LAURI NUMMENMAA

TILASTOTIETEEN KÄSIKIRJA

TAMMI

TILASTOTIETEEN
KÄSIKIRJA

TAMMI

Tammi on osa Werner Söderström Osakeyhtiötä
Graafinen suunnittelu ja taitto Atte Kalke, Vitale Ay
ISBN 978-952-04-0138-2
Painettu EU:ssa

SISÄLLYS

LUKIJALLE

Maailma ei koostu numeroista. Luonnon, tekniikan, talouden ja yhteiskunnan ilmiöiden käsitteleminen numeroina kuitenkin helpottaa niiden jäsentämistä. Numeroiden avulla voidaan kuvailla asioiden välisiä suuruuksia, suhteita, yleisyyttä tai todennäköisyyttä. Tilastotieteen tehtävänä on auttaa tällaisten numeeristen päätelmien tekemisessä. Tilastotiedettä tarvitaan kaikilla tieteenaloilla tutkimusaineiston keräämiseen, kuvailemiseen ja mallintamiseen. Lisäksi monet sovellusalat kuten kaupankäynti, media-ala sekä terveydenhoito ja teknologia tarvitsevat tilastotieteen menetelmiä riskien arvioimiseen ja ennusteiden laatimiseen. Tilastotieteen perusteiden hallinta kuuluukin nykyään kaikkien tieteenalojen yleissivistykseen.

Tämä kirja on kirjoitettu tilastotieteen käyttäjille – sekä opiskelijoille että tutkijoille – yhdistäen tilastotieteen teoriaa ja sovelluksia. Kirja on lähtökohtaisesti soveltavan tilastotieteen perusteos, mutta keittokirjamaisen esityksen sijaan asiat esitellään havainnollisesti tilastotieteen teoriaan perustuen. Satunnaisuuden ja todennäköisyyslaskennan perusteista sekä yksinkertaisimmista perusmenetelmistä on laadittu käsinlaskuesimerkit, koska tällaisia menetelmiä on vielä mielekästä harjoitella ilman tietokoneen apua. Monimutkaisemmista menetelmistä on esitetty niiden teoreettinen perusta intuitiivisesti, ja keskitytty analyysiin liittyviin käytännön kysymyksiin sekä tilastollisen ohjelmoinnin perusteisiin. Kirja etenee tutkimuksen tekemisen perusteista tutkimusaineiston graafiseen ja numeeriseen kuvaamiseen. Tämän jälkeen esitellään tilastollisen päättelyn periaatteita ja yksinkertaisimpia tilastollisia testejä yhtäsuuruuksiin ja riippuvuuksien väliseen tarkastelemiseen. Seuraavaksi käsitellään tilastollisen mallintamisen ja luokittelun perusteita, ja kirjan loppuosa keskittyy aineistoa tiivistävien tilastollisten monimuuttujamenetelmien käsittelemiseen.

Mitä paremmin ymmärrämme luonnon ja yhteiskunnan toimintaa, sitä monimutkaisempaa huomaamme sen olevan. Lisäksi mittalaitteiden ja mittaustapojen kehittyminen lisää saatavilla olevan datan määrää tehden aineistojen analysoimisesta vuosi vuodelta vaativampaa. Onneksi tilastolliset ohjelmistot pysyvät mukana kehityksessä. Nykypäivänä yksi tilastollinen ohjelmisto on merkittävästi kilpailijoitaan parempi. Avoimeen lähdekoodiin perustuva, ilmainen ja valtavan kansainvälisen verkoston kehittämä ja tukema **R** (https://cran.r-project.org) on 2000-luvulta alkaen ollut tilastotieteen *lingua franca* -yleiskieli jota tilastotieteilijät ympäri maailmaa puhuvat äidinkielenään, mutta jota myös ”muunkieliset” toisten alojen tutkijat ymmärtävät sujuvasti. Tämän vuoksi tämän kirjan analyysiesimerkit perustuvat R-koodin käyttöön. Vaikka R perustuu komentokieleen ja tilastolliseen ohjelmointiin, ei kannata pelästyä. Ohjelmointia voi lähestyä kuten aloitteleva ruuanlaittaja keittokirjaa. Hyvästä keittokirjasta löytyy helppotajuiset reseptit, joita noudattamalla aloittelijakin onnistuu ruoanlaitossa, ja voi vähitellen alkaa muuntelemaan ja parantelemaan alkuperäisiä reseptejä. Samalla tavalla ohjelmointiakin voi opetella aluksi kopioimalla ja

muuntelemalla olemassa olevaa ohjelmakoodia. Jos et vielä tätä kirjaa lukiessasi osaa ohjelmoida, niin nyt on erinomainen syy aloittaa opiskelu! Harvasta asiasta on niin paljon hyötyä kaikilla elämän alueilla kuin ohjelmointitaidosta, koska se opettaa ajattelemaan algoritmeja, tiedon hallintaa sekä prosessien toteuttamista niiden automatisoimisen näkökulmasta.

