Komulainen, Erkki & Karma, Kai

Tilastollisen kuvauksen perusteet käyttäytymistieteissä

Toinen laitos
(Versio 2.2, 1.1.2002)

Helsingin yliopisto

Kasvatustieteen laitos

Sisällys

ISBN 952-10-0288-3 (Word)
ISBN 952-10-0289-1 (pdf)

Käyttäjälle

Ensimmäisen painoksen alkusanoja (jotka ovat elokuulta 1979) ei juuri tarvitse muuttaa. Yhä on syytä korostaa, että havaintoaineiston analyysi ei ole matematiikkaa eikä tilastotiedettä. Kaavojen maailmaa tärkeämpää on ymmärtää havaintoaineiston ja saadun tulokset yhteys. Laskutoimitusten suorittaminen sujui jo tuolloin varsin vaivatta. Tietokoneympäristö on laajentunut tuosta ajasta jokaisen käyttäjän pöydälle. Yhä on keskeistä ymmärtää, mitä saatu tulos merkitsee. Konkreettinen (taskulaskurilla suoritettu) harjoitustehtävien tekeminen on ymmärtämisen kannalta välttämätöntä.

Tilastollisiin tarkasteluihin sisältyy myös aina sovellustilanteen mukanaan tuoma konteksti. Sen vuoksi keittokirjamaisia reseptiohjeita sovelluksille on mahdotonta laatia. Tätä pyritään korostamaan ilmauksilla "on ehkä hyvä", "sopii tässä tapauksessa" jne.

Tekstin korjailu kirjan toiseen laitokseen on suoritettu kesän 2001 aikana. Kiitämme kustantajaa luvasta käyttää aikaisempaa tekstipohjaa sellaisenaan. Varsinaista sähköistä oppimateriaalia ei ole pyritty synnyttämään, verkkoa hyödynnetään vain jakelukanavana. Teksti perustuu siis Kai Karman vuonna 1980 ilmestyneeseen samannimiseen kirjaan. Uuden laitoksen muokkaustyön on tehnyt Erkki Komulainen.

Otamme mielellämme palautetta ja suoritamme sen edellyttämiä korjauksia tarpeen mukaan.

Materiaalia saa käyttää vapaasti ei-kaupallisessa yliopistojen ja avoimen yliopiston opetuksessa.

Syyskuussa 2001

Erkki Komulainen (Erkki.Komulainen@Helsinki.Fi)
Kai Karma (Kai.Karma@Siba.Fi)

Versioon 2.2, 1.1.2002 on tehty pieniä korjauksia

1.Kuvauksen lähtöaineisto

Tieteen tehtävänä on uuden tiedon hankkiminen. Käyttäytymistieteet tutkivat elollisten olioiden käyttäytymistä voidakseen ymmärtää sitä tai ainakin löytääkseen siitä säännönmukaisuuksia; tällöin sitä voidaan ennustaa ja mahdollisesti vaikuttaa siihen. Tutkittavien käyttäytymisestä hankitut kuvaukset voivat olla varsin monenlaisia. Näistä voidaan kuitenkin erottaa kaksi tyyppiä: sanalliset kuvaukset siitä, mitä tutkittavat ovat tehneet jossakin tietyssä tilanteessa, sekä luvuin tai helposti luvuilla korvattavissa olevin symbolein ilmaistut tiedot.

Edellisistä voidaan ottaa esimerkeiksi vaikkapa seuraavat: "oppilaat vaikuttivat aktiivisemmilta ja halukkaammilta osallistumaan silloin, kun he olivat saaneet vaikuttaa opetuksen suunnitteluun", "lapsi- ja nuorisoryhmissä on tavallista, että niihin muodostuu "syntipukki", ryhmän jäsen, jonka asema on alhainen ja johon muut helposti purkavat aggressioitaan". Numeerisia, luvuin ilmaistuja kuvauksia olisivat esimerkiksi seuraavat: "Maija sai älykkyystestissä 125 pistettä", "koeeläimet oppivat juoksemaan sokkelon läpi keskimäärin kymmenennellä kerralla".

Molemmilla kuvaustavoilla on sekä hyvät että huonot puolensa. Numeeriset kuvaukset ovat yleensä tarkempia ja objektiivisempia, ts. ne eivät yleensä ole kovin riippuvaisia siitä, kuka ne on hankkinut, eivätkä eri henkilöt yleensä tulkitse niitä kovin eri tavoin. Toisaalta on asioita, joiden kuvaaminen numeroin tai muin yksinkertaisin symbolein on vaikeaa tai mahdotonta tai. tuottaa tulokseksi näennäis-tarkkaa tietoa, josta puuttuvat oleelliset ja mielenkiintoiset näkökohdat. On jokseenkin toivotonta yrittää numeroin kuvata esim. "mitä Pekka ajatteli tehdessään matematiikan tehtäviä " tai "miksi riitainen tai välinpitämätön koti saattaa aiheuttaa sopeutumattomuutta koulun järjestykseen ".Tällaisessa tapauksessa on esim. haastattelu tai tutkijan yritys ymmärtää tilannetta osallistumalla siihen arvokkaampi kuin pinnalliset numerotiedot asiaan mahdollisesti vaikuttavista seikoista.

Valinta näiden kuvaustapojen välillä on tutkijan ensimmäisiä tehtäviä sen jälkeen, kun tutkimuksen ongelma on selvitetty ja rajattu. Valinta on suoritettava sen mukaan, kummalla tavalla uskotaan saatavan arvokkaampaa ja luotettavam-

paa tietoa, eikä esim. sen pohjalta, että käytössä on tietokone tai että haastatteleminen on helpompaa kuin laskeminen tms. Usein päästään parhaaseen tulokseen yhdistämällä verbaalinen ja numeerinen kuvaustapa samaan tutkimukseen. Koska tässä yhteydessä on tarkoitus käsitellä tilastollisia menetelmiä, joiden lähtökohtana ovat luvut, siirrymme käsittelemään nimenomaan numeerista kuvausta; lukijan ei pidä kuitenkaan unohtaa, että se on vain osakäytettävissä olevista mahdollisuuksista.

Nimitämme tässä mittaamiseksi kaikkia sellaisia toimenpiteitä, joilla tutkittavien ominaisuuksia kuvataan numeroin. Mittaamista voisivat siis olla esim. testaaminen, tiettyjen käyttäytymisen piirteiden määrän laskeminen, arviointi jollakin asteikolla, esim. kouluarvosanoina jne. Mittariksi sanomme sitä välinettä, jolla tieto hankitaan, esim. testiä, asenneasteikkoa, kyselylomaketta jne. Tällainen käsitys mittaamisesta on laajempi kuin arkikielen vastaava käsite, mutta se on tarkoituksenmukainen käyttäytymistieteissä.

Olettakaamme, että olemme kiinnostuneita peruskoulun neljäsluokkalaisten suorituksista erilaisissa älykkyystehtävissä, erityyppisten tehtävien suhtautumisesta toisiinsa sekä sukupuolen vaikutuksesta suoritustasoon eri tehtävissä. Tätä varten meidän täytyy kerätä joukolta peruskoulun neljäsluokkalaisia heidän suorituksensa ko. tehtävissä. Tämä joukko, jota tutkimme, on otos. Se joukko, josta olemme kiinnostuneet, tässä tapauksessa peruskoulun neljäsluokkalaiset, on populaatio. Jotta voisimme tehdä päätelmiä populaatiosta eikä vain siitä otoksesta, joka on tutkittavana, otoksen täytyy olla edustava. Toisin sanoen sen täytyy olla oleellisissa suhteissa riittävän samanlainen kuin populaatio. Samanlaisuus pyritään takaamaan otantamenetelmillä, joita emme kuitenkaan lähde tässä tarkemmin kuvailemaan.

Oheiseen taulukkoon on kerätty viidenkymmenen oppilaan tulokset yhdeksässä erityyppisessä älykkyystestissä. Tällaista taulukkoa nimitetään matriisiksi. Koska luvut ovat käsittelemättömiä, sellaisina kuin ne on alunperin hankittu, sanotaan taulukkoa raakapiste-, havainto- tai primäärimatriisiksi. Ominaisuuksia, joita on mitattu ja jotka siis saavat erilaisia arvoja eri koehenkilöiden kohdalla, nimitetään muuttujiksi eli variaabeleiksi. Kullakin matriisin rivillä on siis yhden koehenkilön saamat arvot kymmenellä eri variaabelilla (sukupuoli + yhdeksän testiä). Niinpä esim. ensimmäinen koehenkilö on ollut tyttö, joka on saanut yhteenlaskutehtävissä 22 pistettä, havaintonopeustehtävissä 29 pistettä, vas-

takohtatehtävissä 9 pistettä jne. Kahdeskymmenes testattu on ollut poika, joka on saanut yhteenlaskutehtävissä 21 pistettä, havainto- nopeustehtävissä 22 pistettä jne. Matriisin sarakkeet taas edustavat muuttujia "yli koehenkilöiden", yksi sarake sisältää yhden muuttujan arvot kullekin koehenkilölle. Niinpä esim. kuudennella muuttujalla (lauseentäydennystehtävät) on ensimmäinen koehenkilö saanut 17 pistettä, toinen 10, kolmas 11, neljäs 16 jne.

Matriisin muuttujat:

1 sukupuoli
2 yhteenlaskut
3 havaintonopeus
4 vastakohdat
5 neliötäydennys
6 lauseentäydennys
7 looginen järjestys
8 sanaryhmät
9 matemaattiset tehtävät
10 peilitesti

kh	1	2	3	4	5	6	7	8	9	10
1	0	22	29	09	11	17	20	26	10	12
2	0	19	24	21	03	10	11	08	03	15
3	0	19	24	22	03	11	15	14	06	17
4	0	29	32	33	08	16	20	24	09	16
5	0	25	22	30	09	12	14	22	07	26
6	0	15	27	18	13	09	10	17	03	20
7	0	13	28	35	09	15	15	16	09	18
8	0	11	25	21	05	13	02	16	05	07
9	0	08	23	01	10	12	01	08	05	00
10	0	01	22	16	06	16	14	19	08	17
11	0	12	28	25	08	15	09	22	07	18
12	0	12	34	21	14	19	14	19	03	11
13	0	19	12	21	01	09	06	14	05	04
14	0	14	24	20	02	12	15	18	07	11
15	0	13	21	22	11	14	10	21	05	13
16	0	15	34	29	05	11	11	17	07	08
17	0	08	23	13	06	10	12	13	06	06
18	1	28	30	27	13	13	13	18	10	16
19	1	15	20	06	11	06	05	05	08	16
20	1	21	22	13	05	06	03	07	09	04
21	1	30	16	11	05	10	10	15	12	17
22	1	54	30	35	15	17	17	18	13	16
23	1	18	20	20	12	08	06	10	08	14
24	1	04	26	15	11	12	06	14	08	11
25	1	14	24	10	13	10	13	17	08	20
26	1	15	23	19	07	13	12	12	11	07
27	1	28	32	34	15	14	14	15	22	09
28	1	34	16	24	06	17	13	22	06	07
29	1	12	34	20	14	08	09	12	07	09
30	1	28	24	14	03	06	04	05	04	04
31	1	22	22	28	13	13	15	20	08	08
32	0	08	20	21	05	08	10	12	04	10
33	0	13	21	24	08	09	10	09	07	06
34	0	11	24	24	08	07	14	16	03	21
35	0	24	27	36	13	12	15	24	05	24
36	0	21	32	36	11	14	13	16	08	11
37	0	18	32	41	12	13	20	18	11	19
38	0	02	22	28	01	11	08	11	03	02
39	0	16	24	20	07	04	09	12	02	16
40	0	01	23	16	06	05	01	06	05	06
41	0	19	12	40	08	15	12	19	07	04
42	0	17	28	32	07	07	14	26	05	07
43	1	16	24	35	11	16	17	19	12	09
44	1	07	23	26	12	12	13	10	06	15
45	1	14	18	21	06	08	04	07	03	11
46	1	00	15	21	05	10	07	14	06	12
47	1	19	12	17	08	07	03	12	11	07
48	1	07	25	23	06	14	08	14	04	06
49	1	23	27	29	12	11	18	17	08	17
50	1	18	24	10	08	03	03	11	05	10

Raakapistematriisissa olevaa tietoa ei voida lisätä enää millään menetelmällä, se voidaan ainoastaan saattaa helpommin ymmärrettävään muotoon. Paitsi eri tes-

tien tuloksia sellaisenaan, matriisiin sisältyvät myös testien väliset suhteet, sukupuolen yhteys erityyppisissä tehtävissä menestymiseen jne. Periaatteessa, jos ihminen kykenisi pitämään mielessään suuren joukon lukuja ja niiden välisiä suhteita, ei raakapistematriisia tarvitsisi käsitellä lainkaan, vaan kaikki siinä oleva tieto, informaatio, olisi saatu, kun primääripisteet olisi luettu. Näin ei kuitenkaan ole, vaan suuressa lukujoukossa vallitsevia säännönmukaisuuksia ei normaali ihminen kykene lainkaan riittävästi havaitsemaan ja pitämään mielessään. Joitakin seikkoja voi kuitenkin oheisesta matriisista havaita pelkästään selailemallakin. On esimerkiksi suhteellisen helppo todeta, että neliötäydennystehtävissä ovat pistemäärät useammin pienempiä kuin havaintonopeustehtävissä. Paljonvaikeampaa on jo sen toteaminen, onko neliötäydennystehtävissä ja matemaattisissa tehtävissä tällaista eroa. Vielä hankalampaa on esim. todeta onko näissä testeissä eroa eri sukupuolten menestymisessä tai onko vaikkapa matemaattisissa tehtävissä ja havaintonopeustehtävissä menestymisen välillä yhteyttä. Jos kuvitellaan, että matriisissa olisikin viidenkymmenen koehenkilön sijasta tiedot viidestäsadasta ja kymmenen muuttujan tilalla viisikymmentä, havaitaan, että mielekkään ja oikean tiedon löytäminen pelkästään alkuperäisiä testipisteitä tarkastelemalla olisi ylivoimaista. Tarvitaan menetelmiä, joiden avulla primääriaineistossa oleva tieto voitaisiin saada sellaiseen muotoon, että sen voi ymmärtää ja hallita. Aineistossa vallitsevat säännönmukaisuudet tulisi voida ilmaista joillakin suhteellisen harvoilla symboleilla, jotka keskittyisivät oleellisiin seikkoihin. Juuri tähän pyritään tilastollisella kuvaamisella, jonka usein sanotaankin "tiivistävän" aineistossa olevaa informaatiota.

Tilastollisessa kuvauksessa ei yleensä käsitellä varmoja, yksiselitteisiä seikkoja, vaan pikemminkin joidenkin ilmiöiden taipumusta olla jonkin suuntaisia, niiden todennäköisyyksiä sattua tietyissä tilanteissa, niiden keskimääräistä esiintymistä suurissa joukoissa tai pitkinä ajanjaksoina jne. Useimmat käyttäytymistieteiden käsittelemät seikat ovat juuri tällaisia. Yksilön kohdalla voi olla hyvin vaikea sanoa mitään erityisen varmaa, mutta suuremmissa joukoissa saattavat tietyt säännönmukaisuudet tulla esiin hyvinkin selkeästi (kun verrataan sitä sattumaan). Niinpä on luonnollista, että käyttäytymistieteissä käytetään lukujen käsittelyssä juuri tilastomatematiikkaa.

2. Yksiulotteiset jakaumat

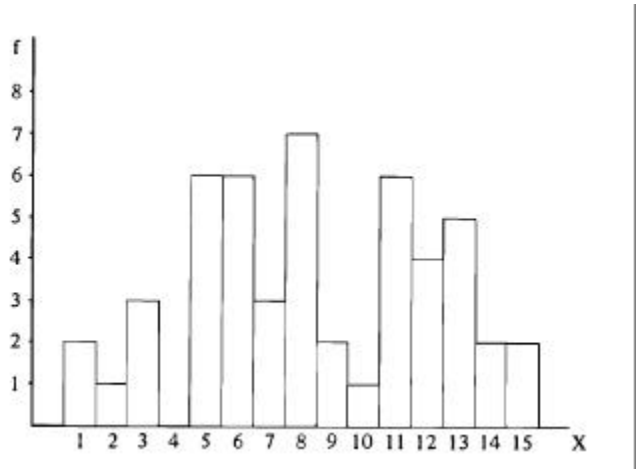
Miten sitten voisimme lähteä "tiivistämään" esim. aiemmin esitetyn matriisin sisältämää informaatiota? Eräs yksinkertainen tapa on lähteä siitä, että variaabeleilla on arvoja, jotka useampi kuin yksi koehenkilö on saanut. Voimme tehdä näistä huomattavasti primääritietoa kätevämmän taulukon luettelemalla saadut arvot vain kerran ja merkitsemällä kunkin kohdalle, kuinka moni henkilö on kyseisen arvon saanut. Testissä saatuja pistemääriä (raakapisteitä, primääripisteitä) merkitään isolla X:llä ja niiden henkilöiden määrää, jotka ovat kunkin pistemäärän saaneet, pienellä f:llä (frekvenssi, engl. frequency). Jos vielä laskemme kunkin frekvenssiluvun prosenttiosuuden koehenkilöiden koko määrästä, saamme melko hyvän kuvan siitä, miten pistemäärät jakaantuvat koehenkilöiden kesken. Esimerkiksi neliötäydennystehtävien tuloksista saamme seuraavan taulukon :

X	f	%
1	2	4
2	1	2
3	3	6
4	0	0
5	6	12
6	6	12
7	3	6
8	7	14
9	2	4
10	1	2
11	6	12
12	4	8
13	5	10
14	2	4
15	2	4
N = 50 = Σf		100

Yhden pisteen on siis neliötäydennystestissä saanut kaksi koehenkilöä, kaksi pistettä on saanut yksi, kolme pistettä kolme jne. Nämä ovat vastaavasti 4, 2 ja 6 prosenttia testattujen koko määrästä. Alimman ja ylimmän saadun pistemäärän väliltä merkitsemme kaikki arvot näkyviin; niinpä X-arvo 4 on mukana, vaikka sitä ei ole kukaan saanut, ts. sen frekvenssi on nolla. 8 pistettä on tavallisin tulos tässä testissä, sen on saanut 7 koehenkilöä eli 14 prosenttia. Seuraavina ovat pistemäärät 5, 6 ja 11; kunkin on saanut 6 koehenkilöä.

