Brynolf Pedantziuksesta data-analyytikkoon

Antti Penttinen
tilastotieteen professori, Jyväskylän yliopisto

Tilastotiede on monelle lähes tuntematon tiede, lähinnä se mielletään kuivaksi alaksi. Tätä käsitystä vahvistaa Toivo Särkän v. 1962 ohjaama ja Suomi Filmin tuottama romanssi ”Ihana seikkailu”. Siinä pääosassa on Brynolf Pedantzius (Esko Mannermaa), silkkihansikkaissa kulkeva, alun perinkin kosiskelupuuhissaan tunariksi tuomittu. Hänestä oli tehty tosi kuiva tyyppi. Jos katsoja ei Pedantziuksen nimestä satu heti saamaan tätä kuvaa, niin hänen ammatikseen oli varmuuden vuoksi laitettu Tilastoviraston kamreeri!

Onko niin, että tilastotiede on tylsä ala ja tilastotieteilijät törttöjä fakki-idiootteja?

Vaikka tilastotiede on huonosti tunnettu tieteenala, on se ollut pitkään läsnä arkielämässämme. Itse asiassa sana ”tilasto” on otettu käyttöön suomen kielessä jo v. 1864. Ehkä tunnetuin esimerkki, jo ajalta ennen tietokoneita, on henki-, palo-, tapaturma- ja vahinkovakuutukset. Nämä ovat tilastotieteeseen perustuvia turvaverkkoja. Toinen ja nykyisinkin keskustelua aiheuttava ala on eläkejärjestelmät. Työeläkkeet tulivat suomalaiseen yhteiskuntaan v. 1961 ja niiden isäksi mielletään professori, vakuutusmatemaatikko Teivo Pentikäinen (1917–2006). Eläkejärjestelmää tosin yritettiin pystyttää jo aikaisemmin, mutta 2. maailmansota, sitä seurannut inflaatio, ja erityisesti työeläkejärjestelmän vaatiman vakuutusmatemaattisen osaamisen vähäisyys siirsivät uudistusta. Myös kansantalouden kirjanpito nojautuu tilastotieteeseen, samoin luonnonvarojen inventointi tieteellisin menetelmin alkaen valtakunnan metsäinventoinnista 1920-luvulla. Arkielämän ulkopuolella tilastotiede on ollut tieteellisen tutkimuksen työkalu lähes kaikilla empiirisiä aineistoja käyttävillä tieteenaloilla, mm. psykologian, maatalouden ja lääketieteen tutkimuksessa.

Tilastotieteen läpimurto yhteiskunnassa on tapahtunut melko hitaasti tietokoneiden aikakaudella 1960-luvulta alkaen, kuitenkin rajusti kasvaen noin viimeisen 15 vuoden aikana ja kasvuvauhti on kiihtyvää. Otan joitakin esimerkkejä tämänhetkisistä haasteista: maailmanlaajuiset ilmasto-ongelmat, EU:n kalastuskiintiöt, rahoitustoiminta ja riskienhallinta (pankki ja vakuutus), kaupan asiakasprofilointi ja ennusteet, täsmämainonta, väestön terveys, teollisuuden prosessien hallinta ja optimointi sekä eri alojen seurantajärjestelmät. Ulkomaisia esimerkkejä seuraten myös tiedotusvälineet tulevat lisäämään jalostetun tilastotiedon käyttöä uutisoinnissa. Yhteistä näille sovelluksille on se, että havainnoista tai rekistereistä saadun informaation perusteella ratkotaan ongelmia ja parannetaan tuottavuutta. Ei tuulipuistoa perusteta tekemättä riittävää määrää tuulimittauksia tuottavuusarvioita varten. Siitä ei selvitä ilman tilastollista mallinnusta.

Tilastotieteellisen läpimurron seuraus on, että käsite ”tilasto” tai ”tilastoaineisto” on häviämässä korvautuen uusilla käsitteillä kuten ”data” tai ”tietovaranto”. Myös tilastoaineisto on monimuotoisempi sisältäen digitaalisia kuvia (joita myös analysoidaan), karttoja ja jopa tekstejä, ei pelkästään numeroita. Myös ammattinimike ”tilastotieteilijä” on häviämässä ja tilalle on tullut ”data-analyytikko” (tai data scientist).

