3 Mittaamisen taso ja tilaston keskiluvut

Tämä tutkimus on sellainen, että (jos nyt jänisten laskua voidaan mittaamiseksi kutsua) mittaamisessa on eroteltavissa neljä erilaista mittaamisen tasoa, mittausasteikkoa. Näistä tasoista on toisilla niin sanotusti syvyyttä enemmän kuin toisilla. Mittausasteikosta riippuen tulokset kertovat enemmän tai vähemmän tutkittavasta asiasta, mutta toisaalta tutkimuskohdekin saattaa asettaa rajat, millä syvyydellä kiinnostuksen kohdetta voidaan tutkia.

- **LUOKITUS – eli NOMINAALIASTEIKKO**: tilastoyksiköstä voidaan vain todeta, mihin luokkaan se mitattavan ominaisuuden puolesta kuuluu. Itse mitattavaa ominaisuutta ei yleensä voida ilmaista lukuarvona.

Esim. 1 Ihmisen veriryhmä, sukupuoli, äidinkieli, sukan väri, purjealuksen tyyppi.

- **JÄRJESTYS – eli ORDINAALIASTEIKKO**: mittaustulokset voidaan asettaa järjestykseen, mutta ei ilmaista luvulla.

Esim. 2 Maastojuoksukilpailun tulokset, kun käytettävissä ei ole kelloa: voittaja Aili, toinen Jaani, kolmas Jaakko, ...

- **VÄLIMATKA- eli INTERVALLIASTIKKO**: tilastoyksiköihin liittyviä mittaustuloksia voidaan verrata erotuksiensa puolesta, mutta asteikolta puuttuu absoluuttinen nollapiste.

Esim. 3 Edellisen esimerkin, ilmeisesti yhteislähdöllä alkaneen maastojuoksukilpailun tulokset ilmoitetaan, kuinka monta sekuntia osanottaja tuli voittajaa jälkeen maaliin: Aili 0, Jaani 3.4, Jaakko 13.2, ...

Jos maastojuoksukilpailu toistetaan vaikka samoissakin olosuhteissa, kilpailujen tuloksia tai yksittäisen juoksijan henkilökohtaisia juoksuaikoja ei voida verrata toisiinsa. Tietysti voidaan kysyä, miten on mahdollista mitata maaliinsaapumisaikoja alkaen siitä, kun voittaja on urakkansa tehnyt, mutta ei saada kaikkien osanottajien kokonaisjuoksuaikoja.

Tällaista nollapisteetöntä asteikkoa käytetään kuitenkin lähes jokaista aika ajoin kiinnostavassa lämpötilan mittauksessa. Vastaväite ”Onhan Celsius-asteikolla nollapiste” voidaan heti kumota toteamalla, että se

ei ole niin sanotusti absoluuttinen. Kerran oli eräässä sanomalehdessä talvisen, pitkäkestoisen pakkassäällä sattuneen sähkökatkoksen seurauksia kuvaileva uutinen, jossa sanottiin asunnon lämpötilan laskeneen 11 Celsiusasteseen, kun asukkaat tavallisesti oleskelivat 22:ssa. Toimittaja kirjoitti: ”Kyllä minäkin huolestuisin, jos asuntoni lämpötila laskisi puoleen normaalista.” Mitähän kyseinen toimittaja olisi kirjoittanut, jos asunnon lämpötila olisi laskenut tasan nollaan (Celsiusasteeseen) ? Lisäksi on fysiikasta kiinnostuneille syytä huomauttaa, että ideaalikaasun tilayhtälöä sovellettaessa on erittäin kohtalokasta syöttää lämpötila celsiusasteina. Mitä lähempänä lämpötila on jään sulamispistettä (normaalipaineessa), sitä hullumpia tuloksia saadaan.

- **SUHDEASTEIKKO**: Mitta-asteikolla on absoluuttinen nollapiste. Tämä on täydellisin mittaamisen taso. Esimerkiksi kelpaa maastojuoksukilpailun loppu-tulos täydellisin juoksuajoin.

Monet opetusharjoittelun käyneet kritisoivat ankarasti kasvatustieteen opintoja hyödyttömiksi, mutta eivät ne ole kaikilta osin merkityksettömiä. Eräässä kirjassa nimittäin kritisoitiin, kuinka ihmiset arjessa käyttävät sanontoja, joiden merkitys ei yllä edes alimman tilastotieteellisen mittaamisen tasolle.