Tilastollisten menetelmien käyttäminen vaatii usein improvisointia. Jokainen tutkimusaineisto voidaan analysoida lukuisilla erilaisilla tilastollisilla menetelmillä, eikä mikään näistä ole välttämättä ainoa oikea tai absoluuttisesti paras. Samoin luonto ajautuu usein törmäyskurssille tilastotieteen teorian kanssa. Tässä kirjassa on koetettu lähestyä tutkimusaineiston analysoimista mahdollisimman realistisesta näkökulmasta. Siistien ja ongelmattomien aineistoesimerkkien sijaan kirjassa käytetään todellisia tutkimusaineistoja, jotka eivät läheskään aina ole täydellisiä. Aineistoja ja esimerkkejä on kerätty useilta tieteenaloilta. Kiitän lämpimästi kaikkia omia aineistojaan esimerkkikäyttöön luovuttaneita tutkijoita, joiden täydellinen lista löytyy kirjan lopusta. Kaikki kirjassa käytetyt esimerkkiaineistot ja -koodit voi ladata omalle koneelleen osoitteesta **emotion.utu.fi/tilastotiede**.

Tutkimus ja tilastollinen analyysi on luovaa työtä siinä missä soittaminen, kirjoittaminen tai maalaaminenkin. Tilastotiede on tutkijan työkalu samoin kuin asteikot ja harmonia ovat muusikon työkaluja, tai väripaletti maalarin työväline. Ilman työkalujen hyvää hallintaa ei itse työstäkään tule mitään. Toisaalta myöskään pelkkä työkalujen hallinta ei riitä kovin pitkälle. Teknisesti ja harmonisesti täydellinen musisointi on puuduttavaa kuunneltavaa, jos siitä puuttuu musiikillinen idea. Samoin tilastotieteen näkökulmasta virheettömästi suoritettu tutkimus voi olla tyhjänpäivästä, jos tutkijalta puuttuu oivallus ja näkemys siitä, mikä maailmassa on kiinnostavaa ja merkityksellistä. Hyvä tutkija hallitseekin sekä oman tutkimusalansa sisällöt puhtaan teknisen tilastollisen osaamisen lisäksi. Yhdessä kirjassa on mahdotonta esittää kattavasti kaikkia mahdollisia tilastotieteen osa-alueita. Tässä kirjassa esiteltyjen menetelmien avulla voi kuitenkin hallita ja käsitellä jo hyvin monipuolisia tutkimusaineistoja. Tilastotieteen teoria ja menetelmät päivittyvät jatkuvasti, ja tämän kirjan tiedoilla selviää erinomaisesti myös tulevaisuudessa uusia tilastomenetelmiä opetellessa ja soveltaessa.

Turussa 8.6.2021

Tekijä

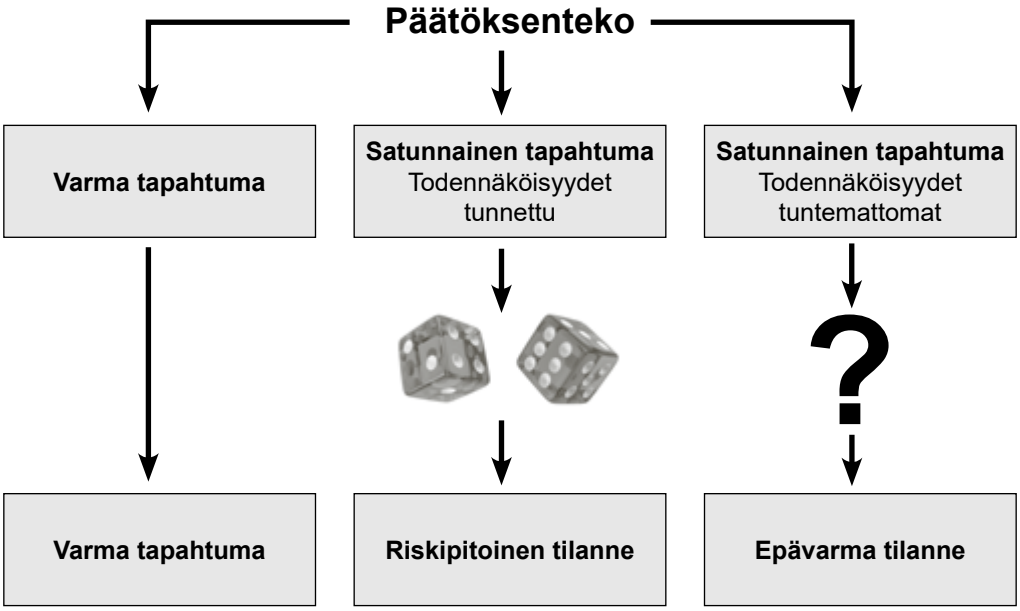
1. EPÄVARMUUDEN VÄHENTÄMINEN

Kaikki tutkimus perustuu tavalla tai toisella mittaustulosten numeeriseen tai tilastolliseen käsittelemiseen. Tilastollisten menetelmien tunteminen on tarpeellista paitsi tutkimusta tehdessä myös tutkimusraportteja lukiessa. Tieteellisten artikkelien lisäksi sanomalehdet, kirjat ja nettisivut ovat täynnä tilastollista tietoa – esimerkiksi talouden kehittymistä kuvaavia lukuja sekä työllisyyden kehitystä tai ilmaston lämpenemistä esittäviä tilastollisia kuvaajia. Tässä luvussa kerrotaan, kuinka tilastollisten menetelmien avulla voidaan tiivistää suuria tietomassoja ja kuinka niiden avulla voidaan tehdä luonnon ja yhteiskunnan ilmiöitä koskevia päätelmiä.

