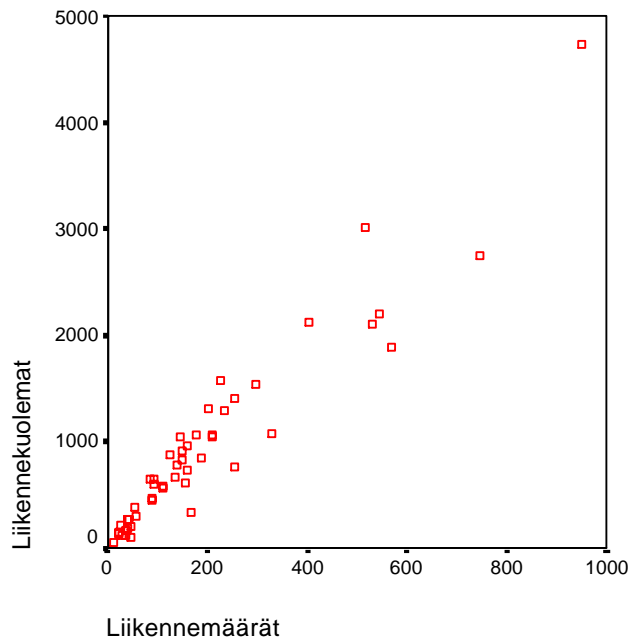
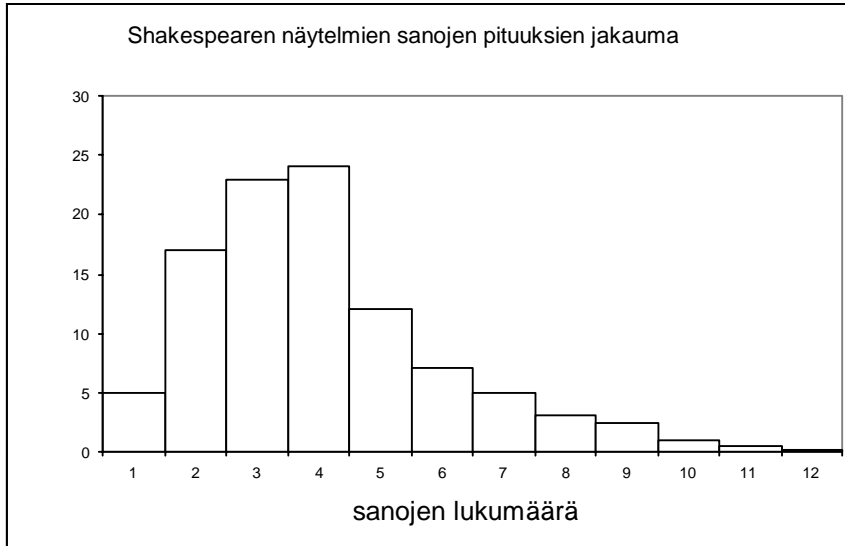


Tilastotieteen johdantokurssi [TILTP1]

<http://www.uta.fi/~strale/tiltp1/index.html>

Luentorunko



Raija Leppälä <http://www.uta.fi/~strale>
Informaatiotieteiden yksikkö
Tampereen yliopisto
Puh. (03) 3551 6301
Sähköposti raija.leppala@uta.fi

Sisällysluettelo

1	JOHDANTO	3
2	TILASTOLLINEN TUTKIMUS JA SEN TYÖVAIHEET	5
3	HAVAINTOAINEISTO	8
4	MITTAAMINEN	12
5	EMPIIRISET JAKAUMAT	15
5.1	Yksiulotteinen jakauma.....	15
5.1.1	Frekvenssijakauma	15
5.1.2	Frekvenssijakaumien graafiset esitykset.....	18
5.1.3	Yksiulotteisen jakauman tunnuslukuja	22
5.2	Kaksiulotteinen jakauma	31
5.2.1	Pisteparvi	31
5.2.2	Ristiintaulukko	32
5.2.3	Kaksiulotteisen jakauman tunnuslukuja	34
6	AIKASARJOISTA	42
6.1	Aikasarjan graafinen esitys	42
6.2	Otosautokorrelaatio	48
7	TILASTOLLISEN PÄÄTTELYN PERUSTEITA	52
7.1	Satunnaisilmiö ja tapahtuma.....	52
7.2	Klassinen todennäköisyys	53
7.3	Satunnaismuuttuja ja todennäköisyysjakauma	54
7.4	Normaalijakauma.....	56
7.5	Satunnaisotos, otossuure ja otantajakauma	60
7.6	Piste-estimointi ja luottamusvälejä.....	61
7.6.1	Prosenttiosuuden luottamusväli.....	63
7.6.2	Populaation odotusarvon luottamusväli	64
7.6.3	Kahden populaation odotusarvojen erotuksen luottamusväli	66
7.7	Hypoteesien testausta	67
7.7.1	Prosenttiosuuden testaus	68
7.7.2	Odotusarvon testaus	71
7.7.3	χ^2 -riippumattomuustesti	73
7.7.4	Odotusarvojen yhtäsuuruuden testaaminen t-testillä.....	76
7.7.5	Lineaarinen riippuvuus	78
8	LOPUKSI	79
LIITE 1	Oheiskirjallisuutta	81
LIITE 2	Kaavakokoelma ja taulukot.....	82
LIITE 3	HOTDOG-aineisto.	84
LIITE 4	PULSSI-aineisto.	85
LIITE 5	Opintojakson TILTP1 kurssiarviointilomakkeet.....	86
LIITE 6	Menetelmien ja testien ryhmittelystä muuttujan roolin mukaan.	87

1 JOHDANTO

Tilastotiede on menetelmätiede, joka käsittelee tietojen hankinnan suunnittelua (otantamenetelmät, koejärjestelyt, kyselylomakkeet), tietojen keruuta, tietojen esittämistä (kuvailevaa tilastotiedettä) ja tietojen analysointia (johtopäätelmien tekoa erilaisten analysointimenetelmien avulla).

Sovelijat käyttävät tilastotieteilijöiden kehittämia menetelmiä tietoaineiston keruuseen, kuvailuun ja analysointiin. Tilastotiedettä käytetään hyväksi aina, kun käsitellään empiiristä tietoaineistoa.

Tilastollinen analyysi voidaan karkeasti jakaa kuvailevaan analyysiin (*descriptive statistics*) ja tilastolliseen päättelyyn (*statistical inference*).

Kuvaileva osuus pyrkii kuvailemaan tietoaineistoa erilaisten graafisten esitysten ja tunnuslukujen sekä taulukoiden avulla. Tilastollinen päättely eli inferenssi käsittelee johtopäätelmien tekoa aineiston (otoksen) perusteella. Johtopäätelmät tehdään erilaisten todennäköisyyslaskentaan perustuvien tilastollisten testien ja analysointimenetelmien avulla.

Kehittynyt tietotekniikka luo mahdollisuudet "helppoon" aineiston analysointiin; tietokone hoitaa laskennan tehokkaasti sekä tarjoaa analysoinnin tuloksia käyttäjälle. Käyttäjän on kuitenkin tiedettävä mitä milloinkin voi tehdä: mitä analysointimenetelmiä voi käyttää, miten niitä tulkitaan ja millä tavalla tehdään johtopäätelmiä. Tietoaineiston analysointia ei voida tehdä siten, että syötetään aineisto tietokoneeseen ja saadaan nappia painamalla automaattisesti tutkimustulokset!

Tietojenkäsittelyä ja matematiikkaa voidaan pitää eräänlaisina tilastotieteen apuvälineinä. Itse tilastotieteen teorian tutkiminen vaatii melko vankkaa matematiikan tuntemusta.

Tällä opintojaksolla käsitellään aluksi empiirisen aineiston hankintaan liittyviä asioita sekä havaintoaineiston sisältämän tiedon esittämistä kuvailevan analyysin keinoin. Kuvailevan osuuden jälkeen tutustutaan esimerkinomaisesti joihinkin todennäköisyysjakauksiin, testiteorian alkeisiin sekä muutamiin tilastollisiin testeihin.

Tässä monisteessa lähes jokaiseen asiakokonaisuuteen liittyen on lyhyesti esitetty myös se, miten kyseinen analyysi saadaan tehtyä tilastolaskentaohjelmalla SPSS.

Ks. SPSS tarkemmin sivulta

<http://www.fsd.uta.fi/menetelmaopetus/SPSS/spss.html> ,

<http://www.uta.fi/~strale/spsslink.html> , jossa erityisesti hyödyllinen

<http://joyx.joensuu.fi/~ek/SPSS/spss.html>

Opiskelun tukena voi käyttää liitteessä 1 esitettyä oheiskirjallisuutta sekä tässä monisteessa olevia linkkejä, joista tärkeimpinä mainittakoon Yhteiskuntatieteellisen tietoarkiston kvantitatiivisten tutkimusmenetelmien oppimisympäristö <http://www.fsd.uta.fi/menetelmaopetus/intro.html> , Juha Purasen (Helsingin yliopisto) tilastotieteen kurssien termistöstä koottu Tilastosanasto

<http://www.mathstat.helsinki.fi/tilastosanasto/> ja Tilastokeskuksen verkkokoulu
<http://tilastokeskus.fi/tup/verkkokoulu/index.html> .

2 TILASTOLLINEN TUTKIMUS JA SEN TYÖVAIHEET

Tilastollinen tutkimus kohdistuu tutkimusobjektien muodostamaan joukkoon, jota kutsutaan perusjoukoksi eli populaatioksi (*population*). Lähes aina empiirinen tutkimus joudutaan suorittamaan käyttäen vain osaa populaatiosta (otos (*sample*)). Populaation yksiköitä kutsutaan tilastoyksiköiksi eli havaintoyksiköiksi.

Empiirinen havaintoaineisto (*data*) saadaan mittaamalla tilastoyksiköiden ominaisuuksia. Tilastoyksikön ominaisuuksia kutsutaan tilastollisiksi muuttujiksi (*variable, statistical variable*). Voidaan yleisesti merkitä $x, y, z, \dots, x_1, x_2, x_3, \dots$

Tilastolliset analyysimenetelmät ovat välineitä havaintoaineiston tutkimiseksi sekä johtopäätelmien tekemiseksi populaatiosta aineiston perusteella.

Esim. 2.1. Opintojaksolle ilmoittautuminen.

Ks. sivulla http://mtl.uta.fi/tilt/enlist.php?cfg=tiltp1_2011

- tilastoyksikkö opiskelija
- muuttujia
 - opiskelijan pääaine/koulutusohjelma, sukupuoli, harjoitusryhmätoive, opintojen aloitusvuosi

Aineiston perusteella saadaan selville vaikkapa miesten ja naisten lukumäärät, opiskelijoiden lukumäärät opintojen aloitusvuoden perusteella, miesten ja naisten prosentuaaliset määrät pääaineittain. Näin menetellen aineiston sisältämää tietoa tiivistetään. Saadaan kuvattua aineistoa sekä voidaan tutkia erilaisia riippuvuuksia.

Esim. 2.2. Opintojakson tenttiin osallistujat

- tilastoyksikkö opiskelija
- populaatio esim. kaikki opintojakson opiskelijat
- muuttujia esim. opiskelijan pääaine, saatu tenttipistemäärä

Aineiston perusteella voidaan tutkia esimerkiksi tenttimenestymistä opiskelijan taustan mukaan laskemalla tenttipistemäärän keskiarvot pääaineittain.

Esim. 2.3.

- a) Populaationa Suomen kunnat
- tilastoyksikkö kunta
 - muuttujia esim.
 - kunnan asukasluku, asuntojen keskikoko, lääni, jossa kunta sijaitsee

Aineiston perusteella voidaan kuvata esim. kuntien asukasmäärää, asuntojen kokoa ja vertailla vaikkapa lääneittäin kunnan asukasmäärien keskiarvoja.

Ks. esimerkkiaineisto http://mtl.uta.fi/tilasto/tiltp_aineistoja/kunnat_2.sav tai http://mtl.uta.fi/tilasto/tiltp_aineistoja/kunnat_2.xls ja muuttujien kuvaus http://mtl.uta.fi/tilasto/tiltp_aineistoja/kunnat_2.PDF

- b) Populaationa (tai otoksena) Eduskunta 2007
 – tilastoyksikkö kansanedustaja
 – muuttujia edustajan ikä, puolue, äänimäärä, ammatti

Eduskunta-aineiston perusteella saadaan selville mm. naisten prosentuaalinen osuus, ikäjakauma, ikäjakauma naisilla ja miehillä, naisten prosentuaaliset määrät puolueittain. Voidaan myös selvittää, miten hyvin eduskunta edustaa koko kansaa. Onko ikäjakauma samanlainen kuin kaikkien äänioikeutettujen? Onko naisten suhteellinen osuus populaatiota vastaava?

Ks. aineisto http://mtl.uta.fi/tilasto/tiltp_aineistoja/eduskunta_2007.sav tai http://mtl.uta.fi/tilasto/tiltp_aineistoja/eduskunta_2007.xls

Esim. 2.4. Tapahtuma tilastoyksikkönä.

- a) Synnytys, jolloin voidaan tutkia, ovatko tytöt ja pojat keskimäärin samanpainoisia syntyessään, ovatko esikoiset yhtä hyväkuntoisia kuin ei-esikoiset, mikä on ensisynnyttäjän keski-ikä.
 ks. esimerkkiaineisto http://mtl.uta.fi/tilasto/tiltp_aineistoja/saidit.sav tai http://mtl.uta.fi/tilasto/tiltp_aineistoja/saidit.xls
- b) Liikenneonnettomuus, jolloin muuttujina voisi olla esim. onnettomuuspaikka ja –aika, loukkaantuneiden määrä, onnettomuuden osapuolten lukumäärä.
- c) Työtapaturma, josta kirjataan itse tapaturmaan liittyviä monenlaisia tietoja sekä tietoja tapaturmasta johtuvista hoitotoimenpiteistä ja –kuluista.
- d) Jääkiekko-ottelu, jolloin muuttujia voisi olla yleisömäärä, ottelun lopputulos, jäähyjen yhteismäärä.

Tilastollisen tutkimuksen (siis myös opintojakoon liittyvän harjoitustyön) työvaiheet voidaan esittää seuraavasti:

1. Suunnittelu
 - tutkimuskohteen ja aiheen valinta (tilastoyksikkö, muuttujat)
 - tutkimuksen suorittamisen suunnittelu (kyselylomake, otantamenetelmä, koejärjestely jne.)
2. Aineiston keruu ja tallennus analysointia varten
 - suunnitellun havaintoaineiston hankinta
 - tallennus ja muokkaus analysointia varten
3. Aineiston kuvailu
 - kuvailevan tilastotieteen keinoin aineiston sisältämän tiedon tutkimista (myös virheiden ja poikkeavien tietojen etsintä)
4. Tilastolliset mallit ja testaukset
 - populaatiosta tehtyjen erilaisten väittämien testaukset otoksen (aineiston) perusteella
 - todennäköisyysteoriaan perustuvien tilastollisten mallien sovittaminen havaintoaineistoon
5. Raportointi
 - johtopäätelmien teko ja niiden esittäminen ja tulkinta

Ks. tarkemmin

http://mtl.uta.fi/tilasto/tiltp1/syksy2003/moniste_1.pdf ja

<http://mtl.uta.fi/tilasto/tiltp1/syksy2011/htyop111.pdf> harjoitustyön ohjeistus.

3 HAVAINTOAINEISTO

Tutkimuksen suunnitteluvaiheessa on siis selvitetty tutkimuksen tavoite, määritelty tilastoyksikkö ja populaatio sekä muuttujat, joiden avulla tutkittaviin ongelmiin pyritään löytämään ratkaisuja.

Suunnitteluvaiheen jälkeen on vuorossa aineiston hankinta. Päätetään tehdäänkö otanta- vai kokonaistutkimus. Populaation osajoukko on satunnaisotos (*random sample*), jos se on valittu todennäköisyysotannalla eli tiettyjen sääntöjen mukaan satunnaisesti siten, että tutkijan subjektiivinen harkinta ei vaikuta. Otantatutkimuksessa tutkitaan vain otos; kokonaistutkimuksessa koko populaatio. Harvoin on mahdollista suorittaa kokonaistutkimusta. Käytännössä tehdään siis otoksen perusteella johtopäätelmiä populaatiosta, jolloin johtopäätelmän tekemiseen liittyy epävarmuutta.

Esim. 3.1.

a) Jos arvioidaan esimerkiksi tietyn puolueen kannatusosuutta, niin voidaan ilmoittaa arvioon liittyvä virhemarginaali, ks. esim. Taloustutkimuksen tekemiä tutkimuksia <http://www.taloustutkimus.fi> (→ Tuotteet ja palvelut → Puolueiden kannatusarviot). Tällöin on muodostettu otoksen avulla luottamusväli todelliselle kannatusosuudelle.

b) Jos halutaan arvioida suomalaisten naisten keskipituutta, niin lasketaan otoksesta keskipituus ja arvioidaan virhettä, joka liittyy päättelyyn. Tässäkin voidaan muodostaa keskipituudelle luottamusväli.

Tapoja, joilla otos voidaan tehdä populaatiosta, kutsutaan otantamenetelmiksi. Otantamenetelmiä ovat yksinkertainen satunnaisotanta (YSO), systemaattinen otanta (SO), ositettu otanta (OO) sekä ryväotanta (RY).

Yksinkertaisessa satunnaisotannassa tilastoyksiköt arvotaan otokseen (voidaan tehdä joko palauttaen tai palauttamatta valittu alkio populaatioon ennen seuraavan arvontaa).

Systemaattinen otanta soveltuu menetelmäksi populaation alkioden ollessa järjestyksessä, esim. kortistossa, autojonossa. Olkoot esimerkkinä 12 alkion populaatio, josta halutaan tehdä neljän alkion otos. Jaetaan populaation kolmen alkion ryhmiin $a_1 a_2 a_3 | a_4 a_5 a_6 | a_7 a_8 a_9 | a_{10} a_{11} a_{12}$ ja valitaan ensimmäisestä ryhmästä satunnaisesti yksi ja sitten joka kolmas. Näin saadaan yksi otos. Kaikki mahdolliset 4 alkion systemaattisella otannalla tehdyt otokset ovat

a_1, a_4, a_7, a_{10}

a_2, a_5, a_8, a_{11}

a_3, a_6, a_9, a_{12}

Satunnaistaminen tehtiin siis siten, että valittiin satunnaisesti kolmesta ensimmäisestä yksi ja sitten joka 3. Poimintaväli tässä on siis $12/4 = 3$.

Ositetussa otannassa populaatio jaetaan osiin jonkun ominaisuuden suhteen ja tehdään ositteista otanta jollain otantamenetelmällä.

Ryväotannassa otosalkiona on joukko tilastoyksiköitä.

Ks. tarkemmin otantamenetelmistä sivuilta

<http://www.fsd.uta.fi/menetelmaopetus/otos/otantamenetelmat.html> ja
<http://tilastokeskus.fi/tup/verkkokoulu/data/tt/01/09/index.html>

Joskus aineisto voi olla valmiina olemassa tai se saadaan yhdistelemällä useammasta tietolähteestä (lehtitiedot, kortistot, tapahtumat, kokoelmat, tilastojulkaisut mm. Tilastokeskuksen julkaisemat valtion tilastot, kuntien tilastot, jne..., erilaiset tietokannat, www, ...).

Monissa tutkimustilanteissa aineisto on kuitenkin hankittava itse. Voidaan tehdä joko kysely- tai haastattelututkimus (erilaiset mielipidetutkimukset, markkinointitutkimukset) tai kokeellinen tutkimus. Kokeellisessa tutkimuksessa pyritään usein koejärjestelyn avulla selvittämään jonkun käsittelyn vaikutusta tilastoyksiköiden ominaisuuksiin. Tällaisia ovat mm. viljelykokeet, lääketieteelliset kokeet, oppimiskokeet, käytettävyysskokeet.

Kerätty aineisto muokataan ja esitetään taulukkona, jota kutsutaan havaintomatriisiksi (*data matrix*). Mm. tilastolliset ohjelmistot käyttävät havaintomatriisiesitystä aineiston "tallennustapana".

Olkoon aineistossa n tilastoyksikköä ja p muuttujaa. Merkitään tilastoyksiköitä yleisesti $a_1, a_2, a_3, \dots, a_n$ (toki ne voidaan myös nimetä) ja muuttujia $x_1, x_2, x_3, \dots, x_p$ (tarvittaessa nimetään tilanteen mukaan). Havaintomatriisi sisältää muuttujien arvot jokaiselta tilastoyksiköltä muodossa

	x_1	x_2	...	x_j	...	x_p
a_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
a_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
.						
.						
a_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
.						
.						
a_n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

missä x_{ij} on i . tilastoyksikön mittaluku ominaisuudelle x_j . Havaintomatriisi on siis mittaluvuista (ks. mittaaminen) muodostettu taulukko, jossa i . vaakarivillä on i . tilastoyksikön mittaluvut (kutsutaan havaintovektoriksi, havainnoksi) ja j . pystyrivillä eli sarakkeella on j . muuttujan mittaluvut tilastoyksiköittäin. Sarakkeen luvut muodostavat kyseisen muuttujan jakauman.

Aineistossa ei aina ole jokaiselta tilastoyksiköltä kaikkia muuttujan arvoja. Tällöin on kyse puuttuvista tiedoista, jotka eivät välttämättä hankaloita aineiston analysointia tai vääristä tutkimustuloksia, mutta saattavat näin tehdäkin.

Ks. tarkemmin <http://www.fsd.uta.fi/menetelmaopetus/puuttuvat/puuttuvat.html> .

Esim. 3.2. Jos muuttujat on nimetty vaikkapa x , y ja z , niin havaintomatriisi voidaan tällöin kirjoittaa muodossa:

	x	y	z
a_1	x_1	y_1	z_1
a_2	x_2	$y_2 \dots$	z_2
\cdot	\cdot		
\cdot	\cdot		
a_n	x_n	y_n	z_n

Tämä esitystapa on joskus mukavampi (esim. tunnuslukujen määritelmien yhteydessä).

Ks. lisätietoja havaintomatriisista

<http://tilastokeskus.fi/tup/verkkokoulu/data/tt/01/06/index.html>

Esim. 3.3. Havaintomatriisi, tilastoyksikkönä myynnissä oleva asunto.

Selling price of homes in Gainesville, Florida, January 1996. P = selling price in thousands of dollars, S = size of home in thousands of square feet, BE = number of bedrooms, BA = number of bathrooms, New = whether new (1 = yes, 0 = no). Data provided by Jane Myers, Coldwell-Banker Realty.

P	S	Be	Ba	New
61,5	1,01	3	2	0
60,0	1,34	3	2	0
65,9	1,22	3	1	0
86,0	1,15	2	2	1
86,9	1,58	3	2	1
86,9	1,58	3	2	1
67,9	1,28	3	2	0
68,9	1,29	3	2	0
69,9	1,52	3	2	0
70,5	1,25	3	2	0
72,9	1,28	3	2	0
72,5	1,28	3	1	0
72,0	1,36	3	2	0
71,0	1,20	3	2	0
76,0	1,46	3	2	0
72,9	1,56	4	2	0
73,0	1,22	3	2	0
70,0	1,40	2	2	0
76,0	1,15	2	2	0
69,0	1,74	3	2	0

Esim. 3.4. Havaintomatriisi opetuksen ja oppimisen arviointi -aineiston (kyselylomake <http://mtl.uta.fi/arviointi/tiltp1.htm>). perusteella.

	Pääaine	Sukup.	Opiskeluv.	Kurssin työläys ...	Lisätyö
Opiskelija 1	1	1	1	2	2
Opiskelija 2	3	2	1	4	1
Opiskelija n	2	1	3	5	1

Esim. 3.5. Havaintomatriisi liitteessä 3 sisältää 54 tilastoyksikköä ja 5 muuttujaa. Havaintomatriisiin voidaan lisätä uusi muuttuja, joka kertoo hotdogin kilohinnan euroina. Koska 1 unssi on 28,35 g ja 1 dollari on n. 0,79 euroa), niin Kilohinta = $(\$/oz) \times 0,79 / 0,02835$.

Esim. 3.6. Havaintomatriisit esimerkeissä 2.3 ja 2.4 sekä liite 4.

SPSS Havaintomatriisin avaaminen tai uuden luominen tapahtuu valikosta

```

File
  New>
    Data...
    uuden luominen
  Open>
    Data...
    vanhan avaaminen (oletusarvoisesti näkyvät .sav-tunnisteella olevat).
```

Usein tarvitaan uusia laskennallisia muuttujia. Uuden muuttujan tekeminen havaintomatriisissa olemassa olevien muuttujien avulla (esimerkiksi summat, suhteet, mittayksikkövaihdot) suoritetaan valikosta

```

Transform
  Compute...
  Avautuvassa ikkunassa nimetään uusi muuttuja (Target Variable) ja
  määritellään laskukaava (Numeric Expression).
```

4 MITTAAMINEN

Mittaaminen tarkoittaa menettelyä (sääntöä), jolla tilastoyksikköön liitetään sen tiettyä ominaisuutta kuvaava luku, mittaluku. Joidenkin muuttujien kohdalla voidaan valita käytetäänkö mittalukuna merkkiä vai numeroa. Monet tilastolliset ohjelmistot sallivat myös merkkitiedon käytön.

Mittaamisen yhteydessä esiintyy usein mittausvirhettä johtuen mittarin tarkkuudesta tai mittaukseen liittyvistä häiriötekijöistä. Mittausmenetelmä saattaa olla sellainen, että toisistaan riippumattomat samalle tilastoyksikölle tehdyt mittaukset antavat huomattavasti poikkeavia tuloksi. Tällöin sanotaan mittarin reliabiliteetin olevan alhainen. Mittari saattaa olla myös huonosti laadittu ts. mittari ei mittaakaan sitä ominaisuutta, mitä sen olisi tarkoitus mitata. Sanotaan, että mittari ei ole validi. Ks. tarkemmin <http://www.fsd.uta.fi/menetelmaopetus/mittaaminen/mittaaminen.html>.

Osa muuttujista on suoraan mitattavissa ja tulkittavissa; esim. pituus, paino, kunnan asukasluku ovat suoraan mitattavissa eikä muuttujien tulkinnassa ole vaikeuksia. Mutta esim. muuttujat älykkyys, musikaalinen lahjakkuus, uskonnollisuus, asenne johonkin, www-sivun käytettävyyks eivät ole suoraan mitattavissa eikä niiden määrittäminen ole yksikäsitteistä. Näitä nk. teoreettisia muuttujia mitataan usean indikaattorimuuttujan avulla. Uskonnollisuutta voidaan mitata esim. kirkossa käyntien määrällä, uskonnollisen kirjallisuuden lukemisella, jne. Asenne-/mielipidemittauksissa asetetaan asennetta/mielipidettä peilaavia väitteitä. Väittämien suunnittelussa on oltava huolellinen. Väittämät on osattava asettaa siten, että ne mittaavat sitä mitä tutkija haluaa niiden mittaavan.

Väittämä asetetaan usein siten, että vastaaja valitsee esimerkiksi vaihtoehdoista

- 1 täysin samaa mieltä
- 2 jokseenkin samaa mieltä
- 3 ei osaa sanoa
- 4 jokseenkin eri mieltä
- 5 täysin eri mieltä

Ks. myös kyselylomakkeen laatimisesta

<http://www.fsd.uta.fi/menetelmaopetus/kyselylomake/laatiminen.html> sekä postikyselyn tekemisestä <http://www.fsd.uta.fi/menetelmaopetus/postikysely/postikysely.html>.

Esim. 4.1. Liite 5.

Muuttujia voidaan luokitella monellakin tavalla. Ensinnäkin tilastolliset muuttujat voidaan jakaa kahteen päätyyppiin: kategoriisiin (*categorical*) eli kvalitatiivisiin ja numeerisiin (*numerical*) eli kvantitatiivisiin. Lisäksi muuttujat voidaan luokitella nk. mitta-asteikkojen perusteella.

Mitta-asteikot sanelevat mitä tilastollisia menetelmiä saa käyttää, mitä tunnuslukuja laskea, jne.

Kategoriset I. kvalitatiiviset muuttujat jakavat tilastoyksiköt tarkasteltavan ominaisuuden suhteen luokkiin; esim. henkilön siviilisääty, opiskelijan koulutusohjelma, kaupungin sijaintilääni.

Jos kategorisen muuttujan arvoja ei voida asettaa järjestykseen (esim. paremmuus, suuruus, kovuus) sanotaan, että muuttuja on luokittelu- eli laatuero- eli nominaaliasteikollinen; esim. siviilisääty.

Jos arvot voidaan asettaa mielekkääseen järjestykseen on kyse järjestys- eli ordinaaliasteikosta; esim. asennekysymykset.

Kategoriset muuttujat voidaan koodata numeerisesti, mutta numeroarvoilla ei ole määrällistä tulkintaa; ne ovat vain luokkien nimiä tai kertovat luokkien "suuruusjärjestyksen".

Numeerisen I. kvantitatiivisen muuttujan arvo on jo mitattaessa reaalinen. Mitataan lukumäärää tai suoritetaan mittaus mittayksikköä käyttäen. Numeerisessa mittauksessa sama mittalukujen erotus vastaa kaikkialla samansuuruisia ominaisuuksien eroja. Esimerkiksi lasten lukumäärä perheessä, lapsen paino kiloina, pituus sentteinä, 100 m juoksuaika, tehopisteet jääkiekossa ovat kvantitatiivisia muuttujia.

