

Tilastotieteen johdantokurssi

Luentorunko

Christina Gustafsson

SISÄLLYSLUETTELO

1. JOHDANTO	2
1.1. Mitä tilastotiede on?	2
1.2. Tilastotieteen historiaa.....	3
2. HAVAINTOAINEISTON HANKINNASTA JA MITTAAMISESTA	5
2.1. Havaintoaineisto, havaintomatriisi ja mittaaminen	5
2.2. Havaintoaineiston hankinnasta ja otantamenetelmistä	9
3. YKSIULOTTEINEN EMPIIRINEN JAKAUMA	12
3.1. Frekvenssijakauman peruskäsitteitä ja luokitus	12
3.2. Graafinen esitys.....	16
3.3. Yksiulotteisen jakauman tunnusluvut	21
3.3.1. Keskiluvut	21
3.3.2. Hajontaluvut	27
3.3.3. Yksiulotteisen jakauman muita tunnuslukuja.....	32
4. KAKSIULOTTEINEN EMPIIRINEN JAKAUMA	33
4.1. Ristiintaulukko ja kontingenssikerroin	33
4.2. Korrelaatiodiagrammi ja korrelaatio	38
4.3. Järjestyskorrelaatio.....	42
4.4. Regressio.....	44
5. TODENNÄKÖISYYSLASKENTAA	49
5.1. Kombinatoriikka.....	49
5.2. Todennäköisyyden määrittely.....	51
5.3. Ehdollinen todennäköisyys ja riippumattomuus.....	53
6. TEOREETTISISTA JAKAUMISTA	56
6.1. Satunnaismuuttujista	56
6.2. Diskreeteistä teoreettisista jakaumista	56
6.3. Jatkuvista teoreettisista jakaumista	58
7. TILASTOLLISESTA PÄÄTTELYSTÄ	64
7.1. Estimoinnista.....	64
7.2. Hypoteesien testauksesta	68
7.2.1. Testauksen pääpiirteet	68
7.2.2. Keskiarvotestejä	71
7.2.3. Prosenttilukutestejä.....	76
7.2.4. Riippuvuustutkimukseen liittyviä testejä	78
7.2.5. χ^2 -yhteensopivuustesti	81

1. JOHDANTO

1.1. Mitä tilastotiede on?

Tilasto on empiiristä ilmiötä kuvaava usein taulukkona esitetty numeerinen aineisto. **Tilastointi** tuottaa tällaisia eri ilmiöitä kuvaavia aineistoja. Erilaisia empiirisiä ilmiöitä kuvaavissa aineistoissa esiintyy samantyyppisiä ongelmia, joiden tutkimisessa tilastotieteestä on apua ja muodostetut tilastot ovat **tilastollisen tutkimuksen** materiaalina.

Tilastotiede on oppiaine, jonka aiheena voidaan pitää sattuman ja vaihtelun hallintaa, informaation suodattamista datasta sekä mallintamista monille eri tieteenaloilla.

Tilastotiede on empiirislouentoisten tietojen

hankinnan suunnittelua

keräämistä

järjestämistä

esittämistä

sekä

analysointia

tulkintaa

koskeva tiede.

**deskriptiivinen eli
kuvaileva tilastotiede**

**tilastollinen päättely eli
inferenssi *)**

*) Tilastollinen päättely on luonteeltaan induktiivista, jolloin osajoukkoa koskevat tulokset yleistetään koskemaan koko perusjoukkoa.

Tilastotiede on ns. **menetelmätiede**, jonka tehtävänä on kehittää menetelmiä muiden tieteiden (esim. talous-, luonnon- ja yhteiskuntatieteiden) empiirisiä ilmiöitä kuvaavien tietojen analysointia varten. Empiirinen ilmiö voi olla sellainen, johon vaikuttavat vain systemaattiset tekijät (**deterministinen ilmiö**) tai sellainen, johon systemaattisten tekijöiden lisäksi vaikuttaa myös sattuma (**satunnaisilmiö**). Sattuman käsitteellä tarkoitetaan satunnaisilmiön sitä käyttäytymisen osuutta, jota ei voida etukäteen tarkkaan ennakoida. Usein kuitenkin sattuman käyttäytyminen noudattaa omia lakejaan. Tilastotiedettä käytetään erityisesti satunnaisilmiöiden tutkimiseen.

Tilastotieteen lisäksi menetelmätieteitä ovat myös matematiikka ja tietotekniikka. Tilastotiede soveltaa menetelmiä kehittäessään matematiikan teoriaa, erityisesti todennäköisyyslaskennan teoriaa, siksi tilastotiedettä usein pidetäänkin sovelletun matematiikan eräänä osa-alueena (**matemaattinen eli teoreettinen tilastotiede**). Tilastotieteen ja tietotekniikan yhteistä aluetta sanotaan **tilastolliseksi tietojenkäsittelyksi**. Usein tilastollisten menetelmien kehittämisvaiheessa niihin liittyy

vaatimus sovellettavuudesta ja käsitys sovellustilanteesta. Onkin käynyt usein niin, että raja tilastotieteen ja soveltavien tieteiden välillä on hämärtnyt, jolloin soveltavien tieteiden piirissä on raja-aluetta alettu nimittää omalla nimellä (esim. epidemiologia, biometria, enviometriikka, psykometriikka, demometria ja ekonometria). Tilastotiedettä voidaan kuitenkin soveltaa lähes mihin tahansa tieteeseen, koska tilastotieteen teoria on yleistä.

Esim. 1 Deskriptiiviseen eli kuvailevaan tilastotieteeseen törmätään päivittäin

- osakkeiden hinnanmuutoksissa
- työttömyysluvuissa
- puolueiden kannatusluvuissa
- lämpötiloissa yms.

Esim.2 Tilastollista päättelyä käytetään mm.

- tulevaisuuden ennustamisessa
- vakuutusyhtiön arvioidessa vakuutuksen hintaa
- laaduntarkkailussa

Tilastollisissa analyysissä tutkittavat ongelmat pelkistyvät usein seuraavanlaisiksi kysymyksiksi:

- Millainen tilanne on keskimäärin?
- Kuinka suuri on prosentuaalinen osuus?
- Kuinka suurta on ominaisuuden vaihtelu?

- Onko eroa?
- Onko samanlaisuutta?
- Onko muutosta?

- Onko riippuvuutta?
- Millaista riippuvuus on?

- Miten tulevaisuudessa?

1.2. Tilastotieteen historiaa

Laajassa mielessä tilastotiedettä harrastettiin systemaattisten tietojen keräyksen muodossa jo muinaisessa Kiinassa ja Egyptissä (väestökirjanpito). Modernin tilastotieteen juuret voidaan ajoittaa 1600-luvulle, jolloin eurooppalaisten yhteiskuntien kehittyessä tarvittiin

luotettavaa tietoa talouden ilmiöistä (= poliittinen taloustiede, jonka erästä osa-aluetta sanottiin yliopistostatistiikaksi) sekä valtion ja väestön tilasta (= poliittinen aritmetiikka). Sanan tilasto saksan- ja englanninkieliset vastineet *Statistik* ja *statistics* viittaavatkin sanan alkuperäiseen merkitykseen: valtion kuvaus. Vuonna 1662 julkaistiin Englannissa tilastollisen tutkimuksen uranuurtajan **John Grautin** teos *Natural and Political Observations on the Bills of Mortality*.

Merkittävästi tilastotieteen syntyyn ja kehitykseen ovat vaikuttaneet myös uhkapeliongelmien. Uhhapeliuharrastusten lisääntymisen myötä alettiin 1600-luvulla tutkia todennäköisyyslaskentaa erityisesti Ranskassa.

Vielä 1700-luvulla ja sen jälkeenkin havaintoaineistoja käsiteltiin varsin alkeellisin menetelmin (yksinkertaisia menetelmiä, lähinnä kuvailevaa tilastotiedettä). Analysoivan tilastotieteen rinnalla kulki siitä erillisenä hallinnollinen tilastointi. Nämä yhdistyivät jossain määrin 1800-luvulla, kun matematiikan voimakas kehittyminen loi tilastotieteelle selkeän teoreettisen pohjan. 1800-luvulla alettiin luonnon-, yhteiskunta- ja käyttäytymistieteissä kiinnostua tilastotieteen menetelmistä. Tältä ajalta ovat peräisin esim. Gregor Mendelin perinnöllisyyskokeet. Myös matemaattinen tilastotiede alkoi kehittyä voimakkaasti 1800-luvun loppupuolella, esimerkiksi korrelaatioteorian ja regressiolain perusteet esitettiin v. 1888.

1900-luvun alkupuolella syntyivät monet tilastotieteen perusmenetelmistä. Viime vuosikymmeninä tilastotieteen teoria ja sovellusalueet ovat laajentuneet valtavasti. Tähän on erityisesti vaikuttanut tietojenkäsittelymahdollisuuksien kehittyminen.

Suomenkielinen sana *tilasto* otettiin käyttöön 1840-luvulla. Ruotsi-Suomi oli ensimmäinen valtio, jossa alettiin säännöllisesti laatia väestötilastoja, ensimmäiset tiedot koskivat vuotta 1749. Tuolloin Ruotsi-Suomen väkiluku oli 2 132 619 henkeä. Ensimmäinen suomenkielinen tilastokirja *Suomen Suuriruhtinaan Nykyinen Tilasto* julkaistiin vuonna 1848. Vuonna 1865 perustettiin **Tilastollinen toimisto** (nyk. **Tilastokeskus**). Vuonna 1905 **Karl Willgren** julkaisi ensimmäisen suomalaisen tilastotieteen oppikirjan. Ensimmäinen tilastotieteen professuuri saatiin Helsingin yliopistoon vuonna 1945.

2. HAVAINTOAINEISTON HANKINNASTA JA MITTAAMISESTA

Havaintoaineisto on tilastollisen analyysin perusta, joten on tärkeää, että se on huolella koottu ja esikäsitelty.

2.1. Havaintoaineisto, havaintomatriisi ja mittaaminen

Tilastollinen tutkimus kohdistuu aina joidenkin tutkimusobjektien muodostamaan joukkoon, joka on tutkimuksen **perusjoukko** eli **populaatio** eli **kohdejoukko**. Populaation rajaaminen on tutkimuksen ensimmäisiä vaiheita. Populaation alkioita eli tutkimusobjekteja kutsutaan **tilastoyksiköiksi**, joista käytetään merkintää $a_1, a_2, a_3, \dots$. Jos tutkittavana on konkreettinen aineisto, tilastoyksiköt nimetään omalla nimellään.

Esim.3 Tutkittavana on 20 kpl Suomen kuntia, joista tiedetään veroprosentti. Tilastoyksikkönä on kunta, mutta mikä on populaatio?

- em. kuntien joukko, jos tutkitaan vain näitä kuntia (kokonaistutkimus)
- kaikki Suomen kunnat (otantatutkimus)
- tietyn läänin kunnat (otantatutkimus)
- ...

Huom. Tutkittavista tilastoyksiköistä tehtävät johtopäätökset ulottuvat vain määrättyyn populaatioon (vrt. superpopulaatio).

Tilastoyksikköön liittyviä ominaisuuksia kutsutaan **tilastollisiksi muuttujiksi**, joita merkitään usein $x, y, z, \dots$ tai $x_1, x_2, x_3, \dots$. Jotta tilastollisia menetelmiä voidaan soveltaa, on tutkittavan ilmiön ominaisuudet voitava esittää numeerisesti. Tämä tehdään mittaamalla tilastoyksiköiltä **muuttujien arvot** eli **havaintoarvot**.

Mittaamisella tarkoitetaan menettelyä (operaatiota, sääntöä), jolla tutkittavaan tilastoyksikköön liitetään jotakin sen ominaisuutta kuvaava luku eli **mittaluku**. Kun tilastoyksikön tarkastelunalainen ominaisuus mitataan ja saadaan mittaustulos, sanotaan tätä tulosta **muuttujan arvoksi**.

Käytetyt mittaluvut ovat tilastollisen tutkimuksen lähtökohta, johon tutkimuksen onnistuminen perustuu. On huolehdyttävä siitä, että muuttujalla on korkea **validiteetti** (asianmukaisuus) eli muuttuja mittaa sitä ominaisuutta, jota sen olisi tarkoitus mitata. Esimerkiksi kysymys ”Kuinka monta kertaa syöt viikossa porkkanaraastetta?” ei mittaa sitä, pidätkö porkkanaraasteesta vai et. Myös muuttujan **reliabiliteetin** (pysyvyyden, ei-sattumanvaraisuuden) täytyy olla korkea, eli toisistaan riippumattomien samalle tilastoyksikölle tehtyjen mittausten tulokset pitäisi olla samat. Erityisesti käyttäytymis- ja yhteiskuntatieteissä ovat muuttujan validiteetti ja reliabiliteetti tärkeitä käsitteitä.

Tilastolliset muuttujat voivat olla **suoraan mitattuja** tai **teoreettisia muuttujia**. Teoreettisten muuttujien (esim. älykkyyden) mittaamisessa käytetään apuna **indikaattoreita**. Älykkyyden indikaattoreita voisivat olla esim. menestyminen erilaisissa testeissä, joiden tulokset yhdistetään esim. yhdeksi muuttujaksi laskemalla eri testien pistemäärät yhteen.

Tilastollinen muuttuja on **jatkuva**, jos se voi periaatteessa saada minkä tahansa reaalitylukuarvon joltain (järkevältä) väliltä. Vaikka muuttuja olisikin periaatteessa jatkuva, on käytännössä mittaustarkkuus aina äärellinen. Jatkuvuuden käsite perustuukin ajatukseen, että mittaustarkkuutta voidaan parantaa rajatta. Muuttuja on **diskreetti** eli **epäjatkuva**, jos sen arvoina voivat olla vain jotkin erilliset lukuarvot jollakin välillä.

Kun tutkittavilta tilastoyksiköiltä mitataan halutut tutkittavat ominaisuudet, saadaan **havaintoaineisto**. Havaintoaineisto esitetään usein **havaintomatriisina**, joka siis koostuu tilastoyksiköiden ominaisuuksia kuvaavista luvuista. Aineisto on siten **kvantitatiivinen** eli ns. **kova aineisto**.

$$\begin{array}{c}
 \begin{matrix} & x_1 & x_2 & \cdots & x_j & \cdots & x_k \\
 \begin{matrix} a_1 \\ a_2 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{matrix} & \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{j1} & \cdots & x_{k1} \\
 x_{12} & x_{22} & \cdots & x_{j2} & \cdots & x_{k2} \\
 \vdots & \vdots & \ddots & & \ddots & \\
 x_{1i} & x_{2i} & & x_{ji} & & x_{ki} \\
 \vdots & \vdots & \ddots & & \ddots & \\
 x_{1n} & x_{2n} & & x_{jn} & & x_{kn} \end{bmatrix}
 \end{matrix}
 \end{array}$$

Tilastoyksiköitä tässä havaintomatriisissa on **n** kpl (eli vaakarivien lukumäärä). Yhden tilastoyksikön (**a_i**) eri ominaisuudet esitetään yhdellä vaakarivillä. Tätä vaakariviä sanotaan ko. tilastoyksikön **havaintovektoriksi** eli **profiiliksi**. Muuttujia havaintomatriisissa on **k** kpl (eli sarakkeiden lukumäärä). Yhdellä sarakkeella esitetään siten kaikkien tilastoyksiköiden sama ominaisuus (**x_j**). Sarake muodostaa siten ko. muuttujan **jakaumavektorin**.

Mitta-asteikot

Havaintomatriisissa olevat havaintoarvot näyttävät tavallisilta reaalityluvilta. Näillä arvoilla on kuitenkin myös toinen sisältö. Ne kuvaavat jotakin ominaisuutta, ja käytetty esitystapa on vain väline ilmiön tutkimisessa. Tavallisia reaalitylukuja voidaan laskea yhteen, jakaa keskenään, niistä voidaan ottaa logaritmeja jne. Myös havaintoaineistolle tehtävät tilastolliset operaatiot perustuvat tällaisiin laskutoimituksiin, mutta näitä

operaatioita tehtäessä on aina pidettävä mielessä, että saatu tulos on voitava tulkita empiirisesti mielekkäällä tavalla. Tulkinna mielekkyys riippuu muuttujan tilastollisesta **mitta-asteikosta**. Muuttujan mitta-asteikon tunteminen on tärkeää, koska erilaisille muuttujille sopivat vain tietyt tilastolliset tunnusluvut ja analysointimenetelmät. Mitä korkeampi on mittaustaso, sitä enemmän on käytössä analyysimenetelmiä. Seuraavassa esitellään mitta-asteikkojako, jossa muuttujat jaetaan neljään ryhmään, jotka esitetään alhaisimmasta korkeimpaan.

1° Nominaali- eli luokittelu- eli laatueroasteikko

Jos tilastoyksiköt ainoastaan jaetaan ominaisuuden x perusteella luokkiin, mitataan ominaisuutta **nominaaliasteikolla**. Tällöin esimerkiksi kahdesta tilastoyksiköstä voidaan sanoa ainoastaan, että ne ovat joko *samanlaisia* tai *erilaisia* ominaisuuden x suhteen. Nominaaliasteikolla muuttujan luokkien nimet voidaan korvata numerokoodilla, jotka voidaan valita vapaasti. Luokkien tai numerokoodien järjestyksen vaihtaminen ei vaikuta mitenkään saataviin tuloksiin. Muuttuja-arvojen välillä aritmeettisilla laskutoimituksilla ei ole mielekästä tulkintaa. Ainoastaan lukumääriin perustuvat laskennat ovat järkeviä.

Esim. 4 sukupuoli: mies = 1
nainen = 2
ammatti: pappi = 1
lukkari = 2
kanttori = 3

Esim. 5 Liisa on pappi ja Leena on kanttori. Liisalla ja Leenalla *on eri* ammatit. Liisalla ja Leenalla *on sama* sukupuoli.

2° Ordinaali- eli järjestysasteikko

Ordinaaliasteikolla voidaan nominaaliasteikollisen luokittelun lisäksi jokin järjestysrelaatio, joka voidaan ilmaista sanoilla "parempi", "vaikeampi", "kauniimpi", ... Luokittelun lisäksi luokat voidaan asettaa järjestykseen ominaisuuden x mukaan. Muuttujan arvojen välillä vallitsee jokin järjestysrelaatio. Mitään lukua ei vertailuun voida kuitenkaan ottaa mukaan. Peruslaskutoimitukset eivät ole sallittuja ordinaaliasteikolla. Ordinaaliasteikollisen muuttujan arvojen koodaus on muuten vapaata, kunhan olemassa oleva järjestys tulee yksikäsitteisesti määrätyksi.

Esim. 6 arvosana: tyydyttävä = 1
hyvä = 2
kiitettävä = 3

suhtautuminen tiettyyn väitteeseen:

täysin eri mieltä = 1

jokseenkin eri mieltä = 2

ei eri mieltä eikä samaa mieltä = 3

jokseenkin samaa mieltä = 4

täysin samaa mieltä = 5

sijoitus maastajuoksun piirimestaruuskilpailuissa

Esim. 7 Matti sai tentistä arvosanan hyvä ja Liisa sai arvosanan kiitettävä. (Matti ja Liisa saivat eri arvosanan.) Liisan arvosana on *parempi kuin* Matilla.

3° Intervalli- eli välimatka-asteikko

Intervalliasteikolla voidaan luokittelun ja järjestykseen asettamisen lisäksi vertailla ominaisuuden x lisäysten suuruutta keskenään lukujen avulla. Kahden tilastoyksikön a_i ja a_j välistä eroa ominaisuuden x osalta vastaa muuttuja-arvojen x_i ja x_j erotus. Muuttuja-arvojen yhteen- ja vähennyslasku on sallittu. Muunnos $f(x) = a + bx$, kun $b > 0$, säilyttää intervalliasteikon rakenteen. Asteikon nollapiste on sopimuksenvarainen (keinotekoinen). Jotkut intervalliasteikolla mitatut muuttujat voivat saada negatiivisiakin arvoja.

Esim. 8 lämpötila Celsius- tai Fahrenheit-mittarilla mitattuna
 (x Celsius, y Fahrenheit; lineaarinen muunnos $y = 32 + 1.8x$)
 kalenterin mukaan mitattava aika
 leveys- ja pituusasteet

Esim. 9 Vaasa lämpötila on $-6\text{ }^{\circ}\text{C}$ ja Helsingin $+2\text{ }^{\circ}\text{C}$. (Vaasassa ja Helsingissä on eri lämpötila. Helsingissä on lämpimämpää kuin Vaasassa.) Helsingissä $8\text{ }^{\circ}\text{C}$ *lämpimämpää kuin* Vaasassa.

4° Suhdeasteikko

Jos intervalliasteikon vaatimukset ovat voimassa ja lisäksi on olemassa absoluuttinen nollapiste, jossa tarkasteltava ominaisuus "häviää" eli ominaisuuden määrä on todella nolla, on muuttujan mitta-asteikko **suhdeasteikko**. Yhteen- ja vähennyslaskun lisäksi voidaan muuttuja-arvoja kertoa ja jakaa keskenään. Muunnos $f(x) = ax$, kun $a > 0$, on sallittu. Suhdeasteikolla voidaan esimerkiksi vertailla, kuinka moninkertainen tietyn tilastoyksikön muuttuja-arvo on toisen tilastoyksikön muuttuja-arvoon verrattuna.

Esim. 10 pituus cm
 paino kg

Esim. 11 Matti painaa 90 kg ja Liisa 45 kg. (Matti ja Liisa ovat eri painoisia. Matti on painavampi kuin Liisa. Matti painaa 45 kg enemmän kuin Liisa.) *Matin paino on kaksinkertainen Liisan painoon verrattuna.*

Muuttujan mitta-asteikko ilmoitetaan sen toteuttaman korkeimman asteikon perusteella

Mitta-asteikot jaotellaan vielä kahteen luokkaan: nominaali- tai ordinaaliasteikon muuttujia sanotaan **kvalitatiivisiksi** eli **laadullisiksi** muuttujiksi. Intervalli- tai suhdeasteikon muuttujia sanotaan **kvantitatiivisiksi** eli **määrällisiksi** muuttujiksi.

Asteikkotyypin määrittäminen ei ole välttämättä helppoa eo. tunnusmerkkien avulla. Joissakin tilanteissa muuttujan mitta-asteikosta esiintyy erilaisia näkemyksiä. Tyypillisesti tällainen muuttuja mittaa mielipidettä. Tarkasti ottaen ko. muuttuja voi olla esim. järjestysasteikon muuttuja, mutta joskus sen ajatellaan olevan välimatka-asteikon muuttuja. Viimeksi mainittu tulkintatapa johtuu siitä, että aineiston käsittelijä mieltää muuttuja-arvojen erotuksen numeerisen erotuksen mukaiseksi.

2.2. Havaintoaineiston hankinnasta ja otantamenetelmistä

Usein tutkimuksessa on vaivalloisin ja aikaa ja rahaa vaativin vaihe havaintoaineiston hankkiminen ja muokkaaminen käyttökelpoiseksi tilastollista analyysia varten.

Kaikkien tilastoyksiköiden joukko muodostaa siis tutkimuksen populaation eli perusjoukon. Perusjoukon osajoukkoa kutsutaan **satunnaisotokseksi**, jos jokaisella perusjoukon alkiolla on tunnettu positiivinen todennäköisyys tulla valituksi otokseen. Jos ko. ehto ei ole voimassa, on kyseessä **näyte**.

Mikäli tiedot kerätään jokaisesta perusjoukon alkiosta, on kyseessä **kokonaistutkimus**. Jos tiedot kerätään otoksen avulla, on kyseessä **otantatutkimus**.

Poikkileikkaustutkimuksessa tiedot kerätään useasta eri tilastoyksiköstä tietyllä ajanhetkellä. **Pitkittäisleikkaustutkimuksessa** tilastoaineisto kerätään mittaamalla samaa tai muutamaa tilastoyksikköä eri ajankohtina (aikasarja, erilaiset seurantatutkimukset).

Havaintoaineisto saadaan valmiista tilastolähteistä ja/tai keräämällä tiedot itse. Valtion osalta tilastoinnin hoitaa **Tilastokeskus**, jonka toiminnasta ja julkaisuista saa tarkempia tietoja esim. [www-sivulta http://www.stat.fi/](http://www.stat.fi/). Myös muut laitokset, kunnat, yritykset ja järjestöt tekevät tilastollisia selvityksiä toimialansa asioista. Osa selvityksistä julkaistaan, osa on tarkoitettu vain sisäiseen käyttöön. Lisäksi näillä elimillä on rekistereitä ja muita tietoaaineistoja, joista usein on saatavissa tietoja mm. tutkimusta varten. Useimmilla mailla on omat tilastokeskuksensa, joiden [www-sivuihin](http://www.stat.fi/) löytyy linkkejä Tilastokeskuksen [www-sivuilta](http://www.stat.fi/).

Mikäli havaintoaineistoa ei ole valmiina, joudutaan se kokoamaan tilastoyksikköihin kohdistuvina mittauksina. Otantatutkimus on usein halvempi ja nopeampi kuin kokonaistutkimus. Koska tilastoyksiköitä on otoksessa vähemmän kuin koko populaatiossa, voidaan mittaukset suorittaa huolellisemmin. Jos populaatio on ääretön tai jos tilastoyksikkö joudutaan tuhoamaan mittaustilanteessa, ei kokonaistutkimusta voida tehdä.

Otantatutkimuksen suunnittelusta

1. Tavoitteiden määrittely
 - täsmentäminen
 - tietojen käyttötarkoitus
 - tulosten täsmällisyysvaatimus
2. Ongelman muuntaminen tilastolliseksi
 - perusjoukon täsmentäminen
 - tilastoyksiköt / otantayksiköt
 - muuttujat
 - kohdepopulaatio/kehikkopopulaatio
 - mittausmenetelmät
 - kyselykaavakkeet
3. Otoksen poimiminen
 - otantamenetelmän valinta
 - otoskoon määrittäminen
4. Aineiston kerääminen
 - kysely tai muunlainen mittaaminen
5. Aineiston käsittely
 - tietojen muokkaus ja analysointi
6. Raportointi
 - tulosten esittäminen selväkielisesti

Otantamenetelmistä

Otantamenetelmän valintaan liittyy otoksen koosta päättäminen, johon vaikuttaa haluttu tulosten tarkkuus, tutkimukselle varattu aika ja tietojen keruutapa. Otoksen optimikoon määrittämiseksi on kehitetty erilaisia laskentakaavoja, jotka nyt sivuutetaan. Karkeasti ottaen tutkimustulosten luotettavuus paranee tiettyyn rajaan asti otoskoon kasvaessa, mutta samalla kustannukset lisääntyvät voimakkaasti. Mielipidetutkimuksissa, jossa kartoitetaan esimerkiksi puolueiden kannatusta, käytetään usein 1000–3000 henkilön otoskokoja. Jos tutkitaan vaikkapa jonkin yrityksen imagoa tietyllä alueella, voi otoskoko olla n. 200–300 henkilöä.

