Artikkeli

Numeeristen tietojen tiivistäminen kuviksi

Numeerisen tiedon visualisointi perustuu suurelta osin muutamiin hyviin peruskuvatyyppeihin ja niiden muunnelmiin. Tärkeimpiä näistä ovat palkki-, pylväs- ja pistekuvat, viivakuvat, hajontakuvat ja laatikkokuvat. Hyvän peruskuvan merkitys on, että se välittää monipuolisesti ja ymmärrettävästi erilaisia tietosisältöjä ajatellulle kohderyhmälleen, mutta antaa myös laajemmalle yleisölle mahdollisuuden ymmärtää monimutkaisiakin ilmiöitä selkeästi. Tilastograafisten kuvien laatimista koskevat selvät säännöt, joiden yksityiskohdat vaihtelevat kuvatyypeittäin. Sääntöjen yhteisenä tavoitteena on se, että kuva ei saa vääristää tietoja. Visuaalisen viestin voima on valtava, ja niinpä virheellisesti laadittu kuva voi johtaa pahasti harhaan. Tämä artikkeli auttaa alkuun sopivan peruskuvatyypin tunnistamisessa ja valinnassa sekä neuvoo näitä kuvia koskevien yksityiskohtaisempien sääntöjen hahmottamisessa.

RAILI SALMELIN, KIMMO VEHKALAHTI

JOHDANTO

Tilastografiikka eli numeerisen, yleensä tilastollisin menetelmin tuotetun aineiston (tilastollisen aineiston) tiivistäminen kuviksi on helposti ymmärrettävä ja tehokas tapa kuvata, tarkastella ja tehdä yhteenvetoa hyvin suurestakin numerojoukosta (Foley ja Van Dam 1983, Tufte 1983). Tilastograafinen kuva, jota tästä eteenpäin kutsutaan yksinkertaisuuden vuoksi vain kuvaksi, ellei ole tarpeen erottaa sitä muista kuvatyypeistä, voi todellakin kertoa enemmän kuin tuhat sanaa tai lukua tekstissä tai taulukossa. Tutkimusaineistoja analysoitaessa esimerkiksi poikkeavan tiedon tai muuttujien riippuvuuden epälineaarisuuden havaitseminen kuvan avulla on helppoa mutta ilman sitä hyvin vaikeaa. Alansa tieteellisiä lehtiä selatessaan lääkärilukijoista noin puolet tarkastelee – otsikon ja tiivistelmän lisäksi – kuvia ja taulukoita; loput lukevat tekstiä, lähinnä tulos- ja pohdintaosaa (Salmelin 1997). Samankaltainen jakauma lienee muissakin koulutetuissa ryhmissä. Näin ollen tieteellisen artikkelin keskeiset tulokset pitäisi esittää sekä kuvina tai taulukoina että, tiivistettyinä, tekstissä, jos haluaa kaikentyyppisten lukijoiden ne huomaavan.

Taulukko sopii kuvaa paremmin useiden yksittäisten, pienehköjen ja kuvailevien tietojen, esim. tutkimusryhmän sosiodemografisten tietojen kuvailuun sekä silloin, kun tarvitaan tarkkoja lukuarvoja. Suurten tietomäärien vertailuun sekä tiedon rakenteiden tai trendien havaitsemiseen kuva sen sijaan on ylivoimainen. Tällaisten kuvien käyttö tieteellisissä lehdissä on kuitenkin yllättävän vähäistä (Salmelin 1997). Sosiaalilääketieteellisen aikakauslehdenkin numeroissa 1/2012 – 1/2014 oli 17 kvantitatiivisia tuloksia esitellyttä artikkelia ja niistä vain seitsemässä oli hyödynnetty tilastograafisia kuvia. Useissa artikkeleissa oli kuitenkin suuria taulukoita tai tekstissä esitettyjä numeerisia tuloksia, jotka olisivat olleet kuvina helpommin omaksuttavissa.

Koska se, mitä näemme kuvana, herättää enemmän mielenkiintoa ja jää paremmin mieleen kuin se, minkä kuulemme tai luemme tekstinä ja numeroina, on kuvien tärkein ominaisuus totuudenmukaisuus, ts. ne eivät saa johtaa katsojaa harhaan. Kuva voi antaa väärän vaikutelman monesta syystä: siinä esitettävä numeerinen tieto on tuotettu väärin tai valittu huonosti, kuvatyyppi on sopimaton ko. tiedon esittämiseen tai kuva on toteutettu harhaanjohtavalla tavalla.

Kuvantekoprosessi alkaa esitettävän tiedon eli tietoalkion valinnalla (esim. sosiaalisten suhteiden mittarin arvo ennen ja jälkeen intervention

interventio- ja verrokkiryhmässä, Joronen ym. 2013). Seuraavaksi valitaan esitettävän tiedon ominaisuuksien ja pääviestin kannalta sopivin kuvatyyppi ja lopuksi tehdään kuva. Jos kuvaa käyttää vain tutkija ja hänen ryhmänsä, riittää toteutuksessa totuudenmukaisuudesta huolehtiminen. Laajemmalle yleisölle tieteellisissä lehdissä, postereissa, esitelmissä yms. esitettävien kuvien on lisäksi oltava pääviestin nopeasti esiin tuovia, kohdeyleisön ominaisuudet huomioon ottavia sekä kaikin puolin hyvin tehtyjä; tällaisilla kuvilla pitää myös olla sisältöä kattavasti kuvaava otsikko. Tavallisimmat, usein toimistoohjelmapaketteihin kuuluvat taulukkolaskenta- ja grafiikkaohjelmat oletusasetuksin käytettyinä eivät välttämättä tuota hyvää tai edes totuudenmukaista lopputulosta, koska ne on tehty kaupalliseen, ei tieteelliseen käyttöön. Kuvan tekijän on siis oltava tietoinen sekä hyvän kuvan vaatimuksista että käyttämänsä ohjelman ominaisuuksista.

AINEISTO JA MENETELMÄT

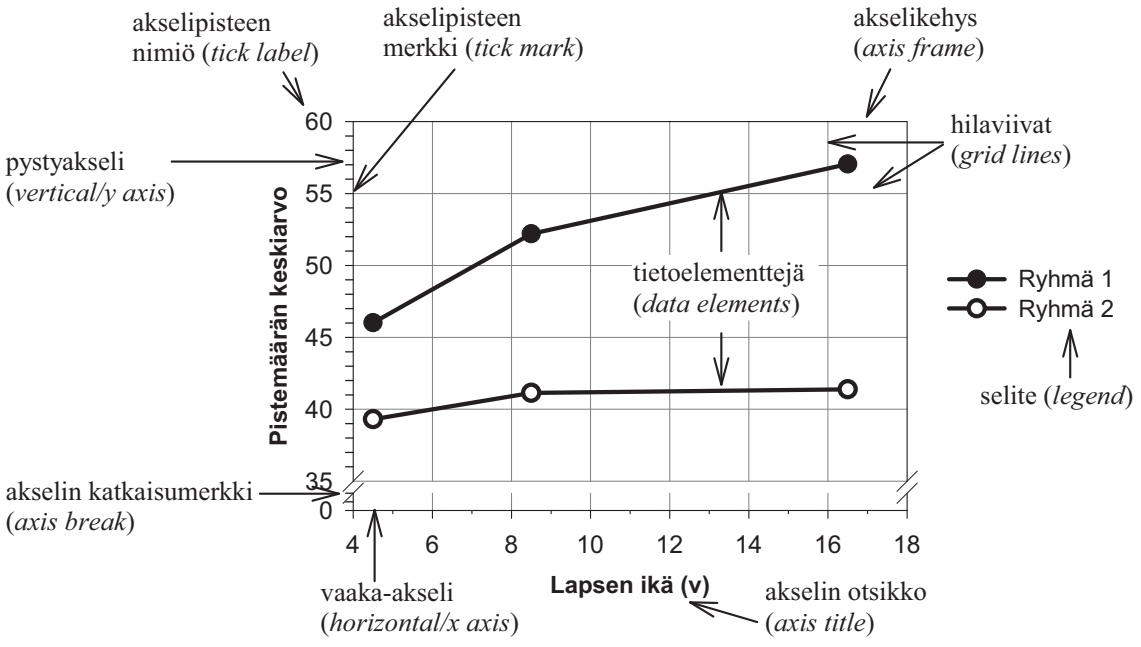
Tässä artikkelissa keskitytään varsinaiseen kuvantekoprosessiin. Esitettävän tiedon tuottamisen ja valinnan virhemahdollisuuksia ei käsitellä.

Esiteltävät kuvatyypit valittiin Sosiaalilääketieteellisen aikakauslehden numeroissa 1/2012 – 1/2014 julkaistujen empiiristen alkuperäisartikkeleiden perusteella. Mukaan valittiin kuvatyypit, joita yleisimmin käytetään tai jotka sopisivat yleisimpien analyyseissä ja tuloksissa esiintyvien tietoalkiotyyppien esittämiseen: yksinkertaiset ja ryhmitellyt palkki- ja pistekuvat sellaisinaan tai vaihteluvälijanoilla täydennettyinä, jaetut palkkikuvat, histogrammi, hajontakuva sekä laatikkokuva. Kustakin kuvatyypistä käydään läpi sen soveltamisen kannalta keskeisiä tekijöitä; kuvatyyppien tavallisin englanninkielinen nimitys mainitaan ohjelmien käytön helpottamiseksi.

Myös yksinkertaisten ja ryhmiteltyjen palkki- ja pistekuvien, jaettujen palkkikuvien ja histogrammin havainnollistamiseen käytettyjen esimerkkikuvien pohjana olevat tiedot valittiin samoista lähteistä. Hajonta- ja laatikkokuvissa esitetyt tiedot ovat peräisin European Social Survey -tutkimuksen kuudennen kierroksen aineistosta (ESS 2012). Kuviin valittiin tiedot siten, että kukin niistä parhaiten sopi juuri kyseisen kuvatyypin ominaisuuksien esittelyyn. Esitetyt kuvat ovat esimerkkejä oikein tehdyistä kuvista, lukuun ottamatta kuviota 3b.

Koska artikkelissa joudutaan toistuvasti viittaamaan tilastograafisen kuvan rakenneosiin, on kuviossa 1 määritelty näille osille nimet. Nimi-

Kuvio 1.
Tilastograafisen kuvan rakenneosat ja niistä tässä artikkelissa käytetyt nimitykset. Kuvan ja kuvatekstin muodostamaan numeroituun kokonaisuuteen viitataan tässä artikkelissa sanalla kuvio. Englanninkielisten grafiikkaohjelmien käyttöä silmälläpitäen on kursivoituna suluissa annettu kunkin käsitteen englanninkielinen nimitys. (Kuva on mukailtu artikkelin Korhonen ym. (2014) kuvasta 2a.)



Kuvio. Äidin eri tutkimusajankohtina täyttämän CBCL:n kokonais-T-pistemäärän keskiarvo valituissa ryhmissä.
kuvion otsikko (*figure caption*)

tykset eivät ole kaikilta osin vakiintuneet sen enempää suomen kuin englannin kielessäkään, joten kirjallisuudessa voi esiintyä muitakin käsitteitä.