Esim. 4 a) ”Mieheni viettää kanssani aikaa liian vähän. Netissä vaan pyörii.”
b) ”Teidän Tiinalla on liian tiukat farkut”
c) ”Suomessa on avioeroja aivan liikaa.”

Jos Sinulla on mahdollisuus osallistua tällaiseen keskusteluun, esitä varovainen (?) vastakysymys:

a) Montako tuntia (viikossa) vaatisit henkilökohtaista aikaa?
b) Mikähän tulisi Tiinan farkkujen numero olla? Lökäpöksyt vai?
c) Kuinka monta avioeroa vuodessa olisi sopiva määrä Suomen oloissa?

Voin vakuuttaa, että keskustelu terävöityy melko paljon vastakysy-myksesi jälkeen. Voit ääritapauksessa varautua jopa pahoinpitelyyn ellet pääse karkuun. Kun sanotaan, että avioeroja on aivan liikaa, niin eihän ilmaisu täytä edes luokitusasteikon tasovaatimusta. Kun ihmisestä määritetään hänen veriryhmänsä, tulos on yksinkäsitteisen selvä ja merkittävissä tilastoon. Joku terveysministeri voi asian-tuntijalausuntoihin tukeutuen asettaa, tai lähinnä suositella alkoholin

käytön rajat, mutta loppujen lopuksi kuka määrittelee yksikäsitteiset rajat vuosittaisille avioeroille asteikolla i) liian vähän ii) sopivasti ja iii) liian paljon.

Millä mittaamisen tasolla Puttosen jänislaskenta oli suoritettu?

Esim. 5 Osa jotakin laajempaa terveystutkimusta saattaisi olla kysely, jossa tiedustellaan kohdehenkilöiden viikoittain aktiivisen liikuntaan (tai matematiikan kotitehtävien suorittamiseen) käyttämää aikaa. Tällöin voitaisiin tutkia, onko liikunnan määrä (tai sen puute) yhteydessä jonkin sairauden esiintymiseen kohdejoukossa (tai onko kotitehtäviin käytetyllä ajalla kuinkakin tiukka yhteys matematiikassa menestymiseen). Alla oleva taulukko 2 on laadittu 26 henkilön viikoittaisesta aktiiviliikunnasta, ajat ovat tunteja, pyöristetyt 30 minuutin tarkkuuteen.

Taulukko 2. Viikoittainen liikunta

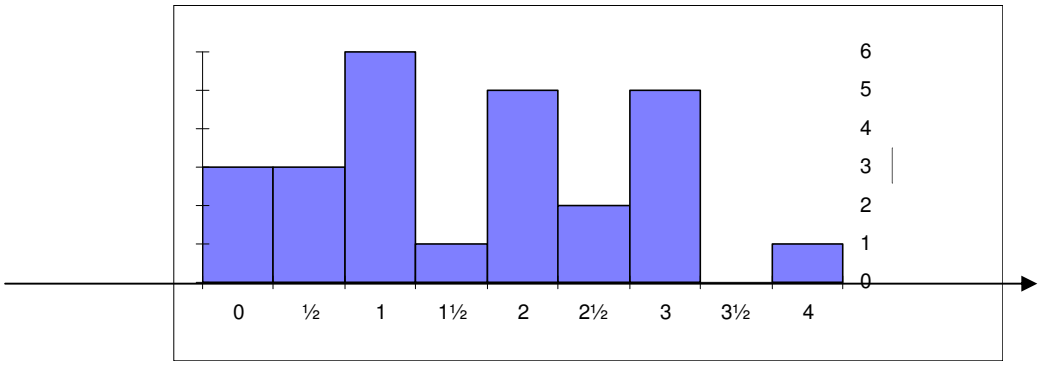
1½	3	1	0	0
3	2	1	1	1
½	2½	2½	3	4
2	2	3	2	1
½	0	½	2	3
1				

Luokitellaan ja esitetään graafisesti:

LUOKKA	tukkimiehen kirjanpito	f
0	|||	3
½	|||	3
1	|||||	6
1½	|	1
2	|||||	5
2½	||	2
3	|||||	5
3½		0

4	|	1

Histogrammi on pylväsdiagrammi, jossa pylväät ovat kiinni toisissaan. Huomaa pylväiden sijainti vaaka-akseliin nähden. Liikuntaan käytetty aika on esimerkiksi yksi tunti sellaisilla henkilöillä, jotka ulkoilevat enemmän kuin 45 minuuttia mutta alle 1 h 15 minuuttia.