Yksinkertainen satunnaisotanta (YSO) on otantamenetelmien perusmenetelmä. Se on käyttökelpoinen, jos perusjoukosta ei ole käytettävissä ennakkoinformaatiota. Sen etuna on myös tulosten laskentahelppous, sillä tilastolliset ohjelmistot olettavat yleensä, että otos on koottu YSO:lla tai systemaattisella otannalla. YSO:ssa jokaisella samankokoisella

otoksella on sama todennäköisyys tulla valituksi, ja myös jokaisella tilastoyksiköllä on sama todennäköisyys tulla valituksi. YSO:lla ei välttämättä saada edustavaa otosta, jos perusjoukko on jakaantunut keskenään heterogeenisiin ryhmiin, joissa alkiomäärät ovat hyvin erilaiset. YSO toteutetaan usein käyttämällä satunnaislukuja. Perusjoukon alkiot numeroidaan, sen jälkeen generoidaan satunnaislukuja, ja näitä lukuja vastaavat tilastoyksiköt valitaan otokseen.

Systemaattinen otanta (SO) sopii käytettäväksi silloin, kun perusjoukkoa ei tarkkaan pystytä määrittämään (esim. liikkeen asiakastutkimus) tai jos perusjoukon alkiot on listattu (esim. valmis rekisteri). SO:ssa poimitaan tilastoyksiköt tasavälein läpi koko populaation. Kun perusjoukosta valitaan joka k. tilastoyksikkö, on poimintaväli k. Poiminta aloitetaan valitsemalla k:n ensimmäisen tilastoyksikön joukosta ensimmäinen otokseen tuleva alkio, ja tämän jälkeen poimitaan systemaattisesti joka k. tilastoyksikkö. Jos perusjoukossa on N tilastoyksikköä ja otoksen koko on n yksikköä, niin poimintaväli $k = N/n$. Systemaattisen otannan etu yksinkertaiseen satunnaisotantaan on se, että se on helpompi ja nopeampi suorittaa. Systemaattinen otanta aiheuttaa kuitenkin virhettä, jos otantayksiköiden arvot kasvavat tai pienenevät systemaattisesti (esim. tilastoyksiköt ovat tutkittavan muuttujan suhteen suuruusjärjestyksessä) tai jos poimintavälin pituus on sama kuin jonkin systemaattisen jakson pituus.

Ryväsotannassa (RO) perusjoukko jaetaan kiinteisiin ryhmiin eli ryppäisiin jonkin ominaisuuden mukaan. Ryppään voi muodostaa esim. kunta, oppilaitos, jne. Ryppäistä poimitaan esimerkiksi YSO:lla tutkimukseen tulevat ryppäät. Lopullisen otoksen muodostavat joko ryppäiden kaikki tilastoyksiköt tai niistä voidaan edelleen poimia osa esimerkiksi yksinkertaisella satunnaisotannalla.

Ositetussa otannassa (OO) pyritään hyödyntämään käytettävissä olevaa taustatietoa tutkittavan ominaisuuden käyttäytymisestä perusjoukossa. Jos populaatio on jakautunut keskenään heterogeenisiin ryhmiin siten, että ryhmät ovat sisäisesti homogeenisia ja eroteltavissa toisistaan, kannattaa käyttää ositettua otantaa. Kutakin ryhmää sanotaan ositteeksi ja jokaisesta ositteesta poimitaan erikseen satunnaisotos käyttäen esim. YSO:ta tai SO:ta. Eri ositteista otokseen valittavien alkioiden lukumäärä määritetään erikseen. Ositetussa otannassa otoksen kokonaismäärän jakamista eri ositteiden kesken kutsutaan **kiintiöinniksi**. Kiintiöinti voidaan suorittaa usealla eri tavalla. **Tasainen kiintiöinti** tarkoittaa sitä, että jokaisesta ositteesta otetaan yhtä monta alkioita otokseen. **Suhteellisella kiintiöinnillä** jokaisesta ositteesta poimitaan otokseen ositteen suhteellista osuutta vastaava määrä alkioita. **Optimaalisella kiintiöinnillä** pyritään pitämään otoksen perusteella suoritettava estimointi mahdollisimman tarkkana, jolloin kiinnitetään huomiota mm. tarkasteltavan ominaisuuden ositehajontoihin. Toisinaan otetaan huomioon myös otantakustannukset.

3. YKSIULOTTEINEN EMPIIRINEN JAKAUMA

3.1. Frekvenssijakauman peruskäsitteitä ja luokitus

Jos tutkittavien **tilastoyksiköiden lukumäärä n** on suuri, ei havaintomatriisi aina riittä muuttujien yleispiirteiden selvittämiseksi. Muuttujan yleiset ominaisuudet hukkuvat yksityiskohtien joukkoon. Aineistoa on järjestettävä ja tiivistettävä. Havaintomatriisiin sisältämää tietoa voidaan tiivistää esimerkiksi muodostamalla muuttujan (luokiteltu, suora, yksiulotteinen) **frekvenssijakauma**.

Frekvenssijakauman muodostamiseksi muuttujan x saamat arvot jaetaan erillisiin **luokkiin**, merk. $E_1, E_2, \dots, E_k$, missä k on **luokkien lukumäärä**. Luokkaan E_i kuuluvien x :n arvojen lukumäärää sanotaan luokan E_i **frekvenssiksi**, merk. f_i . Kun muuttujan x luokat ja luokkia vastaavat frekvenssit tunnetaan, niin silloin tunnetaan x :n **frekvenssijakauma**.

Usein absoluuttisten frekvenssien sijasta esitetään frekvenssit, jotka on suhteutettu havaintojen kokonaismäärään n . Näitä suhteutettuja frekvenssejä voidaan käyttää esimerkiksi kahden eri havaintoaineiston frekvenssijakaumien vertailuun. Lukua $p_i = \frac{f_i}{n}$ sanotaan luokan E_i **suhteelliseksi frekvenssiksi** ja lukua **$100p_i$** sanotaan **prosentuaaliseksi frekvenssiksi**.

Jos muuttuja on epäjatkuva eli diskreetti, on luokkien määrittely yleensä selvää. Luokkina käytetään muuttujan arvoja joko sellaisenaan tai niitä vastaavia koodilukuja. Jos muuttujan luokilla on jokin vakiintunut esittämisjärjestys tai muuttuja on järjestysasteikolla mitattu, on luokat esitettävä vastaavassa järjestyksessä.

Esim. 12 Vuoden 2009 alussa Suomen kuntien läänijakauma oli seuraava:
(Aineiston peruslähde on Tilastokeskuksen tietokannat)

Lääni	f_i	p_i	$100p_i$
Etelä-Suomen	72	0.207	20.7
Länsi-Suomen	142	0.408	40.8
Itä-Suomen	54	0.155	15.5
Oulun	43	0.124	12.4
Lapin	21	0.060	6.0
Ahvenanmaan	16	0.046	4.6
Yhteensä	348	1.000	100.0

Jos luokkia tulee hyvin paljon ja suuri osa frekvensseistä on pieniä, kannattaa luokkia yhdistellä. Tällöin luokat on yhdisteltävä niin, että samaan luokkaan tulevat arvot kuuluvat mahdollisimman loogisesti yhteen.

Jos muuttuja on jatkuva-arvoinen, on sen luokittelu hankalampaa, koska tällaisen muuttujan arvot voivat olla mitä tahansa reaalitykijöitä joltain väliltä, ja kaikki mitatut arvot voivat olla erisuuruisia. Jos muuttujasta halutaan muodostaa tiivis frekvenssijakauma, on luokkien oltava välejä, jotka kattavat muuttujan arvot. Jatkuvan muuttujan luokittelussa tietoa häviää, koska nyt ei enää ilmoiteta muuttujan havaittuja arvoja vaan luokka, johon havaintoarvo kuuluu. Luokitellun aineiston esitystapa on kuitenkin usein selvempi kuin luokittelemattoman, koska jatkuva-arvoisen muuttujan jakauman esittäminen tilastokuviona perustuu usein luokitteluun. Jatkuvan muuttujan luokittelua voidaan hahmottaa seuraavasti:

Oletetaan, että luokiteltavia **havaintoja** on **n** kpl ja ne on pyöristetty jollekin **mittaustarkkuudelle**, merk. **d**. (Jos mittaustulokset ovat kokonaislukuja, on $d = 1$, jos mittauksissa on käytetty yhtä desimaalia, niin $d = 0.1$).

- 1° Etsitään **pienin arvo**, merk. $x_{(1)}$, ja **suurin arvo**, merk. $x_{(n)}$. Muuttujan x arvojen **vaihteluvälin** muodostaa väli $(x_{(1)}, x_{(n)})$. **Vaihteluvälin pituus** $w = x_{(n)} - x_{(1)}$.
- 2° Päätetään, käytetäänkö tasavälistä vai epätasavälistä luokitusta. Luokitus on tasavälinen, jos kaikki luokat ovat yhtä leveitä. Jos vain voidaan, kannattaa käyttää tasavälistä luokitusta varsinkin silloin, jos luokittelua käytetään esimerkiksi frekvenssihistogrammin perustana.
- 3° Valitaan **luokkien lukumäärä k**, ($k \approx \sqrt[3]{n}$ tai) $2^k \geq n$. (Jos $n = 125$, niin $k \approx 5 - 7$.) Yleensä luokkia on 4 - 10 kpl.
- 4° Tasavälisessä luokituksessa määritetään arvio **luokkavälin pituudelle c** siten, että $c > \frac{w}{k}$. Luokkien rajojen on oltava selkeitä, ja siksi c valitaan usein hiukan suuremmaksi kuin edellinen suhde.
- 5° Muodostetaan luokat siten, että ne peittävät koko vaihteluvälin. Ensimmäisen luokan **pyöristetyn alarajan** pitäisi olla pienempi tai yhtä suuri kuin $x_{(1)}$. Muut luokat määritellään **pyöristettyjen luokkarajojen** avulla, jotka esitetään samalla mittaustarkkuudella kuin muuttujakin on mitattu.
- 6° Tutkitaan jokainen arvo, ja määrätään luokkien frekvenssit. Yksittäinen havainto voi kuulua vain yhteen luokkaan.

Esim. 13 Erään nettimyymäyriyriksen syksyn 2009 farkkumallistosta tehdyssä otoksessa farkkujen myyntihinnat euroina ovat suurusjärjestyksessä: 43, 49, 53, 60, 64, 69, 70, 73, 73, 79, 80, 80, 85, 89, 90, 90, 90, 99, 99, 99, 100, 100, 109, 109, 110, 110, 115, 119, 120, 129, 135.

Farkkujen eli tilastoyksiköiden määrä eo. aineistossa on 31. Myyntihinta on esitetty kokonaisiksi euroiksi pyöristettynä, joten mittaustarkkuus $d = 1$.

Myyntihinta-muuttujan mitta-asteikko on suhdeasteikko. Muuttujan pienin arvo on 43 ja suurin arvo on 135. Vaihteluvälin pituus on siten $w = 135 - 43 = 92$.

Luokitellaan aineisto viiteen tasaväliseen luokkaan (k , koska $2^5 \geq 31$), siis $k = 5$.

Määritetään luokkavälin pituus: $c > \frac{92}{5} = 18.4$ ja valitaan luokkavälin

pituudeksi $c = 20$. Ensimmäisen luokan pyöristetyksi alarajaksi on valittu luku 40, koska se pienintä arvoa pienempi tasaluku. Toisen luokan pyöristetty alaraja on luokkavälin pituuden etäisyydellä ensimmäisen luokan alarajasta. Ensimmäisen luokan pyöristetty yläraja on mittaustarkkuuden verran pienempi kuin toisen luokan pyöristetty alaraja.

Absoluuttisten frekvenssien lisäksi jakaumassa on esitetty prosentuaaliset frekvenssit.

Farkkujen myyntihinta (€)	f_i	$100 p_i$
40 - 59	3	9.7
60 - 79	7	22.6
80 - 99	10	32.2
100 - 119	8	25.8
120 - 139	3	9.7
Yhteensä	31	100.0

Mittaustarkkuus d näkyy frekvenssijakaumataulukossa siten, että se on tietyn luokan pyöristetyn alarajan ja sitä edeltävän luokan pyöristetyn ylärajan erotus.

Taulukossa näkyvät pyöristetyt luokkarajat ovat luokkien symboleja. Tasavälisessä luokituksessa edeltävän luokan ja seuraavan luokan pyöristettyjen alarajojen (ja myös ylärajojen) välinen etäisyys vastaa luokkavälin pituutta. Luokan **todellinen alaraja** on ko. luokan pyöristetyn alarajan ja sitä edeltävän luokan pyöristetyn ylärajan välinen **pyöristysraja**. Ko. todellinen alaraja on samalla edeltävän luokan **todellinen yläraja**. Luokan E_i todellisesta alarajasta käytetään merkintää L_i ja todellisesta ylärajasta merkintää U_i . Todellisia luokkarajoja käytetään mm. graafisissa esityksissä sekä tunnuslukujen laskemisessa.

Luokan E_i **luokkavälin pituus c_i** on luokan todellisen ylä- ja alarajan erotus eli $c_i = U_i - L_i$. Tasavälisessä luokituksessa luokkavälin pituus on kaikilla luokilla sama ja tällöin siitä voidaan käyttää merkintää c .

Luokan E_i **luokkakeskus** m_i on luokan keskipiste eli $m_i = \frac{L_i + U_i}{2}$. Koska luokittelussa usein katoaa tilastoyksiköiden tarkat muuttuja-arvot, tulkitaan luokkakeskus usein ko. luokan havaintojen keskiarvona. Luokkakeskuksia käytetään mm. tilastokuvioissa.

Jos muuttuja on epäjatkuva, intervalli- tai suhdeasteikolla mitattu ja jos muuttujalla on paljon erilaisia arvoja, voidaan muuttujaa kohdella kuin se olisi jatkuva.

Jos muuttuja on mitattu vähintään järjestysasteikolla, voidaan muuttujalle määrittää **summafrekvenssi eli kumulatiivinen frekvenssi** F_i , joka ilmaisee, kuinka monta tilastoyksikköä (havaintoa) kuuluu luokkaan E_i tai sitä edeltäviin luokkiin yhteensä eli

$$F_i = \sum_{j=1}^i f_j$$

eli

$$F_1 = f_1$$

$$F_2 = f_1 + f_2 = F_1 + f_2$$

$$F_3 = f_1 + f_2 + f_3 = F_2 + f_3$$

⋮

$$F_k = f_1 + f_2 + \dots + f_k = F_{k-1} + f_k = n$$

Edelleen saadaan **suhteellinen summafrekvenssi** $P_i = \frac{F_i}{n}$ ja **prosentuaalinen summafrekvenssi** $100P_i$.

Esim. 14 Seuraavassa taulukossa on esitetty farkkujen hinnan frekvenssijakauman lisäksi summafrekvenssit, prosentuaaliset summafrekvenssit, todelliset luokkarajat ja luokkakeskukset.

Farkkujen myyntihinta €	f_i	F_i	$100 P_i$	L_i	U_i	m_i
40 - 59	3	3	9.7	39.5	59.5	49.5
60 - 79	7	10	32.3	59.5	79.5	69.5
80 - 99	10	20	64.5	79.5	99.5	89.5
100 - 119	8	28	90.3	99.5	119.5	109.5
120 - 139	3	31	100.0	119.5	139.5	129.5
Yhteensä	31					

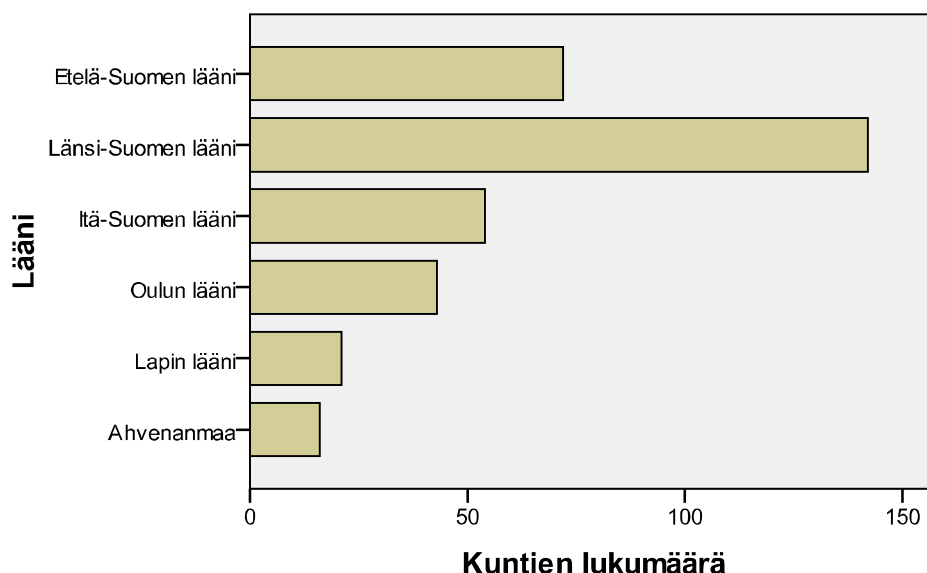
3.2. Graafinen esitys

Frekvenssijakauman voi esittää myös graafisesti. Usein käytetty kuviotyyppi on **pylväskuvio**. Pylväskuviot muodostuvat joko vaaka- tai pystypylväistä. Pylväiden pinta-alat (ja tasalevyisten pylväiden pituudet) kuvaavat määriä, joten pylvään pituutta osoittavan asteikon on hyvä lähteä luvusta 0.

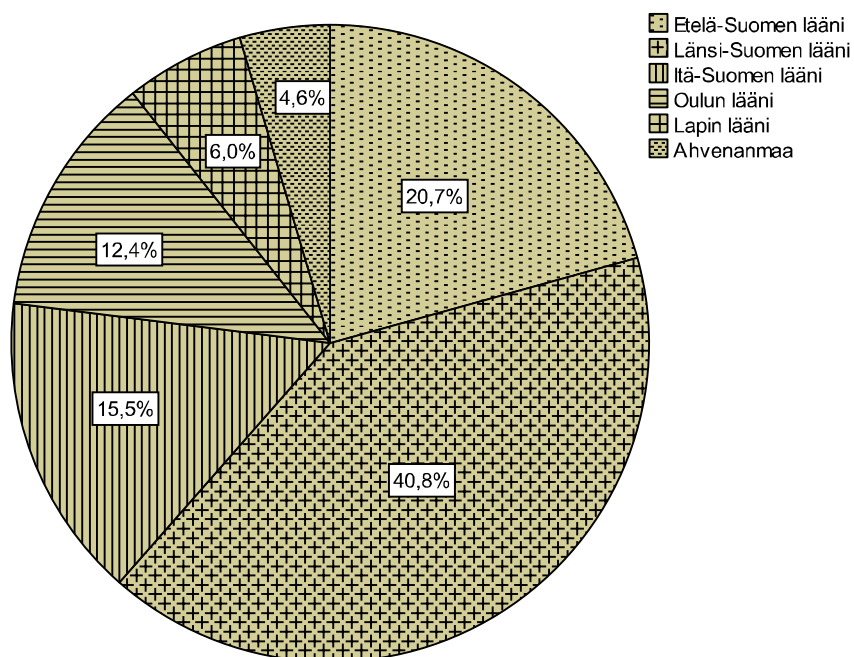
Vaakapylväskuvioita tulisi käyttää silloin, kun kuvataan laadullisen muuttujan jakaumaa. Muuttujan luokat esitetään pystyakselilla ja vaaka-akselilla kuvataan frekvenssit (absoluuttiset, suhteelliset tai prosentuaaliset). Jos muuttuja on nominaaliasteikolla mitattu, esitetään aineisto niin, että ylin pylväs on pisin ja muut pylväät piirretään pituusjärjestyksessä. Pylväiden väliin jätetään pienet raot. Jos muuttuja on järjestysasteikollinen, esitetään pylväät luokkia vastaavassa järjestyksessä.

Sektoridiagrammia (ympyräkuvio, piirakkakuvio) käytetään laadullisen muuttujan jakauman esittämisessä erityisesti silloin, kun halutaan havainnollistaa jonkin kokonaisuuden jakautumista osiin. Jokaisen luokan kokoa edustaa sektorin pinta-ala, joka on suoraan verrannollinen luokan kokoon. Sektorikuvion sijasta kannattaa käyttää vaakapylväsesitystä erityisesti silloin, jos halutaan esittää, että kahden (tai useamman) melko samankokoisen ryhmän välillä on kuitenkin eroavuutta havaintomäärässä.

Esim. 15 Suomen kuntien läänijakauma vaakapylväskuviona



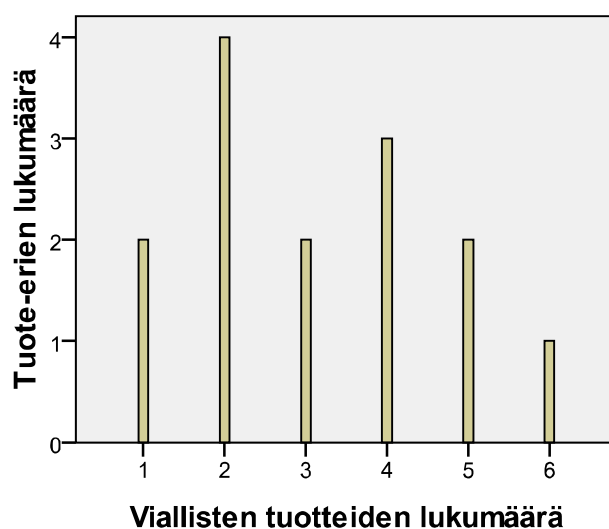
Esim. 16 Suomen kuntien läänijakauma sektorikuviona



Määrällisen epäjatkuvan muuttujan jakaumaa voidaan kuvata **janakuviolla**, joka on **pystypylväskuvio**. Janadiagrammi piirretään niin, että koordinaatistoon piirretään muuttujan arvojen kohdalle kyseisten arvojen frekvenssien korkuiset janat tai pylväät.

Esim. 17 Viallisten tuotteiden lukumääräjakauma tuote-erissä esitettynä taulukkona ja janakuviona

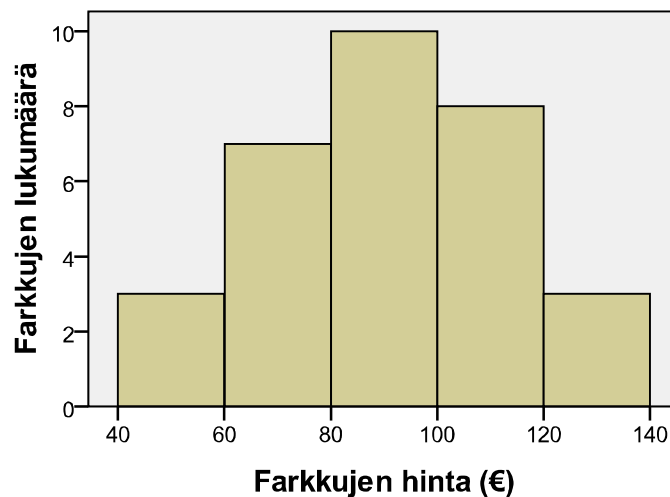
viallisten lkm	f_i
1	2
2	4
3	2
4	3
5	2
6	1



Frekvenssihistogrammi on **pystypylväskuvio**, jota käytetään määrällisille jatkuville muuttujille. Kun luokitus on tasavälinen, histogrammi muodostuu pylväistä, joiden leveys on luokkavälin pituus c , korkeus luokan E_i frekvenssi f_i ja kantojen kärkipisteinä vaakakselilla ovat todelliset luokkarajat. Yleensä kuitenkin todellisten luokkarajojen sijasta

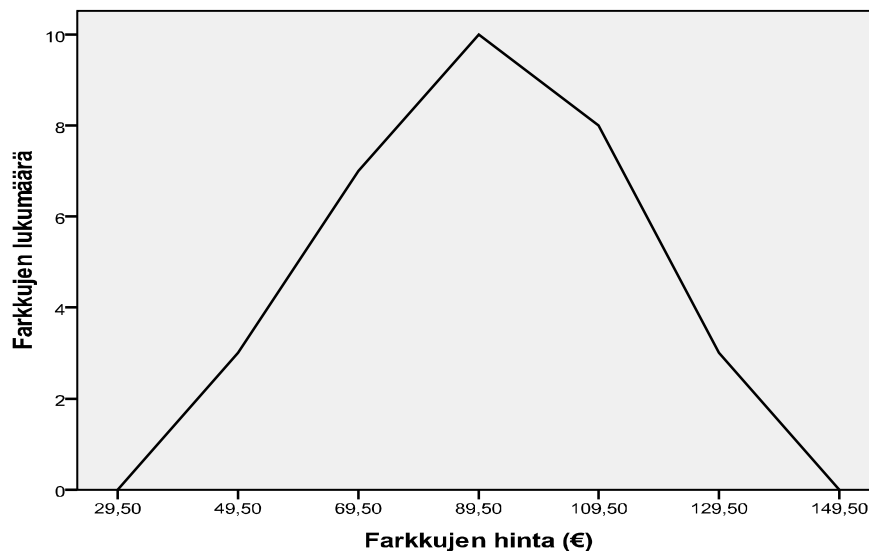
merkitään vaak-akselille näkyviin "siistit" luvut, jotka ovat lähellä todellisia luokkarajoja (tai luokkakeskuksia). Histogrammissa on pylvään pinta-ala tärkeämpi kuin korkeus, joten kuvio olisi piirrettävä niin, että luokan frekvenssi on suoraan verrannollinen pylvään pinta-alaan. Tämä vaatimus toteutuu helposti tasavälisen luokituksen yhteydessä.

Esim. 18 Farkkujen myyntihintajakauma frekvenssihistogrammina



Jatkuvan määrällisen muuttujan frekvenssijakauma voidaan esittää myös **frekvenssimonikulmion** avulla. Jokaisen luokkakeskuksen kohdalle piirretään piste frekvenssin (tai suhteellisen tai prosentuaalisen frekvenssin) korkeudelle ja peräkkäiset pisteet yhdistetään toisiinsa janoilla. Frekvenssimonikulmion päätepisteet ovat x-akselilla ns. nollaluokkien (= luokituksen alkuun ja loppuun lisättävien ylimääräisten luokkien) luokkakeskuksissa. Jos nollaluokkia ei voi määrittää, ei frekvenssimonikulmiota voi piirtää.