PALKKI-, PISTE- JA PYLVÄSKUVAT

Palkkikuva muodostuu yksittäisistä palkeista eli vaakapylväistä (Kuvio 3) tai palkkiryhmistä (Kuvio 5). Vaakapistekuva puolestaan koostuu palkkien pituutta vastaavaan kohtaan sijoitetuista pisteistä (Kuvio 2) ja mahdollisesti niitä täydentävistä viivoista tai varsista (Kuvio 4). Molemmat kuvatyypit soveltuvat jakauman muodon tutkimiseen sekä erilaisten ryhmien ja jakaumien vertailuun. Kuviosta 2 käy selvästi ilmi mm., että suomalaisten kirjoittajien määrä vaihtelee suuresti lehdittäin. Kuviosta 3 puolestaan näkyy, että vauvan päiväunen määrä ei kasva tai vähene systemaattisesti äidin koulutuksen myötä. Kuvio 4 kertoo mm., että hiv-positiivisten kohderyhmässä miehiä oli selvästi enemmän kuin naisia ja että molemmilla sukupuolilla yleisimmät kuolinsyyt liittyivät hiv:iin ja huumeiden käyttöön. Visuaalisesti ilmavampaa pistekuvaa kannattaa käyttää, kun palkkeja olisi niin paljon, että niistä tulisi hyvin kapeita ja kuva näyttäisi ahtaalta. Jatkossa palkki- ja (vaaka)pistekuvista käytetään nimitystä *pp-kuvat*, sillä niitä koskevat paljolti samat ohjeet. Kuvatyypit mainitaan erikseen vain, kun ohjeet poikkeavat toisistaan.

Pp-kuvalla esitetään jäljempänä tarkemmin määriteltäviä numeerisia arvoja – kuten frekvenssejä, mediaaneja tai muita tilastollisia tunnuslukuja – yhden tai useamman luokittelu- tai järjestysasteikollisen ja siten ei-numeerisen *diskreetin* (vain suhteellisen harvoja eri arvoja saavan) muuttujan, *selittäjän* ja mahdollisten *parametrimuuttujien* luokille; esim. viimeisimmän satunnaisen seksikumppanin tapaamismuotojen frekvenssit (Kylmä ym. 2014) tai vauvan päiväunen pituuden mediaani äidin koulutuksen mukaan (Kuvio 3). Kuvissa voidaan myös käyttää vaihteluvälijanoja (Kuvio 6) kuvaamaan palkin tai pisteen edustamaan tilastosuureeseen liittyviä luottamusvälejä, kvartiileja (ks. kohtaa Laatikkokuva), hajontaa tms.

Jos diskreetti selittäjä on numeerinen, siis luokiteltu jatkuva muuttuja kuten luokiteltu ikä tai vain kokonaislukuarvoja saava muuttuja, esim. lasten lukumäärä, käytetään pystysuuntaisia pylväitä tai pisteitä. Tähän ryhmään kuuluvat kuvatyypit soveltuvat pp-kuvia harvemmin Sosiaalilääketieteellisessä aikakauslehdessä tyypillisille aineistoille eikä niitä siksi käsitellä tässä artikkelissa tarkemmin. Poikkeuksena on histogrammi, jota esitellään lyhyesti. Pylväskuviin pätevät pitkälti samat ohjeet kuin pp-kuviin.

Tilanteessa, joissa palkki- tai pylväskuva tai niiden pistemuunnelma on oikea kuvatyyppi, ts. kun esitettävässä tietoalkiossa selittäjä on muu kuin aidosti jatkuva muuttuja, ei tietoalkion esittämiseen pääsääntöisesti saa käyttää viivakuvaa. Jatkuvana ja kulmakertoimien tulkintaan perustuvana se johtaa siinä tilanteessa helposti katsojaa harhaan.

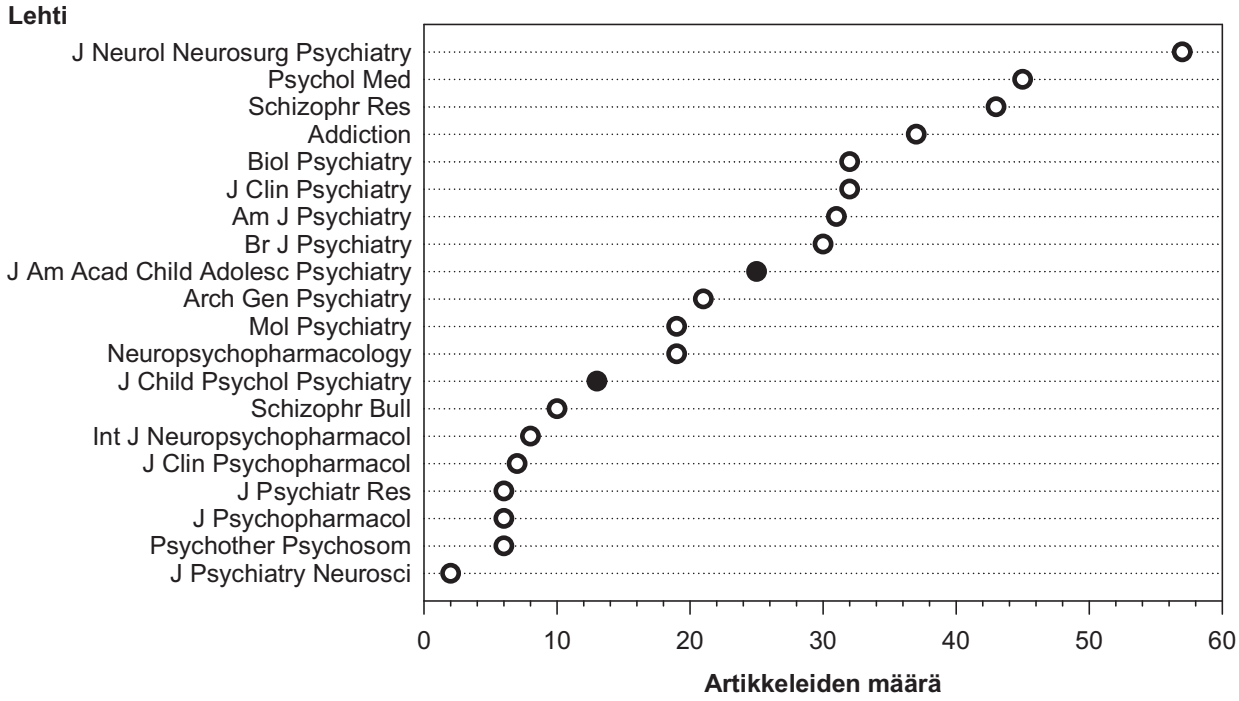
Pp-kuvissa kuhunkin selittäjän luokkaan (Kuviossa 2 lehtiin ja Kuviossa 3 koulutusluokkiin) liittyvän pisteen sijainti tai palkin pituus vastaa kyseisen luokan numeerista arvoa, jonka kertoo numeerinen vaaka-akseli. Selittäjän luokkien nimet, jotka ovat muuttujan luonteen vuoksi sanallisia, ovat ei-numeerisella pystyakselilla vaakasuuntaisina, jolloin ne mahtuvat hyvin ja ovat helppoja lukea toisin kuin vinoon tai suorastaan pystysuuntaan kirjoitetut nimiöt.

Palkki- ja pylväskuvia on kolmea päätyyppiä: yksinkertaiset, ryhmitellyt ja jaetut tai kerrostetut. Kaksi ensimmäistä voidaan toteuttaa myös pisteversioina ja niihin voidaan liittää vaihteluvälijanat. Seuraavaksi kerrotaan kullekin tyypille ominaisista piirteistä, ja lopuksi kuvien toteutukseen liittyviä seikkoja, jotka ovat yhteisiä osalle tai kaikille palkki-, piste- ja pylväskuville.

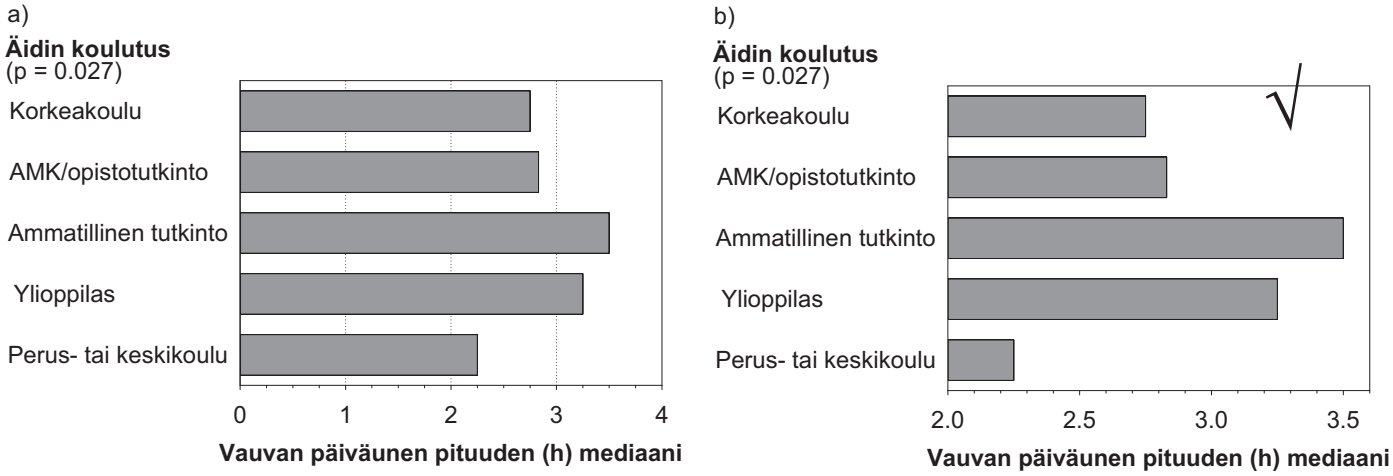
YKSINKERTAINEN PALKKI- JA VAAKAPISTEKUVA (SIMPLE BAR CHART & HORIZONTAL DOT PLOT)

Yksinkertainen pp-kuva sopii 1) yhden ei-numeerisen muuttujan luokkien, esim. viimeisimmän satunnaisen seksikumppanin tapaamismuotojen frekvenssien eli muuttujan frekvenssijakauman esittämiseen (Kylmä ym. 2014), 2) yhden muuttujan (*selitettävä*n eli riippuvan muuttujan) frekvenssien vertaamiseen toisen muuttujan (selittäjän) luokissa, esim. suomalaisia kirjoittajia sisältäneiden artikkeleiden määrä (selitettävä) psykiatrian ydinlehdissä (selittäjä; Kuvio 2) ja 3) yhden muuttujan tilastollisen tunnusluvun kuten mediaanin tai keskiarvon vertailuun selittäjän luokissa, esim. vauvan päiväunen pituuden mediaani (selitettävä) äidin koulutuksen luokissa (selittäjä; Kuvio 3a). Näistä pelkkä yhden muuttujan frekvenssien esittäminen on analyysivaiheessa erittäin hyödyllinen muuttujan jakauman muodon selvittämiseksi, mutta artikkeleissa suhteellisen harvoin tarvittu, koska useimmiten julkaistavissa tuloksissa kuvaillaan kahden tai useamman muuttujan välisiä yhteyksiä.