Jos numeerisen muuttujan arvo nolla vastaa tarkasteltavan ominaisuuden "häviämistä", absoluuttista nollapistettä, niin on kyse suhdeasteikollisesta muuttujasta; esim. paino kiloina, pituus metreinä. Jos muuttujan arvolla nolla ei ole tätä tulkintaa, niin on kyse välimatka- eli intervalliasteikosta; esim. lämpötila Celsius-asteina. Suhdeasteikolla muuttujan arvojen suhteilla on mielekäs tulkinta. Intervalliasteikolla voidaan vertailla arvojen eroja, mutta ei suhteita.

Välimatka- ja suhdeasteikolla mittayksikköä ei ole kiinnitetty. Absoluuttinen asteikko on kyseessä, jos suhdeasteikollinen muuttuja voidaan mitata vain tiettyä mittayksikköä käyttäen; esim. lukumäärät.

Numeeristen muuttujien yhteydessä lähes samat menetelmät ja tunnusluvut käyvät kaikille kolmelle mitta-asteikolle; esim. keskiarvon laskenta edellyttää vähintään intervalliasteikon, samoin korrelaatiokertoimen käyttö jne.

Numeerinen muuttuja voi olla joko diskreetti (epäjatkuva) tai jatkuva. Diskreetti muuttuja voi saada arvokseen äärellisen määrän erisuuria arvoja tai äärettömän määrän siten, että arvot ovat numeroitavissa positiivisia kokonaislukuja käyttäen.

Esim. 4.2. Tehdään liikuntaa harrastaville kysely liikuntamääristä.

Kvantitatiivisesti mitattuna:

Harrastan liikuntaa (väh. 30 min/kerta) keskimäärin ____ kertaa viikossa.
Keskimäärin kerralla liikun ____ min.

Kvalitatiivisesti mitattuna:

Harrastan liikuntaa (väh. 30 min/kerta) keskimäärin
____ alle 3 kertaa viikossa.
____ 3–4 kertaa viikossa.
____ enemmän kuin 4 kertaa viikossa.

Keskimäärin kerralla liikun

____ alle tunnin
____ 1–2 tuntia
____ yli 2 tuntia

Ks. lisätietoja mittaamisesta

<http://www.fsd.uta.fi/menetelmaopetus/mittaaminen/ominaisuudet.html>,

<http://tilastokeskus.fi/tup/verkkokoulu/data/tt/01/05/index.html>

<http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/asteikko>

http://mtl.uta.fi/tilasto/tiltp1/syksy2003/moniste_2.pdf

5 EMPIIRISET JAKAUMAT

5.1 Yksiulotteinen jakauma

Empiirisen aineiston eritysmuotona käytetään siis havaintomatriisia, jossa n tilastoyksikön p muuttujan arvot esitetään tilastoyksiköittäin seuraavasti:

	X_1	X_2	...	X_j	...	X_p
a_1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}
a_2	X_{21}	X_{22}	...	X_{2j}	...	X_{2p}
.						
.						
a_i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ip}
.						
.						
a_n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}

Havaintomatriisissa sarakkeilla on siis muuttujien $x_1, x_2, x_3, \dots, x_p$ jakaumat. Nämä yksiulotteiset (yhden muuttujan) jakaumat eivät sellaisenaan ole havainnollisia vaan niiden sisältämää tietoa on syytä tiivistää. Yksi tapa on muodostaa luokiteltu yksiulotteinen jakauma eli frekvenssijakauma.

5.1.1 Frekvenssijakauma

Yhden muuttujan frekvenssijakauma muodostetaan siten, että jaetaan (tarvittaessa) muuttujan arvot luokkiin ja ilmoitetaan jokaisen luokan osalta kuinka monta on sellaista tilastoyksikköä, joilla tarkasteltavan muuttujan arvo kuuluu tähän luokkaan.

Ks. myös <http://tilastokeskus.fi/tup/verkkokoulu/data/tt/01/12/index.html>

Esim. 5.1.1. Pulssi-muuttujan frekvenssijakauma (aineisto liite 4).

<u>Pulssi</u> <u>-muuttujan</u> <u>luokat</u>	<u>Tilastoyksiköiden</u> <u>määrä</u> <u>luokassa</u>
42–52	4
53–63	7
64–74	31
75–85	25
86–96	12
97–107	1

Luokkaan i kuuluvien havaintojen määrää kutsutaan i . luokan frekvenssiksi, merkitään f_i .

Esim. 5.1.2. Esimerkki frekvenssijakaumasta
<http://tilastokeskus.fi/tup/verkkokoulu/data/tt/02/02/index.html>

Esim. 5.1.3. Yritykset toimialoittain.

Toimiala	Frekvenssi
Elintarviketeollisuus	7
Ilmailu	18
IT	22
Lääketeollisuus	12
Paperinjalostus	11
Öljynjalostus	27
	97

Jos muuttuja on järjestysasteikollinen, esitetään luokat "suuruus" järjestyksessä.

Esim. 5.1.4. Opetuksen ja oppimisen arviointi -aineisto (kyselylomake <http://mtl.uta.fi/arviointi/tiltp1.htm>).

Luennoilla käynti		
lähes joka kerta		47%
noin joka toinen kerta		31%
silloin tällöin		14%
en ollenkaan		8%
Opintojakson työläys		
vähätöinen	1	2%
	2	21%
	3	32%
	4	39%
työläs	5	6%

Ks. muut jakaumat <http://www.uta.fi/%7Estrale/p1pal02.html>

Kun siis ilmoitetaan muuttujan x luokat E_1, E_2, \dots, E_k ja näiden luokkien frekvenssit f_1, f_2, \dots, f_k , niin on muodostettu muuttujan x yksiulotteinen luokiteltu jakauma eli suora jakauma eli frekvenssijakauma (*frequency distribution*). Luokat ovat käyttäjän määrättävissä. Jos numeerisen muuttujan kaikki luokat ovat samanpituisia, niin kyseessä on tasavälinen luokitus.

Numeeristen muuttujien yhteydessä määritellään pyöristetyt luokkarajat ja todelliset luokkarajat. Pyöristetyt luokkarajat ovat mittaustarkkuudella ilmoitettuja rajoja. Jos $d =$ mittaustarkkuus, niin luokan todellinen yläraja = pyöristetty yläraja + $d/2$ ja todellinen alaraja = pyöristetty alaraja - $d/2$.

Esim. 5.1.5. Pulssi muuttujan jakauma esimerkissä 5.1.1 (aineisto liite 4).

pyöristetyt luokkarajat	todelliset luokkarajat	frekv.
42–52	41,5–52,5	4
53–63	52,5–63,5	7
64–74	63,5–74,5	31
75–85	74,5–85,5	25
86–96	85,5–96,5	12
97–107	96,5–107,5	1

Luokan E_j suhteellinen frekvenssi on $p_j = f_j/n$, prosentuaalinen frekvenssi $(f_j/n)100$, summafrekvenssi $F_j = f_1 + f_2 + \dots + f_j$ ja suhteellinen summafrekvenssi $P_j = p_1 + p_2 + \dots + p_j$

Numeerisen muuttujan luokkakeskus on luokan keskikohta. Luokan pituus lasketaan todellisten luokkarajojen erotuksena (tasavälisessä luokituksessa myös luokkakeskusten erotuksena).

Esim. 5.1.6. (jatkoa esim. 5.1.5).

Pulssi	luokkak.	f_j	F_j
42–52	47	4	4
53–63	58	7	11
64–74	69	31	42
75–85	80	25	67
86–96	91	12	79
97–107	102	1	80

Tässä on käytetty tasavälistä luokitusta, jossa luokan pituus 11.

Ehdollinen frekvenssijakauma saadaan muodostamalla tarkasteltavasta muuttujasta frekvenssijakaumat toisen muuttujan eri luokissa. Voidaan selvittää, miten tämä ehdollistettu muuttuja vaikuttaa tarkasteltavan muuttujan jakaumaan. Ehdollisia jakaumia vertailtaessa käytetään suhteellisia tai prosentuaalisia frekvenssejä.

Esim. 5.1.7. Lepopulssi miehillä ja naisilla erikseen (aineisto liite 4).

	Mies	Nainen	
Pulssi	42–52	0	4
		9,1%	0,0%
	53–63	1	7
		13,6%	2,8%
	64–74	13	31
		40,9%	36,1%
	75–85	12	25
	29,5%	33,3%	
86–96	9	12	
	6,8%	25,0%	
97–107	1	1	
	0%	2,8%	
	44	36	80

SPSS Frekvenssijakauman saa taulukkona valikosta

Analyze
 Descriptive Statistics>
 Frequencies...

Numeerisen muuttujan yhteydessä luokituksen tekeminen tai kategoristen muuttujien tapauksessa luokkien yhdistäminen tapahtuu tekemällä uusi muuttuja havaintomatriisiin uudelleen koodauksen kautta. Koodaus tapahtuu valikosta

 Transform
 Recode >
 Into Different Variables...
 jossa annetaan luokiteltava muuttuja (Input Variable), luokituksen seurauksena syntyvän muuttujan nimi (Output Variable) sekä koodauksen (luokituksen) määrittely (Old and New Values...);

Ehdollisten jakaumien (tai yleensä ehdollistamisen) teon yhteydessä ilmoitetaan ohjelmistolle, että jatkossa halutaan analysoinnit tehtävän jonkun muuttujan (tai muuttujien) eri luokissa erikseen (esimerkiksi miehillä ja naisilla erikseen) antamalla ehdollistava muuttuja valikossa

 Data
 Split file...
 vaihtoehto Compare groups ja valitsemalla muuttujaluettelosta ryhmittelymuuttuja;
 ryhmittelyn purkaminen vaihtoehto Analyze all cases.

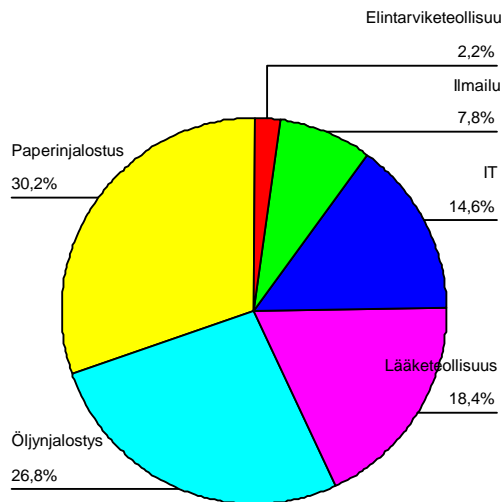
Tämän määrittelyn jälkeen tehtävät analyysit tapahtuvat erikseen kaikissa ehtomuuttujan ryhmissä (myös puuttuvien tietojen ryhmässä!) erikseen. Ehto on päällä siihen asti kun se otetaan pois.

5.1.2 Frekvenssijakaumien graafiset esitykset

Frekvenssijakaumat esitetään usein graafisesti käyttäen tilanteeseen sopivaa esitystapaa. Graafinen esitys on havainnollisempi ja "luettavampi" kuin taulukko, jos luokkia on useampia. Tilastolliset ohjelmistot tarjoavat erilaisia graafisia esitystapoja, joista käyttäjä voi itse valita parhaiten tilanteeseen sopivan.

Luokittelu- ja järjestysasteikollisten muuttujien yhteydessä käytetään yleisesti piirakkakuvioita eli sektoridiagrammeja tai pylväsdigrammeja kuvaamaan graafisesti frekvenssijakaumaa. Graafinen esitys voidaan tehdä frekvensseihin, prosentuaalisiin tai suhteellisiin frekvensseihin perustuen.

Esim. 5.1.8. Esimerkin 5.1.3 jakauma graafisesti piirakkakuvion avulla.



Esim. 5.1.9. Joidenkin kvalitatiivisten muuttujien jakaumia ja graafisia esityksiä ks. http://mtl.uta.fi/tilasto/tiltp7/moniste_2.pdf .

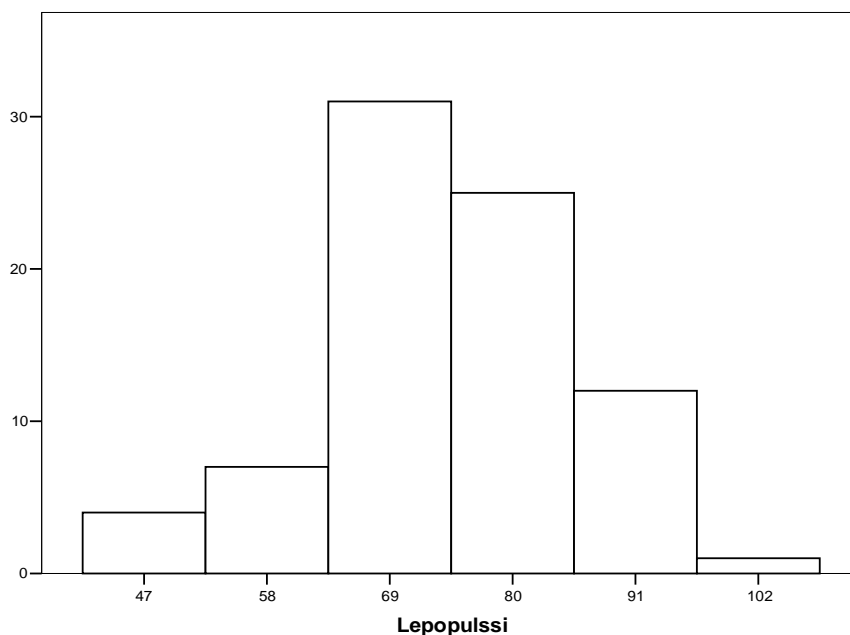
Kvantitatiivisten muuttujien yhteydessä voidaan, tilanteesta riippuen, erottaa diskreetin ja jatkuvan muuttujan eritystavat toisistaan. Jos muuttuja on diskreetti saaden vain "jokusia" arvoja (esim. opiskelijan meneillään oleva opiskeluvuosi, sisarusten lukumäärä), niin kategoristen muuttujien yhteydessä käytetty pylväsdiagrammi sopii käytettäväksi. Pylväsdiagrammin sijaan voi käyttää janadiagrammia. Muuten numeeristen muuttujien esityksessä käytetään frekvenssihistogrammia tai frekvenssimonikulmiota, jotka voi piirtää käyttäen joko absoluuttisia, suhteellisia tai prosentuaalisia frekvenssejä.

Histogrammi piirretään siten, että se muodostuu suorakulmioista, joiden leveys on luokan pituus, korkeus (tasavälisessä luokituksessa) luokan frekvenssi ja kantojen kärkipisteet ovat todellisissa luokkarajoissa. Frekvenssimonikulmio piirretään yhdistäen pisteet, jotka muodostuvat luokakeskuksista ja luokkafrekvensseistä. Ajatellaan ensimmäisen luokan alapuolelle ja viimeisen luokan yläpuolelle luokat, joiden frekvenssi on nolla. Näin kuvio lähtee vaaka-akselilta ja palaa sinne. Kuvion ja x-akselin väliin jäävät pinta-alat ovat samoja sekä histogrammissa että monikulmiossa. Frekvenssihistogrammi/-monikulmio kuvaa jakauman muotoa.

Ks. myös <http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/histogramma>

http://mtl.uta.fi/tilasto/tiltp7/moniste_3.pdf

Esim. 5.1.10. Esimerkin 5.1.5. jakauma graafisesti histogrammin avulla.



Usein kvantitatiivinen muuttuja on nk. normaalijakaumaa noudattava. Tällöin on kyse symmetrisestä, yksihuippuisesta jakaumasta, joka levittäytyy keskiarvon ympärille tietyllä tavalla, ks. luku 7.4.

Summakäyrä piirretään summafrekvensseistä (tai suhteellisista tai prosentuaalisista). Summakäyrää piirrettäessä yhdistetään pisteet, jotka muodostuvat todellisista luokkarajoista sekä summafrekvensseistä (ks. http://mtl.uta.fi/tilasto/tiltp7/moniste_3.pdf). Summakäyrän avulla voidaan arvioida mikä on se muuttujan arvo, jota pienempiä arvoja on p % tai kuinka monta prosenttia arvoista on pienempiä kuin luku a .

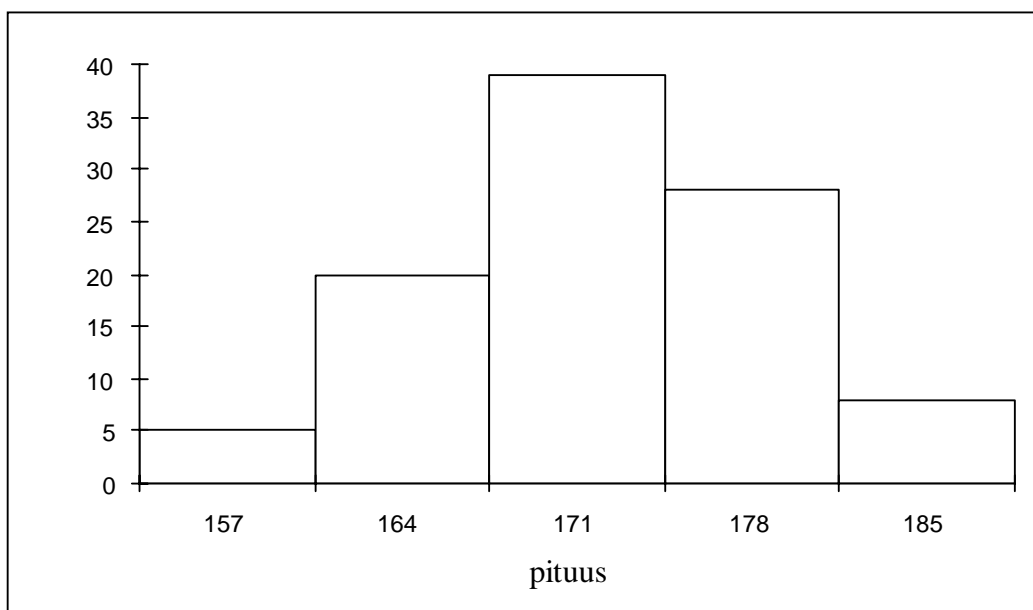
Histogrammi, monikulmio ja summakäyrä piirretään edellä esitetyllä tavalla silloin, kun käytetään tasavälistä luokitusta. Jos käytetään ei-tasavälistä luokitusta tulee frekvenssit suhteuttaa luokan pituuteen pylvään korkeutta määrättäessä.

Esim. 5.1.11. Eräästä aineistosta laskettuna miesopiskelijoiden lukumäärät pituuden mukaan ovat:

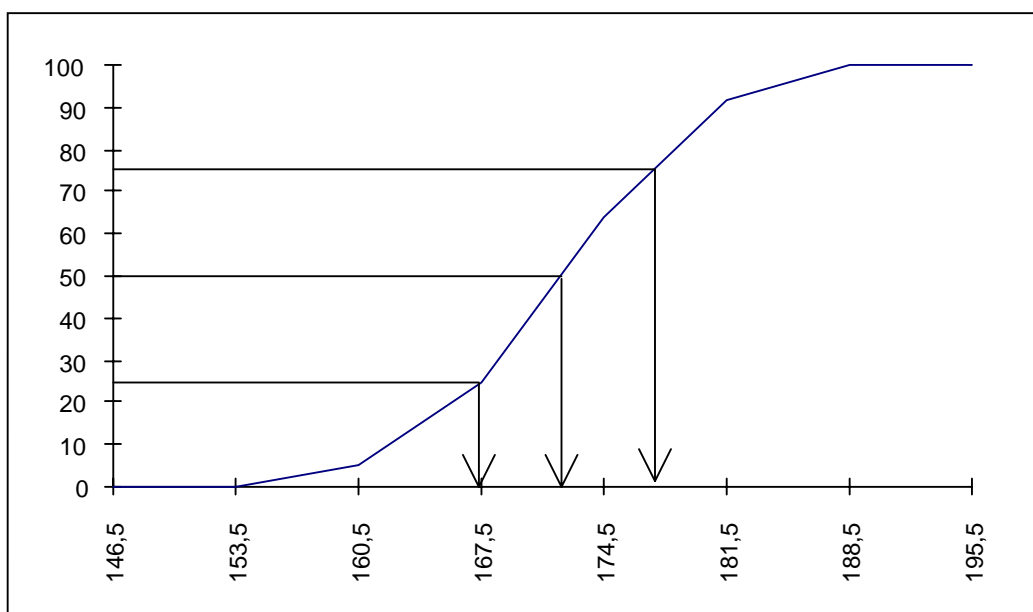
Pituusluokka (cm)	154–160	161–167	168–174	175–181	182–188
Frekvenssi	5	20	39	28	8

Käytetyssä luokituksessa luokkakeskukset ovat 157, 164, 171, 178, 185; todelliset luokkarajat 153,5; 160,5; 167,5; 174,5; 181,5; 188,5 sekä prosentuaaliset frekvenssien kertymät 5 %, 25 %, 64 %, 92 %, 100 %.

Koska kyseessä on jatkuva muuttuja, niin jakauman graafinen esitys on histogrammi:



Summakäyrä on (piirretään pisteistä $(153,5; 0)$, $(160,5; 5)$,...)

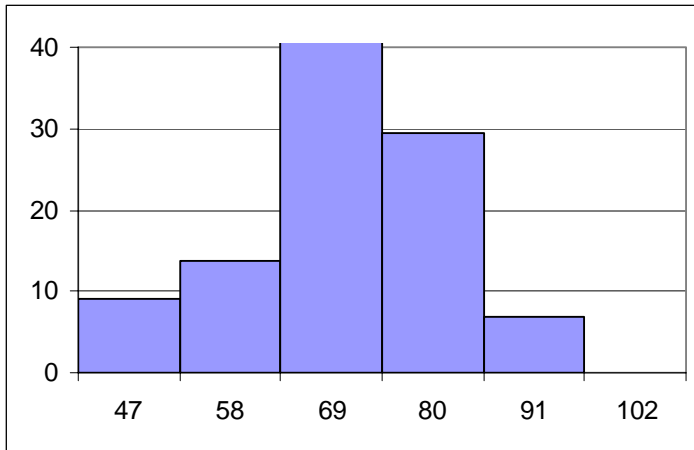


Summakäyrästä voidaan arvioida, että 25 % opiskelijoista oli alle 167 cm, puolet oli pituudeltaan alle 172 cm ja alle 177 cm pitkiä oli 75%.

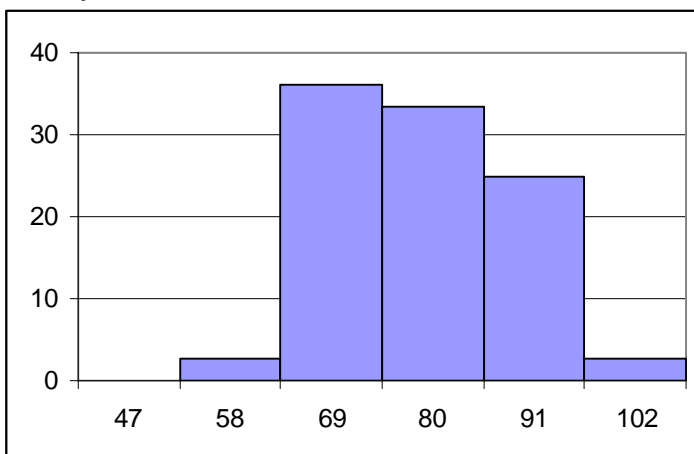
Esim. 5.1.12. Ks. http://mtl.uta.fi/tilasto/tiltp7/moniste_4.pdf esimerkkejä erilaisista jakaumista.

Esim. 5.1.13. Pulssi-muuttujan histogrammit miehillä ja naisilla esimerkin 5.1.7 taulukosta. Histogrammit piirretään käyttäen prosentuaalisia frekvenssejä, jotta vertailu olisi mahdollista.

Lepopulssin jakauma miehillä



Lepopulssin jakauma naisilla



Onko jakaumissa eroa?

Ks. lisätietoja graafisista esityksistä

<http://www.fsd.uta.fi/menetelmaopetus/kuviot/kuviot.html>

SPSS Graafiset esitykset löytyvät valikosta

Graphs

Bar...	pylväs- ja janadiagrammit,
Pie...	piirakat,
Histogram...	frekvenssihistogrammit;
	esityksen valinnan jälkeen annetaan muuttuja(t), jolle graafinen esitys tehdään.

5.1.3 Yksiulotteisen jakauman tunnuslukuja

Tunnusluvun avulla pyritään kuvaamaan muuttujan jakaumaa muuttujan arvoista lasketulla luvulla. Kuvataan esimerkiksi jakauman sijaintia sopivan keskiluvun avulla tai muuttujien arvojen vaihtelua hajontaluvun avulla. Esiteltävät tunnusluvut ovat otoksen

perusteella laskettuja. Niillä voidaan arvioida (estimoida, ks. luku 7.6) populaation vastaavia tunnuslukuja.

Mittaustaso rajoittaa sitä, millaista tunnuslukua jakauman piirteiden kuvaamiseen voidaan käyttää. Syynä tähän tietenkin on se, että mitta-asteikosta riippuu millaisilla matemaattisilla operaatioilla on empiirinen tulkinta. Seuraavassa esityksessä ilmoitetaan aina alin mitta-asteikko, joka vaaditaan tarkasteltavan tunnusluvun käyttöön.

Sijainnin tunnuslukuja

Jakauman sijaintia kuvaavilla tunnusluvuilla mitataan useimmiten muuttujan arvojen "keskimääräistä" suuruutta tai laatua. Tällöin on kyse keskiluvuista, joita ovat moodi, mediaani ja keskiarvo.

KESKILUKUJA

Moodi eli tyypiarvo (*mode*) on se muuttujan arvo/luokka, joka esiintyy useimmin/jossa on eniten havaintoja. Moodin (merk. Mo) ominaisuuksia: ei yksikäsitteinen, ei aina jakauman keskellä, voidaan käyttää jo luokittelutasoisen muuttujan yhteydessä. Jatkuvan muuttujien yhteydessä moodi on harvemmin käyttökelpoinen, tällöin voidaan ilmoittaa moodi-luokka eli luokka, jossa on eniten havaintoja.

Mediaani (*median*) on sellainen muuttujan arvo, jota pienempiä ja suurempia arvoja on yhtä paljon. Mediaanin (merk. Md) ominaisuuksia: ei herkkä poikkeaville arvoille, usein "paras" keskiluku, voidaan käyttää kun järjestyksellä on tulkinta eli kun muuttujan on vähintään järjestysasteikollinen, jos havaintoja on parillinen määrä voidaan mediaani määrittellä numeeristen muuttujien yhteydessä kahden keskimmäisen keskiarvona. Järjestysasteikolla vastaavassa tilanteessa mediaani ei ole yksikäsitteinen.

Seuraavassa otetaan käyttöön summamerkintä. Tarvittaessa voi tutustua summan käyttöön http://mtl.uta.fi/tilasto/tiltp1/syksy2003/moniste_5.pdf

Olkoon muuttujan x arvot tilastoyksiköittäin x_1, x_2, \dots, x_n .

Tällöin muuttujan x keskiarvo (*mean*)

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

Keskiarvo on herkkä poikkeaville arvoille (varsinkin pienissä aineistoissa) ja se on sallittu vasta numeeristen muuttujien yhteydessä.

Summamerkinnän käyttöön voi tarvittaessa tutustua http://mtl.uta.fi/tilasto/tiltp1/syksy2003/moniste_5.pdf

Esim. 5.1.14. Lepopulssi-muuttujan (liite 4) mediaani 74, keskiarvo 73,75.

Esim. 5.1.15. Keskiluvut symmetristen ja vinojen jakaumien tapauksissa, ks. http://mtl.uta.fi/tilasto/tiltp7/moniste_4.pdf .

Ks. lisätietoja ja esimerkkejä keskiluvuista
<http://tilastokeskus.fi/tup/verkkokoulu/data/tt/02/index.html>
<http://www.fsd.uta.fi/menetelmaopetus/keskiluvut/keskiluvut.html>
<http://www.mathstat.helsinki.fi/tilastosanasto/sanasto>

Usein halutaan vertailla muuttujan arvoja tilastoyksiköittäin keskiarvoon, jolloin suoritetaan muuttujan keskistäminen.

Olkoon muuttujan x arvot tilastoyksiköittäin x_1, x_2, \dots, x_n . Keskistetty muuttuja y määritellään

$$y_i = x_i - \bar{x}, \quad i = 1, 2, \dots, n.$$

Jos muuttujalle x tehdään lineaarinen muunnos

$$y_i = ax_i + b, \quad i = 1, 2, \dots, n,$$

niin

$$\bar{y} = a\bar{x} + b.$$

Esim. 5.1.16. Mittayksikkö vaikuttaa keskiarvoon.

Esim. 5.1.17. Voidaan osoittaa, että keskistetyn muuttujan keskiarvo on nolla.

Tarkastellaan ryhmäkeskiarvojen avulla keskiarvon määrittämistä. Olkoon muuttujan x keskiarvot ryhmittäin $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ ja vastaavasti havaintojen määrät n_1, n_2, \dots, n_k

Tällöin koko aineistosta laskettu muuttujan x keskiarvo

$$\bar{x} = (n_1 \bar{x}_1 + \dots + n_k \bar{x}_k) / (n_1 + \dots + n_k).$$

Ryhmäkeskiarvoja voidaan käyttää riippuvuuden tutkimisessa. Vertaillaan tarkasteltavan muuttujan keskiarvoja (tai mediaaneja) ryhmittäin ja tehdään johtopäätelmät mahdollisista eroista. Näistä ryhmäkeskiarvoista käytetään myös nimitystä ehdolliset keskiarvot. Ne on siis laskettu toinen muuttuja ehdollistettuna. Kun lasketaan keskiarvoja, niin muuttujan tulee olla kvantitatiivinen!