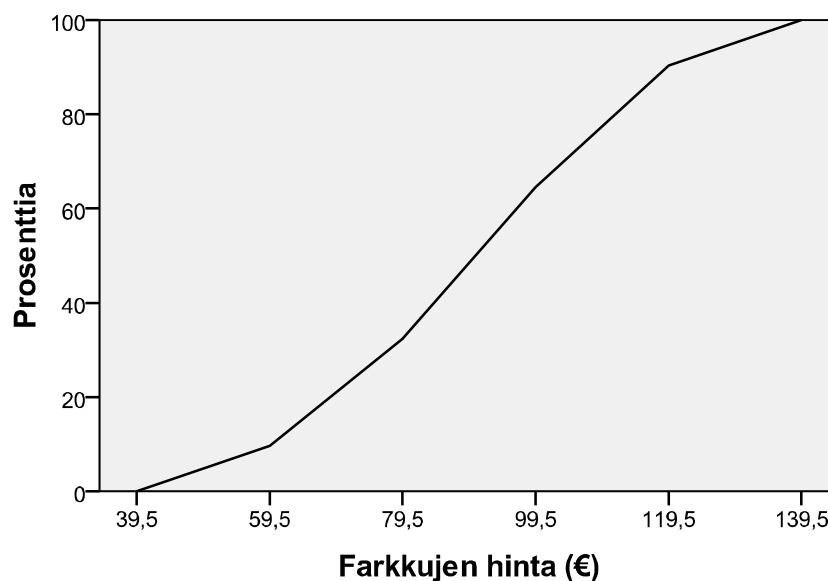
Esim. 19 Farkkujen myyntihinnan jakauma frekvenssimonikulmiona



Myös summafrekvenssijakauma voidaan esittää kuviona. Jatkuvan määrällisen muuttujan summafrekvenssijakaumaa kuvataan **summakäyrällä**. Jokaisen luokan todellisen ylärajan kohdalle piirretään piste summafrekvenssin (tai suhteellisen tai prosentuaalisen summafrekvenssin) korkeudelle ja peräkkäiset pisteet yhdistetään toisiinsa janoilla. Summakäyrä lähtee vaaka-akselilta ja nousee **n**:ään asti. Jos summakäyrä muodostetaan prosentuaalisesta summafrekvenssijakaumasta, voidaan käyrän avulla selvittää mm.

- kuinka monta % havaintoarvoista on pienempiä kuin luku **a**
- mikä on se muuttujan arvo, jota pienempiä havaintoarvoja on **p** %.

Esim. 20 Farkkujen myyntihinnan prosentuaalinen summakäyrä

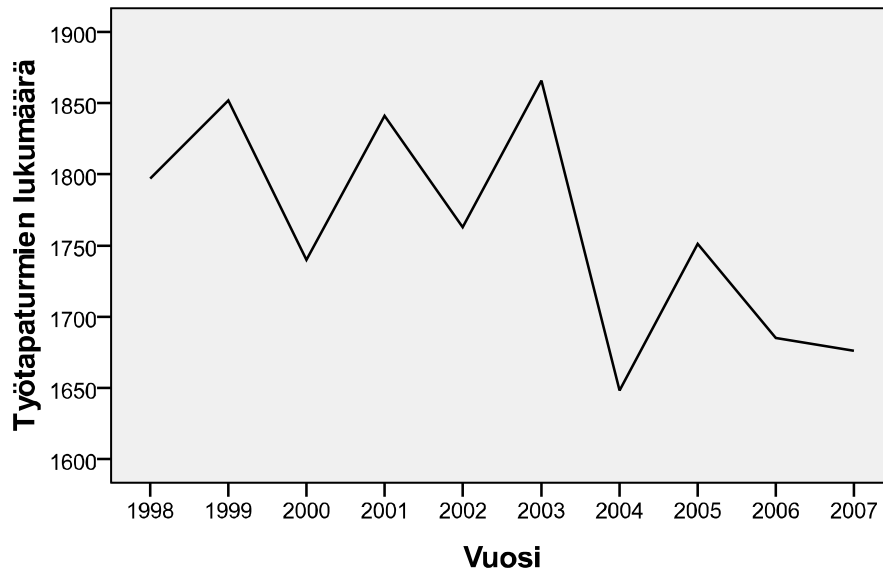


Diskreetin määrällisen muuttujan summafrekvenssijakaumaa vastaava summakäyrä on **porrasfunktio**. Vaaka-akselille merkitään muuttujan arvot ja piirretään käyrä, joka saa arvon kohdalla sen frekvenssin suuruuden hyppäyksen ja pysyy arvojen välillä edellisen arvon kohdalla saamallaan tasolla.

Viivakuviota käytetään ennen kaikkea aikasarjojen graafiseen esittämiseen. Tällöin muuttuja x kuvaa yleensä yhden tilastoyksikön yhtä ominaisuutta eri ajankohtina. Viivadiagrammissa vaaka-akselilla kuvataan aika ja pystyakselilla kuvataan muuttujan x arvot. Sekä vaaka- että pystyakselin voi katkaista.

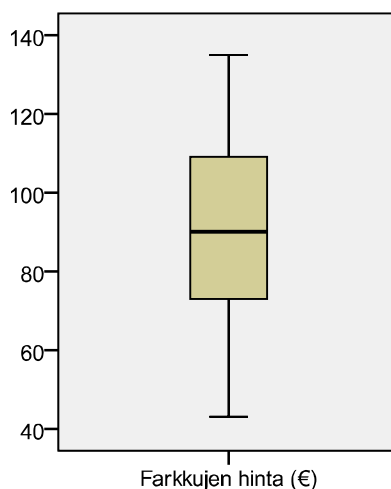
Esim. 21 Hallinto- ja toimistotyössä olevien palkansaajien työtapaturmien lukumäärä vuosina 1998–2007 (Lähde: Tilastokeskuksen PX-Web-tietokannat)

vuosi	1998	1999	2000	2001	2002	2003	2007	2005	2006	2007
liikevaihto	1797	1852	1740	1841	1763	1866	1648	1751	1685	1676



Jos muuttuja on mitattu vähintään järjestysasteikolla, voidaan sen havaintoarvojen jakautuminen esittää **laatikko-viikset-** eli **box-plot**-kuviona. Tässä kuviossa ei esitetä luokitteluun perustuvaa jakaumaa, vaan kuviosta ilmenee muuttujan tunnuslukujen arvoja. Kuviossa piirretään laatikko, jonka pohja on alakvartiilin korkeudella ja kansi on yläkvartiilin korkeudella. Muuttujan mediaani merkitään laatikkoon poikkiviivalla. Laatikon pohjasta ja kannesta piirretään viikset kummallekin puolella laatikkoa. Viiksien piirtämisessä on useita käytäntöjä, viiksien päätepisteinä voivat olla esim. pienin arvo ja suurin arvo. Viiksien päätepisteinä voivat olla myös 10 %:n ja 90 %:n fraktiilit, jolloin kuvaan voidaan vielä erikseen merkitä ne havainnot, jotka ovat kauempana jakauman keskikohdasta kuin em. fraktiilit.

Esim. 22 Farkkujen myyntihinnan laatikko-viikset-kuvio. Kuviossa viiksien päätepisteinä ovat suurin ja pienin arvo.



Määrällisen muuttujan jakaumaa voidaan esittää **runko-lehti** -kuviolla. Muuttuja-arvoista jätetään esittämättä tietty määrä oikeanpuoleisia numeroita. Jäljelle jäävistä muodostetaan

esityksen runko, jonka arvot esitetään perättäisinä kokonaislukuina pienin luku ylimmällä rivillä ja suurin alimmalla rivillä. Runkoarvojen perään kirjoitetaan lehdet yleensä siten, että havainnoista pois jätetyn numero-osuuden ensimmäiset numerot tulevat oikealle riville suuruusjärjestyksessä.

Esim. 23 Seuraavassa on farkkujen myyntihintaesimerkin runko-lehti-kuvio, jossa rungon leveys on 10 €:

4:	39
5:	3
6:	049
7:	0339
8:	0059
9:	000999
10:	0099
11:	0059
12:	09
13:	5

3.3. Yksiulotteisen jakauman tunnusluvut

Frekvenssijakaumien laatimisella yritetään saada muuttujan keskeiset ominaisuudet helpommin hahmotettaviksi. Usein muuttujan havaintoarvojen sisältämä informaatio halutaan tiivistää vieläkin voimakkaammin. Tällöin lasketaan havainnoista tilastollisia tunnuslukuja.

Sijaintia kuvaavia tilastollisia tunnuslukuja sanotaan **keskiluvuiksi**. **Hajontaluvuilla** puolestaan kuvataan havaintoarvojen vaihtelua eli "hajaantumista" jakauman keskikohdan ympärille. On olemassa myös muita jakauman muotoa kuvaavia tilastollisia tunnuslukuja.

3.3.1. Keskiluvut

Muuttujan arvojen keskimääräistä suuruutta ja jakauman sijaintia muuttuja-akselilla kuvataan keskilukujen avulla.

Moodi (Mo) eli **tyyppiarvo** on se muuttujan arvo tai luokka, jonka frekvenssi on suurin. Moodi sopii kaikille mitta-asteikoille, mutta se ei ole aina yksikäsitteinen. Vähintään intervalliasteikollisen muuttujan luokitellussa aineistossa moodi voidaan tulkita moodiluokan luokkakeskukseksi.

Esim. 24 Lääni-muuttujan moodi on Länsi-Suomen lääni, koska kuntia on eniten Länsi-Suomen läänissä.

Esim. 25 Farkkujen myyntihinnan moodiluokka on kolmas luokka: 80–99. Moodin voidaan nyt tulkita olevan moodiluokan luokkakeskus eli 89.5 € eli n. 90 €. (Alkuperäisistä havainnoista tarkasteltuna moodi ei ole yksikäsitteinen: moodiarvoja on kaksi: sekä 90 että 99.)

Esim. 26 Erään tilastotieteen kurssin opiskelijoista valitussa 19 henkilön otoksessa olivat opiskelijoiden iät suuruusjärjestyksessä: 19, 20, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 23, 23, 25, 26, 29, 42 ja 46.

Iän moodiarvo on 21 vuotta.

Mediaani (Md) eli **keskusarvo** on se havaintoarvo, jota pienempiä ja suurempia havaintoarvoja on yhtä paljon. Mediaania ei voi laskea nominaaliasteikollisesta muuttujasta.

Jos havainnot on asetettu nousevaan suuruusjärjestykseen ja kyseessä on luokittelematon aineisto, niin mediaani määrätään seuraavasti:

1° **n** pariton: Md on keskimäinen havaintoarvo $x_{(k)}$, missä $k = \frac{n+1}{2}$

2° **n** parillinen: etsitään kumpikin keskimäisistä arvoista. Jos muuttuja on ordinaaliasteikolla mitattu, on mediaani kumpikin näistä arvoista. Jos muuttuja on määrällinen, on mediaani keskimäisten havaintojen keskiarvo eli $\frac{x_{(k)} + x_{(k+1)}}{2}$, missä $k = \frac{n}{2}$.

Esim. 27 Edellisen esimerkin ikä-muuttujan mediaani sijaitsee suuruusjärjestyksessä sijalla $k = \frac{19+1}{2} = 10$. Sijalla 10 oleva havaintoarvo on 21 vuotta, joka on siis mediaani.

Luokitellulle aineistolle mediaanin määrittämiseksi on kaksi tapaa. Jos muuttuja on ordinaaliasteikollinen tai diskreetti kvantitatiivinen, niin mediaani määrätään kuten edellä. Jatkuvan luokitellun kvantitatiivisen muuttujan mediaani lasketaan kaavalla

$$Md = L_M + \frac{c}{f_M} \left(\frac{n}{2} - F_{M-1} \right),$$

missä

L_M = Md-luokan todellinen alaraja

f_M = Md-luokan frekvenssi

F_{M-1} = Md-luokkaa edeltävän luokan summafrekvenssi

c = luokkavälin pituus
n = havaintojen lkm.

Mediaaniluokka on ensimmäinen sellainen luokka, jossa $F_i \geq \frac{n}{2}$.

Mediaani voidaan määrätä myös summakäyrän avulla.

Esim. 28 Farkkujen myyntihinnan mediaaniluokka on kolmas luokka: 80 – 99. Mediaani

$$Md = 79.5 + \frac{20}{10} \left(\frac{31}{2} - 10 \right) = 90.5 \text{ €} \approx 91 \text{ €}$$

(Alkuperäisistä havainnoista tarkasteltuna mediaani on suuruusjärjestyksen sijalla 16, joten tarkka mediaani on 90 €.)

Mediaani on **fraktiilien** erikoistapaus. Fraktiilit ovat jakauman "sijaintia" kuvaavia lukuja, vaikka ne eivät yleisesti kuvaakaan keskikohtaa. Muuttujan x **p:n prosentin fraktiili** $x^{(p)}$ on sellainen havaintoarvo, jota pienempiä muuttujan arvoista on p %. Tärkeimpiä fraktiileja ovat

alakvartiili	$Q_1 = x^{(25)}$
yläkvartiili	$Q_3 = x^{(75)}$
mediaani	$Md = x^{(50)}$
desiilit	$x^{(10)}, x^{(20)}, \dots, x^{(90)}$

Fraktiilit voidaan määritellä muille paitsi nominaaliasteikon muuttujille. Kvartiilien ja fraktiilien määrittämisessä käytetään apuna mm. summakäyrää. Fraktiilin $x^{(p)}$ määrittäminen voidaan toteuttaa myös seuraavasti. Lasketaan ensin fraktiilin $x^{(p)}$ sijaluku nousevassa suuruusjärjestyksessä: $(n+1) \cdot p/100 = k.d$, missä k on kokonaisosa ja d on desimaaliosa ja lopuksi ko. fraktiili saadaan kaavasta

$$x^{(p)} = x_{(k)} + 0.d \cdot (x_{(k+1)} - x_{(k)}).$$

Esim. 29 Opiskelijoiden ikähavaintoja on 19 kpl. Alakvartiilin sijaluku on $((19+1) \cdot 25/100 =) 5.0$, joten

$$Q_1 = x^{(25)} = 20 + 0.0 \cdot (21 - 20) = 20 \text{ vuotta.}$$

Yläkvartiilin sijaluku on $((19+1) \cdot 75/100 =) 15.0$, joten

$$Q_3 = x^{(75)} = 25 + 0.0 \cdot (26 - 25) = 25 \text{ vuotta.}$$

Esim. 30 Farkkujen myyntihinnan prosentuaalisesta summakäyrästä arvioituna hinnan alakvartiili $Q_1 \approx 73 \text{ €}$ ja yläkvartiili $Q_3 \approx 108 \text{ €}$.

(Alkuperäisistä havainnoista tarkasteluna $Q_1 = 73 \text{ €}$ ja $Q_3 = 109 \text{ €}$.)

Aritmeettinen keskiarvo voidaan laskea intervalli- tai suhteasteikollisesta muuttujasta.

Luokittelemattomalle aineistolle keskiarvo saadaan kaavasta

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Esim. 31 Ikä-muuttujan keskiarvo

$$\bar{x} = \frac{1}{19} (19 + 20 + 20 + 20 + 20 + 21 + 21 + \dots + 46) = \frac{1}{19} \cdot 462 \approx 24 \text{ vuotta}$$

Luokitellulle aineistolle aritmeettinen keskiarvo saadaan kaavalla

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i m_i ,$$

missä

- f_i = luokan E_i frekvenssi
- m_i = luokan E_i luokkakeskus
- k = luokkien lkm
- n = havaintojen lkm

Huom. Edellä olevaa kaavaa voidaan käyttää, vaikka muuttuja x olisi diskreetti. Tällöin luokkakeskukset m_i korvataan muuttujan x arvoilla ja luokkien frekvenssit f_i korvataan yksittäisten arvojen frekvensseillä.

Esim. 32 Farkkujen myyntihinnan aritmeettinen keskiarvo luokitellun aineiston perusteella:

$$\begin{aligned} \bar{x} &= \frac{1}{31} (3 \cdot 49.5 + 7 \cdot 69.5 + 10 \cdot 89.5 + 8 \cdot 109.5 + 3 \cdot 129.5) \\ &= \frac{1}{31} \cdot 2794.5 = 90.145... \approx 90 \text{ €} \end{aligned}$$

(Alkuperäisistä havainnoista laskettuna tarkka aritmeettinen keskiarvo on 90 €.)

Olkoon n tilastoyksikköä jaettu k :hon ryhmään, joissa on $n_1, n_2, \dots, n_k$ tilastoyksikköä, ja joissa muuttujan x keskiarvot ovat $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$. Koko aineiston keskiarvo on

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \bar{x}_i .$$

Esim. 33 Eräällä työpaikalla on naisia 400 ja miehiä 500. Keskituntiansiot ovat vastaavasti 26.58 € ja 34.59 €. Mikä on työntekijöiden keskituntiansio?

$$\bar{x} = \frac{400 \cdot 26.58 + 500 \cdot 34.59}{400 + 500} = 31.03\text{€}$$

Keskiarvo on eniten käytetty keskiluku, joka on kuitenkin herkkä poikkeaville havainnoille. Varsinkin pienissä havaintoaineistoissa yksikin muista selvästi poikkeava arvo vetää keskiarvoa puoleensa. Joskus äärimmäisen isot ja pienet muuttuja-arvot halutaan jättää tarkastelun ulkopuolelle. Tällöin voidaan laskea esimerkiksi 5 %:n **leikattu keskiarvo**, jolloin 5 % pienimmistä ja suurimmista arvoista jätetään pois ja lopuista havainnoista lasketaan ”tavallinen” keskiarvo.

Geometrinen keskiarvo käytetään suhteasteikolla mitatun muuttujan keskiarvona silloin, kun halutaan kuvata keskimääräistä suhteellista muutosta. Geometrinen keskiarvo voidaan laskea muuttujasta, jonka kaikki havaitut arvot ovat positiivisia. Geometrinen keskiarvo saadaan laskettua kaavasta

$$G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}.$$

Esim. 34 Tuotteen hinta 1.5-kertaistui ensimmäisen vuoden aikana, toisena vuotena se 5-kertaistui ja viimeisenä vuotena 4-kertaistui. Hinnan suhteellisten muutosten geometrinen keskiarvo on

$$G = \sqrt[3]{1.5 \cdot 5 \cdot 4} \approx 3.1$$

Harmoninen keskiarvo käytetään myös suhteasteikolla. Harmoninen keskiarvo saadaan laskettua kaavasta

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}.$$

Esim. 35 Matkan ensimmäinen kolmannes ajettiin vauhtia 50 km/h, toinen kolmanneksella 25 km/h ja viimeisellä 100 km/h. Mikä on keskimääräinen vauhti koko matkalla? (Ts. millä vauhdilla nämä välit olisi ajettava, jotta koko matkaan menisi sama aika kuin todella meni, ja jokaisella kolmanneksella vauhti on sama?) Lasketaan harmoninen keskiarvo

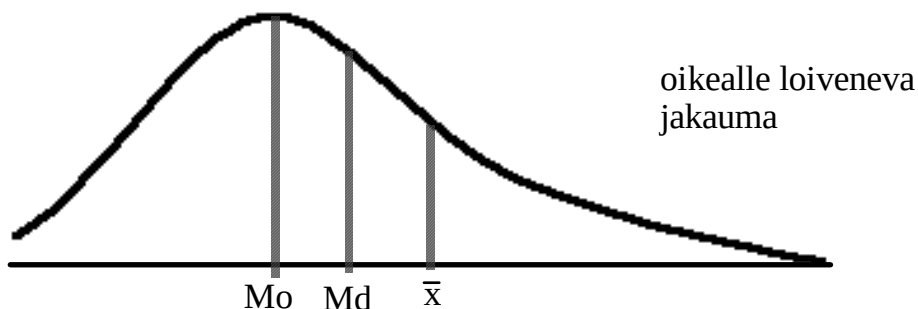
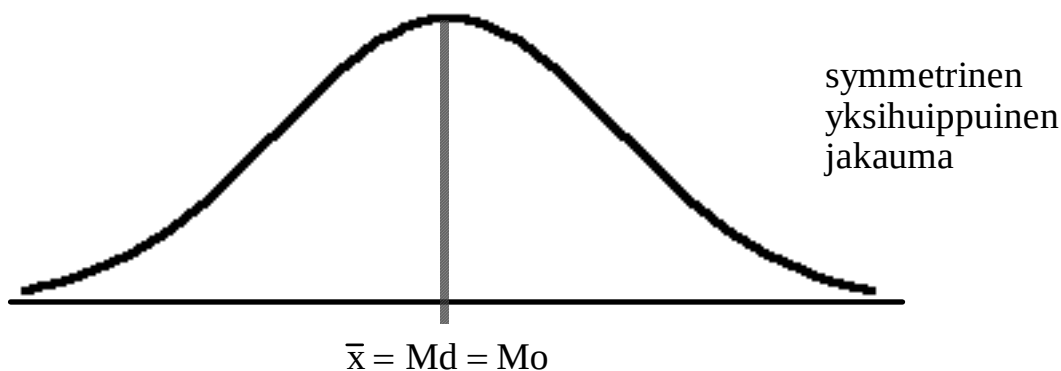
$$H = \frac{3}{\frac{1}{50} + \frac{1}{25} + \frac{1}{100}} = \frac{3}{0.02 + 0.04 + 0.01} = 42.9$$

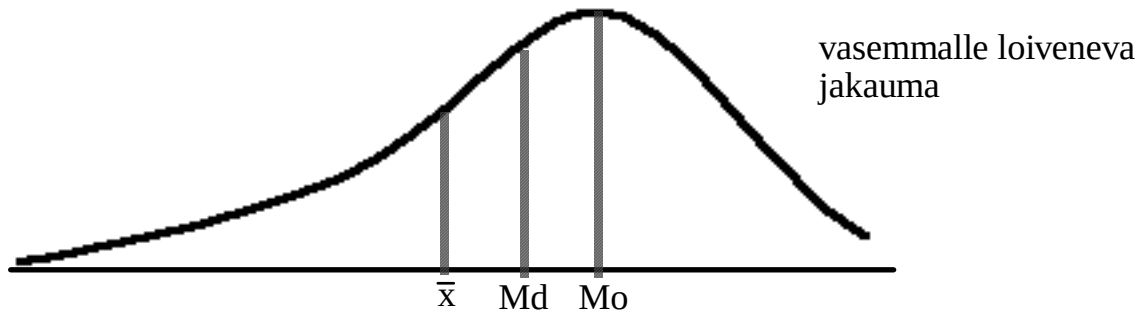
Keskilukujen vertailua

Aritmeettinen keskiarvo on tärkein keskiluku, koska se on helppo laskea. Aritmeettinen keskiarvo on herkkä poikkeaville havainnoille, se ei ole siis robusti keskiluku. Aritmeettinen keskiarvo ei ole välttämättä tyypillisin tai yleisin havaintoarvo. Jos samoista muuttujan arvoista lasketaan kaikki edellä esitetyt keskiarvot (mikä ei yleensä ole mielekäästä), ovat tulokset aina järjestyksessä $H \leq G \leq \bar{x}$.

Mediaani on helppo ymmärtää. Se on vakaa keskiluku, joka ei ole herkkä poikkeaville havainnoille. Jos muuttujan jakauma on vino, kuvaa mediaani usein aritmeettista keskiarvoa paremmin havaintojen jakaumaa. Mediaania ei kuitenkaan käytetä paljoakaan pitkälle menevissä tilastollisissa operaatioissa. Mediaani ei ole herkkä poikkeaville havainnoille, se on robusti keskiluku.

Moodi soveltuu kaikille mitta-asteikoille, mutta se on karkea keskiluku. Se ei ole aina yksikäsitteinen. Jos muuttujan jakauma on monihuippuinen, kuvaa moodi usein mediaania ja aritmeettista keskiarvoa parempi havaintojen jakaumaa. Moodi on myös robusti keskiluku.





3.3.2. Hajontaluvut

Muuttujan arvojen keskimääräistä suuruutta kuvaavat luvut eivät riitä kuvaamaan kaikkea havaintoarvojen ominaisuuksista. On myös pystyttävä kuvaamaan sitä, kuinka suurta on muuttujan arvojen vaihtelu.

Entropia eli **satunnaisuusaste** mittaa sitä, kuinka selvästi tai voimakkaasti havaitut muuttujan arvot keskittyvät yhteen tai vain muutamaan luokkaan. Entropia voidaan laskea kaavasta

$$H = - \sum_{i=1}^k p_i \log_2 p_i = -3.32193 \sum_{i=1}^k p_i \log_{10} p_i ,$$

missä

p_i = luokan E_i suhteellinen frekvenssi
 k = luokkien lkm.

Entropia soveltuu kaikille mitta-asteikoille. Se on suurimmillaan silloin, kun eri luokkien frekvenssit ovat yhtä suuret eli silloin, kun vaihtelu on suurinta. Entropian arvosta on vaikeaa nähdä suoraan, kuinka suuresta vaihtelusta on kyse, koska siihen vaikuttaa luokkien lukumäärä. Laskettua arvoa voidaan verrata entropian maksimiarvoon $H_{\max} = 3.32193 \log_{10} k$.

Esim. 36 Kuntien läänijakauman entropia

Lääni	p_i	$\log_{10} p_i$	$p_i \log_{10} p_i$
Etelä-Suomen	0.207	-0.68403	-0.14159
Länsi-Suomen	0.408	-0.38934	-0.15885
Itä-Suomen	0.155	-0.80967	-0.12550
Oulun	0.124	-0.90658	-0.11242
Lapin	0.060	-1.22185	-0.07331
Ahvenanmaan	0.046	-1.33724	-0.06151
Yhteensä	1.000		-0.67318

$$H = -3.32193 \cdot (-0.67318) \approx 2.236$$

$$H_{\max} = 3.32193 \cdot \log_{10} 6 \approx 2.585$$

Vaihteluväli on pienimmän ja suurimman havaintoarvon määräämä väli ($\mathbf{x}_{(1)}$, $\mathbf{x}_{(n)}$). Vaihteluväliä ei voi käyttää nominaaliasteikolla. Luokitellussa aineistossa vaihteluvälin muodostavat ensimmäisen luokan pyöristetty alaraja ja viimeisen luokan pyöristetty yläraja.