Kuvio 2.
Psykiatrian ydinlehdissä julkaistut artikkelit, joiden kirjoittajaluettelossa on mainittu ainakin yksi suomalainen taustaorganisaatio. Lehtien nimistä on käytetty Index Medicus -lyhenteitä. Lastenpsykiatrian alan lehdet on merkitty täytetyllä pisteellä. (Kuvassa esitetyn tiedon lähde: Nieminen ja Miettunen (2012).) Kuva on esimerkki yksinkertaisesta vaakapistekuvasta.



Kuvio 3.
Vauvan päiväunen pituuden ja äidin koulutuksen välinen riippuvuus vauvan ollessa 12 kk ikäinen. (Kuvassa esitetyn tiedon lähde: Korhonen ym. (2013).) Kuva on esimerkki yksinkertaisesta palkkikuvasta. Kohdan a kuva on oikein tehty, kun taas kohdassa b nollaa suuremmasta arvosta aloitettu vaaka-akseli vääristää palkkien pituuserot ja johtaa siten harhaan.



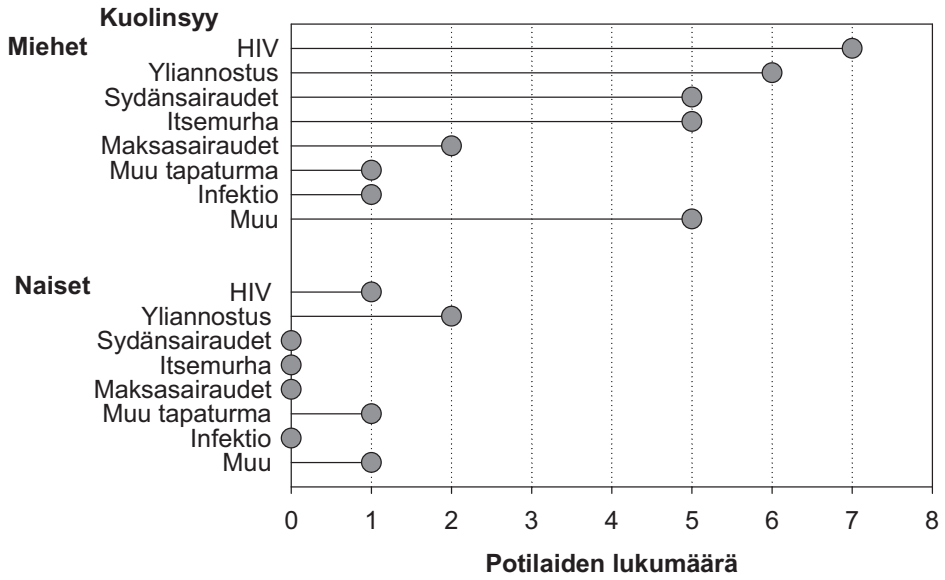
Kuhunkin selittäjän luokkaan liittyy yksi palkki ja palkkien välissä on tyhjää tilaa, ts. palkit eivät ole kiinni toisissaan. Mikäli käytetään pisteitä, on niiden oltava kyllin suuria. Pisteiden kohdalla on hyvä käyttää vaaka-akselin suuntaisia hilaviivoja kuten kuviossa 2, jotta pisteiden ja akselipisteiden nimiöiden (selittäjän luokkien nimien) välinen yhteys olisi helppo havaita. Jos selittäjä on luokittelutasoinen, ts. sen luokilla, kuten Kuviossa 2 psykiatrian alan lehdillä, ei ole luonnollista suuruus-, paremmuus- tai aikajärjestystä, järjestetään luokat ja niihin liittyvät tietoelementit elementin pituuden tai sijainnin eli selitettävän muuttujan arvon mukaan nousevaan tai laskevaan järjestykseen. Jos selittäjän luokilla on jokin luontainen, sisällöllinen järjestys kuten Kuviossa 3 koulutusluokilla, järjestetään tietoelementit tämän järjestykseen mukaan, niiden suuruudesta riippumatta. Jos halutaan helpottaa tietoelementtien pituuden arvioimista, voidaan käyttää pystysuuntaisia hilaviivoja.

Yleensä kaikkien palkkien täyttötapa tai kaikkien pisteiden tyyppi on sama, esim. palkeissa sama harmaan sävy tai samanlainen musta ympyrä kaikissa pisteissä. Jos kuitenkin jokin yksittäinen tietoelementti halutaan erottaa muista tai kiinnittää katsojan huomio siihen, voi ko. palkin tehdä eri sävyllä kuin muut. Pistekuvassa voi käyttää eri pistetyyppiä kuten Kuviossa 2, jossa lastenpsykiatrian lehdet on merkitty täytetyillä pisteillä. On myös mahdollista lihavoida tai kursivoida kyseisen luokan nimi akselilla tai erottaa jokin tietoelementti muista tavallista suuremmalla tyhjällä tilalla. Yksinkertaisessa pp-kuvassa ei kuitenkaan edes yhden tai muutaman luokan korostusta käytettäessä tarvita erillistä selitettä, jossa kahden täyttötavan tai pistetyypin merkitys selitettäisiin, vaan maininta kuvan otsikossa riittää.

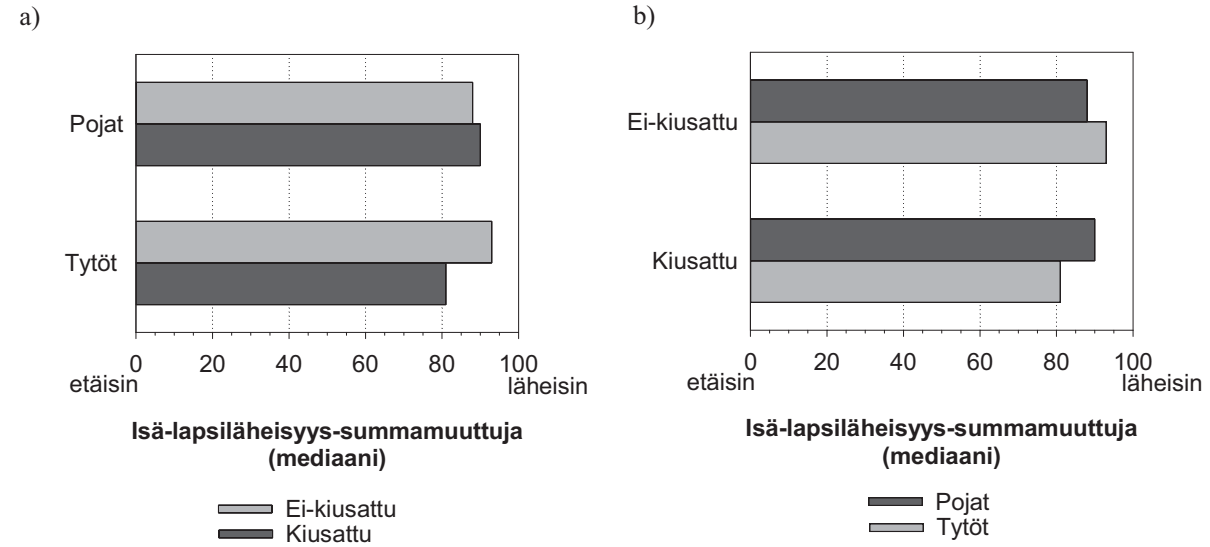
RYHMITELTY PALKKI- JA PISTEKUVA (GROUPED BAR CHART & HORIZONTAL DOT PLOT)

Ryhmitelty pp-kuva on sopiva kuvatyyppi, kun on tarpeen verrata 1) yhden ei-numeerisen selitettävän muuttujan luokkien frekvenssejä ei-numeerisen selittäjän luokissa, esim. hiv-positiivisten potilaiden kuolinsyiden jakaumaa miehillä ja naisilla (Kuvio 4) tai 2) numeerisen, jatkuvaluonteisen selitettävän muuttujan mediaania tms. tilastollista tunnuslukua kahden muun, ei-numeerisen muuttujan (selittäjä ja parametrimuuttuja) luokissa, esim. isä-lapsi-suhteen läheisyyttä kuvaavan muuttujan mediaani ei-kiusatuilla ja kiusatuilla tytöillä ja pojilla (Kuvio 5). Tässä kuvatyypissä on sama määrä tietoelementtiryhmiä kuin parametrimuuttujalla on luokkia (Kuviossa 5a ryhmä pojille ja tytöille) ja kussakin ryhmässä on yhtä monta tietoelementtiä kuin selittäjällä on luokkia (Kuviossa 5a palkki ei-kiusatuille ja kiusatuille). Aina ei ole itsestään selvää, kumpi ei-numeerisista muuttujista on selittäjä ja kumpi parametrimuuttuja. Tällöin on vierekkäisiksi elementeiksi sijoitettava ne, joiden vertailua tutkija pitää ensisijaisena, ja ryhmiksi toissijaisen muuttujan luokat.

Kuvio 4.
Seuranta-aikana kuolleiden hiv-positiivisten potilaiden kuolinsyiden jakauma erikseen miehille (n = 32) ja naisille (n = 5). (Kuvassa esitetyn tiedon lähde: Niemi ym. (2013).) Kuva on esimerkki ryhmitellystä varrellisesta vaakapistekuvasta.



Kuvio 5.
Isän ja lapsen välisen suhteen läheisyyttä mittaavan summamuuttujan mediaani a) ei-kiusatuilla ja kiusatuilla lapsilla, erikseen tytöille ja pojille sekä b) pojille ja tytöille, erikseen ei-kiusatuille ja kiusatuille. (Kuvassa esitetyn tiedon lähde: Söderlund ja Joronen (2013).) Kuva on esimerkki ryhmitellystä palkkikuvasta ja ryhmittelytavan käyttämisestä ensisijaisen vertailun osoittamiseksi (a:ssa ei-kiusattu, kiusattu ja b:ssä pojat, tytöt).



Kuviossa 5a tärkeimpänä vertailuna olisi kiusatuksi tuleminen tai ilman sitä selviäminen ja toissijaisena sukupuolten vertailu, Kuviossa 5b sen sijaan ensisijaista olisi tyttöjen ja poikien vertailu. Yhdessä tietoelementtiryhmässä saisi olla enintään neljä palkkia tai pistettä, muuten ryhmien vertailu ja rakenteiden hahmottaminen vaikeutuu.

Ryhmitellyssä palkkikuvassa kunkin palkkiryhmän palkit, jotka siis vastaavat selittäjän luokkia, ovat kiinni toisissaan mutta palkkiryhmien välissä on tyhjää tilaa (Kuvio 5). Vastaavassa pistekuvassa pisteryhmien välille jätetään suurempi tyhjä tila kuin ryhmään kuuluvien pisteiden välille (Kuviot 4 ja 6). Jos selittäjän luokilla ei ole luonnollista suuruus-, paremmuus- tai aikajärjestystä, järjestetään ensimmäisen tietoelementtiryhmän elementit kyseisen elementin (Kuviossa 4 miesten kuolinsyyn) pituuden mukaan nousevaan tai laskevaan järjestykseen ja muissa ryhmissä (Kuviossa 4 naisilla) vertailtavuuden varmistamiseksi samaan järjestykseen kuin ensimmäisessä ryhmässä, pituudesta riippumatta. Tietoelementtiryhmät (parametrimuuttujan luokat), mikäli niillä ei ole luonnollista järjestystä, puolestaan järjestetään yhden, yleensä ensimmäisen, ryhmään kuuluvan elementin suuruuden mukaan nousevaan tai laskevaan järjestykseen. Jos selittäjän tai parametrimuuttujan luokilla on jokin luontainen järjestys, järjestetään tietoelementit tämän järjestyksen mukaan, niiden pituudesta riippumatta.

