Raija Leppälä

Ohjeita tilastollisen tutkimuksen toteuttamiseksi IBM SPSS Statistics -ohjelmiston avulla

TAMPEREEN YLIOPISTO

INFORMAATIOTIETEIDEN YKSIKÖN RAPORTTEJA 55/2017

TAMPERE 2017

TAMPEREEN YLIOPISTO
INFORMAATIOTIETEIDEN YKSIKÖN RAPORTTEJA 55/2017
KESÄKUU 2017

Raija Leppälä

Ohjeita tilastollisen tutkimuksen toteuttamiseksi IBM SPSS Statistics -ohjelmiston avulla

ISBN 978-952-03-0501-7 (pdf)

ISSN-L 1799-8158
ISSN 1799-8158

Aluksi

Tämä opas on tarkoitettu tilastollisen tutkimuksen tekemisen oppaaksi, kun toteutuksessa käytetään IBM SPSS Statistics -ohjelmistoa. Tässä oppaassa on päivitetty aiemmin tehtyä opasta *Ohjeita tilastollisen tutkimuksen toteuttamiseksi SPSS for Windows-ohjelmiston avulla*, joka on tehty vuosien varrella Tampereen yliopistossa pitämieni tilastotieteen peruskurssien yhteydessä tapahtuneiden SPSS-ohjelmiston käytön opetuksen sekä erillisten SPSS-kurssien pohjalta. Oppaassa kiinnitetään erityistä huomiota analyysin oikeaan valintaan ja tulosten tulkintaan. Eri toimintojen teknistä toteuttamista ei esitellä kovin yksityiskohtaisesti, vaan annetaan ohjeet siitä, mistä haluttu analyysi löytyy. Ohjeistus on tehty ohjelmistoversion 23 mukaisesti, mutta soveltuu hyvin myös muille versioilla. Tarvittaessa teknisiä oppaita sekä online-kursseja löytyy lukuisia.

Lukijalta edellytetään perustietoja tilastotieteestä, mutta toisaalta opasta voi käyttää vaikka tiedot olisivatkin melko vähäiset. Opas tarjoaakin mahdollisuuden tietojen ajan tasalle saattamiseksi sekä täydentämiseksi, sillä oppaassa esitellään lyhyesti myös tilastotieteen perusteita. Opas soveltuu tilastollisen analyysin tekemisen tueksi silloin, kun käytetään tavanomaisempia tilastollisia analyysejä. Tilastotieteen teorian perinpohjainen hallitseminen ei siis ole oppaan käytön kannalta tarpeellista. Opas on paremminkin pyritty kirjoittamaan soveltajan näkökulma huomioiden. Pyritään siis avustamaan tutkijaa menetelmien valinnassa ja ohjelmiston antamien tulosten tulkinnassa.

Esimerkeissä käytetyt aineistot kuvauksineen ovat saatavilla (ks. liite), joten lukija voi itse tehdä esimerkkien analyysit. Opas on kirjoitettu siten, että parhaan hyödyn siitä saa tekemällä kaikki oppaassa olevat esimerkit peräkkäin aloittaen esimerkistä 1. Jos lukijalla ei ole käytössään SPSS-ohjelmistoa, hän voi ladata IBM SPSS Statistics -kokeiluversion 14 päivän koekäyttöön sivulta https://www.ibm.com/analytics/us/en/technology/spss/spss-trials.html.

Tämä opas on saatavana myös versiona, jossa joihinkin tässä esiteltyihin esimerkkeihin on lisätty sekä R- että MATLAB-toteutukset. Lukija voi halutessaan siis katsoa, millä tavalla SPSS-ohjelmistolla tehdyn analyysin voi tehdä näillä ohjelmilla.

Jarmo Niemelä on auttanut erinomaisella ammattitaidolla raportin LaTeX-muotoon työstämisessä.

Tampereella 19. kesäkuuta 2017

Raija Leppälä

Sisällysluettelo

1 Johdanto

Tilastollinen analyysi voidaan karkeasti jakaa kuvailevaan analyysiin ja tilastolliseen inferenssiin (päättelyyn). Kuvaileva osuus pyrkii kuvailemaan tietoaineistoa erilaisten graafisten esitysten ja tunnuslukujen sekä taulukoiden avulla. Tilastollinen päättely käsittelee johtopäätelmien tekoa populaatiosta aineiston (otoksen) perusteella. Inferenssi perustuu todennäköisyysjakaumiin ja niiden hyväksi käyttöön erilaisten testien ja analyysien yhteydessä.