Vaihteluvälin pituus w soveltuu intervalli- ja suhteasteikon muuttujille. Se on suurimman ja pienimmän havaintoarvon erotus eli $w = \mathbf{x}_{(n)} - \mathbf{x}_{(1)}$. Luokitellussa aineistossa se on viimeisen luokan ylärajan ja ensimmäisen luokan alarajan erotus. Vaihteluvälin pituus on helppo laskea, mutta se ei ole yksistään käytettynä hyvä hajontaluku, koska se ottaa huomioon vain muuttujan äärimmäiset arvot.

Esim. 37 Läänin vaihteluväliä ei voida määrittää.

Esim. 38 Farkkujen myyntihinnan vaihteluväli on (43, 135) ja vaihteluvälin pituus on $w = 135 - 43 = 92$.

(Luokitellusta aineistosta: farkkujen hinnan vaihteluväli on (40, 139) ja vaihteluvälin pituus 99.)

Esim. 39 Opiskelijoiden iän vaihteluväli on (19, 46) ja vaihteluvälin pituus on 27 vuotta.

Muuttujan vaihtelua voidaan kuvata **kvartiilivälillä** (Q_1 , Q_3), joka ilmaisee havaintoarvojen keskipaikkeilta sellaisen välin, jossa on 50 % keskimmaisista arvoista. **Kvartiilivälin pituus** saadaan erotuksena $Q_3 - Q_1$. **Kvartiilipoikkeamalla** tarkoitetaan lukua

$$Q = \frac{Q_3 - Q_1}{2}.$$

Kvartiilipoikkeama on vaihteluvälin pituutta vakaampi hajontaluku ja kertoo, kuinka pitkällä välillä aineiston keskellä olevat 25 % havainnoista sijaitsevat.

Kvartiiliväli voidaan määrätä ordinaaliasteikolliselle muuttujalle, mutta kvartiilivälin pituus ja kvartiilipoikkeama vasta intervalliasteikolla. Muuttuja-arvojen hajaantumista voidaan pelkän kvartiilivälin tarkastelun sijasta tarkastella paremmin vertailemalla kvartiiliväliä ja vaihteluväliä toisiinsa.

Esim. 40 Farkkujen myyntihinnan prosentuaallisen summakäyrän perusteella arvioituna $Q_1 \approx 73$ € ja $Q_3 \approx 108$ €, joten kvartiilivälin pituus on noin 35 € ja kvartiilipoikkeama $Q = 17.5$ €.

Esim.41 Opiskelijoiden iän kvartiiliväli on (20, 25). Kvartiilivälin pituus on 5 vuotta ja kvartiilipoikkeama 2.5 vuotta.

Käytetyimpiä hajontalukuja ovat **varianssi** s^2 ja **keskihajonta** s , vaikka niiden tulkinta ei ole niin yksinkertaista kuin em. hajontaluvuilla. Varianssi ja keskihajonta voidaan laskea intervalli- tai suhdeasteikollisesta muuttujasta. Keskihajonta on varianssin positiivinen neliöjuuri eli $s = \sqrt{s^2}$.

Varianssi kertoo, kuinka tiiviisti havaintoarvot ovat keskittyneet keskiarvon ympärille. Jos kaikki mittaustulokset ovat samoja, on $s^2 = 0$, muulloin $s^2 > 0$. Keskihajonnassa ja varianssissa muuttujien arvojen vaihtelu ilmaistaan rakenteellisesti samalla tavalla. Keskihajonta on kuvailussa havainnollisempi, koska sillä on sama laatu kuin muuttujan arvoilla, ja se kertoo, kuinka kaukana keskimäärin havainnot ovat keskiarvosta. Varianssi on taas parempi teoreettisissa tarkasteluissa.

Luokittelemattoman aineiston varianssi voidaan laskea kaavalla

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right].$$

Esim. 42 Opiskelijoiden iän varianssin ja keskihajonnan laskenta:

$$\sum x_i = 19 + 20 + 20 + 20 + 20 + 21 + \dots + 46 = 462$$

$$\sum x_i^2 = 19^2 + 20^2 + 20^2 + 20^2 + 20^2 + 21^2 + \dots + 46^2 = 12214$$

$$s^2 = \frac{1}{19-1} \left[12214 - \frac{462^2}{19} \right] \approx 54.450 \dots \text{ vuotta}^2$$

$$s = 7.379 \dots \text{ vuotta} \approx 7 \text{ vuotta}$$

Luokitellun aineiston varianssi on

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (m_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^k f_i m_i^2 - \frac{\left(\sum_{i=1}^k f_i m_i \right)^2}{n} \right]$$

missä

- f_i = luokan E_i frekvenssi
- m_i = luokan E_i luokkakeskus
- k = luokkien lkm
- n = havaintojen lkm

Esim. 43 Farkkujen myyntihinnan varianssi ja keskihajonta

$$s^2 = \frac{1}{31-1} \left[(3 \cdot 49.5^2 + 7 \cdot 69.5^2 + 10 \cdot 89.5^2 + 8 \cdot 109.5^2 + 3 \cdot 129.5^2) - \frac{2794.5^2}{31} \right]$$

$$= \frac{1}{30} \cdot \left(267497.75 - \frac{2794.5^2}{31} \right) \approx 519.57 \text{ €}^2$$

$$s = 22.79 \dots \text{ €} \approx 23 \text{ €}$$

(Alkuperäisistä havainnoista tarkasteluna: $s^2 = 549.267$ ja $s = 23.4$.)

Varianssin ja myös keskihajonnan arvot riippuvat muuttujan mittayksiköstä. Jos muuttujalle tehdään lineaarinen muunnos $y = a + bx$, niin $s_y^2 = b^2 s_x^2$ ja $s_y = |b|s_x$.

Esim. 44 Pituus x on mitattu tuumina ja x :n varianssi on 5. Jos pituus mitataan cm:nä eli x :lle tehdään muunnos $y = 2.54 x$, niin y :n varianssi $s_y^2 = 2.54^2 \cdot 5 = 32.36$.

Keskiarvoa ja keskihajontaa voidaan hyödyntää esimerkiksi muuttujan x havaintoarvojen **standardoinnissa**:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Standardoitu arvo z_i kertoo, kuinka monen keskihajonnan etäisyydellä havaintoarvo x_i on keskiarvosta $\bar{x}$. Standardoiduille arvoille z_i pätee aina, että niiden keskiarvo $\bar{z} = 0$ ja keskihajonta $s_z = 1$. Standardoitu muuttuja z on pelkkä luku; se on siis riippumaton

alkuperäisen muuttujan x mittayksiköstä. Standardoituja havaintoarvoja voidaan käyttää mm. kun eri havaintoaineistojen tilastoyksiköitä verrataan toisiinsa.

Esim. 45 Opiskelija osallistui tilastotieteen tenttiin ja sai pistemääräksi 36. Hän osallistui myös talousmatematiikan tenttiin ja sai pistemääräksi 30. Tilastotieteen tentin pistemäärän keskiarvo oli 29 ja keskihajonta 6, talousmatematiikan tentissä vastaavat luvut olivat 22 ja 8.

Tenttitulokset standardoituna ovat

$$z_{tt} = \frac{36 - 29}{6} \approx 1.2 \quad \text{ja} \quad z_{tm} = \frac{30 - 22}{8} \approx 1.0$$

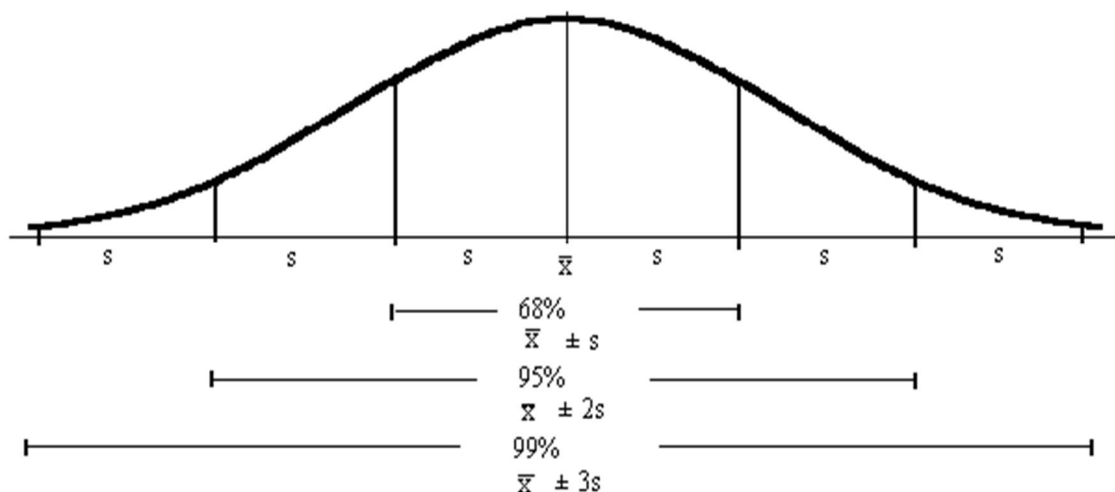
Opiskelija menestyi tilastotieteessä suhteellisesti paremmin.

Variaatiokerroin $V = \frac{s}{\bar{x}}$ on mittayksiköstä riippumaton hajontaluku, jota voidaan käyttää suhdeasteikolla. Variaatiokerroin ilmaisee muuttujan suhteellisen vaihtelun. Usein variaatiokerroin ilmaistaan prosenttilukuna, jolloin luku $100V$ kertoo, kuinka monta % keskihajonta on keskiarvosta. Variaatiokerrointa käytetään vertailtaessa mittayksiköiltään erilaisia aineistoja.

Esim. 46 Opiskelijoiden iän variaatiokerroin $V = \frac{7.379...}{24.315...} \approx 0.30$. Iän hajonta on siten 30 % iän keskiarvosta.

Esim. 47 Farkkujen myyntihinnan variaatiokerroin $V = \frac{22.79...}{90.145...} \approx 0.25$. Hinnan hajonta on siten 25 % hinnan keskiarvosta

Huom. Käytännössä on havaittu:



3.3.3. Yksiulotteisen jakauman muita tunnuslukuja

Jakauman sijaintia ja vaihtelua kuvaavien tunnuslukujen lisäksi voidaan mitata jakauman symmetriasta poikkeamista eli **vinoutta** sekä keskittymisen terävyyttä tai tylsyyttä eli **huipukkuutta**. Ko. tunnuslukuja määritetään yleensä intervalli- ja suhdeasteikon muuttujille.

Jos muuttujan arvot ovat keskittyneet voimakkaasti alimpiin luokkiin, ja jakaumalla on pitkä häntä oikealle päin, sanotaan muuttujan jakaumaa oikealle vinoksi tai oikealle loivenevaksi. Jos taas muuttujan arvot ovat keskittyneet ylimpiin luokkiin, on muuttujan jakauma vasemmalle loiveneva tai vino. Symmetrisessä jakaumassa keskiarvo ja mediaani ovat yhtä suuret, ja jakauman muoto oikealle ja vasemmalle keskipisteestä saadaan peilikuvana.

Frekvenssijakauman **vinouden** mitta on suure

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}.$$

Jos jakauma on täsmälleen symmetrinen on $g_1 = 0$ (esim. normaalijakauma); jos jakauma on vasemmalle loiveneva, on $g_1 < 0$; jos jakauma on oikealle loiveneva, on $g_1 > 0$. Peukalosääntönä pidetään usein seuraavaa: symmetrisenä jakaumana pidetään jakaumaa, jolle $-0.5 < g_1 < 0.5$.

Jakaumaa voi tutkia myös huipukkuuden avulla. **Huipukkuuden** mittana on suure

$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - 3.$$

Jos muuttujan arvot ovat keskittyneet pariin luokkaan, ja muissa luokissa on vain vähän havaintoja, on jakauma huipukas eli **leptokurtinen**. Huipukkaalla jakaumalla $g_2 > 0$. Laakeassa eli **platykurtisessa** jakaumassa $g_2 < 0$. Jos $g_2 = 0$, on jakauman muoto jotain tältä väliltä eli **mesokurtinen**. Mm. normaalijakauman huipukkuus $g_2 = 0$. Peukalosääntönä pidetään usein seuraavaa: mesokurtisena jakaumana pidetään sellaista jakaumaa, jolla $-0.5 < g_2 < 0.5$.

4. KAKSIULOTTEINEN EMPIIRINEN JAKAUMA

Kahden tilastollisen muuttujan arvoilla voi olla taipumus liittyä toisiinsa niin, että muuttujan x tietyn arvon yhteydessä esiintyy suhteellisesti ottaen muita useammin tietty muuttujan y arvo. Usean muuttujan samanaikaisen tarkastelun lähtökohtana onkin usein juuri muuttujien välisen mahdollisen yhteyden olemassaolon, voimakkuuden ja luonteen selvittäminen.

Täydellinen (matemaattinen, funktionaalinen) riippuvuus on kyseessä silloin, kun vuorovaikutussuhde voidaan ilmaista yksikäsitteisesti jollakin kaavalla.

Esim. 48 Neliön pinta-ala A riippuu neliön sivun pituudesta x kaavan $A = x^2$ mukaan.

Tilastollisen eli epätäydellisen riippuvuuden kuvaamiseen käytetään mm. erilaisia tilastollisia **riippuvuustunnuslukuja**.

4.1. Ristiintaulukko ja kontingenssikerroin

Oletetaan, että muuttujan x arvot on jaettu luokkiin, joita on J kappaletta ja muuttujan y arvot on jaettu luokkiin, joita on I kappaletta. Muuttujat x ja y voivat olla mitta-asteikoiltaan mitä tahansa. Merkitään muuttujan x eri luokkia symboleilla $E_1, E_2, \dots, E_J$ ja muuttujan y eri luokkia symboleilla $G_1, G_2, \dots, G_I$. Muuttujien x ja y välinen **ristiintaulukko** (frekvenssitaulukko, kontingenssitaulukko, x :n ja y :n yhteisjakauma) on silloin

x	E_1	E_2	...	E_J	yhteensä
y					
G_1	f_{11}	f_{12}	...	f_{1J}	$f_{1\bullet}$
G_2	f_{21}	f_{22}	...	f_{2J}	$f_{2\bullet}$
...					
...			f_{ij}		
...					
G_I	f_{I1}	f_{I2}	...	f_{IJ}	$f_{I\bullet}$
yhteensä	$f_{\bullet 1}$	$f_{\bullet 2}$	...	$f_{\bullet J}$	$f_{\bullet\bullet} = n$

missä

- x on yleensä syy,
- y on yleensä seuraus,
- f_{ij} on sellaisten tilastoyksiköiden lukumäärä joiden x -muuttujan arvo kuuluu luokkaan E_j ja joiden y -muuttujan arvo kuuluu luokkaan G_i ; sanotaan, että f_{ij} on solun (G_i, E_j) **havaittu frekvenssi**

- $f_{i\bullet}$ on i:nneen vaakarivin frekvenssien summa eli **rivisumma**;
 $f_{i\bullet} = f_{i1} + f_{i2} + \dots + f_{ij}$
- $f_{\bullet j}$ on j:nneen sarakkeen frekvenssien summa eli **sarakesumma**;
 $f_{\bullet j} = f_{1j} + f_{2j} + \dots + f_{Ij}$
- $f_{\bullet\bullet}$ on tilastoyksiköiden kokonaismäärä eli **kokonaissumma**;
 $f_{\bullet\bullet} = n = f_{\bullet 1} + f_{\bullet 2} + \dots + f_{\bullet J} = f_{1\bullet} + f_{2\bullet} + \dots + f_{I\bullet}$

Luvut $f_{1\bullet}, f_{2\bullet}, \dots, f_{I\bullet}$ muodostavat muuttujan y **reunajakauman**, joka on itse asiassa muuttujan y yksiulotteinen jakauma. Vastaavasti luvut $f_{\bullet 1}, f_{\bullet 2}, \dots, f_{\bullet J}$ muodostavat muuttujan x reunajakauman.

Kahden muuttujan **suhteellinen yhteisjakauma** saadaan jakamalla jokainen solufrekvenssi havaintojen kokonaismäärällä n. **Prosentuaalinen yhteisjakauma** saadaan kertomalla suhteelliset solufrekvenssit luvulla sata.

Esim. 49 Eräällä tilastotieteen kurssilla opiskelijoilta kysyttiin, ovatko he ansiotyössä ja kuinka he kokevat opintojensa edistyvän muihin saman alan saman vuosikurssin opiskelijoihin verrattuna. Saatiin ristiintaulukko

Onko ansiotyössä? Opintojen edistyminen	kyllä	ei	Yht.
keskimääräistä hitaammin	16	11	27
keskimääräisesti	25	75	100
keskimääräistä nopeammin	3	14	17
Yhteensä	44	100	144

Ristiintaulukosta laskettu prosentuaalinen yhteisjakauma on

Onko ansiotyössä? Opintojen edistyminen	kyllä	ei	Yht.
keskimääräistä hitaammin	11 %	8 %	19 %
keskimääräisesti	17 %	52 %	69 %
keskimääräistä nopeammin	2 %	10 %	12 %
Yhteensä	30 %	70 %	100 %

Kahden muuttujan yhteisjakaumasta voidaan määrätä lisäksi **ehdollisia jakaumia**. Jokainen sarake muodostaa y-muuttujan ehdollisen jakauman tietyllä muuttujan x arvolla, ja jokainen vaakarivi muodostaa x-muuttujan ehdollisen jakauman tietyllä muuttujan y arvolla. Jakamalla ehdollisen jakauman frekvenssit niiden summalla saadaan **suhteellinen ehdollinen frekvenssijakauma**. Kertomalla ne edelleen sadalla, saadaan **prosentuaalinen ehdollinen frekvenssijakauma**.

Erityisesti prosentuaaliset ehdolliset jakaumat ovat käyttökelpoisia, kun selvitetään muuttujien välisen riippuvuuden luonnetta.

Esim. 50 Seuraavassa on esitetty y-muuttujan (= opintojen edistyminen) prosentuaaliset ehdolliset jakaumat x-muuttujan (=onko ansiotyössä) eri luokissa:

Onko ansiotyössä?	kyllä	ei	Yht.
Opintojen edistyminen			
keskimääräistä hitaammin	36 %	11 %	19 %
keskimääräisesti	57 %	75 %	69 %
keskimääräistä nopeammin	7 %	14 %	12 %
Yhteensä	100 %	100 %	100 %

Nyt muuttujan y jakaumat ovat erilaiset x-muuttujan eri luokissa.

Jos ristiintaulukosta määrätyt prosentuaaliset ehdolliset jakaumat ovat samanlaiset eri sarakkeilla (tai riveillä), sanotaan muuttujien olevan **tilastollisesti riippumattomia**. Jos muuttujat eivät ole riippumattomia, sanotaan niiden olevan **tilastollisesti riippuvia**.

Tilastollisen riippuvuuden voimakkuutta voidaan mitata ristiintaulukosta seuraavasti: lasketaan riippumattomuustilannetta vastaavat **odotetut** eli **teoreettiset frekvenssit** e_{ij} , jotka saadaan määrättyä havaitun ristiintaulukon rivi- ja sarakesummien avulla kaavalla

$$e_{ij} = \frac{f_{i\bullet} \cdot f_{\bullet j}}{n}.$$

Nämä voidaan asettaa samanlaiseen taulukkomuotoon kuin havaitut frekvenssit:

x	E ₁	E ₂	...	E _J	yhteensä
y					
G ₁	e ₁₁	e ₁₂	...	e _{1J}	f _{1•}
G ₂	e ₂₁	e ₂₂	...	e _{2J}	f _{2•}
⋮					
⋮					
⋮					
G _I	e _{I1}	e _{I2}	...	e _{IJ}	f _{I•}
yhteensä	f _{•1}	f _{•2}	...	f _{•J}	f _{••} = n

Havaittujen frekvenssien ja odotettujen frekvenssien välinen ero kuvaa muuttujien x ja y riippuvuuden määrää. Tämän riippuvuuden suuruuden eräänä mittana voidaan käyttää χ^2 -arvoa (lue: khiin neliö)

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(f_{ij} - e_{ij})^2}{e_{ij}}.$$

χ^2 -arvo saa arvon nolla, jos havaitut frekvenssit ovat täsmälleen riippumattomuus-tilannetta vastaavia. Sen arvo on sitä suurempi, mitä enemmän havaitut frekvenssit poikkeavat riippumattomuuden tilanteesta. χ^2 -arvo ei sellaisenaan sovellu riippuvuuden mitaksi, koska mm. ristiintaulukon koko vaikuttaa sen suuruuteen. χ^2 :n **maksimiarvo**

$$\chi^2_{\max} = (k - 1) n,$$

missä k on pienempi luvuista I ja J, ja n on havaintojen määrä.

Usein riippuvuuslukuna käytetään **kontingenssikerrointa C**, joka saadaan normeeraamalla χ^2 -arvo seuraavasti:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

Jos x ja y ovat riippumattomia, on C:n arvo nolla. Muuttujien välisen riippuvuuden lisääntyessä kasvaa C:n arvo. Kontingenssikertoimen maksimiarvo on

$$C_{\max} = \sqrt{\frac{k-1}{k}},$$

missä k on pienempi luvuista I ja J.

Koska sekä χ^2 -arvo että kontingenssikertoimen arvo riippuvat ristiintaulukon koosta, ei erikokoisten ristiintaulukoiden riippuvuustunnuslukuja pitäisi verrata toisiinsa. Vertailtavuutta voidaan parantaa, kunhan havaitut riippuvuusluvut jaetaan vastaavilla maksimiarvoilla.

Esim. 51 Seuraavassa on esitetty odotetut frekvenssit ja tunnuslukuja

Onko ansiotyössä? Opintojen edistyminen	kyllä	ei	Yht.
keskimääräistä hitaammin	8	19	27
keskimääräisesti	31	69	100
keskimääräistä nopeammin	5	12	17
Yhteensä	44	100	144

$$\chi^2 = \frac{(16-8)^2}{8} + \frac{(11-19)^2}{19} + \frac{(25-31)^2}{31} + \frac{(75-69)^2}{69} + \frac{(3-5)^2}{5} + \frac{(14-12)^2}{12} = 14.2$$

$$C = \sqrt{\frac{14.2}{14.2 + 144}} = 0.30$$

$$\chi^2_{\max} = (2 - 1) \cdot 144 = 144$$

$$C_{\max} = \sqrt{\frac{2-1}{2}} \approx 0.71$$

Kontingenssikertoimen perusteella voidaan sanoa muuttujien välillä olevan riippuvuutta. Riippuvuutta voi luonnehtia ehdollisten jakaumien avulla: ansiotyössä käyvistä opiskelijoista 36 % koki opintojensa edistyvän keskimääräistä hitaammin, kun taas muiden opiskelijoiden joukossa vastaava prosenttiosuus oli vain 11 %.

Huom. Ristiintaulukkoa, jossa on kaksi vaakariviä ja kaksi saraketta, sanotaan nelikentäksi. Nelikentästä voidaan χ^2 -arvo laskea kaavasta

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1\bullet} f_{2\bullet} f_{\bullet 1} f_{\bullet 2}}$$

Esim. 52 Muuttuja x saa arvot M ja N sen mukaan kumpi puolisoista yleensä hoitaa puhumisen, ja muuttuja y saa arvot M ja N sen mukaan kumpi puolisoista useammin päättää yhteisistä asioista. Satunnaisesti poimituista 34 avioparista saatiin seuraava ristiintaulukko:

y \ x	x		Yhteensä
	M	N	
M	13	6	19
N	5	10	15
Yhteensä	18	16	34

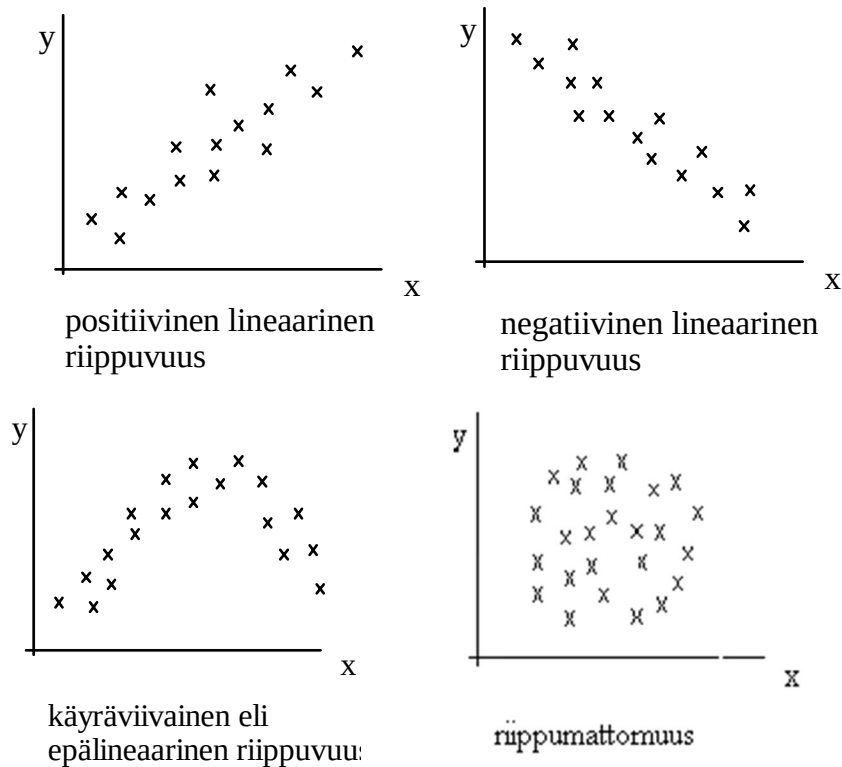
$$\chi^2 = \frac{34(13 \cdot 10 - 5 \cdot 6)^2}{18 \cdot 16 \cdot 19 \cdot 15} = 4.1 \quad \text{ja} \quad \chi^2_{\max} = (2-1) \cdot 34 = 34$$

$$C = \sqrt{\frac{4.1}{4.1 + 34}} = 0.33 \quad \text{ja} \quad C_{\max} = \sqrt{\frac{2-1}{2}} \approx 0.71$$

Voidaan siis päätellä, että puhumisen ja yhteisistä asioista päättämisen välillä näyttää tilastollista riippuvuutta olevan siten, että se joka puhuu usein myös päättää.