Kuva 2. Histogrammi, liikuntaharrastuksen jakauma

Tilastoa luonnehditaan eritasoisilla tunnusluvuilla, jotka jaetaan tavallisesti ns. keskilukuihin ja hajontalukuihin. Tilaston laadinnassa käytetty mittaamisen taso määrää, minkätasoisia tunnuslukuja voidaan määrittää.

Tilaston keskilukuja on kolme

- **Aritmeettinen keskiarvo** saadaan jakamalla havaintoarvojen summa niiden lukumäärällä

- **Mediaani** (M) on havaintoarvoista keskimmäinen, kun havaintoarvot on asetettu suuruusjärjestykseen. Jos havaintoarvoja on parillinen määrä, mediaani on kahden keskimmäisen keskiarvo (milloin keskiarvo voidaan

laskea) taikka sitten joidenkin teoriakirjojen mukaan kumpi tahansa keskimmäisistä arvoista.

- **Tyyppiarvo eli moodi** (Mo) on se havaintoarvo, jonka frekvenssi on suurin eli jota tilastossa esiintyy eniten. Luokitellusta aineistosta tyyppiarvoa ilmoitettaessa voi esittää frekvenssiltään suurimman luokan luokkakeskuksen tai koko luokan luokkarajoin ilmaistuna.

Esim. 6 Edellisen esimerkin tilaston tyyppiarvo Mo = 1 h, koska sen frekvenssi on suurin.

Koska havaintoarvoja (tilastoyksiköitä) on 26 kpl, niistä keskimmäisiä on kaksi, 13. ja 14. sen jälkeen, kun lukuajat ovat suuruusjärjestyksessä. Luokissa 0, ½ ja 1 h on yhteensä 12 havaintoarvoa, joten 13. havaintoarvo on 1½ h ja 14. on 2 h.
Tilaston **Md** on nyt joko 1½ tai 2 tunnin lukuaika tai näiden keskiarvo 1h 45 min.

Aritmeettisen **keskiarvon** osannee jokainen laskea.

Mikäli tilaston mittaaminen on suoritettu luokitusasteikon tasolla, sille voi keskiluvuista ilmoittaa ainoastaan tyyppiarvon. Järjestysasteikon tasolla mitatulle aineistolle pystyy ilmoittamaan sekä tyyppiarvon että mediaanin. Välimatka- ja suhdeasteikon tasolla mitatulle aineistolle voidaan ilmoittaa kaikki kolme keskilukua.

Tilaston keruussa jonkin seikan mitatut lukuarvot saattavat olla mitatut niin tarkasti, ettei ole mielekästä määrittää jokaisen mittaustuloksen frekvenssiä, ja tällöin käytetään ns. aineiston luokittelua. Havaintoaineisto koko laajuudeltaan jaetaan yleensä tasalevyisiin luokkiin, jolloin luokitellussa frekvenssitaulukossa yksittäisen tiedon tarkka informaatio katoaa, ja sen, esimerkiksi keskiarvoa laskettaessa, katsotaan sijaitsevan luokan keskuksessa. Kurssin loppupuolella vastaantulevan todennäköisyysjakauman kertymäfunktiota ajatellen on tässä yhteydessä syytä tutustua myös ns. summafrekvensseihin.

Esim. 7 Erään kokeen pistemäärät jakaantuivat seuraavalla sivulla olevan taulukon 3 mukaisesti. Aineisto on siinä jo luokiteltu. Sikäli kun koepisteet ovat olleet kokonaislukuja, esimerkiksi luokkaan $40 < x \leq 45$ kuuluvat 13 arvoa saattavat olla vaikka jokainen mitä tahansa

