavoin kausaalisuutta, eli sitä kumpi aiheuttaa kumman, jos kumpikaan.⁸(<https://www.tylervigen.com/spurious-correlations>) mitä moninaisimpia esimerkkejä kahdenvälisistä nk. *näennäisistä* korrelaatioista.
 - Esimerkiksi “jäätelön syönti ja hukkumiskuolemat” -tapauksessa havainnollisesti todetaan jäätelönkulutuksen ja hukkumiskuolemien lukumäärän korreloivan keskenään, mutta taustalla vaikuttava tekijä onkin lämmin kesä, joka vaikuttaa molempiin.
- Vaikkei näiden esimerkkien oikeellisuutta ole syytä kiistää, niin tilastollisen tiedon arvioinnissa on kuitenkin syytä päästä syvemmälle.

Kritiikki matemaattisuutta kohtaan

⁸Tyler Vigen on kerännyt verkkosivuilleen

- Ehkä merkittävin kritiikki tilastollisia menetelmiä kohtaan kohdistuu kritiikin näkökulmasta perusteettomaan, tai ainakin liian vahvaan, matemaattisuuden tuomaan itsevarmuuteen. Voidaankin siis perustellusti kysyä, että **onko tieteellisyys = matemaattisuus**?
- Useat tieteenalat käyttävät tutkimuksessaan edistyneitäkin tilastollisia menetelmiä siitä huolimatta, että tutkijoiden tilastomatematiinen pohjakoulutus ei välttämättä ole riittävällä tasolla kyseisten menetelmien kokonaisvaltaiseen ymmärtämiseen.
 - * Helppokäyttöisistä tilasto-ohjelmistoista on riittävät perustaidot omaaville käyttäjille erittäin paljon hyötyä mutta koneiden ja ohjelmien käytön opettelu ei kuitenkaan ole varsinaista tilastotiedettä (tarvitaan enemmän tilastotieteen opintoja).
 - * Laskentatehon ja modernin tietojenkäsittelyteknologian ansiosta monimutkaisiakin tilastollisia analyysejä on kuitenkin mahdollista tehdä vaikka tutkijalla olisi tilastotieteestä vain perustiedot, jos sitäkään.
 - * Pahimmillaan tämä saattaa johtaa siihen, että analyyseja tehdään ymmärtämättä mitä itse asiassa ollaan tekemässä.
- Tilastollisten analyyysien hyödyllisyyden ja järkevyyden ehtona on kuitenkin käytettävien menetelmien, aineiston ja tutkittavan ilmiön pintaa syvemmälle ulottuva tuntemus.
 - * Käytettävien tilastollisten menetelmien oletukset on osattava ottaa huomioon ja toisaalta odottamattomien tulosten syyt on pystyttävä jäljittämään.
 - Teknistä esitystä käyttävää tutkijaa saatetaan pitää erityisen uskottavana, koska hän kykenee käyttämään vaikeita menetelmiä. Tästä huolimatta tutkimusongelma ei saisi päästä unohtumaan.
 - Tutkijan tulisikin varmistua siitä, että käytettävät menetelmät todella vastaavat asetettuihin tutkimuskysymyksiin ja että tutkimusongelma on ratkaistavissa käytettävillä menetelmillä.
 - Tekninen esitys ei takaa onnistunutta tilastollista tutkimusta eri näkökulmista katsoen. Monet tilastolliset menetelmät ovat vaikeita ja vaativat soveltaajiltaan paljon.
 - Lisäksi on hyvä muistaa, että käytettävien menetelmien lähtökohdat ja oletukset eivät matemaattisuudesta huolimatta ole välttämättä neutraaleja!
 - * Kaikkia tieteentekijöitä ei voida velvoittaa opiskelemaan edistynyttä abstraktia tilastotieteen teoriaa (tilastomatematiikkaa), mutta menetelmien oikeaoppinen käyttö kuitenkin vaatii riittävästi ymmärrystä.

