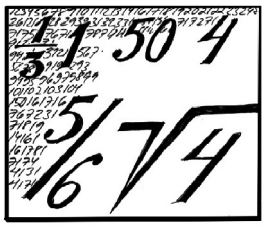


Tunnuslukujen keski- ja epämääräisyyksiä tutkailemassa

Anne-Maria Ernvall-Hytönen
Åbo Akademi

Kimmo Vehkalahti
Helsingin yliopisto

Todennäköisyyslaskennassa ja tilastotieteessä esiintyy monia *tunnuslukuja*, joilla voidaan kuvata jonkin joukon olemusta. Tällaisia ovat muun muassa kolme erilaista keskiarvoa sekä mediaani ja moodi. Käytännössä ylivoimaisesti useimmin näistä kohdataan **aritmeettinen keskiarvo**, jossa joukon alkiot lasketaan yhteen ja jaetaan niiden lukumäärällä. Harvemmin vastaan tulevat **geometrinen** ja **harmoninen keskiarvo**, vaikka kaikki kolme tunnettiin jo antiikin Kreikassa.

Mediaani on *keskimmäinen* arvo, kun joukon alkiot on laitettu suuruusjärjestykseen. **Moodi** puolestaan tarkoittaa joukon *tyypillisintä* eli useimmin esiintyvää arvoa. Siitä käytetään myös nimeä *tyyppiarvo*.

Joukon keskimääräistä olemusta kuvaavia tunnuslukuja tuntuu olevan kovin paljon. Tarvitaanko niitä kaikkia oikeasti? Tutkiskellaanpa tätä aihepiiriä muutaman esimerkin avulla.

Esimerkki 1: kouluarvosanat

Kuvitellaan, että koulutodistuksessa on 11 arvosanaa, joista kuusi on kymppejä, yksi on seiska ja loput neljä viitosia. Todistuksen (aritmeettinen) keskiarvo on siis

$$\frac{6 \cdot 10 + 7 + 4 \cdot 5}{11} = \frac{87}{11} \approx 7,9.$$

Kun laitetaan arvosanat järjestykseen {10, 10, 10, 10, 10, 10, 7, 5, 5, 5, 5}, nähdään helposti, että mediaani ja moodi ovat molemmat kymppejä. Jos todistuksessa olisikin neljä kymppiä ja kuusi viitosta (ja seiska), olisi keskiarvo

$$\frac{4 \cdot 10 + 7 + 6 \cdot 5}{11} = \frac{77}{11} = 7,$$

kun taas mediaani ja moodi olisivat molemmat viitosia. Tämän perusteella tuntuu, että aritmeettinen keskiarvo on varsin fiksu tapa kuvata arvosanajoukon olemusta tiivistetysti. Sen sijaan mediaani ja moodi eivät ehkä vaikuta tässä suhteessa kovin luotettavilta.

Ehkä tavallinen keskiarvo tosiaan riittää. Vai pitäisikö sittenkin tutkia asiaa vielä toisen esimerkin valossa?

Esimerkki 2: keskinopeudet

Turusta Alastarolle ajaa noin tunnissa nopeudella 70 km/h. Jos siis ajaa Turusta Alastarolle ja takaisin tällä nopeudella, kestää matkanteko kaksi tuntia. Kuvitellaanpa, että ajetaankin Turusta Alastarolle nopeudella 35 km/h. Jos takaisin ajaisi nopeudella 105 km/h, niin taittuisiko koko matka jälleen kahdessa tunnissa?

Lukujen 35 ja 105 keskiarvo on ilmiselvästi 70:

$$\frac{35 + 105}{2} = \frac{140}{2} = 70,$$

mutta tämä ei kerrokaan koko totuutta. Jo pelkkä menomatka vie kaksi tuntia, jos nopeus on 35 kilometriä tunnissa, jolloin riippumatta siitä miten pahaa ylinopeutta kaahaa paluumatkan, ei koko matka voi millään taittua kahdessa tunnissa. Mikä tässä nyt meni pieleen? Miksi aritmeettinen keskiarvo ei nyt toimikaan?