Tässä monisteessa esitellään lyhyesti joitain analysointimenetelmiä, annetaan ohjeita menetelmän valinnasta ja analyysin suorittamisesta IBM SPSS Statistics -ohjelmiston avulla sekä tulkitaan esimerkeissä saatuja tuloksia. Lähdetään liikkeelle aineiston tallennuksesta, muokkauksesta ja kuvailusta. Sitten tutustutaan joihinkin testeihin ja menetelmiin, joita voidaan käyttää tilastollisen tutkimuksen teossa.

IBM SPSS Statistics -ohjelmisto on helppokäyttöinen, valikko-ohjattu tilastollinen ohjelmisto, jolla on mahdollista suorittaa empiirisen aineiston tallennus ja muokkaus sekä tilastolliset analyysit graafisine esityksineen. Ohjelmisto sisältää hyvin laajan valikoiman analysointimenetelmiä aina aineiston kuvailuun liittyvistä menetelmistä epäparametrisiin testeihin, monimuuttujamenetelmiin, epälineaarisiin malleihin ja aikasarja-analyysiin.

Vaikka tässä oppaassa oheistus on tehty ohjelmistoversion 23 mukaisesti, niin se soveltuu hyvin myös muille versioille.

2 SPSS-ympäristö

Kun SPSS-ohjelmisto käynnistetään, niin voidaan luoda tyhjä havaintomatriisipohja (New Dataset). Tällöin avautuu Data Editor -näkymä, jossa on kaksi välilehteä: Data View (havaintomatriisi), Variable View (muuttujien määritysikkuna). Lisäksi saadaan automaattisesti Output-ikkuna, jonne tulostuvat kaikki tehtyjen analyysin tulokset.

Päävalikko, jonka avulla käyttäjä pyytää ohjelmistoa suorittamaan toiminnot, sisältää mm. seuraavat kohdat:

File Havaintomatriisin luominen, avaaminen, tallennus, tulostaminen, ohjelmiston käytön lopetus

Data Havaintomatriisiin liittyvien määritysten teko, kuten ehdollistaminen

Transform Muunnosten teko muuttujille, uusien muuttujien määrittäminen olemassa olevien muuttujien avulla, muuttujien luokitusten teko

Analyze Valitaan haluttu analyysi

Graphs Graafisten esitysten tekeminen (esim. jakaumat, pisteparvet, laatikko-jana -kuviot)

Seuraavassa esitellään näiden valikoiden käyttöä tilastollisen tutkimuksen teon edetessä aineiston tallennuksesta analysointeihin. Lähdetään siis liikkeelle havaintoaineiston talletuksesta ja muokkauksesta. Kun aineisto on talletettu, voidaan aineiston analysointi aloittaa jakaumien teolla ja tunnuslukujen laskulla. Kuvailevan osuuden jälkeen on vuorossa tilastollisten analysointien teko riippuvuuksien

selvittämiseksi. Suoritettaessa analyysiä valitaan tilanteeseen sopiva komento, jonka jälkeen ohjelmisto pyytää tarvittavat tiedot. Annetaan muuttuja tai muuttujat, joita analyysissä käytetään. Muuttujat valitaan esillä olevasta muuttujaluettelosta.

3 Havaintomatriisin luominen, muokkaaminen ja ehdollistaminen

Empiirisen aineiston eritysmuotona käytetään havaintomatriisia. Jos tilastoyksiköiden lukumäärä on n ja muuttujien lukumäärä on p, niin havaintomatsiisi muodostetaan muuttujien arvoista tilastoyksiköittäin seuraavasti:

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

missä x_{ij} on i. tilastoyksikön mittaluku ominaisuudelle x_j. Muuttujan x_j jakauma on j. pystyrivi eli sarake havaintomatriisissa. Tilastoyksikön a_i havaintovektori on i. vaakarivi.