Esim. 5.1.18. Lepopulssi-muuttujan keskiarvo naisilla 77,6 ja miehillä 70,6. $M_d = 74$, $M_{dM} = 70$, $M_{dN} = 77,5$ (aineisto liite 4).

Esim. 5.1.19. Pyritään selvittämään vaikuttaako viljelymenetelmä saatavaan satomäärään. Voidaanko satomäärän vaihtelua selittää käytetyllä menetelmällä? Tällöin sanotaan, että

satomäärä = selitettävä eli riippuva muuttuja (usein merk. y) ja
 viljelymenetelmä = selittävä eli riippumaton muuttuja (usein merk. x).

Jos oletetaan, että satomäärä on mitattu kvantitatiivisesti, niin eräs mahdollisuus riippuvuuden selvittämisessä on satomäärän keskiarvojen vertailu ryhmittäin (viljelymenetelmittäin) eli siis ehdollisten keskiarvojen käyttö. Lasketaan satomäärämuuttujasta keskiarvot eri viljelymenetelmillä ja vertaillaan keskiarvoeroja. Jos nämä ehdolliset keskiarvot (ryhmäkeskiarvot) poikkeavat (tarpeeksi) toisistaan sanotaan, että viljelymenetelmä-muuttujalla voidaan selittää satomäärä-muuttujan vaihtelua. Sanotaan, että satomäärä-muuttuja riippuu viljelymenetelmä-muuttujasta. Jos ehdolliset keskiarvot ovat lähes samoja, niin riippuvuutta ei ole. Milloin poikkeama on riittävän suuri? Päättely vaatii tilastollisen testauksen tai sopivan luottamusvälin määrittämisen. Testauksessa päätellään otoksen perusteella se, voidaanko eri viljelymenetelmillä saatujen satomäärien keskiarvot (populaatiossa) olettaa yhtä suuriksi. Luottamusvälin yhteydessä lasketaan otoksen perusteella väli, jolle populaatioiden keskiarvon erotuksen arvellaan kuluva. Luvuissa 7.6 ja 7.7.4 esitellään luottamusväli ja testaus tilanteisiin, joissa selittävä muuttuja on kaksiluokkainen.

Esim. 5.1.20. Satomäärät istutustiheyksittäin ja lajikkeittain (L1, L2, L3).

	istutustiheys (1000 tainta/ha)			
	10	20	30	40
L1	7,9;9,2;10,5	11,2;12,8;13,3	12,1;12,6;14,0	9,1;10,8;12,5
L2	8,1;8,6;10,1	11,5;12,7;13,7	13,7;14,4;15,4	11,3;12,5;14,5
L3	15,3;16,1;17,5	16,6;18,5;19,2	18,0;20,8;21,0	17,2;18,4;18,9

Sadon ehdolliset keskiarvot:

	istutustiheys (1000 tainta/ha)				
	10	20	30	40	
L1	9,2	12,4	12,9	10,8	11,3
L2	8,9	12,6	14,5	12,8	12,2
L3	16,3	18,1	19,9	18,2	18,1
	11,5	14,4	15,8	13,9	13,9

MUITA SIJAINNIN TUNNUSLUKUJA

Alakvartiili (*lower quartile*) ja yläkvartiili (*upper quartile*) ovat mediaanin kaltaisia tunnuslukuja, jotka kuvaavat jakauman sijaintia. Alakvartiili on luku, joka jakaa muuttujan arvot kahteen osaan siten, että korkeintaan 25% havaituista arvoista on pienempiä kuin alakvartiili. Yläkvartiili on luku, joka jakaa muuttujan arvot kahteen osaan siten, että korkeintaan 75% arvoista on pienempiä kuin yläkvartiili.

Alakvartiili, mediaani ja yläkvartiili jakavat muuttujan arvot neljään havaintomääriltään yhtä suuriin osiin. Yhdessä näitä kolmea tunnuslukua kutsutaan kvartiileiksi.

Muuttujan arvot voidaan jakaa viiteen, kuuteen, jne. havaintomääriltään yhtä suureen osaan. Yleisesti näitä osiin jakavia tunnuslukuja kutsutaan fraktiileiksi.

Esim. 5.1.21. Esimerkistä 5.1.11 alakvartiili ≈ 167 , mediaani ≈ 172 ja yläkvartiili ≈ 177 .

Esim. 5.1.22. Erään tilastollisen ohjelmiston tuottamat tunnusluvut Pulssi-muuttujalle (aineisto liite 4).

Quantiles

maximum	100,0%	100,00
	99,5%	100,00
	97,5%	94,00
	90,0%	88,00
quartile	75,0%	81,50
median	50,0%	74,00
quartile	25,0%	66,50
	10,0%	58,00
	2,5%	50,00
	0,5%	45,00
minimum	0,0%	45,00

Moments

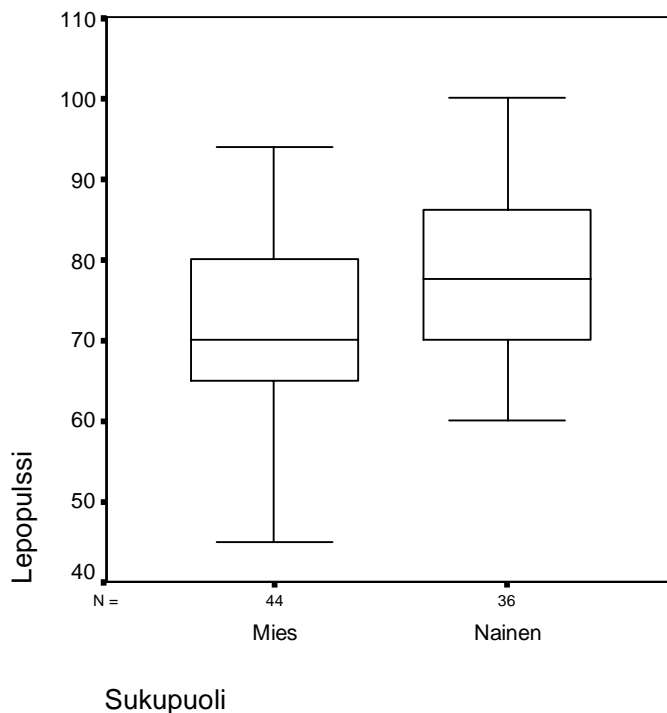
Mean	73,75000
Std Dev	11,12814
Std Err Mean	1,24416
upper 95% Mean	76,22645
lower 95% Mean	71,27355
n	80

LAATIKKO-JANA-KUVIO

Numeeristen muuttujien yhteydessä jakaumaa tai ehdollisia jakaumia voidaan havainnollistaa (käyttäen erilaisia fraktiileja) laatikko-jana -kuviolla (*box and whiskers display, boxplot*).

Ks. myös <http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/jana-laatikko-diagramma> , <http://www.fsd.uta.fi/menetelmaopetus/kuviot/kuviot.html>

Esim. 5.1.23. Laatikko-jana -kuvio Pulssi-muuttujasta (liite 4) miehillä ja naisilla.



Laatikko-jana -kuviossa laatikon keskimäinen viiva on pulssin mediaanin kohdalla ja laatikon ylä- ja alareunat ylä- ja alakvartileissa. "Laatikot" sisältävät 50 % ryhmän havainnoista. Kuviossa nähdään, että miesten jakauma on alempana kuin naisten.

Vaihtelua mittaavia tunnuslukuja

Muuttujan arvot vaihtelevat tilastoyksiköstä toiseen. Vaihtelun voimakkuutta pyritään mittaamaan erilaisia tunnuslukuja käyttäen (vrt. keskiluvut). Rajoitutaan tässä tarkastelemaan vain numeeristen muuttujien yhteydessä käytettäviä jakauman hajontaa mittaavia tunnuslukuja. Kategoristen muuttujien yhteydessä hajonnan käsite vaikeampi mieltää ja hajontalukujen käyttö käytännössä melko harvinaista.

Numeeristen muuttujien yhteydessä yleensä ilmoitetaan muuttujan vaihteluväli ja lasketaan nk. varianssi.

Olkoon muuttujan x arvot tilastoyksiköittäin x_1, x_2, \dots, x_n . Muuttujan x (otos)varienssi s^2 (*variance*) määritellään

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

Kaavakokoelmassa (liite 2) kaavassa (2) on varianssin, usein käsinlaskuja helpottava, laskukaava. Varianssin (positiivinen) neliöjuuri s on nimeltään keskihajonta (*hajonta*) (*standard deviation*).

Varianssi mittaa kuinka tiiviisti muuttujien arvot ovat keskittyneet keskiarvon ympärille. Jos kaikki mittaustulokset ovat samoja, niin $s^2 = 0$, muulloin $s^2 > 0$. Varianssi on riippumaton jakauman sijainnista, mutta riippuu käytetystä mittayksiköstä.

Tarvittaessa merkitään muuttujan x hajontaa s_x ja muuttujan y hajontaa s_y (ja varianssit vastaavasti alaindeksoiden).

Esim. 5.1.24. Ohessa lapsen Sisarusten lukumäärä –muuttujan frekvenssijakauma.

Sisarusten lukumäärä	Frekv.
0	56
1	39
2	13
3	10
4	5
5	2
6	1
Yht.	126

Aineistossa lapsella on keskimäärin 1,04 sisarusta, koska $\bar{x} = (0 \cdot 56 + 1 \cdot 39 + \dots + 6 \cdot 1) = 131/126 \approx 1,04 \approx 1$.

Sisarusten lukumäärän varianssi $s^2 \approx (56(0-1)^2 + 39(1-1)^2 + \dots + (6-1)^2)/(126-1) \approx 1,69$

Esim. 5.1.25. Tutkittiin kahden lisäaineen (A ja B) vaikutusta teräksen kovuuteen. Koska teräksen tuote-erien laatu vaihtelee, poimittiin näytteet 10 tuote-erästä, joista kukin jaettiin edelleen kahtia. Toiseen osaan lisättiin lisäainetta A ja toiseen lisäainetta B. Mitattiin kovuusindeksi (ja laskettiin erotukset)

Tuote-erä	1	2	3	4	5	6	7	8	9	10
Lisäaine A	22	26	29	22	31	34	31	20	33	34
Lisäaine B	27	25	31	27	29	41	32	27	32	34
EROTUS	-5	1	-2	-5	2	-7	-1	-7	1	0

Onko perusteita pitää toista lisäainetta parempana kuin toista? Lasketaan kovuusindeksien erotukset tuote-erittäin ja sitten erotuksen keskiarvo ja hajonta. Erotuksista laskettu keskiarvo on $-2,3$ (tai $2,3$, jos erotukset laskettu toisin päin) ja $s^2 \approx ((-5-(-2,3))^2 + (1-(-2,3))^2 + \dots + (0-(-2,3))^2)/(10-1) \approx 11,79$, joten hajonta $3,4$.

Millainen havaintomatsiisi on?

Esim. 5.1.26. Tutkitaan sairauden vuoksi työstä poissaoloja. Halutaan vertailla päivätöitä ja yötyötä tekeviä. Valitaan satunnaisesti molemmista ryhmistä yöntekijöitä ja lasketaan heidän sairauspoissaolopäivien lukumäärät kuluneen vuoden ajalta saadaan oheiset tulokset:

Yötyö	15	10	10	7	7	4	9	6	10	12
Päivätyö	8	9	2	0	0	9	9	7	3	3

Nyt $n_{yö} = n_{päivä} = 10$, $\bar{y}_{yö} = 90/10$, $\bar{y}_{päivä} = 50/10$, $SS_{yö} = 900 - 90^2/10 = 90$,
($15+10+10+7+7+4+9+6+10+12=90$ ja $15^2+10^2+10^2+7^2+7^2+4^2+9^2+6^2+10^2+12^2=900$)

$$SS_{päivä} = 378 - 50^2/10 = 128,$$

($8+9+2+0+0+9+9+7+3+3=50$ ja $8^2+9^2+2^2+0^2+0^2+9^2+9^2+7^2+3^2+3^2=378$).

Poissaolopäivien ehdolliset hajonnat ovat siten $s_{yö} = \sqrt{90/9}$ ja $s_{päivä} = \sqrt{128/9}$. Millainen havaintomatsiisi on? (Aineisto: Ott & Mendenhall (1985) *Understanding Statistics*.)

Ks. lisätietoja ja esimerkkejä hajontaluvuista

<http://tilastokeskus.fi/tup/verkkokoulu/data/tt/02/index.html>

<http://www.fsd.uta.fi/menetelmaopetus/hajontaluvut/hajontaluvut.html>

<http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/varianssi>

Jos muuttujan arvoista vähennetään sen keskiarvo ja erotus jaetaan hajonnalla on muuttuja standardoitu; muuttujan x standardointi:

$$z_i = (x_i - \bar{x})/s, \quad i = 1, 2, \dots, n.$$

Jos muuttujalle x tehdään lineaarinen muunnos

$$y_i = ax_i + b, \quad i = 1, 2, \dots, n,$$

niin $s_y = |a|s_x$.

Ks.

http://mtl.uta.fi/tilasto/tiltp1/syksy2003/moniste_6.pdf ,
http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/standardoitu_muuttuja

Esim. 5.1.27. Voidaan osoittaa, että standardoidun muuttujan keskiarvo on 0 ja keskihajonta 1.

Esim. 5.1.28. Opiskelija osallistui tentteihin A ja B saaden pisteet 25 ja 24. Tentissä A tuloksen keskiarvo oli 20 ja keskihajonta 4. Vastaavat luvut tentissä B olivat 20 ja 2. Kummassa tentissä opiskelija menestyi suhteellisesti paremmin? Standardoidut arvot ovat 5/4 ja 2, joten menestyminen tentissä B oli parempi.

Esim. 5.1.29. Normaalijakauma, ks. http://mtl.uta.fi/tilasto/tiltp7/moniste_6.pdf

Variaatiokerroin (*coefficient of variation*) on hajonnan tunnusluku, joka määritellään s/\bar{x} . Variaatiokerroin edellyttää suhteasteikollista mittauksia. Variaatiokerroin on riippumaton mittayksiköstä. Variaatiokerroin soveltuu myös tilanteisiin, jossa vertaillaan kahden tai useamman sijainniltaan tai havaintoarvoiltaan täysin erilaisen aineiston hajontoja keskenään, esim. elefanttien ja hiiren painojakauman vertailu.

Ks. tarkemmin <http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/variaatiokerroin>

Muita tunnuslukuja

Voidaan mitata esimerkiksi jakauman huipukkuutta ja vinoutta.

Ks. tunnusluvuista myös <http://tilastokeskus.fi/tup/verkkokoulu/data/tt/02/index.html>

SPSS

Jakaumaa kuvaavia erilaisia tunnuslukuja saadaan mm. seuraavilla tavoilla:

```
Analyze
  Descriptive Statistics>
    Frequencies...
      saadaan halutuista muuttujista mm. keskiarvo, mediaani,
      fraktiilit, moodi, keskihajonta, varianssi, pienin arvo, suurin arvo
    Descriptives...
      saadaan halutuista muuttujista
      mm. keskiarvo, keskihajonta, varianssi,
      pienin arvo, suurin arvo, vaihteluväli
    Explore...
      saadaan mm. keskiarvo, keskihajonta,
      varianssi, pienin arvo, suurin arvo, vaihteluväli sekä
      tunnusluvut ehdollisina antamalla (kvalitatiivinen) selittäväksi
muuttujaksi
  Compare Means>
    Means...
      saadaan tunnusluvut ehdollisina antamalla ehtomuuttuja
      (kvalitatiivinen) selittäväksi muuttujaksi.
```

Ehdollisia jakaumia voidaan havainnollistaa myös laatikko-jana-kuvion (boxplot) avulla. Kuvio perustuu eri fraktiileihin ja saadaan tehdyksi valikosta

Graphs

Boxplot...

antamalla Variable-kohtaan tutkittava muuttuja ja
Category-kohtaan ryhmittelymuuttuja.

5.2 Kaksiulotteinen jakauma

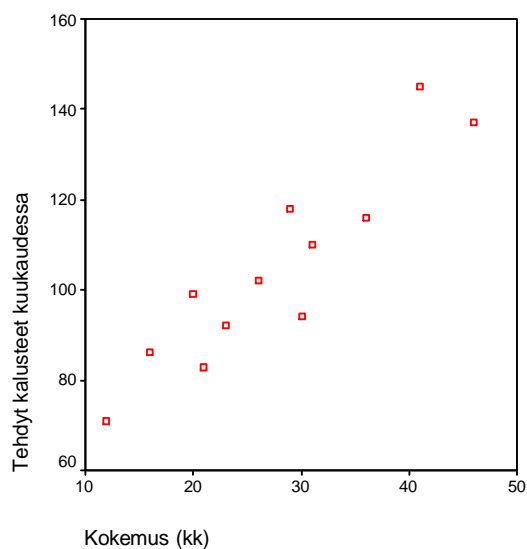
5.2.1 Pisteparvi

Esim. 5.2.1. Eräs pieni tehdas työllistää 12 työntekijää, jotka tekevät puutarhakalusteita. Työntekijät saavat palkan, joka perustuu tehtyjen kalusteiden lukumäärään. Tehtaan omistajalla on käsitys, että työntekijän kokemus vaikuttaa merkittävästi siihen, paljonko hän ehtii tehdä kalusteita. Erään tyypillisen kuukauden aikana kirjattiin tulokset ja saatiin oheinen aineisto.

Työntekijän kokemus kuukausina (x)	Työntekijän tekemien kalusteiden lukumäärä kuukauden aikana (y)
36	116
26	102
23	92
16	86
29	118
20	99
21	83
31	110
30	94
12	71
41	145
46	137

Esimerkissä 5.2.1 on kyse kaksiulotteisesta jakaumasta. Kaksiulotteisessa jakaumassa tarkastellaan kahta muuttujaa samanaikaisesti. Kaksiulotteisen jakauman avulla pyritään selvittämään voidaanko y :n vaihtelua selittää x :n avulla eli riippuuko y -muuttuja x -muuttujasta. Kaksiulotteinen jakauma voidaan esittää graafisesti pisteparven eli korrelaatiodiagrammin (*scatter plot*, *scatter diagram*) avulla silloin, kun selitettävä muuttuja on numeerinen. Pisteparvi muodostetaan piirtämällä pisteet $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ koordinaatistoon selitettävän muuttujan ollessa pystyakselilla.

Esim. 5.2.2. Pisteparvi esimerkin 5.2.1 aineistosta.



Pisteparvi antaa hyvän yleissilmäyksen mahdollisesta riippuvuudesta. Esimerkin 5.2.2 tilanteessa riippuvuus on hyvin suoranomaista.

Esim. 5.2.3. Erilaisia riippuvuuksia numeeristen muuttujien yhteydessä, ks. http://mtl.uta.fi/tilasto/tiltp7/moniste_8.pdf

Pisteparven avulla riippuvuutta selvitetäessä on tulkinnassa huomioitava x-muuttujan mitta-asteikko.

Esim. 5.2.4. Riippuvuuksia, kun x-muuttuja kategorinen, ks. http://mtl.uta.fi/tilasto/tiltp7/moniste_9.pdf

SPSS Pisteparvi saadaan valikosta
 Graphs
 Scatter...
 antamalla selitettävä y pystyakselille ja selittävä x vaaka-akselille.

5.2.2 Ristiintaulukko

Esim. 5.2.5. Automallien lukumäärät auton koon ja valmistusmaan mukaan.

		Valmistusmaa			
		USA	Eur.	Japani	
Koko	Iso	36	4	2	42
	Kesk.	53	17	54	124
	Pieni	26	19	92	137
		115	40	148	303

Koko-muuttujan ehdolliset prosenttijakaumat eli koko-muuttujan prosentuaaliset jakaumat erikseen valmistusmaittain:

		Valmistusmaa		
		USA	Eur.	Japani
Koko	Iso	31,30%	10,00%	1,35%
	Kesk.	46,09%	42,50%	36,49%
	Pieni	22,61%	47,50%	62,16%
		100%	100%	100%

Esimerkissä 5.2.5 oli tarkasteltu havaintojen määrää kahden muuttujan luokissa. Tämän nk. ristiintaulukon avulla voidaan tutkia kahden muuttujan välistä yhteyttä. Tutkitaan sitä, onko auton koko riippuvainen valmistusmaasta. Riippuvuustarkastelu tehdään vertailemalla selitettävän muuttujan ehdollisia prosenttijakaumia keskenään. Jos jakaumat poikkeavat toisistaan, niin sanotaan selitettävän muuttujan riippuvan selittäjästä (tässä koon riippuvan valmistusmaasta). Jos ehdolliset prosenttijakaumat ovat lähes samanlaiset, niin riippuvuutta ei ole. Tässäkin halutaan tehdä otoksen perusteella yleistys koko populaatioon, joten ristiintaulukon lisäksi tilastollinen testaus on tarpeen, ks. luku 7.7.3.

Olkoon muuttujan x arvot jaettu luokkiin E_1, E_2, \dots, E_J ja y muuttujan arvot luokkiin F_1, F_2, \dots, F_I . Tällöin kaksiulotteinen frekvenssijakauma eli ristiintaulukko on

		x			
		E_1	E_2	\dots	E_J
	F_1	f_{11}	$f_{12} \dots$	f_{1J}	$f_{1\cdot}$
	F_2	f_{21}	$f_{22} \dots$	f_{2J}	$f_{2\cdot}$
y	\vdots	\vdots	\vdots	\vdots	\vdots
	F_I	f_{I1}	$f_{I2} \dots$	f_{IJ}	$f_{I\cdot}$
		$f_{\cdot 1}$	$f_{\cdot 2} \dots$	$f_{\cdot J}$	$f_{\cdot \cdot}$

missä f_{ij} = solun (F_i, E_j) solufrekvenssi, $f_{i\cdot}$ on i . rivin frekvenssien summa ja $f_{\cdot j}$ on sarakkeen j frekvenssien summa. Reunafrekvenssit $f_{1\cdot}, f_{2\cdot}, \dots, f_{I\cdot}$ muodostavat y -muuttujan frekvenssijakauman ja reunafrekvenssit $f_{\cdot 1}, f_{\cdot 2}, \dots, f_{\cdot J}$ muuttujan x frekvenssijakauman. Sarakkeiden frekvenssit muodostavat y :n ehdolliset jakaumat. Sarakkeiden prosenttifrekvenssit muodostavat y :n ehdolliset prosenttijakaumat.

Ristiintaulukosta, joka siis on kaksiulotteisen luokiteltu jakauma, käytetään myös nimityksiä $I \times J$ -frekvenssitaulu (kontingenssitaulu). Ristiintaulukko voidaan muodostaa vaikka x ja y olisivat vain luokitteluasteikollisia. 2×2 -frekvenssitaulusta käytetään nimitystä nelikenttä.

Esim. 5.2.6. Tutkittiin onko tavaratalon koolla (mitattuna henkilöstömäärällä) vaikutusta markkinointisuunnitelman olemassaoloon. Ohessa tutkimustulokseen liittyvä ristiintaulukko.

	Markkinointisuunnitelma			
		on	ei	
Henkilöstö-	alle 100	13	10	23
määrä	100 – 500	18	12	30
	yli 500	32	6	38

Koska henkilöstön määrä on selittävä muuttuja, lasketaan tässä prosenttijakaumat riveittäin. Saadaan ehdolliset prosenttijakaumat

	Markkinointisuunnitelma			
		on	ei	
Henkilöstö-	alle 100	56,6 %	43,5 %	100 %
määrä	100 – 500	60,0 %	40,0 %	100 %
	yli 500	84,2 %	15,8 %	100 %

Henkilöstömäärällä näyttäisi olevan vaikutusta markkinointisuunnitelman olemassaoloon. Suurista yrityksistä suunnitelma oli 84,2 %:lla kun taas alle 100 hengen yrityksillä vain 56,6 %:lla.

Kolmas muuttuja voidaan ottaa mukaan riippuvuustarkasteluihin esimerkiksi ehdollisten pisteparvien tai ehdollisten ristiintaulukoiden avulla. Voidaan myös laskea ristiintaulukon "soluissa" keskiarvot kolmannen muuttujan arvoista (ks. esim. 5.1.20). On muistettava, että keskiarvon lasku edellyttää kvantitatiivista mittaamista.

Ks. ristiintaulukon muodostamisesta

<http://www.fsd.uta.fi/menetelmaopetus/ristiintaulukointi/ristiintaulukointi.html>

SPSS Ristiintaulukointi ja testaus tehdään valikosta

Analyze

Descriptive Statistics>

Crosstabs... annetaan sarake- ja rivimuuttujat, lisämääreinä

Statistics... -painike>Chi-square, χ^2 -testisuure

Cells... -painike, ehdolliset prosenttijakaumat, "suunta" valitaan siten, että saadaan selitettävän prosenttijakaumat selittäjän luokissa.

SPSS muodostaa ristiintaulukon siten, että molempien muuttujien jokainen arvo on omana luokkanaan. Jos on tarve yhdistellä muuttujien arvoja, tehdään se muodostamalla uusi muuttuja havaintomatriisiin (Transform>Recode>). Kvantitatiivista muuttujaa voi halutessaan käyttää ristiintaulukoinnissa, kunhan sen ensin luokittelee tekemällä uuden muuttujan havaintomatriisiin.

5.2.3 Kaksiulotteisen jakauman tunnuslukuja

Kaksiulotteisten jakaumien tunnusluvuilla pyritään mittaamaan riippuvuuden voimakkuutta. Tunnuslukuja määritellään joko luokiteltuun tai luokittelemattomaan jakaumaan perustuen. Yleisesti käytössä olevat kaksiulotteisen jakauman tunnusluvut ovat numeeristen muuttujien yhteydessä lineaarista riippuvuutta mittaava korrelaatiokerroin, järjestysasteikkosilta muuttujilta järjestyskorrelaatiokertoimet ja ristiintaulukosta laskettava kontingenssikerroin.

Jos muuttujat x ja y ovat numeerisia, niin niiden välistä lineaarista eli suoranomaista riippuvuutta voidaan mitata korrelaatiokertoimen avulla. Korrelaatiokerroin mittaa miten tiiviisti pisteparvi on keskittynyt pisteparveen ajatellun suoran ympärille.

Lineaarinen riippuvuus voi olla joko positiivista, negatiivista tai täydellistä.

Olkoon muuttujan x arvot x_1, x_2, \dots, x_n ja y muuttujan arvot y_1, y_2, \dots, y_n . Tällöin muuttujien välinen korrelaatiokerroin r (*coefficient of correlation*), Pearsonin tulomomenttikorrelaatiokerroin r , määritellään

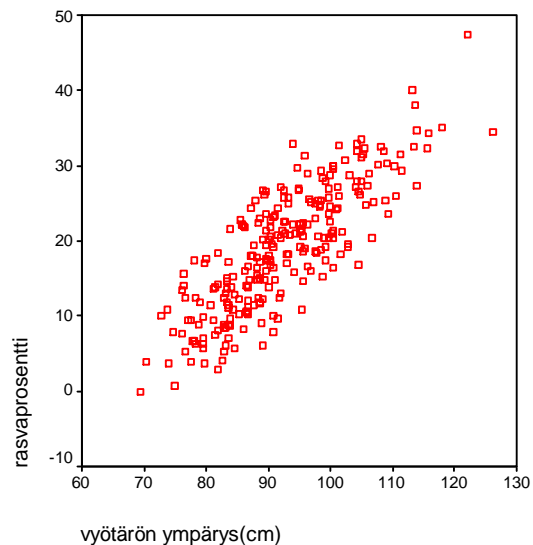
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Kaavakokoelmassa (liite 2, kaava (4)) on korrelaatiokertoimen, usein käsinlaskua helpottava, laskukaava. Tarvittaessa merkitään r_{xy} , r_{zy} ...

Korrelaatiokertoimen ominaisuuksia:

- 1) $-1 \leq r \leq 1$
- 2) Jos pisteet kaikki samalla nousevalla suoralla $r = 1$. Jos pisteet samalla laskevalla suoralla $r = -1$.
- 3) Jos lineaarinen riippuvuus on negatiivista $r < 0$. Jos lineaarinen riippuvuus positiivista $r > 0$. Jos lineaarista riippuvuutta ei ole $r \approx 0$.
- 4) HUOM!! Jos $r \approx 0$, niin tarkasteltavien muuttujien välillä ei ole *lineaarista* riippuvuutta, mutta voi toki olla muunlaista riippuvuutta.