Järjestystä kaaoksen keskellä

Maailma on monimutkainen paikka. Ihmisaivot tekevät jatkuvasti parhaansa ennustaakseen ja ymmärtääkseen ympäristömme ja sosiaalisen pelikenttämme lainalaisuuksia, ja useimmiten ne toimivat erinomaisesti. Osaat varmasti sanoa kokemuksesi perusteella, ovatko miehet yleensä naisia pidempiä (ovat), kesäpäivät talvipäiviä lämpimämpiä (ovat), tai ovatko lapset aikuisia vahvempia (eivät ole). Toisaalta mitkään edellisistä väittämistä eivät pidä *täysin* varmasti paikkaansa. Tunnet varmasti monta miestä, jotka ovat lyhyempiä kuin jotkut tuntemasi naiset. Suomen kesäöinä lämpötila voi joskus laskea alemmas kuin poikkeuksellisen leutoina talvipäivinä, ja joillain lapsilla voi olla enemmän voimia kuin esimerkiksi huonokuntoisilla aikuisilla tai vanhuksilla. Kuinka voimme kuitenkin toimia ja tehdä erilaisia itseämme ja muita koskevia päätöksiä, jos emme voi olla täysin varmoja siitä, mitä päätöksistämme seuraa?

Päätöksenteko voidaan jakaa karkeasti kolmeen erilaiseen tilanteeseen (**Kuva 1.1**). Jos tiedämme täysin varmasti mitä päätöksestämme seuraa, päätös on turvallinen tehdä. Tiedämme esimerkiksi, että vapaasti palavien puiden synnyttämä lämpö ei riitä sulattamaan terästä, joten voimme valmistaa teräksestä turvallisen kiukaan saunaan. Toisaalta tiedämme, että sama palavien puiden lämpö, joka ei riitä sulattamaan terästä, aiheuttaa ihoomme kivuliaan palovamman, joten ymmärrämme olla koskematta kuumaan kiukaaseen. Toisessa ääripäässä olemme tuntemattoman tai hallitsemattoman ilmiön edessä, emmekä tiedä



Kuva 1.1. Päätöksenteko turvallisessa, riskipitoisessa ja epävarmassa tilanteessa.

lainkaan mitä päätöksestä seuraa. Esimerkiksi rahamarkkinoiden heiluessa rajusti sijoittaja ei voi välttämättä kuin arvailla, kannattaako hänen myydä osakkeitaan vai ostaa niitä lisää, koska ennakkotieto ei riitä tulevan kurssikehityksen ennustamiseen.

Tällainen satunnaisuus kuuluu ihmiselämään sekä kaikkeen mitä luonnossa tapahtuu. Satunnaisuus ei kuitenkaan tarkoita samaa kuin kaoottisuus. Vaikkemme voikaan täysin varmasti tietää mitä erilaisista päätöksistä seuraa, emme useimmiten joudu toimimaan täysin sokkona. Suurin osa tapahtumista onkin näiden kahden ääripään – täysin varman ja enemmän tai epävarman – välissä. Useimmat tapahtumat ovat nimittäin – toiset enemmän ja toiset vähemmän **ennustettavia:** tiedämme useimmiten, kuinka **todennäköisiä** eri lopputulokset ovat. Tiedämme esimerkiksi, että voimme lentokoneella matkustaessamme joutua onnettomuuteen ja kuolla. Mutta – koska tiedämme myös, että lento-onnettomuuteen liittyvä riski on häviävän pieni (noin yksi 11 miljoonasta), emme juurikaan murehdi lentomatkustamista. Vastaavasti tiedämme, että todennäköisyys voittaa suomalaisessa lotossa on vielä pienempi (noin yksi 15 miljoonasta), joten haaveillessamme lottovoitosta ymmärrämme, että unelmamme ei kovinkaan varmasti toteudu. Tupakoinnin taas tiedämme lisäävän keuhkosyövän riskiä niin paljon (noin 10–50 kertaa verrattuna siihen, ettemme tupakoi), että jätämme mielellämme savukkeet sytyttämättä.

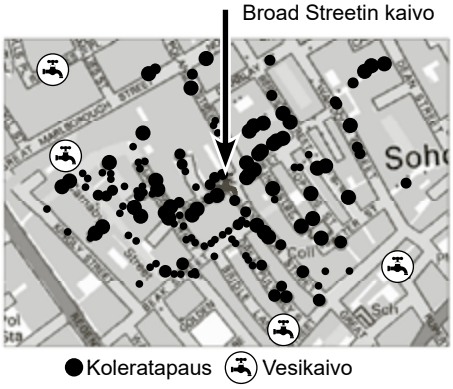
Kaikissa edellisissä tapahtumissa molemmat lopputulokset ovat mahdollisia: Useimmilla kierroksilla joku ihminen voittaa lotossa ja suurin osa ainoastaan menettää rahansa, joka vuosi miljoonien turvallisten lentomatkojen lisäksi sattuu harvoja lento-onnettomuuksia, ja aina jotkut onnekkaat tupakoitsijat onnistuvat välttämään keuhkosyövän ankarasta sauhuttelustaan huolimatta. Vaikkemme tietäisikään mitä itsellemme tulee tapahtumaan, kun astumme koneeseen, täytämme lottokupongin tai jätämme savukkeet kauppaan,

todennäköisyyksien tunteminen auttaa meitä riskipitoisissa päätöksentekotilanteissa. Mutta kuinka voimme määritellä erilaisiin luonnon, yhteiskunnan ja talouden ilmiöihin liittyviä todennäköisyyksiä ja käyttää niitä päätöksenteon apuna?