Taulukon alla f:n edessä oleva merkki on kreikkalainen kirjain, iso sigma, joka tarkoittaa summaa. Merkintä "sigma f" luetaan "frekvenssien summa" ja sitä tässä edustava luku, 50, saadaan siis laskemalla f-sarakkeen kaikki luvut yhteen. Symboli N tulee sanasta numerus ja tarkoittaa kaikkien mitattujen yksiköiden (jotka tässä ovat ihmisiä ja siis yksilöitä tai henkilöitä) yhteismäärää. Koska kaikki kunkin pistemäärän saaneet henkilöt ovat f-sarakkeessa mukana, on heidän yhteenlasketun määränsä oltava sama kuin koko koehenkilöjoukon eli siis 50. Prosenttisarakkeen summa on luonnollisestikin 100.

Usein havainnollisemmin ja nopeammin käsitettävästi voidaan sama tulos esittää graafisesti, piirroksena. Jos on verrattava useaa jakaumaa, on vertailu helpointa juuri graafisen esityksen perusteella. Tällaisessa kuvauksessa sijoitamme pisteet suorakulmaiseen koordinaatistoon, jonka pystyakseli edustaa frekvenssejä ja vaaka-akseli saatuja mittaustuloksia, pistemääriä. Äskeinen taulukko näyttää graafisena esityksenä seuraavalta:



Tällaista kuvaustapaa sanotaan pylväsdiagrammiksi tai histogrammiksi. Se muodostuu kutakin havaintoluokkaa X-akselilla edustavan matkan levyisistä pylväistä, joiden korkeus edustaa tähän luokkaan kuuluvien tapausten lukumäärää. Havaintoluokat ovat tässä tapauksessa yhden pisteen, siis kokonaisen numeron, mittaisia. Luokkaväli on yksi yksikkö.

Kuviosta on helppo nähdä, että nelosten kohdalla on tyhjä paikka, ts. nelosten frekvenssi on nolla, kukaan ei ole saanut tätä pistemäärää. Nopeakin vilkaisu

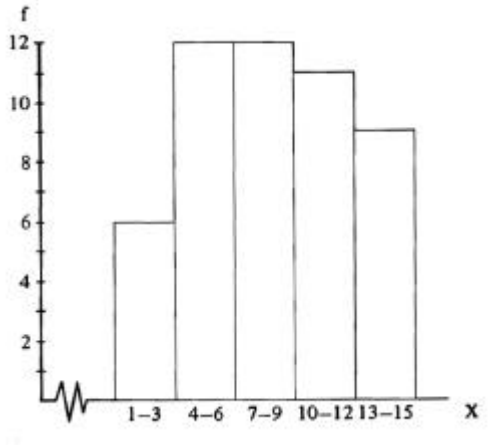
osoittaa myös esim. sen, että yhdeksikköjä ja kymppejä on hyvin vähän. Tavallisin arvo (8) on myös helppo paikallistaa koska sen kohdalla on kuvaaja korkein.

Kuvion esittämä jakauma on suhteellisen epätasainen, siinä on epäsäännöllisin välein korkeampia ja matalampia kohtia. Joskus tällainen epätasaisuus on todellista ja mielenkiintoista tietoa, mutta usein - ja ilmeisesti myös tässä tapauksessa - on kysymys vain siitä, että koehenkilöitä on ollut suhteellisen vähän verrattuna saataviin pistemääriin. Jos numerus olisi ollut suurempi, olisivat epätasaisuudet todennäköisesti huomattavasti tasoittuneet. Tällaisessa tapauksessa voidaan kuvausta tarkoituksellisesti tehdä karkeammaksi yhdistämällä viereisiä luokkia keskenään.

Siihen, kuinka monta alkuperäistä pistettä muodostavat uuden luokan, ei ole mitään ehdotonta sääntöä. Riittää, kun suorittaa yhdistämisen, luokittelun, siten että alkuperäisen jakauman turha, sattumanvarainen vaihtelu häviää, mutta jakauman oleelliset piirteet jäävät jäljelle. Nyt käsiteltävinä olevat neliötäydennystestin pistemäärät voitaisiin luokitella vaikkapa niin, että aina kolme vierekkäistä alkuperäistä pistearvoa muodostaa yhden luokan. Tällöin siis ensimmäiseen luokkaan tulevat ykköset, kakkoset ja kolmoset, joiden määrä on yhteensä kuusi, toiseen luokkaan neloset, viitoset ja kuutoset, joiden yhteismäärä, luokkafrekvenssi, on kaksitoista jne. Seuraavassa on aikaisempi taulukko täydennettynä luokkien rajoja osoittavilla viivoilla sekä luokkien frekvensseillä:

X	f	lf (luokkafrekvenssi)
1	2	
2	1	6
3	3	
4	0	
5	6	12
6	6	
7	3	
8	7	12
9	2	
10	1	
11	6	11
12	4	
13	5	
14	2	9
15	2	
N = 50 = Σf		50 = N = Σlf

Luokiteltu aineisto voidaan luonnollisesti esittää myös graafisesti. X-akselille voidaan merkitä joko kuhunkin luokkaan kuuluvat pistemäärät (siis 1-3, 4-6 jne.) tai luokkien rajakohdat (0.5, 3.5, 6.5 jne.). Tässä tapauksessa, kun testissä on saatu vain kokonaisia pisteitä, on ehkä selkeintä merkitä ne näkyviin kokonaislukuina eikä absoluuttisina luokkarajoina. Seuraavassa kuviossa on luokiteltu aineisto esitetty histogrammina:



Tarkasteltaessa nyt käytettävässä raakapistematriisissa olevia muuttujia voidaan helposti todeta, että ensimmäinen muuttuja, sukupuoli, on erikoisasemassa muihin nähden. Mm. voitaisiin ykköset ja nollat yhtä hyvin vaihtaa keskenään siten, että tyttöjä merkittäisiinkin ykkösellä ja poikia nollalla. Kun tiedetään, mitä näiden lukujen on sovittu merkitsevän, voidaan jokaisen koehenkilön kohdalla luotettavasti tietää, onko ko. henkilö tyttö vai poika. Samoin voitaisiin symboleina käyttää muitakin kuin numeroita, esim. kirjaimia T ja P. Näitä muutoksia ei voitaisi haitatta tehdä muilla muuttujilla. Kunkin testin pistemäärä kullakin henkilöllä muodostuu oikein ratkaistujen tehtävien lukumäärästä, joten se ei ole sopimuksenvaraista. Variaabelia, jota edustavat luvut symboloivat määrää (kvantiteettia), sanotaan kvantitatiiviseksi. Aina kun voimme ilmaista jotakin olevan enemmän tai vähemmän, paljon tai vähän, olemme tekemisissä kvantitatiivisen variaabelin kanssa. Tällaisia ovat esim. älykkyys (jollakulla voi olla "enemmän älykkyyttä" kuin toisella), sijoitus kilpailuissa (esim. juoksun suorittaminen vähemmässä ajassa), ekonominen asema (paljon tuloja tai omaisuutta) jne.

Jos symboleja käytetään sopimuksen mukaan erottamaan eri yksilöitä tai ryhmiä toisistaan, eikä kyseessä ole minkään ominaisuuden määrä vaan laatu (kvali-

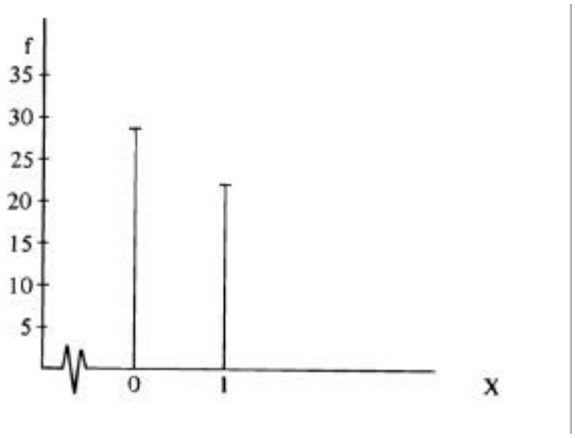
teetti), on kyseessä kvalitatiivinen variaabeli. Tällainen muuttuja on esim. edellä mainittu sukupuoli, jota on kahta laatua, joilla kummallakin on oma sovittu symbolinsa. Koska kysymys ei ole määrästä, ei ole myöskään välttämätöntä käyttää symboleina juuri numeroita, vaan periaatteessa mitkä tahansa toisistaan erotettavat merkit, esim. kirjaimet, kelpaavat yhtä hyvin. Usein on kuitenkin tarkoituksenmukaista, esim. aineiston tietokonekäsittelyn takia, käyttää numeroita. Kvalitatiivisia muuttujia ovat esim. pankkitilien numerot (joilla eri tilit erotetaan toisistaan), värit, koehenkilöiden kotipaikkakunta, ammatti jne. Samoin voidaan kvalitatiivisena muuttujana pitää vaikkapa aiemmin esitetyn havaintomatriisin koehenkilöiden numeroita. Niillähän pyritään juuri erottamaan yksilöitä toisistaan eikä ilmaisemaan minkään ominaisuuden määrää. Niiden sijasta voisi yhtä hyvin olla Leena, Maija, Virpi jne.

Sukupuoli eroaa primäärimatriisin muista muuttujista muussakin suhteessa kuin siinä, että se on kvalitatiivinen muiden ollessa kvantitatiivisia. Se voi saada vain kaksi arvoa eikä mitään niiden väliltä; onhan näitä symboleja käytettäessä järjetöntä sanoa esim. , jonkun koehenkilön sukupuoli on 0.65. Sen sijaan voidaan hyvin kuvitella esim., että joku saisi yhteenlaskutestissä 17.25 pistettä, jos testissä annettaisiin muitakin kuin kokonaisia pisteitä.

Muuttujaa, joka voi tietyllä alueella saada vain rajallisen määrän arvoja eikä mitään niiden väliltä, sanotaan epäjatkuvaksi variaabeliksi. Tällaisia ovat kaikki kvalitatiiviset muuttujat (ammatti, kotipaikkakunta jne.) kuten edellä mainittiin sekä esim. järjestysluvut (ensimmäinen, toinen jne.) ja vaikkapa jossakin koetilanteessa olevien henkilöiden määrä, joka vaihtelee vain kokonaislukuina.

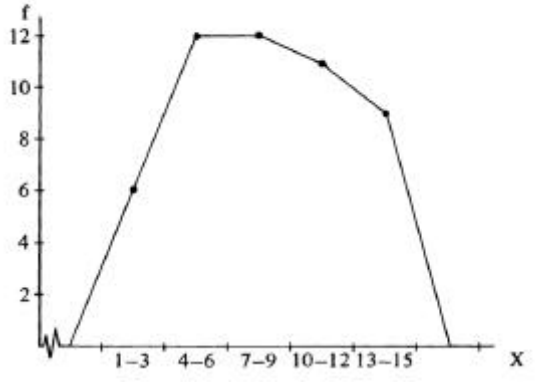
Muuttuja, joka voi saada periaatteessa kuinka monta arvoa tahansa jollakin välillä, on jatkuva. Jatkuvia muuttujia ovat esim. pituus ja paino, jotka eivät mitenkään luonnostaan vaihtele hyppäyksittäin, vaan voivat muuttua liukuvasti, periaatteessa kuinka vähän tahansa. Usein mittari on sellainen, ettei sillä voida mitata kovin hienoja eroja (esim. testi, josta voi saada vain kokonaisia pisteitä). Tämä mittauksen epäjatkuvuus ei kuitenkaan merkitse sitä, etteikö itse variaabeli, esim. testillä mitattu ominaisuus, voisi olla jatkuva. Muuttujan jatkuvuus/epäjatkuvuus sekä sen kvantitatiivisuus/kvalitatiivisuus vaikuttavat siihen, minkälaiset graafiset kuvaukset ovat havainnollisimpia. Jos halutaan korostaa erityisesti kuvatun muuttujan epäjatkuvuutta, voidaan pylväät piirtää toisistaan erilleen tai käyttää pelkkiä pystysuoria janoja havaintoluokkien kuvaamiseen.

Näin saisi esim. primäärimatriisissa oleva sukupuolimuuttuja seuraavan graafisen kuvauksen:



Jos muuttuja on, kuten sukupuoli tässä, kvalitatiivinen, on eri luokkien järjestys periaatteessa yhdentekevä asia. Voimme esityksen tarkoituksen mukaan esim. asettaa pylväät suurimmasta pienimpään tms. , kunhan vain on selkeästi esitetty, mitä kukin niistä kuvaa. Tämän valinnanvaraisuuden vuoksi on selvää, ettei kvalitatiivisen muuttujan yhteydessä voida puhua jakauman muodosta; sehän vaihtelee tutkijan valitseman esitysjärjestyksen mukaan. Kvantitatiivisen muuttujan ollessa kyseessä on sen sijaan jakauman muoto usein kiinnostava seikka eikä luokkien järjestystä saa mielivaltaisesti muuttaa. Graafisessa esityksessä X-arvot kasvavat vasemmalta oikealle ja frekvenssien, siis pylväiden korkeuden jakauma määräytyy aineistosta eikä siis ole tutkijan valittavissa.

Jos mitattu variaabeli on ainakin periaatteessa jatkuva, voimme kuvitella, että saadut arvot ovat "kiinnekohtia", joiden väliltäkin voisi saada pisteitä, jos vain mittaus olisi tarkempaa. Tällöin on usein frekvenssipolygoni käyttökelpoinen, vaikkakaan ei välttämätön kuvaustapa. Sitä tehtäessä mennään kunkin luokan keskikohdasta X-akselilta ylöspäin ko. luokan frekvenssin verran ja yhdistetään näin saadut pisteet suorin viivoin toisiinsa. Näitä viivoja voidaan pitää estimaatteina, arvioina, siitä miten pisteet jakaantuisivat saatujen pisteiden välillä, jos mittaus olisi tarkempaa. Koska kyseessä on arvio eikä vain saadun tuloksen esittäminen sellaisenaan, on usein selkeintä jättää käsiteltävänä olevalle otokselle tyypillinen sattumanvarainen vaihtelu pois ja käyttää jakauman yleistä muotoa, ts. piirtää polygoni luokitellusta aineistosta. Ohessa on tällä tavoin tehty kuvaus neliötäydennystestin tuloksista:



Ykkösten, kakkosten ja kolmosten luokkaan kuuluvia arvoja on siis saatu 6 kpl, seuraavaan luokkaan sijoittuu 12 arvoa jne. Kuvaajan päät ovat X-akselilla aineiston ylä- ja alapuolelle kuviteltujen tyhjien luokkien keskellä (luokkakeskus). Ne siis ilmaisevat, ettei näissä luokissa ole yhtään tapausta.

Esitetyt jakaumat kertoivat, kuinka moni tai kuinka suuri osa koehenkilöitä oli saanut jonkin tietyn pistemäärän. Usein tämä ei ole kuitenkaan mielenkiintoisin tieto pisteiden jakaantumisesta, vaan saatamme olla kiinnostuneita esim. siitä, kuinka moni ylti johonkin tiettyyn suoritukseen tai siitä yli, mikä pistemäärä jäi saavuttamatta puolelta testatuista jne. Tällaista tietoa tarvitaan esim. silloin, kun joudumme. karsimaan opetukseen, työhön tms. pyrkijöitä. Tällöin on avuksi laskea kumulatiiviset eli kasautuvat frekvenssit.

Kumulatiivisia frekvenssejä laskettaessa aloitetaan (yleensä) pienimmän pistemäärän, X-arvon, frekvenssistä, edetään suurempia luokkia kohti ja kerätään kaikki frekvenssit, mitä "matkan varrella" on. Kunkin luokan kumulatiivinen frekvenssi on luokan oma frekvenssi ja edellisten luokkien frekvenssit yhteensä. Kumulatiivinen frekvenssi siis ilmaisee, kuinka moni on saanut tietyn tai sitä alemman piste- arvon. Viimeisen luokan (suurimman X-arvon) kumulatiivisessa frekvenssissä on siis mukana koko mitattu joukko eli se on yhtä kuin numerus. Neliötäydennystestistä saadaan seuraavat kumulatiiviset frekvenssit (symboli F):

X	f	F
1	2	2
2	1	3
3	3	6
4	0	6
5	6	12
6	6	18
7	3	21
8	7	28
9	2	30
10	1	31
11	6	37
12	4	41
13	5	46
14	2	48
15	2	50 = N

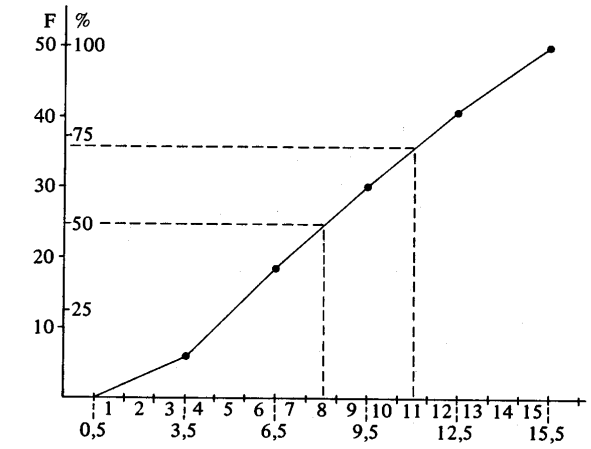
Pistemäärän 1 on siis saanut 2 koehenkilöä, pistemäärän 2 tai alle on saanut 3 (2+1), kolmosen tai alle on saanut 6 (2+1+3) henkilöä jne. Pistearvon 4 kumulatiivinen frekvenssi on myös 6, koska nelosten oma frekvenssi on nolla. Vajaa puolet koehenkilöistä, 21 kappaletta, ei ole yltänyt seitsemää pistettä parempaan suoritukseen. 13 pisteen alle on jäänyt n. 4/5 kaikista eli 41 henkilöä.