Ryhmitellyssä pp-kuvassa kunkin ryhmän tietoelementit on useimmiten (paitsi Kuvion 4 tyyppisessä tilanteessa) tarpeen erottaa toisistaan täyttötavalla tai pistemerkillä; esim. palkit voivat olla vaaleampaa ja tummempaa harmaata kuten kuviossa 5 tai pistemerkkeinä voidaan käyttää esim. harmaan sävyjä edustavia ympyröitä (Kuvio 6). Jos selittäjän luokilla on jokin sisällöllinen järjestys tai merkitys, kannattaa erottelutapa valita vastaamaan sisältöä, esim. vaaleampi harmaa vähemmän ja tummempi enemmän oireilevien ryhmälle. Jonkin tietoelementtiryhmän korostamiseen voi käyttää samoja tehokeinoja kuin yksinkertaisissakin pp-kuvissa. Ryhmitellyssä pp-kuvassa käytetään erillistä selitettä kertomaan, mikä palkin täyttötapa tai pistetyyppi vastaa mitäkin selittäjän luokkaa.

VAIHTELUVÄLIJANALLA (ERROR BAR) TÄYDENNETTY PISTEKUVA (PISTE-VIIKSET-KUVA)

Tilastosuureita kuten mediaaneja tai ristitulosuhteita (OR) esittäviin yksinkertaisiin tai ryhmiteltyihin pp-kuviin on usein hyödyllistä lisätä vaihteluvälijanat kuvaamaan kyseisen tilastosuureen vaihtelua tai luotettavuutta (Kuvio 6). Ryhmien välisten erojen hahmottaminen kuvasta on huomattavasti helpompaa kuin mahdollisesti pitkästäkin listasta kolmen luvun sarjoja. Esimerkiksi Kuvion 6 tiedot sisältäneessä alkuperäisessä taulukossa (Kinnunen ym. 2013) oli seitsemän OR + 95 % luottamusväli -lukusarjaa ja lisäksi neljä ykköstä. Kuvasta on taulukkoa helpompi todeta mm., että suurimman riskin työntekijän uupumustasoiselle väsymykselle aiheuttaa vähäinen

autenttisen johtajuustyylin käyttö; lähes yhtä suuri riskitekijä on loukkaava tai epäoikeudenmukainen johtajuustyyli.

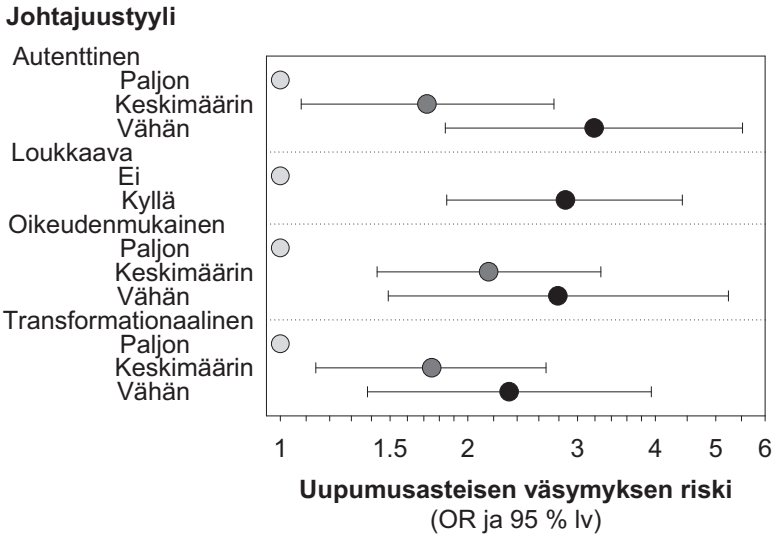
Vaihteluvälijana muodostuu raportoitavaa tilastosuuretta edustavasta pisteestä oikealle ja vasemmalle piirrettävistä ”viiksistä”. Mediaanin yhteydessä esitetään yleensä kvartiilit (ks. Laatikkokuva-kohta), ristitulosuhteen yhteydessä 95 prosentin luottamusväli ja keskiarvon tapauksessa joko hajonta (tai mieluimmin kaksi hajontaa) tai luottamusväli riippuen siitä, halutaanko kuvata oman aineiston hajontaa vai havaitun keskiarvon kykyä ennustaa kohdepopulaation todellista arvoa; kuvan yhteydessä on selkeästi kerrottava, mitä suureita kuva esittää. Yleensä jana kannattaa piirtää molempiin suuntiin, jottei lukijan tarvitse kuvitella toista janaa. Jana voidaan kuitenkin, esim. päällekkäisyyksien välttämiseksi merkitä vain toiseen suuntaan siinä tapauksessa, että se edustaa symmetristä mittaa kuten luottamusväliä tai hajontaa; epäsymmetriset mitat kuten kvartiilit on aina piirrettävä molempiin suuntiin.

Tilastograafisen kuvan kannalta vaativin ja Sosiaalilääketieteellisen aikakauslehden artikkeleissa usein käytetty tietotyyppi, johon vaihteluvälijanalla täydennetty pistekuva soveltuu, muodostuu ristitulosuhteista ja niiden 95 prosentin luottamusväleistä. Itse pistekuva, jossa pisteet edustavat ristitulosuhteita, on samanlainen kuin muitakin tilastosuureita tai frekvenssejä esittävät vaakapistekuvat. Kuvan vaativuus syntyy siitä, että ristitulosuhde luottamusväleineen on luonteeltaan logaritminen; esim. OR 2 (= 2^1) ja 0.5 (= 2^{-1}) kuvaavat yhtä suurta vaikutusta mutta eri suuntaan; toinen on riski, toinen suojaava tekijä, sisällöllisestä tulkinnasta riippuen. Tästä seuraa, että ristitulosuhteiden ja niiden luottamusvälien suuruutta kuvaavan numeerisen, jatkuva-arvoisen vaaka-akselinkin asteikon pitää olla logaritminen, sillä muuten suuruussuhteet vääristyvät. Erot lineaarisen ja logaritmisen asteikon välillä eivät ole kovin suuria pienehköillä (< 10) ristitulosuhteilla, jollaisia tutkimuksissa havaitut arvot usein ovat, mutta oikein tehtyä asteikkoa on hyvä tottua aina käyttämään.

JAETTU PALKKIKUVA (SUBDIVIDED BAR CHART)

Jaettu palkkikuva on käyttökelpoinen kuvatyyppi, kun halutaan näyttää, miten kokonaisuus jakaantuu osiin, ja verrata tuota jakaantumista yhden tai useamman ei-numeerisen muuttujan luokissa. Näissä kuvatyypeissä siis selitettävänä on yksi, toisensa poissulkevista luokista muodostuva muuttuja ja selittäjänä ja mahdollisena parametrimuuttujana ei-numeerinen muuttuja; esim. missä määrin (erittäin tai melko paljon, jonkin verran tai hiukan, ei lainkaan) tietyt tekijät vaikuttivat ETA-alueen ulkopuolella tutkintonsa suorittaneiden lääkäreiden lähtöön Suomeen (Kuvio 7). Jokaista selittäjän luokkaa kohti on yksi jaettu palkki. Esitettävä tieto on aina prosentteina ilmaistuja frekvenssejä ja kaikki yhtä pitkät palkit summaavat 100 prosenttiin; kunkin palkin yhteydessä on hyvä antaa se havaintoyksiköiden lukumäärä, josta prosentit on laskettu. Kuvatyypin kerrostetuksi palkkikuvaksi (*stacked bar chart*)

Kuvio 6.
Vastaajien esimiesten johtajuustyylien yhteys vastaajien uupumusasteiseen väsymykseen. Logistisessa regressiossa on käytetty kovariaatteina sukupuolta, ikää ja terveyttä. (Kuvassa esitetyn tiedon lähde: Kinnunen ym. (2013).) Kuva on esimerkki logaritmiasteikkoisesta piste-viikset-kuvasta.



kutsutussa muunnelmassa selitettävän muuttujan luokkien frekvenssit eli palkin osat ovat havaintoyksiköiden, esim. henkilöiden lukumääriä, jolloin palkin kokonaispituus kertoo kyseiseen selittäjän luokkaan kuuluvien havaintoyksiköiden kokonaismäärän. Kerrostettua palkkikuvaa kuitenkin tarvitaan paljon harvemmin kuin jaettua.

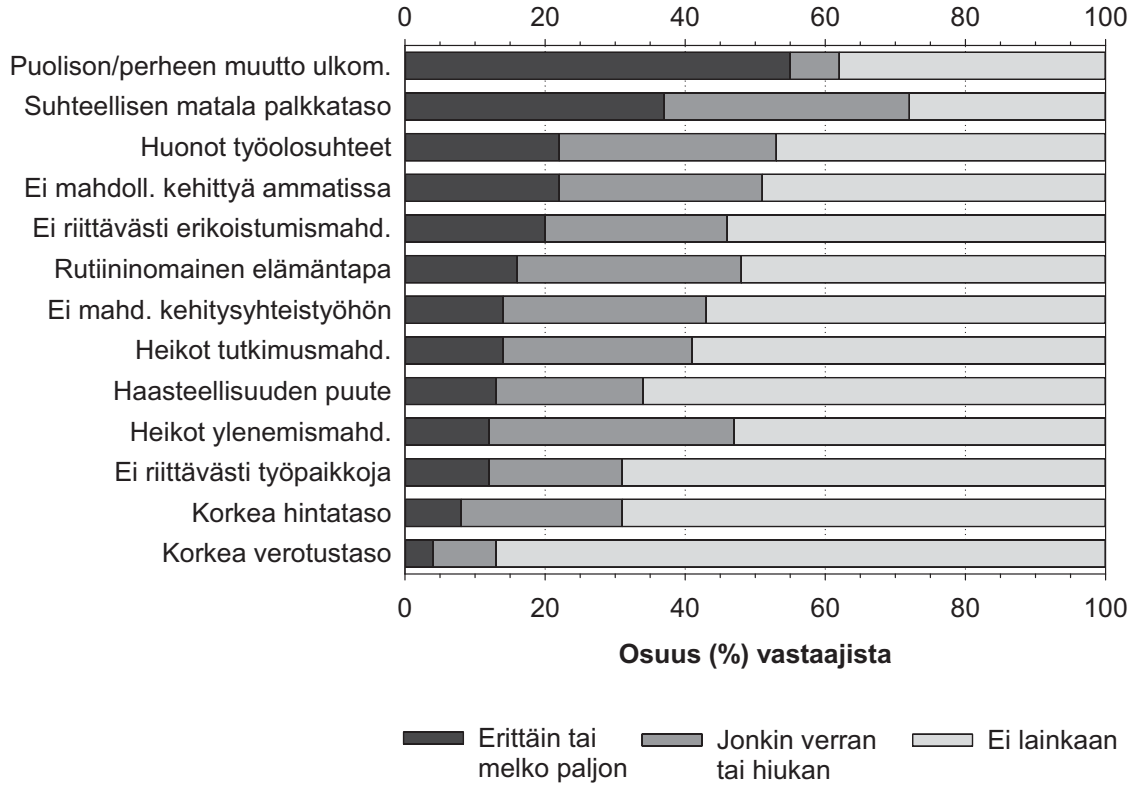
Jaetussa palkissa pitäisi olla mieluimmin 3–4 ja ehdottomasti enintään viisi osaa eli selitettävällä muuttujalla korkeintaan viisi luokkaa, muuten palkkien osoittama jakaumien vertailu vaikeutuu. Jaetut palkit erotetaan toisistaan tyhjällä tilalla kuten yksinkertaiset palkitkin. Jos selitettävän muuttujan luokilla ei ole luonnollista suuruus-, paremmuus- tms. järjestystä, niitä vastaavat palkkien osat järjestetään niin, että osat, joissa on eniten vaihtelua palkkien välillä, ovat äärimmäisinä vasemmalla tai oikealla. Näiden palkkien osien vertailu on nimittäin helpointa, koska ne alkavat arvosta 0 tai päättyvät arvoon 100. Keskellä olevien osien vertailu on selvästi vaikeampaa. Mikäli selittäjän luokilla ei ole luonnollista järjestystä, on niitä vastaavat jaetut palkit hyvä järjestää niin, että äärimmäisenä vasemmalla olevan osan pituus (eli ko. luokan frekvenssi) on nousevassa tai laskevassa järjestyksessä kuten Kuviossa 7, josta näkyy myös äärimmäisten palkkien osien hyvä vertailtavuus. Kuvasta on helppo todeta, että ylivoimaisesti yleisimmät ulkomaille muuton syyt olivat puolison tai perheen muutto sekä kotimaan suhteellisen matala palkkataso ja vähäisin syy korkea verotustaso. Jos selitettävän muuttujan tai selittäjän luokilla on luonnollinen järjestys, palkin osat ja palkit järjestetään sen mukaan.