4.2. Korrelaatiodiagrammi ja korrelaatio

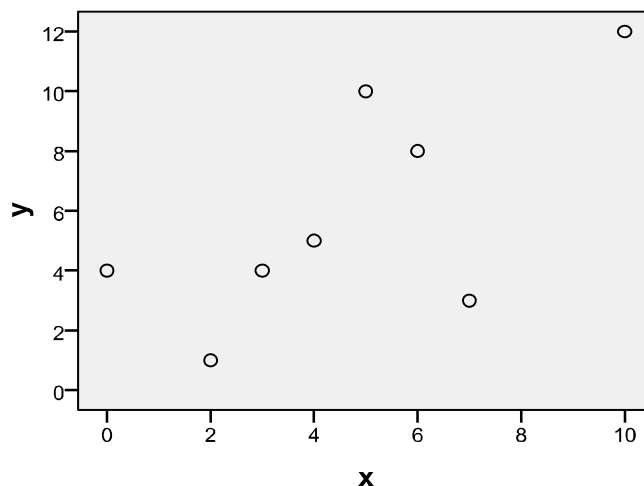
Tarkastellaan kahta muuttujaa x ja y , jotka ovat vähintään intervallasteikolla mitattuja. Muuttujien x ja y välisellä **korrelaatiodiagrammilla** eli **pisteparvella** eli **sirontakuviolla** tarkoitetaan sellaista graafista esitystä, missä n havaintoparia (x_i, y_i) on merkitty koordinaatistoon. Pisteparven muodon avulla voidaan selvittää, onko muuttujien x ja y välillä **lineaarista** eli **suoraviivaista riippuvuutta**. Mitä suoranomaisempaan muotoon pisteparven pisteet ovat ryhmittyneet, sitä voimakkaampaa on x :n ja y :n välinen lineaarinen riippuvuus. **Positiivisella** lineaarisella riippuvuudella tarkoitetaan sitä, että x :n arvojen kasvaessa myös y :n arvot kasvavat tasaisesti; vastaavasti **negatiivisella** lineaarisella riippuvuudella tarkoitetaan sitä, että x :n arvojen kasvaessa y :n arvot pienenevät tasaisesti.

Esim. 53

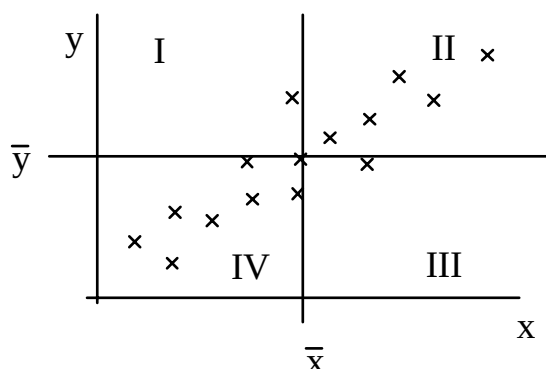
Esim. 54 Tutkitaan erään virvoitusjuoma viimeisimmän mainoskampanjan ja ko. juoman myynnin välistä riippuvuutta. Kahdeksalta henkilöltä kysyttiin, kuinka monta ko. juoman mainosta he olivat nähneet viime aikoina (= muuttuja x), ja kuinka monta ko. merkistä virvoitusjuomapulloa he olivat ostaneet (= muuttuja y). Saatiin havaintoaineisto

Henkilö:	1	2	3	4	5	6	7	8
x :	5	10	4	0	2	7	3	6
y :	10	12	5	4	1	3	4	8

Muuttujien x ja y välinen pisteparvi on



Lineaarisen riippuvuuden voimakkuuden mittaamiseksi jaetaan muuttujien x ja y välinen pisteparvi neljään osaan muuttujien keskiarvojen perusteella:



Lineaarista riippuvuutta mittaava tilastollinen tunnusluku perustuu tuloon $(x_i - \bar{x})(y_i - \bar{y})$. Jos havainto (x_i, y_i) on lohkoissa II tai IV, on tulon $(x_i - \bar{x})(y_i - \bar{y})$ arvo positiivinen, kun taas havainnon ollessa lohkoissa I tai III ko. tulo saa negatiivisen arvon. Nyt edellisessä kuviossa havaintoja on paljon lohkoissa II ja IV, joten eo. tulon termeistä suurin osa on positiivisia ja siten $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ saa positiivisen arvon. Eo.

summa mittaa x :n ja y :n välisen lineaarisen riippuvuuden voimakkuutta. Kun tämä lauseke jaetaan termillä $(n - 1)$, saadaan x :n ja y :n **kovarianssiksi** kutsuttu riippuvuustunnusluku

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{1}{n - 1} \left[\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n} \right].$$

Kovarianssi ei kuitenkaan ole sopivin x :n ja y :n lineaarisen riippuvuuden mitta, koska sen arvoon vaikuttavat mm. muuttujien mittayksiköt, ja sitä on usein hankala tulkita. Sen sijaan **Pearsonin (tulomomentti)korrelaatiokerroin**

$$r_{xy} = r = \frac{s_{xy}}{s_x s_y},$$

missä

- s_{xy} = muuttujien x ja y välinen kovarianssi
- s_x = muuttujan x keskihajonta
- s_y = muuttujan y keskihajonta,

arvoon muuttujien mittayksiköt eivät vaikuta.

Korrelaatiokerroin voidaan myös laskea seuraavan kaavan avulla:

$$r_{xy} = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}\right] \left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}\right]}}$$

Korrelaatiokertoimen arvo on aina välillä $[-1, +1]$. Positiivinen korrelaatiokertoimen arvo kuvaa positiivista lineaarista riippuvuutta ja negatiivinen arvo vastaavasti negatiivista lineaarista riippuvuutta. Mitä suurempi korrelaatiokertoimen itseisarvo on, sitä voimakkaammasta lineaarisesta riippuvuudesta on kyse. Arvo -1 kuvaa täydellistä negatiivista ja arvo $+1$ täydellistä positiivista lineaarista riippuvuutta. Jos muuttujat ovat lineaarisesti riippumattomia, on korrelaatiokertoimen arvo nolla. Jos korrelaatiokertoimen arvo on nolla, voi muuttujien välillä kuitenkin esiintyä epälineaarista riippuvuutta. Korrelaatiokerrointa voidaan yrittää ymmärtää regressiosuoran avulla ja tähän palataan myöhemmin.

Korrelaatiokerrointa laskettaessa ja käytettäessä on suotavaa tarkastella myös pisteparvea, koska silloin välttyy helpommin tulkinnallisilta virheiltä. Esimerkiksi epälineaarisen riippuvuuden ja poikkeushavainnot voi havaita melko helposti pisteparvesta.

Esim. 55 Lasketaan virvoitusjuomaesimerkin korrelaatiokertoimen arvo.

Henkilö	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	5	10	25	100	50
2	10	12	100	144	120
3	4	5	16	25	20
4	0	4	0	16	0
5	2	1	4	1	2
6	7	3	49	9	21
7	3	4	9	16	12
8	6	8	36	64	48
Yhteensä	37	47	239	375	273

$$r = \frac{273 - \frac{37 \cdot 47}{8}}{\sqrt{\left(239 - \frac{37^2}{8}\right)\left(375 - \frac{47^2}{8}\right)}} = 0.68$$

Korrelaatiokertoimen perusteella muuttujien välillä näyttää olevan positiivista lineaarista riippuvuutta.

Kun korrelaatiokertoimia tulkitaan, kannattaa pohtia seuraavia asioita:

1° Yksikin havainto saattaa muuttaa korrelaatiokertoimen arvoa paljon, jos tämän havainnon x- ja/tai y-arvot poikkeavat huomattavasti niiden tavanomaisista arvoista. Jos tällaiselle **poikkeushavainnolle** löytyy järkevä selitys (esim. mittausvirhe, tilastoyksikkö peräisin eri populaatiosta kuin muut tms.), voidaan se poistaa havaintoaineistosta.

2° Ryhmien yhdistämisen jälkeen laskettu korrelaatio saattaa poiketa huomattavasti yksittäisissä ryhmissä lasketuista korrelaatioista.

3° Vaihteluvälin pienentäminen aiheuttaa yleensä myös korrelaatiokertoimen itseisarvon pienenemisen.

4° Korrelaatiokertoimen suurikaan arvo ei takaa, että muuttujien välillä on jokin todellinen riippuvuussuhde. Taustalla voi olla jokin kolmas tekijä z, joka vaikuttaa sekä muuttujaan x että y. Tällaisen tekijän vaikutus yritetään usein poistaa laskemalla **osittaiskorrelaatioita**.

4.3. Järjestyskorrelaatio

Edellä esitetyn Pearsonin tulomomenttikorrelaatiokertoimen käyttöä rajoittaa kaksi seikkaa: sitä voidaan käyttää vain vähintään intervalliasteikollisten muuttujien tapauksessa ja toisaalta se mittaa vain lineaarista riippuvuutta. Jos molemman tarkasteltavat muuttujat ovat vähintään järjestysasteikollisia, voidaan riippuvuuden tutkimiseen käyttää erilaisia järjestyskorrelaatiokertoimia, esim. **Spearmanin järjestyskorrelaatiokerrointa** r_s .

Havainnon **järjestysluku** eli **sijaluku** $R(x_i)$ ilmaisee, monesko havainto x_i on suuruusjärjestyksessä. Jos kahdella tai useammalla tilastoyksiköllä on sama muuttujan arvo, sanotaan, että tilastoyksiköiden välillä on **sidos**. Näille havainnoille annetaan järjestyslukuksi niiden järjestyslukujen keskiarvo, mitkä nämä havainnot olisivat hieman erisuuruusina saaneet.

Tutkittaessa nyt kahta vähintään järjestysasteikollista muuttujaa x ja y saadaan kutakin tilastoyksikköä kohti järjestyslukupari $\{R(x_i), R(y_i)\}$. Muuttujien välistä riippuvuutta mittaa **järjestyslukujen erotukseen**

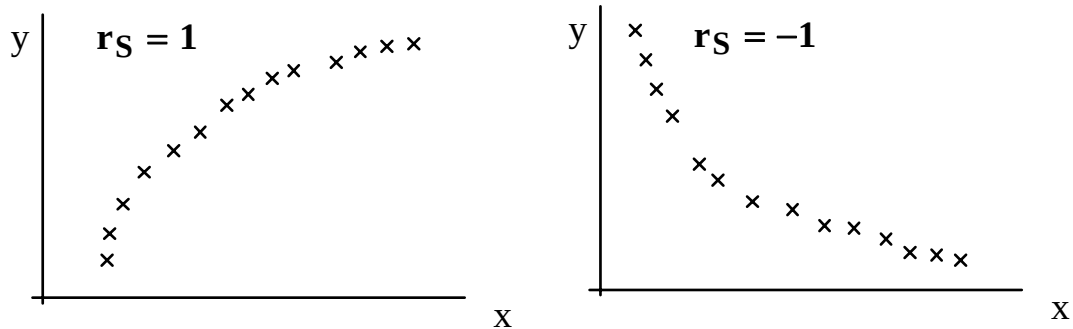
$$d_i = R(x_i) - R(y_i)$$

perustuva **Spearmanin järjestyskorrelaatiokerroin**

$$r_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n},$$

joka on itse asiassa järjestyslukupareille $\{R(x_i), R(y_i)\}$ määrätty Pearsonin tulomomentti-korrelaatiokerroin.

Esim. 56



Spearmanin järjestyskorrelaatio on arvoltaan aina välillä $[-1, +1]$, ja arvo on itseisarvoltaan sitä suurempi, mitä voimakkaammasta **monotonisesta** riippuvuudesta on kyse. Järjestyskorrelaatio mittaa kahden eri muuttujan arvojen järjestysten samanlaisuutta. Jos muuttujien arvojen järjestys on täysin sama, kertoimen arvo on $+1$, täysin päinvastainen järjestys antaa kertoimelle arvon -1 . Äärimmäiset arvot esiintyvät silloin, kun muuttujien välillä vallitsee täydellinen monotoninen riippuvuus. Kun muuttujien arvojen järjestys on toisiinsa nähden satunnainen, kertoimen arvo on nolla.

Esim. 57 Lasketaan Spearmanin järjestyskorrelaation arvo seuraavasta aineistosta, missä x kuvaa ennen koulutusjaksoa suoritettun testin arvosana ja y kuvaa koulutusjakson jälkeen suoritettun testin arvosana.

x_i	$R(x_i)$	y_i	$R(y_i)$	d_i	d_i^2
10	1	23	2.5	-1.5	2.25
11	2	27	5	-3	9
12	3	18	1	2	4
13	4	29	7	-3	9
15	5	23	2.5	2.5	6.25
16	6	25	4	2	4
17	7	28	6	1	1
18	8	31	8	0	0
					35.5

$$r_s = 1 - \frac{6 \cdot 35.5}{8^3 - 8} \approx 0.577$$

Tulosta voisi tulkita niin, että testien pistemäärillä näyttäisi olevan positiivinen riippuvuussuhde siten, että jos menestyi hyvin ennen koulutusjakson alkua pidetyssä testissä, menestyi hyvin myös koulutusjakson jälkeen pidetyssä testissä.

Huom. Jos aineistossa on paljon sidoksia, kannattaa Spearmanin järjestyskorrelaatio-kertoimen sijasta käyttää Kendallin τ -b -kerrointa, jonka tulkinta on samankaltainen kuin Spearmanin korrelaatiolla.

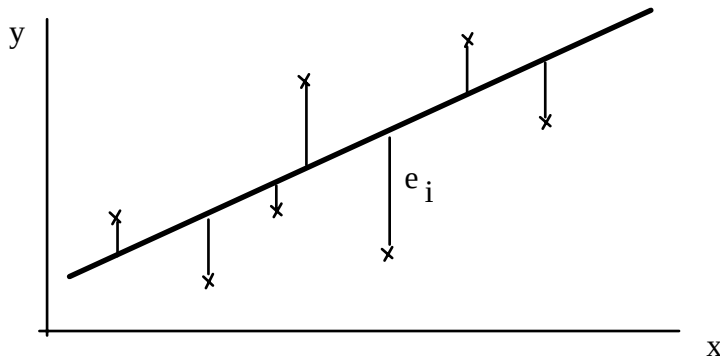
4.4. Regressio

Tilastollisten tarkastelujen yhtenä tavoitteena on pyrkiä kuvamaan ilmiöitä ja löytämään niiden välillä oleva mahdollinen **syy-seuraussuhde** siten, että jos tietyt ehdot ovat voimassa, niin niistä on tuloksena tiettyjä seuraamuksia.

Tilastollisen (epätäydellisen) syy-seuraussuhteen kuvaamiseen voidaan käyttää mm. **regressioanalyysiä**, jonka päämääränä onkin löytää muuttujien välillä mahdollisesti vallitseva yhteys ja kuvata sitä matemaattisen mallin avulla. Jos muuttujia on kaksi, sanotaan toista muuttujaa **selittäväksi** ja toista **selitettäväksi**. Selitettävää muuttujaa merkitään y :llä ja selittävää x :llä.

Tulomomenttikorrelaatiokerrointa tarkasteltaessa todettiin, että se mittaa lineaarista riippuvuutta. Jos muuttujaa x voidaan pitää syynä ja y :tä seurauksena, ja on havaittu, että muuttujien välillä vallitsee jonkinlainen lineaarinen riippuvuus, voidaan muuttujien x ja y välistä riippuvuutta kuvata suoralla. Tämä ns. **regressiosuora** $\hat{y} = b_0 + b_1x$ voidaan määrittää **pienimmän neliösumman menetelmällä**. PNS-menetelmän perusidea on se, että pistejoukkoon sovitetaan sellainen suora, joka kuvaa sitä mahdollisimman luontevasti. Tällöin suoran ylä- ja alapuolelle jää suunnilleen yhtä monta pistettä. Jokaista havaintoparia (x_i, y_i) kohden lasketaan sen pystyakselin suuntainen **etäisyys (jäännös, residuaali)** e_i mallina olevasta suorasta. Merkitään x_i :n arvolla laskettua suoran pistettä $(x_i, \hat{y}_i)$:llä. Nimensä mukaisesti PNS-menetelmässä haetaan ratkaisua lausekkeelle

$$\text{Min} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \text{Min} \sum_{i=1}^n e_i^2 .$$



Minimoinnin tuloksena saadaan regressiosuoran $\hat{y} = \mathbf{b}_0 + \mathbf{b}_1\mathbf{x}$ parametreille laskukaavat

$$\mathbf{b}_1 = \frac{s_{xy}}{s_x^2}$$

$$\mathbf{b}_0 = \bar{y} - \mathbf{b}_1\bar{x}.$$

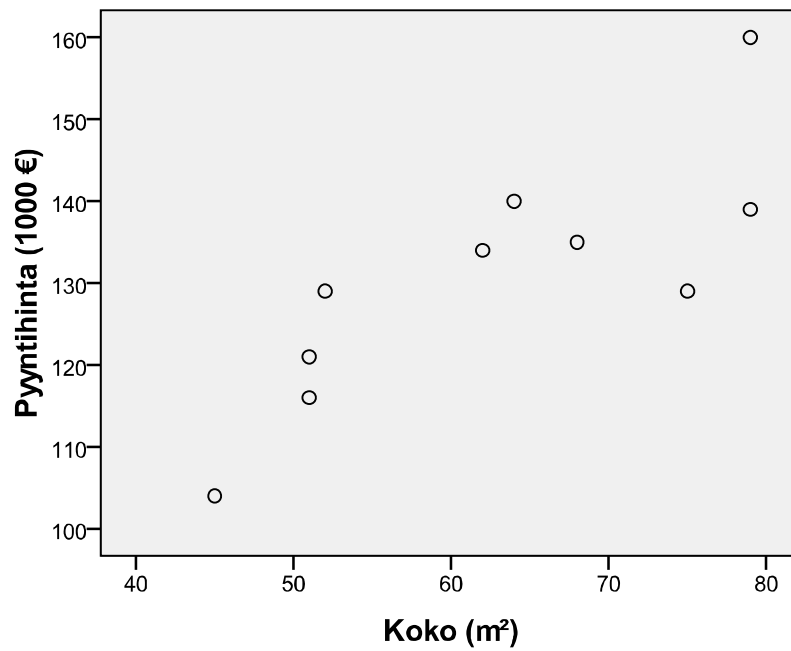
Vakioita $\mathbf{b}_0$ ja $\mathbf{b}_1$ sanotaan **regressiokertoimiksi**. Vakio b_0 ilmoittaa pisteen, jossa suora leikkaa y-akselin. Vakioilla b_0 on selkeä tulkinta vain silloin, kun muuttuja x voi saada arvon nolla. Vakio b_1 on regressiosuoran kulmakerroin. Vakio b_1 ilmaisee, kuinka paljon ja mihin suuntaan y :n arvo keskimäärin muuttuu, kun x kasvaa yhden yksikön verran.

Edellä olevia kaavoja voi käyttää, olipa x :n ja y :n arvot mitkä tahansa. Ennen laskemista on järkevää tarkastaa kuvion avulla, onko pisteparvea mielekästä kuvata suoralla.

Esim. 58 Elokuun 2009 viimeisellä viikolla Vaasan keskustassa myynnissä olevista vanhoista kerrostalokaksioista ja -kolmioista otetussa otoksessa olivat asuntojen koot (m^2) ja pyyntihinnan (1000 €) seuraavat:

Asunto:	1	2	3	4	5	6	7	8	9	10
Koko:	79	79	75	51	52	62	64	51	68	45
Hinta:	160	139	129	121	129	134	140	116	135	104

Muuttujien välinen pisteparvi on



Pisteet ovat ryhmittyneet melko suoranomaisesti, joten muodostetaan regressiomalli, jossa asunnon hinnan vaihtelua selitetään asunnon koon avulla.

Asunto	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	79	160	6241	25600	12640
2	79	139	6241	19321	10981
3	75	129	5625	16641	9675
4	51	121	2601	14641	6171
5	52	129	2704	16641	6708
6	62	134	3844	17956	8308
7	64	140	4096	19600	8960
8	51	116	2601	13456	5916
9	68	135	4624	18225	9180
10	45	104	2025	10816	4680
Yhteensä	626	1307	40602	172897	83219

$$\bar{x} = \frac{626}{10} = 62.6 \quad s_x^2 = \frac{1}{10-1} \left(40602 - \frac{626^2}{10} \right) = 157.155...$$

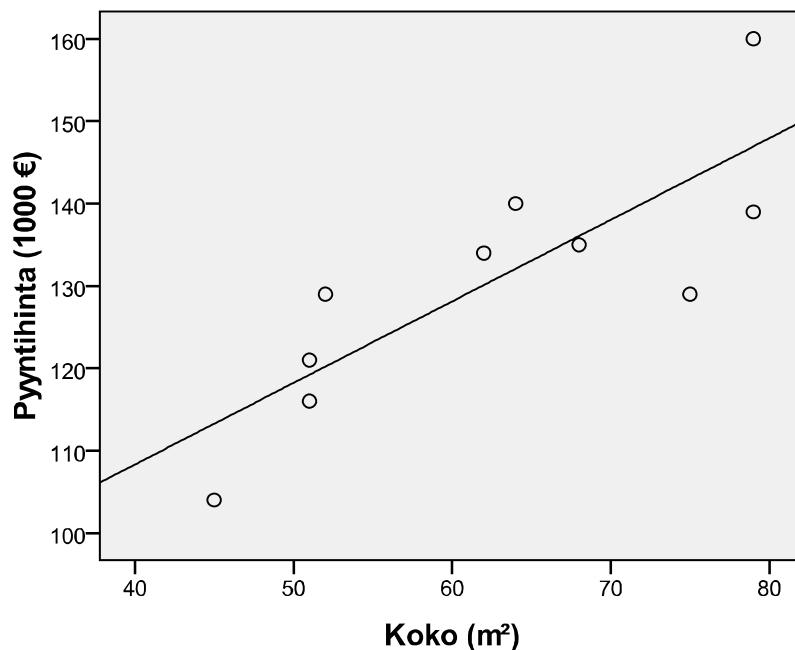
$$\bar{y} = \frac{1307}{10} = 130.7 \quad s_y^2 = \frac{1}{10-1} \left(172897 - \frac{1307^2}{10} \right) = 230.233...$$

$$s_{xy} = \frac{1}{10-1} \left(83219 - \frac{626 \cdot 1307}{10} \right) = 155.644...$$

$$b_1 = \frac{155.644...}{157.155...} = 0.990...$$

$$b_0 = 130.7 - 0.990... \cdot 62.6 = 68.701....$$

Suoran yhtälö on siis $\hat{y} = 68.7 + 0.990x$. Saatua mallia voidaan tulkita mm. siten, että yksi lisäneliö kerrostalohuoneiston koossa maksaa keskimäärin 0.99 tuhatta €. Piirretään lopuksi regressiosuora pistejoukkoon.



Yksi syy tilastollisen mallin käyttämiselle on se, että mallin avulla voidaan laatia ennusteita. Kun malliin sijoitetaan x :n paikalle tarkasteltava arvo, saadaan mallista lasketuksi tarkasteltavaa x :n arvoa vastaava y :n estimaatti eli ennuste y :n arvolle. Tätä mallin avulla laskettua **ennustetta** merkitään $\hat{y}$:lla. Ennusteen laatiminen on kuitenkin mahdollista vain silloin, kun ennusteeseen käytetty x :n arvo on suhteellisen lähellä muita x :n havaittuja arvoja.

Esim. 59 Jos asunnon koko on 60 m^2 , niin ennuste asunnon hinnalle saadaan sijoittamalla:

$$\hat{y} = 68.7 + 0.990 \cdot 60 = 128.1$$

eli huoneiston, jonka koko on 60 m^2 , keskimääräinen hinta on 128 tuhatta €.

Mallin hyvyys vaikuttaa osaltaan siihen, kuinka luotettavina mallin avulla laskettuja ennusteita voidaan pitää. Yksi tapa arvioida ennusteen luotettavuutta on laskea **selityskerroin** eli **selitysaste R^2** , joka mittaa mallin kykyä kuvata selitettävän muuttujan y vaihtelua. Suuri selityskertoimen arvo ei kuitenkaan yksin takaa tarkkoja ennusteita.

Yhden selittäjän lineaarisen regressiomallin tilanteessa selityskerroin saadaan korrelaatiokertoimen avulla

$$R^2 = r^2.$$

Usein selityskerroin ilmaistaan prosenttilukuna $100R^2$ %, jolloin se ilmaisee kuinka monta prosenttia selitettävän muuttujan y vaihtelusta voidaan selittää pelkästään selittävän muuttujan x avulla.

Esim. 60 Kerrostalohuoneistojen koon ja hinnan välinen korrelaatio on

$$r = \frac{155.644...}{\sqrt{157.155...} \cdot \sqrt{230.233...}} \approx 0.818$$

ja mallin selitysaste on siten

$$R^2 = 0.818^2 \approx 0.670,$$

joten asunnon koko selittää n. 67 % asunnon hinnan vaihtelusta

5. TODENNÄKÖISYYSLASKENTAA

Yksinkertaisten ilmiöiden todennäköisyyksiä on mahdollista arvioida tiettyjä sääntöjä noudattaen. **Satunnaisilmiöllä** tarkoitetaan sellaista ilmiötä, jonka tulosta ei varmasti tiedetä ennen kuin se on havainnoitu. Erilaisten tulosten esiintymistodennäköisyyksiä voidaan kuitenkin usein tarkastella jo ennen ilmiön havainnointia.

5.1. Kombinatoriikkaa

Todennäköisyyksien yhteydessä joudutaan usein laskemaan erilaisten tapahtumien lukumääriä. Erilaisten mahdollisuuksien lukumäärien laskemista koskevaa matematiikan aluetta sanotaan **kombinatoriikaksi**.

Joukon alkioden **permutaatioiksi** kutsutaan mitä tahansa kyseisten alkioden järjestettyä jonoa.

Esim. 61 Joukon $A = \{a, b, c\}$ kaikki kolmen alkion permutaatiot ovat:
 $(a, b, c), (a, c, b), (b, a, c), (b, c, a), (c, a, b), (c, b, a)$

Permutaatioita tarkasteltaessa oleellista on alkioden järjestys. Permutaatiot ovat samat vain jos niissä on samat alkiot samassa järjestyksessä.

Kombinaatio on sellainen alkioden joukko, jossa alkioden järjestyksellä ei ole merkitystä. Kombinaatiot ovat samoja, jos niissä on samat alkiot.

Esim. 62 (a, b, c) ja (b, a, c) ovat samoja kombinaatioita.