Arviointien suorittaminen pistearvojen väliltä on mukavinta kumulatiivisten frekvenssien graafisen kuvaajan, kumulatiivisen frekvenssipolygonin, avulla. Jakaumassa olevan sattumanvaraisen epätasaisuuden määrästä riippuu, kumpi on parempi lähtökohta, alkuperäinen vai luokiteltu aineisto. Neliötäydennystestin tulosten perusteella tapahtuvaan arviointiin on jo todettu luokitellun aineiston sopivan ilmeisesti paremmin. Saamme tästä seuraavat kumulatiiviset frekvenssit (F):

X	f	F
1–3	6	6
4–6	12	18
7–9	12	30
10–12	11	41
13–15	9	50 = N

Kumulatiivisessa frekvenssipolygonissa eivät ole kiinnekohtina luokkien keskukset vaan luokkarajat. Kuvattavanahan on periaatteessa jatkuva variaabeli,

joka voi saada arvoja kokonaisten pisteiden väliltäkin. Vasta kun olemme käyneet koko luokan läpi, alarajalta ylärajalle, tiedämme että kaikki luokkaan kuuluvat tapaukset ovat varmasti mukana. Loogisinta on pitää luokkien (luokkakeskusten) puolivälissä sijaitsevaa kohtaa luokkarajana. Luokiteltujen neliötäydennystestin pisteiden luokkarajoiksi saamme siis 0.5, 3.5, 6.5, 9.5, 12.5 ja 15.5. Tämän aineiston kumulatiivinen frekvenssipolygoni näyttää seuraavalta:

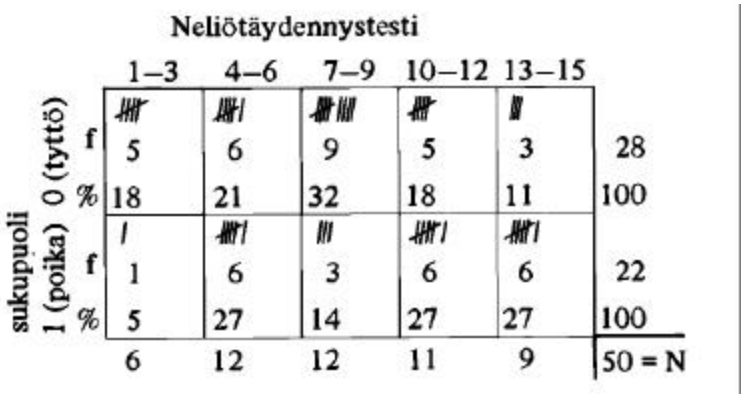


Alle 0.5 pisteen ei ole yhtään tapausta (kuvaaja on X-akselilla), alle 3.5 pisteen on 6 tapausta, alle 6.5 pisteen 18 jne. F-akselille, joka siis kuvaa kumulatiivisia frekvenssejä, on piirretty samalla myös vastaavat prosenttiarvot, 50 on 100 %; puolet siitä eli 25 on 50 % jne. Näin voidaan arvioita suorittaa sekä frekvensseinä että prosentteina saman tien. Kuvioon on pilkkuviivoin piirretty kaksi pisteiden (X-arvojen) sekä frekvenssien (tai prosenttien) vastaavuutta. Jos siis haluamme arvioida, minkä pistearvon alapuolelle jäisi 50 % koehenkilöistä vastaavassa joukossa, jos mittaus olisi jatkuvaa, siirrytään prosenttiakselilta 50:n kohdalta vaakasuoraan piirtämällemme kuvaajalle ja siitä pystysuoraan alas X-akselille, jolloin saamme arvoiksi hieman yli 8. Jos taas haluaisimme tietää, kuinka moni jäisi pistearvon 11 alapuolelle, tekisimme vastaavan operaation X-akselilta alkaen ja saisimme arvoiksi n. 36 (runsaat 70 %). Tällä tavoin voidaan muuntaa raakapiste prosenttipisteeksi tai päin vastoin. Arviotarkoituksiin tarkkuus riittää.

3. Kaksiulotteiset jakaumat: ristiintaulukointi

Edellä esitetyn kaltaisilla jakaumilla kuvataan aina yhtä variaabelia kerrallaan. Tieteen mielenkiinto suuntautuu kuitenkin hyvin usein useampien muuttujien välisiin yhteyksiin. Haluamme esim. tietää, ovatko asenteet ja sosiaaliluokka yhteydessä toisiinsa, onko sukupuolella yhteyttä joihinkin tuloksiin, voidaanko menestyminen arvioida yhdellä alalla, jos se tunnetaan toisella jne. Yhteyttä voidaan tutkia monenkin muuttujan välillä yhtaikaa, mutta keskitymme tässä edellä esitetyn kaltaisiin kahden muuttujan välisiin yhteyksiin.

Yksinkertainen ja usein riittävä esitystapa on muuttujien ristiintaulukointi. Olkoon esimerkkinä vaikkapa sukupuolen ja neliötäydennystestissä menestymisen yhteys.



Taulukon ruudut sisältävät niiden tapausten määrän, joilla on yhtaikaa jokin määriteltyarvo kahdella muuttujalla. Ensimmäinen ruutu ylhäällä vasemmalla sisältää ne tapaukset, joilla on arvo nolla (tyttö) sukupuolimuuttujalla ja jokin arvoista 1-3 neliötäydennystestissä. Näiden määrä on viisi. Pistemäärän 4-6 saaneita tyttöjä on 6 jne. Alemmassa rivissä ovat vastaavasti pojat, joista I on saanut 1-3 pistettä, 6 on saanut 4-6 pistettä jne. Reunajakaumissa ovat frekvenssit laskettuina riveittäin ja sarakkeittain. Niinpä taulukon alla on testipisteiden jakauma, kun sukupuolet on laskettu yhteen (5+1, 6+6 jne.). Tämähän on jo tuttu neliötäydennystestin jakauma luokitetusta aineistosta. Oikeana on sukupuolten

jakauma, kun eri testipisteet on laskettu yhteen; aineistossa on siis 28 tyttöä ja 22 poikaa. Reunafrekvenssien summa on sekä pystysuoraan että vaakasuoraan laskettuna 50, joka on havaintoyksikköjen kokonaismäärä (numerus).

Taulukon perusteella näyttää siltä, että sukupuolena ja neliötäydennystestinä on hieman yhteyttä. Tämä on helpointa havaita vertaamalla prosenttilukuja sarakkeittain. Frekvenssithän eivät ole suoraan verrattavissa, koska tyttöjen ja poikien määrät eivät ole samat. Tässä testissä ovat pojat olleet hieman parempia. Kahdessa ylimmässä luokassa heidän prosenttiosuutensa on suurempi, alimmassa luokassa taas pienempi. Poikkeuksen muodostaa toiseksi alin luokka (4-6), jossa poikia on hiukan enemmän.

Ristiintaulukoitavat muuttujat voivat olla joko kvalitatiivisia tai kvantitatiivisia ja ne ovat aina epäjatkuvia tai sellaisina esitettyjä. Jatkuvien muuttujien esittämisestä puhumme enemmän myöhemmin korrelaation yhteydessä. Jos taulukkoon jää paljon tyhjiä ruutuja (soluja) tai niissä on hyvin pieniä frekvenssejä, on usein aiheellista luokitella materiaali, kuten edellä esitetyssä taulukossakin tehtiin.

Varsin tavallinen ristiintaulukointi on kahden kaksiarvoisen (dikotomisen) muuttujan yhteyden esitys 2*2 -taulukkona (nelikenttä). Kuvitellaan, että jossakin oppilaitoksessa on järjestetty vapaaehtoista lisäopetusta, jonka yhteydestä opintomenestykseen ollaan kiinnostuneita. Näistä voimme muodostaa kaksi dikotomista muuttujaa: osallistunut/ei osallistunut sekä selvinnyt tentistä/ei selvinnyt. Voisimme saada esim. seuraavan tuloksen:

selviytyminen \ osallistuminen	ei	kyllä	
kyllä	20	85	105
ei	45	15	60
	65	100	165 = N

Taulukosta näkee, että osallistumisen ja selviytymisen välinen yhteys on varsin selvä, suurin osa tapauksista sijoittuu ruutuihin "ei osallistunut/ei selvinnyt" sekä "osallistunut/selvinnyt". On huomattava, että kyseessä on yhteys; osallistumisen

tai selviämisen syiden selvittäminen on huomattavasti monimutkaisempi asia eivätkä ne ole luettavissa suoraan taulukosta. Ne ovat asian suhteen tehtyjä päätelmiä, johon sovelletaan kaikkea tietämystä asiasta, jota tutkitaan. Reunajakaumista näemme, että kaikkiaan on selviytyneitä 105 ja reputtaneita 60, yhteensä 165. Osallistuneita on 100 ja ei-osallistuneita 65 , joiden summa on luonnollisesti myös 165, koska kyseessä ovat samat henkilöt eri suuntaan yhteen lasketuina.

Mielenkiintoista tietoa saattaa tarjota kahden useampiluokkaisen, kvantitatiivisen muuttujan ristiintaulukointi. Käsittelemästämme aineistosta voimme ottaa esimerkiksi loogisen järjestyksen ja sanaryhmien testit, joiden tulokset luokiteltuina ja ristiintaulukoituina ovat seuraavanlaiset:

looginen järjestys \ Sanaryhmät	5–10	11–16	17–22	23–28	
16–20	0	0	4	2	6
11–15	2	7	10	2	21
6–10	2	9	3	0	14
1–5	6	3	0	0	9
	10	19	17	4	50

Tarkasteltaessa mihin suuntaan muuttujien arvot kasvavat (vasemmalta oikealle ja alhaalta ylös) huomataan, että henkilöt, joilla on molemmissa testeissä hyvä tulos, sijoittuvat oikealla ylhäällä oleviin ruutuihin ja sellaiset, joiden tulos molemmissa on huono, joutuvat vasemmalle alas. Vastaavasti sellaiset, joiden tulos on toisessa testissä hyvä ja toisessa huono, tulevat vasemmalle ylös ja oikealle alas. Viimeksi mainituissa ruuduissa on vähän tai ei lainkaan tapauksia. Näillä kahdella testillä on siis sellainen ominaisuus, että niissä molemmissa pyrkii sa-

ma henkilö menestymään suunnilleen samalla tavalla, toisen testin tuloksesta voi karkeasti arvata (oikeammin: ennustaa) toisen testin tuloksen. Tämä johtuu

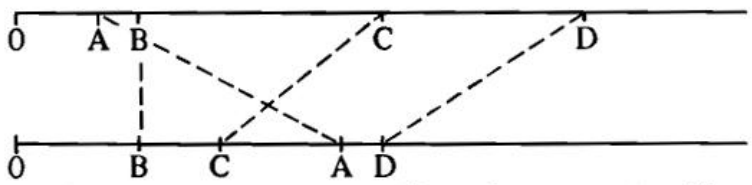
ilmeisesti siitä, että ne mittaavat samantapaista ominaisuutta, kielellistä järkeilykykyä.

Tietokoneohjelmat tuottavat ristiintaulukon yleensä kuitenkin siten, että muuttujan koodiarvot kasvavat oikealle ja alas.

4. Asteikkotyypit

Siitä väljästi määritellystä mittaamisen käsitteestä, joka aikaisemmin esitettiin, johtuu että mittausten tuloksina esitetyt luvut saattavat sisältää varsin eri määrän informaatiota, tietoa. Olemme jo todenneet eron laadullisten ja määrällisten variaabelien välillä. Silloin kun mittaustulosten sisältämää informaatiota tarkastellaan yksityiskohtaisemmin, puhutaan tavallisesti mitta-asteikoista tai skaalatyypeistä. Yleensä näitä erotetaan neljää tyyppiä.

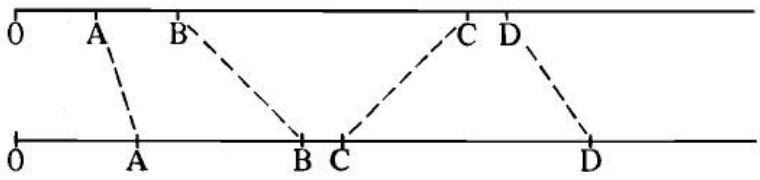
Ensimmäisenä, vähiten informaatiota sisältävänä asteikkotyyppinä voidaan pitää laatuero- eli nominaaliasteikkoa. Luvut, joilla kuvataan kvalitatiivista muuttujaa, ovat tällaisella asteikolla. Luvut ovat siis nominaaliasteikollisia, jos niiden tehtävänä on vain osoittaa yksilöiden tai ryhmien eroavuutta toisistaan ilman, että niihin sisältyy tietoa minkään ominaisuuden määrästä. Tällaisia ovat kaikki ne tapaukset, jotka aikaisemmin esitettiin esimerkkeinä kvalitatiivisista muuttujista, koehenkilöiden ja tilien numerot yms.



Nominaaliasteikko: muuttujien arvoilla ei ole määrättyä paikkaa eikä järjestystä, ylempi ja alempi kuvaus samoista neljästä muuttujasta ovat samanarvoisia.

Kvantitatiivisista asteikoista ensimmäinen on ordinaali- eli järjestysasteikko. Tällä asteikolla olevat luvut kertovat määrästä vähimmän mahdollisen: sen, onko jotakin ominaisuutta enemmän tai vähemmän. Tietoa siitä, kuinka paljon enemmän jotakin ominaisuutta on, ei tällaiseen asteikkoon sisälly. Järjestysasteikolla ovat luonnollisestikin puhtaat järjestysluvut. Jos vaikkapa kilpailuissa joku on tullut ensimmäiseksi, joku toiseksi jne., emme tiedä tästä muuta kuin, että ensimmäisen suoritus on parempi. Sitä, kuinka paljon parempi se on, eivät järjestysluvut ilmaise.

Meidän kannaltamme mielenkiintoisempi on sellainen tapaus, joka usein esiintyy käyttäytymistieteissä: jotkut luvut voivat olla järjestysasteikolla tai lähellä sitä huolimatta siitä, että ne näyttävät olevan enemmän kuin järjestyslukuja. Ajatellaanpa vaikka uutta testiä, jota ei ole aikaisemmin kokeiltu. Eri henkilöt ratkaisevat eri määrän tehtäviä ja saavat siis erilaiset pistemäärät. Toiset henkilöt näyttävät näiden pisteiden perusteella suoriutuneen hyvin samankaltaisesti toisten välisen eron taas ollessa suurempi. Näiden erojen ei kuitenkaan tarvitse välttämättä johtua mitattavasta ominaisuudesta, jolloin ne olisivat todellisia mitattuja eroja eri koehenkilöiden välillä. Ne voivat yhtä hyvin johtua testistä, siitä että sen osatehtävät vaikeutuvat epätasaisesti. Jos testissä on ryhmä samankaltaisia tehtäviä, sen jälkeen vaikeustasossa "aukko" ja taas uusi ryhmä selvästi vaikeampia tehtäviä, ryhmittyvät koehenkilöiden tulokset, vaikka heidän välisensä kykyerot olisivat tasaisesti jakaantuvia. Ainoa johtopäätös, jonka voimme tällaisen testin tuloksista hyvällä omallatunnolla tehdä, on se, että enemmän pisteitä saaneet ovat parempia, jolloin tieto on ordinaaliasteikolla.



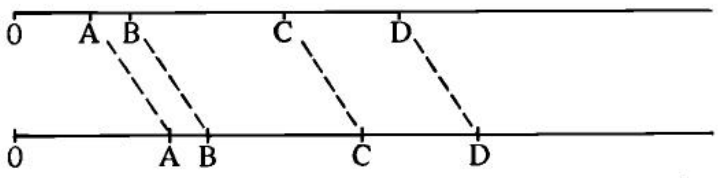
Ordinaaliasteikko: muuttujien arvojen järjestys on määrätty, paikat eivät.