Jaetun palkin osat erotetaan toisistaan täyttötavalla, esim. harmaan sävyillä. Jos selitettävän luokilla, joita palkkien osat kuvaavat, on jokin sisällöllinen järjestys tai merkitys, kannattaa erottelutapa valita vastaamaan sisältöä kuten ryhmitellyissä pp-kuvissakin. Jonkin palkin korostamiseen voi käyttää samoja tehokeinoja kuin yksinkertaisissakin pp-kuvissa. Myös jaetuissa palkeissa tarvitaan erillinen selite kertomaan, mikä palkin osa vastaa mitäkin selitettävän muuttujan luokkaa.

HISTOGRAMMI (HISTOGRAM)

Histogrammi on pylväskuvan erikoistapaus, jota käytetään, kun esitetään tasavälisesti luokitellun jatkuvan muuttujan luokkien frekvenssit (Kuvio

Kuvio 7.
ETA-alueen ulkopuolella tutkintonsa suorittaneiden lääkäreiden vastausten jakauma kyselytutkimuksessa vuonna 2009 esitettyyn kysymykseen: ”Missä määrin seuraavat tekijät (’epäkohdat’) vaikuttivat päätökseesi lähteä kotimaastasi/opiskelumaastasi Suomeen?”. (Kuvassa esitetyn tiedon lähde: Haukilahti ym. (2012).) Kuva on esimerkki jaetusta palkkikuvasta.



8). Histogrammi siis kuvaa jatkuvan muuttujan luokiteltua jakaumaa ja on siten hyödyllinen erityisesti analyysivaiheessa. Tilastollisten menetelmien valinnassahan on usein oleellista tietää, sopiiko normaalijakauma kuvaamaan muuttujan vaihtelua. Kuviosta 8 näkyy selkeästi (jo ilman normaalijakauman kellokäyrääkin), että EPDS-jakauma on varsin vino ja tässä aineistossa lisäksi kaksihuippuinen.

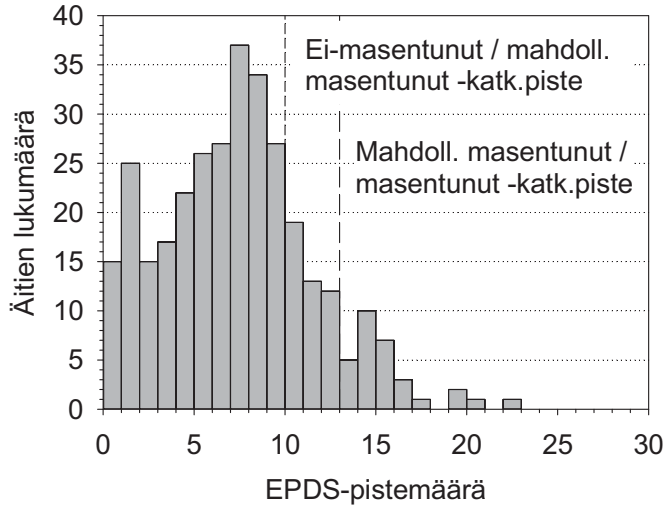
Jatkuvan muuttujan tasavälisten luokkien määrä tai kunkin luokan suuruus riippuu muuttujan luonteesta, arvojen vaihteluvälistä ja havaintoyksiköiden määrästä. Luokkia kannattaa kuitenkin olla riittävästi jakauman muodon (normaali, vino, kaksihuippuinen tms.) saamiseksi näkyviin. Tilastolliset ohjelmistot tarjoavat valmiita ehdotuksia, jotka ovat usein käyttökelpoisia sellaisinaan. Histogrammissa pylväät ovat kiinni toisissaan ja ne kuvaavat yleensä lukumääriä, kun taas tavallisessa pylväskuvassa, jossa selittäjänä on epätasavälisesti luokiteltu tai diskreetti jatkuva muuttuja, pylväiden välissä on tyhjä tila ja frekvenssit kuvataan usein prosentteina kuten palkkikuvassakin. Histogrammin kaikissa pylväissä käytetään samaa täyttötapaa, esim. harmaasävyä.

KAIKISSA PALKKI-, PISTE- JA PYLVÄSKUVISSA MUISTETTAVAA

Pp-kuvissa palkkien pituutta, pisteen sijaintia tai pylvään korkeutta kuvaavan numeerisen akselin on ehdottomasti alettava nollasta. Jos akselin arvot alkavat nollaa suuremmasta arvosta, vääristyvät tietoelementtien mittasuhteet (pituudet, pinta-alat), joiden tehtävä on nimenomaan välittää viesti selittäjän luokkien välisten erojen suuruudesta (Kuvio 3b). Logaritmiasteikko on poikkeus, koska siinä ei voi esiintyä nollaa.

Pp-kuvissa numeerisen vaaka-akselin asteikon on oltava tasavälinen tai, esitettävän tilastosuureen niin vaatiessa, logaritminen. Akselilla pitää myös olla asianmukainen asteikko nimiöityine pistemerkkeineen. Myös pp-kuvan ei-numeeriselle pystyasteikolle kirjoitetaan akselipisteiden nimiöt eli selittäjän ja mahdollisen parametrimuuttujan luokkien nimet, joihin tietoelementit liittyvät, mutta akselipistemerkkejä ei tarpeettomina käytetä (Kuviot 2 ja 3). Pylväskuvassa molemmilla akseleilla pitää olla asianmukainen asteikko nimiöityine pistemerkkeineen. Hilaviivat – vaaka-akselin suuntaisina – ovat välttämättömiä vain runsaasti pelkkiä pisteitä sisältävissä vaakapistekuvissa (Kuvio 2). Muissa palkkikuvissa voidaan käyttää pystyakselin (Kuviot 3–5 ja 7) ja pylväs-

Kuvio 8.
Tamperelaisten ensisynnyttäjä-äitien 1989 raskauden viimeisellä kolmanneksella täyttämän Edinburgh Postnatal Depression Scale (EPDS) -masennusseulan pistemäärien jakauma sekä katkaisupisteet, jotka erottavat äidit ei-masentuneiksi (0-9 pistettä), mahdollisesti masentuneiksi (10-12 pistettä) ja masentuneiksi (≥ 13 pistettä). (Kuvassa esitetyn tiedon lähde: prof. Tuula Tamminen, henk.koht. tiedonanto.) Kuva on esimerkki histogrammista.



kuvissa vaaka-akselin (Kuvio 8) suuntaisia hilaviivoja, jos niiden katsotaan parantavan luettavuutta. Ne sijoitetaan tietoelementtien alle/taakse ja niiden on oltava ohuita, jotteivät ne vie liikaa huomiota tärkeämmiltä tietoelementeiltä.

Pp-kuvan tietoelementtien (selittäjän luokkien) järjestystä päätettäessä kannattaa muistaa, että aakkosjärjestys ei ole luonnollinen järjestys. Silti sen käyttäminen voi joskus olla perusteltua, jos on odotettavissa, että lukija todennäköisesti etsii tietoa sen avulla, esim. oman maansa sijoittumista monien maiden joukossa. Palkkien päähän tai sisään tai jaetun palkin osien sisään ei yleensä tulisi laittaa ko. elementin tai sen osan esittämää lukuarvoa. Kuvan tarkoitus on ensisijaisesti antaa käsitys suuruussuhteista. Jos tarkkoja arvoja tarvitaan, kannattaa useimmiten käyttää taulukkoa.

Palkki- ja pylväskuvassa palkkien ja pylväiden tulee olla selvästi leveämpiä kuin niiden väliin jäävän tyhjän tilan. Hyvä sääntö on, että palkin tai pylvään leveyden ja tyhjän tilan suhde on 2:1. Suurikin tilannekohtainen vaihtelu on kuitenkin mahdollista.

Palkeissa, pylväissä tai niiden jaettujen versioiden osissa voidaan käyttää harmaan sävyjä ja postereissa yms. värejä. Jos värejä ei voida käyttää ja eroteltavia elementtejä on niin monta, ettei riittävän hyvin erottuvia harmaan sävyjä, valkoinen ja musta mukaan lukien, ole mahdollista löytää, voidaan muutamissa elementeissä harkitusti käyttää tiheää mustavalkoista rasterointia eli erilaisia viivoituksia tai pilkutuksia. Jos kuva julkaistaan harmaan sävyillä toteutettuna kuten useimmissa tieteellisissä lehdissä, mukaan lukien Sosiaalilääketieteellinen aikakauslehti, kannattaa kuvissa käyttää alun alkaen harmaan sävyjä, koska muiden värien muuttuminen harmaan sävyiksi on vaikea ennakoida. Pistekuvassa on yleensä helppo löytää suurikin määrä erilaisia pistetyyppejä, koska vaihtoehtoja voi käyttää avoimina tai täytettyinä. Erilaisten palkkien ja pisteiden merkitykset määrittelevä selite kannattaa sijoittaa akselikehyksen sisälle, mikäli siellä on sopivasti tyhjää tilaa, koska tällöin akselikehys ei pienene. Muussa tapauksessa selite sijoitetaan kuvan alle tai joskus sen viereen.