Data-analyysin kasvu luo työpaikkoja. Tarvitaan mallintajia, aineiston hankinnan osaajia ja tietokantaeksperttejä, kuvankäsittelijöitä, kokeiden suunnittelijoita ja tilastoanalyysin ohjelmistojen tuntijoita. Perinteisen tilastotieteilijän vahvuus on siinä, että hän ottaa huomioon sen, miten aineisto on syntynyt, osaa käsitellä aineistossa olevia oikkuja kuten puuttuvaa tietoa ja osaa tulkita tuloksia. Osaajista on pula. Oma arvioni on, että tilastotieteen ja data-analyysin osaajien koulu-

tusta voitaisiin lisätä 2,5–3-kertaiseksi ja työmarkkinat edelleen vetäisivät. Kun otetaan huomioon tiedonhankinnan menetelmien huima kehitys mm. automaation ja internetin ansiosta, tämä arvioni voi olla aivan liian varovainen.

Miten sitten pitäisi kouluttautua ammattiin. Koska ala on jossain määrin piilossa, tarvitaan ”kipinä”. Jokaisella tilastotieteen opiskelijalla on oma tarinansa. Itse koin tällaisen elämyksen lukiolaisena joskus 1960-luvun lopussa. Naapurin poika teki suomen kielen kirjallisuudesta pro gradu -tutkielmaa, siis humanistista tutkimusta. Hän oli mitannut Keski-Suomi-lehdestä kirjallisuusarvostelujen määrää palstamillimetreinä ja tietysti oli sisällötkin pannut muistiin. Hän pyysi minua esittämään graafisesti mittaamansa palstamillimetrien määrän kehittymisen ajassa, jolloin hänen mielestään tärkeää informaatiota voitaisiin saada helpommin luettavaksi. Tein työtä käskettyä, vaikka en ollut kuullut tilastotieteestä yhtään mitään. Graafisen aineistoesityksen tulos oli jännittävä: lehden perustamisen alkuaikoina 1870-luvun puolivälissä kirjallisuusarvosteluja oli poikkeuksellisen paljon, samoin v. 1886. Gradunvääntäjän avulla syykin selvisi: ensimmäisellä aktiivisella jaksolla toimitusta avusti Minna Canth ja toisella Juhani Aho toimi lehden päätoimittajana. Tämä oli hieno juttu, ja mieleeni jäi kytemään ajatus siitä, mitä kaikkea datojen avulla voisikaan saada selville. Tämä alkeellinen graafinen esitys muuten julkaistiin Keski-Suomi-lehden seuraajassa, Keskisuomalaisessa.

Jonkinlaisesta ”tilastollisesta heräämisestä” huolimatta aloin opiskella fysiikkaa ja matematiikkaa, sekä sivuaineena tilastotiedettä. Kävi selväksi että fyysikko en ole ja matemaatikkoa minusta ei tule, jotain muuta pitää keksiä. Niinpä ajauduin tilastotieteeseen – aika monella kollegalla on samantapainen rekrytointihistoria. Ratkaisevaa on ollut se, että tilastotieteilijällä (tai data-analyytikolla) on aina asiakas, jolla on mielenkiintoinen ongelma, ja tämän ongelman ratkaisemista varten hän on hankkinut tai hankkimassa empiiristä aineistoa. Olen työskennellyt usean alan asiantuntijoiden kanssa ja alojen luettelo on pitkä: ekologia, metsäntutkimus, psykologia, kasvatustiede, teollisuuden prosessit, lääketiede, epidemiologia, kielitiede, tietojenkäsittely, taloustiede, musiikkitiede, ja lähihistoriassa ja -tulevaisuudessa taloushistoria sekä arkeologia. Mukaan mahtuu myös hyvinkin teoreettisia ja matemaattisia projekteja, tietojenkäsittelyä ja ennen kaikkea paljon kansainvälistä yhteistyötä. Innostavat hankkeet eivät lopu kesken, kyvyt pikemminkin. Tämä taitaa olla uteliaan ja itsepäisen ihmisen unelmahommaa! On mistä valita.

Opettajan roolissani mielelläni ohjaan lukijaa lisätietojen hankkimiseen. Suosikkini tilastotieteen kuvaukseen on Peter Digglen ja Amanda Chetwyndin kirjoittama englanninkielinen kirja ”Statistics and scientific method: An introduction for students and researchers” (Oxford University Press, 2011). Tilastotieteen luonnetta ja opiskelua avannee oman laitokseni verkkolinkki `https://www.jyu.fi/maths/opiskelijavalinta/abit/tilastotiede`, jossa mm. tilastotieteen opiskelijat kertovat kokemuksistaan.

En koe olevani Brynolf Pedantzius, enkä missään tapauksessa koe, että datojen ja tilastollisten mallien kanssa työskentely olisi kuivaa. Jo kohtuullisella opiskelupanoksella päästään mielenkiintoisiin haasteisiin ja, kiinnostuksen mukaan, varmasti löytyy se oikea duuni, jota jaksaa tehdä.