Katsotaanpa. Nopeus on matkan ja ajan suhde:

$$v = \frac{s}{t},$$

missä v on nopeus, s matka ja t aika. Tästä voimme ratkaista, että

$$t = \frac{s}{v}.$$

Jos siis ajamme matkan s ensin nopeudella v_1 ja sitten nopeudella v_2, kestää matkanteko yhteensä

$$t = \frac{s}{v_1} + \frac{s}{v_2}.$$

Keskinopeus on luonnollisestikin kokonaismatkan $2s$ suhde kokonaisaikaan $\frac{s}{v_1} + \frac{s}{v_2}$, eli

$$\frac{2s}{\frac{s}{v_1} + \frac{s}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}},$$

eli lukujen v_1 ja v_2 harmoninen keskiarvo.

Jos siis Turusta Alastarolle ajaa nopeudella 35 km/h ja takaisin nopeudella 105 km/h, on keskinopeus, eli nopeuksien harmoninen keskiarvo

$$\frac{2}{\frac{1}{35} + \frac{1}{105}} = 52,5,$$

joka myös vastaa todellisuutta: Matka Turusta Alastarolle ja takaisin nopeudella 52 km/h kestää

$$\frac{140 \text{ km}}{52,5 \text{ km/h}} = 160 \text{ minuuttia},$$

ja toisaalta, matka Turusta Alastarolle nopeudella 35 km/h kestää kaksi tuntia ja paluumatka nopeudella 105 km/h kestää 70 km/105 km/h = 40 minuuttia, eli yhteensä 160 minuuttia (kun jätetään huomiotta se aika, joka kuluu, kun poliisi pysäyttää ylinopeuden vuoksi ja kirjoittaa sakot).

Nopeuksissa siis harmoninen keskiarvo peittoaa aritmeettisen keskiarvon. Tämäkään esimerkki ei kuitenkaan riitä perustelemaan mediaanin tai moodin käyttöä, joten otetaan vielä kolmas esimerkkitapaus.

Esimerkki 3: varat ja velat

Tarkastellaan suomalaisten nettovarallisuutta eli varojen ja velkojen erotusta. Tilastokeskuksen mukaan vuonna 2013 kotitalouksien nettovarallisuuden mediaani oli 110 000 euroa, kun taas keskiarvo oli 195 332 euroa, siis lähes kaksinkertainen määrä. Kumpi tunnusluku – keskiarvo vai mediaani – tässä tapauksessa on luotettavampi ja miksi?

Likemmäksi selvyyttä vie nettovarallisuustaulukko [1], johon on koottu mediaanin ja keskiarvon ohella *fraktiileiksi* kutsuttuja tunnuslukuja. Niistä tyypillisimpiä ovat *ala- ja yläkvartiili*, jotka osoittavat 25 %:n ja 75 %:n kohdalla olevan arvon järjestetystä joukosta lukuja, vastaavasti kuin mediaani osoittaa puolivälin eli 50 %:n kohdan.

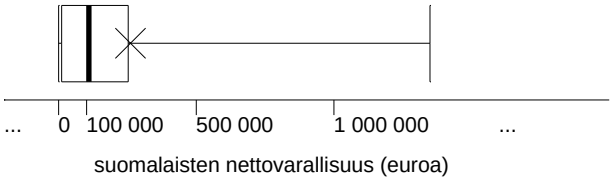
Nettovarallisuuden alakvartiili on 10 000 euroa, mikä siis tarkoittaa, että neljäsosalla kotitalouksista on nettovarallisuutta alle 10 000 euroa. Varakkaimmalla neljäsosalla on yläkvartiilin perusteella nettovarallisuutta yli 252 116 euroa. Ala- ja yläkvartiilin rajaamaan väliin, jonka keskelle mediaani asettuu, kuuluu puolet (75 % − 25 % = 50 %) suomalaisista kotitalouksista.

Keskiarvo näyttää nyt olevan lähempänä yläkvartiilia kuin mediaania. Mistä tämä johtuu? Tunnusluvut vaikuttavat välillä keskimäärin kovin epämääräisiltä!