Kuinka näkymätön saadaan näkyväksi?

Kolera oli vielä 1800-luvun puolivälissä vaarallinen tartuntatauti, jonka aiheuttajaa ei tunnettu. Teollisen vallankumouksen johdosta Lontoo oli nopeasti kasvava suurkaupunki, jossa erityisesti työläisten olot olivat tiiviin asutuksen ja huonon hygienian vuoksi surkeat. Suuri osa jätteistä laskettiin surutta Thames-jokeen, josta myös otettiin vettä kotitalouksien käyttöön. Koleran ja muiden tartuntatautien tarttumismekanismia ei vielä tunnettu. Tutkijat ymmärsivät tautien kulkeutuvan ihmisestä toiseen, mutta tutkijat olivat jakautuneet kahteen leiriin – osa uskoi koleran leviävän ilmateitse ”miasmojen”, myrkyllisten huurujen välityksellä. Toiset tutkijat taas ajattelivat tauteja aiheuttavien mikrobien leviävän ulosteiden kautta vaatteisiin ja käyttöveteen. Vuonna 1853 Lontoon kolmannen suuren koleraepidemian aikaan brittiläinen lääkäri John Snow ryhtyi suorittamaan tilastollista tutkimusta koleratapauksista selvittääkseen pitääkö mikrobiteoria paikkansa. Hän vertaili koleratapausten esiintymistä kahdella Lontoon kaupunkialueella, jotka olivat muuten samanlaisia, mutta jotka saivat juomavetensä eri vesiyhtiöiltä. Snow huomasi, että koleratapauksia oli huomattavasti vähemmän alueella, jonne vesi toimitettiin Thamesin yläjuoksulta, minne jätevesiä ei juurikaan laskettu. Sen sijaan koleratapauksia oli paljon sellaisen vesiyhtiön jakelualueella, joka otti vetensä keskusta-alueelta, jossa siihen pääsi sekoittumaan viemärivettä. Kesken tutkimuksen Snow’n kotiseudulla Lontoon Sohossa puhkesi yllättäen raju epidemia. Snow haastatteli asukkaita ja laati haastattelujen perusteella kartan sairastuneiden potilaiden osoitteista (**Kuva 1.2**). Hän havaitsi tautitapausten keskittyvän alueelle, jossa ihmiset ottivat vetensä Broad Streetin yleisestä kaivosta, johon pääsi valumaan vettä läheisestä suuresta likakaivosta. Snow taivutteli kaupunginvaltuuston poistamaan Broad Streetin saastuneen kaivon pumppukahvan, jotta epidemia saataisiin talttumaan.

Myöhemmin Snow esitti havaintonsa kuuluisassa kartassaan, jossa koleratapaukset on merkitty Sohon alueen karttaan pisteinä (**Kuva 1.2**). Pisteet keskittyvät selvästi Broad Streetin kaivon ympäristöön. Tämä tutkimussarja sai Snown vakuuttuneeksi siitä, että taudit voivat tarttua nesteissä olevien mikrobien mukana. Snow analysoi kaivon vesinäytteitä sekä kemiallisesti että mikroskoopin avulla, mutta ei pystynyt aukottomasti osoittamaan vedessä olevan taudinaiheuttajia. Sen sijaan hän päätteli tautitapausten esiintymisen ja kaivon sijainnin välisestä **yhteydestä**, että veden



Kuva 1.2. John Snow’n kartta koleratapausten esiintymisestä Lontoon Sohossa. Mitä lähempänä Broad Streetin kaivoa talot sijaitsivat, sitä enemmän koleratapauksia havaittiin. Muiden samanlaisten kaivojen yhteydessä samanlaista tautikeskittymää ei havaita.

käyttäminen oli tautitapauksia yhdistävä tekijä: Broad Streetin pumpun lähistöllä sijaitsevissa taloissa koleratapauksia oli merkittävästi enemmän kuin muilla vastaavilla alueilla ja samanlaisten kaivojen läheisyydessä. Myöhemmin 1800-luvun lopulla Louis Pasteurin ja Robert Kochin tutkimukset todistivat tiettyjen tautien olevan elävien olentojen, mikrobien, aiheuttamia. Snow'n tekemät tarkat **havainnot** tautitapausten ilmenemisen **lainalaisuuksista** sekä niiden **esittäminen** karttana osoittivat kuitenkin **tilastollisesti** yhteyden likaisen veden ja sairastumisen välisen yhteyden, ja tutkimusta pidetään eräänä ensimmäisenä epidemiologisena eli tautien tartuntamekanismeja selvittävänä tutkimuksena. Karttansa avulla Snow teki kirjaimellisesti näkymättömästä näkyvän – hän pystyi osoittamaan, kuinka ihmissilmälle näkymättömien (ja tuolloin tuntemattomien!) kolerabakteerien leviäminen noudatti selkeitä lainalaisuuksia.