Kritiikki yksinkertaistuksia kohtaan

- Edellisiä kohtia yleisemmin tilastotiedettä on kritisoitu siitä, että se ei kykene riittävällä tasolla huomioimaan reaali maailman kompleksisuutta.
 - Merkittävässä osassa tilastollisia analyyseja lähtökohtana on usko “todellisen” maailman ja näin ollen aineistoa generoivien mekanismien olemassaoloon.
 - * Tätä saatetaan usein pitää kuitenkin kyseenalaisena: voiko “tosielämän stokastiikasta” muka todella löytyä säännönmukaisuuksia?
 - * Tämä kysymys on kuitenkin pitkälti tieteenfilosofinen ja palautuu lopulta sovellusalaan sekä tutkimusongelmaan ja -kysymykseen: tilastollisten menetelmien toimivuutta voidaan helposti testata esimerkiksi simulaatiokokeilla.
 - Tilastotiedettä on myös kritisoitu sen “sokeudesta” sosiaaliseen vuorovaikutukseen liittyviin subjektiivisiin kokemuksiin kuten tunteisiin, kokemuksiin ja ei-numeerisiin havaintoihin.
 - * Tämä kritiikki ei kuitenkaan suoranaisesti ole tilastotieteen kritiikkiä, vaan jälleen sovellusalaan liittyvä ja erityisesti tutkimuskysymyksen asettelua koskeva ongelma.
 - Tuntemuksia ja kokemuksia voidaan hyvin testata tilastollisin menetelmin, mikäli tutkija osaa uskottavasti määritellä niille numeerisen mittauksen kriteeristöt!
 - Tämä on kuitenkin vaikeaa, sillä aivan kaikkea ei voida kvantifioida: kirjoitetun tekstin tai sosiaalisten merkitysten tulkinnan sekä elämysten kuten musiikin ja taiteen aiheuttamien mielikuvien ja tunteiden voidaan perustellusti nähdä olevan hyvin haastavia kvantifioida.
 - * Näiden aiheiden tulkinta, ymmärtäminen ja tutkiminen ulottuu kvantitatiivisen tutkimuksen ulkopuolelle.
 - Mikäli tutkittavasta ilmiöstä pystyy kvantitatiivisilla mittauksilla 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 käyttökelpoisia.

Temppukokoelmakritiikki

- Eräs ehkä osin implisiittinen kritiikki tilastotiedettä kohtaan on sen pitäminen nk. **“temppukokoelmana”**.
 - Tilastotieteen voi nähdä koostuvan numeeristen tietojen jalostamisen menetelmistä. Tämä näkemys, joka on usein tahaton, pelkistää tilastotieteen *vain* **menetelmäkokoelmaksi**, vailla omaa teoriaa.
 - Eri tutkimusalojen empiirisessä työssä (liian) usein vain kerätään aineisto ja vasta sitten mietitään mitä sillä voitaisiin tehdä.
 - Usein apuun haetaan tilastotieteilijä, jonka odotetaan loihdivan (tilastollisen) ratkaisun ongelmaan kuin ongelmaan.
 - * Joskus tämä toki onnistuukin, mutta useimmiten ei.
 - * Tilastotiede ei siis ole “työkalupakki”, josta valitsemalla oikeanlaisen menetelmän voi vastata mihin tahansa tutkimuskysymykseen!
 - Tilastolliset menetelmät tulee ymmärtää ja niitä tulee soveltaa kaikissa soveltavan tutkimuksen vaiheissa, jotta tutkimusongelmaan kyetään vastaamaan eikä turhaa työtä tule tehdyksi.
 - Karkeasti luokitellen tilastotieteilijät kehittävät menetelmiä, joita soveltajat käyttävät.
 - * Soveltavia tilastotieteilijöitä löytyy kuitenkin yhä kiihtyvissä määrin! Erityisesti eri rajatieteiden alueilla, kuten alaluvussa 3.6 lyhyesti esitellään.

Tilastotieteen väärinkäyttö

- Tilastotiedettä on myös mahdollista käyttää väärin monin eri tavoin, joka edelleen altistaa koko tieteenalan (perusteettomalle) kritiikille!
 - Tilastoja ja tilastotiedettä käytetään paljon väärin, mutta tämä on usein tahatonta (esim. puutteellisesta koulutuksesta johtuvaa).
 - * Joskus kuitenkin näkee tarkoituksellista tilastojen vääristelyä tai tahallista tilastollisten menetelmien väärinkäyttöä!
 - * Kansalaisten tiedelukutaidon ja tilastollisten menetelmien tuntemuksen merkitys on kasvanut viime vuosikymmeninä ja kasvane jatkossa yhä, kun esimerkiksi erilaiset “vaihtoehtotieteet” ovat nousseet suosituimmiksi.
 - * Tilastotieteen ymmärrys auttaa itse kutakin tunnistamaan virheellisiä tai puutteellisin tiedoin tehtyjä päätelmiä ja täten helpottaa tietoyhteiskunnassa toimimista ja kriittistä ajattelua!
 - Yleisiä tilastollisten menetelmien väärinkäyttötapoja ovat esimerkiksi seuraavat:
 - **“Kolmannen tyypin virhe”**: kun tilastollisia menetelmiä käyttämällä saadaan oikeita vastauksia, mutta väärin kysymyksiin! Esimerkiksi jos tutkija ei täysin ymmärrä minkälaisia vastauksia käytävissä olevasta aineistosta ja valitulla menetelmällä voidaan saada,