Usein kombinatoriikassa on kyse siitä, kuinka monella tavalla joukon alkiot voidaan järjestää, tai kuinka monella tavalla annetusta joukosta voidaan poimia tietyn kokoinen tietyn tyyppisiä alkioita sisältävä osajoukko. Näihin ongelmiin löytyy yleensä vastaus kerto- ja yhteenlaskuperiaatteen avulla. Seuraavassa esitettyjä kaavoja tarvitaan määrättäessä todennäköisyyksiä ns. symmetrisessä tapauksessa.

Kertolasku- eli tuloperiaate

Tarkastellaan koetta, joka voidaan toteuttaa n :ssä eri vaiheessa. Oletetaan, että

1. vaiheessa on k_1 tulosvaihtoehtoa
2. vaiheessa on k_2 tulosvaihtoehtoa
- ⋮
- n . vaiheessa on k_n tulosvaihtoehtoa

edellisen vaiheen tuloksista *riippumatta*. Tällöin koko kokeessa on $k_1 \cdot k_2 \cdot \dots \cdot k_n$ tulomahdollisuutta.

Huom. Jos $k_1 = k_2 = \dots = k_n = k$, niin tulomahdollisuuksia on k^n kpl.

Esim. 63 A:sta B:hen on 3 reittiä ja B:stä C:hen on 4 reittiä. Siten A:sta C:hen on $3 \cdot 4 = 12$ eri reittiä.

Yhteenlaskuperiaate

Oletetaan, että kokeessa on n toisensa *poissulkevaa* tulosvaihtoehtojoukkoa siten, että

1. joukossa on k_1 tulosvaihtoehtoa
2. joukossa on k_2 tulosvaihtoehtoa
- $\vdots$
- n . joukossa on k_n tulosvaihtoehtoa.

Nyt koko kokeessa on $k_1 + k_2 + \dots + k_n$ tulomahdollisuutta.

Esim. 64 A:sta pääsee B:hen 3 junalla ja 4 autolla. Ts. A:sta pääsee B:hen $3 + 4 = 7$ tavalla.

Esim. 65 Perinteinen 13 kohteen vakioveikkausrivi voidaan muodostaa $3^{13} = 1\,594\,323$ tavalla.

Esim. 66 Kuinka monta sellaista 13 kohteen veikkausriviä on, jossa on 12 oikein? Väärä kohde voi olla mikä tahansa 13 kohteesta ja se voi olla kahdella tavalla väärin. Siten erilaisia 12 oikein veikkausrivejä on $\underbrace{2 + 2 + \dots + 2}_{13 \text{ kpl}} = 26$ kpl.

Olkoon n positiivinen kokonaisluku. Luvun n **kertoma** $n! = n \cdot (n - 1) \cdot (n - 2) \dots 3 \cdot 2 \cdot 1$. Lisäksi $0! = 1$.

Esim. 67 $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$

Huom. Tuloperiaatteen perusteella n :n erilaisen alkion **permutaatioita** on $n!$ kappaletta.

Huom. Tuloperiaatetta voidaan soveltaa myös tilanteeseen, jossa n :n erilaisen alkion joukosta muodostetaan k :n alkion ($0 \leq k \leq n$) järjestettyjä jonoja, joita sanotaan **variaatioiksi** tai **k -permutaatioiksi**. Tällaisten jonojen lukumäärä on

$$n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot (n - (k - 1)) = \frac{n!}{(n - k)!}.$$

Esim. 68 Kuinka monta erilaista 2 kirjaimen jonoa voidaan muodostaa sanasta LAHTI?

$$\frac{5!}{(5-2)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 5 \cdot 4 = 20$$

Huom. Joukossa on n erilaista alkioita. Sen alkioista muodostettujen $k:n$ alkion **kombinaatioiden** ($0 \leq k \leq n$) lukumäärä on

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad ("n \text{ yli } k:n"). \text{ Lukuja } \binom{n}{k} \text{ sanotaan } \mathbf{binomikertoimiksi}.$$

Esim. 69 Erilaisia lottorivejä on

$$\binom{39}{7} = \frac{39!}{7!32!} = \frac{39 \cdot 38 \cdot 37 \cdot 36 \cdot 35 \cdot 34 \cdot 33}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 15\,380\,937 \text{ kpl}$$

5.2. Todennäköisyyden määrittely

Merkintöjä

Ω tai E	perusjoukko eli kaikkien tulostmahdollisuuksien eli alkeistapausten joukko; varma tapahtuma
$\emptyset$	tyhjä joukko; mahdoton tapahtuma
$A, B, C, \dots$	perusjoukon osajoukkoja; tapahtumat
$a, b, c, \dots$	alkioita eli alkeistapaukset; yksittäiset tulostmahdollisuudet
$A \cup B$	yhdiste; tapahtuu A tai B tai molemmat
$A \cap B$	leikkaus; sekä A että B tapahtuvat
A^C	A:n komplementti; A ei tapahdu
$A \cap B = \emptyset$	A ja B ovat erillisiä; A ja B eivät voi tapahtua yhtä aikaa

Klassinen eli teoreettinen todennäköisyys (a priori, ennalta annettu)

Tarkastellaan jotakin satunnaiskoetta tai -ilmiötä (esim. rahan heitto, pallojen nostaminen urnasta jne.), jonka tulostmahdollisuuksia eli **alkeistapauksia** on **äärellinen määrä n** . Oletetaan vielä, että kaikki **tulostmahdollisuudet** ovat **symmetrisiä** eli kukin alkeistapaus esiintyy yhtä usein. Jos tapahtumaan A johtaa näistä alkeistapauksista k kappaletta, määritellään tapahtuman A todennäköisyydeksi

$$P(A) = \frac{k}{n} = \frac{\text{A : lle suotuisten alkeistapausten lkm}}{\text{kaikkien alkeistapausten lkm}}.$$

Esim. 70 Nopan heitto:

Perusjoukko: $\Omega = \{1, 2, 3, 4, 5, 6\}$,

Olkoot tapahtumat: $A = \{2\}$ ja $B = \{1, 3, 5\}$

$$P(A) = \frac{1}{6} \quad \text{ja} \quad P(B) = \frac{3}{6} = \frac{1}{2}$$

Klassisen todennäköisyyskäsitteen soveltamisala on suppea: sen käyttöä rajoittaa ennen muuta alkeistapausten symmetrisyyden vaatimus. Klassista todennäköisyyskäsitettä voidaan soveltaa lähinnä peliteoriaan. Käytännön ongelmien ratkaisemiseen tarvitaankin yleisempää todennäköisyyden käsitettä.

Empiirinen eli tilastollinen todennäköisyys (a posteriori, kokemukseen perustuva)

Frekvenssiteoria:

Oletetaan, että jokin satunnaisilmiö tai -koe voi tapahtua **mielivaltaisen monta kertaa samanlaisissa olosuhteissa**. Lisäksi oletetaan, että ilmiöstä tai kokeesta voidaan havaita erilaisia tapahtumia. Jos koe toistuu n kertaa, ja havaitaan, että tapahtumaan A johtaa näistä f_A kappaletta, saadaan **A :n suhteelliseksi osuudeksi** eli **suhteelliseksi frekvenssiksi**

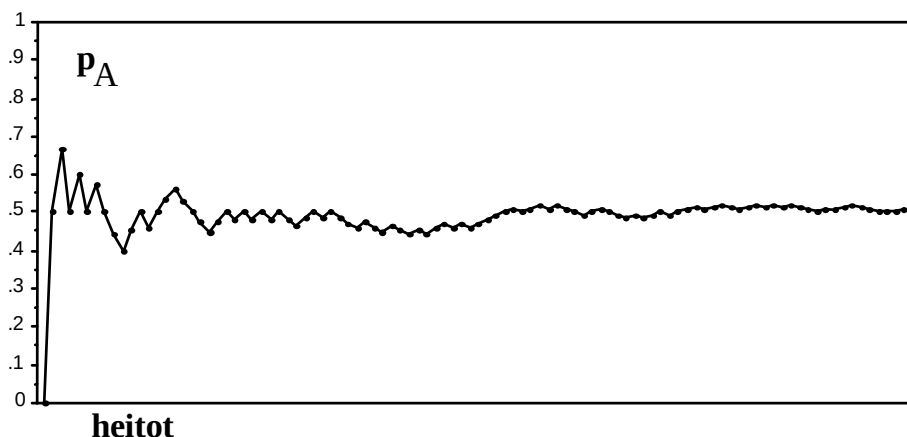
$$p_A = \frac{f_A}{n}.$$

Pitkässä koesarjassa (= kun n on iso) p_A antaa hyvän likiarvon tapahtuman A todennäköisyydelle, siis $p_A \rightarrow P(A)$.

Tilastollisen todennäköisyyden käsite on teoreettinen. Käytännössä suhteellisen frekvenssin vaihtelu tasoittuu kokeen tai ilmiön toistuessa riittävän monta kertaa. Tällä tavoin saadaan riittävän tarkka likiarvo tilastolliselle todennäköisyydelle.

Esim. 71 Sadassa nastan pudotuksessa saatiin 51 kertaa tulos " piikki ylöspäin", joten

$$P(\text{"piikki ylöspäin"}) \approx \frac{51}{100} = 0.51$$



Todennäköisyyden ominaisuuksia

Todennäköisyyden teoreettinen määrittely tapahtuu aksiomaattisesti, ts. todennäköisyydelle asetetaan perusvaatimukset eli aksioomat. **Aksioomajärjestelmän** (1° - 3°) esitti Kolmogorov v. 1933.

- 1° $0 \leq P(A) \leq 1$
- 2° Jos $A \cap B = \emptyset$, niin $P(A \cup B) = P(A) + P(B)$
(ns. erillisten tapahtumien yhteenlaskusääntö)
- 3° $P(\Omega) = 1$
- 4° $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (yhteenlaskusääntö)
- 5° $P(A^C) = 1 - P(A)$
- 6° $P(\emptyset) = 0$
- 7° Jos $A \subseteq B$, niin $P(A) \leq P(B)$

Esim. 72 Todennäköisyys saada veikkauksessa ainakin 12 oikein on

$$P(12 \text{ tai } 13 \text{ oikein}) = P(12 \text{ oikein}) + P(13 \text{ oikein}) = \frac{26}{313} + \frac{1}{313} = 0.000016935$$

5.3. Ehdollinen todennäköisyys ja riippumattomuus

Todennäköisyyslaskennassa tapahtumien vaikutus toisiinsa näkyy siten, että tieto jonkin tapahtuman esiintymisestä muuttaa siitä riippuvien tapahtumien esiintymistodennäköisyyttä. Kahden tapahtuman A ja B **vaikutus toisiinsa** ilmenee mm. **ehdollisen todennäköisyyden kautta**.

Esim. 73 Tarkastellaan kahden rahan heittoa, jolloin $\Omega = \{(kr,kr), (kr,kl), (kl,kr), (kl,kl)\}$

Olkoot tapahtumat

$$A = \{\text{molemmat samat}\} = \{(kr,kr), (kl,kl)\}; \quad P(A) = \frac{1}{2}$$

$$B = \{\text{ainakin yksi kr}\} = \{(kr,kr), (kr,kl), (kl,kr)\}; \quad P(B) = \frac{3}{4}$$

Todennäköisyys sille, että A tapahtuu, jos B on jo tapahtunut, on $\frac{1}{3}$, koska (kl,kl) ei voi tapahtua ja (kr,kr) voi, ja (kr,kr) on yksi B:n kolmesta alkeistapauksesta.

Ehdollisella todennäköisyydellä $P(A \mid B)$ tarkoitetaan todennäköisyyttä

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0 \quad ("A:n \text{ todennäköisyys ehdolla } B")$$

Ehdollisen todennäköisyyden määrittelevä kaava voidaan kirjoittaa myös **yleisen kertolaskusäännön** muodossa:

$$P(A \cap B) = P(B) P(A \mid B)$$

tai

$$P(A \cap B) = P(A) P(B \mid A).$$

Esim. 74 Mikä on todennäköisyys saada korttipakasta ensin ässä ja sitten punainen kuningas?

Merk. $A = \{1. \text{ kortti ässä} \}$ ja $B = \{2. \text{ kortti punainen kuningas} \}$.

$$P(A \cap B) = P(A) P(B \mid A) = \frac{4}{52} \cdot \frac{2}{51} = \frac{2}{663}$$

Tapahtumat **A** ja **B** ovat **tilastollisesti riippumattomia**, jos

$$P(A \mid B) = P(A) \quad \text{tai} \quad P(B \mid A) = P(B).$$

Tällöin **kertolaskusääntö** voidaan kirjoittaa muodossa:

$$P(A \cap B) = P(A) P(B)$$

Esim. 75 Heitetään kahta noppaa peräkkäin, ja ensimmäisen heiton tulos ei vaikuta toisen heiton tulokseen. Kahden nopan heitossa on 36 tulosvaihtoehtoa.

Havainnollistetaan silmälukujen summaa ”noppakoordinaatiostossa”:

2. noppa	6	7	8	9	10	11	12
	5	6	7	8	9	10	11
	4	5	6	7	8	9	10
	3	4	5	6	7	8	9
	2	3	4	5	6	7	8
	1	2	3	4	5	6	7
		1	2	3	4	5	6
							1. noppa

Tarkastellaan tapahtumaa ”kahden nopan silmälukujen summa on 5”. Ko. tapahtuman todennäköisyys voidaan laskea esimerkiksi käyttäen em. laskusääntöjä:

$$\begin{aligned}
 P(\text{summa on } 5) &= P((1. \text{ noppa} = 1 \text{ ja } 2. \text{ noppa} = 4) \\
 &\quad \text{tai } (1. \text{ noppa} = 2 \text{ ja } 2. \text{ noppa} = 3) \\
 &\quad \text{tai } (1. \text{ noppa} = 3 \text{ ja } 2. \text{ noppa} = 2) \\
 &\quad \text{tai } (1. \text{ noppa} = 4 \text{ ja } 2. \text{ noppa} = 1)) \\
 &= P(1. \text{ noppa} = 1) \cdot P(2. \text{ noppa} = 4) \\
 &\quad + P(1. \text{ noppa} = 2) \cdot P(2. \text{ noppa} = 3) \\
 &\quad + P(1. \text{ noppa} = 3) \cdot P(2. \text{ noppa} = 2) \\
 &\quad + P(1. \text{ noppa} = 4) \cdot P(2. \text{ noppa} = 1) \\
 &= \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} \\
 &= \frac{4}{36}
 \end{aligned}$$

Esim. 76 Isäntä ja renki ampuvat yhteislaukauksella naapurin lehmän hirvenä. Syntyy väittely ampujasta. Olkoot

$$\begin{aligned}
 I &= \{\text{isäntä osuu}\} & R &= \{\text{renki osuu}\} & Y &= \{\text{lehmässä yksi osuma}\} \\
 P(I) &= 0.8 & P(R) &= 0.6
 \end{aligned}$$

$$\begin{aligned}
 P(Y) &= P[(I \cap R^C) \cup (I^C \cap R)] = P(I \cap R^C) + P(I^C \cap R) \\
 &= P(I) P(R^C) + P(I^C) P(R) = 0.8 \cdot 0.4 + 0.2 \cdot 0.6 = 0.44
 \end{aligned}$$

$$P(I \cap Y) = P(I \cap R^C) = P(I) P(R^C) = 0.8 \cdot 0.4 = 0.32$$

$$\text{Nyt } P(I | Y) = \frac{0.32}{0.44} = 0.727 \quad \text{ja} \quad P(R | Y) = 1 - P(I | Y) = 0.273$$

6. TEOREETTISISTA JAKAUMISTA

6.1. Satunnaismuuttujista

Satunnaismuuttujat, merk. $\underline{x}$, $\underline{y}$, ... ovat empiiristen muuttujien matemaattisia vastineita. Käytännössä satunnaismuuttujalla tarkoitetaan jotakin satunnaiskokeen tulokseen liittyvää numeerista suuretta, jonka arvo vaihtelee koetta toistettaessa. Satunnaismuuttuja on satunnainen siinä mielessä, että etukäteen ei voida varmasti tietää, mikä ilmiön tuloksista realisoituu.

Esim. 77 Nopanheitossa saatava silmäluku on satunnaismuuttuja.
Vuosittaisten liikenneonnettomuuksien lkm on satunnaismuuttuja.

Matemaattisesti määriteltynä satunnaismuuttuja $\underline{x}$ on kuvaus (funktio) alkeistapausten joukolta reaalilukujoukolle. Itse funktio ei ole satunnainen, vaan se on täysin määrätty. Satunnaismuuttujan $\underline{x}$ arvo x voidaan määrätä vasta, kun satunnaiskokeen tulos on selvillä.

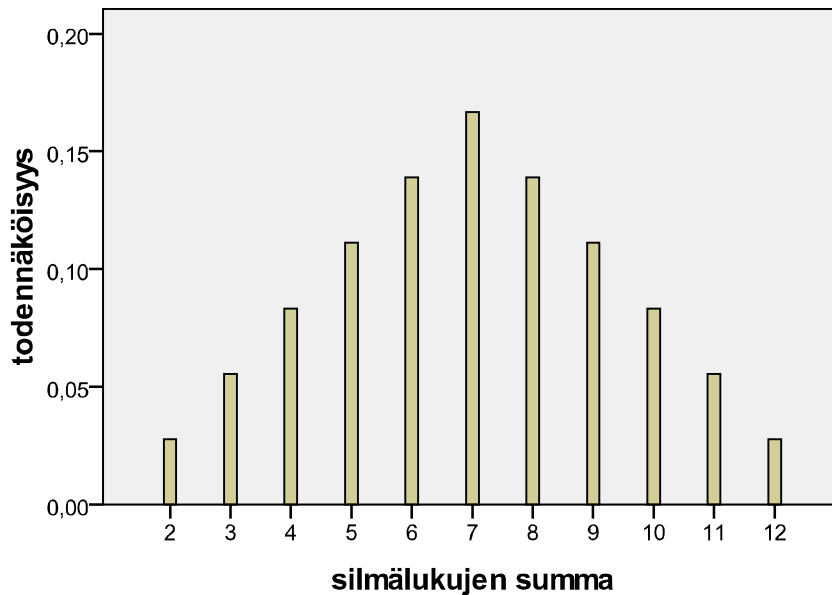
Satunnaismuuttuja on **epäjatkuva** eli **diskreetti**, jos se voi saada vain tiettyjä arvoja. Vastaavasti **jatkuva** satunnaismuuttuja voi saada minkä arvon tahansa tietyltä reaalilukuväliltä.

6.2. Diskreeteistä teoreettisista jakaumista

Diskreetin satunnaismuuttujan jakauman eli ns. **diskreetin todennäköisyysjakauman** muodostavat satunnaismuuttujan $\underline{x}$ arvot x_i , joita on k kpl, sekä arvojen todennäköisyydet p_i , joille on voimassa, että $p_1 + p_2 + \dots + p_k = 1$.

Esim. 78 Kahden nopan heitossa silmälukujen summa on diskreetti satunnaismuuttuja. Eri tuloksiin johtavien alkeistapausten lukumäärien perusteella voidaan laskea eri tuloksia vastaavat todennäköisyydet ja esittää todennäköisyysjakauma taulukkona ja kuvionakin.

Summa	2	3	4	5	6	7	8	9	10	11	12
Toden- näköisyys	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Edellä suurin todennäköisyys osuus summalle 7. Se saadaan tulokseksi keskimäärin joka kuudennella heitolla, kun muut arvot esiintyvät sitä harvemmin. On siis perusteltua käyttää lukua 7 silmälukujen summan ennusteena, kun heitetään kahta noppaa.

Toinen käytetty **ennuste** on satunnaismuuttujan **odotusarvo**, joka on satunnaismuuttujan saamien arvojen keskimääräinen taso pitkässä koesarjassa, jossa toistot ovat toisistaan riippumattomia. Distreetin satunnaismuuttujan odotusarvo $E(\underline{x})$ (, joka siis vastaa aritmeettista keskiarvoa) määritellään seuraavasti:

$$E(\underline{x}) = \sum_{i=1}^k p_i x_i = p_1 x_1 + p_2 x_2 + \dots + p_k x_k.$$

Esim. 79 Silmälukujen summan odotusarvo on

$$E(\text{silmälukujen summa}) = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \dots + \frac{1}{36} \cdot 12 = 7$$

Huom. Odotusarvo ei ole aina satunnaismuuttujan mahdollinen arvo, vaan se ilmaisee arvon, jonka läheisyydessä keskimäärin ollaan.

Odotusarvon sopivuuteen ennusteena vaikuttaa tarkasteltavan satunnaismuuttujan arvojen hajaantuminen, jota voidaan mitata **varianssilla** $\text{Var}(\underline{x})$, joka määritellään diskreetille satunnaismuuttujalle seuraavasti:

$$\begin{aligned} \text{Var}(\underline{x}) &= \sum_{i=1}^k p_i (x_i - E(\underline{x}))^2 \\ &= p_1 (x_1 - E(\underline{x}))^2 + p_2 (x_2 - E(\underline{x}))^2 + \dots + p_k (x_k - E(\underline{x}))^2 \end{aligned}$$

Jos satunnaismuuttujan varianssi on pieni, on satunnaisuutta vähän, ja satunnaismuuttujan jakauma on muodoltaan kapea. Jos taas varianssi on suuri, on satunnaisuutta paljon, ja jakauma on muodoltaan leveä.

Esim. 80 Kahden nopan heiton silmälukujen summan varianssi on

$$\text{Var}(\text{summa}) = \frac{1}{36} \cdot (2 - 7)^2 + \frac{2}{36} \cdot (3 - 7)^2 + \frac{3}{36} \cdot (4 - 7)^2 + \dots + \frac{1}{36} \cdot (12 - 7)^2 \\ \approx 5.8$$

Huom. Satunnaismuuttujan **keskihajonta** $D(\underline{x}) = \sqrt{\text{Var}(\underline{x})}$.

6.3. Jatkuvista teoreettisista jakaumista

Jos satunnaismuuttuja on jatkuva, ei sen yksittäisiä arvoja voi luetella. Jatkuvalle satunnaismuuttujalle on ominaista se, että kaikki **pistetodennäköisyydet** ovat **nollia**, ts. $P(\underline{x} = a) = 0$. Jakauman muotoa voidaan kuitenkin kuvata ns. **tiheysfunktion**, merk. **f(x)**, avulla. Tiheysfunktiota voidaan pitää frekvenssimonikulmion "raja-arvona". Tiheysfunktio on aina arvoltaan ei-negatiivinen, ja tiheysfunktion kuvaajan ja vaakakselin väliin jäävän alueen kokonaispinta-ala on suuruudeltaan 1. Jatkuvan satunnaismuuttujan tiheysfunktio ilmaisee jakauman muodon (kuvaajan), ei todennäköisyyksiä!

Myös jatkuvalle satunnaismuuttujalle $\underline{x}$ voidaan määrittää odotusarvo ja varianssi määrätyn integraalin avulla.

Erityisesti jatkuvien satunnaismuuttujien todennäköisyystarkasteluissa hyödynnetään **kertymäfunktiota** **F(x)**, joka määritellään seuraavasti:

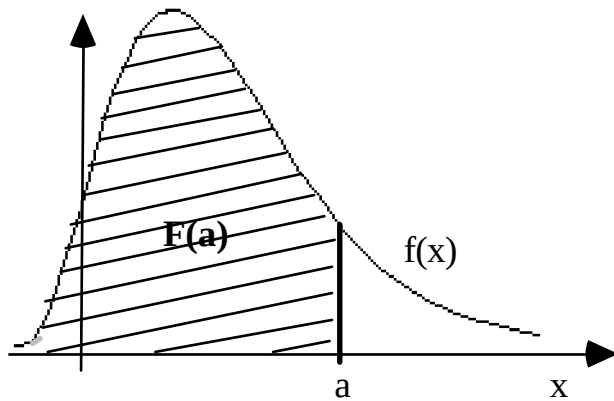
$$F(x) = P(\underline{x} \leq x).$$

F(x) ilmaisee siis todennäköisyyden, että satunnaismuuttuja $\underline{x}$ saa enintään arvon x. Kertymäfunktio voidaan määrittää niin diskreeteille kuin jatkuvillekin satunnaismuuttujille. Diskreetin satunnaismuuttujan kertymäfunktio määritellään yhteenlaskun avulla. Jatkuvan satunnaismuuttujan kertymäfunktio määritetään määrätyn integraalin avulla.