Tutkimustiedon tilastollinen käsittely

Nykypäivänä eräs tärkeimmistä keinoista epävarmuuden vähentämiseen ja lainalaisuuksien havaitsemiseen on **tilastotiede**. Tilastotieteestä tulee helposti mieleen tilastotietojen mekaaninen tallentaminen ja käsitteleminen, ja tämä onkin yksi tilastotieteen tehtävistä. Huolellisen tilastokirjanpidon vuoksi tiedämme esimerkiksi, kuka on pelaaja tehnyt NBA-uralla eniten pisteitä (Kareem Abdul-Jabbar), kenellä rocklaulajalla on laajin tunnettu ääniala (W. Axl Rosella) tai mikä on kaikkien aikojen myydyin musiikkialbumi (Michael Jacksonin *Thriller*). Tämä on kuitenkin vain pieni osa tilastotieteen monentyyppisistä tehtävistä ja sovelluksista. **Tilastotiede** on menetelmätiede, joka tutkii sitä, miten erilaisiin havaintoihin ja mittauksiin perustuva tutkimus pitää suorittaa. Tilastotieteen tutkimusalueisiin kuuluu mittaustulosten analysointiin, kuvaamiseen ja soveltamiseen liittyvien menetelmien kehittäminen ja päätöksentekoa koskevan epävarmuuden vähentäminen näiden menetelmien avulla. Tilastotieteen menetelmiä tarvitaankin kaikilla tieteenaloilla aina soveltavasta fysiikasta taloustieteeseen ja humanistisiin tieteisiin. Tilastotiede ei kuitenkaan ole mikään muiden tieteenalojen aputiede, vaan oma itsenäinen tieteenalansa. Koska eri tieteenalat tarvitsevat erilaisia tilastollisia analyysimenetelmiä, tilastotieteilijät tekevät usein tiivistä yhteistyötä muiden alojen tutkijoiden kanssa. Monet nykypäivänä käytetyistä tilastomenetelmistä ovat syntyneet, kun uudet tutkimuskohteet ovat vaatineet uudenlaisten analyysimenetelmien kehittämistä.

Maailma on täynnä asioita, olioita ja tapahtumia. Ihmiset havainnoivat jatkuvasti ympäristöään ja tekevät tulkintoja ilmiöiden syistä, seurauksista ja lopputuloksista. Erilaisten havaintojen tekeminen ja tulkitseminen kuuluu myös tieteelliseen tutkimukseen. Tutkijan havainnointi poikkeaa kuitenkin merkittävästi arkihavainnoinnista. Tieteellisen tutkimuksen tavoitteena on kerätä **järjestelmällisesti** tietoa luonnon toiminnasta ja muodostaa kerätyn tiedon perusteella teorioita asioiden ja ilmiöiden välisistä yhteyksistä. Teoriat ovat **malleja** tai yksinkertaistuksia todellisuudesta, jotka mahdollistavat tulevien tapahtumien ennustamisen. Muodostetut teoriat puolestaan ohjaavat tulevien tutkimusten suorittamista. Vastaavanlaista mittaamisen, havaintojen tulkinnan ja teorianmuodostuksen menetelmää voidaan soveltaa monissa asiantuntijatehtävissä – ratkaistavat ongelmat vain vaihtelevat sovellusalueittain. Ensimmäiseksi ongelma muotoillaan, seuraavaksi kerätään ongelman-

ratkaisun kannalta tarpeellista tietoa, ja viimeiseksi hankitun tiedon perusteella tehdään tulkintoja ja pyritään muodostamaan vastaus ongelmaan.

Kaikissa mittaustuloksissa esiintyy vaihtelua. Ihmisten verenpaine, tulehdusarvot tai aivojen koko vaihtelevat tutkittavasta ihmisestä toiseen. Osa vaihtelusta on satunnaista. Pienet muutokset ihmisen vireystilassa tai elimistön perustoiminnoissa voivat aiheuttaa hetkittäisiä, terveyden kannalta merkityksettömiä muutoksia verenpaineessa. Osa vaihtelusta on puolestaan systemaattista. Tupakointi, vähäinen liikunta ja ylipaino ovat kaikki järjestelmällisesti yhteydessä kohonneeseen verenpaineeseen ja sitä kautta sairastumisriskiin. Tilastollisten menetelmien yksi tarkoitus on tällaisen – satunnaisen ja systemaattisen vaihtelun – kuvaileminen ja tarkasteleminen. Mutta kuinka voimme erottaa satunnaisen vaihtelun systemaattisesta vaihtelusta? Monet tutkimusaineistot ovat niin suuria, ettemme pysty hahmottamaan niitä paljaalla silmällä lainkaan. Ja vaikka joskus voisimmekin, on pelkkien käsittelemättömien mittaustulosten ymmärtäminen useimmiten mahdotonta. Tällaisten ongelmien ratkaisemiseen tarvitaan tilastotiedettä ja erilaisia tilastollisia menetelmiä. Kaikkien tilastollisten menetelmien perustavoite on sama – ilmiöitä koskevan epävarmuuden vähentäminen. Tilastotiede käyttää tähän viidenlaisia menetelmiä:

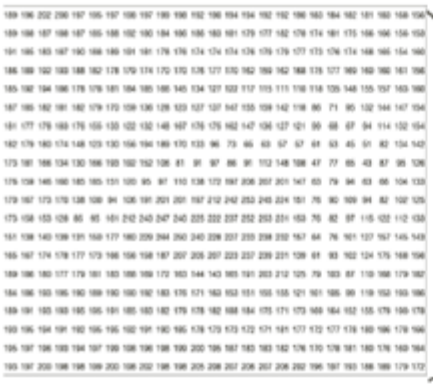
1. **Aineiston rakenteen yksinkertaistaminen ja kuvaileminen**. Mittalaitteilla kerätyt aineistot ovat usein liian monimutkaisia suoraan ymmärrettäväksi. Aineiston ymmärtämiseksi on usein välttämätöntä tiivistää aineistoa ja yksinkertaistaa sen rakennetta. Tämän vuoksi aineistojen yleisiä ominaisuuksia voidaan esittää yksinkertaistetusti **tilastollisten tunnuslukujen** ja erilaisten **kuvaajien** avulla.
2. **Muuttujien välisten yhteyksien tutkiminen.** Suuri osa tutkimuksista keskittyy tavalla tai toisella erilaisten ilmiöiden välisten säännönmukaisten yhteyksien tutkimukseen. Erilaiset muuttujien välisiä riippuvuussuhteita arvioivat menetelmät ovat kenties tilastotieteen käytetyin menetelmäjoukko.
3. **Ilmiöitä koskevien hypoteesien testaaminen ja ilmiöiden yleistyvyyden arvioiminen.** Maailma on täynnä ihmisiä, eläimiä ja molekyylejä. Emme voi mitata millään niitä kaikkia, vaan tutkimuksen mittaukset voidaan suorittaa vain pienelle osajoukolle. Hypoteesien testaamisen avulla voimme arvioida, kuinka todennäköisesti mittaustuloksissamme esiintyvä ilmiö **toistuu** ja **yleistyy** myös muihin tilanteisiin.
4. **Lajittelu ja ryhmittely.** Monet tutkimusongelmat liittyvät asioiden tai mittauskohteiden muodostamien luokkien määrittelemiseen – esimerkiksi siihen, millaisia eläinlajeja tai ihmisryhmiä aineistomme sisältää. Ryhmittelyyn perustuvien tilastollisten menetelmien avulla voimme yrittää löytää mittaustulosten sisältämiä **luokkarakenteita**.
5. **Ennusteiden laatiminen ja mallintaminen.** Pelkkä nykyhetken kuvaileminen on tärkeää, mutta useimmiten haluamme tutkimustulosten avulla myös ennustaa tulevaisuutta. Tilastollisten mentelmien avulla voidaan muodostaa erilaisia **malleja** joita voidaan käyttää tulevaisuuden tapahtumien ennakoimiseen.

Aineiston kuvaileminen ja yksinkertaistaminen

Ihmisen kyky käsittää suuria tietomääriä on rajallinen. Katso esimerkiksi **kuvan 1.3A** numerojoukkoa. Sen sisältämästä tiedosta on lähes mahdotonta saada minkäänlaista selkoa, vaikka numerot sisältävätkin merkityksellistä tietoa – numeromatriisi esittää nimittäin pientä osaa **kuvasta 1.3B,** tarkalleen ottaen tytön oikeaa silmää (**1.3C**). Kun vertaat matriisia ja sen alla olevaa silmän kuvaa, pystyt huomaamaan yhtäläisyyksiä – niissä kohdissa, missä on silmän kovakalvoa, on suuria numeroita, koska kuva on näistä kohdin kirkas. Hieman pienemmät numerot puolestaan kuvaavat kasvojen ihoa, joka on vähemmän kirkas kuin silmän kovakalvo. Kaikkein pienimmät numerot puolestaan kuvaavat tummia kulmakarvoja ja silmän pupillia. Itse asiassa digitaalikamera tallentaa kuvan samaan tapaan pikselikohtaisina numeroarvoina. Vaikka kuvassa ja matriisissa on täsmälleen yhtä paljon tietoa, niiden käsittely vaatii aivan erilaista ponnistelua – suurten lukujoukkojen käsitteleminen on ihmiselle vaikeaa myös silloin, kun tiedämme etukäteen mitä numerot kuvaavat. Ja useimmiten tutkimuksessa ei etukäteen tiedetä, mitä tutkimukseen liittyvät mittaustulokset pitävät sisällään. Tämän vuoksi tarvitsemme samanlaisia työkaluja kuin digitaalisen kuvan sisältämän numerotiedon muuttaminen kuvaksi. Niiden avulla voimme tehdä ihmissilmälle alun perin **näkymättömistä ilmiöistä näkyviä**. Ja kuten valokuvaa katsottaessa, myös tilastolliset menetelmät auttavat meitä katsomaan dataa **riittävän kaukaa.** Näemme **kuvan 1.3B** tarkkana, koska emme katso sitä liian läheltä. Sen sijaan **kuvassa 1.3C** kuvan pikselit erottuvat, ja silmää on hankala erottaa. Mutta jos siirrät kirjaa riittävän kauas, myös silmän kuva on helppo hahmottaa. Tilastollinen kuvaileminen toimii samaan tapaan – sen avulla voimme ottaa tarpeeksi etäisyyttä tutkimusaineiston numerotietoon, jotta voisimme kirjaimellisesti nähdä metsän puilta.