Sama taulukko sisältää kvartiilien lisäksi muitakin fraktiileja, jotka auttanevat ymmärtämään, mistä on kysymys. Lähempänä joukon äärirajoja ovat 10 %:n ja 90 %:n fraktiilit, joista nähdään, että varakkaimmalla 10 %:lla on nettovarallisuutta yli 458 673 euroa, kun taas varattomimmilla vain alle 50 euroa. (Nettovarallisuus ei näytä todellakaan jakautuvan kovin tasaisesti.) Varakkaimmalla prosentilla (sadasosalla) suomalaisista on nettovarallisuutta yli 1,3 miljoonaa euroa.

Luvassa laatikollinen lukuja (kuvassa)

Hyvä tilastollinen kuva tiivistää tiedot vielä paremmin kuin pelkät tunnusluvut. Oheinen *laatikkokuva* näyttää, miten epätasaisesti suomalaisten nettovarallisuus on jakautunut. Laatikon vasen reuna on alakvartiilin ja oikea reuna yläkvartiilin kohdalla (ks. luvut edellä).



Välissä oleva paksumpi pystyviiva osoittaa mediaanin paikan. Laatikon reunoista lähtevät vaakasuorat viivat ulottuvat vain 10 %:n ja 90 %:n fraktiileihin asti, sillä pienintä ja suurinta nettovarallisuuden arvoa taulukko ei paljasta. (Arvatenkin ne ovat melko suuria negatiivisia ja varsin suuria positiivisia lukuja.)

Yläkvartiilin tietämillä näkyvä rasti osoittaa, missä jakauman keskiarvo huitelee. Tässä se on kaukana mediaanista. Mitä enemmän se siitä poikkeaa, sitä *vinompi* jakauma on. Tällöin keskiarvo ei ole todellakaan nimensä veroinen vaan se saattaa jopa räikeästi vääristää käsitystämme tutkittavan ilmiön luonteesta. Tilastoilla valehtelu [2] on törkeä rike, jolla voi olla huomattavan kauaskantoisia ja vakavia seurauksia.

Myös esimerkiksi palkkajakaumat ovat tyypillisesti varsin vinoja, joten keskiarvo ei ole hyvä keskipalkankaan mitta. Mediaani sen sijaan kertoo luotettavasti keskimmäisen arvon, vaikka ääriarvot olisivat kuinka pieniä tai suuria tahansa.

Varallisuustietoja voi tarkastella myös ikäryhmittäin: alle 25-vuotiaat ovat varattomimpia (alakvartiili jopa negatiivinen eli yli neljäsosalla tästä ikäryhmästä on enemmän velkoja kuin varoja) ja 65–74-vuotiaat varakkaimpia (alakvartiili 80 000 euroa, lähes kaikkien suomalaisten mediaanin verran).

Tarkemmin lukuja ja niiden kuvaamaa todellisuutta valottaa Tilastokeskuksen pääjohtaja *Marjo Bruun* vuonna 2013 julkaistussa haastattelussa [3]. Siitä voi myös katsoa ja vertailla tässä esitettyjä lukuja vuoden 2009 tilanteeseen, jolloin nykyinen talouden alamäki oli jo alkanut. Tunnusluvut antanevat osviittaa taantuman mahdollisista vaikutuksista nettovarallisuuteen.

Johtopäätöksiä ja pohdiskelua

Kaikissa edellä esitetyissä esimerkeissä luvuista saattoi laskea erilaisia tunnuslukuja, kunhan oli tarkkana niistä tekemiensä tulkintojen kanssa. Mediaani ja muut fraktiilit eivät kuitenkaan edellytä laskettavuutta. Niiden käyttöön riittää, että tutkittavan joukon alkiot voidaan järjestää suuruusjärjestykseen.

Se, miksi mediaani tuntui käyttäytyvän oudosti esimerkissä 1, johtui vain siitä, että arvosanoja oli niin vähän. Mediaani ynnä muut *järjestystunnusluvut* toimivat paremmin isommilla aineistoilla. Tällöin ei ole myöskään merkitystä, vaikka lukuja olisi parillinen määrä, jolloin yksikäsitteistä keskimmäistä lukua ei ole. Mediaaniksi kelpaa silloin kumpi hyvänsä keskimmäisistä luvuista (tai niiden keskiarvo, jos laskeminen on mielekästä).