Jos pistemäärien välisistä eroista on kohtuullisen luotettavaa tietoa, olemme siirtyneet välimatka- eli intervalliasteikollisen informaation esittämistapaan. Tällä asteikolla voimme sanoa, että A:n ja B:n etäisyys on tietyn suuruinen tai että A:n

ja B :n etäisyys on esimerkiksi kaksi kertaa niin suuri kuin B :n ja C:n etäisyys. Sen sijaan emme voi sanoa, kuinka monta kertaa suurempi A on B:tä. Klassinen esimerkki intervalliasteikosta on lämpömittarin lukema Celsius-asteina. Voimme sanoa, että +2 ja +4 ovat kahden asteen päässä toisistaan. Samoin on loogista sanoa, että etäisyys + 2:sta +4:een on kaksi kertaa niin suuri kuin etäisyys +4:sta +5:een. Sen sijaan emme voi sanoa, että +4:ssa on lämpötila kaksi kertaa niin suuri kuin + 2:ssa. Tämä johtuu siitä, että nolla astetta C ei merkitse lämmön loppumista, vaan nollakohta on sovittu. On tapana sanoa, että intervalliasteikolta puuttuu absoluuttinen nollakohta, tieto siitä, missä mitattua ominaisuutta ei enää ole lainkaan. Useimmat käyttäytymistieteiden käyttämät mitat yltävät korkeintaan intervalliasteikolle. Tällaisia ovat esim. hyvät, suurella koehenkilöjoukolla standardoidut testit, joista on voitu melko suurella varmuudella eliminoida se mahdollisuus, että mittari aiheuttaisi tulosten ryhmittymistä epätasaisesti. Usein tämä tieto ei ole kuitenkaan kovin luotettavaa ja niinpä sanotaankin,

että tällaiset mitat ovat "hyvällä ordinaaliasteikolla" tai "huonolla intervalliasteikolla". Usein puhutaan myös pseudointervalliasteikosta.

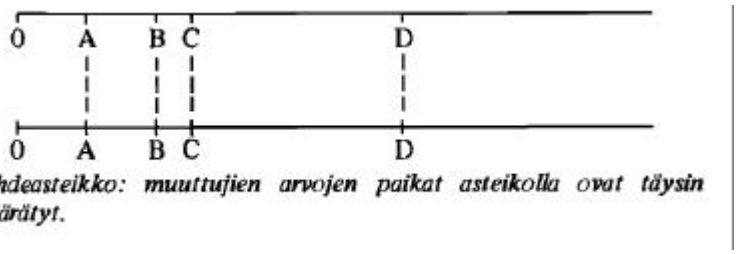
Se, että näistä mitoista puuttuu tieto absoluuttisesta nollakohdasta, tarkoittaa mm. sitä, että mitatun ominaisuuden ei suinkaan tarvitse olla kokonaan olematon silloinkaan, kun numeerisesti ilmaistu testitulos on nolla. Testi voi olla vain sillä tavalla tehty, että tarvitaan jonkin verran kykyä tai taitoa jo siihen, että "yltää asteikolle" eli saa lainkaan pisteitä. Se, että esim. alussa esittämässämme primäärimatriisissa on koehenkilön n:o 9 peilitestin tulos nolla, ei merkitse, ettei hänellä olisi lainkaan kykyä tajuta peilikuvioita. Voidaan kyllä sanoa. esim. , että kymmenen pistettä saaneella on kaksi kertaa niin paljon tehtäviä oikein kuin viisi pistettä saaneella, mutta ei voida sanoa, että hänellä olisi mitattua ominaisuutta (vaikkapa verbaalista lahjakkuutta) kaksi kertaa niin paljon.



Intervalliasteikko: muuttujien arvojen järjestys ja etäisyydet ovat määrätyt, paikka ei.

Silloin, kun edellisten tietojen lisäksi voidaan myös sanoa, että mitattu ominaisuus loppuu, kun mittaluvut saavuttavat arvon nolla, ollaan suhdeasteikolla. Tällaisia ovat kaikki fysikaalisia ominaisuuksia, pituutta, painoa, tilavuutta yms. ilmaisevat luvut. Voidaan hyvin sanoa, että joku on 1.5 kertaa niin pitkä kuin toinen, toisen paino on kaksi kertaa niin suuri kuin toisen jne. Jos jonkin asian paino on nolla, sillä ei tätä ominaisuutta, painoa, ole lainkaan.

Käyttäytymistieteissä käytetään silloin tällöin lukuja, jotka ovat todella suhdeasteikon vaatimukset täyttäviä. Voidaan esim. sanoa, että joku on ollut poissa koulusta kaksi kertaa niin paljon kuin toinen, joku on oppinut kaksi kertaa enemmän vieraan kielen sanoja ulkoa kuin toinen jne. Todella suhdeasteikollisina näitä voidaan kuitenkin pitää vain silloin, kun lukumäärät ovat sellaisinaan mielenkiinnon kohteena eikä niitä pidetä jonkin muun ominaisuuden edustajina. Jos esim. tunnilla puhumisen määrää pidetään aktiivisuuden mittana, ei voida sanoa että oppilaalla, joka ei puhu lainkaan, ei ole lainkaan aktiivisuutta.



Suhdeasteikko: muuttujien arvojen paikat asteikolla ovat täysin määrätyt.

Kvalitatiivisuuden/kvantitatiivisuuden, jatkuvuuden/epäjatkuvuuden sekä asteikkotyyppien suhteita voidaan kuvata seuraavalla kaaviolla:

epäjatkuva	nominaaliasteikko	kvalitatiivinen
	ordinaaliasteikko	
jatkuva tai epäjatkuva	intervalliasteikko	kvantitatiivinen
	suhdeasteikko	

Mittaus voi tapahtua karkein hyppäyksin tai sitten pienemmin jaotuksin. Tämän asian nimittäminen jatkuvuudeksi/epäjatkuvuudeksi on hiukan ongelmallista. Empiiriset, mitatut muuttujat ovat aina käytännössä epäjatkuvia. Havaintoyksikkö kuuluu tiettyyn arvoluokkaan, jolla on luokkakeskus ja luokkarajat ja luok-

kaväli. Teoreettiset jakaumat ovat puolestaan yleensä jatkuvia: normaalijakauma, t-jakauma jne. Kun teoreettinen jakauma voi saada vain tiettyjä äärellisiä pistearvoja, se on epäjatkuva kuten esim. binomijakauma. Oletettu mittauskohde saattaa olla jatkuva muuttuja, sen indikaattorina toimiva mittaustapa tuottaa epäjatkuvia arvoja. Käytännössä jaottelu tarkoittaa usein seuraavaa. Useampiluokkainen kvantitatiivinen mittaus, josta uskalletaan tehdä tasavälisyyttä koskeva toteamus on jatkuva. Kategoriseksi sanotaan kvantitatiivista mittausta, jonka luokat ovat suuruusjärjestyksessä, mutta josta ei voi tehdä tasavälisyysoletusta. Esim. elokuvissa käynnit edeltävinä neljänä viikkona on indikaattorina "jatkuva" (vaikkakin se saa vain kokonaislukuja arvokseen) latentille piirteelle kiinnostus elokuvia, joka on jatkuva muuttuja, mutta jota on vaikea mitata tarkasti. Jatkuvuus ei siis ole erityisen tärkeä kriteeri mittauksen luonteelle. Jopa muuttujaa, jonka vaihtoehdot ovat sanallisesti: ei juuri koskaan, melko usein, lähes aina, voidaan käsitellä kolmiluokkaisena kvantitatiivisena muuttujana intervalliasteikon tapaan (eli jatkuvana) ja laskea sen koodeista (1, 2, 3) keskiarvoja ja hajontoja, jotka edellyttävät pseudo-intervallisuutta. On tärkeä hahmottaa ero kvantitatiivisuuden ja kvalitatiivisuuden välillä.

Asteikkotyypit ovat tärkeitä siksi, että niistä riippuu, minkälaisia johtopäätöksiä luvuista voidaan tehdä ja minkälaisia laskutoimituksia niillä voidaan suorittaa. Sallituista laskutoimituksista puhutaan eri tilastollisten menetelmien yhteydessä myöhemmin, mutta jo tieto siitä, kuinka erilaista lukujen antama informaatio on, auttaa käyttämään niitä mielekkäästi.

5. Keskiluvut

Kaikkein pisimmälle on informaation tiivistämisessä menty silloin, kun otosta kuvataan vain yhdellä luvulla, joka mahdollisimman hyvin edustaa kaikkia otoksen arvoja. Tällaisia lukuja nimitetään keskiluvuiksi. Niistä käsitellään tässä yhteydessä kolmea: moodia, mediaania ja aritmeettista keskiarvoa.

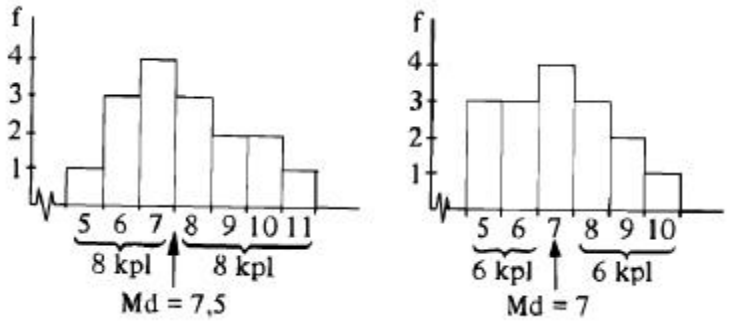
Keskiluvut ovat hyvä esimerkki siitä, miten tilastollisten menetelmien yhteydessä kätevyys ja lyhyys saadaan aikaan uhraamalla muuta informaatiota. Jos kuvaamme otosta yhdellä luvulla, joka edustaa sen kaikkia arvoja, menetämme tiedon siitä, missä eri yksilöt jakaumassa sijaitsevat, minkälainen on jakauman muoto, kuinka laajalle alueelle arvot hajaantuvat jne. Tämän haitan vähentämiseksi käytetään usein keskilukujen yhteydessä hajaantumislukuja, joita esitellään myöhemmin.

Moodi eli tyyppiarvo (Mo) on yksinkertaisin keskiluku. Se on yhtä kuin eniten esiintyvä muuttujan arvo, siis se, jonka frekvenssi on suurin. Niinpä graafisessa kuvauksessa tyyppiarvo on se X:n arvo, jonka kohdalla jakauma on korkein (suurin frekvenssi), mikä tekee sen helpoksi paikallistaa. Jos jälleen tutkimme aluksi esitettyä primäärimatriisia ja siitä tehtyjä jakaumia, voimme esim. todeta, että sukupuolen tyyppiarvo on 0, ts. tyttöjä on enemmän kuin poikia. Neliötäydennystestin tyyppiarvo on 8, tätä arvoahan esiintyi jakaumassa eniten, 14 kpl.

Kvalitatiivista muuttujaa kuvatessamme emme voi käyttää muuta keskilukua, luvuillahan ei ole mitään määrättyä järjestystä eikä mikään arvo ole sen enempää "keskellä " kuin muutkaan. Oikeastaan on harhauttavaa edes nimittää moodia keskiluvuksi laadullisessa muuttujassa. Kvantitatiivisten muuttujien kuvaamiseen voidaan moodia käyttää silloin, kun halutaan nopea arvio otoksen kuvaamiseksi eikä jakauma ole kovin epäsäännöllinen. Tällöin ovat kaikki keskiluvut lähellä toisiaan. Jos kvantitatiivisen muuttujan arvot on luokiteltu, pidetään suurimman luokan keskimmäistä arvoa, luokkakeskusta, moodina.

Jos kuvattava lukujoukko on vähintään ordinaaliasteikolla ja lukujen järjestys toisiinsa nähden on siis täysin määrätty, voidaan käyttää keskilukuna mediaania (Md). Mediaani on se variaabelin arvo, jonka kummallekin puolelle jää 50 % kaikista tapauksista. Silloin kun mediaani sattuu juuri luokan keskelle tai kahden

luokan väliin, ei sen määrääminen tuota vaikeuksia. Näin on seuraavissa esimerkeissä:



Usein nähdään jakaumaa tai lukuja tarkastelemalla vain se; mihin luokkaan mediaani sijoittuu, mutta sen sijainti tämän luokan sisällä jää epävarmaksi. Varsinkin luokitellussa materiaalissa jää epävarmuusalue tällöin kovin suureksi. Tällöin voidaan arvo täsmentää laskemalla. Tämä tulee kyseeseen varsinkin silloin, kun arvioimme periaatteessa jatkuvaa variaabelia epäjatkuvan mittauksen tulosten perusteella. Laskukaava on seuraava:

$$Md = X_o + \frac{\frac{N}{2} - F_{m-1}}{f_m} \cdot \omega$$

Kaavan symbolien selitykset ovat:

X_o = Mediaaniluokan alaraja
f_m = Mediaaniluokan frekvenssi
F_{m-1} = Mediaaniluokkaa edeltävän luokan kumulatiivinen frekvenssi
ω (omega) = Luokkaväli

Ensin joudutaan arvioimaan missä luokassa mediaani on X-asteikolla. Sen jälkeen luokan alarajaan lisätään se osuus luokan pinta-alasta, mikä tarvitaan 50 %:iin pääsemiseksi.

Laskemista varten tarvitaan siis luokkarajat, luokkien frekvenssit sekä kumulatiiviset frekvenssit. Nämä meillä ovat valmiina neliötäydennystestin tuloksista, joten voimme käyttää sitä esimerkkinä:

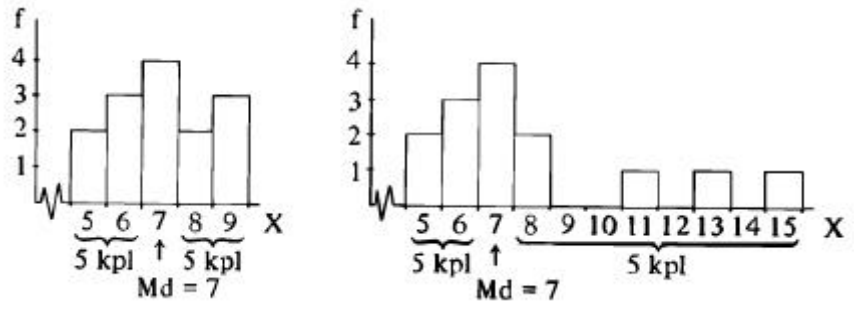
X	luokka-rajat	f	F
	0.5		
1–3	3.5	6	6
4–6	6.5	12	18
7–9	9.5	12	30
10–12	12.5	11	41
13–15	15.5	9	50 = N

Koska periaatteessa mediaanin kummallekin puolelle sijoittuu puolet koko aineistosta, on meidän tässä etsittävä kohta, jonka kummallakin puolella olisi 25 tapausta. Sen täytyy sijaita keskimmäisessä luokassa, jossa ovat pistemäärät 7 - 9, koska ennen kuin tähän luokkaan tullaan, on kumulatiivinen frekvenssi vasta 18 ja luokan ylärajalle tultaessa se on jo 30. Saamme siis mediaaniluokan alarajaksi 6.5, sen frekvenssiksi 12 ja edellisen luokan kumulatiiviseksi frekvenssiksi 18. Luokkaväli on etäisyys yhden luokan rajalta toiselle eli tässä tapauksessa 3. Sijoittamalla luvut kaavaan ja laskemalla saamme mediaaniksi 8.25:

$$Md = 6.5 + \frac{\frac{50}{2} - 18}{12} \cdot 3 = 6.5 + 1.75 = 8.25$$

Mediaani määritellään siis kohdaksi, joka jakaa jakauman pinnan kahteen yhtä suureen osaan. Jakauman pinta-alahan on suoraan verrannollinen tapausten lukumäärään. Tämä kohta voidaan mukavasti määritellä kumulatiivisen frekvenssipolygonin avulla. Tämähän on jo aikaisemmin tehty, kun etsittiin sitä pistelukua, jonka alle jää 50 % kaikista tapauksista. Tällöin totesimme tuloksen olevan jonkin verran yli 8.

Selvästi ordinaaliasteikollisessa muuttujassa mediaani on paras keskiluku, jota voidaan käyttää. Intervalli- ja suhdeasteikollakin sitä voidaan joskus käyttää mikäli aritmeettista keskiarvoa ei tarvita tai se on hankalammin hankittavissa. Samoin jos jakaumassa on kaukana muusta joukosta olevia ääriarvoja, joiden ei haluta vaikuttavan keskilukuun, on mediaani sovelias. Mediaaniinhan eivät vaikuta arvojen etäisyydet, vaan ainoastaan frekvenssit, kuten esim. seuraavat jakaumat havainnollistavat:



On huomattava, että tilasto-ohjelmat laskevat mediaanin arvon siten, että se on se luokkakeskus, jossa mediaani sijaitsee. Yllä oleva on enemmänkin sitä, että tutustutaan käsitteisiin luokka, luokkakeskus, luokkaväli. Halutessasi tarkastella mediaania ja tutustua jakauman muotoon pienillä aineistoilla on esitystapa stem-and-leaf varsin käyttökelpoinen. Se löytyy mm. Spss:n valikoimasta. Aluksi se näyttää sekavalta, mutta on kuitenkin käyttökelpoinen kuten myös esitystapa box-plot (etsi sopivasta kirjasta tai tutustu Spss-ohjelmalla).

Intervalli- tai suhdeasteikollisessa lukujoukossa tarvitaan mahdollisimman täydellisen informaation saamiseksi keskilukua, joka ottaa huomioon myös pistelukujen etäisyydet. Tällainen on aritmeettinen keskiarvo, josta usein käytetään. pelkästään keskiarvo-nimitystä. Sen symboleina ovat joko viiva-X tai M (engl. mean). Aritmeettinen keskiarvo on keskiluku, jota käytetään paljon arkielämänkin tilanteissa, jolla lasketaan vaikkapa todistusten keskiarvo tms. Se määritellään kaavalla:

$$\bar{X} = \frac{\Sigma X}{N}$$

Primääriarvot (raakapisteet, X-arvot) siis lasketaan yhteen ja summa jaetaan tapausten lukumäärällä (N). Niinpä esim. lukujen 1, 3, 4, 3, 5 ja 6 keskiarvo saadaan jakamalla näiden lukujen summa (22) niiden määrällä (6). Tulokseksi tulee 3.6666, pyöristettynä 3.67.