Muuttujia on kahdenlaisia: kvalitatiivisia (kategorisia) ja kvantitatiivisia (numeerisia). Kvalitatiivinen mittaaminen on vain laadullista mittaamista ja se voidaan jakaa nominaali- eli luokitteluasteikolliseksi ja järjestys- eli ordinaaliasteikolliseksi mittaamiseksi. Kvantitatiivinen mittaaminen on numeerista mittaamista, mitta-asteikkoina intervalli- ja suhdeasteikko sekä absoluuttinen asteikko.

Käsiteltävä aineisto on aluksi saatettava havaintomatriisimuotoon, joka siis on kaksiulotteinen taulukko, jossa määritellään sarakkeille muuttujat ja kirjataan riveille tilastoyksiköittäin mittaustulokset. Tilastoyksiköitä ei havaintomatriisiin tarvitse nimetä, mutta identifioiva tunnusmuuttuja (vaikka juokseva numero) on useimmiten syytä olla, jotta tarvittaessa löydetään vastaavuus aineiston ja talletetun havaintomatriisin välillä.

Uuden havaintoaineiston talletus aloitetaan muuttujien määrityksillä Variable View -ikkunassa. Määritellään muuttujan nimi (lyhyesti, ei erikoismerkkejä), tyyppi (yleensä numeerinen), näkyvien desimaalien lukumäärä, selitteet muuttujalle ja sen koodeille.

Kun muuttujat on määritelty, syötetään arvot muuttujille jokaiselta tilastoyksiköltä Data View -välilehdellä. Jos tietoja puuttuu, niin kyseiset solut jätetään tyhjiksi. Tällöin ohjelmisto tulkitsee sen puuttuvaksi tiedoksi eikä käyttäjän yleensä tarvitse huolehtia puuttuvan tiedon käsittelystä, koska ohjelmisto jättää sen pois analyyseistä. Joissain graafisissa esityksissä oletusarvoisesti tulee puuttuvan tiedon ryhmä mukaan. Sen saa lisämäärityksillä pois.

Olemassa olevan havaintomatriisin avaaminen ja myös uuden luominen voidaan tehdä valikosta

File ▸

New ▸ **Data...** uuden luominen

Open ▸ **Data...** vanhan avaaminen (oletusarvoisesti näkyvät .sav-tunnisteella olevat, voi avata myös Excel-tiedostot)

Read Text Data... tekstitiedoston lukeminen

Aineistoa Excel- tai tekstitiedostosta luettaessa voivat muuttujien nimet olla ensimmäisellä rivillä, jolloin tieto tästä annetaan tiedoston avaamisen yhteydessä.

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

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

Muuttujille voidaan tehdä muunnoksia myös käytettävissä olevien erilaisten funktioiden avulla. Joissain tilanteissa käyttökelpoinen muunnos on logaritmonti.

Esimerkki 1. Tarkastellaan Rasvaprosentti-aineistoa. Muutetaan paino kilogrammoiksi ja pituus metreiksi sekä lasketaan painoindeksit. Tällöin tehdään kolme uutta muuttujaa: Target Variable on *Paino_kg* ja Numeric Expression $0.454 * paino$, Target Variable on *Pituus_m* ja Numeric Expression $0.0254 * pituus$, Target Variable on *Painoindeksi* ja Numeric Expression $(Paino\_kg)/(Pituus\_m * Pituus\_m)$. ■

Joskus halutaan analysoida aineistoa valitsemalla käsittelyyn mukaan vain tietyt tilastoyksiköt (esimerkiksi aineistossa olevat kaksiot). Voidaan myös haluta tehdä analyysit jonkun muuttujan ryhmissä (esimerkiksi keskustassa ja keskustan ulkopuolella oleville asunnoille). Nämä ehdollistamiset tehdään ennen analyysien tekoa. Ilmoitetaan ohjelmistolle, että jatkossa halutaan analysoinnit tehtävän jonkun muuttujan (tai muuttujien) eri luokissa erikseen tai tietyn ehdon täyttäville tilastoyksiköille.

Ryhmittäinen tarkastelu määritellään valikosta

Data ▸ **Split File...** vaihtoehto Compare groups ja valitsemalla muuttujaluettelosta ryhmittelymuuttuja; ryhmittelyn purkaminen vaihtoehto Analyze all cases, do not create groups.