Esim. 5.2.7. Erääseen tutkimukseen osallistuneiden miesten rasvaprosentin ja vyötärön ympärysmittan välinen riippuvuus. Korrelaatiokerroin $r = 0,825$



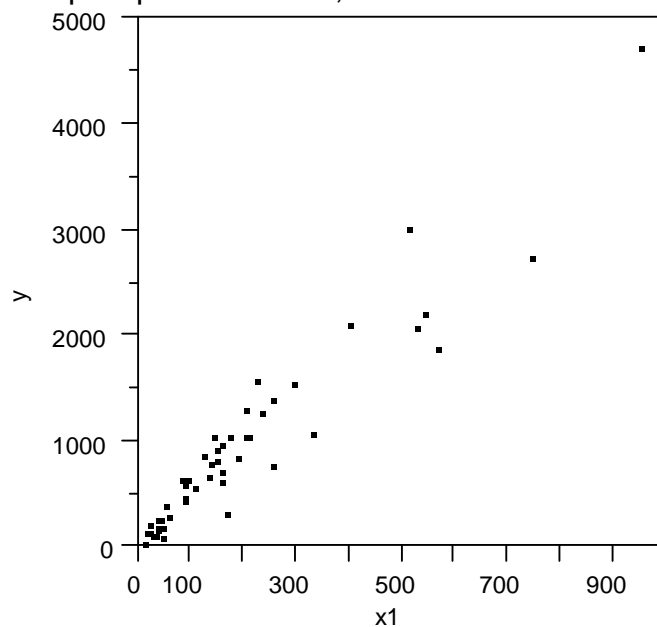
Esim. 5.2.8. Korrelaatiokertoimen lasku, esimerkin 5.2.1 aineisto.

x_i	y_i	x_i^2	$x_i y_i$	y_i^2
36	116	1296	4176	13456
26	102	676	2652	10404
23	92	529	2116	8464
16	86	256	1376	7396
29	118	841	3422	13924
20	99	400	1980	9801
21	83	441	1743	6889
31	110	961	3410	12100
30	94	900	2820	8836
12	71	144	852	5041
41	145	1681	5945	21025
46	137	2116	6302	18769
331	1253	10241	36794	136105

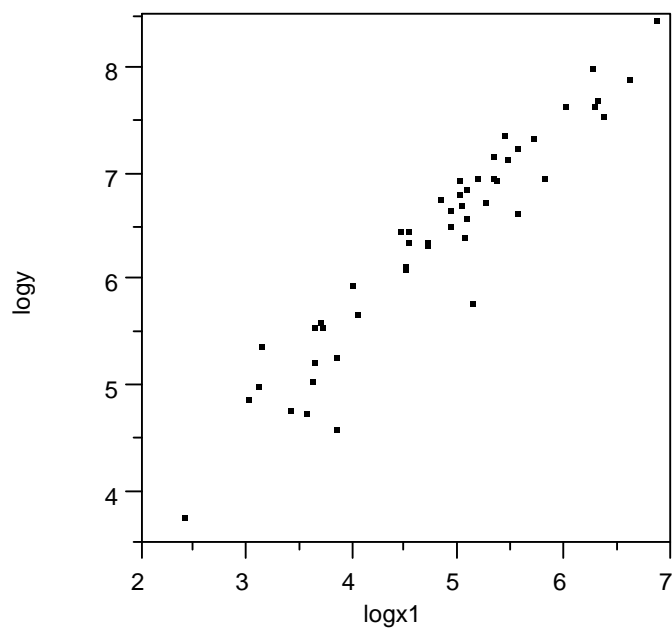
$$r = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i) / n}{\sqrt{(\sum x_i^2 - (\sum x_i)^2 / n)(\sum y_i^2 - (\sum y_i)^2 / n)}}$$

$$= \frac{36794 - 331 \cdot 1253 / 12}{\sqrt{(10241 - 331^2 / 12)(136105 - 1253^2 / 12)}} = 0,922$$

Esim. 5.2.9. Oheisesta pisteparvesta $r = 0,9559$.



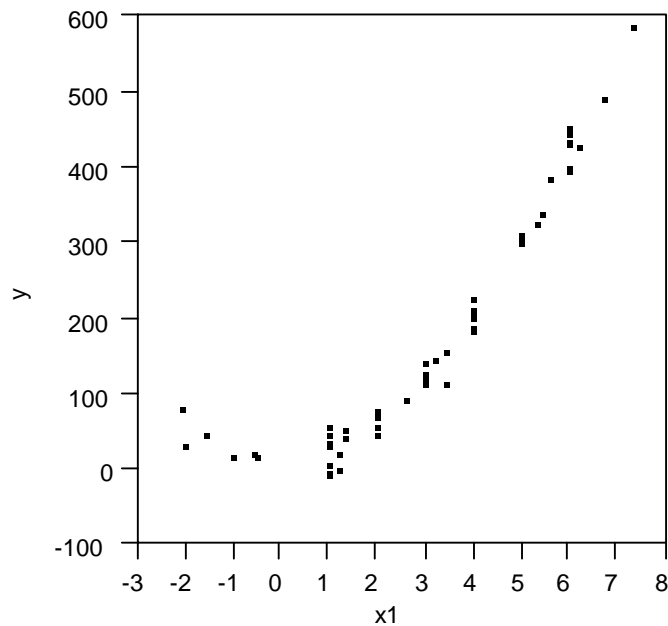
Esim. 5.2.10. Esimerkin 5.2.9 tilanne, kun logaritmoidaan muuttujat. Saadaan $r = 0,9537$.



Korrelaatiomatriisi:

Variable	y	x1	logy	logx1
y	1,0000	0,9559	0,8446	0,8407
x1	0,9559	1,0000	0,7967	0,8566
logy	0,8446	0,7967	1,0000	0,9537
logx1	0,8407	0,8566	0,9537	1,0000

Esim. 5.2.11. Muuttujien välillä riippuvuutta, joka ei ole lineaarinen.



Esim. 5.2.12. Pisteparvia ja korrelaatiokertoimia,
http://mtl.uta.fi/tilasto/tiltp7/moniste_8.pdf .

Muuttujien väliset korrelaatiokertoimet esitetään usein korrelaatiomatriisina

	X_1	X_2	...	X_p
X_1	1	r_{12}	...	r_{1p}
X_2	r_{21}	1	...	r_{2p}
.
.
.
X_p	r_{p1}	r_{p2}	...	1

missä r_{ij} ($= r_{ji}$) on muuttajien x_i ja x_j välinen korrelaatiokerroin. Ks. esim. 5.2.10.

Jos muuttujille x ja y tehdään lineaariset muunnokset

$$u_i = ax_i + b, \quad i = 1, 2, \dots, n,$$

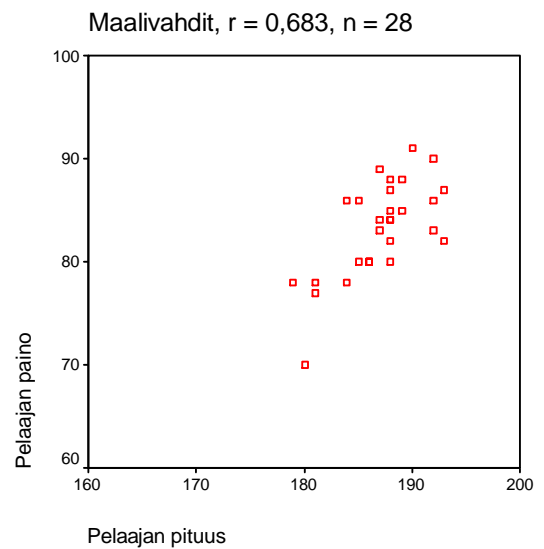
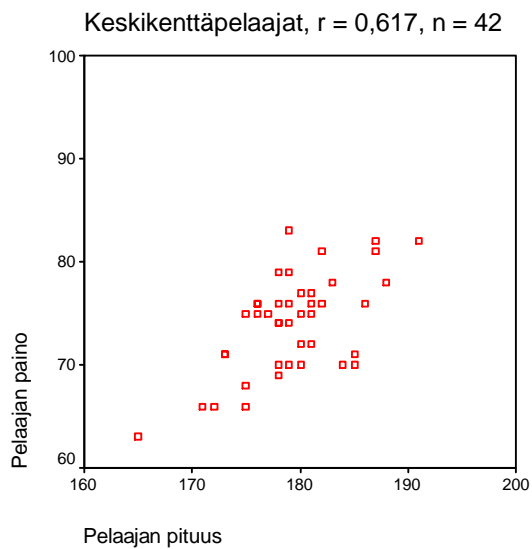
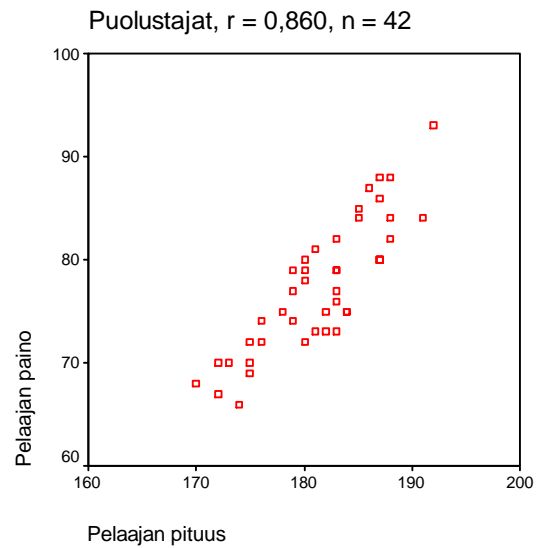
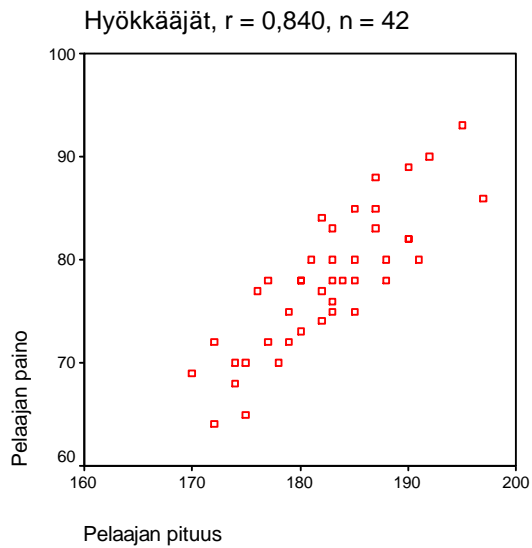
$$v_i = cy_i + d, \quad i = 1, 2, \dots, n,$$

niin $r_{uv} = r_{xy}$, jos $ac > 0$ ja $r_{uv} = -r_{xy}$, jos $ac < 0$.

Esim. 5.2.13. Mittayksikön vaihdokset eivät siis vaikuta korrelaatiokertoimiin.

Pisteparvia ja korrelaatiokertoimia voidaan tarkastella myös ehdollisina.

Esim. 5.2.14. Voidaan tarkastella painon riippuvuutta pituudesta tytöillä ja pojilla erikseen.

Esim. 5.2.15. Veikkausliigan pelaaja 2006.Aineisto [Jalkapalloilijat_2006.sav](#) tai [Jalkapalloilijat_2006.xls](#). sivulla<http://www.uta.fi/%7Estrale/tiltp1/aineistoja.html>

Kahden muuttujan x ja y välistä lineaarista riippuvuutta, kun kolmas muuttuja z on vakioita (vaikutus poistettu), mitataan osittaiskorrelaatiokertoimella

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

Esim. 5.2.16. Tarkastellaan eri ikäisiä (peruskoululaisia ja lukiolaisia) poikia (n=152). SPSS-ohjelman antamia analyysituloksia.

Correlations

		ikä	pituus	paino	cooper
ikä	Pearson Correlation	1	,807**	,768**	,399**
	Sig. (2-tailed)		,000	,000	,000
	N	152	152	152	152
pituus	Pearson Correlation	,807**	1	,892**	,236**
	Sig. (2-tailed)	,000		,000	,003
	N	152	153	153	153
paino	Pearson Correlation	,768**	,892**	1	,102
	Sig. (2-tailed)	,000	,000		,210
	N	152	153	153	153
cooper	Pearson Correlation	,399**	,236**	,102	1
	Sig. (2-tailed)	,000	,003	,210	
	N	152	153	153	153

** . Correlation is significant at the 0.01 level (2-tailed).

Correlations

Control Variables			cooper	paino	pituus
ikä	cooper	Correlation	1,000	-,349	-,160
		Significance (2-tailed)	.	,000	,050
		df	0	149	149
	paino	Correlation	-,349	1,000	,719
		Significance (2-tailed)	,000	.	,000
		df	149	0	149
	pituus	Correlation	-,160	,719	1,000
		Significance (2-tailed)	,050	,000	.
		df	149	149	0

$r_{\text{cooper, paino}} = 0,102$, $r_{\text{cooper, paino} \cdot \text{ikä}} = -0,349$, $r_{\text{cooper, pituus}} = 0,236$, $r_{\text{cooper, pituus} \cdot \text{ikä}} = -0,16$, $r_{\text{ikä, pituus}} = 0,807$ $r_{\text{ikä, paino}} = 0,768$.

Lineaarisen riippuvuuden mittarina voidaan siis käyttää korrelaatiokerrointa. Jos tarkasteltavat muuttujat ovat vähintään järjestysasteikollisia, voidaan riippuvuuden tunnuslukuna käyttää järjestyslukuihin perustuvia tunnuslukuja. N_s järjestyskorrelaatiokertoimia voidaan pitää korrelaatiokertoimen vastineina järjestysasteikollisten muuttujien tapauksessa. Spearmanin järjestyskorrelaatiokerroin r_s sekä Kendallin järjestyskorrelaatiokerroin r_k mittaavat monotonista riippuvuutta ja ovat arvoiltaan aina välillä $[-1, 1]$, ja arvo itseisarvoltaan sitä suurempi mitä voimakkaammasta monotonisesta riippuvuudesta kyse.

Ristiintaulukosta riippuvuuden voimakkuutta voidaan mitata kontingenssikertoimen C avulla. Kontingenssikerroin mittaa minkä tyyppistä riippuvuutta hyvänsä.

Ks. lisätietoja korrelaatiokertoimista

<http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/korrelaatiokerroin>

<http://www.fsd.uta.fi/menetelmaopetus/korrelaatio/korrelaatio.html> ja

SPSS Korrelaatiokertoimen (korrelaatiomatriisin) voi laskea valikosta
Analyze
Correlate>
Bivariate... (Pearson)
antamalla halutut muuttujat.

Tällä opintojaksolla riippuvuustarkastelut esiteltiin aluksi kuvaileviin analyysihin perustuen. Luvussa 7 luodaan esimerkinomaisesti silmäys tilastolliseen testaukseen. Vasta seuraavilla perusopintojaksoilla esitellään tarkemmin, laajemmin ja perustellummin testaukset ja näin saadaan paremmat valmiudet empiirisen tutkimuksen tekemiseen.

6 AIKASARJOISTA

Aikasarja on joukko havaintoja, jotka on saatu tarkasteltavasta ilmiöstä peräkkäisinä ajankohtina t_1, t_2, \dots, t_n . Näihin ajankohtiin liittyvät mittaustulokset $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ muodostavat aikasarjan, jonka pituus on n . Usein mittaustulokset ovat tasavälein esim. vuosittain, kuukausittain, päivittäin.

Esim. 6.1. a) Tampereen kaupungin asukasluku vuosina 1954-2005, b) työttömyysaste kuukausittain vuosilta 1980-94 c) keskilämpötilat vuosittain, kuukausittain, päivittäin, d) yhtiön myynti kuukausittain 1960-92, e) erilaiset indeksisarjat.

Tavanomaisissa empiirisissä aineistoissa havainnot ovat toisistaan riippumattomia ja havaintojen järjestyksellä ei ole merkitystä analysoinnin kannalta. Aikasarjojen yhteydessä lähtökohta on toisenlainen. Aikasarja on jono havaintoarvoja mitattuna tarkasteltavasta ilmiöstä nimenomaan peräkkäisinä ajankohtina. Mitattaessa tiettyä ilmiötä peräkkäisinä ajankohtina, ovat havaintoarvot toisistaan jollain tavalla riippuvia. Esimerkiksi Tampereen kaupungin asukaslukusarjassa vuoden 2004 asukasluku riippuu suuresti vuoden 2003 vastaavasta luvusta.

Aikasarjojen analysointiin on kehitetty omat menetelmät, jotka poikkeavat tavanomaisista tilastollisista menetelmistä. Aikasarjojen käsittelyssä voi olla monenlaisia tavoitteita. Voidaan haluta ainoastaan kuvailla tarkasteltavaa ilmiötä. Tällöin piirretään sarjan kuvaaja, lasketaan joitain tunnuslukuja, arvioidaan sarjan kehityssuuntaa, jne. Usein halutaan ennustaa sarjan tulevia arvoja. Ennustaminen voi perustua vaikkapa havaitun sarjan avulla laskettuihin erilaisiin painotettuihin summiin.

Tarkastellaan jatkossa tasavälein mitattuja aikasarjoja. Mittaustuloksen oletetaan olevan kvantitatiivinen.

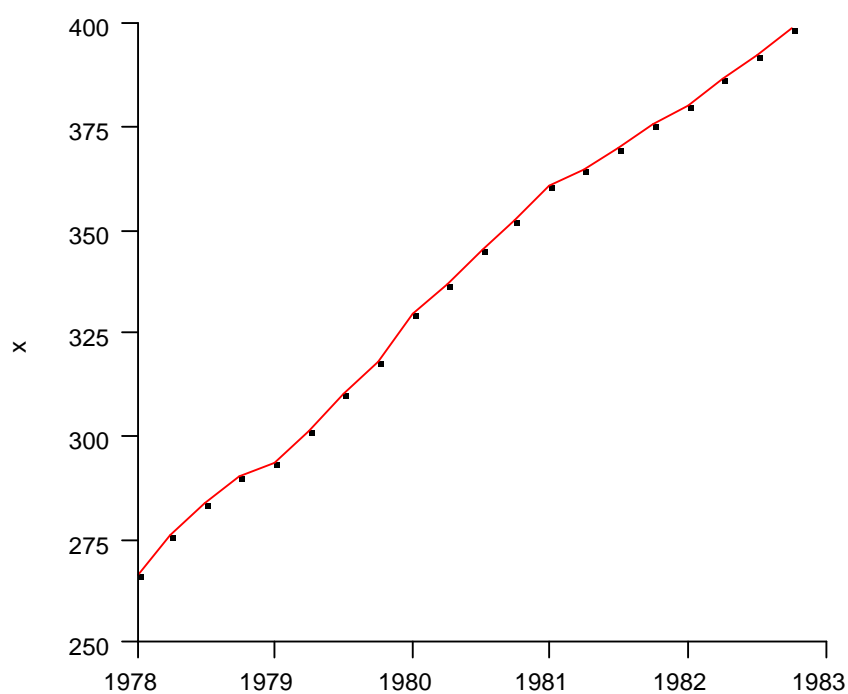
6.1 Aikasarjan graafinen esitys

Aikasarjan kuvaaja piirretään siten, että vaaka-akselilla on mittausajankohdat ja pystyakselilla aikasarjan arvot. Havaintopisteet (t_i, x_{t_i}) , $i = 1, \dots, n$, yhdistetään janoihin toisiinsa, jolloin kuvaajasta saadaan yhtäjaksoinen.

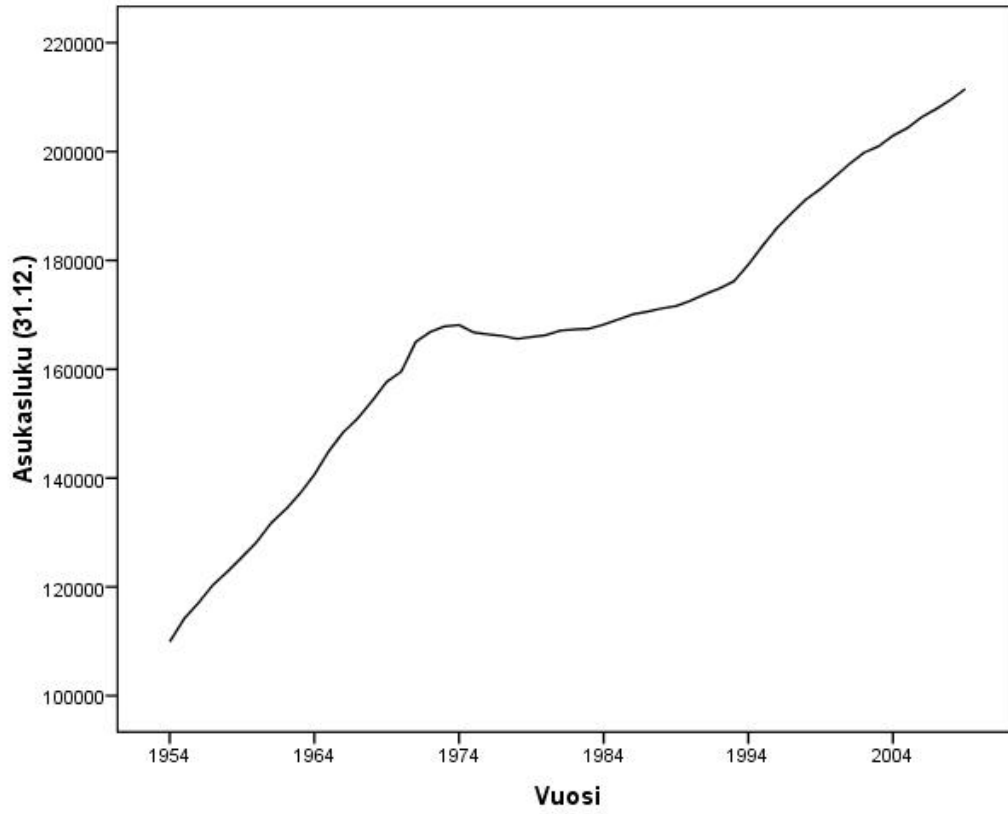
Esim. 6.1.1. Oheinen aikasarja graafisesti esitettynä.

Year	Quarter	x_t	Year	Quarter	x_t
1978	I	266,2		III	345,2
	II	276,0		IV	352,8
	III	284,2	1981	I	361,1
	IV	290,6		II	365,0
1979	I	293,4		III	370,1
	II	301,6		IV	375,7
	III	310,4	1982	I	380,4
	IV	318,3		II	386,6
1980	I	329,6		III	392,7
	II	337,2		IV	398,9

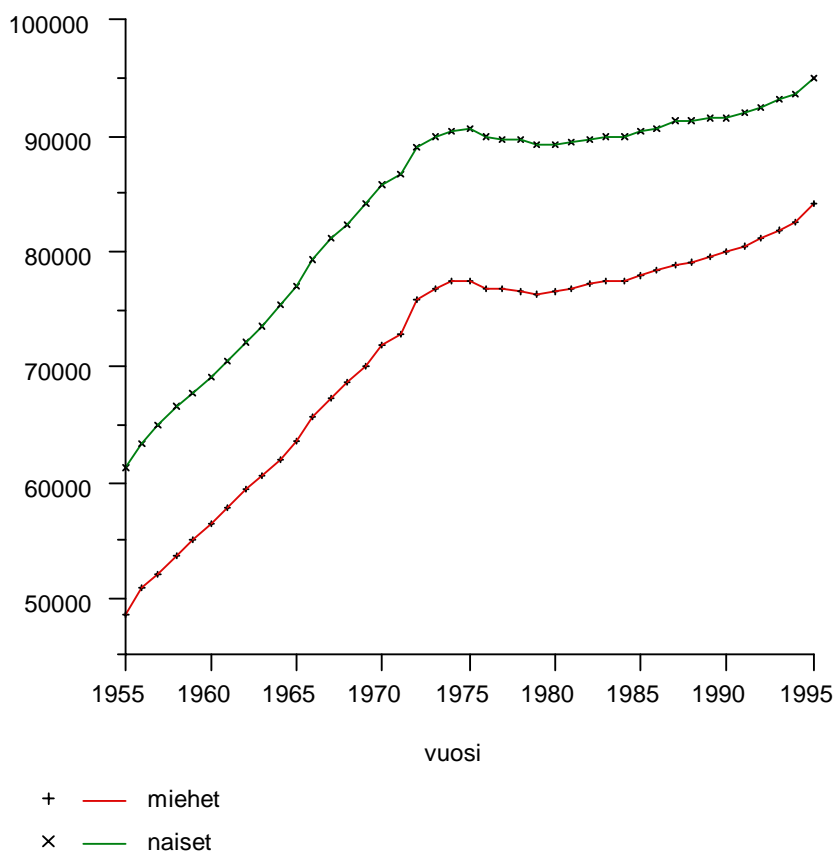
(State and Local Government Purchases, quartely, 1978 thru 1982, x_t = annualized rate (\$ billion))



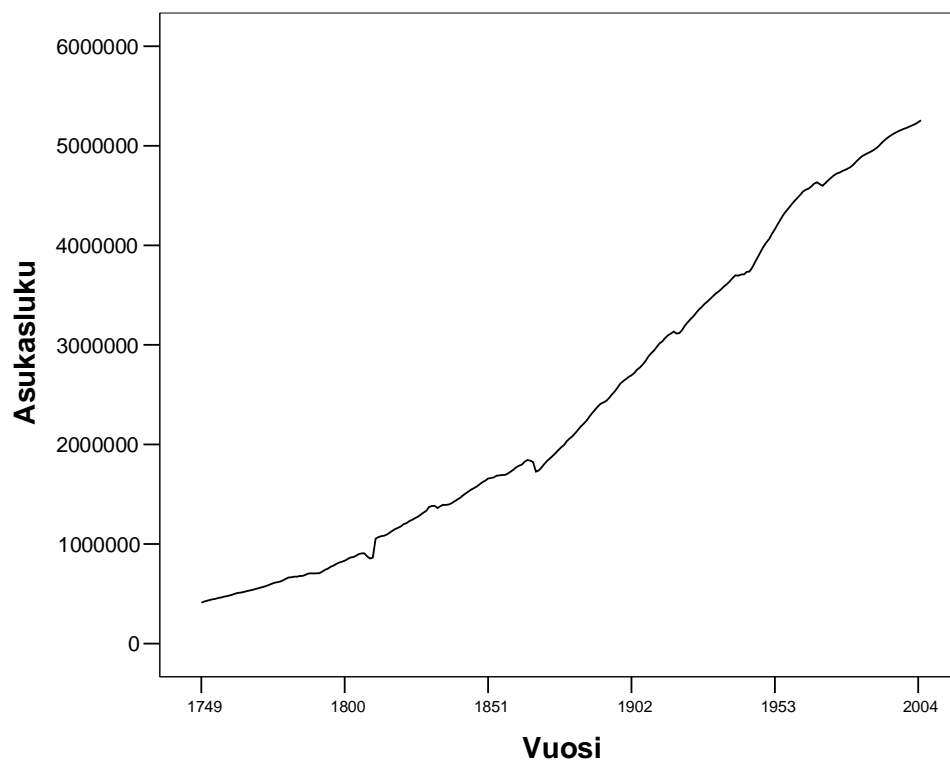
Esim. 6.1.2. Tampereen kaupungin asukasluku vuosina 1954-2009.



Esim. 6.1.3. Naisten ja miesten lukumäärät Tampereella vuosina 1955-1995.



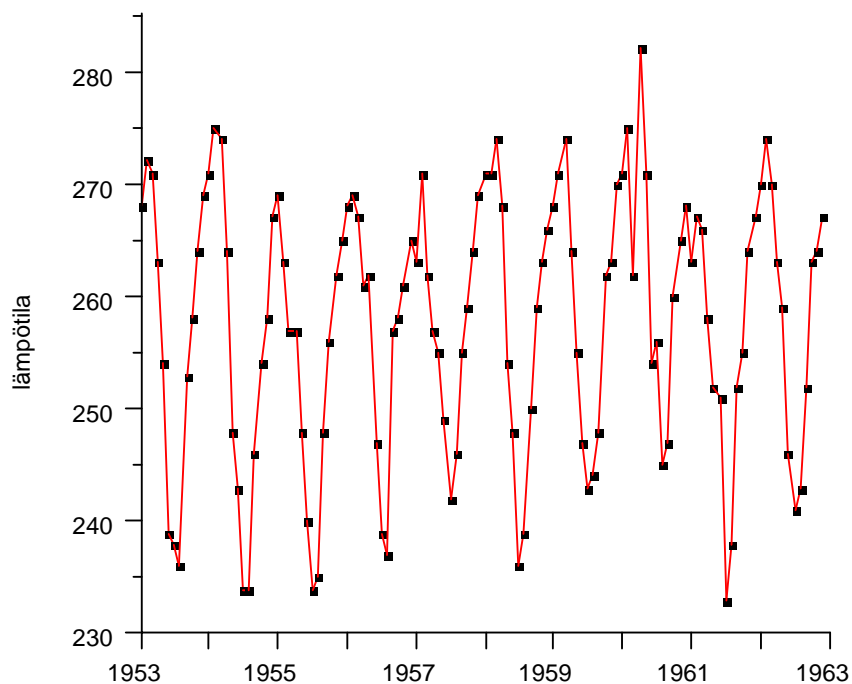
Tampereen kaupungin asukasluvussa on ollut voimakasta nousua aina 70-luvulle asti. Tällainen yleinen kehityssuunta (suoranomainen tai muunlainen) on nimeltään trendi. Hyvin usein aikasarjoissa on voimakas nouseva trendi.

Esim. 6.1.4. Suomen asukasluvun kehitys 1749 - 2005.

Asukaslukuun on vaikuttanut mm. sotavuodet 1789 - 1791, 1808 - 1809, 1918, 1940 - 1944, koleravuosi 1836, nälkävuodet 1866 - 1868, siirtolaisuus Ruotsiin 1969 -1970.

Vuodenajat aiheuttavat toisinaan aikasarjaan jaksottaista vaihtelua. Tätä vaihtelua kutsutaan kausivaihteluksi.