Taulukoidun materiaalin käsittelyssä voidaan käyttää kaavaa:

$$\bar{X} = \frac{\Sigma fX}{N}$$

Tämä tulkitaan siten, että kukin X-arvo (luokkakeskus) kerrotaan sitä vastaavalla frekvenssillä ja tulojen (fX) jaetaan numeruksella.

Olkoon esimerkkinä vaikkapa seuraava:

X	f	fX
5	5	25
6	5	30
7	20	140
8	8	64
	N = 38 =Σf	259 = ΣfX

$$\bar{X} = \frac{259}{38} = 6.8$$

Jos kyseessä on luokiteltu materiaali, käytetään luokkakeskuksia X-arvoina. Esimerkkinä on neliötäydennystesti, erikseen tytöille ja pojille:

tytöt:

X	luokka-keskus	f	fX
1–3	2	5	10
4–6	5	6	30
7–9	8	9	72
10–12	11	5	55
13–15	14	3	42
		N = 28	209 = ΣfX

$\bar{X}$ (tytöt) = $\frac{209}{28}$ = 7.46

pojat:

X	luokka-keskus	f	fX
1–3	2	1	2
4–6	5	6	30
7–9	8	3	24
10–12	11	6	66
13–15	14	6	84
		N = 22	206 = ΣfX

$\bar{X}$ (pojat) = $\frac{206}{22}$ = 9.36

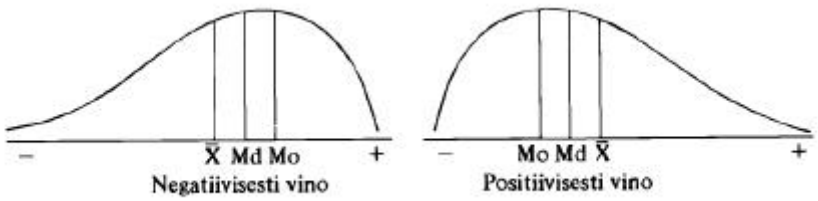
Olemme aikaisemmin taulukoineet ristiin sukupuolen ja neliötäydennystestin, jolloin totesimme, että jakauman mukaan näyttävät pojat menestyvän tässä testissä hieman paremmin kuin tytöt. Tässä on nyt sama esitettynä keskiarvoin. On syytä huomata, kuinka tiivistettyä keskiarvon antama informaatio on; jos jakauman muoto sisältää mielenkiintoista tietoa, on syytä esittää keskiarvojen lisäksi myös em. ristiintaulukointi.

Aritmeettinen keskiarvo on sopivin ja tässä esitetyistä eniten informaatiota sisältävä keskiluku intervalli- ja suhdeasteikolle. Se on myös niistä sopivin lähtökoh-

ta jatkotoimenpiteille. Vaatimus intervalliasteikosta ei ole aivan ehdoton, olemmehan todenneet hyvin monien käyttäytymistieteellisten mittausten tulosten sijoittuvan jonnekin ordinaaliasteikon ja intervalliasteikon välimaille. Jos kuitenkin aineistossa on keskimääräisestä paljon poikkeavia lukuja, joiden etäisyys ei saisi vaikuttaa, on syytä käyttää esim. mediaania. Tilanne voi olla tällainen esim. silloin, kun poikkeavien lukujen etäisyyden voidaan epäillä johtuvan virheestä tai mittarista eikä mitattavasta ominaisuudesta, ts. mittaus on lähellä ordinaaliasteikkoa.

Jos jakauma on yksihuippuinen ja symmetrinen, ovat kaikki keskiluvut samassa kohdassa jakauman keskellä. Tavallinen poikkeama säännöllisestä jakaumasta on vinous. Vinon jakauman suurin frekvenssi ei ole keskellä; toisessa reunassa on suhteellisen harvoja mutta etäällä muusta joukosta olevia arvoja. Jakauma on negatiivisesti tai positiivisesti vino sen mukaan, missä: nämä ääriarvot sijaitsevat. Vinous aiheuttaa sen, että keskiluvut poikkeavat toisistaan. Aritmeettinen keskiarvo siirtyy ääriarvojen suuntaan ja tyyppiarvo taas

vastakkaiseen suuntaan, jossa frekvenssi on suurin. Suhdetta kuvaavat seuraavat jakaumat:



Vertailukohteena on usein normaalijakauma, joka on yksihuippuinen ja symmetrinen (teoreettinen) jakauma. Mittareita laaditaan usein siten, että ne sovitetaan tuottamaan kohdejoukossa normaalijakaumaa lähestyviä jakaumia otoskohtaisesti (esim. koulusaavutuskokeet). Jakaumaan voidaan vaikuttaa mittarin osioiden ominaisuuksilla. Se ei siis ole mitattavan piirteen ominaisuus.

6. Hajontaluvut

Edellä on jo mainittu, että keskilukujen sisältämän informaation vähyys voidaan osittain kompensoida käyttämällä lisänä hajontalukuja. Kun keskiluku pyrkii mahdollisimman hyvin yhdellä luvulla kuvaamaan koko otoksen arvojen absoluuttista kokoa, pyritään vastaavasti hajontaluvulla kuvaamaan arvojen suhteita, sitä kuinka kaukana ne ovat toisistaan. Jos jakauma ei ole kovin epäsäännöllinen, kuvaavat keskiluku ja hajontaluku otosta jo melko hyvin. Periaatteessa hajontalukuja voidaan käyttää vain intervalli- ja suhdeasteikolla, eihän voida puhua hajaantumisesta, jos etäisyyksistä ei tiedetä mitään. Tämäkään vaatimus ei ole kovin jyrkkä, sillä jo nominaaliasteikollisessa aineistossa voidaan hajaantuminen tulkita siten, että vaihtelua on enemmän, jos tapaukset edustavat useampaa luokkaa. Samoin voidaan sanoa, että sellaisessa aineistossa, jossa eri luokat ovat suhteellisen tasaisesti edustettuina, on enemmän hajontaa kuin sellaisessa, jossa suuri osa tapauksista sijoittuu vain harvoihin luokkiin.

Hajontaluvuista yksinkertaisin on vaihteluväli (engl. range), joka ilmaisee suurimman ja pienimmän pistearvon välisen etäisyyden. Esim. neliötäydennystestin pienin saatu pistearvo oli 1 ja suurin 15. Vaihteluväliksi tulee tällöin 15 - 1 eli 14 pistettä. Jo tällä karkealla mitalla voimme todeta, että esim. yhteenlaskutestin vaihteluväli on huomattavasti suurempi, 54-0 eli 54 pistettä. Samalla käy ilmi vaihteluvälin suuri puute hajaantumisen ilmaisemisessa. 54 pistettä on nimittäin yhteenlaskutehtävissä poikkeuksellisen hyvä suoritus, toiseksi parhaalla on pistemäärä 34. Jos jakaumaa ei esitetä, saa pelkästä vaihteluvälistä helposti hyvin harhaisen kuvan irrallisten ääriarvojen vuoksi. Tämän virheen välttämiseksi voidaan hajaantumisluku muodostaa siten, että siinä tulevat laitimmaisten lukujen lisäksi huomioonotetuiksi kaikkien lukujen etäisyydet

toisistaan. Näin on menetelty keskimääräisessä poikkeamassa ja keskihajonnassa. Sen sijaan että laskettaisiin kaikkien lukujen etäisyydet kaikkiin muihin, päästään vähemmällä työllä oleellisesti samaan informaatioon laskemalla kunkin luvun etäisyys keskiarvosta. Keskimääräinen poikkeama (engl. average deviation) onkin sananmukaisesti lukujen keskimääräinen etäisyys keskiarvosta. Tämä voidaan esittää kaavana:

$$KP = \frac{\Sigma\,|X-\bar{X}|}{N}$$

Toisin sanoen, jokaisen luvun ja keskiarvon etäisyyden itseisarvot lasketaan yhteen ja tämä summa jaetaan tapausten määrällä. Esim. pistelukujen 1, 2, 2, 3, 5, 4, 4 ja 3 keskimääräiseksi poikkeamaksi saamme arvon 1. Aritmeettinen keskiarvo on 3 ja poikkeamien itseisarvojen summa on 8.

Koska keskimääräinen poikkeama on sananmukaisesti lukujen ja keskiarvon etäisyyksien keskiarvo, se on tulkinnallisesti selkeä. Se soveltuu kuitenkin huonosti jatkotarkastelujen pohjaksi, joten sen käyttö on jäänyt melko vähäiseksi. Poikkeama keskiarvosta (deviaatio) tarjoaa kuitenkin lähtökohdan hajonnalle (standardipoikkeama eli keskihajonta, engl. standard deviation).

Aloitamme neliösummasta (engl. sum of squares). Kyseiset poikkeamat korotetaan toiseen potenssiin ja lasketaan yhteen. Kun neliösummavaihtelu jatetaan termillä N - 1 päädytään keskineliöön (engl. mean square, N - 1 termiä kutsutaan vapausasteiksi). Se tunnetaan huomattavasti paremmin nimellä varianssi. Kun varianssista otetaan positiivinen neliöjuuri, meillä on lukuarvona hajonta. Sen havainnollisuus ja käyttökelpoisuus liittyy hyvin selkeästi normaalijakaumaan. Jos pistearvot jakautuvat normaalijakauman tapaan, yksittäisen pistemäärän sijainti jakaumassa (suhteessa muihin tapauksiin) voidaan määrittää kohtuullisen hyvin. Kun tarkastelet keskihajonnan kaavaa löydät sieltä kyseiset komponentit:

$$s = \sqrt{\frac{\Sigma\,(X-\bar{X})^2}{N-1}}$$

Yhteys normaalijakaumaan tekee hajonnasta jonkinlaisen mittayksiköistä riippumattoman yleismitan, joka tekee hyvin erilaisten muuttujien vertailun mahdolliseksi. Varianssin ja standardipoikkeaman laskemisen havainnollistamiseksi voimme laskea ne esim. seuraavasta lukujoukosta: 5, 5, 6, 5, 3, 4, 6, 7, 4, 5 (N=10).

X	$X-\bar{X}$	$(X-\bar{X})^2$
5	0	0
5	0	0
6	+1	1
5	0	0
3	−2	4
4	−1	1
6	+1	1
7	+2	4
4	−1	1
5	0	0
		$12 = \Sigma(X-\bar{X})^2$

Neliösumma on 12. Varianssi (eli keskineliö) on 1.33 ja hajonta 1.15. Varianssi ei siis ole tarkasti ottaen aritmeettinen keskiarvo neliösummasta.

Jos aineisto on jo taulukkomuodossa, käytetään X:n arvoina luokkakeskuksia (tämä on tärkeä muistaa, luokkaväli voi olla muukin kuin 1). Koska luokassa on useita tapauksia, kyseisiä poikkeamia on luokkafrekvenssin määrä. Seuraava esimerkki selventää asiaa:

X	f	fX	$X-\bar{X}$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
2	2	4	−3.5	12.25	24.5
3	4	12	−2.5	6.25	25.0
4	4	16	−1.5	2.25	9.0
5	8	40	−0.5	0.25	2.0
6	10	60	+0.5	0.25	2.5
7	6	42	+1.5	2.25	13.5
8	5	40	+2.5	6.25	31.3
	$N = 39$ $= \Sigma f$	214 $= \Sigma fX$			107.8 $= \Sigma f(X-\bar{X})^2$

Keskiarvoksi saadaan 5.49, mikä on pyöristetty arvoksi 5.5. Neliösumma on 107.8. Varianssin arvoksi tulee 2.76 ja hajonnaksi 1.66. Näin varsin harvoin asia lasketaan. Yleensä muuttujien keskiarvot ja hajonnat lasketaan ohjelmalla suoraan raakapisteaineistosta siinäkin tapauksessa, että jakauma esitetään luokitettuna jollain karkeammalla tavalla, jolloin luokkaväli muuttuu alkuperäisestä suuremmaksi.

Raakapisteistä laskettaessa lienee kätevintä laskea neliösumman (SS) kaavalla:

$$SS_{total} = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \qquad SS = \sum X^2 - \sum X^2 / N$$

Tarvitsemme raakapisteiden summan ja raakapisteiden neliöt ja niiden summan. Edellinen tulee jo keskiarvon laskun yhteydessä. SS jaettuna N-1:llä antaa varianssin ja sen neliöjuurena tulee hajonta. Tämän kaavan alaindeksin annetaan jo tässä vaiheessa vihjata siitä, että neliösummavaihtelu voidaan paloitella selitettyyn (between) ja jäännökseen (residuaali) esimerkiksi varianssianalyysin yhteydessä (myöhemmissä opinnoissa).

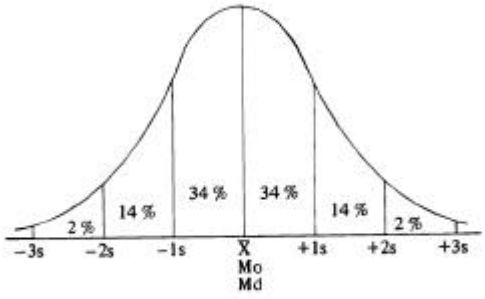
Ehkä on syytä vielä mainita, että hajontaluvut ovat samaa yksikköä kuin primääriarvotkin. Jos siis alkuarvot ovat metrejä, on hajontalukukin metrejä, jos lähtöarvot ovat oikein ratkaistujen tehtävien lukumääriä, ovat hajaantumisluvutkin näitä.

7. Normaalijakauma ja standardipisteet

Aiemmin olemme esittäneet joitakin variaabelin jakaumia histogrammien ja frekvenssipolygonien muodossa. Jos kuvittelemme, että mittaamme varsin tarkasti ja jatkuvaksi katsottavissa olevalla tavalla jotakin sattumanvaraisesti jakaantuvaa ominaisuutta hyvin suurella koehenkilöjoukolla, saamme symmetrisen jakauman, josta histogrammille ominaiset hyppäykset ovat pyöristyneet pois. Tällainen jakauma on lähellä normaalijakaumaa eli Gaussin kellokäyrää, jolla on tilasto- tieteessä yhä keskeinen paikkansa asema. Varusmiesten pituudet voisivat olla melko hyvä esimerkki tästä.

Normaalijakauman puitteissa käytetään mittayksikkönä standardipoikkeamaa eli keskihajontaa, joka esiteltiin hajaantumislukujen yhteydessä. Standardipoikkeama jakaa normaalijakauman pinnan tarkoin määriteltävissä oleviin osiin, joiden ulkopuolelle jäävä pinta-ala pienenee edettäessä keskiarvosta poispäin, ts. mitä kauempana keskiarvosta ollaan, sitä vähemmän tapauksia ko. kohdalla ja kasautuvasti kauempana on.

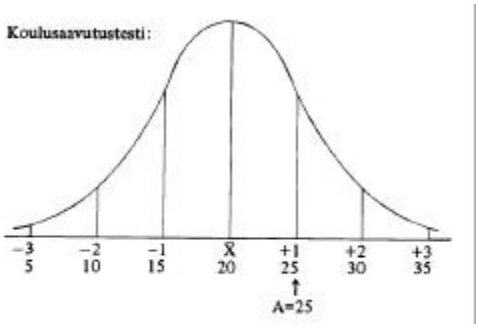
Käyrä on asymptoottinen, ts. se ei koskaan leikkaa X-akselia, mutta käytännössä melkein koko sen pinta-ala on kuuden standardipoikkeaman sisällä (-3 - +3). Koska jakauman pinta-alan osat ovat suorassa suhteessa frekvensseihin, ovat pintojen osuudet koko jakaumasta samalla jakaumasta sattumanvaraisesti valitun yksilön todennäköisyyksiä osua ko. alueelle. Normaalijakauman pinta-ala jakaantuu standardipoikkeamittain suunnilleen seuraavasti:

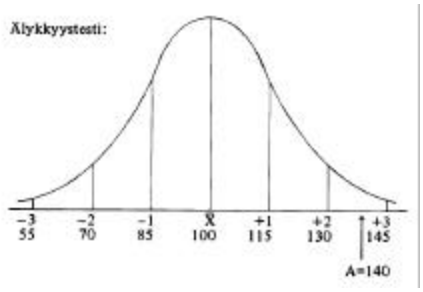


Sattumanvaraisesti valittu yksilö on siis suurella todennäköisyydellä suhteellisen lähellä keskiarvoa. Alueelle, joka sijaitsee keskiarvosta korkeintaan yhden standardipoikkeaman verran molempiin suuntiin, jää 68 % kaikista tapauksista jne. Voimme vaikka kuvitella populaation, jossa ihmisten keskipituus on 170 cm ja pituuden hajonta 5 cm. Tällöin populaatiosta sattumanvaraisesti valittu yksilö sijaitsee 68 %:n todennäköisyydellä välillä 165-175 cm. Kuten kuvioon on merkitty, standardipoikkeamat lasketaan keskiarvosta kumpaankin suuntaan, keskiarvoa pienemmät arvot ovat standardipoikkeamina negatiivisia ja suuremmat positiivisia. Jakauman symmetrisyyden ja säännöllisyyden takia kaikki keskiluvut sattuvat samaan kohtaan.

Koska standardipoikkeaman suuruus on normaalijakaumassa aina tarkoin määrättävissä, se muodostaa mitan, jolla yksittäinen arvo voidaan suhteuttaa suurempaan joukkoon. Niinpä voidaan eri mittareilla hankittuja lukuja verrata suoraan keskenään, kun ne ilmaistaan alkuperäisen mitan (esim. testipisteet, pituus, todistuksen arvosanat) sijasta standardipoikkeamina. Kuvitellaanpa vaikka että henkilö A on saanut älykkyystestissä tuloksen 140 ja koulusaavutustestissä 25. Tiedetään, että älykkyystestin keskiarvo on 100 ja keskihajonta 15. Koulusaavutustestin vastaavat arvot on 20 ja 5.