HAJONTAKUVA (SCATTER PLOT)

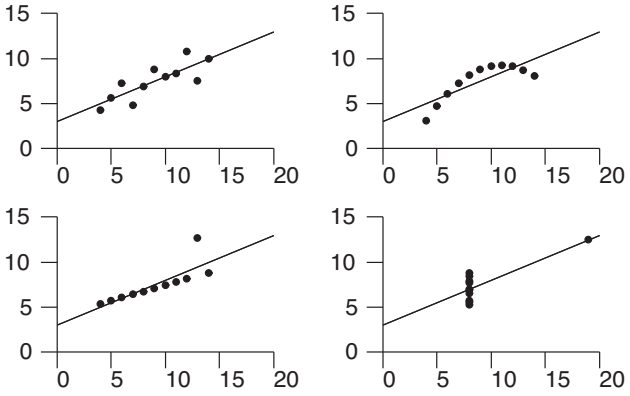
Hajontakuva eli pisteparvikuva on yksinkertainen mutta tehokas tapa tilastollisen aineiston sisältämän vaihtelun ja riippuvuuden yhtaikaiseen tarkasteluun. Perusmuodossaan kaksiulotteinen kuva esittää kahden muuttujan (x ja y) yhteistä jakaumaa niin, että kutakin havaintoarvojen paria (x, y) vastaa yksi piste koordinaatistossa, jonka akselit kuvaavat muuttujien vaihteluvälejä. Kukin piste kuvaa siis yhtä havaintoyksikköä, esimerkiksi yksittäistä ihmistä, ryhmää, yritystä tai valtiota.

On tyypillistä olettaa vähintään toisen muuttujista olevan luonteeltaan melko jatkuva. Usein molemmat muuttujat ovat jatkuvia, mutta myös diskreetit muuttujat soveltuvat kuvattaviksi hajontakuvan avulla. Muuttujien on kuitenkin oltava vähintään järjestysasteikollisia, jotta niiden arvoja on mielekästä esittää joillakin vaihteluväleillä. Piirrettäessä diskreettejä muuttujia voi olla hyödyllistä ”täristää” *(jitter)* hajontakuvaa (eli arpoa pisteille uudet sijainnit alkuperäisen pisteen läheisyydestä, vrt. Kuvio 13), muuttaa pisteen kokoa siihen osuvien havaintopisteiden määrän perusteella tai käyttää muuta piirto-ohjelman tarjoamaa keinoa. Näin voidaan välttää epäinformatiivinen kuva, jossa suuri osa havaintopisteistä osuu samoihin kohtiin. Hajontakuvaa voidaan usein selkiyttää myös sijoittamalla pisteen paikalle havainnon nimi tai tunnus, mikäli sellainen on käytettävissä. Hieman pidemmälle kuvaa voi monipuolistaa kytkemällä pisteen tyyppi, koko tai väri riippumaan jonkin kolmannen muuttujan arvoista.

Kaikkiaan hajontakuva avaa hyödyllisiä näkymiä aineistoon ja tutkittaviin ilmiöihin. Se on usein ylivoimaisesti paras tapa hahmottaa jatkuvien muuttujien välisiä riippuvuuksia, olivat ne sitten suoraviivaisia (lineaarisia) tai käyräviivaisia (epälineaarisia) tai jotain muuta. Lisäksi hajontakuva paljastaa armotta erilaiset poikkeavat havainnot, jotka voivat johtua tallennus-, koodaus- tai mittausvirheistä mutta yhtä hyvin voivat ilmentää myös todellista, joskus yllättävänkin äärimmäistä vaihtelua aineistossa.

On syytä painottaa, että pelkästä korrelaatiokertoimesta ei voi päätellä juuri mitään, joten riippuvuustarkastelut, joissa ainakin toinen muuttujista on jatkuva, on aina aloitettava piirtämällä hajontakuvia. Tällä kohtaa kannattaa muistaa Anscomben (1973) aikoinaan laatima havainnollinen kuvasarja (Kuvio 9), jossa kaikki neljä hajontakuvaa ovat aivan erilaisia, vaikka niitä vastaavista aineistoista lasketut korrelaatio- ja regressiokertoimet ovat täysin samat. Kuvat paljastavat yhdellä silmäyksellä, mistä kulloinkin on kyse: lineaarisuus, epälineaarisuus, poikkeava havainto ja vaihtelun surkastuma. Korrelaation

Kuvio 9.
Anscomben (1973) "kvartetti", jossa hajontakuva paljastaa tehokkaasti aineiston erikoisuudet tilanteissa, joissa tyypilliset tilastolliset tunnusluvut eivät eroa lainkaan. Esimerkiksi kaikkien x-muuttujien keskiarvo on 9 ja varianssi 11 sekä x- ja y-muuttujien korrelaatio 0.82. Regressiokerroin (regressiosuoran kulmakerroin) on vastaavasti joka kuvassa 0.5.



käyttäminen riippuvuuden mittana on perusteltua vain ensimmäisessä tilanteessa.

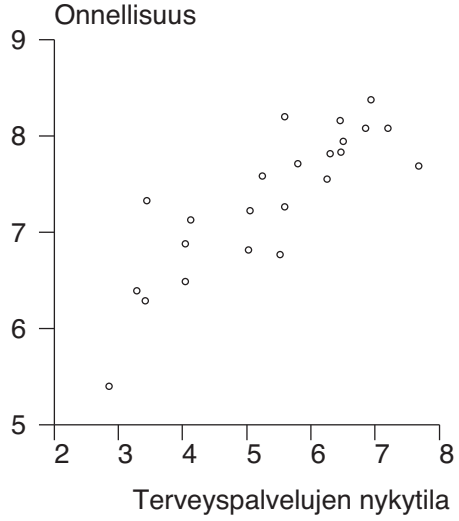
"Anscomben kvartetista" on syytä panna merkille myös toinen asia. Nimittäin jos samassa yhteydessä esitetään useampia hajontakuvia, on välttämätöntä säätää kuvien asteikot samoiksi, muuten kuvien vertailu toisiinsa on täysin mahdotonta. Tämä pätee luonnollisesti muihinkin kuvatyyppeihin kuin hajontakuviin.

Esimerkkinä yksittäisen hajontakuvan piirtämisestä tarkastellaan European Social Survey -tutkimuksen kuudennen kierroksen aineiston (ESS 2012) kahta muuttujaa: maan terveyspalvelujen nykytilaa ja vastaajan onnellisuutta. Näitä asioita on kysytty (lukuisten muiden tietojen ohella) tutkimukseen valituilta vastaajilta kaikkiaan 23 Euroopan maassa.

Molemmat muuttujat on mitattu asteikolla 0–10 (äärimmäisen huono/hyvä ja äärimmäisen onneton/onnellinen). Havainnollisuuden vuoksi alkuperäisestä aineistosta (N = 44243) on tässä laskettu vain maakohtaiset keskiarvot ja tiivistetty siten aineisto havaintomäärältään ainoastaan 23 havainnon kokoiseksi.

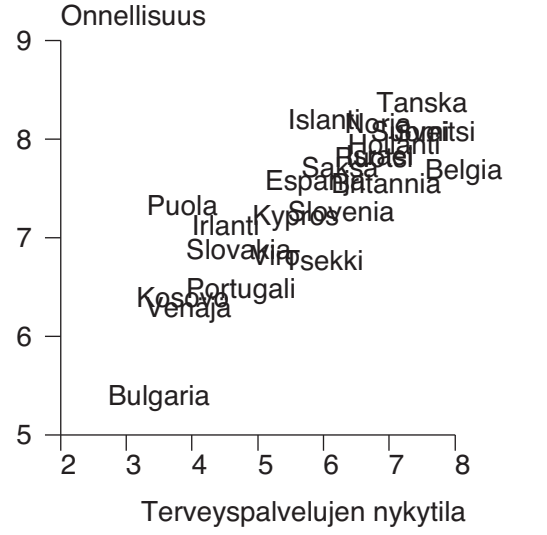
Kuviossa 10 on piirretty muuttujat vastakkain koordinaatistoon siten, että joka maata vastaa yksi piste. Kuvasta nähdään heti, että muuttujat riippuvat toisistaan: maissa, joissa käsitys terveyspalvelujen tilasta on keskimäärin korkeammalla, esiintyy myös keskimäärin enemmän onnellisuutta. Yhteys vaikuttaa melko suoraviivaiselta eli lineaariselta, joten sitä voisi ilmentää myös korrelaatiokertoimella, joka tässä tapauksessa olisi n. 0.84, toisin sanoen n. 70 prosenttia (100 x 0.84^2) toisen muuttujan vaihtelusta on selitettävissä toisella. Selityksen suuntaa ei korrelaatio eikä hajontakuvakaan sinällään voi kertoa, mutta hajontakuvassa on tapana sijoittaa mahdollinen selittäjä vaaka-akselille ja selitettävä pystyakselille. Tässä siis ajatus olisi niin päin, että terveyspalvelujen tila selittäisi (osaltaan) onnellisuutta.

Kuvio 10.
ESS-hajontakuvan lähtökohta: pisteet koordinaatistossa. Kuvassa esitetyn tiedon lähde: ESS (2012).)



Kun kaikkia maita kuvaa samanlainen piste (Kuvio 10), mielenkiinto kohdistuu kuvan yleisnäkymään: riippuvuuden luonteeseen ja mahdollisiin poikkeavuuksiin. Toisessa versiossa (Kuvio 11) pisteiden tilalle on asetettu maiden nimet siten, että nimi alkaa siltä kohtaa, jossa aiemmin

Kuvio 11.
ESS-hajontakuvan seuraava versio: pisteiden paikalla maiden nimet.



oli piste. Yleisnäkymä muuttujien riippuvuudesta on yhä nähtävissä, mutta nyt huomio kiinnittyy enemmän yksittäisiin maihin ja niiden keskinäiseen sijoittumiseen pisteparvessa. Nähdään mm., että Bulgaria on molempien muuttujien suhteen keskimäärin negatiivisin, Belgia terveyspalvelujen kärjessä ja Tanska onnellisuuden huipulla.

Kuviossa 11 saattaa häiritä se, että osa maista ei erotu, kun tekstit menevät osittain päällekkäin. Siitä ei kannata liikaa välittää, sillä pääasiallisena tarkoituksena on riippuvuuden yleisen hahmon kuvaamisen lisäksi nostaa esiin poikkeavuuksia, jotka erottuvat helposti massasta. Tämän tyyppisiä kuvia tutkijan on syytä piirtää runsaasti tutustuakseen aineistoonsa. Aikaa ei kannata siinä vaiheessa hukata liikaa kuvien viimeistelyyn. Sen aika on vasta, kun kuvia aiotaan julkaista.