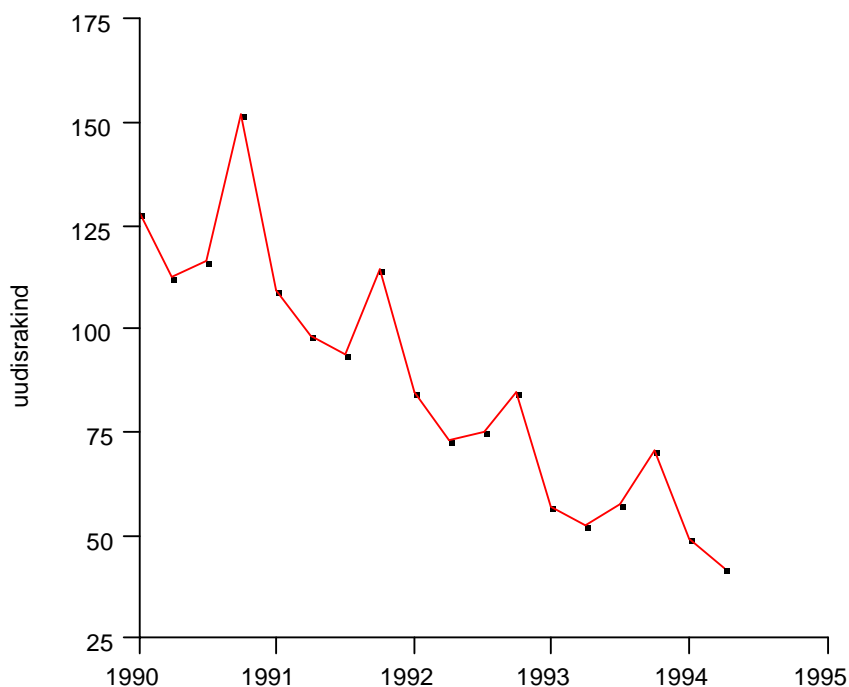
Esim. 6.1.5. Kuukausikeskilämpötilat 1953-1962, Recife, Brasilia.



Kuukausikeskilämpötiloissa on vaihtelua vuodenajan mukaan. Korkeimmillaan lämpötilat ovat talvella ja alhaisimmillaan kesällä. Eri vuosien samojen kuukausien lämpötilat muistuttavat toisiaan. On kyse kausivaihtelua sisältävästä aikasarjasta, jossa kausivaihtelujakson pituus on 12.

Esimerkissä 6.1.6 on aikasarja, joka sisältää sekä trendin että kausivaihtelua.

Esim. 6.1.6. Uudisrakentamisen volyyymi-indeksi neljännesvuosittain (1985=100).



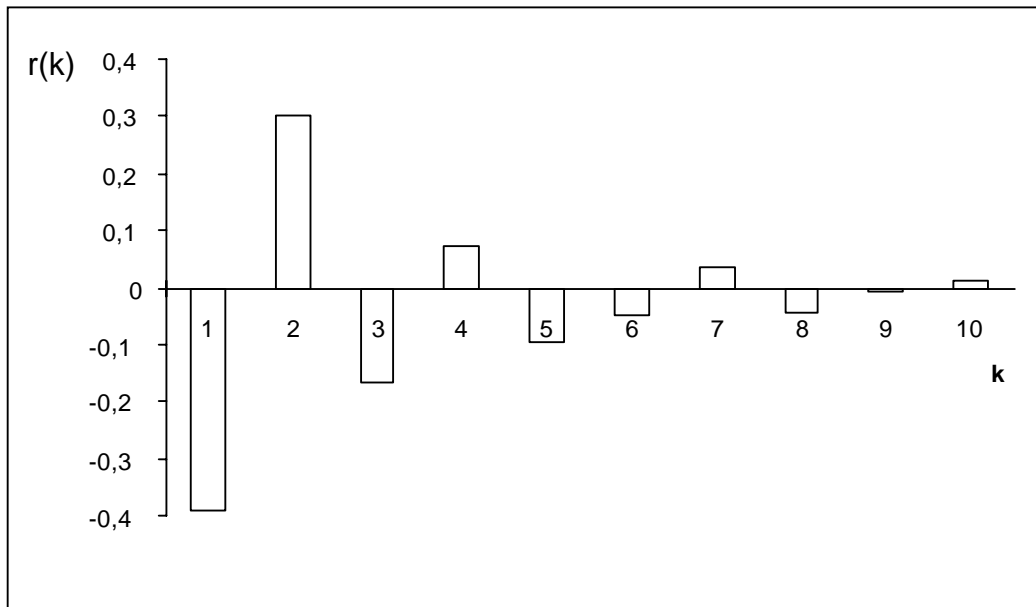
Kausivaihtelun ja trendin lisäksi sarja voi sisältää syklisiä vaihtelua, muusta kuin vuodenajoista johtuvaa, sekä muuta epäsäännöllistä, satunnaista vaihtelua. Aikasarjan voidaan ajatella muodostuva näistä komponenteista.

6.2 Otosautokorrelaatio

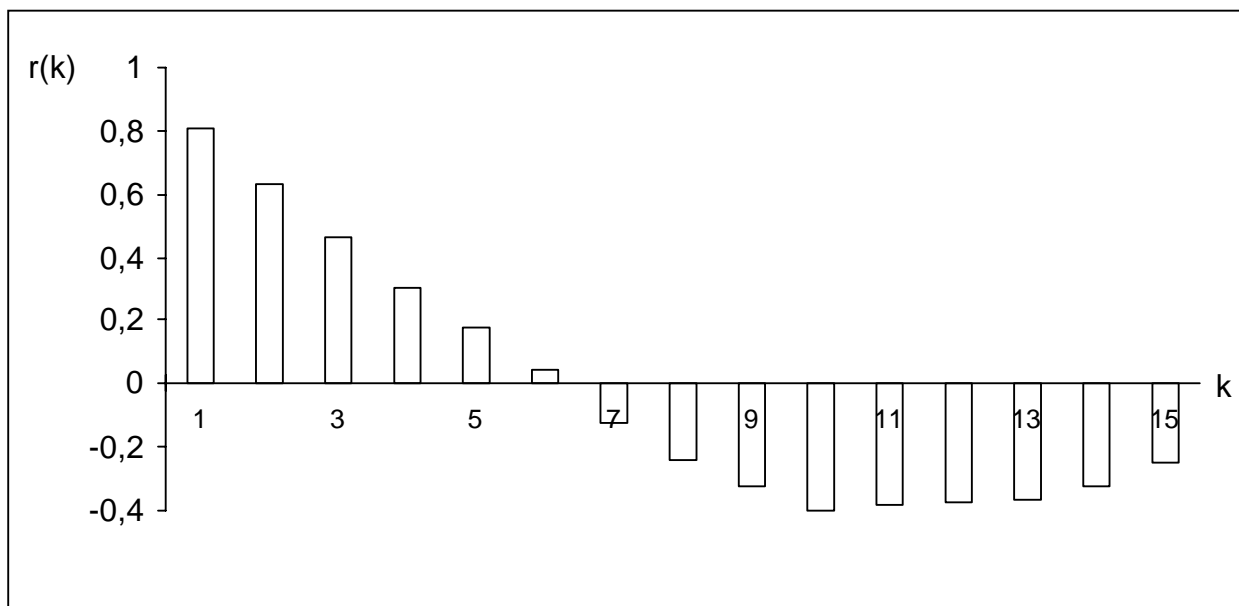
Kuten jo on todettu aikasarjan peräkkäiset havainnot ovat usein toistensa kaltaisia. Tätä samankaltaisuuden astetta kuvataan korrelaatiokertoimen tapaisella tunnusluvulla r_k otosautokorrelaatiokertoimella. Otosautokorrelaatiokerroin viiveellä k , r_k , kuvaa havaintojen $(x_1, x_{1+k}), (x_2, x_{2+k}), \dots$ välistä korrelaatiota. Autokorrelaatio viiveellä 1 kuvaa peräkkäisten havaintojen korreloituneisuutta, autokorrelaatio viiveellä 2 kertoo miten joka toinen arvo korreloi keskenään jne. Otosautokorrelaatiolle on oma laskukaava, mutta sitä voidaan arvioida tavanomaisen korrelaatiokertoimen avulla.

Otosautokorrelaation graafinen erityys näkyy kuviossa 6.2.1. Usein pylväiden sijaan käytetään janoja.

Kuvio 6.2.1. Otosautokorrelaatiofunktio graafisesti.



Jos sarjassa on trendi, niin peräkkäiset arvot ovat voimakkaasti korreloituneita, jolloin otosautokorrelaatiofunktio voisi näyttää vaikka kuvion 6.2.2. mukaiselta.

Kuvio 6.2.2. Otosautokorrelaatiofunktio trendisarjasta esimerkissä 6.2.3.

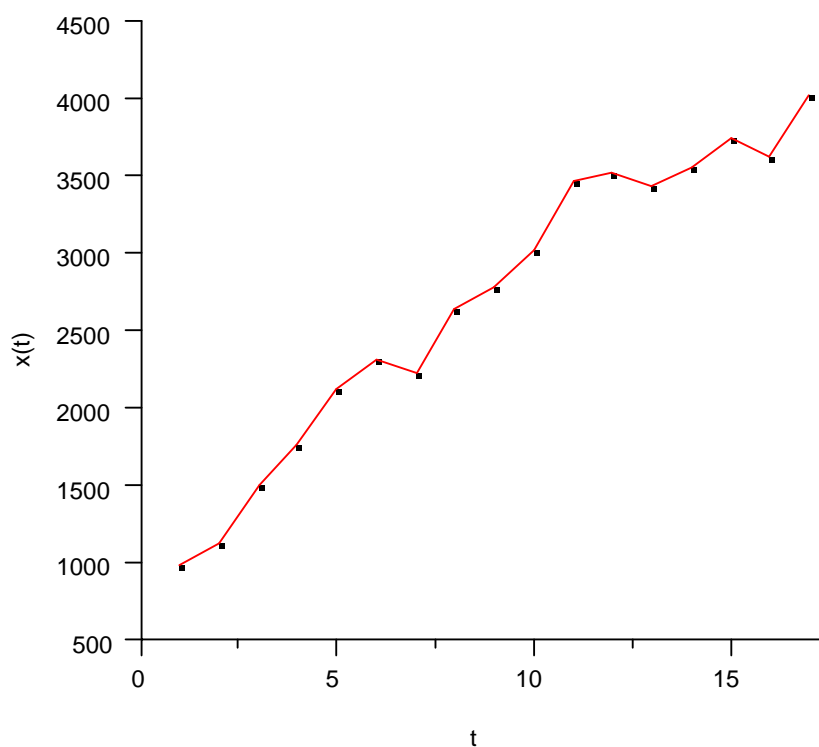
Jos sarjassa on kausivaihtelua ilman trendiä, niin autokorrelaatiota on kausivaihtelujakson viiveellä, sen kerrannaisilla sekä mahdollisesti näiden viiveiden ympärillä.

Sarjassa, jossa on sekä trendi että kausivaihtelu, otosautokorrelaatiofunktio muistuttaa trendisarjan vastaavaa sisältäen lisäksi suuria autokorrelaation arvoja kausivaihtelujakson (sekä mahdollisesti sen kerrannaisilla) viiveillä (mahdollisesti myös näiden viiveiden ympäristössä).

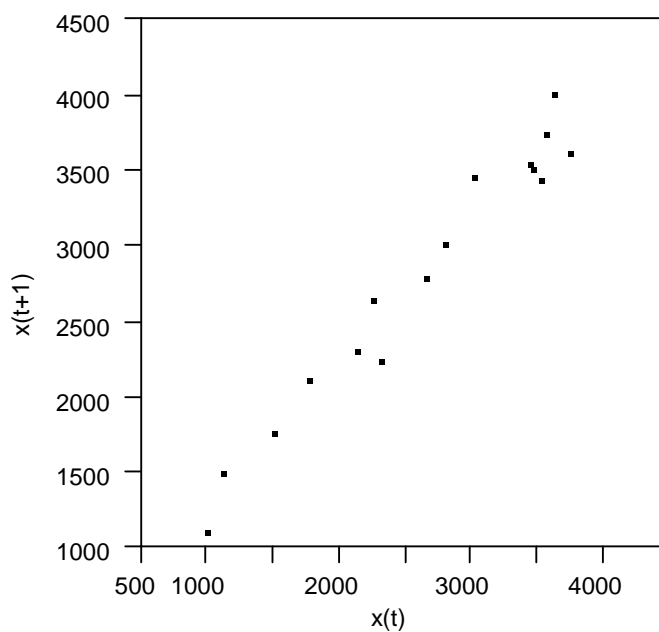
Esim. 6.2.3. Autokorrelaation arviointi pisteparven perusteella.

t	X_t	X_{t+1}	X_{t+2}
1	999	1123	1503
2	1123	1503	1762
3	1503	1762	2126
4	1762	2126	2315
5	2126	2315	2239
6	2315	2239	2655
7	2239	2655	2787
8	2655	2787	3024
9	2787	3024	3467
10	3024	3467	3528
11	3467	3528	3441
12	3528	3441	3558
13	3441	3558	3746
14	3558	3746	3628
15	3746	3628	4021
16	3628	4021	-
17	4021	-	-

Aikasarjan kuvaaja

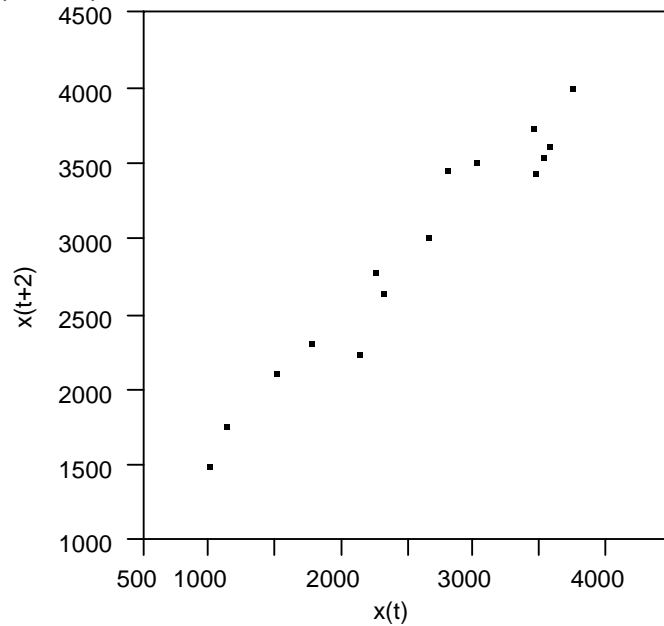


Pisteparvi pisteistä (x_t, x_{t+1})



Tästä voidaan arvioida, että autokorrelaatio viiveellä 1 on positiivinen, koska pisteparvi on nouseva ja korrelaatio selvästi positiivinen.

Pisteparvi pisteistä (x_t, x_{t+2})



Voidaan arvioida, että autokorrelaatio viiveellä 2 on myös positiivinen, koska pisteparvi on nouseva. Autokorrelaatio kuitenkin pienempi kuin viiveellä 1 tarkasteltuna.

Korrelaatiomatriisi

Variable	x(t)	x(t+1)	x(t+2)
x(t)	1,0000	0,9821	0,9738
x(t+1)	0,9821	1,0000	0,9764
x(t+2)	0,9738	0,9764	1,0000

SPSS Aikasarjan piirtäminen ja autokorrelaatioiden laskeminen

Graphs

Sequence... sarjan kuvaaja,
Time Series>

...

Autocorrelations
autokorrelaatiofunktio

7 TILASTOLLISEN PÄÄTTELYN PERUSTEITA

Kun havaintoaineisto on alustavasti tutkittu käyttäen kuvailevan analyysin keinoja, suoritetaan varsinaiset tilastolliset analyysit; tutkitaan mm. erilaisia riippuvuuksia sopivien tilastollisten menetelmien ja testien avulla ja tehdään päätelmiä populaatiosta aineiston (otoksen) perusteella. Tilastollinen analyysi pohjautuu siis otokseen, jonka perusteella tuloksia halutaan yleistää koko populaatioon. Tähän yleistyksen liittyy aina epävarmuutta.

Tilastolliset analyysit voidaan jakaa monellakin tavalla ryhmiin. Voidaan ryhmitellä esim. sen mukaan millaiset roolit tarkasteltavilla muuttujilla on, mikä on muuttujien mittaustaso, mitä oletuksia voidaan tehdä, onko kyse aikasarjoista, pitkittäistutkimuksesta, laaduntarkkailusta, onko selitettäviä muuttujana yksi vai useampia, ovatko ne kvantitatiivisia vai kvalitatiivisia, jne.

Menetelmä voi olla hyvin yksinkertainenkin. Esimerkiksi voidaan haluta selvittää milloin on mahdollista sanoa ehdollisten otoskeskiarvojen perusteella, että populaatioissa keskiarvot poikkeavat toisistaan. Voidaan myös arvioida tiettyä prosenttiosuutta populaatiossa (esimerkiksi selvitetään tietyn puolueen kannatusta ks. esim. 3.1) tai arvioida vaikkapa populaation keskiarvoa. Toisaalta voidaan myös tutkia useampia muuttujia samanaikaisesti. Tässä luvussa esitellään vain muutama yksinkertainen menetelmä ja testi lähinnä joidenkin esimerkkien avulla. Liitteessä 6 on lueteltu useampia menetelmiä ja niiden ryhmittelytapoja. Esimerkkiluettelon on vain tarkoitus antaa yleiskuva olemassa olevien menetelmien laajasta kirjosta ja niistä vain jokusiin tutustutaan tilastotieteen perusopinnoissa.

Ks. myös <http://www.fsd.uta.fi/menetelmaopetus/menetelma/menetelmatyyppit.html>

Tilastollisten päätelmien teko perustuu satunnaisotoksesta (ks. 7.5) määriteltyjen tunnuslukujen (kuten esim. otoskeskiarvojen tai prosenttiosuuksien) todennäköisyysjakaumiin. Johtopäätelmät tehdään erilaisten tilastollisten testien ja analysointimenetelmien avulla. Tällaiseen päättelyyn sisältyy tiettyä epävarmuutta, jota pyritään hallitsemaan käyttäen hyväksi todennäköisyyslaskentaa ja erilaisia todennäköisyysjakaumia.

Seuraavassa tutustutaankin hyvin lyhyesti, lähinnä vain esimerkkien avulla, todennäköisyyslaskentaan ja todennäköisyysjakaumiin sekä tilastollisen päättelyn peruskäsitteisiin. Luodaan katsaus otosjakaumiin ja niiden käyttöön tilastollisessa päättelyssä. Käydään läpi estimointiin liittyviä käsitteitä sekä hypoteesien testausta.

7.1 Satunnaisilmiö ja tapahtuma

Esim. 7.1.1. Heitettäessä rahaa ei tiedetä saadaanko kruunu vai klaava. Tiedetään, että molemmat vaihtoehdot ovat yhtä todennäköisiä. Heitettäessä noppaa tiedetään, että saadaan silmäluku 1, 2, 3, 4, 5 tai 6, mutta ei tiedetä etukäteen silmälukua. Tiedetään, että jokaisen silmäluvun todennäköisyys on sama. Kortin vetäminen sekoitetusta korttipakasta, lottoaminen, veikkaaminen, bussin saapuminen pysäkillä ja päivän sää ennustaminen ovat myös esimerkkejä ilmiöistä, joihin liittyy epävarmuutta.

Satunnaisilmiö on mikä tahansa ilmiö, johon liittyy useita eri tulosmahdollisuuksia sekä epävarmuutta ilmiön tuloksesta. Puhutaan myös satunnaiskokeesta.

Satunnaisilmiöön liittyvien kaikkien mahdollisten tulosten joukkoa kutsutaan perusjoukoksi (otosavaruudeksi) E . Käytännössä ollaan kiinnostuneita joistain perusjoukon osajoukoista (sekä niiden esiintymistodennäköisyyksistä). Perusjoukon osajoukko on nimeltään tapahtuma. Tapahtumia merkitään A, B, C, \dots

Esim. 7.1.2.

Rahanheitto

E = "kaikki mahdolliset tulokset" = {kruunu, klaava}

tapahtumia:

A = "saadaan kruunu" = {kruunu}

B = "saadaan klaava" = {klaava}

Nopanheitto

E = {1,2,3,4,5,6}

tapahtumia:

A = "saadaan parillinen" = {2,4,6}

B = {1}

C = {1,2,3}

D = "saadaan suurempi kuin 4" = {5,6}

Kortin vetäminen sekoitetusta korttipakasta

E = "kaikki kortit"

tapahtumia:

A = "saadaan pata"

B = "saadaan kuningas"

C = "saadaan punainen ässä"

7.2 Klassinen todennäköisyys

Olkoon tarkasteltavan satunnaisilmiön perusjoukossa n tulosta, jotka ovat kaikki yhtä mahdollisia. Olkoon tapahtumaan A liittyviä tuloksia k kappaletta ($0 \leq k \leq n$). Tällöin tapahtuman A todennäköisyys $P(A) = k/n$. Todennäköisyys voidaan ilmoittaa myös prosentteina.

Esim. 7.2.1.

Rahanheitto

A = "saadaan kruunu"

$P(A) = 1/2$

Nopanheitto

A = "saadaan parillinen" = {2,4,6}

$P(A) = 3/6$

$B = \{1\}, P(B) = 1/6$

$D = \text{"suurempi kuin 4"} = \{5,6\}, P(D) = 2/6.$

Klassisen todennäköisyyden (voidaan liittää vain äärellisiin perusjoukkoihin) yhteydessä lukujen n ja k määrittäminen ei aina ole yksinkertaista. Usein joudutaan käyttämään hyväksi kombinatoriikkaa, jota ei tässä yhteydessä kuitenkaan käydä läpi.

Tapahtuman A todennäköisyys voidaan myös määritellä arvoksi, jota tapahtuman suhteellinen frekvenssi lähestyy satunnaiskoetoistojen määrää kasvatettaessa.

Tapahtumat A ja B ovat (tilastollisesti, stokastisesti) riippumattomia, jos B:n tapahtuminen tai tapahtumatta jääminen ei vaikuta A:n tapahtumisen todennäköisyyteen ja A:n tapahtuminen tai tapahtumatta jääminen ei vaikuta B:n tapahtumisen todennäköisyyteen.

7.3 Satunnaismuuttuja ja todennäköisyysjakauma

Funktiota, joka liittää yksikäsitteisen reaaliarvon jokaiseen tarkasteltavan satunnaisilmiön perusjoukon tulokseen, sanotaan satunnaismuuttujaksi. Eri tuloksiin liittyviä reaaliarvoja sanotaan satunnaismuuttujan arvoksi. Jatkossa merkitään (useimmiten) satunnaismuuttujia isoin kirjaimin (X, Y, Z, \dots) ja satunnaismuuttujan arvoja pienin kirjaimin (x, y, z, \dots).

Esim. 7.3.1. Satunnaisilmiö nopanheitto. Satunnaismuuttuja X = saatu silmäluku.

Esim. 7.3.2. Heitetään kolikkoa neljä kertaa. Määritellään satunnaismuuttuja X = klaavojen lukumäärä heittosarjassa. Etukäteen ei tiedetä montako klaavaa saadaan, mutta voidaan laskea eri arvojen todennäköisyydet. Tässä satunnaismuuttujan X mahdolliset arvot ovat 0, 1, 2, 3 ja 4. Erilaisia heittosarjoja on kaikkiaan 16.

	klaavojen lukumäärä		klaavojen lukumäärä
Kl,Kl,Kl,Kl	4	Kr,Kl,Kl,Kr	2
Kr,Kl,Kl,Kl	3	Kl,Kr,Kl,Kr	2
Kl,Kr,Kl,Kl	3	Kr,Kl,Kr,Kl	2
Kl,Kl,Kr,Kl	3	Kl,Kr,Kr,Kr	1
Kl,Kl,Kl,Kr	3	Kr,Kl,Kr,Kr	1
Kl,Kl,Kr,Kr	2	Kr,Kr,Kl,Kr	1
Kr,Kr,Kl,Kl	2	Kr,Kr,Kr,Kl	1
Kl,Kr,Kr,Kl	2	Kr,Kr,Kr,Kr	0

$$P(X=0) = 1/16, P(X=3) = 4/16, P(X=1) = 4/16, P(X=4) = 1/16, P(X=2) = 6/16$$

Esim. 7.3.3. Satunnaisilmiönä vakioveikkaaminen (13 kohdetta, joissa jokaisessa 3 vaihtoehtoa). Tällöin voidaan määritellä satunnaismuuttuja X = oikein veikkattujen kohteiden lukumäärä. X voi saada arvoja 0,1,2,...,13. Näiden arvojen todennäköisyydet voidaan laskea nk. binomijakauman avulla.

Esim. 7.3.4. Tehdään kahden alkion otos luvuista 1, 2, 3, 4, 5, 6 käyttäen systemaattista otantaa. Tällöin mahdolliset otokset ovat 1, 4; 2, 5; 3, 6 ja näistä lasketut keskiarvot 2,5; 3,5 ja 4,5. Siis $P(\bar{X} = 2,5) = P(\bar{X} = 3,5) = P(\bar{X} = 4,5) = 1/3$. Tässä on määrätty ko. tilanteeseen liittyen otoskeskiarvon (erään otossuureen) todennäköisyysjakauma.

Esimerkissä 7.3.2 ja 7.3.4 ilmoitettiin tarkasteltavan satunnaismuuttujan mahdolliset arvot ja eri arvojen todennäköisyydet. Tällöin muodostettiin satunnaismuuttujan todennäköisyysjakauma.

Satunnaismuuttuja voi olla joko jatkuva tai diskreetti. Edellisissä esimerkeissä satunnaismuuttujat olivat diskreettejä. Satunnaismuuttujaa sanotaan diskreetiksi, jos se voi saada arvokseen äärellisen määrän erisuuria arvoja tai äärettömän määrän siten, että arvot ovat numeroitavissa positiivisia kokonaislukuja käyttäen. Muulloin satunnaismuuttuja on jatkuva. Jatkossa tilastolliseen päättelyyn liittyen tarkastellaan vain kahta jatkuvaa jakaumaa. Jakaumien havainnollistamiseksi on edellä kuitenkin tarkasteltu muutamia diskreettejä jakaumia.

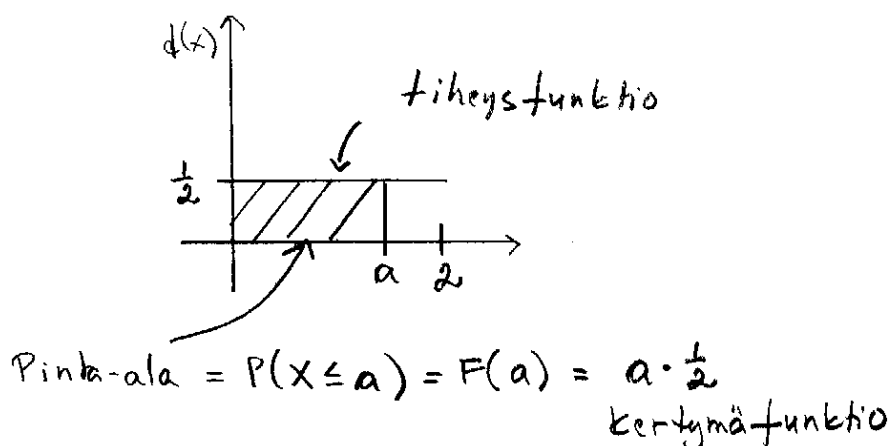
Diskreetin satunnaismuuttujan todennäköisyysjakauma voidaan usein (ainakin periaatteessa) muodostaa kuten esimerkissä 7.3.2. Jatkuvien muuttujien yhteydessä todennäköisyysjakauma määritellään jatkuvan funktion avulla. Funktiota, joka määrittää satunnaismuuttujan todennäköisyysjakauman kutsutaan tiheysfunktioiksi, merkitään $f(x)$. Jotta f olisi jatkuvan satunnaismuuttujan tiheysfunktio on oltava $f(x) \geq 0$ ja $f(x)$:n ja x -akselin väliin jäävä pinta-ala = 1. Tiheysfunktio kuvaa siis ykkösen suuruisen todennäköisyysmassan jakaumaa. Diskreetin muuttujan yhteydessä puhutaan myös pistetodennäköisyyksistä. Näiden todennäköisyyksien summa on 1; siis diskreetin satunnaismuuttujan kaikkien arvojen todennäköisyyksien summa on yksi. Tiheysfunktion voidaan ajatella kuvaavan populaation jakaumaa (vrt. frekvenssimonikulmio empiiristen jakaumien yhteydessä).

Satunnaismuuttujan X kertymäfunktio F määritellään

$$F(x) = P(X \leq x).$$

Kertymäfunktion arvo pisteessä x kertoo siis todennäköisyyden sille, että satunnaismuuttujan X arvo on $\leq x$.

Esim. 7.3.5. Esimerkki erään jatkuvan muuttujan tiheys- ja kertymäfunktioista



Empiiristen jakaumien yhteydessä jakaumaa voitiin kuvailla tunnuslukujen avulla. Myös todennäköisyysjakaumiin liitetään samantyyppisiä tunnuslukuja, jotka määrittävät todennäköisyysjakauman avulla.

Empiirisen jakauman keskiarvoa vastaavaksi tunnusluvuksi todennäköisyysjakauman (populaation) yhteydessä määritellään jakauman odotusarvo (populaation keskiarvo) sekä otosvarianssia ja keskihajontaa vastaaviksi (populaation) varianssi ja keskihajonta. Odotusarvo kuvaa jakauman keskikohtaa ja varianssi mittaa miten tiiviisti todennäköisyysmassa on keskittynyt odotusarvon ympärille.

Satunnaismuuttujan X odotusarvoa merkitään $E(X) = \mu$ sekä varianssia $\text{Var}(X) = \sigma^2$, jonka neliöjuuri σ on nimeltään keskihajonta.

Satunnaismuuttujien riippumattomuus määritellään vastaavalla tavalla kuin tapahtumien riippumattomuuskin.