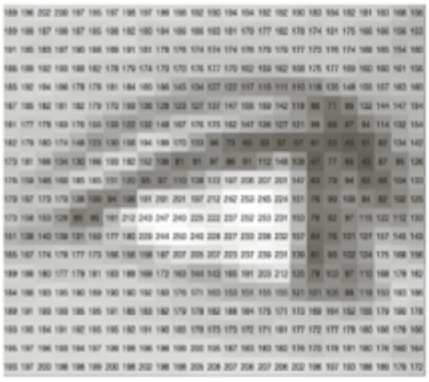
Suuri osa tilastollisen tutkimuksen tekemisestä liittyy mittaustulosten **kuvailemiseen** mahdollisimman helposti ymmärrettävällä tavalla. Kun pystymme ymmärtämään mittaustuloksien sisältämät ilmiöt paremmin, pystymme tekemään parempia tulevaisuutta koskevia päätöksiä. Esimerkiksi köyhyyden poistaminen on tärkeä ihmisten hyvinvointia edistävä tekijä, mutta miten köyhyydelle on maailmassa käynyt? Uutisista saamme jatkuvasti kuulla tuloeroista, eriarvoisuudesta ja luonnonkatastrofeista. Tällainen tietotulva saa meidät helposti **kokemaan**, että asiat ovat menossa maailmassa huonompaan suuntaan. Tutkija ei kuitenkaan luota kokemukseen vaan **järjestelmällisesti kerättyyn tietoon**. Talous toimii hyvinvoinnin polttoaineena, joten Maailmanpankki (http://www.worldbank.org/) kerää jatkuvasti esimerkiksi köyhyyteen liittyvää maakohtaista tietoa, minkä perusteella voimme helposti arvioida, mitä köyhyydelle on todellisuudessa tapahtunut. Emme millään pystyisi hahmottamaan aineistoa, jossa olisi jok'ikisen maapallon asukkaan tulot. Niinpä onkin käytännöllistä, että Maailmanpankki on jo käyttänyt tilastollisia menetelmiä **tiivistääkseen** köyhyyttä koskevaa tietoa. **Taulukossa 1.1** on esitetty absoluuttisen köyhyysrajan (alle $2 päivässä) alla elävien ihmisten lukumäärä vuositasolla muutamalla maapallon köyhimmällä alueella. Tällaisessa taulukossa on miljardien havaintojen sijaan vain muutamia kymmeniä havaintoa. Taulukkoa silmäilemällä vaikuttaisi siltä, että näillä alueilla lukumääräisesti köyhiä on ollut eniten Itä- ja Etelä-Aasiassa ja Saharan eteläpuolisessa Afrikassa. Yhteenlaskettujen köyhyydessä elävien ihmisten määrän perusteella näyttäisi siltä, että köyhyys on

A) Kuvan osa numeroina



B) Alkuperäinen kuva

C) Kuvan osa



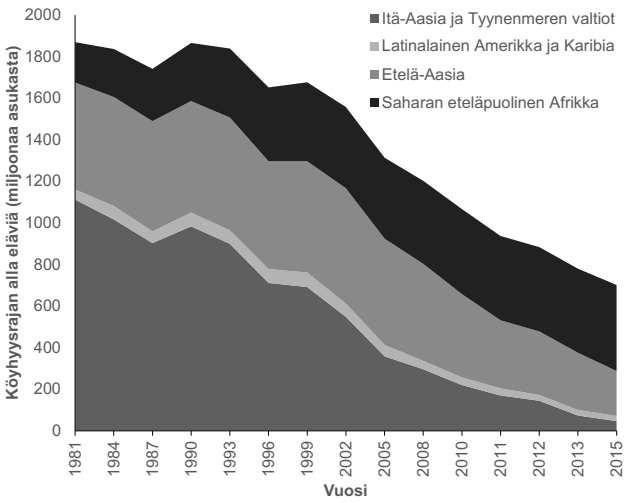
Kuva 1.3. Kuvan hahmottaminen numerojoukkona ja pikselien kirkkautena.

kokonaisuudessaan vähentynyt merkittävästi – seurantajaksolla yli **miljardi (!)** ihmistä on noussut pois absoluuttisesta köyhyydestä! Tällä mittarilla arvioituna ihmiskunnan kehitys ollut erittäin myönteistä.