Vastaako saavutustestin suoritus henkilö A :n älykkyystasoa? Näemme heti, että tulos kummallakin variaabelilla on yli keskiarvon. Voimme kuitenkin tarkentaa tätä tietoa huomattavasti. Asia selvinnee lisää, jos piirrämme molempien variaabelien jakaumat ja tutkimme mihin kohtaan A:n suoritukset niissä sijoittuvat:





Älykkyystestissä yksi hajonta keskiarvosta ylöspäin vastaa 100+15=115 pistettä, kaksi 100+15+15=130 pistettä jne. A:n saama pistemäärä 140 on siis yli 2.5 hajonnan mittaa keskiarvon yläpuolella. Tämä on varsin korkea arvo, sillä sen paremmalle puolelle sijoittuu vain n. 0.5 % kaikista arvoista. Koulusaavutustestin kohdalla +1 standardipoikkeamaa sijoittuu 20+5=25 pisteen kohdalle, +2 on 20+5+5 =30 pistettä jne. A:n saama arvo, 25, on siis +1 standardipoikkeamaa. Tämäkin arvo on selvästi keskiarvon yläpuolella, mutta vielä n. 16 % kaikista arvoista on sitä parempia. Standardipoikkeamina ilmaistuja arvoja voidaan verrata suoraan keskenään ja todeta, että henkilö A ei ole menestynyt saavutustestissä tavalla, jota älykkyystestin perusteella olisi voinut odottaa.

Sama operaatio voidaan suorittaa laskemallakin, jolloin siis muunnetaan primääripisteet standardipoikkeamiksi ja verrataan näitä suoraan toisiinsa. Muuntamiskaava on:

$$Z = \frac{X - \bar{X}}{s}$$

Muunnettuja pisteitä nimitetään standardipisteiksi ja niitä merkitään Z:lla. Myöhempi Pearsonin tulomomenttikerroin (korrelaatio) perustuu tällä tavalla muodostettuihin lukupareihin. Äskeisessä esimerkissä saamme siis seuraavat arvot:

älykkyys:	koulusaavutus:
$Z = \frac{140 - 100}{15} = +2.67$	$Z = \frac{25 - 20}{5} = +1$

Jos primääriarvo on keskiarvoa pienempi, saamme negatiivisen Z-pistemäärän. Esimerkiksi: X = 55 kun keskiarvo = 70 ja hajonta = 17, saamme:

$$Z = \frac{55 - 70}{17} = -.88$$

Standardipistemäärinä ilmaistut mittausten tulokset ovat monessa mielessä käyttökelpoisia. Ne ovat suoraan toistensa kanssa vertailukelpoisia, niistä näkee heti, ovatko ne keskiarvon ylä- vai alapuolella, ja ne ovat helposti suhteutettavissa todennäköisyyteen saada ko. arvo. Asiaan perehtymätöntä negatiiviset luvut ja useat desimaalit kuitenkin helposti hämäävät. Tämän haitan vähentämiseksi käytetään joskus niiden lineaarisia muunnoksia vaikkapa T-pistemääriä, jotka saadaan seuraavan kaavan avulla: T=50+10Z. Toisin sanoen tunnettu standardipistemäärä kerrotaan kymmenellä ja siihen lisätään 50. Tällöin tullaan asteikkoon, jonka keskiarvo on 50 eikä negatiivisia lukuja esiinny lainkaan. Esimerkiksi Z-arvosta +0.6 tulee 50+10(0.6)=56 T-pistettä, Z-arvosta -1.2 saamme 50+10(-1.2)=38 T -pistettä.

Jos mittaamme jotakin sattumanvaraisesti jakaantuvaa ominaisuutta vähintään intervalliasteikolla ja tarpeeksi suurta otosta käyttäen, saamme suunnilleen normaalin jakauman. Näin on esim. ihmisten pituuden ja painon laita: hyvin lyhyitä ja kevyitä on vähän, keskimääräisiä eniten ja hyvin painavia ja pitkiä taas vähän. Usein saadaan kuitenkin käytännössä jakaumia, jotka poikkeavat normaalista jopa huomattavastikin. Tämä johtuu periaatteessa kahdesta tekijästä tai niiden yhdistelmästä. Ensiksikin ominaisuus on sellainen, ettei se jakaannu sattumanvaraisesti, vaan pyrkii keskittymään esim. sellaisille yksilöille, joilla sitä on jo ennestäänkin. Voidaan esim. ajatella jonkin tietoalueen hallintaa, joka tulee sitä helpommaksi, mitä enemmän taustatietoja yksilöllä on. Tällaisessa tilanteessa voidaan odottaa vinoa jakaumaa. Kasautuvuus voi tulla näkyviin myös siten, että jakaumassa korostuvat laidoilla olevat arvot. Esimerkiksi asenteet jotakin tunnepitoista ongelmaa, vaikkapa rotukysymystä, kohtaan voivat olla itseään vahvistavia: sekä negatiivisista että positiivisista näkemyksistä pidetään lujasti kiinni ja neutraaleja asenteita on vähän. Toinen syy jakauman ei-normaaliuteen on mittari tai sen käyttötilanne: jos mittari, esim. testi, on liian helppo, saavat monet hyviä pisteitä ja jakaumasta tulee negatiivisesti vino. Vaikeassa testissä saavat vain harvat korkeita arvoja ja jakauma on positiivisesti vino.

Edellisessä tapauksessa, jossa jakauman poikkeaminen normaalista on informaatiota jostakin aineistossa vallitsevasta tekijästä tai ominaisuudesta, on tämä poikkeaminen luonnollisesti tuotava esiin tutkimustuloksena eikä jakauman muotoa saa ruveta mielivaltaisesti muuttamaan. Toisessa tapauksessa, jossa poikkeama normaalista johtuu mittauksen tai mittarin ominaisuuksista, voidaan jakauma muuttaa lähelle normaalia, normalisoida. Normalisointi voidaan edelliseen perustuen siis tehdä

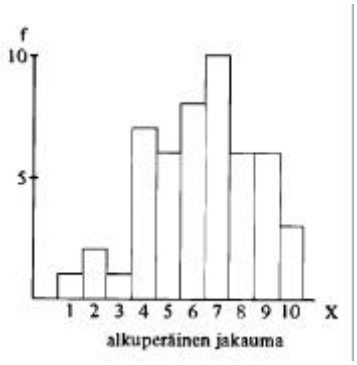
a) jos on perusteita uskoa ominaisuuden jakaantuvan normaalisti,

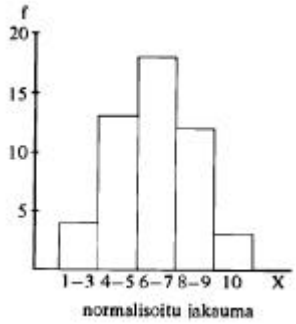
b) jos on aihetta uskoa mittauksen tai mittarin aiheuttaneen poikkeamia normaalista.

On selvää, ettei useinkaan voida olla varmoja siitä, onko ehdot täytetty vai ei. Jos normalisointi tällöin suoritetaan, on siitä ilmoitettava, jotta lukija osaa ottaa tämän huomioon.

Normalisoinnissa yhdistellään vierekkäisiä pistemääriä siten, että saatujen luokkien frekvenssit noudattavat mahdollisimman hyvin normaalijakaumaa. Jos on mahdollista ja aiheellista muodostaa kuusi luokkaa, voidaan ohjearvoina käyttää niitä prosenttilukuja, jotka normaalijakaumassa syntyvät standardipoikkeamiin jaettaessa. Jos viisi luokkaa tuntuu sopivalta, voidaan pyrkiä suunnilleen jakaumaan 7, 24, 38, 24, 7 prosenttia jne. Seuraavassa esimerkki, jossa aineisto on normalisoitu viisiluokkaiseksi :

X	f	luokkaf	%
1 2 3	1 2 1	4	8
4 5	7 6	13	26
6 7	8 10	18	36
8 9	6 6	12	24
10	3	3	6
N = 50			





Alkuperäinen, hieman negatiivisesti vino jakauma on saatu lähes symmetriseksi ja tasaiseksi. Täsmälleen haluttuihin prosenttilukuihin ei tässä normalisoinnissa päästä; jos vaikkapa ainoastaan ykköset ja kakkoset olisi katsottu samaan luokkaan kuuluviksi, jolloin siis pienimmässä luokassa olisi ollut 3 tapausta (6 %), olisi seuraavan luokan prosenttiosuus joko noussut 28:aan (1+7+6=14 tapausta) tai jäänyt 16:een (I+ 7=8 tapausta) jne. Normalisoituja arvoja käsitellään kuten muitakin luokiteltuja arvoja; tässä tapauksessa siis ykköset, kakkoset ja kolmoset katsotaan samanarvoisiksi, neloset ja viitoset samoin jne. Alkuperäinen mittaustarkkuus karkeistuu ja informaatiota siten menetetään. Samoin on syytä huomata, että tavallisesti alkuperäisiä luokkia yhdistellään siten, että yhdistettävien raakapisteluokkien lukumäärä vaihtelee. Kyseessä on epälineaarinen muunnos, jolla ei ole selvää matemaattista muotoa.

8. Korrelaatio

Korrelaation käsite on käyttäytymistieteissä erittäin keskeinen. Korrelaatio sisältyy käsitteenä tai tilastollisena menetelmänä valtaosaan käyttäytymistieteellistä tutkimusta, joten tekniikan hallitseminen on välttämätöntä tutkimuksiin tutustuttaessa. Samoin se muodostaa lähtökohdan suurelle osalle kehittyneempiä tilastollisia kuvauskeinoja, ennen kaikkea monimuuttujamenetelmille, joilla analysoidaan usean muuttujan suhteita yhtaikaa. Näistä syistä korrelaatiota pyritään

tässä esityksessä käsittelemään suhteellisen laajasti. Korrelaation käsitettä selventänevät seuraavat esimerkit (jotka ovat täysin keksittyjä eivätkä siis kuvaa ko. variaabelien todellisia suhteita).

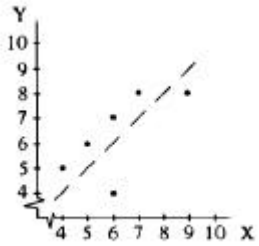
Kuvitellaan, että tutkijaa kiinnostaa englannin ja ruotsin kielen suhde koulussa menestymisen kannalta, ts. se voidaanko toisessa menestymisestä päätellä jotakin siitä, miten sama henkilö onnistuu toisessa. Tätä varten tutkija hankkii otoksen ja kultakin otoksessa olevalta henkilöltä todistuksen numeron sekä englannissa että ruotsissa. Käsittelemme tässä hyvin pieniä otoksia, jotta niissä olevat suhteet olisi helppo nähdä; todellisuudessa otosten tulisi olla huomattavasti suurempia. Otokseen tulleella henkilöllä A on englannissa 9 ja ruotsissa 8, B:llä on englannissa 6 ja ruotsissa 4 jne. Näistä arvoista voimme tehdä seuraavan taulukon:

koehenkilö	englannin arvosana (=X)	ruotsin arvosana (=Y)
A	9	8
B	6	4
C	5	6
D	4	5
E	7	8
F	6	7

Lukuja tutkimalla huomaa helposti, että englannin ja ruotsin numeroiden välillä vallitsee riippuvuussuhde: jos toisessa on hyvä numero, on toisessakin, vaikkakaan ei välttämättä sama numero. Samoin pyrkivät huonot numerot keskittymään samoille henkilöille. Sanomme, että tämän otoksen perusteella ruotsin ja englannin kouluarvosanat korreloivat positiivisesti, niiden välillä vallitsee positiivinen korrelaatio. Korrelaatio on siis riippuvuussuhde kahden variaabelin välillä; riippuvuuden ei tarvitse olla täydellistä, vähäinen taipumuskin riippuvuuteen on korrelointia.

Korrelaation voimakkuus ilmaistaan korrelaatiokertoimen avulla. Kerroin on siis kahden variaabelin välisen yhteyden voimakkuuden (ja suunnan, kuten seuraavassa näemme) mitta. Kerroin on konstruoitu siten, että se voi saada arvoja vain väliltä -1 - + 1. Itseisarvoltaan ykkösen arvoiset korrelaatiot ilmaisevat täydellistä yhteyttä, nolla taas täydellistä yhteyden puutetta.

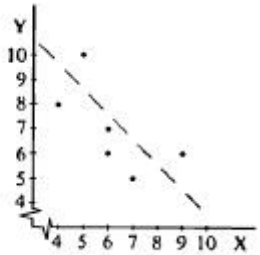
Äskeisessä esimerkissämme oli englannin ja ruotsin arvosanojen välinen korrelaatio + .69, joka on suhteellisen voimakas yhteys. Positiivisista kertoimista jätetään yleensä +-merkki pois, joten kerroin on siis .69. Korrelaatio voidaan havainnollistaa myös graafisesti korrelaatiotauluna. Suorakulmaiseen koordinaatistoon piirretään jokaista lukuparia kuvaava piste siten, että se on niin paljon vasemmalla tai oikealla kuin X-arvo osoittaa, ja niin ylhäällä tai alhaalla kuin Y-akselin vastaava arvo. Edellisestä esimerkistä saamme seuraavan korrelaatiotaulun, jossa henkilö A:ta kuvaava piste X=9, Y=8 on ylhäällä oikealla jne.:



Pisteet sijaitsevat siten, että niiden joukkoa voidaan suhteellisen hyvin kuvata oikealle kallellaan olevalla suoralla. jos suora on oikealle kallellaan, on korrelaatio positiivinen, jos vasemmalle, korrelaatio on negatiivinen. Korrelaatio on sitä voimakkaampi mitä lähempänä pisteet ovat suoraa.

Seuraavaksi kuvitellaan, että samat henkilöt ovat jälleen otoksena, mutta nyt on ruotsin numeron sijalla liikunnan numero. X on siis englanti ja Y voimistelu. Saamme seuraavat tulokset taulukkona ja graafisena esityksenä:

kh	X	Y
A	9	6
B	6	7
C	5	10
D	4	8
E	7	5
F	6	6

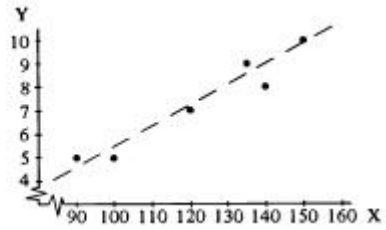


Nyt on yhteys kääntynyt toisinpäin: suuri numero toisella variaabelilla pyrkii saamaan parikseen pienen toisella, ts. jos on hyvä englannissa, on luultavasti huono liikunnassa ja päinvastoin. Yhteyden voimakkuus on suunnilleen sama

kuin äskeisessä esimerkissä, mutta sen suunta on muuttunut: vallitsee negatiivinen korrelaatio, tässä tapauksessa on kertoimen suuruus -.65. Graafisessa kuvauksessa ovat pisteet suunnilleen yhtä etäällä suorasta kuin äsken, mutta suora on kallistunut vasemmalle.

Seuraava otos on kerätty älykkyysosamäärän (X) ja matematiikan arvosanan (Y) välisen yhteyden selvittämiseksi. Ensimmäisen henkilön älykkyysosamääräksi tuli 120 ja matematiikan arvosana oli 7. Seuraavat luvut olivat 100 ja 5 jne. Saamme seuraavan taulukon:

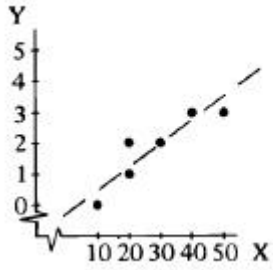
kh	X	Y
A	120	7
B	100	5
C	135	9
D	90	5
E	150	10
F	140	8



Yhteys on tässä nyt erittäin voimakas: korrelaatiokertoimen arvoksi tulee .96, lukupareja kuvaavat pisteet ovat lähes jonossa. Olennaisempaa kuin yhteyden voimakkuus on tässä kuitenkin se, että primääriluvut, älykkyysosamäärät ja matematiikan arvosanat, ovat aivan eri suuruisia, "eri laatua ". Tämä ei millään tavoin haittaa korrelaation määrittämistä. Korrelaatio on siis yhteisen vaihtelun eikä absoluuttisen samankaltaisuuden mitta. Yhteinen mittakaava saadaan Z-pisteiden avulla.

Seuraavaksi esitämme tilanteen, jossa tutkija piloillaan laski joidenkin kylien haikaranpesien ja kylissä asuvien pikkulasten määrän välisen korrelaation. Ensimmäisessä kylässä oli 50 lasta ja kolme pesää, toisessa 10 lasta eikä yhtään pesää jne. Saatiin seuraavat tulokset:

kylä	lapset X	pesät Y
A	50	3
B	10	0
C	40	3
D	20	1
E	30	2
F	20	2



Tutkijan hämmästykseksi muuttujien välillä vallitsee voimakas yhteys .91. Tästä esimerkistä opimme kaksi asiaa. Ensinnäkin sen, ettei korrelaation olemassaolo kerro mitään syy-yhteydestä variaabelien välillä. Korrelaatiosta näkee vain tilastollisen yhteyden lukuparisarjan sisällä, syyn ja seurauksen joutuu päättelemään loogisin, ei matemaattisin keinoin. Tässä tapauksessa voisi voimakas yhteys johtua vaikkapa siitä, että lapsiperheissä pidetään haikaroista ja tarjotaan niille pesimäpaikkoja, ruokaa tms. Toinen tärkeä asia on se, että tilastollisena yksikkönä, jota siis on kuusi (N=6), on kylä eikä sen siis tarvitse olla henkilö, kuten usein on laita. Onkin asiallisempaa puhua yksiköistä (tilastoyksikkö, havaintoyksikkö) kuin yksilöistä tilastollisen kuvaamisen yhteydessä. Nyt esillä oleva yksikkö on lisäksi ns. aggregoitu yksikkö. Sellaisia ovat tilastoyksiköt, joiden tunnusluvut saadaan keskiarvoina niihin kuuluvista yksilöistä (kunnat, koulut, luokat yms., puhutaan myös ekologisesta yksiköstä). Tällaisiin korrelaatioihin liittyy omia erityisiä ongelmiaan.