Tämän määrityksen jälkeen tehtävät analyysit tapahtuvat erikseen kaikissa ehtomuuttujan ryhmissä (myös puuttuvien tietojen ryhmässä) erikseen. Jos ehtomuuttuja on kvantitatiivinen, se on ensin luokiteltava halutulla tavalla. Tämä määritys järjestää havaintomatriisin uudelleen ryhmittelymuuttujan mukaan. Tästä saattaa olla haittaa, jos aineistoon ei ole talletettu havaintoja identifioivaa muuttujaa.

Vain tiettyjen tilastoyksiköiden mukaan ottaminen analysointeihin määritellään valikosta

Data ▸ **Select Cases...** valitsemalla If condition is satisfied ja määrittelemällä sopivan if-ehdon. Ehdon purku All cases.

4 Muuttujien jakaumat ja tunnusluvut

4.1 Jakaumat

Kun havaintomatriisi on kunnossa, voidaan aineiston analysointi aloittaa. Ensin muodostetaan muuttujien frekvenssijakaumat (suorat jakaumat) joko graafisesti tai taulukkona. Muuttujien jakaumista voidaan huomata mahdollisesti tehtyjä tallennusvirheitä.

Frekvenssijakauman graafinen esitys valitaan muuttujan mitta-asteikon perusteella. Frekvenssihistogrammeja käytetään vähintään intervalliasteikollisen muuttujan jakauman esittämiseen, pylväitä tai janoja yleensä järjestysasteikolliselle muuttujalle ja piirakkakuvioita luokitteluasteikollisen muuttujan tapauksessa.