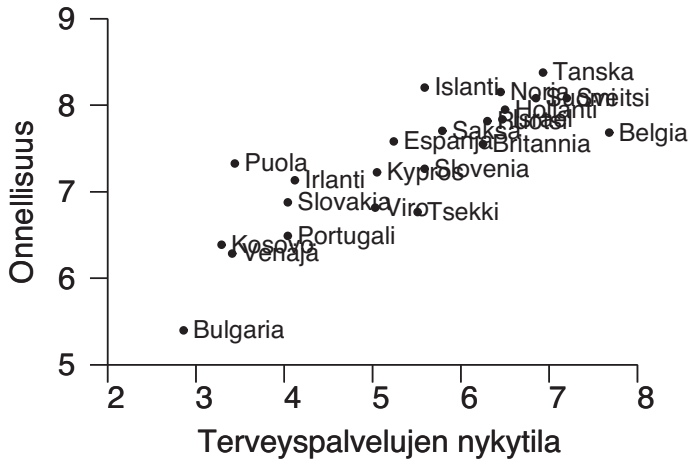
Hajontakuvan julkaisemista varten tärkeää on huolehtia *kuvasuhteesta* eli siitä, että vaaka- ja pystyakselien suhde toisiinsa on järkevä. Tässä muuttujat oli mitattu samanlaisella, 11-portaisella asteikolla, mutta keskiarvotasolla onnellisuudessa on selvästi vähemmän vaihtelua kuin terveyspalvelujen tilassa. Niinpä kuvaa ei ole perusteltua julkaista neliön muotoisena kuten edellä, vaan sitä on säädettävä matalammaksi niin, että akselipisteiden etäisyys vaaka- ja pystyakselilla vastaa toisiaan (Kuvio 12). Huomaa, että akselien numeeriset asteikot valitaan niin, että ne järkevästi kuvaavat muuttujien vaihteluvälejä aineistossa. Hajontakuvassa ei siis ole asteikkojen katkaisuun liittyviä ongelmia. Tehdään samalla kaksi muuta muutosta: palautetaan pisteet kuvaan, mutta jätetään myös maiden nimet sopivalle etäisyydelle havaintopisteistä. Tämä voi auttaa hahmottamaan paremmin myös kuvan yleisilmettä. (Osa havainnoista menee edelleen päällekkäin, mutta siitä ei tarvitse edelleenkään välittää.) Toiseksi sijoitetaan pystyakselin otsikko akselin viereen pystysuoraan, kuten on usein suositeltavaa, jotta on selvää, mihin otsikko liittyy.

Tavallinen, kahden muuttujan hajontakuva on tärkeä työkalu myös monien, pidemmälle menevien tilastollisten menetelmien soveltamisen yhteydessä, sillä menetelmien tuloksena saadaan usein uusia, alkuperäisiä tiivistetympiä ja jatkuvampia muuttujia. Tällaisia uusia muuttujia kannattaa piirtää vastakkain, jolloin näkee taas syvemmälle aineistoon ja ymmärtää sen välittämää tietoa monipuolisemmin.

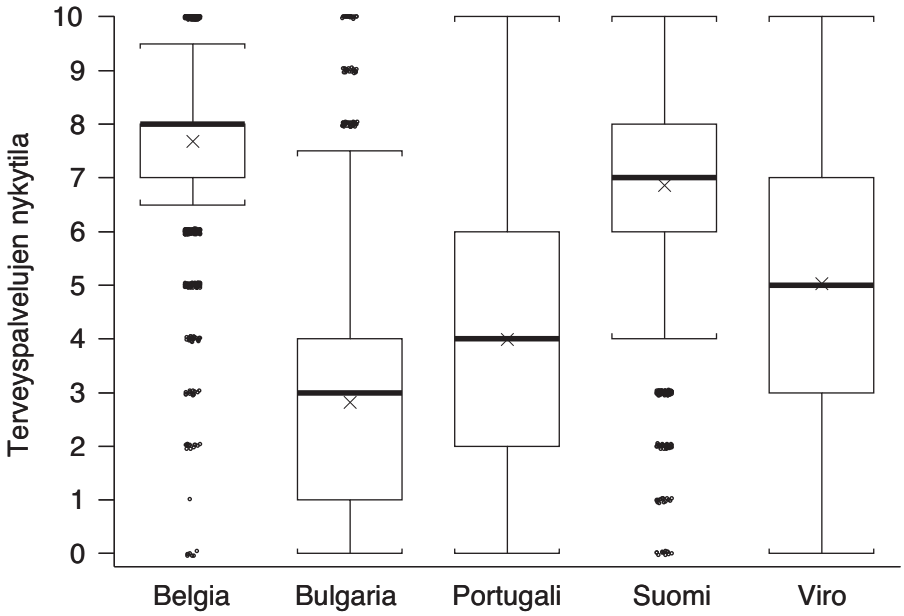
LAATIKKOKUVA (BOX PLOT, BOX-AND-WHISKERS PLOT)

Laatikkokuva, pidemmältä nimeltään laatikko-janviikset -kuva edustaa huomattavasti uudempaa kuvatyyppiä verrattuna pylväisiin, palkkeihin tai viivakuviin, jotka periytyvät vuosisatojen takaa (Tufte 1983, Spence 2005). Laatikkokuvan kehitti *John W. Tukey* osana tilastolliseen aineistoon tutustumiseen laajemmin tähtäävää eksploratiivista analyysia (Tukey 1977). Kuvan ideana on visualisoida jatkuvan muuttujan jakaumaa perustuen viiteen tunnuslukuun, jotka voidaan laskea, kun aineisto on järjestetty tämän muuttujan suhteen. Nämä ns. järjestystunnusluvut ovat *minimi* (pienin arvo), *alakvartiili* (arvo 25 % kohdalla), *mediaani* (arvo 50 % kohdalla eli keskimmäinen arvo), *yläkvartiili* (arvo 75 % kohdalla) ja *maksimi* (suurin arvo). Tunnusluvut eivät välttämättä

Kuvio 12.
ESS-hajontakuvan julkaisukelpoisempi versio: kuvasuhde kunnossa ja sekä pisteet että maiden nimet esitettyinä.



Kuvio 13.
ESS-laatikkokuva: terveyspalvelujen nykytila viidessä eri Euroopan maassa.



ole aina yksikäsitteisiä (esim. parillisilla havaintomäärillä tai pienillä aineistoilla), mutta se on melko epäolennaista. Parhaimmillaan järjestystunnusluvut ovat joka tapauksessa suuremmilla aineistoilla.

Paras hyöty laatikkokuvasta saadaan irti, kun piirretään samaan kuvaan useampia laatikoita (”viiksineen”) kuvaamaan jatkuvan muuttujan vaihtelua jonkin diskreetin muuttujan luokissa. Tällöin päästään helposti vertailemaan luokkia toisiinsa. Tarkastellaan esimerkkinä European Social Survey -tutkimuksen kuudennen kierroksen aineiston (ESS 2012) muuttujaa maan terveyspalvelujen nykytilasta viidessä eri maassa (Kuvio 13). Maat on valittu niin, että samaan kuvaan on saatu mahdollisimman erilaisia, kuvatyyppiä monipuolisesti edustavia näkymiä. Diskreetti muuttuja on siis nyt maa, kun taas terveyspalvelukäsitykset tulkitaan jatkuvaksi muuttujaksi. Tosiasiassa sekin on melko diskreetti muuttuja (vain 11 mahdollista arvoa).

Analyysivaiheessa laatikkokuvia voi piirtää vaaka- tai pystysuuntaisina. Esitysvaiheessa on ei-numeerisen selittäjän tapauksessa hyvä suosia vaakasuuntaa (vrt. palkki- ja pylväskuvat), jolloin muuttujan luokkien nimet mahtuvat paremmin ja ovat helpompia lukea. Pystysuunnassa (kuten analyysivaiheen tilannetta havainnollistavassa Kuviossa 13) on tosin helpompi muistaa, että laatikon alareuna vastaa alakvartiilia ja yläreuna yläkvartiilia (vaakasuunnassa vastaavasti vasen ja oikea reuna). Joka tapauksessa laatikko sisältää aina 50 prosenttia kyseisen luokan (tässä maan) havainnoista, toisin sanoen jakauman keskiosan, jota kutsutaan myös *kvartiiliväliksi*.

Laatikon sisällä oleva paksumpi viiva ilmaisee kyseisen luokan mediaanin sijainnin. Laatikosta lähtevät viivat (eli ”viikset”) voidaan piirtää minimiin ja maksimiin saakka, jolloin ne kuvaavat koko vaihteluväliä. Usein ne kuitenkin sijoitetaan sellaiselle etäisyydelle (esim. 1.5 kertaa kvartiiliväli mediaanista molempiin suuntiin), että jakaumasta n. 95 prosenttia sijoittuu viivojen väliin. Tämä 95 prosentin sääntö pätee sitä paremmin, mitä symmetrisempi muuttujan jakauma on, mutta sitä ei ole tarkoitus ottaa liian tiukasti. Olennaista on, että viivojen osoittaman välin ulkopuolelle jäävät mahdolliset poikkeavammat arvot, jotka on tapana havainnollistaa pisteinä. Näihin pisteisiin (kuten yleensäkin poikkeavampiin havaintoihin) on syytä kiinnittää huomiota, koska ne voivat paljastaa aineistoon sisältyviä virheellisiä tai muulla tavoin yllättäviä tietoja.

Kuviosta 13 voidaan päätellä, että belgialaisilla on parhaat käsitykset maansa terveyspalveluista (vrt. hajontakuva Kuviossa 11). Johtopäätös perustuu siihen, että Belgian mediaanitaso on selvästi korkein, mutta myös vaihtelua on vähemmän kuin muissa maissa. Siitä huolimatta, että Belgiankin osalta yksittäisiä käsityksiä on koko muuttujan vaihteluvälin laajuudelta, on jakauma keskittynyt voimakkaasti arvojen 7 ja 8 tietämille. Koska muuttuja ei ole kovin jatkuva vaan saa

ainoastaan 11 eri arvoa, mediaani ja yläkvartiili ovat Belgian osalta samat. Yksittäisiä pisteitä on ”täristetty” erilleen (vrt. Hajontakuva), jolloin paljastuu, että yksittäisiä pisteitä on jakauman alapäässä paljon vähemmän kuin lähempänä jakauman keskiosaa symboloivaa laatikkoa. (Ilman täristystä pisteet osuisivat täysin samoihin, diskreetteihin kohtiin eikä niiden määrästä voisi päätellä mitään.)

Edelleen Kuviosta 13 nähdään, että Viron ja Portugalin jakaumat muistuttavat toisiaan, Viron mediaani vain on pykälän korkeammalla. Yksittäisiä pisteitä ei kummassakaan esiinny, joten vaihtelua esiintyy runsaammin koko vaihteluvälillä. Tällöin myös niitä vastaavat laatikot ovat korkeampia eli kvartiilivälit laveampia. Asiaan vaikuttaa jälleen myös muuttujan diskreettiys: tällä mittaustarkkuudella 95 prosentin väli kattaa jo koko vaihteluvälin. Suomen ja Bulgarian kohdalla yksittäisiä pisteitä sen sijaan esiintyy, Bulgarialla jakauman yläpäässä ja Suomella jakauman alapäässä. Suomen osalta myös vaihtelu on hieman vähäisempää kuin Bulgarialla. Tasoerot mediaanien osalta ovat ilmiselvät.