Jos asiasta ei erikseen mainita, on korrelaatiokerroin Pearsonin tulomomenttikerroin, jonka symboli on r. Siitä on hyvinkin monenlaisia laskukaavoja, mutta tutustumme ensin sen periaatteelliseen kaavamuotoon Z-pisteinä.

$$r = \frac{\sum Z_x Z_y}{N - 1}$$

Palautamme mieleen, että Z-pistemääräksi muunnettu muuttuja saa keskiarvo 0 ja hajonnan 1. Kun poikkeamat keskiarvosta kerrotaan keskenään ja lasketaan niiden keskiarvo (vapausasteita käyttäen) niin päädytään siihen, mitä nimitetään tulomomenttikertoimeksi. Siitä ilmenee missä määrin poikkeamat menevät samaan suuntaan. Yksittäinen havainto ei sitä määritä kuin pieneltä osin. Se on

myös aina otokseen sidoksissa oleva arvo. Otoksen keskiarvoja ja hajontoja käytetään Z-pistemääriksi muuntamiseen. Kun Z-pisteistä mennään raakapisteisiin on korrelaatiokertoimen laskukaava seuraava:

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]\ [N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Kaava on pahemman näköinen kuin se loppujen lopuksi on, kunhan sen osaa hahmottaa osiinsa. Tämä hahmottaminen tapahtunee parhaiten esimerkin avulla. Oletetaan, että kymmenen henkilön otos on hankittu koulusaavutusten ja introversion suhteen tutkimista varten. Molempia on mitattu testeillä; ensimmäinen koehenkilö on saanut koulusaavutustestissä pistemääräksi 7ja introversiopistemääräksi 3 jne.:

koehenkilö	koulusaavutus-testi (X)	introversio-testi (Y)
1	7	3
2	8	4
3	9	3
4	5	2
5	7	2
6	8	4
7	6	3
8	9	4
9	7	2
10	6	3

Laskutoimitus ja sitä varten tehty taulukko ovat seuraavanlaiset (laske annetusta datasta keskiarvo ja hajonta, muunna kukin raakapiste Z-pistemääräksi, laske niiden tulo, tulojen summa ja jaa se N-1:llä harjoituksena):

X	Y	XY	X^2	Y^2
7	3	21	49	9
8	4	32	64	16
9	3	27	81	9
5	2	10	25	4
7	2	14	49	4
8	4	32	64	16
6	3	18	36	9
9	4	36	81	16
7	2	14	49	4
6	3	18	36	9
72	30	222	534	96
ΣX	ΣY	ΣXY	ΣX^2	ΣY^2

$$r = \frac{10 \cdot 222 - 72 \cdot 30}{\sqrt{(10 \cdot 534 - 72^2)(10 \cdot 96 - 30^2)}}$$

$$r = \frac{2220 - 2160}{\sqrt{(5340 - 5184)(960 - 900)}}$$

$$r = \frac{60}{\sqrt{156 \cdot 60}} = \frac{60}{\sqrt{9360}} = \frac{60}{97} = .62$$

Numerus on lukuparien, siis tässä tapauksessa koehenkilöiden määrä (10). Sitä tarvitaan osoittajan alussa ja molemmissa sulkeissa neliöjuuren alla. SummaXY on keskenään kerrottujen lukuparien summa. Siis 7*3=21 , johon lisätään 8*4=32 jne., kunnes koko summaksi saadaan 222. SummaXSummaY taas tarkoittaa X-arvojen summaa ja Y-arvojen summaa kerrottuna keskenään. Laskemme siis yhteen kaikki X-arvot (72) ja kaikki Y -arvot (30) ja kerromme ne keskenään. Neliöjuuren alla taas on kummallekin variaabelille lauseke, joka on tuttu neliösumman (hajonnan yhteydessä) kaavasta. SummaX ja summaY meillä jo on (72 ja 30). Ne korotetaan toiseen. Numeruksen jäljessä olevat kerrottavat saamme korottamalla kummankin variaabelin arvot toiseen ja laskemalla ne yhteen (534 ja 96). Kertoimeksi saimme .62, jonka mukaan introvertit henkilöt menestyvät koulussa paremmin. Sama lopputulos pitäisi tulla Z-pisteiden avulla.

Seuraavaksi käsittelemme esimerkkiä, jossa kerroin saa negatiivisen arvon. Kerätään kuudelta oppitunnilta opettajan positiivisten reaktioiden määrä (X) ja oppilaiden häiritsevien toimien määrä (Y). Esimerkki on laskettu samoin kuin edellinenkin, kertoimeksi tulee tällä kertaa -.67, jonka mukaan opettajan positiiviset reaktiot ja oppilaiden häirintä pyrkivät odotetusti olemaan kääntäen verrannollisia: kun toista on paljon, on toista vähän.

X	Y	XY	X^2	Y^2
6	5	30	36	25
15	4	60	225	16
7	5	35	49	25
2	6	12	4	36
11	6	66	121	36
8	5	40	64	25
4	6	24	16	36
53	37	267	515	199

$$r = \frac{7 \cdot 267 - 53 \cdot 37}{\sqrt{(7 \cdot 515 - 53^2)(7 \cdot 199 - 37^2)}}$$

$$r = \frac{1869 - 1961}{\sqrt{19104}}$$

$$r = \frac{-92}{138} \qquad r = -.67$$

Jos korrelaatiota laskettaessa lähtöarvot ovat järjestyslukuja tai ne muutetaan sellaisiksi, voidaan laskussa käyttää seuraavaa kaavaa. Järjestyskorrelaatio on aivan tavallinen tulomomenttikorrelaatio. Lukuparit ovat vain järjestyssijoja. Se voitaisiin laskea edelliseen tapaan. Seuraavalla kaavalla laskeminen on kuitenkin yksinkertaisempaa.

$$r_j = 1 - \frac{6 \cdot \Sigma d^2}{N\ (N^2 - 1)}$$

Kaavaa nimitetään Spearmanin järjestyskorrelaatioksi kaavan johdon suorittaneen psykometriikan klassikkonimen mukaan. Tätä kaavaa käytettäessä ei muuttujissa saisi olla sidoksia (engl. tied ranks).

Esimerkkinä järjestyskorrelaatiosta oletamme, että seitsemän koehenkilön älykkyys on sekä testattu että pantu arvioimalla järjestykseen. Ensimmäisen henkilön testitulos oli 95 ja hänet arvioitiin kolmanneksi älykkäimmäksi, toisen pistemäärä oli 110 ja arvioitu sijaluku oli 6 jne. Saamme seuraavan taulukon ja laskutoimituksen:

kh	1	2	3	4	5	6	7
X = ÄO	95	110	120	100	110	90	150
Y = arvioitu järjestys	3	6	1	4	5	7	2

kh	X	X_j	Y_j	d	d^2
1	95	6	3	3	9
2	110	3.5	6	2.5	6.25
3	120	2	1	1	1
4	100	5	4	1	1
5	110	3.5	5	1.5	2.25
6	90	7	7	0	0
7	150	1	2	1	1
					$20.5 = \Sigma d^2$

$$r_j = 1 - \frac{6 \cdot 20.5}{7(7^2 - 1)} = 1 - 0.37 = .63$$

Koska kaava vaatii lähtöluvuikseen järjestyslukuja, täytyy älykkyyspisteet ensin muuttaa järjestysluvuiksi (Xj): 150 on ensimmäinen, 120 toinen jne. 110 esiintyy kaksi kertaa, joten sijaluvut 3 ja 4 jaetaan niin, että molemmat saavat sijan

3.5 ja seuraavan arvo on 5. Ainoa uusi symboli kaavassa on d, joka tarkoittaa järjestyslukujen erotusta kussakin parissa. Ensimmäiseltä koehenkilöltä saamme 6-3=3, toiselta 3.5-6=2.5 (etumerkillä ei ole tässä merkitystä) jne. Erotukset korotetaan toiseen potenssiin ja lasketaan yhteen. Numerus on 7 ja kuutonen on vakio, joka kuuluu kaavaan. Korrelaatioksi tulee .63, joten arviointi ja testi täsmäsivät suhteellisen hyvin.

Kuten sanottu, saadaan tulomomenttikertoimella täsmälleen sama tulos. Voimme osoittaa tämän laskemalla edellisen esimerkin uudelleen käyttäen tulomomenttikertoimen kaavaa:

X	Y	XY	X^2	Y^2
6	3	18	36	9
3.5	6	21	12.25	36
2	1	2	4	1
5	4	20	25	16
3.5	5	17.5	12.25	25
7	7	49	49	49
1	2	2	1	4
28	28	129.5	139.5	140

$$r = \frac{7 \cdot 129.5 - 28 \cdot 28}{\sqrt{(7 \cdot 139.5 - 28^2)(7 \cdot 140 - 28^2)}}$$

$$r = \frac{122.5}{\sqrt{192.5 \cdot 196}}$$

$$r = \frac{122.5}{194.24} = .6306$$

Jos sen sijaan tulomomenttikerroin lasketaan suoraan esim. testipisteistä, kouluarvosanoista tms. ja luvut muutetaan järjestysluvuiksi järjestyskorrelaation laskemiseksi, saamme vain likimääräisesti saman tuloksen. Järjestyskorrelaatiota voidaan tällöin pitää tulomomenttikertoimen helpommin laskettavissa olevana arviona (joskus käytetään termiä estimaatti).

Sellaisenaan korrelaatiokerroin ilmaisee yhteyden kahden variaabelin välillä. Yhteys usean muuttujan kesken voidaan esittää kokoamalla korrelaatiot luku-

tauluksi (matriisi, korrelaatiomatriisi). Aineistostamme on laskettu korrelaatiomatriisi. Koska korrelaatio on yhteyden ilmaisimena symmetrinen, ts. X:n korrelaatio Y:hyn on sama kuin Y:n korrelaatio X:ään, tarvitaan kunkin muuttujaparin välille vain yksi korrelaatio ja matriisista voi olla kolmion muotoinen, kun turhat luvut ovat poissa. Alussa esitetyn aineiston korrelaatiomatriisi on seuraava:

	1	2	3	4	5	6	7	8	9
1									
2	.259								
3	−.168	.143							
4	−.185	.303	.291						
5	.277	.271	.467	.214					
6	−.136	.219	.297	.424	.301				
7	−.192	.369	.439	.583	.347	.604			
8	−.297	.292	.294	.480	.284	.630	.706		
9	.392	.446	.187	.243	.412	.338	.339	.181	
10	−.101	.207	.251	.214	.373	.109	.486	.374	.060

Matriisin muuttujat:

1 sukupuoli
2 yhteenlaskut
3 havaintonopeus
4 vastakohdat
5 neliötäydennys
6 kuuseentäydennys
7 looginen järjestys
8 sanaryhmät
9 matemaattiset tehtävät
10 peilitesti

Matriisista näemme esim., että sukupuolella on voimakkain yhteys matemaattisiin tehtäviin (.392), korkein korrelaatio vallitsee loogisen järjestyksen ja sanaryhmien testien välillä (.706), matalin taas matemaattisten tehtävien ja peilitestin välillä (.060) jne.

On syytä huomata, että korrelaatiot on laskettu myös sukupuolen ja muiden muuttujien välille, vaikkei ensin mainittu ole lainkaan kvantitatiivinen variaabeli. Näin voi kuitenkin tehdä, kunhan osaa tulkita tulokset oikein. Laadullisen muuttujan on kuitenkin oltava kaksiarvoinen (dikotominen), jolloin sen arvot voidaan eräässä mielessä tulkita kvantitatiivisiksi (esim. 0 ja 1). Dikotomisen ja useampiluokkaisen kvantitatiivisen muuttujan tulomomenttikerrointa nimitetään usein pistebiseriaaliseksi korrelaatioksi. Voimme ajatella, että muuttuja ei ole "sukupuoli" vaan vaikkapa olla "poika". Poikien koodi 1 on olla "poika" ja koodi 0 "ei-poika" käy laskemiseen. Asia on yleisempikin: menettelyä nimitetään dummy-

miseen. Asia on yleisempikin: menettelyä nimitetään dummy-muuttujaksi. Joku laadullinen asia voidaan purkaa esiin käyttämällä useita koodattuja muuttujia ja luoda niillä haluttuja vertailuja (mutta asia kuuluu sinällään vasta myöhempiin opintoihin).

Korrelaatiomatriisin ensimmäisen sarakkeen lukuarvoista näemme siis sukupuolen yhteyden muihin muuttujiin. Koska pojilla on sukupuolimuuttujalla suurempi koodina käytetty arvo, he ovat parempia niillä variaabeleilla, joiden korrelaatio on positiivinen ja päinvastoin. Tytöt ovat tässä otoksessa parempia kaikissa verbaalisissa tehtävissä, pojat taas matemaattisissa. Olemme aikaisemmin ristiintaulukoinnin ja keskiarvojen avulla osoittaneet, että pojat ovat hieman parempia neliötäydennystestissä. Nyt meillä on kolmas tapa esittää sama tämä asia hieman eri muodossa: sukupuolen ja neliötäydennystestin korrelaatio on positiivinen, .277. Se tarkoittaa: poikien keskiarvo on suurempi kuin tyttöjen keskiarvo. Mikähän olisi ollut tulos, jos olisimmekin koodanneet tytöt kakkosella ja pojat ykkösellä?

Mikä esitystapa kulloinkin on sopivin, riippuu tilanteesta. Jos esim. haluaa esittää, kuinka yhteys muodostuu eri kykytasoilla, on ristiintaulukointi havainnollinen. Jos taas haluaa vain verrata keskimääräisiä suorituksia ja mahdollisesti testata tilastollisesti, onko ero sattumalta saatavaa suurempi (merkitsevyyden testaus), ovat keskiarvot käyttökelpoisia havainnollisuutensa vuoksi. Jos haluaa analysoida muuttujien välisiä suhteita edelleen ja tutkia useiden muuttujien suhteita vaikkapa ns. monimuuttujamenetelmillä, ovat korrelaatiot usein lähtöarvoja. Raportointi saattaa edellyttää monessa tapauksessa sekä ristiintaulukointia että keskiarvoja.

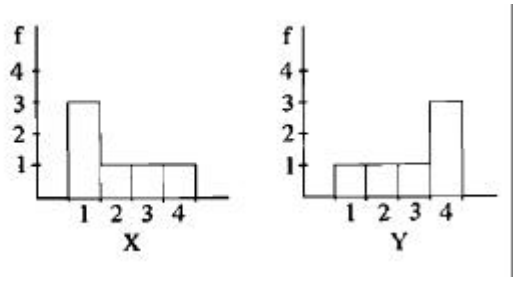
Kun ristiintaulukoinnin yhteydessä esitettiin loogisen järjestyksen ja sanaryhmien testin välinen taulukko, oli mukaan harkitusti valittu tämän muuttujajoukon voimakkaimmin korreloiva pari. Tällöin huomattiin, että tapaukset pyrkivät keskittymään taulukon lävistäjälle, diagonaalille, mikä on itse asiassa aivan sama asia kuin se, että voimakkaan korrelaation graafisessa esityksessä tapaukset asettuvat lähelle suoraa. Ristiintaulukointi oli siis kahden epäjatkuvan muuttujan välinen korrelaatiotaulu.

On joitakin tapauksia, joissa korrelaation maksimiarvo on rajoittunut, jolloin se ei voi saavuttaa arvoa 1 edes suurimmillaan, vaikka yhteys annetuissa oloissa on maksimaalinen. Eräillä muillakin tekijöillä on vaikutusta siihen, millaiseksi korrelaatio voi muodostua. Seuraavassa luvussa näistä tärkeimpiä.

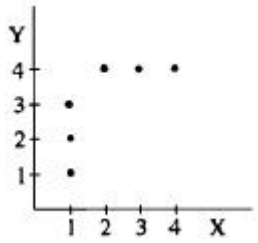
1) käyräviivainen yhteys

Tutkitaanpa seuraavaa taulukkoa:

X	Y
1	1
1	2
1	3
2	4
3	4
4	4



Molemmilla muuttujilla on samoja lukuja, mutta X:llä on useita ykkösiä ja Y:llä taas useita nelosia. X-muuttuja on positiivisesti vino ja Y negatiivisesti. Näissä puitteissa on luvut järjestetty niin suuren korrelaation aikaansaamiseksi kuin mahdollista, suuret luvut ovat suurten pareina ja pienet pienten niin pitkälle kuin niitä variaabelilla riittää. Korrelaatio jää kuitenkin arvoon .75. Sama voidaan esittää myös graafisesti, jolloin ilmiön syy on helpompi selvittää:



Pisteet eivät pyrikään asettumaan lähelle suoraa, vaan niiden yhteyttä voitaisiin kuvata parhaiten käyrän avulla. Sanomme, että muuttujien välillä ei vallitsekaan lineaarinen yhteys vaan käyräviivainen, nonlineaarinen.