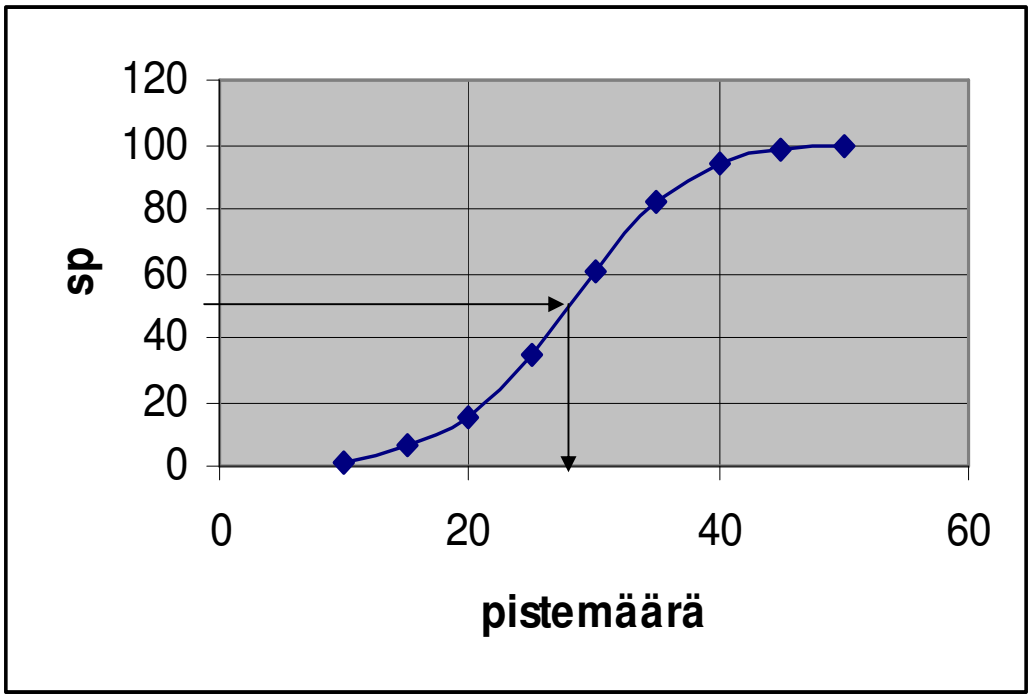
luvuista 41, 42, 43, 44 tai 45, mutta ajatellaan niiden sijaitsevan luokkakeskuksessa, joka on 43 (onko tarkkaan ottaen).

Aiemmin esillä olleissa taulukoissa on ollut vain frekvenssisarake f. Taulukossa 3 näet kolme muutakin saraketta otsikoin sf (summa-frekvenssi), p (prosentuaalinen osuus) ja sp (prosenttien summa).

Taulukko 3. Koearvosanat

x	f	sf	p	sp
$0 \leq x \leq 10$	2	2	0,63	0,63
$10 < x \leq 15$	18	20	5,70	6,33
$15 < x \leq 20$	29	49	9,18	15,51
$20 < x \leq 25$	59	108	18,67	34,18
$25 < x \leq 30$	84	192	26,58	60,76
$30 < x \leq 35$	67	259	21,20	81,96
$35 < x \leq 40$	39	298	12,34	94,30
$40 < x \leq 45$	13	311	4,11	98,42
$45 < x \leq 50$	5	316	1,58	100,00
	316			

Sarakkeessa sf oleva luku, esimerkiksi 259 kertoo sen, että pistemäärän 35 tai tätä pienemmän sai täsmälleen 259 opiskelijaa, ja samalla rivillä sp – sarakkeessa oleva luku 81.96 antaa tiedon, että korkeintaan tämän pistemäärän sai hivenen vajaa 82 prosenttia opiskelijoista. Piirretään suhteellisten frekvenssien summakäyrä:



Kuva 3. Suhteellisten frekvenssien summakäyrä. Mediaanin määritys.

Jos tilaston mittaaminen on suoritettu vähintään välimatka-asteikon tasolla, niin suhteellisten frekvenssien summakäyrältä löydetään arvoa sp = 50% vastaava havainto-arvo (siis mediaani) yllä olevan käytännön mukaisesti: pystyakselin 50% pisteestä summakäyrälle ammutun nuolen ja summakäyrän leikkauspiste projisoidaan vaaka-akselille. Arvio jää tässä kohden hieman karkeaksi, mutta Md lienee 26 – 27 pisteen tuntumassa. Tämä on tietystikin arvio ja olettaa pistemäärien jakaantuneen kuhunkin luokkaan suunnilleen tasaisesti.