voi hän syyllistyä kolmannen tyypin virheeseen. Tällöin voi nimittäin käydä niin, että hän tulkitsee tilastolliset testit täysin oikein, mutta 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 tutkijana haluat tutkia onko kahden eri ikäryhmän ihmisten pituuksissa eroja ja sinulla on käytettävissä edustava otos molempien ikäluokkien edustajista. Tutkit siis onko toisen ryhmän, ryhmän A, keskipituus *pienempi* kuin ryhmän B ja testaat päteekö tämä *yksisuuntaisesti*. Testituloks osoittaa, että voit hylätä nollahypoteesin, jonka mukaan ryhmien keskipituus olisi sama. Kolmannen tyypin virhe syntyy silloin, jos tosiasiallisesti testin hylkääminen johtui siitä, että ryhmän A keskipituus olikin *suurempi* kuin ryhmän B keskipituus, mutta tätä et testin tuloksen perusteella voi tietää!

3.6 Tilastotieteen sovelluskohteita ja “rajatieteitä”

- Yleisenä menetelmätieteenä tilastotiedettä sovelletaan useilla eri tieteenaloilla.
 - Jokaisella sovellusalalla on oma erillinen teoriapohjansa sekä empiiriset käytänteet, joten substanssietous 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 substanssietoutta. Tämän myötä soveltavan tilastollisen tutkimuksen kenttä on laaja ja rikas.
 - Osa näistä sovelluskentistä on kehittynyt vahvassa yhteisvaikutuksessa tilastotieteen ja lähitieteiden (viime aikoina erityisesti koneoppimisen) yhteydessä.
- Usein on pystyttävä arvioimaan ongelmanasettelun ja tulosten tarkoituksenmukaisuutta ja pyrkiä välttämään siltä että tutkijan tieteelliset ja yhteisölliset sitoumukset heijastuisivat tutkimuksen kulkuun.

- 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 kvantitatiivisessa muodossa voi soveltaa/voisi soveltaa/pitäisi soveltaa tilastollisia menetelmiä sekä tutkimusaineistoja kerättyinä 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":
 - Sovellusaloja, joilla tilastotieteen soveltaminen on muodostunut omaksi tutkimuskohteekseen/tieteenlajikseen:
 - * Psykologia: psykometriikka,
 - * Sosiaalitieteet: sosiometria,
 - * Taloustiede: ekonometria,
 - * Kemia: kemometria,
 - * Bio- ja lääketiede: biometria,
 - * Epidemiologia,
- Soveltavan matematiikan tutkimusaloja, jotka ovat osaltaan päällekkäisiä tilastotieteen kanssa
 - Informaatiteoria,
 - Matemaattinen tilastotiede,
 - Todennäköisyyslaskenta,
 - Operaatioanalyysi
- Tietojenkäsittelytieteen alaan (osittain) lukeutuvia tutkimusaloja
 - Laskennalliset menetelmät,
 - Data mining,
 - Knowledge discovery,,
 - Hahmontunnistus,,
 - Tekoäly,,
 - Koneoppiminen,
- Ja paljon muita!

Luku 4

Sattuma ja satunnaisuus tilastotieteessä

- 4.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä
- 4.2 Satunnaisuus ja todennäköisyydet
- 4.3 Tilastolliset mallit, jakaumat ja parametrit
- 4.4 Odotusarvo ja varianssi
- 4.5 Joitain jakaumia
- 4.6 Sattuman rooli tieteenteossa: Vale-emävale-tilasto?

Luku 5

Tilastolliset aineistot, niiden kerääminen ja mittaaminen

5.1 Kertausta: Data eli aineisto

5.2 Otannan idea

5.3 Mittaaminen ja mitta-asteikot

5.4 Kontrolloidut kokeet ja suorat havainnot

5.5 Otantamenetelmät

5.6 Otantaesimerkkejä

5.7 Otannan haasteita vielä kootusti

Luku 6

Otokset ja otosjakaumat: tilastollisen päättelyn näkökulma

- 6.1 Satunnaisotos, yhteisjakauma ja tilastollinen malli
- 6.2 Otosjakauma: Estimaattori ja estimaatti
- 6.3 Otoskeskiarvo ja otosvarianssi (estimaatto-reinta)
- 6.4 Suhteellisen frekvenssin otosjakauma
- 6.5 Muita tunnuslukuja
- 6.6 Luottamusvälit
- 6.7 Otokoko

Luku 7

Tilastollinen riippuvuus ja korrelaatio

7.1 Muuttujien väliset riippuvuudet

7.2 Kahden muuttujan havaintoaineiston kuvaaminen

7.3 Tunnusluvut

7.4 Satunnaismuuttujien kovarianssi ja korrelaatio

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

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 huomioita

Luku 12

Tilastotieteen kehityksen nykytrendejä