7.4 Normaalijakauma

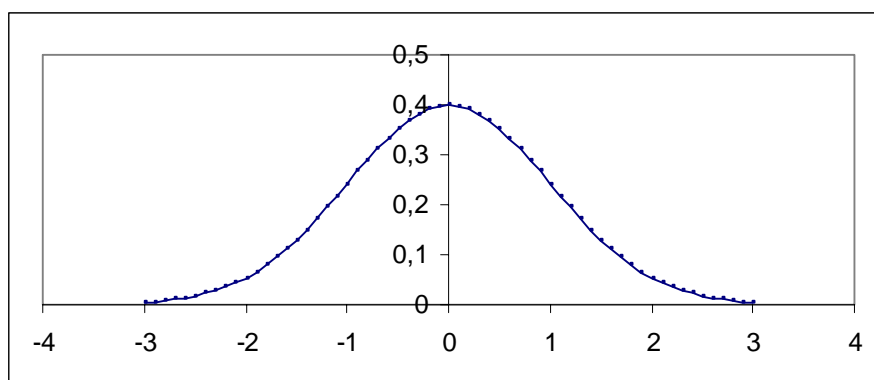
Seuraava todennäköisyysjakauma on tilastotieteessä hyvin keskeinen. Tarkastellaan jatkuvaa satunnaismuuttujaa X , joka voi saada arvokseen kaikki reaalityöt.

Satunnaismuuttuja X noudattaa normaalijakaumaa parametrein μ ja σ^2 ($\sigma > 0$), jos sen tiheysfunktio on μ :n suhteen symmetrinen yksihuippuinen, tietyllä tavalla määritelty jatkuva funktio, (ks.

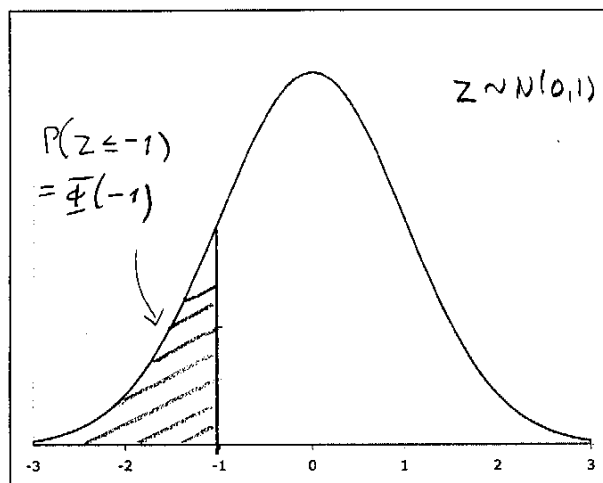
<http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/normaalijakauma>), merkitään $X \sim N(\mu, \sigma^2)$. Tällöin $E(X) = \mu$ ja $\text{Var}(X) = \sigma^2$. Varianssi kertoo jakauman levittäytymisestä odotusarvon μ ympärille ja parametrit μ ja σ^2 määrittävät jakauman yksikäsitteisesti.

Usein tarkastellaan normaalijakaumaa, jonka odotusarvo on nolla ja varianssi yksi. Tätä kutsutaan standardoiduksi normaalijakaumaksi, merkitään $Z \sim N(0, 1)$, ja $F(z) = P(Z \leq z) = \Phi(z)$.

Esim. 7.4.1. Standardoidun normaalijakauman tiheysfunktio.

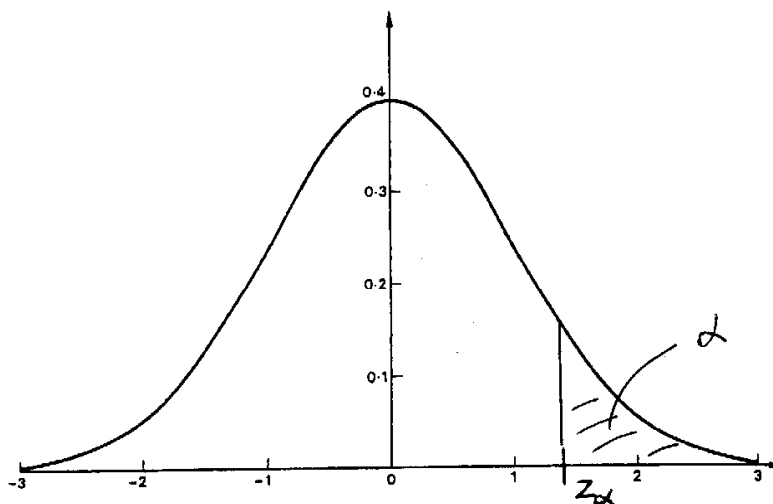


Standardoidun normaalijakauman kertymäfunktion $\Phi(z) = P(Z \leq z)$ arvoja on taulukoitu (ks. liite 2 tai taulukko [http://mtl.uta.fi/tilasto/tiltp7/N\(0_1\).pdf](http://mtl.uta.fi/tilasto/tiltp7/N(0_1).pdf)), kertymäfunktio graafisesti:

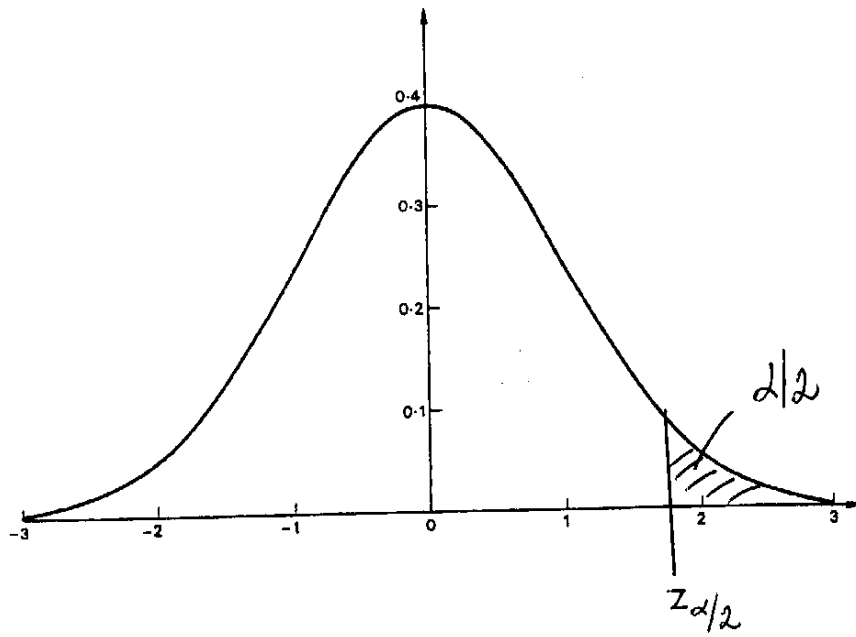


Taulukoiden avulla voidaan määrittää kertymäfunktion arvoja ja laskea erilaisia todennäköisyyksiä, joihin ei kuitenkaan tällä opintojaksolla tarkemmin perehdytä. Myöhemmin, luottamusvälien ja testauksen yhteydessä, tarvitaan kuitenkin nk. harvinaisten arvojen rajaa, johon liittyen määritellään z_α .

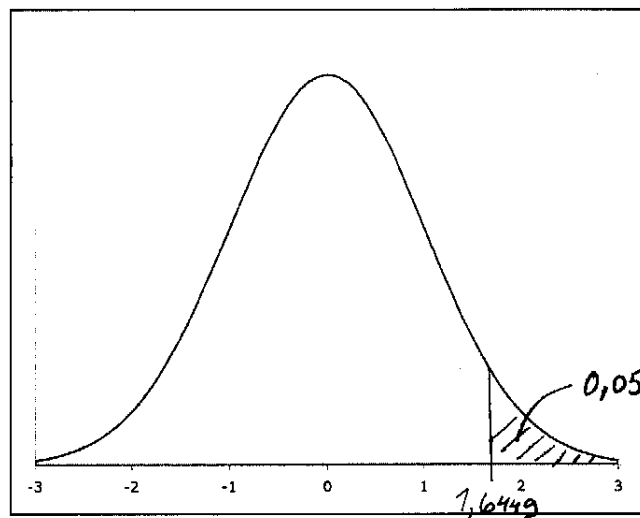
Olkoon $Z \sim N(0,1)$. Määritellään z_α siten, että $P(Z \geq z_\alpha) = \alpha$, graafisesti:



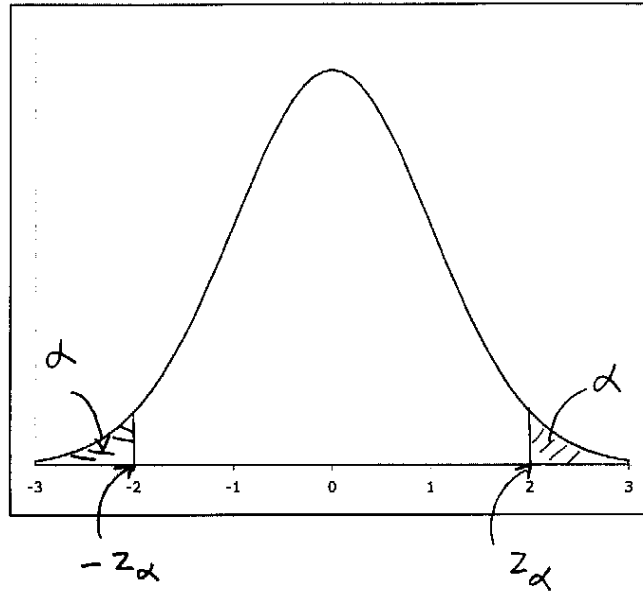
Samoin $z_{\alpha/2}$ siten, että $P(Z \geq z_{\alpha/2}) = \alpha/2$, graafisesti



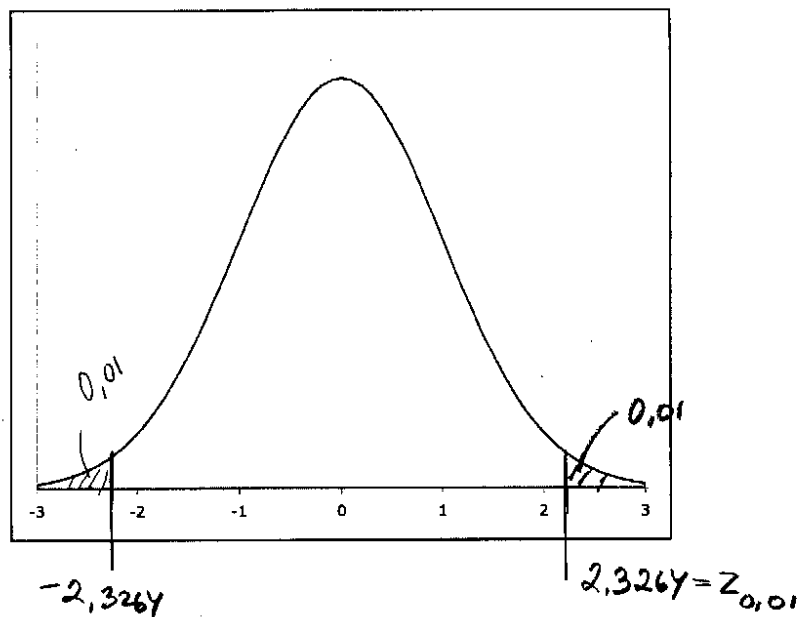
Esimerkiksi $z_{0,05} = 1,6449$ ja $z_{0,05/2} = 1,96$, $z_{0,01} = 2,3264$ (voidaan katsoa taulukosta, liite 2), graafisesti:



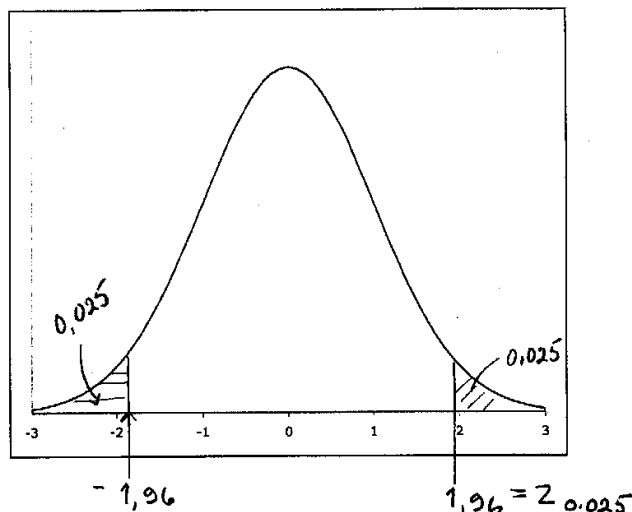
Standardoitu normaalijakauma on symmetrinen nollan suhteen, joten $P(Z \leq -z_{\alpha}) = \alpha$, graafisesti:



Esim. 7.4.2. Olkoon $Z \sim N(0, 1)$. Tällöin $P(Z \geq 2,3264) = 0,01$, $P(Z \leq -2,3264) = 0,01$, graafisesti:



Vastaavalla tavalla $P(Z \geq 1,96) = 0,025$, $P(Z \leq -1,96) = 0,025$,



ja $P(Z \geq 1,6449) = 0,05$, $P(Z \leq -1,6449) = 0,05$

7.5 Satunnaisotos, otossuure ja otantajakauma

Tilastolliset johtopäätelmät, jotka koskevat populaation eli perusjoukon ominaisuuksia tehdään otoksen avulla. Jotta erilaisten otoksesta laskettujen tunnuslukujen luotettavuutta voidaan arvioida, otos valitaan poimimalla se todennäköisyysotannalla. Todennäköisyysotannassa kaikki mahdolliset n alkion otokset voidaan luetella, tunnetaan jokaisen mahdollisen otoksen poimintatodennäköisyys ja otokset poimitaan näiden todennäköisyyksien mukaan. On myös tiedettävä, miten otoksen perusteella yleistetään tulokset koko populaatioon.

Jatkossa tarkastellaan vain yksinkertaisella satunnaisotannalla tehtyyn otokseen liittyviä tuloksia. Lisäksi ollaan kiinnostuneita vain yhdestä populaation alkioihin liittyvästä ominaisuudesta, muuttujasta.

Olkoon X_1, X_2, \dots, X_n n :n satunnaismuuttujan jono. Tätä jonoa sanotaan satunnaisotokseksi, jos X_i :t ovat riippumattomia ja noudattavat samaa jakaumaa.

Sanonta " X_1, X_2, \dots, X_n on satunnaisotos $N(\mu, \sigma^2)$:sta" tarkoittaa sitä, että jokainen $X_i \sim N(\mu, \sigma^2)$ ja X_i :t ovat riippumattomia.

Kun äärettömästä populaatiosta tehdään otanta yksinkertaisella satunnaisotannalla (palauttaen tai palauttamatta) ja tarkastellaan yhtä tiettyä muuttujaa (tilastoyksikön ominaisuutta), on kyse satunnaisotoksesta. Jos populaatio on äärellinen YSO palauttaen johtaa satunnaisotokseen, mutta palauttamatta ei, koska riippumattomuusoletus ei ole voimassa. Kuitenkin, jos populaatio on suuri YSO palauttamattakin johtaa lähes riippumattomiin satunnaismuuttujiin.

Satunnaisotos määritellään siis satunnaismuuttujien perusteella. Nämä satunnaismuuttujat saavat arvot, kun otos on tehty. Siis otoksen tekemisen jälkeen satunnaisotokselle saadaan arvo, joka vaihtelee otoksesta toiseen.

Satunnaisotoksen avulla määriteltyä funktiota, joka myös on satunnaismuuttuja, kutsutaan otossuureeksi. Koska otossuure on satunnaismuuttuja, liittyy siihen todennäköisyysjakauma (ks. esim. 7.3.4). Otossuureen todennäköisyysjakaumasta käytetään nimitystä otanta- tai otosjakauma.

Käyttökelpoisia otossuureita esim. otoskeskiarvo, otosvarianssi, prosenttiosuus otoksessa.

Otossuureen todennäköisyysjakauman avulla saadaan selville miten otossuure voi vaihdella otoksesta toiseen. Tämä taas auttaa, kun olemme kiinnostuneita populaatioon liittyvistä arvioista perustaen arviot otokseen.

Tarkastellaan tilastollisessa päättelyssä kahden tärkeän otossuureen, otoskeskiarvon ja prosenttiosuuden, otosjakaumia.

Otoskeskiarvon jakauma tunnetaan silloin, kun otos on normaalijakaumasta. Jos X_1, X_2, \dots, X_n on satunnaisotos $N(\mu, \sigma^2)$:sta, niin

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

Vaikka X_i :t eivät olisikaan normaalisti jakautuneita, niin otokseen ollessa riittävä \bar{X} on likimain normaalisti jakautunut nk. keskeisen raja-arvolauseen perusteella. Ks. otokseen ja jakauman vaikutus otoskeskiarvon jakaumaan <http://onlinestatbook.com/simulations/CLT/clt.html> (kokeile vaikka 10000 alkion otosta normaalijakaumasta).

Olkoon populaatiossa π % tietyn tyyppisiä alkioita, joita kutsutaan jatkossa ”viallisiksi”, ja olkoon p = ”viallisten” prosenttiosuus otoksessa. Voidaan osoittaa, että

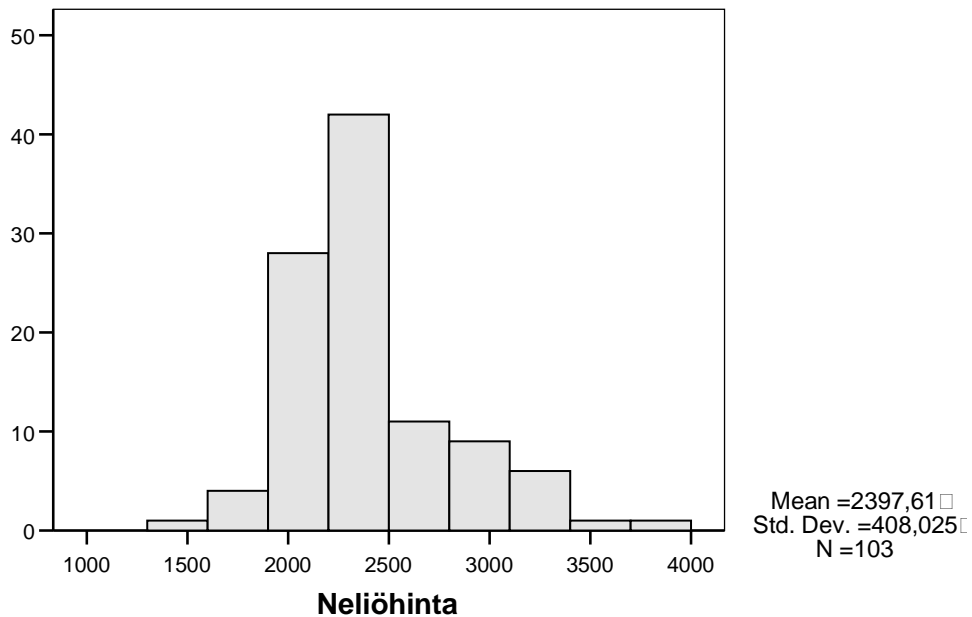
$$p \sim N(\pi, \pi(100-\pi)/n), \text{ likimain.}$$

Tiedetään siis p :n otosjakauma.

7.6 Piste-estimointi ja luottamusväläjä

Estimointi on populaation tuntemattoman parametrin arviointia sopivan otossuureen avulla. Näin tehtäessä puhutaan piste-estimoinnista.

Esim. 7.6.1. Populaation odotusarvoa voidaan estimoida otoskeskiarvolla, populaation varianssia otosvarienssilla. Tarkastellaan kerrostalohuoneistojen neliöhintoja. Eräältä alueelta tehdyssä otoksessa myynnissä olleiden huoneistojen neliöhinnat olivat keskimäärin 2398 euroa keskihajonnan ollessa 408.



Esim. 7.6.2. Puolue haluaa arvioida kannatusprosenttinsa ja kysyy satunnaisesti valitulta sadalta kansalaiselta mielipidettä. Sadan vastaajan joukossa on kannattajia 25%. Todellista kannatusprosenttia π ei siis tiedetä, mutta sitä voidaan arvioida otoksen perusteella. Arvio puolueen kannatusprosentiksi olisi 25.

Olkoon populaatiossa π % "viallisia". Pyritään arvioimaan π :tä otoksen perusteella. Luonnollinen arvio π :lle olisi siis vastaava luku otoksessa eli p = "viallisten" prosenttiosuus otoksessa.

Kun p on otossuure, jolla estimoidaan π :tä, sanotaan, että p on π :n estimaattori. Kun otos on tehty, voidaan p :lle laskea arvo eli estimaatti.

Otossuureen p keskihajontaa $\sqrt{\pi(100 - \pi)/n}$ sanotaan otoksen prosenttiosuuden keskivirheeksi.

Koska p :n keskivirheen lauseke sisältää tuntemattoman parametrin π , on keskivirhekin estimoitava. Se voidaan tehdä käyttämällä p :tä π :n estimaattorina. Näin siis p :n estimoitu keskivirhe on $\sqrt{p(100 - p)/n}$. Keskivirhe kuvaa otossuureen vaihtelua; tässä sitä miten tiiviisti p :n todennäköisyysjakauma on keskittynyt oikean π :n ympärille. Keskivirheen avulla voidaan arvioida sitä epävarmuutta, joka liittyy estimointiin. Keskivirhettä käytetään hyväksi luottamusvälien ja testauksien yhteydessä.

Esim. 7.6.3. Esimerkissä 7.6.2 p :n estimoitu keskivirhe on $\sqrt{25/(100 - 25)/100} = 0,43$.

Riippumatta otossuureesta sen hajontaa kutsutaan keskivirheeksi. Voidaan siis tarkastella esimerkiksi otoskeskiarvon tai otoskeskiarvojen erotuksen keskivirhettä.

Piste-estimointi tuottaa siis (otoksen teon jälkeen) yhden luvun, jolla arvioidaan estimoitavaa parametria. Estimointiin liittyy tietysti aina epävarmuutta. Usein halutaankin määrätä yksittäisen arvon sijaan väli, jolla arvellaan tuntemattoman

parametrin olevan. Tällöin puhutaan väliestimoinnista. Väliestimoinnissa muodostetaan nk. luottamusväli vastaavan piste-estimaattorin ja piste-estimaattorin otantajakauman keskihajonnan eli estimaattorin keskivirheen avulla.

Luottamusväliä määriteltäessä on kyseessä satunnaisvälistä, joka sisältää populaation tuntemattoman, estimoitavan parametrin todennäköisyydellä $1 - \alpha$. Kun otos on tehty, voidaan luottamusvälin ylä- ja alaraja laskea. Näin saadaan väli, joka joko sisältää parametrin tai ei sisällä. Käytännössä ei tiedetä sisältyykö parametri saadulle välille. Koska päättely halutaan kuitenkin tehdä melko suurella varmuudella, valitaan α esim. 0,05 tai 0,01; jolloin on kyse 95 %:n tai 99 %:n luottamusväleistä (0,95, 0,99 luottamustaso).

Luottamusvälejä voidaan muodostaa eri parametreille ja laskukaavat määrittyvät tilanteeseen liittyvän piste-estimaattorin ja sen keskivirheen kautta.

Ks. <http://www.stat.fi/tup/verkkokoulu/data/tt/01/11/index.html>

7.6.1 Prosenttiosuuden luottamusväli

Olkoon p = "viallisten" prosenttiosuus otoksessa. Tällöin 95 %:n luottamusväli π :lle ("viallisten" prosenttiosuudelle populaatiossa) on

$$p \pm z_{0,05/2} \sqrt{p(100 - p)/n}$$

missä $z_{0,05/2} = z_{0,025} = 1,96$. Luottamusvälin lauseke yleisesti kaavakokoelmassa, kaava (8).

Esim. 7.6.4. Luottamusväli esimerkistä 7.6.2

Väli, jolla arvellaan puolueen kannatuksen olevan, on

$$25 \pm 1,96 \sqrt{25(100 - 25)/100}$$

Esim. 7.6.5. Yritys tekee tiettyä komponenttia, jota käytetään auton moottorissa. Tämä komponentti hajoaa joskus heti, kun se on otettu käyttöön. Yritys valvoo tuotantoaan siten, että virheellisten komponenttien osuus ei saisi olla suurempi kuin 4 %. Laaduntarkkailussa tehtiin 500 komponentin otos, jossa 28 komponenttia osoittautui virheellisiksi. Onko tuotanto keskeytettävä?

Lasketaan 95 %:n luottamusväli virheellisten komponenttien prosenttiosuudelle. Saadaan

$$5,6 \pm 1,96 \sqrt{5,6(100 - 5,6)/500}$$

Virheellisten osuuden arvellaan olevan välillä 3,6 % - 7,6 %, joten vaihtelu on sallituissa rajoissa, koska 4 % kuuluu luottamusvälille.

Ks. <http://www.stat.fi/tup/verkkokoulu/data/tt/01/11/index.html> ja <http://www.stat.fi/tup/verkkokoulu/data/tt/01/11/esimerkit.html>

7.6.2 Populaation odotusarvon luottamusväli

Tarkastellaan seuraavaksi populaation odotusarvon arviointia luottamusvälin perusteella.

Esim. 7.6.6 Halutaan arvioida poikien keskimääräistä syntymäpituutta, siis poikapopulaation keskiarvoa. Otoksessa 65 pojan syntymäpituuden keskiarvo oli 50,95 cm ja keskihajonta 1,97 cm (SAIDIT-aineisto). Arvio voidaan tehdä muodostamalla populaation odotusarvon luottamusväli, joka laskussa käytetään otoskeskiarvoa ja otoshajontaa. Tulokseksi saadaan, että poikien keskipituuden arvellaan olevan välillä 50,5 cm – 51,4 cm.

Descriptives

			Statistic	Std. Error
PITUUS	Mean		50,95	,245
	95% Confidence Interval for Mean	Lower Bound	50,47	
		Upper Bound	51,44	
	Std. Deviation		1,972	

Esimerkin 7.6.6 tilanteessa oletetaan, että tehdään satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 tuntematon. Jos populaation varianssi tunnettaisiin, niin otoskeskiarvo noudattaisi normaalijakaumaa, ja luottamusvälin määrittäisiin normaalijakauman perusteella. Tässä esimerkissä (kuten käytännössä aina) populaation varianssi on tuntematon, jolloin luottamusvälin määrittämisessä käytetään hyväksi satunnaismuuttujaa

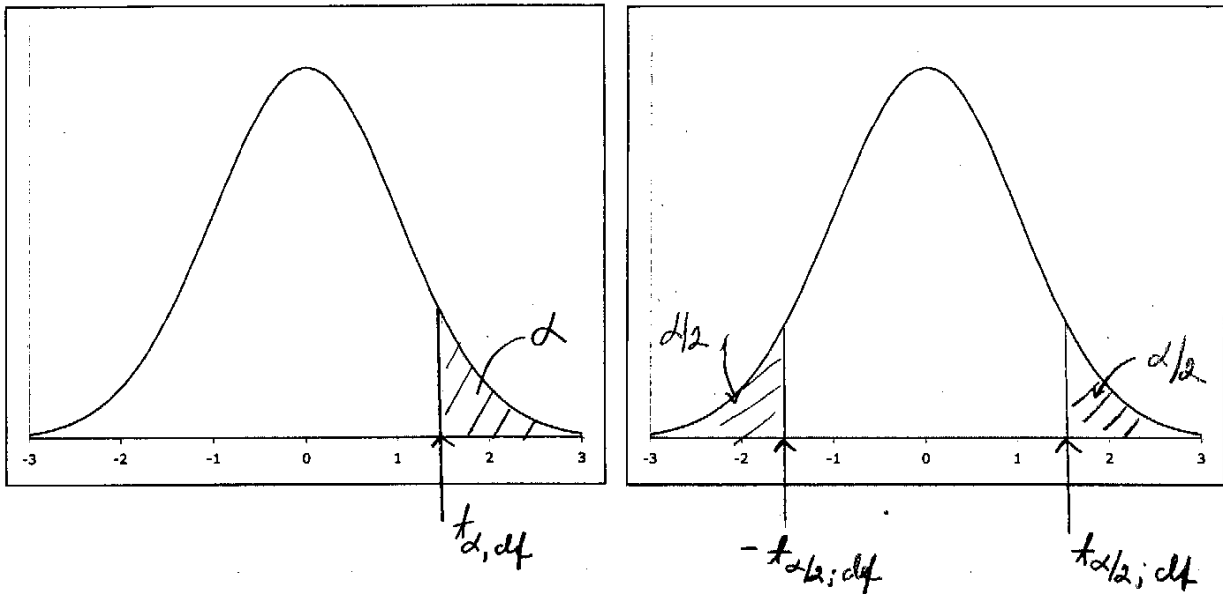
$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

joka noudattaa ns. Studentin t-jakaumaa vapausastein $n-1$.