Tietenkin taulukon tutkiminen on edelleen hankalaa – lukujen keskinäinen vertaileminen ja niiden suuruusluokan arvioiminen vaatii melkoista ponnistelua. Tämän vuoksi tilastollisen tiivistämisen ja kuvaamisen apuna käytetään usein **tilastollista grafiikkaa**. Ihmisaivot ovat paljon taitavampia käsittelemään pinta-aloja, etäisyyksiä ja värejä kuin numeroita. **Taulukossa 1.1.** esitetty aineisto on helpompi hahmottaa, jos se esitetään graafisessa muodossa **kuvaajana (Kuva 1.4)**. Tässä **aluekuvaajassa** x-akselilla on kuvattu aika ja y-akselilla äärimmäisessä köyhyydessä elävien ihmisten määrä. Eri maantieteelliset alueet on esitetty erivärisinä monikulmioina, joidenka korkeus jokaisessa kuvaajan kohdassa kertoo köyhyyden määrän kyseisellä alueella. Monikulmiopinon yhteenlaskettu korkeus puolestaan kertoo köyhien kokonaismäärän. Tästä kuvaajasta näkyy selvästi, että köyhyys on vähentynyt tasaista tahtia ja että väheneminen on ollut kaikkein nopeinta Aasiassa. Sen sijaan Saharan eteläpuolinen Afrikka vaikuttaisi pudonneen kehityksestä – siellä köyhyys on itse asiassa lisääntynyt hieman tarkastelujakson aikana.

Taulukko 1.1. Äärimmäisessä köyhyydessä elävien ihmisten lukumäärä (miljoonaa asukasta) kaikkein köyhimmillä alueilla.

Vuosi	Itä-Aasia ja Tyynenmeren valtiot	Latinalainen Amerikka ja Karibia	Etelä-Aasia	Saharan eteläpuolinen Afrikka	Yhteensä
1981	1112	50	514	194	1869
1984	1017	65	524	230	1836
1987	903	56	530	251	1741
1990	984	65	536	280	1865
1993	899	65	542	332	1838
1996	711	67	518	356	1651
1999	692	69	534	381	1677
2002	549	63	555	391	1558
2005	359	55	510	389	1313
2008	296	40	467	399	1202
2010	221	36	401	409	1066
2011	170	34	328	406	938
2012	144	29	305	406	883
2013	73	28	274	404	779
2015	47	24	217	413	701



Kuva 1.4. Äärimmäisessä köyhyydessä elävien ihmisten lukumäärä (miljoonaa asukasta) kaikkein köyhimmillä alueilla vuosina 1981–2015 Maailmanpankin (http://worldbank.org) tilaston perusteella.

Pelkästään yhden tunnusluvun tarkasteleminen ei tietenkään riitä siihen, että vakuuttuisimme maailman tilan olevan menossa parempaan suuntaan. Siksi onkin hyvä tutkia samaan tapaan myös muita kuin pelkkiä taloudellisia muuttujia. **Kuvaan 1.5** on piirretty lisää maailman tilaa kuvaavia indikaattoreita elämän eri osa-alueilta. On selvää, että lukutaidottomien ihmisten määrä on romahtanut 1940-luvun jälkeen (**A**); samoin lapsikuolleisuus

Tilastollisten menetelmien merkitys tutkimuksessa kasvaa jatkuvasti. Tilastotiedettä tarvitaan kaikilla aloilla tutkimusaineiston keräämiseen, kuvailemiseen ja mallintamiseen sekä riskien arvioimiseen ja ennusteiden laatimiseen.

Teoksessa esitellään kattavasti keskeisimmät tilastolliset analyysimenetelmät alkaen aineiston keräämisestä ja kuvailemisesta sekä perusanalyyseista aina tilastollisiin monimuuttujamenetelmiin ja tilastolliseen luokitteluun asti.

Tilastollisten menetelmien teoria esitetään havainnollisesti, joten aloittelijakin pääsee analysoimaan erilaisia aineistoja itsenäisesti. Havainnolliset laskuesimerkit helpottavat asioiden omaksumista. Monipuoliset ohjelmistoesimerkit perustuvat kattavaan ja ilmaiseen R-ohjelmistoon, ja ne sisältävät selkeät esimerkkikoodit sekä ohjeet tulosteiden tulkinnasta.

Teos soveltuu yliopistojen ja ammattikorkeakoulujen tutkimuksen teon ja tilastollisten menetelmien perusoppikirjaksi lääketieteen, tekniikan, yhteiskunta- ja taloustieteiden sekä käyttäytymis- ja luonnontieteiden aloilla. Kattavuutensa ja monipuolisten esimerkkiensä ansiosta kirja sopii erinomaisesti myös varttuneemman tutkijan hakuteokseksi.

Kattava ja ajantasainen perusteos tilastollisten menetelmien sovelluksista opiskelijoille ja tutkijoille.

www.tammi.fi | 31 | ISBN 978-952-04-0138-2

Lauri Nummenmaa (FT) on maamme johtavia aivotutkijoita ja tilastollisen analyysin asiantuntijoita. Hän toimii lääketieteellisen kuvantamisen ja mallintamisen professorina Turun yliopistossa, jossa Nummenmaan ryhmä tutkii ja mallintaa aivojen toimintaa molekyylitasolta alkaen.