Moodi on vielä vaatimattomampi sen suhteen, mitä se aineistolta olettaa. Riittää, että joukon alkiot voidaan nimetä jollain tunnuksilla (esimerkiksi päärynä, appelsiini ja banaani). Niillä ei tarvitse olla edes mitään järjestystä - moodihan ilmaisee vain, mitä näistä alkioista esiintyy eniten eli mikä on joukon tyypillisin edustaja.

Hyvin usein saatetaan käyttää tunnuksina myös numeroita, mutta tällöin numerot ovat vain koodeja, joilla ei pidä laskea mitään. Vaikka siis edellä mainitut hedelmät koodattaisiin numeroin {1, 2, 3}, ei esimerkin 1 tapaisilla laskelmilla olisi mitään virkaa. Olisikin turvallisempaa käyttää sanallisia koodeja, esimerkiksi {P, A, B}, niin ei tulisi vahingossa tehtyä hölmöyksiä.

Tilastollisessa tutkimuksessa, joka perustuu olennaisesti erilaisten asioiden ja ilmiöiden mittaamiseen, on tärkeää kiinnittää huomiota siihen, miten mitataan. Kyselytutkimuksissa [4] riittää lähes aina kolme *mittaustasoa*: 1) luokittelu, 2) järjestäminen ja 3) (numeerinen) mittaus. Tämä on vähän yksinkertaisempi tapa ajatella kuin kirjallisuudessa yleensä esitetty jako neljään niin kutsuttuun *mitta-asteikkoon*: luokitteluasteikko, järjestysasteikko, välimatka-asteikko ja suhdeasteikko. Käsitteenä "luokitteluasteikko" on hieman onneton, sillä sana "asteikko" antaa ymmärtää, että jotain voitaisiin asettaa (asteikolle) järjestykseen, mikä ei luokittelutasolla nimenomaan ole mahdollista.

On syytä mitata aina mahdollisimman tarkasti, koska tällöin on käytettävissä enemmän erilaisia tunnuslukuja ja muita tilastollisia menetelmiä. Jos tyydytään vain luokittelemaan asioita eri nimisiksi, ei voida käyttää edes mediaania, keskiarvoista puhumattakaan. Pelkkien keskilukujen lisäksi on tarkasteltava, mitä niiden ympärillä tapahtuu, kuten esimerkin 3 tapauksessa.

Erilaisten tunnuslukujen kuten keskiarvon ja mediaanin tärkeä tehtävä on tiivistää tietoa, mutta pelkkien tunnuslukujen tuijottelu on kuin yrittäisi sokeasti hapuilla lukujen seassa: liian paljon jää pimentoon. Luvuista piirretyt tilastolliset kuvat kertovat usein yhdellä vilkaisulla enemmän kuin mitkään tunnusluvut. Yksi *Yogi Berran* viisauksista kuuluukin (vapaasti suomennettuna): *"Voit nähdä paljon pelkästään katsomalla."*

Viitteet

[1] Suomen virallinen tilasto (SVT): Kotitalouksien varallisuus [verkkojulkaisu]. ISSN=2242-3214. Helsinki: Tilastokeskus. `www.stat.fi/til/vtutk/index.html`

[2] `https://fi.wikipedia.org/wiki/Kuinka_tilastoilla_valehdellaan`

[3] `http://www.porssisaatio.fi/blog/2013/10/14/varallisuus-tilastojen-valossa/`

[4] Vehkalahti, Kimmo (2014). *Kyselytutkimuksen mittarit ja menetelmät.* Luku 2: Mittaus ja tiedonkeruu. Helsinki: Finn Lectura.

`https://fi.wikipedia.org/wiki/Aritmeettinen_keskiarvo`
`https://fi.wikipedia.org/wiki/Geometrinen_keskiarvo`
`https://fi.wikipedia.org/wiki/Harmoninen_keskiarvo`
`https://fi.wikipedia.org/wiki/Yogi_Berra`