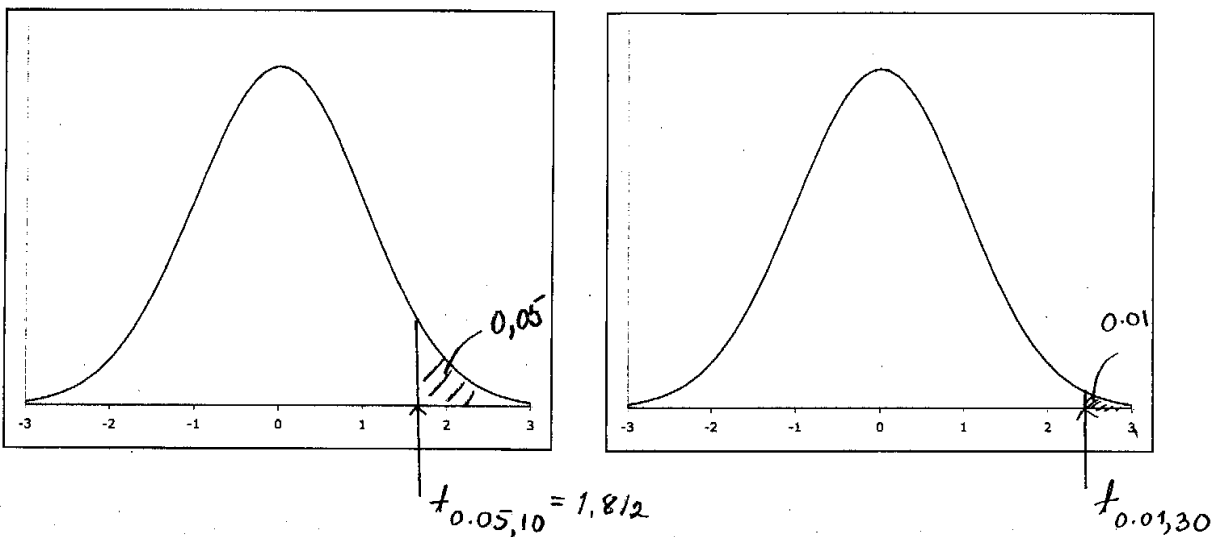
Studentin t-jakauma, joka määritellään nk. vapausastein (df), on jatkuva, origon suhteen symmetrinen jakauma, merkitään t_{df} . Suurilla vapausasteilla t-jakauma lähestyy standardoitua normaalijakaumaa. Ks.

http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/t-jakauma/muodon_riippuvuus_vapausasteista

Olkoon t_{df} Studentin t-jakaumaa noudattava satunnaismuuttuja. Määritellään $t_{\alpha,df}$ ja $t_{\alpha/2,df}$ siten, että $P(t_{df} \geq t_{\alpha,df}) = \alpha$ ja $P(t_{df} \geq t_{\alpha/2,df}) = \alpha/2$, graafisesti:



Näitä arvoja eri vapausastein on taulukoitu (ks. liite 2). Taulukosta saadaan esimerkiksi $t_{0,05;10} = 1,812$ ja $t_{0,01;30} = 2,457$, graafisesti



Odotusarvon luottamusväli määritellään nyt

$$\bar{X} \pm t_{\alpha/2; n-1} s / \sqrt{n},$$

missä s on otoshajonta.

Esim. 7.6.7. Esimerkin 7.6.6 tilanteessa odotusarvon 95 %:n luottamusväli 50,5 cm – 51,4 cm on laskettu $50,95 \pm 2 \cdot 1,972 / \sqrt{65}$, ($t_{0,05/2,65-1} \approx 2$). Esimerkin 7.6.1 tilanteessa neliöhinnan odotusarvoa (keskiarvoa koko populaatiossa) voidaan arvioida luottamusvälillä $2397,61 \pm 1,98 \cdot 408,025 / \sqrt{103}$ ($t_{0,05/2,103-1} \approx 1,98$). Arvellaan siis kyseisellä alueella neliöhinnan keskiarvon olevan välillä 2318 €– 2477 €.

7.6.3 Kahden populaation odotusarvojen erotuksen luottamusväli

Luottamusväli voidaan laskea myös kahden populaation odotusarvojen erotuksella (tällä opintojaksolla ei esitellä laskukaavaa, käytetään ohjelmiston antamia tuloksia). Tällöin ollaan kiinnostuneita siitä, onko kahden populaation keskiarvot samoja.

Esim. 7.6.8. Tarkastellaan Tampereella keväällä 2006 myynnissä olleita kerrostalohuoneistoja (aineisto Asunnot_2006 sivulla <http://www.uta.fi/~strale/tiltp1/aineistoja.html>). Verrataan keskustassa ja lähiöalueella olevien huoneistojen neliöhintoja. Tehdyn analyysin tuloksesta (alla) löytyy 95 %:n luottamusväli (-989,844, -798,862) odotusarvojen erotukselle. Tämän perusteella arvellaan lähiöasuntojen neliöhintojen keskiarvon olevan 799 – 990 euroa alhaisempi kuin keskustan keskineliöhinta.

Group Statistics

Onko keskustassa?	N	Mean	Std. Deviation	Std. Error Mean
Neliöhinta Ei ole	126	1503,2538	325,34129	28,98371
On	103	2397,6072	408,02462	40,20386

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Neliöhinta	1,235	,268	-18,455	227	,000	894,35342	48,46101	-989,844	-798,862
Equal variances assumed									
Equal variances not assumed			-18,045	193,029	,000	894,35342	49,56214	-992,106	-796,601

Esim. 7.6.9. Jos halutaan selvittää, ovatko pojat ja tytöt syntyessään keskimäärin saman painoisia, niin tehdään tyttö- ja poikapopulaatioista satunnaisotokset ja arvioidaan otoskeskiarvojen avulla kahden populaation odotusarvojen yhtäsuuruutta. Nyt voidaan muodostaa kahden populaation odotusarvojen erotuksen luottamusväli. Esimerkin 7.7.6 tuloksista löytyy otoksesta laskettu luottamusväli, jonka perusteella arvellaan odotusarvojen erotuksen olevan välillä 15 g – 363 g. Koska nolla ei kuulu saadulle tuottamuskäytölle, päätellään, että tytöt ja pojat eivät ole syntyessään keskimäärin saman painoisia.

Esimerkin 7.6.8 ja 7.6.9 tilanteissa oletetaan, että populaatiot ovat normaalisti jakautuneita. Tarkasteltiin siis poika- ja tyttöpopulaatioissa syntymäpainoja, keskusta- ja lähiöalueen asuntojen neliöhintoja. Näiden populaatioiden variansseja ei tunneta, mutta oletetaan ne yhtä suuriksi (siis tytöllä ja pojilla, keskustassa ja lähiöalueella).

Luottamusväliä kahden populaation odotusarvojen erotukselle voidaan käyttää, kun selitettävä muuttuja on kvantitatiivinen ja selittäjä on kaksiluokkainen (tai luokiteltu siten). Jos luottamusväli sisältää nollan, niin voidaan tehdä johtopäätelmä, että odotusarvot ovat samoja.

Ks. luottamusväleistä <http://www.fsd.uta.fi/menetelmaopetus/paattely/paattely.html>

SPSS

Ks. luku 7.7.4.

7.7 Hypoteesien testausta

Tilastollinen hypoteesi on väittämä populaatiosta (tai populaatioista), sen jakaumasta ja/tai jakauman parametrissa. Hypoteesin testaus tarkoittaa väittämän tutkimista otoksen perusteella. Väitteen paikkansapitävyyttä tutkitaan otoksen (käytettävissä olevan aineiston) perusteella laskemalla tilanteeseen sopiva otossuure, jota kutsutaan testisuureksi. Tämän testisuureen arvon perusteella joko uskotaan väite tai ei uskota (jolloin vaihtoehtoinen väite hyväksytään). Johtopäätelmän tekeminen perustuu siihen, että selvitetään voidaanko otoksesta laskettua testisuureen arvoa väitteen ollessa tosi pitää "tavanomaisten" arvojen joukkoon kuuluvana vai katsotaanko se harvinaisten arvojen joukkoon kuuluvaksi. Jos testisuureen arvo kuuluu harvinaisten arvojen joukkoon, niin ei uskota väitettä.

Mikä sitten on harvinaista? Testauksessa harvinaisiksi arvoiksi katsotaan sellaisten arvojen joukko, jonka todennäköisyys on melko pieni, esim. 0,05, 0,025, 0,01, 0,001. Kun valitaan tämä todennäköisyys, niin kiinnitetään riskitaso. Harvinaisten arvojen raja eli kriittinen arvo voidaan katsoa testisuureen jakauman taulukosta, mutta testauksessa on kuitenkin usein tapana ilmoittaa nk. p -arvo, joka kertoo todennäköisyyden saada väitteen ollessa tosi otoksesta saatua arvoa harvinaisempi arvo. Tämä on siis pienin riskitaso, jolla asetettu väite voidaan hylätä. Jos siis testaukseen liittyvä p -arvo on pieni, sanotaan vaikkapa 0,001, niin asetettua väitettä ei uskota; se hylätään ja hyväksytään vaihtoehtoinen väittämä. Se milloin p -arvon katsotaan olevan tarpeeksi pieni, riippuu siitä millainen todennäköisyys sallitaan sille, että tehdään väärä johtopäätelmä; väärä siten, että väittämä hylätään vaikka sen on tosi. Tämä virhetodennäköisyys ei saa olla suuri, ja sen halutaan usein olevan suuruusluokkaa pienempi kuin 5 % (melkein merkitsevä), 1 % (merkitsevä), 0,1 % (erittäin merkitsevä).

Hypoteesin testauksessa asetetaankin siis kaksi väittämää, joista toinen on välttämättä voimassa. Nollahypoteesi H_0 , jonka ollessa tosi, testisuuren todennäköisyysjakauma tunnetaan, sekä vaihtoehtoinen hypoteesi H_1 . Nollahypoteesi H_0 tulee aina asettaa käytetyn testin sanelemalla tavalla.

Mikä sitten on testattava hypoteesi H_0 ? Mitä testiä (t-testi, χ^2 -testi, z-testi...) käytetään? Miten johtopäätelmä tehdään?

Testaus voi liittyä populaation (tai populaatioiden) parametriin (tai parametreihin) kuten esim. t-testi kahden populaation odotusarvojen (keskiarvojen) yhtäsuuruuden selvittämiseen, z-testi prosenttiosuuden tutkimiseen (puolueen kannatus). Voidaan myös tutkia kahden populaation välistä lineaarista riippuvuutta (t-testi kahden muuttujan välisen korrelaation tutkimiseen) tai ylipäätään riippuvuutta kahden muuttujan välillä (χ^2 -riippumattomuustesti ristiintaulukon pohjalta).

Ks. <http://www.fsd.uta.fi/menetelmaopetus/hypoteesi/testaus.html>

7.7.1 Prosenttiosuuden testaus

Tarkastellaan lähemmin tilannetta, jossa halutaan tehdä populaatiossa tietyn tyyppisten alkioiden ("viallisten") prosentuaalista osuutta koskeva päättely. Olkoon populaatiossa π % "viallisia". Tehdään satunnaisotos tästä populaatiosta.

Nyt asetetaan

$$H_0: \pi = \pi_0 \text{ ja}$$

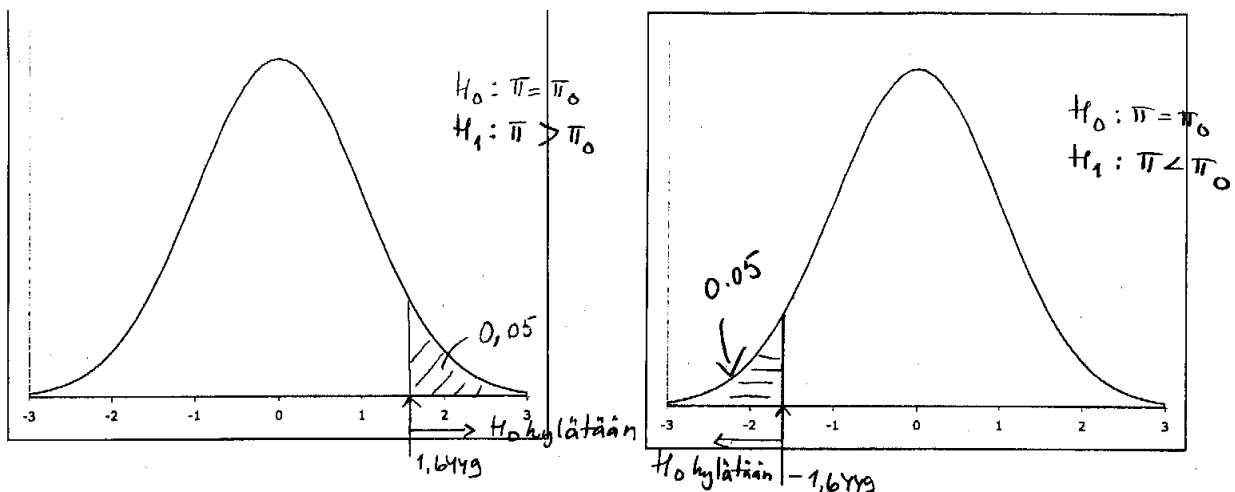
$$H_1: \pi \neq \pi_0 \quad (\text{kaksisuuntainen testi tai vaihtoehtoisesti joko } H_1: \pi > \pi_0 \text{ tai } H_1: \pi < \pi_0, \text{ yksisuuntainen testi}).$$

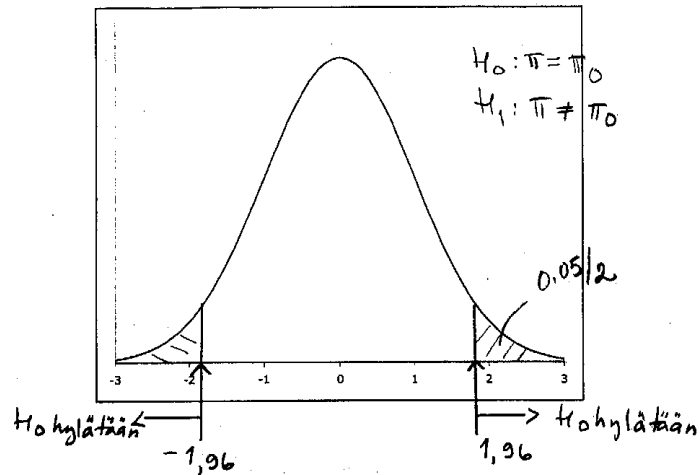
Jos H_0 on tosi, niin testisuure

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}}$$

missä p on otoksessa "viallisten" %-osuus, noudattaa likimain standardoitua normaalijakaumaa. Testaus suoritetaan siten, että lasketaan otoksesta testisuurelle arvo. Sitten katsotaan kuuluuko se standardoidun normaalijakauman tavanomaisten arvojen joukkoon vai onko se harvinaisten arvojen joukkoon kuuluva. Harvinaisten arvojen raja riippuu valitusta riskitasosta (ja myös vaihtoehtoisesta hypoteesista). Jos saadaan harvinaisten arvojen joukkoon kuuluva, niin kyseisellä riskitasolla nollahypoteesi hylätään ja vaihtoehtoinen hyväksytään.

Päättely 5%:n riskitasolla sekä yksi- että kaksisuuntaisissa testeissä:





Riskitason muuttuessa taulukkoarvo muuttuu.

Esim. 7.7.1. Eräs puolue väittää, että suomalaisista 40 % kannattaa sitä. Väitteen tutkimiseksi teit kyselyn 1000 henkilölle, joista 360 ilmoitti kannattavansa kyseistä puoluetta. Onko puolue arvioinut kannatuksensa oikein?

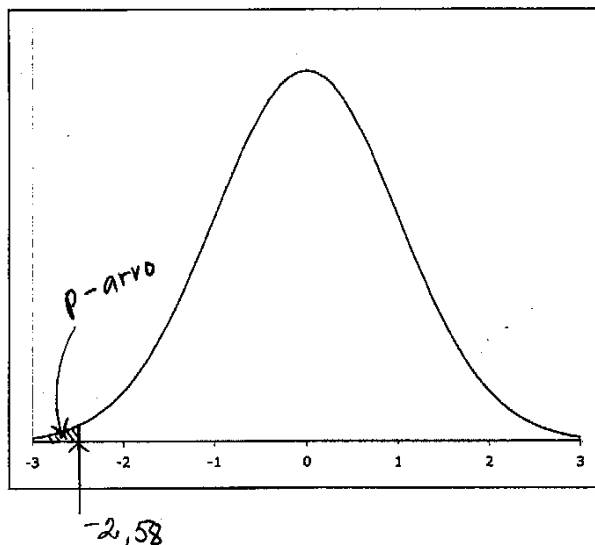
Nyt

$H_0: \pi = 40\%$ ja $H_1: \pi < 40\%$ (yksisuuntainen testi).

Otoksessa kannattajia 36 %. Aineiston perusteella testisuureen arvoksi saadaan

$$z = \frac{36 - 40}{\sqrt{40(100 - 40)/1000}} = -2,58.$$

Onko tämä harvinaisten arvojen joukkoon kuuluja? Jos H_0 tosi, niin $P(Z \leq -2,58) \approx 0,005$ (taulukko, liite 2). Siis pienin riskitaso, jolla H_0 voidaan hylätä on n. 0,005, graafisesti:



Koska p-arvo on näin pieni (esimerkiksi $<0,01$), H_0 hylätään. Päätellään, että puolue on arvioinut kannatuksensa liian suureksi. Jos $H_1: \pi \neq 40\%$ (kaksisuuntainen testi), niin $p = P(Z \leq -2,58) + P(Z \geq 2,58) \approx 0,01$.

Esim. 7.7.2. Tarkastellaan erään liikkeen asiakkaita. Halutaan tutkia, onko asiakkaista yli puolet naisia. Tehdään 200 asiakkaan satunnaisotos, jossa naisia on 113.

Nyt

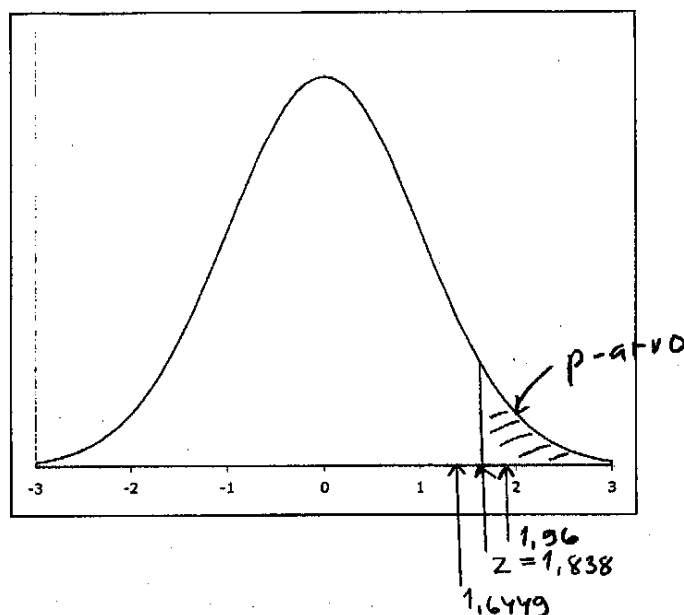
$$H_0: \pi = 50\% \text{ ja}$$

$$H_1: \pi > 50\%.$$

Otoksessa naisia 56,5 %. Aineiston perusteella testisuureen arvoksi saadaan

$$z = \frac{56,5 - 50}{\sqrt{50(100 - 50)/200}} = 1,838.$$

Onko tämä harvinaisten arvojen joukkoon kuuluja? Jos tehdään testaus 5 %:n riskitasolla, niin harvinaisiksi arvoiksi katsotaan lukua 1,6449 suuremmat ja täten hylätään nollahypoteesi ja päätellään asiakkaista olevan naisia enemmän, graafisesti:



Riskitasolla 2,5 % nollihypoteesi hyväksytään, koska harvinaisten arvojen raja on 1,96. Jos H_0 tosi, niin $0,025 < P(Z > 1,838) < 0,05$ (taulukko, liite 2). Siis pienin riskitaso p , jolla H_0 voidaan hylätä on $0,025 < p < 0,05$.

Tilastollisten testin suorittaminen tapahtuu periaatteessa kaikissa tilanteissa esimerkeissä 7.7.1 ja 7.7.2 esitetyllä tavalla. Asetetaan testattava hypoteesi, lasketaan testisuureen arvo ja pienin riskitaso, jolla nollihypoteesi voidaan hylätä. Tämän p -arvon (tai vaihtoehtoisesti kiinnitetyllä riskitasolla katsotun taulukkoarvon) perusteella joko hyväksytään väittämä tai hylätään se. Huomaa, että päättely riippuu myös vaihtoehtoisesta hypoteesista, siis siitä tarkastellaanko testiä yksi- vai kaksisuuntaisena. Eri tilanteissa nollihypoteesi, testisuure ja sen jakauma ovat vain erilaisia.

SPSS

Luottamusväli tai z-testi prosentuaaliselle osuudelle. Ohjelmistolla lasketaan prosentuaalinen osuus aineistossa esim. frekvenssijakauman avulla

Analyze

Descriptive Statistics >

Frequencies...

ja tämän jälkeen voi laskea kyseisen luottamusvälin tai testisuureen (kaavat 8,10).

7.7.2 Odotusarvon testaus

Toisena esimerkkinä tarkastellaan populaation odotusarvon testausta. Oletetaan, että tehdään satunnaisotos normaalijakaumasta, jonka varianssi on tuntematon. Nyt testattava hypoteesi on

$$H_0: \mu = \mu_0$$

ja testisuure

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

joka noudattaa Studentin t-jakaumaa vapausastein $n - 1$, kun H_0 tosi. Tässä p-arvoa voidaan arvioida t-jakauman taulukon avulla.

Edellä esitettyä t-testiä voidaan käyttää suurten otosten tapauksessa vaikka populaatio ei olisikaan normaalin.

Esim. 7.7.3. Kauppias väitti, että kananmunien keskipaino on 50 g. Tehdään 36 alkion satunnaisotos ja saadaan $\bar{x} = 47$, $s = 6$. Onko kauppiaan väittämään uskomista? Nyt

$$H_0: \mu = 50 \text{ g ja}$$

$$H_1: \mu < 50 \text{ g}$$

Saadaan

$$t = \frac{47 - 50}{6 / \sqrt{36}} = -3,$$

joka on pienempi kuin $-t_{0,005;35} \approx -2,75$ (taulukko, liite 2), joten kauppiaan väitettä ei voida uskoa. Hylätään nollahypoteesi 0,5 %:n riskitasolla. Jos haluaa määrittää p-arvon, niin on laskettava $P(t_{35} < -3)$, joka siis on pienempi kuin 0,005. Vaihtoehtoinen hypoteesi voidaan myös asettaa kaksisuuntaisena eli

$$H_1: \mu \neq 50 \text{ g, jolloin } p = P(t_{35} < -3) + P(t_{35} > 3) < 0,01.$$

Esim. 7.7.4. Perunalastujen valmistaja ilmoittaa perunalastupussiensä keskipainoksi 340 g. Tutkitaan väitettä ja tehdään 16 pussin satunnaisotos ja saadaan keskipainoksi 336 g ja keskihajonnaksi 11 g. Voitko uskoa valmistajan väitteen?

Nyt

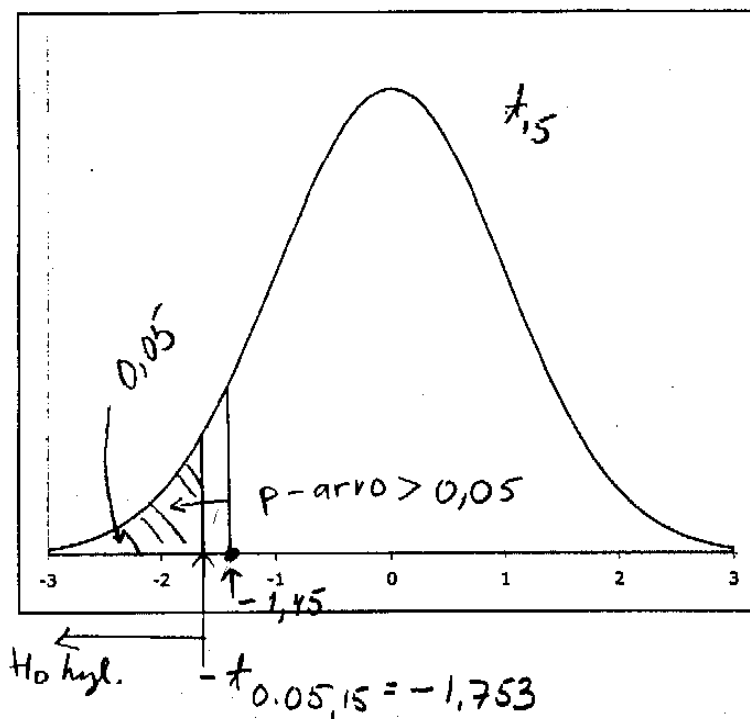
$$H_0: \mu = 340 \text{ g ja}$$

$$H_1: \mu < 340 \text{ g (tai kaksisuuntaisena } H_1: \mu \neq 340 \text{ g)}$$

ja

$$t = \frac{336 - 340}{11 / \sqrt{16}} = -1,45,$$

joka on suurempi kuin $-t_{0,05;15} = -1,753$ (taulukko, liite 2). Valmistajan väite siis uskotaan 5 %:n riskitasolla tarkastellen. Jos haluaa määrittää p-arvon, niin on laskettava $P(t_{15} < -1,45)$, joka siis on suurempi kuin 0,05, sillä taulukosta $P(t_{15} < -1,753) = 0,05$, graafisesti:



Uskotaan valmistajan väite 5 %:n riskitasolla tarkasteltuna. Sama johtopäätelmä tehdään, jos vaihtoehtoinen hypoteesi asetetaan kaksisuuntaiseksi ($t_{0,05/2;15} = 2,131$).

SPSS

Luottamusväli populaation odotusarvolle sekä t-testi odotusarvon testaamiseksi

Analyze

Compare Means >

One-Sample T-test...

tai

Descriptive Statistics->

Explore

Tarkasteltavan muuttujan oltava vähintään intervalliasteikollinen.

7.7.3 χ^2 -riippumattomuustesti

Edellä tarkasteltiin testejä, jotka liittyvät yhden populaation parametriin. Tilastollisen tutkimuksen teossa usein mielenkiintoisempaa ja olennaisempaa on riippuvuustarkasteluihin liittyvät testaukset.

Kahden kvalitatiivisen muuttujan välinen riippuvuustarkastelu voidaan tehdä ristiintaulukon avulla vertailemalla selitettävän muuttujan ehdollisia prosenttijakaumia (ks. luku 5.2.2). Riippuvuuden merkitsevyys voidaan testata. Testisuurena käytetään χ^2 -riippumattomuustestisuuretta ja hypoteesit asetetaan

H_0 : ei riippuvuutta

H_1 : on riippuvuutta

Kun nollahypoteesi on tosi, niin testisuure noudattaa nk. χ^2 -jakaumaa (ks. <http://www.mathstat.helsinki.fi/tilastosanasto/sanasto/chi2-jakauma>). χ^2 -testisuureen arvot ovat ≥ 0 ja harvinaisten arvojen joukko muodostuu "suurista" arvoista.

Testin käyttöön liittyy joitain oletuksia (ei kuitenkaan mitta-asteikkovaatimuksia). Tilanteissa, jossa ristiintaulukointi on tehty siten, että molemmilla muuttujilla on kaksi luokkaa, testiä voidaan käyttää, jos $n > 40$ tai jos $20 \leq n \leq 40$ kaikkien nk. teoreettisten frekvenssien (frekvenssit, jos riippuvuutta ristiintaulukon perusteella ei olisi) oltava ≥ 5 . Muulloin on kaikkien teoreettisten frekvenssien oltava > 1 sekä enintään 20% saa olla < 5 . Jos vaatimukset eivät täyty, on ristiintaulukointi tehtävä uudella luokituksella.

Ks. <http://www.fsd.uta.fi/menetelmaopetus/ristiintaulukointi/ristiintaulukointi.html>

Esim. 7.7.5. Tarkastellaan eräältä kurssilta saatua kurssipalautetta.

(ks. aineisto http://mtl.uta.fi/tilasto/tiltp_aineistoja/arvio.sav ,

http://mtl.uta.fi/tilasto/tiltp_aineistoja/arvio.xls , kuvaus

http://mtl.uta.fi/tilasto/tiltp_aineistoja/arviointi_lomake.pdf).

Halutaan selvittää, onko opintosuunnalla vaikutusta annettuun palautteeseen.

Aineistossa on muuttuja *OPINTOJAKSON TYÖLÄYS*, joka kertoo vastaajan mielipiteen opintojakson työläydestä (työläs/sopiva/vähätöinen) sekä palautteen antajan opintosuunta (*OPSUUNTA*). Nyt asetetaan

H_0 : Opintosuunta ei vaikuta annettuun arvioon

H_1 : Opintosuunta vaikuttaa annettuun arvioon.

Ristiintaulukointi tuottaa taulukon

Opintojakson työläys * OPSUUNTA Crosstabulation

			OPSUUNTA		Total
			hallinto	taloust	
Opintojakson työläys	työläs	Count	13	16	29
		% within OPSUUNTA	68.4%	34.8%	44.6%
	sopiva	Count	5	15	20
		% within OPSUUNTA	26.3%	32.6%	30.8%
	vähätöinen	Count	1	15	16
		% within OPSUUNTA	5.3%	32.6%	24.6%
Total	Count	19	46	65	
	% within OPSUUNTA	100.0%	100.0%	100.0%	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,668 ^a	2	,022
Likelihood Ratio	8,680	2	,013
Linear-by-Linear Association	7,548	1	,006
N of Valid Cases	65		

a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,68.