Kuvioon 13 on viiden järjestystunnusluvun lisäksi sijoitettu maakohtaiset keskiarvot (kuvattu rasteina), jolloin voidaan arvioida jakauman vinouden vaikutusta tunnuslukuihin. Järjestystunnusluvut ovat ns. *robusteja* eli eivät juurikaan häiriinny vinouksista tai poikkeavista havainnoista, kun taas keskiarvo on tällaisille poikkeamille sangen herkkä. Tässä ei olisi suurta vaaraa keskiarvojen käytössä; ainoastaan Belgian kohdalla on pientä eroa mediaaniin. Muuttujissa, joissa on enemmän vaihtelua ja vinouksia kuin tässä (esim. tulotaso tai EPDS-pistemäärä Kuviossa 8), on käytettävä järjestystunnuslukuja kuten mediaania ja vältettävä keskiarvoa.

Kaikkiaan laatikkokuva on yksinkertainen ja ymmärrettävä kuvatyyppi, kun halutaan havainnollistaa jatkuvan muuttujan jakaumaa. Erityisen tehokas se on tilanteessa, jossa tarkastelu tapahtuu jonkin diskreetin muuttujan luokissa. Tällöin laatikkokuva auttaa erilaisten luokkien tai ryhmien visuaalisessa vertailussa sekä jakaumien keskitason että vaihtelun määrän ja laadun suhteen.

LOPUKSI

Tilastollisen aineiston visualisointi perustuu suurelta osin muutamiin hyviin peruskuvatyyppeihin ja niiden muunnelmiin. Tärkeimpiä näistä ovat palkki-, pylväs- ja pistekuvat, viivakuvat, hajontakuvat ja laatikkokuvat. Näiden lisäksi on lukuisia erikoistuneita visualisointikeinoja, jotka vaativat enemmän sekä tekijältä että lukijalta (esim. Chen ym. 2008). Hyvän peruskuvan merkitys on, että se välittää monipuolisesti ja ymmärrettävästi erilaisia tietosisältöjä ajatellulle kohderyhmälleen, mutta antaa myös laajemmalle yleisölle mahdollisuuden ymmärtää monimutkaisiakin ilmiöitä selkeästi.

Tilastograafisten kuvien laatimista koskevat selvät säännöt, joiden yksityiskohdat vaihtelevat kuvatyypeittäin. Sääntöjen yhteisenä tavoitteena on se, että kuva ei saa vääristää esitettäviä tietoja. Tämä pätee myös kuviin, joita tutkija piirtää itselleen tai tutkimusryhmälleen tutustuakseen aineistoonsa. Visuaalisen viestin voima on valtava, ja niinpä virheellisesti laadittu kuva voi johtaa pahasti harhaan. Monet virheistä ovat tahattomia ja johtuvat mm. ohjelmistojen toisinaan huonoista toimintatavoista (esim. asteikkojen mielivaltainen katkaiseminen), mutta toisinaan ei voi vältyä vaikutelmalta tahallisesta harhaanjohtamisesta. Tässä ei ole mitään uutta; ilmiön tiivisti osuvasti jo *Darrell Huff* (1954) klassikollaan tilastoilla valehtelusta.

Tutkijan tehdessä kuvia vain itselleen tai ryhmälleen ei kannata käyttää aikaa kuvien viimeistelyyn vaan keskittyä siihen, mitä tietoa kuvat sisällöllisesti välittävät. Kuvia on syytä piirtää runsaasti, koska ne tarjoavat arvokkaita näkökulmia aineiston sisältämien riippuvuuksien ym. ominaisuuksien ymmärtämiseen. Laajemmalle yleisölle esitettäviksi tarkoitettujen kuvien on kuitenkin totuudenmukaisuuden lisäksi oltava selkeitä, ihmisen havainnointikyvyn huomioon ottavia ja esteettisesti miellyttäviä. Siksi niiden on täytettävä lukuisia muita vaatimuksia. Esimerkiksi näennäistä eli pseudokolmiulotteisuutta, jolla tavoitellaan syvyysvaikutelmaa pylväissä, palkeissa tai muissa tietoelementeissä, ei pidä käyttää, koska se vaikeuttaa tietoelementtien välisten erojen havaitsemista. Kuvassa ei pitäisi olla mitään muutosta, joka ei vastaa muutosta aineistossa, eikä muutenkaan mitään ylimääräistä koristelua tms. Nimiöiden ja muiden tekstien on oltava lopullisessa koossa luettavia ja tietoelementtien erotuttava selkeästi hilaviivoista, akselikehyksestä ja muista apuelementeistä. Yhteen kuvaan ei saisi laittaa liikaa tieto- tai muita elementtejä, kuvan yleisvaikutelman tulisi olla rauhallinen ja otsikon kuvan viestiä tukeva. Periaatteessa kuvan otsikoineen tulisi olla yhteydestään irrotettunakin ymmärrettävä. Kaikesta tästä seuraa, että hyvien kuvien tekemiseen on varattava aikaa.

Haasteellista hyvien tilastograafisten kuvien laadinnassa on se, että siinä tarvitaan niin substanssialan, tilastotieteen kuin visualisoinninkin taitoja. Toisaalta se avaa luontevia mahdollisuuksia eri asioiden osaajien yhteistyölle. Tunnetusti omalle tekstilleen tulee helposti sokeaksi. Sama pätee kuviin: itse on helppo ymmärtää, mitä on kuvalla halunnut viestiä, mutta toisen voi olla yllättävän vaikea nähdä viestiä samoin. Onkin tärkeää pystyä ilmaisemaan kuvan keskeinen sanoma tiiviisti myös sanallisesti tekstissä tai joissakin tilanteissa kuvan ala- tai yläpuolella olevassa otsikossa. Kannattaa muistaa myös, että laadukas kuva välittää yhdellä vilkaisulla keskeisen sanoman mutta ei rajoitu siihen vaan paljastaa tarkemmalla katsomisella enemmän yksityiskohtia.

Tämä artikkeli auttaa alkuun sopivan peruskuvatyypin tunnistamisessa ja valinnassa sekä neuvoo näitä kuvia koskevien yksityiskohtaisempien sääntöjen hahmottamisessa. Hyviä lisätiedon lähteitä ovat asiaa tarkemmin käsittelevät julkaisut kuten Tufte (1983), Cleveland (1994), Kosslyn (1994), Wallgren (1996), Wilkinson (1999), Kuusela (2000), Salmelin (2003), Robbins (2005), Yau (2009) ja Vehkalahti (2014).

Salmelin R, Vehkalahti K. Visualisation of statistical data
Sosiaalilääketieteellinen aikakauslehti – Journal of Social Medicine 2014:51:301–316

Visualisation of statistical data is based on a few good basic statistical graph types and their variations. The most important of them are bar and column charts, dot plots, histograms, line graphs, scatter plots and box plots. The significance of a good graph is that it transmits various data contents to its target group in a versatile and understandable way but also gives wider audience an opportunity to understand even complex phenomena clearly. There are several clear rules for drawing statistical graphs, the details varying by graph type. The common purpose of the rules is to ensure that the graph does not distort information. The power of a visual message is huge and an incorrectly drawn graph can lead badly astray. This article helps in identifying and choosing the suitable basic statistical graph type and in fathoming the more detailed rules concerning these graphs.

KIRJALLISUUS

Anscombe F. Graphs in statistical analysis. Am Stat 1973:27:17–21.

Chen C, Härdle W, Unwin A. (toim.) Handbook of Data Visualization. Springer, New York 2008.

Cleveland WS. The Elements of Graphing Data. Wadsworth Advanced Books and Software, Monterey 1994.

ESS Round 6: European Social Survey Round 6 Data. Data file edition 2.0. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data, 2012.

Foley JD, Van Dam A. Fundamentals of interactive computer graphics. Addison-Wesley, Reading 1983.

Haukilahti R-L, Virjo I, Mattila K. ETA-alueen ulkopuolella perustutkintonsa suorittaneiden lääkärien Suomeen tulon syyt, työllistyminen ja jatkosuunnitelmat. Sosiaalilääk Aikak 2012:49:13–30.

Huff D. How to Lie with Statistics. Victor Gollanz Ltd, Lontoo 1954. (Suom. Kuinka tilastoilla valehdellaan. Otava, Helsinki 1974.)

Joronen K, Konu A, Rankin S, Åstedt-Kurki P. Draamaohjelman vaikutus oppilaiden sosiaalisiin suhteisiin ja kiusaamiskokemuksiin alakoulussa. Sosiaalilääk Aikak 2013:50:139–49.

Kinnunen U, Perko K, Virtanen M. Esimiehen johtamistyylin yhteys työntekijän kokemaan työuupumukseen ja sairaana työskentelyyn. Sosiaalilääk Aikak 2013:50:59–70.

Kosslyn SM. Elements of graph design. W. H. Freeman and Company, New York 1994.

Korhonen E, Salonen AH, Aho AL, Kaunonen M. Vauvan nukkuminen ja tyytyväisyys vanhemmuuteen äidin näkökulmasta. Sosiaalilääk Aikak 2013:50:192–207.

Korhonen M, Luoma I, Salmelin RK, Helminen M, Kaltiala-Heino R, Tamminen T. The trajectories of child's internalizing and externalizing problems, social competence and adolescent self-reported problems in a Finnish normal population sample. Sch Psychol Int 2014, Online first, Supplement Material. DOI: 10.1177/0143034314525511.

Kuusela V. Tilastografiikan perusteet. Edita, Helsinki 2000.

Kylmä J, Sepponen A-M, Pakarinen M, Heikkinen T, Suominen T. Seksuaalikäyttäytyminen miesten välisissä satunnaisissa suhteissa – tietoa seksuaaliterveyden edistämiseen. Sosiaalilääk Aikak 2014:51:32–44.

Niemi P, Tuomola P, Seppä K. Hiv-positiivisten huumeiden käyttäjien kuolleisuus. Sosiaalilääk Aikak 2013:50:51–8.

Nieminen P, Miettunen J. Suomalainen julkaisuaktiivisuus psykiatrian alan ydinlehdissä vuosina 2001–2010. Sosiaalilääk Aikak 2012:49:317–27.

Robbins NB. Creating More Effective Graphs. Wiley-Interscience, Hoboken 2005.

Salmelin R. Graphical representation of statistical results in medical research. University of Tampere, Tampere 1997.

Salmelin RK. Mistä on hyvät tilastokuvat tehty? Duodecim 2003:119:1761–73.

Spence I. No Humble Pie: The Origins and Usage of a Statistical Chart. J Educ Behav Stat 2005:30:353–68.

Söderlund E, Joronen K. Vanhempi-lapsisuhteen läheisyys ja kouluyhteisön sosiaaliset suhteet. Sosiaalilääk Aikak 2013:50:300–11.

Tufte ER. The Visual Display of Quantitative Information. Graphics Press, Chesire 1983.

Tukey JW. Exploratory Data Analysis. Addison-Wesley, Reading 1997.

Vehkalahti K. Kyselytutkimuksen mittarit ja menetelmät. Finn Lectura, Porvoo 2014.

Wallgren A. Statistikens bilder: att skapa diagram. Publica, Stockholm 1996.

Wilkinson L. The Grammar of Graphics. Springer, New York 1999.

Yau N. Data Points: Visualization That Means Something. Wiley, Indianapolis 2009.

Raili Salmelin
FT
Tampereen yliopisto
Biostatistiikka, Terveystieteiden yksikkö

Kimmo Vehkalahti
Dosentti, VTT
Helsingin yliopisto
Yhteiskuntatilastotiede, Sosiaalitieteiden laitos