Korrelaatio on tarkoitettu lineaaristen yhteyksien kuvaamiseen eikä toimi kunnolla muissa tapauksissa. Tämä näkyy siinä, että kertoimen arvo laskee. Ääritapauksissa saattaa korrelaation arvo olla nolla, vaikka nonlineaarista yhteyttä selvästi on, ts. muuttujalta toiselle voidaan tehdä päätelmiä, mutta ei niin, että tasainen kasvu toisella muuttujalla vastaisi tasaista kasvua myös toisella.

2) jakaumien samanmuotoisuus

Tapauksessa, jossa toinen muuttujista on jatkuva ja toinen dikotominen, ei korrelaatio voi koskaan olla ykkösen suuruinen. Jos molemmat muuttujat ovat dikotomisia, riippuu korrelaation maksimiarvo siitä, kuinka samankaltaisesti nämä kaksi arvoa ovat muuttujilla jakaantuneet. Esimerkkeinä ovat seuraavat tapaukset, joissa on kussakin maksimi korrelaatio niiden jakaumien puitteissa, jotka variaabeleilla on:

a)

X	Y
1	1
1	1
1	0
1	0
0	0
0	0

r = .50

b)

X	Y
1	1
1	1
1	1
1	0
0	0
0	0

r = .71

c)

X	Y
1	1
1	1
1	1
0	0
0	0
0	0

r = 1

d)

X	Y
1	1
1	1
0	0
0	0
0	0
0	0

r = 1

Jos siis kumpaakin arvoa on toisella variaabelilla yhtä monta kuin toisellakin, voi korrelaatio olla ykkösen suuruinen. Mitä erilaisemmat määrät suurempaa ja pienempää arvoa (tässä: 1 ja 0) variaabeleilla on, sitä kauempana ykkösestä on korrelaation maksimiarvo.

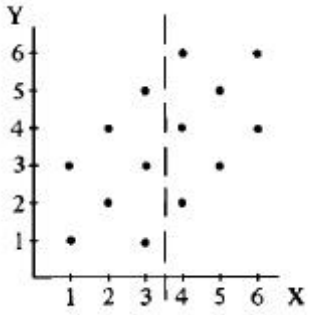
Jakaumien muodon samankaltaisuus pätee myös useampiarvoisten kvantitatiivisten muuttujien tapauksessa. Jos toinen muuttuja on vino ja toinen symmetrinen, korrelaatio ei voi saada maksimiarvonaan ykköstä. Sama koskee huipukkuuden eroja. Itsensä kanssa muuttuja korreloi arvolla 1, jos vain muuttujalla on vaihtelua. Vakiota ei pidetä muuttujana.

3) koehenkilöjoukon valikoituminen

Jos toisen tai molempien variaabeleiden varianssia keinotekoisesti pienennetään, esim. valitsemalla koehenkilöiksi jokin suhteellisen samankaltaisten yksilöiden joukko, pienenee korrelaation arvo. Jos vaikkapa tutkisimme älykkyyden korrelaatiota johonkin muuhun ominaisuuteen, saisimme opiskelijoista kootussa otoksessa todennäköisesti pienemmän yhteyden kuin koko väestöä edustavassa otoksessa. Tämä siksi, että opiskelijat ovat älykkyyden suhteen homogeenisempia, samankaltaisempia, kuin väestö keskimäärin.

Erittäin selvänä tämä ilmiö tulee näkyviin, kun tutkimme esim. jonkin oppilaitoksen pääsykokeen ja opintomenestyksen välistä yhteyttä. Jos karsintaa on,

saamme opintomenestyksen mittausarvon vain sisään päässeiltä, niiltä joilla pääsykoemenestys on ollut varsin samanlainen ja opintomenestystäkin on siis valinnan kautta homogenisoitu hyvin paljon. Korrelaatio voi jäädä aivan olemattomaksi, vaikka pääsykoe olisi hyväkin: meillä ei vain ole tutkittavana kokeeseen tulleiden koko varianssia, vaan voimakkaasti samanlainen sisään päässeiden joukko. Tätä voisi havainnollistaa seuraavalla keksityllä esimerkillä:



Kuvitellaan, että Y on pääsykoemenestys ja X menestys opinnoissa. Koko pistejoukko kuvaa näiden variaabeleiden suhdetta sellaisessa tilanteessa, jossa kaikki pyrkijät otetaan. Tässä vallitsee .55 :n suuruinen korrelaatio. Jos kuitenkin oppilaitos karsii puolet pyrkijöistä, emme voi hankkia tätä korrelaatiota, vaan sen, joka muodostuu, kun pilkkuviivan oikealla puolella olevat tapaukset otetaan huomioon. Tästä joukosta saadaan vain .28:n suuruinen korrelaatio. (Voit tarkistaa nämä arvot laskemalla.)

Edellä sanottu koskee variaabelin luonnollisen vaihtelualueen rajoittamista eikä siis sellaisia tilanteita, joissa sama vaihtelu vain ilmaistaan eri luvuin. Jos toisen tai molempien variaabeleiden varianssi muuttuu siksi, että lukuihin lisätään tai niistä vähennetään vakio tai luvut kerrotaan tai jaetaan vakiolla, ei korrelaatio muutu. Näin esim. jos pituus ilmaistaan metrien sijasta sentteinä tms.

Testeillä saattaa olla ns. katto- tai lattiavaikutuksia, jotka vaikuttavat siihen, että osa variaatiosta häviää. On myös monia muita syitä siihen, että tutkimusjoukko valikoituu tai samankaltaistuu.

Koska korrelaatio on yhteisen vaihtelun mitta, ei variaabeli, jossa ei ole vaihtelua, voi korreloida toiseen variaabeliin.

Aikaisemmin varianssin yhteydessä esitettiin, että sen voi jakaa osiin sen mukaan, kuinka suuri osuus eri tekijöillä on varianssin muodostumiseen. Korrelaatiokertoimen neliö ilmaisee, kuinka suuri osuus variaabeleiden varianssista on yhteistä, selitettävissä toisen variaabelin avulla. Niinpä .20:n korrelaatio kertoo, että $.20^2$=.04 (=4 % varianssista) on yhteistä (= tulee selitetyksi suuntaan tai toiseen), .30:n korrelaatio merkitsee 9 %:n yhteistä osuutta, .80 vastaa 64 %:a ja .90:n korrelaatio kertoo jo 81 %:n yhteisestä osuudesta. Näistä luvuista huomaamme samalla sen, ettei korrelaation "selittävyys" tai "tärkeys" kasva suorassa suhteessa kertoimen kanssa: lähellä ykköstä antaa samansuuruinen nousu enemmän selitystä kuin lähellä nollaa. Olihan selitysprosentin nousu .20:stä .30:een vain 5 %, kun taas nousu .80:sta .90:een nosti selitysosuutta 17 %.

Varianssin katoaminen valinnan tai valikoitumisen johdosta on tärkeä tekijä, joka vaikuttaa otoksesta laskettavan korrelaation suuruuteen.

4) puuttuvan tiedon korvaaminen keskiarvolla

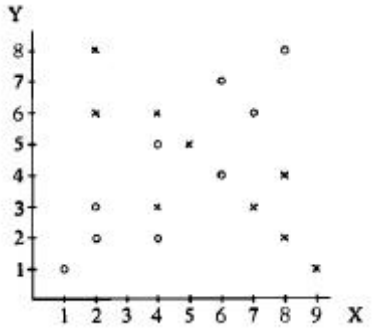
Lukupari, jossa toinen tai molemmat arvot edustavat muuttujan keskiarvoa tuottaa Z-pisteiden tulon 0-arvon. Lukupari, josta toinen puuttuu pitäisi jättää korrelaatiolaskennassa käyttämättä. Tällainen puuttuvan tiedon käsittely saattaa johtaa monen muuttujan yhteisessä tarkastelussa huomattavaan tapausten katoon (ns. listatyyppinen käsittely, kun yksikin puuttuu niin kaikki pois). Mikäli puuttuva tieto korvataan muuttujan keskiarvolla, syntyvä korrelaatio ei muutu. Tämä onkin yksinkertainen ja paljon käytetty keino puuttuvan tiedon käsittelyssä. Puuttuvan tiedon ja muun tiedon hankkimisen kadon kysymystä on tutkimuksessa aina käsiteltävä: onko se systemaattista? Kuinka kato vaikuttaa saatuihin tuloksiin?

5) poikkeava lukupari

Yksikin XY-pari, jossa toinen tai molemmat luvut ovat hyvin selvästi poikkeavia arvoja muiden XY-pisteiden parvesta, vaikuttaa voimakkaasti syntyvään korrelaatiokertoimeen. Korrelaatiota on syytä katsoa myös graafisena esityksenä, jossa tällaiset poikkeavat arvot näkyvät havainnollisesti.

6) epähomogeeniset osaryhmät

Kun aineisto muodostuu osaryhmistä, voi korrelaatiokerroin muodostua muuttujien välille erilaisista kombinaatioista. Ryhmien sisällä voi olla erisuuruinen korrelaatio, jonka yhdistäminen peittää alleen. Ryhmien väliset keskiarvo- ja hajontaerot luonnollisesti vaikuttavat myös lopputulokseen. Seuraavassa on koko pistejoukosta laskettu korrelaatio on .02, mutta ristien joukosta laskettuna se on -.85 ja ympyröistä laskettuna .88:



Pearsonin tulomomenttikertoimesta annetaan usein seuraavat vaatimukset: 1) muuttujien tulee olla määrällisiä, tasavälisellä asteikolla ilmaistuja mittalukuja, 2) otoksen tulee olla umpimähkäinen otos perusjoukosta, jossa muuttujilla on normaali jakautuminen, 3) regressioresiduaalien tulee olla normaalisti jakautuneita ja 4) havaintoyksiköiden tulee olla riippumattomia toisistaan (autokorrelaatiota ei saa olla, esim. aikasarjan tapainen riippuvuus).

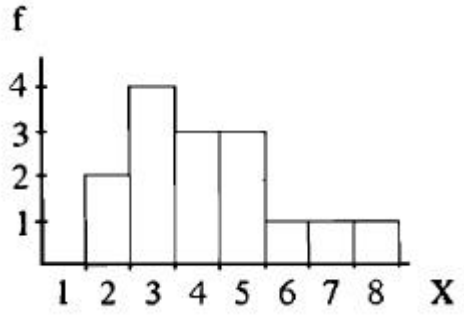
Osa näistä vaatimuksista koskee vain sitä tilannetta, jossa tehdään päätelmiä poikkeaako perusjoukon korrelaatio nollakorrelaatiosta (tilastollinen hypoteesin testaus). Kuten edellä on todettu, poikkeavia käyttötilanteita tästä ihanteesta on hyvin usein. Korrelaatiokertoimen muodostumisen periaate ja sen suuruuteen vaikuttavat tekijät on syytä tuntea perusteellisesti. Korrelaatiokertoimesta ei voida mekanistisesti tehdä tutkimuspäätelmiä ilman vaikuttavien tekijöiden pohtimista.

Luettelosta puuttuu vaikuttavia käyttötilanteita. Tällainen on ns. tekninen korrelaatio. Tällöin X-muuttuja ja Y-muuttuja sisältävät (usein käyttäjän huomaamatta) samaa ainesta. Tyypillinen sovellustilanne on osion korrelaatio osioista muodostuvan summan kanssa. Yksittäinen osio sisältyy myös summaan (tai mahdolliseen keskiarvoon). Testitarkastelussa lasketaan tämän vuoksi ns. puhdistettu osiokorre-

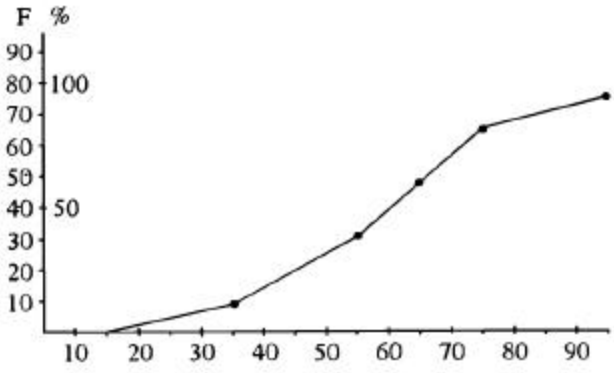
laatio (osion korrelaatio muiden osioiden summaan/keskiarvoon). Muitakin puutteita on.

9. Harjoitustehtäviä

1) Tutkimuksessa saatiin oheinen jakauma. a) Mikä oli N? b) Mikä oli moodi? c) Mikä oli mediaani? d) Mihin suuntaan jakauma on vino?



2) Eräässä tutkimuksessa saatiin oheinen kumulatiivinen frekvenssikäyrä. a) Mikä oli numerus? b) Mikä oli mediaani? c) Kuinka moni suunnilleen on saanut pistemäärän 60 tai alle? d) Kuinka monta prosenttia suunnilleen on jäänyt pistemäärän 70 alapuolelle?



3) Luokittele primäärimatriisin lauseentäydennystesti luokkiin 3-5, 6-8, 9-11, 12-14, 15-17, 18-20. Taulukoi näin saadut luokat ristiin sukupuolen kanssa.

4) Laske edellisen tehtävän luokitellusta lauseentäydennystestistä tyttöjen mediaani ja aritmeettinen keskiarvo.

5) Määritä seuraavan lukusarjan a) vaihteluväli b) keskimääräinen poikkeama c) keskihajonta: 1, 2, 3, 3, 2, 5, 4, 4, 5, 1.

6) Valitse jokin primäärimatriisin variaabeli ja laske sille keskiarvo ja standardipoikkeama.

7) Normalisoi primäärimatriisissa olevan peilitestin jakauma kuuteen luokkaan pyrkien jakaumaan 2 %, 14 %, 34 %, 34 %, 14 %, 2 %.

8) A-variaabelilla (keskiarvo = 6 ja hajonta = 1.5) saatiin pistemäärä 4 ja B-variaabelilla (ka = 70 ja haj = 12) pistemäärä 55. Kummalla variaabelilla on pistemäärä suhteellisesti parempi?

9) Arvioi, minkälainen korrelaatio seuraavissa esimerkeissä vallitsee. Tarkista laskemalla, jos on tarpeen.

a)

X	Y
5	7
1	3
6	7
2	3
5	6

b)

X	Y
3	120
3	0.45
3	10
3	18
3	562

c)

X	Y
1	10
2	5
1	10
2	5
1	10
2	5

d)

X	Y
2	2
5	1
2	2
6	1
2	2
8	1

10) Eräässä karsintatilaisuudessa järjestettiin testi sekä arvioitiin pyrkijöiden paremmuusjärjestys. Saatiin seuraavat tulokset :

pyrkijä	1	2	3	4	5
arvioitu sija	3	1	2	4	5
testitulos	19	22	25	18	20

Minkälainen yhteys vallitsee arvion ja testituloksen välillä? Laske järjestyskorrelaatio.

56

Tehtävien vastaukset

1) a) 15 b) 3 c) 4 d) positiiviseen

2) a) 80 b) noin 63 c) noin 35 d) noin 70%

3)

lauseentäydennys

Sukupuoli	3–5	6–8	9–11	12–14	15–17	18–20	
	2	3	8	8	6	1	28
	1	7	4	7	3	0	22
	3	10	12	15	9	1	50

4)

X	f	F
3–5	2	2
6–8	3	5
9–11	8	13
12–14	8	21
15–17	6	27
18–20	1	28

$$Md = 11.5 + \frac{\frac{28}{2} - 13}{8} \cdot 3$$

$$Md = 11.5 + \frac{1}{8} \cdot 3 = 11.88$$

X	lk	f	fX
3–5	4	2	8
6–8	7	3	21
9–11	10	8	80
12–14	13	8	104
15–17	16	6	96
18–20	19	1	19
		N = 28	328

$$\bar{X} = \frac{328}{28} = 11.71$$

5)

X	$\lvert X-\bar{X}\rvert$	$(X-\bar{X})^2$
1	2	4
2	1	1
3	0	0
3	0	0
2	1	1
5	2	4
4	1	1
4	1	1
5	2	4
1	2	4
	12	20
	$\Sigma\lvert X-\bar{X}\rvert$	$\Sigma(X-\bar{X})^2$

vaihteluväli = 5 − 1 = 4

$\bar{X} = 3$ N = 10

$KP = \frac{12}{10} = 1.2$

$s^2 = \frac{20}{10} = 2$

$s = \sqrt{2} = 1.414$

6)

mja	$\bar{X}$	s^2	s
1	.44	.25	.50
2	16.64	89.19	9.44
3	24.08	30.03	5.48
4	22.66	78.38	8.85
5	8.42	14.00	3.74
6	11.20	13.60	3.69
7	10.76	25.02	5.00
8	15.14	28.20	5.31
9	7.08	11.63	3.41
10	11.80	34.24	5.85

Edellä olevaan taulukkoon liittyy epätarkkuus. Varianssit (ja hajonnat) on laskettu neliösummista käyttämällä jakajana arvoa N arvon N-1 (eli vapausasteet) sijasta.

7)

X	f	lf	%
0	1	1	2
2	1		
3	0		
4	4	9	18
5	0		
6	4		
7	5		
8	2		
9	3	17	34
10	2		
11	5		
12	2		
13	1		
14	1		
15	2	15	30
16	5		
17	4		
18	2		
19	1		
20	2		
21	1		
22	0	7	14
23	0		
24	1		
25	0		
26	1	1	2
$\Sigma f = 50 = N$			

8) Z_A=-1.33 ja Z_B=-1.25 siis B parempi

9) a) .96 b) 0 (X:n varianssi nolla) c) -1.00 d)-.93 (huomaa, että 2 on X-muuttujalla pieni luku ja Y-muuttujalla taas suuri)

10) r=.60