Kertymäfunktion ominaisuuksia

- 1° $F(-\infty) = 0, \quad F(\infty) = 1, \quad F(x) \leq F(y), \text{ jos } x < y.$
- 2° $P(a < \underline{x} \leq b) = P(\underline{x} \leq b) - P(\underline{x} \leq a) = F(b) - F(a)$
- 3° Jos $\underline{x}$ on **jatkuva**, niin $P(\underline{x} = a) = 0$ ja tällöin $F(a) = P(\underline{x} \leq a) = P(\underline{x} < a)$
- 4° $P(\underline{x} > a) = 1 - P(\underline{x} \leq a) = 1 - F(a)$

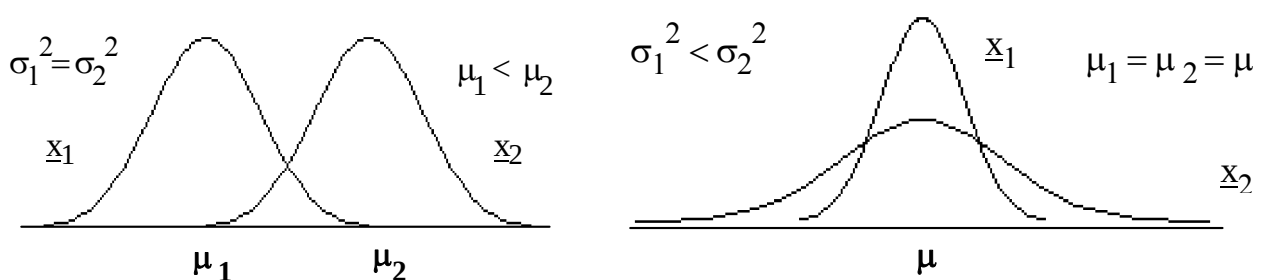
Huom. Jatkuvan satunnaismuuttujan kertymäfunktioita voidaan myös tulkita pinta-alana:



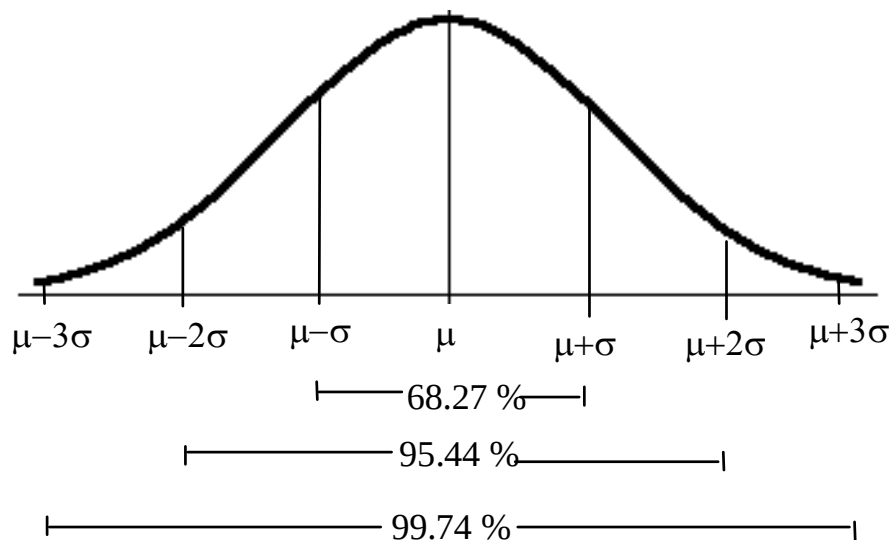
Tilastotieteen kannalta tärkein jatkuva jakauma on **normaalijakauma**. Jos satunnaismuuttuja $\underline{x}$ noudattaa normaalijakaumaa, jonka odotusarvona $E(\underline{x}) = \mu$ ja varianssina $\text{Var}(\underline{x}) = \sigma^2$, käytetään merkintää $\underline{x} \sim N(\mu, \sigma^2)$. Tällöin satunnaismuuttujan $\underline{x}$ tiheysfunktio on muotoa

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Normaalijakauman tiheysfunktion kuvaaja on symmetrinen odotusarvon μ suhteen. Tiheysfunktion $f(x)$ kuvaaja on ns. Gaussin kellokäyrä, joka lähestyy x-akselia μ :n molemmilla puolilla. Tiheysfunktio on yksihiippuinen, joten symmetrisyydestä seuraa, että $\mu = M_o = M_d$. Kuvaajan sijainti määräytyy odotusarvon μ mukaan ja muoto varianssin σ^2 mukaan:



Jatkuvissa jakaumissa tapahtumien todennäköisyydet voidaan määrittää graafisesti pinta-aloina. Kokonaiset keskihajonnat jakavat normaalijakauman seuraavasti:

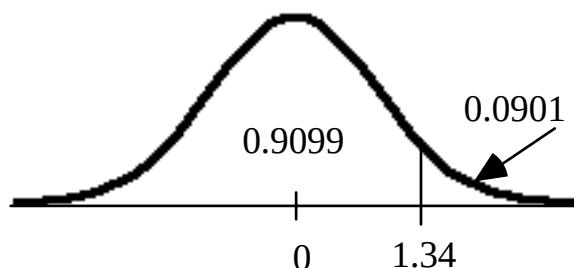


Normaalijakauman keskeinen asema tilastotieteessä perustuu siihen, että useat empiiriset jakaumat ovat likimain normaalisia. Keskeisen raja-arvolauseen mukaan mielivaltaisten riippumattomien satunnaismuuttujien summan jakauma on likimain normaalin, jos yhteenlaskettavien lukumäärä on suuri, ja eräät hyvin yleisluontoiset ehdot ovat voimassa. Käytännön kannalta on merkittävää, että usein jo pienillä n :n arvoilla ($n \geq 30$) päästään normaalijakaumalla hyvään approksimaatioon.

Standardoitu normaalijakauma

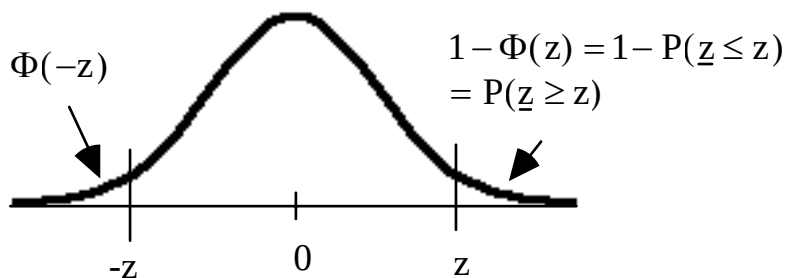
Normaalijakauman tärkein erikoistapaus on ns. **standardoitu normaalijakauma**. Nyt $\mu = 0$ ja $\sigma^2 = 1$, jolloin merk. $\underline{z} \sim N(0, 1)$. Satunnaismuuttujan $\underline{z}$ kertymäfunktion Φ arvoja on taulukoitu, koska tiheysfunktioilla ei ole integraalifunktiota alkeisfunktioiden joukossa.

Esim. 81 $P(\underline{z} \leq 1.34) = \Phi(1.34) = 0.9099$



Useimmissa taulukoissa on kertymäfunktion arvoja esitetty vain z :n positiivisilla arvoilla, joten negatiivisilla arvoilla nojaututaan normaalijakauman symmetrisyyteen ($z > 0$):

$$P(\underline{z} \leq -z) = \Phi(-z) = 1 - \Phi(z).$$



Esim. 82 Jos $\underline{z} \sim N(0, 1)$, niin $P(\underline{z} \leq -1) = \Phi(-1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$

$N(0, 1)$ -jakauman käyttökelpoisuus perustuu seuraavaan tulokseen:

Jos $\underline{x} \sim N(\mu, \sigma^2)$, niin muunnoksella

$$\underline{z} = \frac{\underline{x} - \mu}{\sigma},$$

jota kutsutaan **standardoinniksi**, saadaan $E(\underline{z}) = 0$ ja $\text{Var}(\underline{z}) = 1$. Tällöin $\underline{z} \sim N(0, 1)$.

Siten $P(\underline{x} \leq x) = P(\underline{x} - \mu \leq x - \mu) = P\left(\frac{\underline{x} - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(\underline{z} \leq z) = \Phi(z)$.

Esim. 83 Olkoot $\underline{x} \sim N(3, 4) = N(3, 2^2)$

$$\begin{aligned} \text{a) } P(\underline{x} \leq 0) &= P\left(\frac{\underline{x} - 3}{2} \leq \frac{0 - 3}{2}\right) = P(\underline{z} \leq -1.5) = \Phi(-1.5) = 1 - \Phi(1.5) = 1 - 0.9332 \\ &= 0.0668 \end{aligned}$$

$$\begin{aligned} \text{b) } P(\underline{x} > 1) &= 1 - P(\underline{x} \leq 1) = 1 - P(\underline{z} \leq -1) = 1 - \Phi(-1) = 1 - [1 - \Phi(1)] = \Phi(1) \\ &= 0.8413 \end{aligned}$$

$$\text{c) } P(\underline{x} \leq a) = 0.8, \quad a = ?$$

$$0.8 = P(\underline{x} \leq a) = P\left(\underline{z} \leq \frac{a - 3}{2}\right) = \Phi\left(\frac{a - 3}{2}\right)$$

$$\text{Taulukosta : } \Phi(0.84) = 0.7995 \approx 0.8. \quad \text{Siten } \frac{a - 3}{2} = 0.84 \Leftrightarrow a = 4.68$$

Normaalijakaumasta johdettuja jakaumia

χ^2 -jakauma

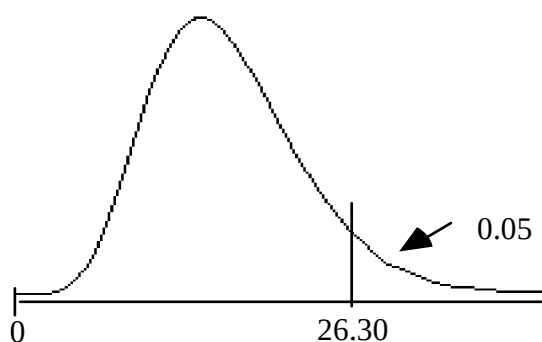
Olkoot $\underline{z}_1, \dots, \underline{z}_n$ $N(0, 1)$ -jakautuneita riippumattomia satunnaismuuttujia. Silloin satunnaismuuttuja

$$\chi^2_n = \chi^2 = \sum_{i=1}^n z_i^2$$

noudattaa χ^2 -jakaumaa, jonka muoto riippuu **vapausasteluvusta** n , merk. $\chi^2 \sim \chi^2(n)$. Odotusarvona on $E(\chi^2) = n$ ja varianssina $\text{Var}(\chi^2) = 2n$. Jakauman häntätodennäköisyyksiä on taulukoitu.

χ^2 -jakaumaa voidaan käyttää riippumattomuus- ja yhteensopivuustesteissä.

Esim. 84 Jos $\chi^2 \sim \chi^2(16)$, niin $P(\chi^2 \geq 26.30) = 0.05$



Studentin t -jakauma

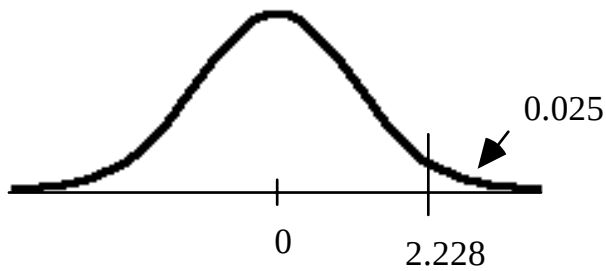
Olkoot $\underline{x}$, z_1 , ..., z_n $N(0, 1)$ -jakautuneita riippumattomia satunnaismuuttujia. Silloin satunnaismuuttuja

$$t_n = t = \frac{\underline{x}}{\sqrt{\frac{1}{n} \sum_{i=1}^n z_i^2}}$$

on **Studentin t-jakautunut** vapausasteella n , merk. $t \sim t(n)$. Odotusarvona on $E(t) = 0$ ja varianssina $\text{Var}(t) = \frac{n}{n-2}$. Myös t-jakauman häntätodennäköisyyksiä on taulukoitu. Kun $n \rightarrow \infty$, t-jakauma lähestyy $N(0, 1)$ -jakaumaa.

Jakauman käyttö tulee jatkossa esiin luottamusvälien, keskiarvotestien sekä korrelaatiokertoimen testauksen yhteydessä.

Esim. 85 Jos $t \sim t(10)$, niin $P(t \geq 2.228) = 0.025$



Fisherin F- jakauma

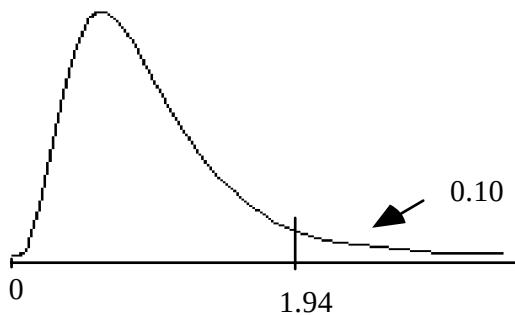
Olkoot χ^2_m ja χ^2_n riippumattomia χ^2 -jakautuneita satunnaismuuttujia vapausasteina m ja n . Silloin satunnaismuuttuja

$$F = \frac{\chi^2_m / m}{\chi^2_n / n}$$

noudattaa **Fisherin F-jakaumaa** vapausastein m ja n , merk. $F \sim F(m, n)$.

F-jakaumaa voidaan käyttää mm. varianssitesteissä.

Esim. 86 Jos $F \sim F(10, 20)$, niin $P(F \geq 1.94) = 0.10$



7. TILASTOLLISESTA PÄÄTTELYSTÄ

Tähän asti olemme käsitelleet menetelmiä, joiden avulla havaintoaineiston sisältämää informaatiota kuvataan käyttäen erilaisia kuvioita sekä tilastollisia tunnuslukuja. Kuvaileva tilastotiede ei ole aina riittävää empiirisessä tutkimuksessa, vaan usein yritetään esimerkiksi selvittää, onko havaintoaineistosta (=otoksesta) laskettu tunnusluku arvoltaan lähellä todellista arvoa, joka saataisiin, jos tutkittaisiin koko populaatio. Havaintoaineiston perusteella yritetään tehdä päätelmiä populaation tilanteesta, eli otoksesta lasketun tunnusluvun avulla estimoidaan (=arvioidaan) populaation parametrin (= tunnusluvun) suuruutta. Päätöksenteossa käytetään apuna todennäköisyyslaskentaa, jonka avulla pystytään arvioimaan mm. millä todennäköisyydellä otoksesta lasketun tunnusluvun arvo saadaan, ja kuinka suuria poikkeamien voidaan olettaa olevan. Tilastollinen päättely yhdistää siis empirian ja todennäköisyyslaskennan.

7.1. Estimoinnista

Otoksesta lasketun tunnusluvun avulla tapahtuvaa perusjoukon **tunnusluvun** eli **parametrin** (esim. μ , σ^2 , Π , ρ) arviointia sanotaan **estimoinniksi**.

Piste-estimoinnilla määrätään yksi ainoa luku, jolla parametria arvioidaan. **Väliestimoinnilla** määrätään **luottamusväli**, joka suurella varmuudella (yleensä 95%, 99%) sisältää todellisen parametrin arvon.

Populaation estimoitavaa tunnuslukua merkitään yleensä θ :lla (theetta) ja $\hat{\theta}$:llä sellaista otoksen havaintojen lauseketta, jolla θ :aa arvioidaan. $\hat{\theta}$:n **laskukaavaa** sanotaan θ :n **estimaattoriksi** ja tiettyä otoksesta **laskettua arvoa** θ :n **estimaatiksi**.

Esim.87 Populaation keskiarvon μ estimaattoriksi sopii otoksesta laskettu aritmeettinen keskiarvo eli ns. otoskeskiarvo. Otoksesta laskettu arvo esim. $\bar{x} = 5$ on μ :n estimaatti.

Jos samasta populaatiosta valitaan useita otoksia havaintoaineistojen muodostamiseksi, on varsin uskottavaa, että esimerkiksi saman muuttujan keskiarvoluvut vaihtelevat otoksittain. Uskottavalta tuntuu myös se, että ison otoksen avulla voidaan paremmin kuvata populaatiota, ja silloin otoskeskiarvo on lähellä populaation keskiarvoa. Eo. yleisesti muotoiltuihin ideoihin on olemassa selvät säännöt, joiden käsittelyyn tarvitaan **otantajakauman** (otosjakauman) käsitettä. Käytännössä otantajakaumaa ei tunneta, mutta teoriassa se voidaan tarkoin määrittää. Esimerkiksi keskiarvon otantajakauman keskikohta on sama kuin populaation keskiarvo. Eri otoksista laskettujen keskiarvojen vaihtelu on pienempää kuin otosten muuttuja-arvojen vaihtelu, joten keskiarvon otantajakauma on muodoltaan korkeampi ja kapeampi kuin otoksen havaintojen jakauma.

Esim. 88 Keskeisin raja-arvolauseen perusteella otoskeskiarvon jakauma on (ainakin likimain) normaalijakauma. Otoksesta määritetyn tietyn tyyppisten alkioden prosenttiosuuden jakauma on likimain normaalijakauma.

Väliestimointi (luottamusvälit)

Väliestimoinnilla määrätään reaalilukuväli, joka suurella varmuudella (yleensä 95 %, 99 %) sisältää todellisen populaatioparametrin arvon. Satunnaisväliä (t_n , T_n) sanotaan parametrin θ **luottamusväliksi luottamustasolla $1 - \alpha$** ($0 < \alpha < 1$), jos

$$P(t_n \leq \theta \leq T_n) = 1 - \alpha.$$

Siten todennäköisyydellä $1 - \alpha$ väli $[t_n, T_n]$ peittää perusjoukon parametrin todellisen arvon. Välin alaraja t_n on jokin havaintojen funktio, ja samoin välin yläraja T_n on jokin toinen havaintojen funktio. Yleensä luottamusväli yritetään pitää mahdollisimman kapeana ja silloin luottamusväli useimmiten muodostetaan niin, että valitaan "hyvä" θ :n estimaattori ja muodostetaan väli, jossa kyseinen estimaattori on keskipisteenä. Jotta luottamusväli voidaan muodostaa, on estimaattorin otantajakauma tunnettava.

1° Populaation keskiarvon eli odotusarvon μ luottamusväli

(a) Oletetaan ensin, että **populaation varianssi σ^2** on **tunnettu** (esim. aikaisempi tutkimus, pitkä seuranta-aika), ja että havainnot $\underline{x}_1, \dots, \underline{x}_n$ ovat otos normaalijakaumasta $N(\mu, \sigma^2)$. Kun havainnoista lasketaan (otos)keskiarvo, tiedetään sen kuvaavan hyvin populaation keskiarvoa. Keskiarvojen arvot vaihtelevat eri otoksissa. Tätä vaihtelua kuvaa keskiarvon keskihajonta eli **keskiarvon keskivirhe**

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

joten otoskeskiarvo $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ ja edelleen standardoimalla saadaan

$$\underline{z} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).$$

Standardoidussa normaalijakaumassa merkinnällä $z_{\frac{\alpha}{2}}$ tarkoitetaan sellaista kohtaa, että sitä suurempia satunnaismuuttujan $\underline{z}$ arvoista on $100 \frac{\alpha}{2} \%$. Luku $z_{\frac{\alpha}{2}}$ katsotaan kertymäfunktio-aulukosta α :n perusteella. Normaalijakauman symmetriaan perustuen saadaan

$$P(-z_{\frac{\alpha}{2}} \leq \underline{z} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$\Leftrightarrow P(-z_{\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$\Leftrightarrow P(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha,$$

Joten **populaation keskiarvon μ $100(1 - \alpha)$ %:n luottamusväli** on nyt

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right).$$

Esim. 89 Muuttujan $\underline{x}$ varianssin tiedetään olevan populaatiossa 9. Oletetaan muuttujan $\underline{x}$ jakauman olevan normaalijakauma. Otokseen ollessa 36 saatiin otoskeskiarvoksi $\bar{x} = 10$. Muodostetaan 95 %:n luottamusväli populaation keskiarvolle. Nyt siis

$$\alpha = 0.05 \text{ ja edelleen } z_{\frac{\alpha}{2}} = z_{0.025} = 1.96.$$

Siten μ :n 95 %:n luottamusväli on

$$(10 - 1.96 \cdot \frac{\sqrt{9}}{\sqrt{36}}, 10 + 1.96 \cdot \frac{\sqrt{9}}{\sqrt{36}}) = (9, 11)$$

Jos sanotaan, että populaation keskiarvo on ko. välillä, ollaan 95 %:n todennäköisyydellä oikeassa, ja otetaan 5 %:n riski väärälle johtopäätökselle.

(b) Jos **populaation varianssi σ^2 on tuntematon**, joudutaan se estimoimaan otoksesta. Sen estimaattorina on otosvarianssi

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right].$$

Nyt otoksen perusteella voidaan **arvioida keskiarvon keskivirhettä**

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

Odotusarvon eli populaation keskiarvon μ $100(1 - \alpha)$ %:n luottamusväliksi saadaan, kun σ^2 korvataan otosvarianssilla s^2 ja $z_{\frac{\alpha}{2}}$ korvataan taulukkoarvolla $t_{\frac{\alpha}{2}}(n - 1)$:

$$\left(\bar{x} - t_{\frac{\alpha}{2}}(n - 1) \frac{s}{\sqrt{n}}, \quad \bar{x} + t_{\frac{\alpha}{2}}(n - 1) \frac{s}{\sqrt{n}} \right).$$

Esim. 90 Kauppias väitti asiakkaalle, että kaupattavien kananmunien keskipaino on 50 g. Asiakas poimi 30 munan otoksen ja sai $\bar{x} = 45$ g ja $s = 6$ g. Tuntuuko kauppiaan väite uskottavalta?

Muodostetaan μ :lle 95 %:n luottamusväli, jolloin

$$t_{\frac{\alpha}{2}}(n - 1) = t_{0.025}(29) = 2.045$$

ja luottamusväli on

$$\left(45 - 2.045 \frac{6}{\sqrt{30}}, 45 + 2.045 \frac{6}{\sqrt{30}} \right) = (42.8, 47.2)$$

Nyt $50 \notin (42.8, 47.2)$, joten kauppiaan väite ei tunnu uskottavalta.

2° Prosenttiosuuden Π luottamusväli

Populaation tietyn tyyppisten alkuiden prosenttiosuutta Π :tä voidaan estimoida otoksesta lasketun prosenttiosuuden $\underline{P}$ avulla. Prosenttiosuus vaihtelee otoksittain, ja prosenttiluvun keskihajontaa kuvaa **prosenttiluvun keskivirhe**

$$s_P = \sqrt{\frac{P(100 - P)}{n}}.$$

Prosenttiluvun Π $100(1 - \alpha)$ %:n **luottamusväli** on

$$\left(\underline{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{\underline{P}(100 - \underline{P})}{n}}, \quad \underline{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{\underline{P}(100 - \underline{P})}{n}} \right).$$

Esim. 91 Tuote-erästä valittiin 64 tuotteen otos, jossa viallisia oli 26. Viallisuusprosentin 95 %:n luottamusväli, kun

$$P = 100 \cdot \frac{26}{64} = 40.625 \text{ ja } z_{\frac{\alpha}{2}} = 1.96,$$

on

$$\left(40.625 - 1.96 \sqrt{\frac{40.625 \cdot 59.375}{64}}, 40.625 + 1.96 \sqrt{\frac{40.625 \cdot 59.375}{64}} \right) = (29, 53).$$

7.2. Hypoteesien testauksesta

7.2.1. Testauksen pääpiirteet

Havaintoaineiston perusteella pyritään päättämään, pitääkö perusjoukon parametreista tms. tehty olettaus eli **hypoteesi** paikkansa. **Tilastollinen hypoteesi** on väite, joka tehdään populaation jakaumasta ja se koskee jotakin populaation parametria. Tällaista väitettä kutsutaan myös **nollahypoteesiksi** H_0 . Usein nollahypoteesi on oletusarvo, joka edustaa aikaisempaa tietämystä. Testin avulla H_0 joko hylätään tai hyväksytään. H_0 :n vaihtoehtoa sanotaan **vastahypoteesiksi** tai **vaihtoehtoiseksi hypoteesiksi** H_1 , joka edustaa usein uutta teoriaa, josta varsinaisesti ollaan kiinnostuneita.

Esim. 92 Maalipurkkeja täyttävän koneen on tarkoitus täyttää 1000 g:n purkkeja. Pitkän seurannan perusteella on havaittu, että paino vaihtelee niin, että painon keskihajontana on 10 g. Koneen toiminnassa tuntuu olevan joitain häiriöitä, joten koneen toiminnan tutkimiseksi asetetaan hypoteesit

$$H_0: \mu = 1000 \text{ g}$$

$$H_1: \mu \neq 1000 \text{ g.}$$

Edellä vastahypoteesi H_1 on **kaksisuuntainen**. **Yksisuuntainen** hypoteesi olisi kyseessä ollut silloin, jos olisi asetettu

$$H_1: \mu > 1000 \text{ g}$$

$$\text{tai } H_1: \mu < 1000 \text{ g}$$

Koska valinta H_0 :n ja H_1 :n välillä tapahtuu otoksen antaman informaation avulla, on muodostettava **testisuure**, jonka avulla päätös tehdään. Tavallisesti testisuure perustuu testattavan parametrin estimaattoriin. Tällaiselle testisuurelle määritetään jakaumamuoto, jota testisuure noudattaa nollahypoteesin ollessa totta. Jotta testisuureen **otantajakauma** voidaan määrätä, edellytetään tiettyjen **jakaumaa koskevien oletusten** olevan voimassa. Oletuksena voi olla esimerkiksi se, että tarkasteltavien määrällisten muuttujien jakaumat ovat normaalijakaumia.

Seuraavaksi määrätään ne testisuureen arvot, joiden katsotaan olevan liian poikkeavia, jotta näyttäisi uskottavalta, että ko. otos olisi peräisin H_0 :n mukaisesta populaatiosta. Tätä arvojoukkoa kutsutaan **kriittiseksi alueeksi C** (nollahypoteesin **hylkäysalue**). Kriittinen alue määräytyy vastahypoteesin muodon, testin jakauman ja testin **merkitsevyys-** eli **riskitason** α perusteella. Merkitsevyystaso ilmaisee todennäköisyyden, jolla tosi H_0 hylätään. Nollahypoteesia testattaessa päätetään siis etukäteen, kuinka suuri riski otetaan sille, että nollahypoteesi hylätään, vaikka sen on tosi. Yleensä käytettyjä merkitsevyystasoja ovat

0.05	tilastollisesti melkein merkitsevä (*)
0.01	tilastollisesti merkitsevä (**)
0.001	tilastollisesti erittäin merkitsevä (***)

Nollahypoteesin hylkääminen merkitsevyystasolla α tarkoittaa, että tehdään väärä johtopäätös todennäköisyydellä α .

Tämän jälkeen lasketaan sitten testisuuren arvo, verrataan sitä kriittiseen alueeseen ja tehdään johtopäätökset.

Hypoteesien testauksen jälkeenkään ei voida olla täysin varmoja siitä, että populaatiota koskeva otoksen perusteella tehty johtopäätös on oikea. Päätöksentekoon liittyy siten mahdollisuus virheen tekemiseen:

		TODELLISUUS	
		H_0 tosi	H_0 epätosi
P Ä Ä	H_0	OK	II lajin virhe eli hyväksymis- virhe
T Ö S	H_1	I lajin virhe eli hylkäämis- virhe	OK

α = I lajin virheen todennäköisyys

β = II lajin virheen todennäköisyys, joka määräytyy otoskoon ja α :n perusteella

Testauksen kulku on siis seuraava:

- 1° Asetetaan nollahypoteesi H_0
- 2° Asetetaan vastahypoteesi H_1
- 3° Määritetään tunnusluku tai testisuure ja sen jakauma H_0 :n vallitessa
- 4° Määritetään merkitsevyystaso α ja sitä vastaava kriittinen alue
- 5° Poimitaan otos ja lasketaan testisuuren arvo ja verrataan sitä kriittiseen alueeseen
- 6° Tulkitaan tulokset ja tehdään johtopäätökset.

Testien jaottelusta

Sopivan testin valintaa voi pohtia vastaamalla seuraaviin kysymyksiin:

- 1° Mistä tunnusluvusta on kyse?

- keskiarvotestit
- varianssitestit
- korrelaatiotestit
- prosenttiosuustestit
- ...

2° Mikä on vertailtavien ryhmien lukumäärä?

- yhden otoksen testit
- kahden otoksen testit
- kolmen tai useamman otoksen testit

3° Jos otoksia on enemmän kuin yksi, ovatko otokset riippumattomia vai riippuvia?