Aluksi huomataan, että *OPINTOJAKSON TYÖLÄYDEN* prosentuaaliset jakaumat opintosuunnittain poikkeavat huomattavasti toisistaan. Mutta ovatko erot riittävän suuria, jotta voidaan tehdä päätelmä riippuvuuden olemassaolosta? Tuloksesta (kohta a.) nähdään ensin, että oletuksen testin käyttöön ovat voimassa (pienin teoreettinen (odotettu) frekvenssi 4,68, alle 5 teoreettisia frekvenssejä 16,7%). χ^2 -riippumattomuustestisuureen arvo (Pearson Chi-Square) on 7,668, joka voidaan katsoa harvinaisten arvojen joukkoon kuuluvaksi, jos harvinaisena pidetään sellaisten arvojen joukkoa, joiden todennäköisyys on esim. pienempi kuin 0,025. Tällöin H_0 hylätään ja H_1 hyväksytään ja tehdään johtopäätelmä, että opintosuunta vaikuttaa annettuun arvioon. Tässä siis p -arvo on 0,022. Jos halutaankin ottaa vain 1 %:n riski, niin silloin H_0 hyväksytään, koska $0,022 > 0,01$.

Esim. 7.7.6. Esimerkin 5.2.5 ristiintaulukosta laskettu χ^2 -riippumattomuustestisuureen arvo on 66,3 ja p-arvo $< 0,0001$. Päätellään, että riippuvuutta on.

SPSS	Ristiintaulukointi ja testaus tehdään valikosta	
Analyze	Descriptive Statistics>	
	Crosstabs...	annetaan sarake- ja rivimuuttujat, lisämääreinä
	Statistics...	-painike>Chi-square, χ^2 -testisuure
	Cells...	-painike, ehdolliset prosenttijakaumat, "suunta" valitaan siten, että saadaan selitettävän prosenttijakaumat selittäjän luokissa.

SPSS muodostaa ristiintaulukon siten, että molempien muuttujien jokainen arvo on omana luokkanaan. Jos on tarve yhdistellä muuttujien arvoja, tehdään se muodostamalla uusi muuttuja havaintomatriisiin (Transform>Recode>).

Kvantitatiivista muuttujaa voi halutessaan käyttää ristiintaulukoinnissa, kunhan sen ensin luokittelee tekemällä uuden muuttujan havaintomatriisiin.

7.7.4 Odotusarvojen yhtäsuuruuden testaaminen t-testillä

Tutkittaessa kvantitatiivisen muuttujan riippuvuutta kvalitatiivisesta muuttujasta, jolla on kaksi luokkaa voidaan käyttää riippumattomien otosten t -testiä kahden populaation keskiarvojen (odotusarvojen) yhtäsuuruuden testaamiseksi. Hypoteesit asetetaan

H_0 : populaation keskiarvot ovat samoja ("ei riippuvuutta")

H_1 : populaation keskiarvot eivät ole yhtä suuria ("on riippuvuutta")

Vaihtoehtoinen hypoteesi voidaan asettaa myös H_1 : toisen populaation keskiarvo on toista suurempi. Riippumattomien otosten t -testissä oletetaan, että käytössä on riippumattomat satunnaisotokset normaalijakaumista, joiden varianssit ovat yhtä suuret, mutta tuntemattomat (tätä oletusta voidaan myös testata). Testisuure, jota käytetään, noudattaa nollahypoteesin ollessa tosi Studentin t -jakaumaa. Siis harvinaisten arvojen joukko muodostuu itseisarvoltaan "suurista" arvoista.

Jos selittävällä muuttujalla on enemmän kuin kaksi luokkaa, niin t -testi voidaan kyllä tehdä pareittainkin, mutta parempi menetelmä olisi yksisuuntainen varianssianalyysi. Tällöin tutkittaisiin myös populaatioiden keskiarvojen yhtäsuuruutta, populaatioita vain voisi olla enemmän kuin kaksi. Varianssianalyysissä saadaan t -testisuureen sijaan nk. F -testisuure. Tätä analyysia ei kuitenkaan tällä opintojaksolla käsitellä.

Esim. 7.7.7. Onko tytöillä ja pojilla eroja syntymäpainossa?

H_0 : Painon keskiarvot samoja molemmissa populaatioissa

H_1 : Painon keskiarvot eivät samoja molemmissa populaatioissa.

Aineistossa http://mtl.uta.fi/tilasto/tiltp_aineistoja/saidit.sav (myös http://mtl.uta.fi/tilasto/tiltp_aineistoja/saidit.xls) on muuttujat *PAINO* (g) ja *SEX*. Nyt siis selitetään muuttujaa *PAINO*, joka on kvantitatiivinen. Selittäjä on *SEX*-muuttuja, joka on

kvalitatiivinen, kaksiluokkainen. Suoritetaan riippumattomien otosten t -testi ja saadaan tulokset

Group Statistics

	SEX	N	Mean	Std. Deviation	Std. Error Mean
PAINO	poika	65	3640.46	438.244	54.357
	tyttö	55	3451.27	523.280	70.559

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
PAINO	Equal variances assumed	.293	.589	2.156	118	.033	189.19	87.765	15.390	362.987
	Equal variances not assumed			2.124	105.703	.036	189.19	89.069	12.595	365.783

Tässä on siis riippumattomat otokset tyttö- ja poikapopulaatioista. Otoskoot 65 ja 55. Syntymäpainon keskiarvojen erotus on 189,19 g. Painon otosvarianssit ($438,24^2$ ja $523,28^2$) poikkeavat toisistaan.

Nyt tuloksista löytyy testisuure (Levene's Test for Equality of Variances) hypoteesille H_0 : Populaatioiden varianssit samoja. Koska tähän liittyvä p -arvo on 0.589 (>0.05), H_0 hyväksytään ja todetaan, että vaatimus varianssien yhtäsuuruudesta voidaan kuitenkin olettaa olevan täytetty. (Jos näin ei olisi, niin t -testin tulokset luettaisiin vastaavasta kohdasta.)

Varsinaisen testisuureen arvo on siis 2,156 ja p -arvo 0,033. Jos riskitasoksi valitaan 5 %, niin nollahypoteesi hylätään (koska $p < 0,05$) ja tehdään päätelmä, että tytöt ja pojat ovat syntyessään keskimäärin eri painoisia. Jos otettaisiin riski, joka olisi pienempi kuin 3,3 % (vaikkapa 1 %) niin tehtäisiin päinvastainen päätelmä.

Tulostuksesta löytyy myös 95% luottamusväli odotusarvojen erotukselle. Testin sijaan voidaan käyttää tätä luottamusväliä johtopäätelmän tekemisessä. Jos luottamusväli sisältää nollan niin populaation keskiarvojen erotus voidaan arvioida olevan nolla (eri tyttö- ja poikapopulaatioissa syntymäpainon keskiarvot samoja). Tässä luottamusväli, jolle populaatioiden keskiarvojen erotuksen arvellaan kuuluvan, on (15,39, 362,99).

Esim. 7.7.8. Esim. 7.6.8, kerrostalohuoneistojen neliöhintojen keskiarvot poikkeavat keskusta- ja lähiöalueella, koska p -arvo pieni ($<0,001$).

SPSS Luottamusväli populaation odotusarvojen erotukselle riippumattomien otosten tilanteessa sekä t-testi odotusarvojen yhtäsuuruudelle

```
Analyze
  Compare Means>
    Independent Samples T-test...
      (riippumattomat otokset) annetaan selitettävä (Test Variable)
      sekä selittävä, ryhmittely -muuttuja (Grouping Variable).
```

Tuloksena saadaan testisuureen lisäksi myös ehdolliset keskiarvot ja varianssit sekä testisuure varianssien yhtäsuuruuden testaamiseksi. Riippuvan muuttujan oltava vähintään intervallasteikollinen; selittävä muuttuja kahdessa luokassa (tai valitaan kaksi luokkaa tarkasteluun). Yksisuuntainen varianssianalyysi löytyy kohdasta Compare Means> One-Way ANOVA...

7.7.5 Lineaarinen riippuvuus

Korrelaatiokerroin r (Pearsonin tulomomenttikorrelaatiokerroin, ks. luku 5.2.3) mittaa lineaarisen riippuvuuden voimakkuutta. Otoksesta laskettua korrelaatiokerrointa käyttäen voidaankin testata, onko populaatiossa kahden muuttujan välinen korrelaatiokerroin nolla. Tällöin

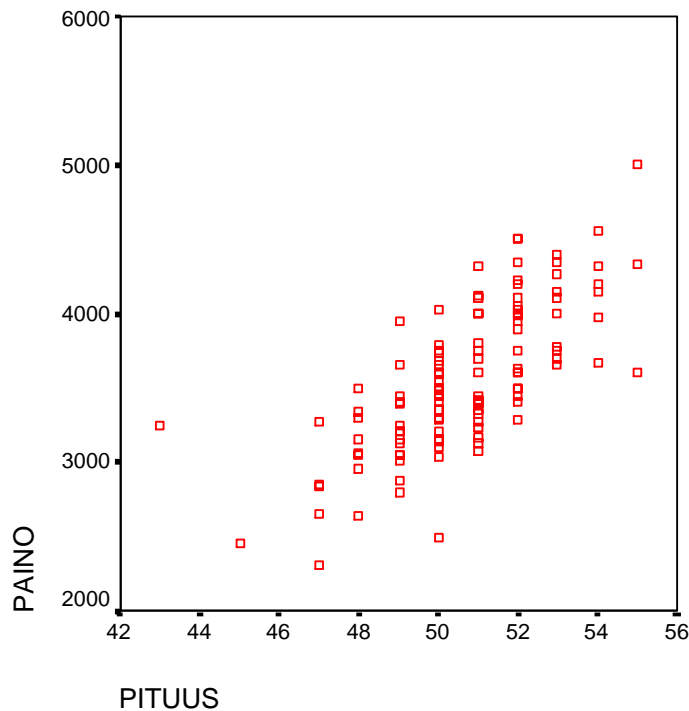
H_0 : populaatiossa kahden muuttujan välinen korrelaatiokerroin on nolla (“ei lineaarista riippuvuutta”)

H_1 : populaation kahden muuttujan välinen korrelaatiokerroin ei ole on nolla (“on lineaarista riippuvuutta”).

Tässä käytetään testisuuretta (ks. kaava (14)), joka noudattaa Studentin t -jakaumaa H_0 :n ollessa tosi. Harvinaiset arvot muodostuvat itseisarvoltaan “suurista” arvoista.

Esim. 7.7.9. Esimerkin 2.4 a) aineistosta

http://mtl.uta.fi/tilasto/tiltp_aineistoja/saidit.sav pisteparvi lapsen syntymäpituuden ja -painon välillä kertoo lineaarisesta riippuvuudesta ja korrelaatiokerroin on 0,72.



Kun testataan hypoteesia H_0 : lapsen paino ja pituus eivät riipu lineaarisesti toisistaan, se hylätään, koska $p = 0,000$. Lineaarista riippuvuutta siis on.

Esim. 7.7.10. Esimerkistä 5.2.7, p-arvo on pienempi kuin 0,001 samoin esimerkistä 5.2.9.

SPSS

Ks. luku 5.2.3

8 LOPUKSI

Tällä opintojaksolla tarkasteltiin empiirisen tutkimuksen aloittamiseen liittyviä asioita. Opintojaksolla esiteltiin aluksi kuvailevan tilastotieteen käyttöä empiirisessä tutkimuksessa. Esiteltiin empiirisen tilastollisen tutkimuksen työvaiheita lähtien havaintoaineiston hankinnasta ja aineiston esittämisestä havaintomatriisimuodossa. Havaintoaineiston sisältämän tiedon tiivistämis- ja havainnollistamistapoihin tutustuttiin muuttujien tunnuslukujen ja frekvenssijakaumien yhteydessä. Muuttujien välisiä riippuvuuksia tutkittiin aluksi ehdollisten tunnuslukujen, pisteparvien ja ristiintaulukoiden avulla. Opintojakson asiat liittyivät suurelta osin kuvailevaan analyysiin ja riippuvuustarkastelujakin esiteltiin aluksi ilman testauksia.

Toinen osa opintojakson asioista liittyi tilastollisen päättelyn lyhyeen ja esimerkinomaiseen esittelyyn. Tilastolliseen päättelyyn ja eri menetelmiin tutustutaan tarkemmin ja laajemmin opintojaksoilla TILTP2

(<http://www.uta.fi/~strale/tiltp2/index.html>) ja TILTP3
(<http://www.uta.fi/~strale/tiltp3/index.html>).

Lopuksi vielä yhteenveto opintojaksolla käsiteltyjen tilastollisten analyysien suorittamisesta *SPSS*-ohjelmalla:

Analyze

Descriptive Statistics>

frekvenssijakaumat, tunnusluvut,
ristiintaulukot,

Compare Means>

t-testit, ryhmäkeskiarvot, muut ehdolliset tunnusluvut

Correlate>

korrelaatiomatriisi

Graphs

Bar... pylväs- ja janadiagrammit

Pie... piirakat,

Boxplot... laatikko-jana-kuviot,

Scatter... pisteparvet,

Histogram... frekvenssihistogrammit.

Sequence... aikasarjan kuvaaja,

Time Series> autokorrelaatiofunktio

LIITE 1 Oheiskirjallisuutta

Muutamia poimintoja hyvin opiskelun tueksi sopivista kirjoista

- Agresti, A. & Finlay, B., *Statistical Methods for the Social Sciences*, 3rd ed., [Prentice Hall](#), 1997.
 Sekä perus- että aineopintotasoisten menetelmien esittelyä (ilman matemaattista johtamista) soveltajan näkökulmasta katsottuna. Paljon esimerkkejä ja tehtäviä.
- Clarke, G.M. & Cooke, D., [A Basic course in Statistics](#), 5th ed., Arnold, 2004.
 Tilastotieteen peruskäsitteistön ja analyysien esittelyä, joka on tasoltaan osittain aineopintotasoista. Paljon kuvia, soveltavia esimerkkejä ja tehtäviä.
- Helenius, H., *Tilastollisten menetelmien perustiedot*. Statcon Oy, 1995.
 Melko laaja teos, jossa tilastotieteen perusasiat käsitellään perusteellisesti. Sisältää paljon esimerkkejä. Kirjasta on varmasti hyötyä myöhemminkin opiskelussa. Kirja niille, jotka kaipaavat tieteellistä, selkeää asioiden käsittelyä.
- Heikkilä, T., *Tilastollinen tutkimus*, 4. painos, Edita, 2002.
 Tiivis yleisteos, jossa tilastotieteen perusasiat ja tilastollisen tutkimuksen tekeminen on selitetty yksinkertaisesti ja melko lyhyesti. Myös joitakin perusopintojen jälkeen tulevia asioita on käsitelty lyhyesti. Kirja sisältää hyvät ohjeet SPSS:n käytöstä ja SPSS -esimerkkejä tulkintoineen. Kirja niille, jotka kaipaavat asioiden "vääntämistä rautalangasta".
- Leppälä, R., [Ohjeita tilastollisen tutkimuksen toteuttamiseksi SPSS for Windows -ohjelmiston avulla](#). Tampereen yliopisto, MTF laitos, B53, 2004.
- Moore, D., [The Basic Practice of Statistics](#), 4th ed. Freeman, 2007.
 Perusopintojen asioita käytännönläheisesti, ilman johtamista.

Lisälukemistoa

- Devore, J. & Peck, R., *Statistics, The Exploration and Analysis of Data*. West Publishing Company, 2005.
- Freund, J., *Modern Elementary Statistics*. [Prentice-Hall](#), 2004.
- Grönroos, M., *Johdatus Tilastotieteeseen - Kuvailu, mallit ja päättely*, Finn Lectura, 2004.
- Karjalainen, L., *Tilastomatemiikka*, 8. painos, [Pii-kirjat](#), 2004.
- Larson, R. & Farber, B., *Elementary statistics: Picturing the World*, [Prentice-Hall](#), 2006.
- Liski, E. & Puntanen, S., *Tilastotieteen peruskurssi I & II*, Tampereen yliopisto, 1987
- Mellin, I., *Johdatus tilastotieteeseen*, 1. kirja, tilastotieteen johdantokurssi. Helsingin yliopisto, 1996
- Moore, D. & Notz, W., [Statistics: Concepts and Controversies](#), 6th ed. Freeman, 2006.
- Moore, D. & McCabe G., [Introduction to the Practice of Statistics](#), 5th Edition. Freeman, 2005.
- Newbold, P., *Statistics for Business and Economics*. [Prentice Hall](#), 2003.
- Siegel, A., *Statistics and Data Analysis An Introduction*, 2nd ed., [John Wiley & Sons](#), 1998.
- Weiss, N., [Elementary Statistics](#), 6th ed., Addison Wesley, 2004.
- Weiss, N., [Introductory Statistics](#), 7th ed., Addison Wesley, 2004.
- Yates, D. & Moore, D. & McCabe G., [The Practice of Statistics](#), 2nd ed., Freeman, 2003.

LIITE 2 Kaavakokoelma ja taulukot

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

$$(2) \quad s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{SS_x}{n-1}$$

$$(3) \quad s_x = \sqrt{s_x^2}$$

$$(4) \quad r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}}$$

$$= \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

$$(5) \quad X \sim N(\mu, \sigma^2), E(X) = \mu, \text{Var}(X) = \sigma^2, Z \sim N(0,1), P(Z \leq z) = \Phi(z)$$

$$(6) \quad E(\bar{X}) = \mu, \text{Var}(\bar{X}) = \sigma^2 / n$$

$$(7) \quad t = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

$$(8) \quad 100(1-\alpha)\%:n \text{ luottamusväli prosentiosuudelle } p \pm z_{\alpha/2} \sqrt{p(100-p)/n}$$

$$(9) \quad 100(1-\alpha)\%:n \text{ luottamusväli odotusarvolle (varianssi tuntematon)}$$

$$\bar{X} \pm t_{\alpha/2; n-1} s / \sqrt{n}$$

$$(10) \quad H_0 : \pi = \pi_0, Z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}} \stackrel{\text{likimain}}{\sim} N(0,1), \text{ kun tosi } H_0.$$

$$(11) \quad H_0 : \mu = \mu_0, t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \sim t_{n-1}, \text{ kun tosi } H_0.$$

$$(12) \quad \text{Ristiintaulukosta riippumattomuuden testaus: } \chi^2 \sim \chi^2_{(I-1)(J-1)}, \text{ kun ei riippuvuutta.}$$

$$(13) \quad H_0 : \mu_1 = \mu_2, t \sim t_{n+m-2}, \text{ kun tosi } H_0 \text{ (oletetaan riippumattomat otokset ja populaatioiden varianssit yhtä suuriksi, mutta tuntemattomiksi).}$$

$$(14) \quad H_0 : \text{populaatioissa kahden muuttujan korrelaatiokerroin } (\rho) \text{ on nolla,}$$

$$t = \frac{r_{xy}}{\sqrt{(1-r_{xy}^2)/(n-2)}} \sim t_{n-2}, \text{ kun } H_0 \text{ tosi.}$$

Standardoidun normaalijakauman taulukkoarvoja.

$Z \sim N(0, 1)$

z	1,6449	1,9600	2,3264	2,5758	3,0902	3,2905
$\Phi(z) = P(Z \leq z)$	0,9500	0,9750	0,9900	0,9950	0,9990	0,9995
$P(Z \geq z) = 1 - P(Z \leq z) = P(Z \leq -z)$	0,0500	0,0250	0,0100	0,0050	0,0010	0,0005

Esimerkiksi $\Phi(1,96) = P(Z \leq 1,96) = 0,975$, $P(Z \geq 1,96) = 0,025$ eli $z_{0,025} = 1,96$, $P(Z \leq -1,96) = 0,025$.

Studentin t-jakauman taulukkoarvoja $t_{\alpha;df}$, joille $P(t_{df} \geq t_{\alpha;df}) = \alpha$.

df	$\alpha = 0,05$	$\alpha = 0,025$	$\alpha = 0,01$	$\alpha = 0,005$
1	6,314	12,706	31,821	63,656
2	2,920	4,303	6,965	9,925
3	2,353	3,182	4,541	5,841
4	2,132	2,776	3,747	4,604
5	2,015	2,571	3,365	4,032
6	1,943	2,447	3,143	3,707
7	1,895	2,365	2,998	3,499
8	1,860	2,306	2,896	3,355
9	1,833	2,262	2,821	3,250
10	1,812	2,228	2,764	3,169
11	1,796	2,201	2,718	3,106
12	1,782	2,179	2,681	3,055
13	1,771	2,160	2,650	3,012
14	1,761	2,145	2,624	2,977
15	1,753	2,131	2,602	2,947
16	1,746	2,120	2,583	2,921
17	1,740	2,110	2,567	2,898
18	1,734	2,101	2,552	2,878
19	1,729	2,093	2,539	2,861
20	1,725	2,086	2,528	2,845
21	1,721	2,080	2,518	2,831
22	1,717	2,074	2,508	2,819
23	1,714	2,069	2,500	2,807
24	1,711	2,064	2,492	2,797
25	1,708	2,060	2,485	2,787
26	1,706	2,056	2,479	2,779
27	1,703	2,052	2,473	2,771
28	1,701	2,048	2,467	2,763
29	1,699	2,045	2,462	2,756
30	1,697	2,042	2,457	2,750
40	1,684	2,021	2,423	2,704
60	1,671	2,000	2,390	2,660
120	1,658	1,980	2,358	2,617
∞	1,645	1,960	2,326	2,576

Esimerkiksi $t_{0,05;10} = 1,812$, siis $P(t_{10} \geq 1,812) = 0,05$. $P(t_{10} \leq -1,812) = 0,05$.

Kaavakokoelman ja taulukot annetaan tentissä käyttöön, löytyvät myös sivulta
<http://www.uta.fi/%7Estrale/tiltp1/materiaali.html>

LIITE 3 HOTDOG-aineisto.

Aineisto mikroluokkien verkossa hakemistossa P:\pub\pk\data

Type	Taste	\$/oz	Calories	Sodium
Beef	Bland	0,11	186	495
Beef	Bland	0,17	181	477
Beef	Bland	0,11	176	425
Beef	Medium	0,15	149	322
Beef	Medium	0,10	184	482
Beef	Medium	0,11	190	587
Beef	Medium	0,21	158	370
Beef	Medium	0,20	139	322
Beef	Medium	0,14	175	479
Beef	Medium	0,14	148	375
Beef	Medium	0,23	152	330
Beef	Medium	0,25	111	300
Beef	Medium	0,07	141	386
Beef	Medium	0,09	153	401
Beef	Medium	0,10	190	645
Beef	Medium	0,10	157	440
Beef	Medium	0,19	131	317
Beef	Medium	0,11	149	319
Beef	Medium	0,19	135	298
Beef	Scrumptious	0,17	132	253
Meat	Bland	0,12	173	458
Meat	Bland	0,12	191	506
Meat	Bland	0,12	182	473
Meat	Bland	0,10	190	545
Meat	Bland	0,11	172	496
Meat	Bland	0,13	147	360
Meat	Medium	0,10	146	387
Meat	Medium	0,09	139	386
Meat	Medium	0,11	175	507
Meat	Medium	0,15	136	393
Meat	Medium	0,13	179	405
Meat	Medium	0,10	153	372
Meat	Medium	0,18	107	144
Meat	Medium	0,09	195	511
Meat	Scrumptious	0,07	135	405
Meat	Scrumptious	0,08	140	428
Meat	Scrumptious	0,06	138	339
Poultry	Bland	0,08	129	430
Poultry	Medium	0,05	132	375
Poultry	Medium	0,07	102	396
Poultry	Medium	0,08	106	383
Poultry	Medium	0,08	94	387
Poultry	Medium	0,07	102	542
Poultry	Medium	0,09	90	359
Poultry	Medium	0,06	99	357
Poultry	Medium	0,07	107	528
Poultry	Medium	0,08	113	513
Poultry	Medium	0,07	135	426
Poultry	Medium	0,07	142	513
Poultry	Medium	0,07	83	358
Poultry	Medium	0,08	143	581
Poultry	Medium	0,06	152	588
Poultry	Medium	0,07	146	522
Poultry	Scrumptious	0,06	144	545

LIITE 4 PULSSI-aineisto.

Miesten ja naisten lepopulsseja, ks. http://mtl.uta.fi/tilasto/tiltp_aineistoja/pulssi.sav
http://mtl.uta.fi/tilasto/tiltp_aineistoja/pulssi.xls

<u>Sukup.</u>	<u>Pulssi</u>	<u>Sukup.</u>	<u>Pulssi</u>
Nainen	88	Mies	80
Mies	80	Mies	60
Mies	68	Nainen	65
Nainen	70	Nainen	90
Mies	80	Mies	56
Mies	75	Nainen	64
Mies	58	Nainen	86
Mies	82	Mies	70
Mies	78	Nainen	64
Nainen	77	Nainen	88
Nainen	86	Nainen	86
Nainen	78	Nainen	68
Mies	70	Nainen	80
Mies	72	Mies	84
Mies	65	Mies	70
Nainen	72	Nainen	80
Mies	94	Nainen	94
Mies	50	Nainen	80
Mies	84	Mies	65
Nainen	86	Mies	70
Nainen	90	Mies	58
Nainen	66	Mies	45
Mies	72	Mies	66
Nainen	60	Mies	50
Nainen	74	Mies	50
Mies	90	Mies	70
Nainen	80	Mies	84
Mies	72	Nainen	64
Nainen	84	Mies	80
Mies	74	Nainen	84
Mies	70	Nainen	72
Mies	74	Nainen	72
Mies	58	Nainen	64
Mies	70	Mies	58
Nainen	70	Mies	70
Mies	75	Nainen	76
Nainen	76	Nainen	76
Nainen	82	Mies	80
Mies	75	Mies	68
Mies	88	Nainen	100

LIITE 5 Opintojakson TILTP1 kurssiarviointilomakkeet.

TILTP1 syksy 2001

kyselylomake ja jakaumat

<http://mtl.uta.fi/tilasto/tiltp1/palaute.pdf>

TILTP1 syksy 2002

kyselylomake ja jakaumat

<http://www.uta.fi/%7Estrale/p1pal02.html>

TILTP1 syksy 2003

kyselomake ja jakaumat

<http://mtl.uta.fi/tilasto/tiltp1/syksy2003/p1pal03.htm>

TILTP1 syksy 2004

kyselomake ja jakaumat

<http://mtl.uta.fi/tilasto/tiltp1/syksy2004/p1pal04.htm>

TILTP1 syksy 2005

Kyselylomake ja jakaumat

<http://mtl.uta.fi/tilasto/tiltp1/syksy2005/p1pal05.html>

TILTP1 syksy 2006

Kyselylomake ja jakaumat

<http://mtl.uta.fi/tilasto/tiltp1/syksy2006/p1pal06.html>

TILTP1 syksy 2007

Kyselylomake ja jakaumat

<http://mtl.uta.fi/tilasto/tiltp1/syksy2007/p1pal07.html>

TILTP1 syksy 2008

Kyselylomake, jakaumat ja yhteenveto

<http://mtl.uta.fi/tilasto/tiltp1/syksy2008/p1palyht08.html>

TILTP1 syksy 2009

Kyselylomake, jakaumat ja yhteenveto

<http://mtl.uta.fi/tilasto/tiltp1/syksy2009/p1pal09.html>

TILTP1 syksy 2010

Kyselylomake, jakaumat ja yhteenveto

<http://mtl.uta.fi/tilasto/tiltp1/syksy2010/p1pal10.html>

LIITE 6 Menetelmien ja testien ryhmittelystä muuttujan roolin mukaan.

Jotkut muuttuja selitettäviä, jotkut selittäviä:

- Regressioanalyysi (yksi kvantitatiivinen selitettävä, yksi tai useampi kvantitatiivinen selittäjä)
- Kanoninen analyysi (kuten edellä, mutta monta selitettävää, monta selittäjää)
- Kovarianssianalyysi (esim. yksi kvantitatiivinen selitettävä, yksi kvantitatiivien selittäjä ja yksi dummy)
- Varianssianalyysi (yksi kvantitatiivinen selitettävä, yksi tai useampi kvalitatiivinen selittäjä)
- t-testi odotusarvojen yhtäsuuruuden testaamiseksi (yksi kvantitatiivinen selitettävä, yksi kvalitatiivinen selittäjä)
- Z-testi prosentiosuuksien testaamiseksi
- χ^2 -riippumattomuustesti

Kaikki muuttujat samanarvoisia:

- Pääkomponenttianalyysi ja
- Faktorianalyysi (mitataan joitakin abstrakteja, ei mitattavissa olevia käsiteittä havaittujen muuttujien avulla)
- Erotteluanalyysi (tutkitaan muuttujien keskiarvoja eri havaintoyksikköryhmissä, erot pyritään kuvaamaan mahdollisimman pienellä muuttujajoukolla, keskiarvojen pääkomponenttianalyysi)

Ks. myös
SPSS

Analyze

- Descriptive Statistics>
 - Esim. ristiintaulukot ja χ^2 -riippumattomuustesti
- Compare Means>
 - Esim. keskiarvotestit (t-testit, varianssianalyysi)
- General Linear Models>
 - Erilaisten mallien rakentaminen (voi olla yksi (->Univariate) tai useampia (->Multivariate) kvantitatiivinen selitettävä)
- Correlate>
 - mm. korrelaatiokertoimen ja niiden testaus
- Regression>
 - mm. regressioanalyysi, logistinen regressio (selitettävä kvalit. selittäjä kvantitat.)
- Loglinear>
 - Malleja, joissa selitettävä ja selittäjä kvalitatiivinen
- Classify>
 - Menetelmiä ryhmitellä tilastoyksiköitä
- Data Reduction>
 - mm. faktorialalyysi
- Scale>
 - mm. Cronbachin alpha
- Nonparametric tests>
 - Ei-parametrisia testejä
- Time Series>
 - Aikasarjojen analysointia

Ks. SPSS Help -> Topics-> Index menetelmien kuvauksia.