- esimerkiksi, jos naisista ja miehistä poimitaan erilliset otokset, ovat otokset riippumattomia
- riippuvia otoksia syntyy mm. ennen-jälkeen-tilanteissa ja vastinparitutkimuksissa

4° Onko kyseessä parametrinen vai epäparametrinen testi?

- parametrisissa testeissä on paljon muuttujien jakaumiin liittyviä oletuksia
- epäparametrisissa testeissä ei jakaumaoletuksia juurikaan ole

Jos tilastollisessa testuksessa käytetään apuna jotakin tilastollista ohjelmistoa, sujuu testaus yleensä yksinkertaisemmin. Usein ei testisuureen otantajakaumaa ja kriittistä aluetta tarvitse pohtia lainkaan, koska useissa tilasto-ohjelmissa testausten yhteydessä tulostuu ns. **p-arvo** (Prob, probability, significance, sig). Se ilmaisee todennäköisyyden sille, että saadaan havaittu tai sitä vielä harvinaisempi testisuureen arvo. Käsineläskentätilanteessa p-arvon pystyy määrittämään helposti, jos testisuureen jakauma on standardoitu normaalijakauma. Muiden jakaumien tilanteessa p-arvoa voi approksimoida jakaumataulukon avulla. P-arvoa sanotaan **havaituksi merkitsevyystasoksi** tai pienimmäksi merkitsevyystasoksi, jolla nollahypoteesi voidaan hylätä. Tilastollisen testin tulos tiivistyy siis yhdeksi luvuksi eli p-arvoksi, jolloin otos antaa todistusta nollahypoteesia vastaan riskillä p. Tätä p-arvoa voidaan suoraan verrata valittuun merkitsevyystasoon α :

jos $p \geq \alpha$, jää testin nollahypoteesi H_0 voimaan

jos $p < \alpha$, hylätään nollahypoteesi H_0 riskitasolla α .

Usein p-arvoa tulkitaan seuraavasti:

- jos $p > 0.05$, sanotaan, että saatu tulos **ei** ole **tilastollisesti merkitsevä**. Tällöin nollahypoteesi hyväksytään.
- jos $0.01 < p \leq 0.05$, sanotaan tuloksen olevan **tilastollisesti melkein merkitsevä**. Nyt nollahypoteesi hylätään 5 %:n merkitsevyystasolla, muttei enää 1 %:n merkitsevyystasolla
- jos $0.001 < p \leq 0.01$, sanotaan tuloksen olevan **tilastollisesti merkitsevä**. Nyt nollahypoteesi hylätään 1 %:n merkitsevyystasolla, muttei enää 0.1 %:n merkitsevyystasolla
- jos $p \leq 0.001$, sanotaan, että saatu tulos on **tilastollisesti erittäin merkitsevä**. Tällöin nollahypoteesi hylätään 0.1 %:n merkitsevyystasolla

7.2.2. Keskiarvotestejä

1° Yksi otos

Yhden otoksen tilanteessa tutkitaan, antaako otoskeskiarvo tukea nollahypoteesissa esitetylle oletukselle, joka on tehty populaation keskiarvosta. Oletuksena testille on se, että havaintojen $x_1, \dots, x_n$ jakauma on normaalijakauma $N(\mu, \sigma^2)$. Nollahypoteesi esitetään nyt muodossa

$$H_0: \mu = \mu_0.$$

(a) Kun populaation varianssi σ^2 on **tunnettu**, on testisuureena

$$\underline{z} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1),$$

kun H_0 on tosi. Kriittinen alue C muodostetaan seuraavasti:

$$H_1: \mu > \mu_0 \Rightarrow C = \{\underline{z} \mid \underline{z} > z_\alpha\}$$

$$H_1: \mu < \mu_0 \Rightarrow C = \{\underline{z} \mid \underline{z} < -z_\alpha\}$$

$$H_1: \mu \neq \mu_0 \Rightarrow C = \{\underline{z} \mid |\underline{z}| > z_{\frac{\alpha}{2}}\}$$

(b) Kun populaation varianssi σ^2 on **tuntematon**, on sille laskettava ensin estimaatti otosvarianssin avulla

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Testisuurena on nyt

$$\underline{t} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t(n-1),$$

kun H_0 on tosi. Kriittinen alue C saadaan t-jakauman avulla:

$$H_1: \mu > \mu_0 \Rightarrow C = \{\underline{t} \mid \underline{t} > t_\alpha(n-1)\}$$

$$H_1: \mu < \mu_0 \Rightarrow C = \{\underline{t} \mid \underline{t} < -t_\alpha(n-1)\}$$

$$H_1: \mu \neq \mu_0 \Rightarrow C = \{\underline{t} \mid |\underline{t}| > t_{\frac{\alpha}{2}}(n-1)\}$$

Esim. 91 Maalipurkkeja täyttävän koneen purkeista otettiin 25 purkin otos. Otoksen purkkien painon keskiarvoksi saatiin 1006 g.

Kriittinen alue, kun $\alpha = 0.05$: $C = \{\underline{z} \mid |\underline{z}| > z_{\frac{\alpha}{2}}\} = \{\underline{z} \mid |\underline{z}| > 1.960\}$

$$\underline{z} = \frac{1006 - 1000}{\frac{10}{\sqrt{5}}} = 3.0$$

$\Rightarrow H_0$ hylätään riskitasolla 0.05 eli $\mu \neq 1000$ g.

Kriittisen alueen sijasta voidaan nyt myös laskea p-arvo:

$$p = P(|\underline{z}| \geq 3.0) = 2 P(\underline{z} \geq 3.0) = 2 [1 - P(\underline{z} \leq 3.0)] = 0.0026.$$

2° Kaksi riippumatonta otosta

Kahden riippumattoman otoksen tilanteessa vertaillaan kahta otoskeskiarvoa ja selvitetään, antavatko otokset tukea nollahypoteesissa esitetylle oletukselle, että eri populaatioiden keskiarvot ovat yhtä suuret. Testin oletuksena on se, että havaintojen $x_{11}, x_{12}, \dots, x_{1n_1}$ ja $x_{21}, x_{22}, \dots, x_{2n_2}$ jakaumat ovat normaalijakaumia, ja että otokset ovat riippumattomia. Nollahypoteesi esitetään nyt muodossa

$$H_0: \mu_1 = \mu_2$$

(a) Jos populaatioiden varianssit σ_1^2 ja σ_2^2 ovat **tunnettuja**, on testisuurena

$$\underline{z} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1),$$

kun H_0 on tosi. Kriittinen alue saadaan muodostettua standardoidun normaali-jakauman perusteella:

$$H_1: \mu_1 > \mu_2 \Rightarrow C = \{\underline{z} \mid \underline{z} > z_\alpha\}$$

$$H_1: \mu_1 < \mu_2 \Rightarrow C = \{\underline{z} \mid \underline{z} < -z_\alpha\}$$

$$H_1: \mu_1 \neq \mu_2 \Rightarrow C = \{\underline{z} \mid |\underline{z}| > z_{\frac{\alpha}{2}}\}$$

(b) Jos populaatioiden varianssit σ_1^2 ja σ_2^2 ovat **tuntemattomia**, mutta ne voidaan olettaa **yhtä suuriksi** eli $\sigma_1^2 = \sigma_2^2 = \sigma^2$, lasketaan kummastakin otoksesta varianssien σ_1^2 ja σ_2^2 estimaattorit

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 \quad \text{ja} \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2$$

sekä populaatioiden yhteisen varianssin σ^2 estimaattori

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Testisuurena on nyt

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

kun H_0 on tosi. Kriittinen alue muodostuu t-jakauman perusteella:

$$H_1: \mu_1 > \mu_2 \Rightarrow C = \{t \mid t > t_{\alpha}(n_1 + n_2 - 2)\}$$

$$H_1: \mu_1 < \mu_2 \Rightarrow C = \{t \mid t < -t_{\alpha}(n_1 + n_2 - 2)\}$$

$$H_1: \mu_1 \neq \mu_2 \Rightarrow C = \{t \mid |t| > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)\}$$

- (c) Jos populaatioiden varianssit σ_1^2 ja σ_2^2 ovat **tuntemattomia**, mutta ne voidaan olettaa **erisuuriksi** eli $\sigma_1^2 \neq \sigma_2^2$, lasketaan kummastakin otoksesta varianssien σ_1^2 ja σ_2^2 estimaattorit s_1^2 ja s_2^2 . Testisuure on

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(df) \text{ likimain,}$$

kun H_0 on tosi. Testin vapausasteet df saadaan

$$\frac{1}{df} = \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1}, \quad \text{missä } c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Kriittinen alue muodostuu t-jakauman perusteella:

$$H_1: \mu_1 > \mu_2 \Rightarrow C = \{t \mid t > t_{\alpha}(df)\}$$

$$H_1: \mu_1 < \mu_2 \Rightarrow C = \{t \mid t < -t_{\alpha}(df)\}$$

$$H_1: \mu_1 \neq \mu_2 \Rightarrow C = \{t \mid |t| > t_{\frac{\alpha}{2}}(df)\}$$

Varianssien σ_1^2 ja σ_2^2 yhtä suuruuden testaus

Tämän testin avulla tehdään päätös siitä, kumpaa t-testiä (2^o(b) vai (c)) käytetään keskiarvojen testaamiseksi, kun populaatioiden varianssit σ_1^2 ja σ_2^2 ovat tuntemattomia. Testin hypoteesit ovat

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Testisuure

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1), \text{ (varianssisuhde muodostetaan siten, että } F > 1),$$

kun H_0 on tosi. Tällöin kriittinen alue muodostuu seuraavasti:

$$C = \{F \mid F > F_{\alpha/2}(n_1 - 1, n_2 - 1)\}$$

Esim. 92 Tutkitaan, onko luovan opetuksen antamisella kehittävä vaikutus.

kontrolliryhmä koeryhmä

$$n_1 = 97 \quad n_2 = 21$$

$$\bar{x}_1 = 9.25 \quad \bar{x}_2 = 11.83$$

$$s_1 = 2.72 \quad s_2 = 2.56$$

Tutkitaan ensin, onko populaatiovariansseilla eroa. Hypoteesit ovat

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2.$$

Valitaan $\alpha = 0.05$. Kriittinen alue on $C = \{F \mid F > F_{0.025}(96, 20) \approx 2.35\}$

$$F = \frac{2.72^2}{2.56^2} = 1.13$$

$\Rightarrow H_0$ jää voimaan eli populaatioiden varianssit voidaan olettaa yhtä suuriksi.

Jatketaan sitten keskiarvotestillä, jolloin hypoteesit ovat

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 < \mu_2$$

Määritetään populaatioiden yhteisen varianssin estimaatti:

$$s^2 = \frac{(97-1) \cdot 2.72^2 + (21-1) \cdot 2.56^2}{97+21-2} = 7.25$$

Valit. $\alpha = 0.01$; $n_1 + n_2 - 2 = 116 \approx 120 \Rightarrow C = \{t \mid t < -t_{0.01}(120) = -2.358\}$

$$t = \frac{9.25 - 11.83}{\sqrt{7.25} \cdot \sqrt{\frac{1}{97} + \frac{1}{21}}} = -3.92 \in C$$

$\Rightarrow H_0$ hylätään eli luovalla opetuksella näyttää olevan kehittävä vaikutus.

(Nyt $p = P(t < -3.92)$, t-jakauman taulukon perusteella arvioituna $p < 0.0005$)

3° Kaksi toisistaan riippuvaa otosta

Riippuvien otosten tilanteessa erotellaan kaksi eri päätyyppiä: vastinparitilanne (koe- ja vertailuryhmä) ja ennen-jälkeen -tilanne (tilastoyksikkö on "itsensä pari")

Nytkin vertaillaan kahta otosta, mutta nyt jokaiselle tilastoyksikköparille lasketaan x - havaintojen erotukset d_i , joiden oletetaan noudattavan normaalijakaumaa $N(\mu_d, \sigma_d^2)$. Näin saatuihin erotuksiin d_i voidaankin soveltaa yhden otoksen keskiarvotesteistä t-testiä. Nollahypoteesina on

$$H_0: \mu_1 = \mu_2 \quad (\text{eli } \mu_d = 0).$$

Esim. 93 Tutkitaan lääkkeiden A ja B vaikutusta 6 henkilön reagointinopeuteen. Mittaukset on suoritettu eri päivinä:

Henkilö	A	B	d_i	d_i^2
1	1	3	-2	4
2	7	1	6	36
3	4	2	2	4
4	3	3	0	0
5	5	1	4	16
6	4	2	2	4

$$\sum_{i=1}^6 d_i = 12; \quad \sum_{i=1}^6 d_i^2 = 64; \quad \bar{d} = \frac{12}{6} = 2; \quad s_d^2 = \frac{1}{6-1} \left(64 - \frac{12^2}{6} \right) = 8$$

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$$\text{Valitaan } \alpha = 0.05 \Rightarrow C = \{t \mid |t| > t_{0.025}(5) = 2.571\}$$

$$t = \frac{2}{\frac{\sqrt{8}}{\sqrt{6}}} = 1.73 \notin C$$

$\Rightarrow H_0$ hyväksytään eli A:n ja B:n vaikutus näyttää olevan sama

(Nyt $p = P(|t| > 1.73)$, t-jakauman perusteella arvioituna $0.10 < p < 0.20$)

7.2.3. Prosenttilukutestejä

1° Yksi otos

Yhden otoksen tilanteessa tutkitaan, antaako otoksesta laskettu tyyppiä A olevien alkuiden prosenttiosuus P tukea nollahypoteesissa esitetylle oletukselle populaation A-tyyppisten alkuiden prosenttiosuudesta Π . Nollahypoteesi esitetään muodossa

$$H_0: \Pi = \Pi_0$$

Testisuure on

$$\underline{z} = \frac{P - \Pi_0}{\sqrt{\frac{\Pi_0(100 - \Pi_0)}{n}}} \sim N(0, 1),$$

kun H_0 on tosi. Kriittinen alue saadaan:

$$H_1: \Pi > \Pi_0 \Rightarrow C = \{\underline{z} \mid \underline{z} > z_\alpha\}$$

$$H_1: \Pi < \Pi_0 \Rightarrow C = \{\underline{z} \mid \underline{z} < -z_\alpha\}$$

$$H_1: \Pi \neq \Pi_0 \Rightarrow C = \{\underline{z} \mid |\underline{z}| > z_{\frac{\alpha}{2}}\}$$

Esim. 94 Puolueen kannatus edellisissä vaaleissa oli 18.4 %. Onko kannatus kasvanut?

$$H_0: \Pi = 18.4$$

$$H_1: \Pi > 18.4$$

$n = 1493$; otoksessa kannatusprosentti $P = 22.7$

$$\text{Valitaan } \alpha = 0.01 \Rightarrow z_\alpha = 2.33 \Rightarrow C = \{\underline{z} \mid \underline{z} > 2.33\}$$

$$\underline{z} = \frac{22.7 - 18.4}{\sqrt{\frac{18.4(100 - 18.4)}{1493}}} \approx 4.3 \in C$$

$\Rightarrow H_0$ hylätään eli puolueen kannatuksen voidaan katsoa kasvaneen.

2° Kaksi otosta

Kahden otoksen tilanteessa vertaillaan A-tyyppisten alkioden prosenttiosuutta P_1 ja P_2 ja selvitetään, antavatko otokset tukea nollahypoteesissa esitetylle oletukselle, että eri populaatioiden prosenttiosuudet Π_1 ja Π_2 ovat yhtä suuret. Nollahypoteesina on

$$H_0: \Pi_1 = \Pi_2.$$

Testisuure on

$$\underline{z} = \frac{\underline{P}_1 - \underline{P}_2}{\sqrt{\underline{P}(100 - \underline{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1),$$

kun H_0 on tosi. Testisuureessa esiintyvä P kuvaa A-tyyppisten alkioden prosenttiosuutta, kun otokset yhdistetään eli

$$\underline{P} = \frac{n_1 \underline{P}_1 + n_2 \underline{P}_2}{n_1 + n_2}.$$

Kriittinen alue saadaan:

$$H_1: \Pi_1 > \Pi_2 \Rightarrow C = \{\underline{z} \mid \underline{z} > z_\alpha\}$$

$$H_1: \Pi_1 < \Pi_2 \Rightarrow C = \{\underline{z} \mid \underline{z} < -z_\alpha\}$$

$$H_1: \Pi_1 \neq \Pi_2 \Rightarrow C = \{\underline{z} \mid |\underline{z}| > z_{\frac{\alpha}{2}}\}$$

Esim. 95 Kahdestasadasta satunnaisesti valitusta naisopiskelijasta tupakoi 80 ja sadasta satunnaisesti valitusta miesopiskelijasta tupakoi 47. Onko tupakoivien prosenttiosuudella eroa?

$$H_0: \Pi_N = \Pi_M$$

$$H_1: \Pi_N \neq \Pi_M$$

$$\text{Valit. } \alpha = 0.05 \Rightarrow z_{0.025} = 1.96 \Rightarrow C = \{\underline{z} \mid |\underline{z}| > 1.96\}$$

$$\begin{aligned} P_N &= 40 & P_M &= 47 \\ n_N &= 200 & n_M &= 100 \end{aligned}$$

$$P = \frac{200 \cdot 40 + 100 \cdot 47}{200 + 100} = 42.3$$

$$z = \frac{40 - 47}{\sqrt{42.3 \cdot 57.7 \left(\frac{1}{200} + \frac{1}{100} \right)}} = -1.16 \notin C$$

$\Rightarrow H_0$ jää voimaa eli eroa ei ole.

(Nyt $p = P(|z| > 1.16) = 2 [1 - P(z \leq 1.16)] = 0.246$)

7.2.4. Riippuvuustutkimukseen liittyviä testejä

1^o χ^2 - riippumattomuustesti

Kurssin alkupuolella tarkasteltiin ristiintaulukoinnin yhteydessä tilannetta, missä havaitun ristiintaulukon solufrekvenssejä verrattiin sellaiseen ristiintaulukkoon, jossa muuttujien välillä ei esiintynyt tilastollista riippuvuutta. Tällainen ristiintaulukko koostui ns. teoreettisista eli odotetuista frekvensseistä. Havaittujen frekvenssien ja teoreettisten frekvenssien avulla muodostettiin ns. χ^2 -arvo, johon muuttujien x ja y välisen riippuvuuden voimakkuuden mittaaminen perustui. Tätä samaista χ^2 -arvoa voidaan pitää myös testisuurena, jos odotetut frekvenssit toteuttavat seuraavat ehdot:

1^o enintään 20 % e_{ij} :stä saa olla < 5

2^o jokainen e_{ij} on > 1 .

Hypoteesipari on riippumattomuustestissä:

H_0 : x ja y ovat riippumattomia

H_1 : x ja y eivät ole riippumattomia.

Testisuure on

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2[(I-1)(J-1)],$$

kun H_0 on tosi. Vapausasteiden kaavassa ovat

I = rivien lukumäärä

J = sarakkeiden lukumäärä.

Kriittinen alue saadaan:

$$C = \{\chi^2 \mid \chi^2 > \chi^2_{\alpha}[(I-1)(J-1)]\}.$$

Esim. 96 Ristiintaulukoinnin yhteydessä tarkasteltiin opiskelijoiden ansiotyössä käymisen ja opintojen edistymisen välistä riippuvuutta. Ristiintaulukossa oli $I = 3$ ja $J = 2$.

H_0 : Ansiotyössä käymisen ja opintojen edistymisen välillä ei ole riippuvuutta

H_1 : Ansiotyössä käymisen ja opintojen edistymisen välillä on riippuvuutta

Valitaan $\alpha = 0.05$; vapausasteet $= (I-1)(J-1) = 2$

$$\Rightarrow C = \{\chi^2 \mid \chi^2 > \chi^2_{0.05}(2) = 5.99\}$$

Odotetut frekvenssit e_{ij} täyttävät niille asetetut ehdot, jolloin laskettua χ^2 -arvoa voidaan pitää testisuureen arvona eli $\chi^2 = 14.2 \in C$, joten H_0 hylätään eli muuttujien välillä on riippuvuutta.

(Nyt $p = P(\chi^2 > 14.2)$, taulukosta arvioituna $p < 0.001$)

Huom. Nelikentän tapauksessa testisuurena voidaan käyttää kaavaa

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1\bullet} f_{2\bullet} f_{\bullet 1} f_{\bullet 2}} \sim \chi^2(1)$$

2° Korrelaatiokertoimen testaus

Pearsonin tulomomenttikorrelaatiokerrointa r käytetään populaation parametrin ρ estimaattorina tarkasteltaessa kahden muuttujan välistä lineaarista riippuvuutta. Jotta lineaarisen riippuvuuden testaaminen korrelaation avulla onnistuisi, on muuttujien x ja y yhteisjakauman oltava kaksiulotteinen normaalijakauma. Nollahypoteesina testissä on

$H_0: \rho = 0$ eli muuttujat ovat lineaarisesti riippumattomia.

Testisuure on

$$\underline{t} = \frac{\underline{r}\sqrt{n-2}}{\sqrt{1-\underline{r}^2}} \sim t(n-2),$$

kun H_0 on tosi. Kriittinen alue saadaan t -jakauman avulla:

$$H_1: \rho > 0 \Rightarrow C = \{\underline{t} \mid \underline{t} > t_{\alpha}(n-2)\}$$

$$H_1: \rho < 0 \Rightarrow C = \{\underline{t} \mid \underline{t} < -t_{\alpha}(n-2)\}$$

$$H_1: \rho \neq 0 \Rightarrow C = \{t \mid |t| > t_{\frac{\alpha}{2}}(n-2)\}$$

Esim. 97 Kurssin alkupuolella tutkittiin korrelaation avulla erään virvoitusjuoman mainonnan ja ostamisen välistä yhteyttä. Tutkittuja tilastoyksiköitä oli 8 ja havaittu korrelaatiokertoimen arvo oli 0.68. Selvitetään nyt, onko mainonnalla positiivista vaikutusta ostamiseen, jolloin

$$H_0: \rho = 0$$

$$H_1: \rho > 0$$

$$\text{Valitaan } \alpha = 0.05 \Rightarrow C = \{t \mid t > t_{0.05}(6) = 1.943\}$$

$$t = \frac{0.68\sqrt{8-2}}{\sqrt{1-0.68^2}} = 2.27 \in C,$$

siten H_0 hylätään eli mainonnan ja ostamisen välillä on positiivista lineaarista riippuvuutta.

Huom. Edellä esitetty t-testisuure sopii myös Spearmanin järjestyskorrelaation testaamiseen, mikäli havaintojen lukumäärä n on riittävän iso. Nyt normaalijakauma-oletustenkaan ei enää täydy olla voimassa. Nollahypoteesi tässä testissä on

$$H_0: \rho_S = 0 \quad \text{eli muuttujien välillä ei ole monotonista riippuvuutta.}$$

3° Regressiokertoimen testaus

Kun regressiosuoraa sovitetaan pisteparveen, on usein kiinnostavaa tietää, onko muuttujalla x todella vaikutusta muuttujan y vaihteluun. Tällöin keskeisessä asemassa on regressiosuoran kulmakertoimen eli regressiokertoimen testaus, koska regressiokerroin ilmoittaa, miten x :n muutos vaikuttaa keskimäärin muuttujaan y . Populaation regressiokertoimen β_1 estimaattori on otoksesta laskettu $\underline{b}_1$. Oletuksena regressioanalyysiin liittyvissä testeissä on taaskin muuttujien normaalijakautuneisuus. Regressiokertoimen testauksessa nollahypoteesiksi voidaan asettaa esimerkiksi

$$H_0: \beta_1 = 0 \quad (x \text{ ja } y \text{ lineaarisesti riippumattomia}),$$

jolloin testisuurena on

$$\underline{t} = \frac{\underline{b}_1}{\underline{s}_{b_1}} \sim t(n-2),$$

kun H_0 on tosi. Kaavassa esiintyvä s_{b_1} tarkoittaa regressiokertoimen b_1 keskivirhettä, jonka arvo saadaan laskettua esimerkiksi seuraavasti

$$s_{b_1} = \frac{s_y}{s_x} \sqrt{\frac{1-r^2}{n-2}}.$$

7.2.5. χ^2 -yhteensopivuustesti

Tarkastellaan satunnaisotosta muuttujasta, jolla on k luokkaa. Merkitään luokkien frekvenssejä $f_1, f_2, \dots, f_k$. Nyt yritetään selvittää, ovatko havaitut frekvenssit f_i sopusoinnussa asetettujen teoreettisten eli odotettujen frekvenssien e_i ($= n p_i$) kanssa. Tarkoitus on selvittää, noudattaako satunnaismuuttuja $\underline{x}$ tiettyä todennäköisyysjakaumaa. Nyt hypoteeseina ovat

H_0 : $\underline{x}$ noudattaa annettua todennäköisyysjakaumaa

H_1 : $\underline{x}$ ei noudata H_0 :n mukaista todennäköisyysjakaumaa

Testisuurena on

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} \sim \chi^2(k - s - 1),$$

kun H_0 on tosi, ja k on luokkien lukumäärä, ja s on estimoitavien parametrien lukumäärä.

Odotettujen frekvenssien tulee täyttää samat ehdot kuin χ^2 -riippumattomuustestissä.

Testin kriittinen alue saadaan seuraavasti:

$$C = \{\chi^2 \mid \chi^2 > \chi^2_{\alpha}(k - s - 1)\}$$

Esim. 98 Testataan, onko pelissä käytettä noppa kunnollinen. Hypoteesit ovat nyt

H_0 : jokaisen silmäluvun todennäköisyys on sama eli $\frac{1}{6}$

H_1 : ainakin jokin silmäluku esiintyy useammin tai harvemmin kuin muut

Noppaa heitettiin 120 kertaa, jolloin kunnollisen nopan tilanteessa jokaisen silmäluvun olisi pitänyt esiintyä $e_i = np_i = 120 \cdot \frac{1}{6} = 20$ kertaa. Havaitut frekvenssit olivat kuitenkin seuraavat:

silmluku	1	2	3	4	5	6	Yhteensä
lukumäärä	12	16	20	17	22	33	120

Valit. $\alpha = 0.05$; $r = 6$, $s = 0 \Rightarrow C = \{\chi^2 \mid \chi^2 > 11.07\}$

$$\chi^2 = \frac{(12 - 20)^2}{20} + \frac{(16 - 20)^2}{20} + \frac{(20 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(22 - 20)^2}{20} + \frac{(33 - 20)^2}{20}$$

$$= 13.1 \in C$$

$\Rightarrow H_0$ hylätään eli noppa ei näytä olevan kunnollinen.