Graafiset esitykset löytyvät valikosta

Graphs ▸ Legacy Dialogs ▸

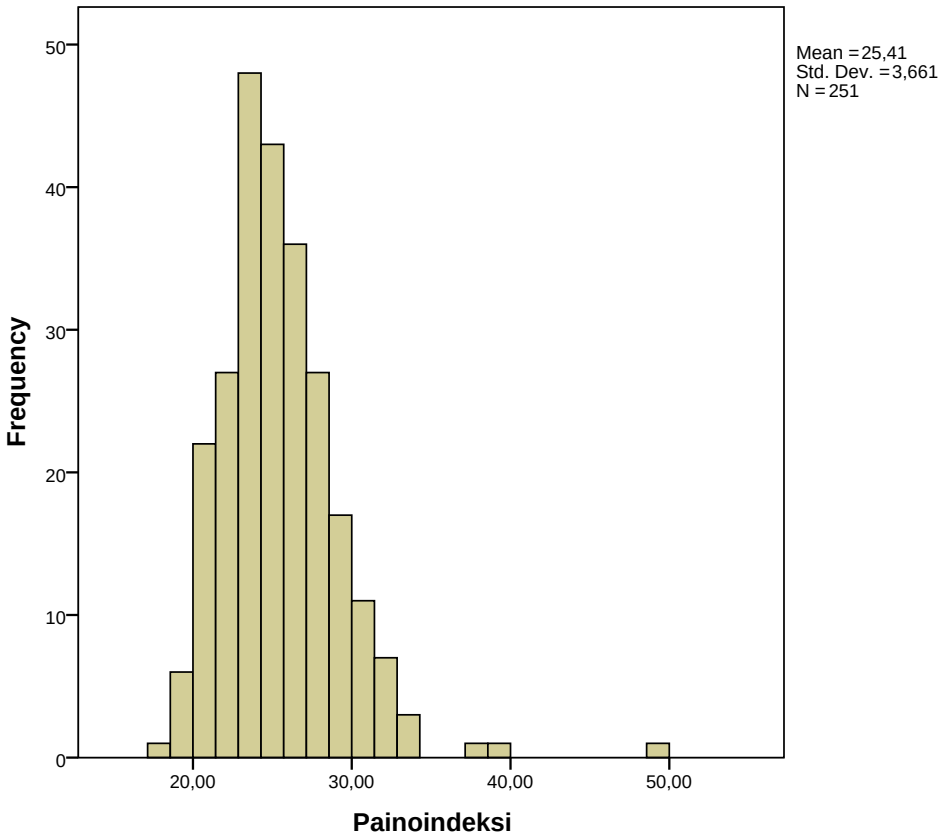
Bar... pylväsdiagrammit

Pie... piirakat

Histogram... frekvenssihistogrammit

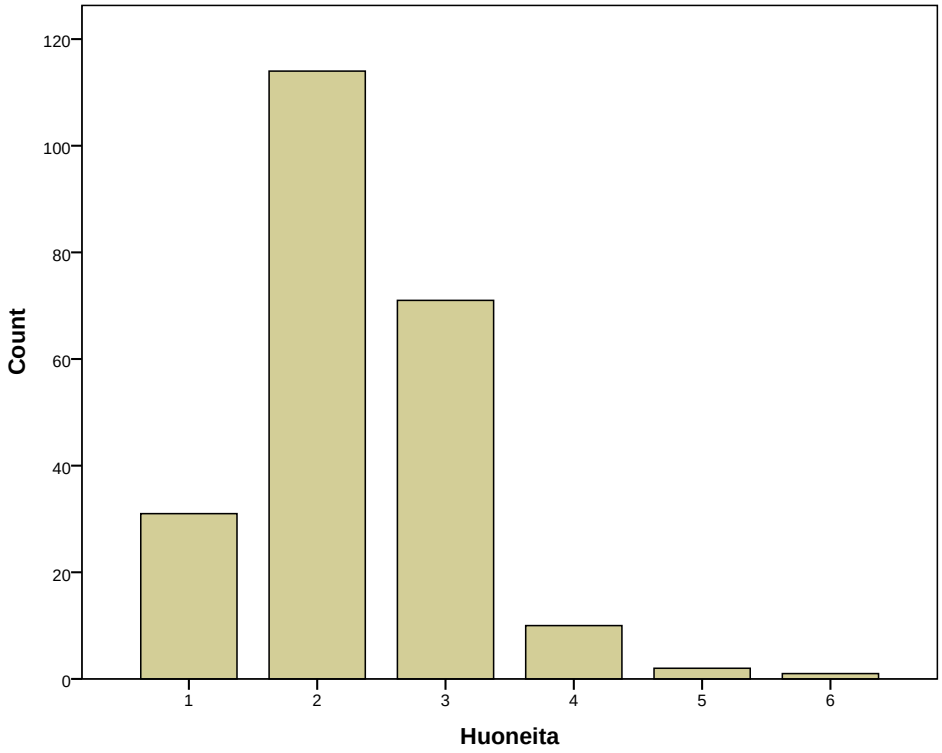
Esityksen valinnan jälkeen annetaan muuttuja, jolle graafinen esitys tehdään.

Esimerkki 2. Miesten painoindeksin (esimerkki 1) frekvenssihistogrammi.



Huomataan, että painoindeksin jakauma on oikealle vino. Lisäksi nähdään, että muutamalla miehellä on hyvin korkea painoindeksin arvo. ■

Esimerkki 3. Huoneiden lukumäärän jakauma aineistosta Asunnot_2006.



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

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

Frekvenssijakauman saa taulukkona valikosta

Analyze ► Descriptive Statistics ► Frequencies...

Esimerkki 4. Luokitellaan asunnot huoneiden lukumäärän perusteella yksiöihin, kaksioihin sekä kaksioita suurempiin. Muodostetaan uusi muuttuja, joka saa arvot (1, 2 ja 3) huoneiden lukumäärän perusteella. Muodostetaan tämän uuden muuttujan frekvenssijakauma. Jos uudelleen koodauksen yhteydessä on annettu koodeille selitteet *Yksiö*, *Kaksio*, *Kaksiota suurempi* sekä uudelle muuttujalle selite *Huoneisto*, saadaan taulukko

Huoneisto

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yksiö	31	13,5	13,5	13,5
	Kaksio	114	49,8	49,8	63,3
	Kaksiota suurempi	84	36,7	36,7	100,0
	Total	229	100,0	100,0	

missä on huoneiden lukumäärän mukaisesti asuntojen lukumäärät (Frequency) ja prosentuaaliset määrät (Valid Percent) sekä kumulatiiviset prosentit (Cumulative Percent). Siis lähes puolet myynnissä olleista asunnoista oli kaksioita. ■

Frekvenssitaulukkoa tehtäessä ohjelmisto luokittelee muuttujan jokaisen arvon omaan luokkaansa riippumatta siitä, montako arvoa muuttujalla on. Tämän vuoksi kvantitatiivisten muuttujien yhteydessä taulukko on useimmiten käyttökelpoinen vasta, kun muuttuja on ensin luokiteltu.

Esimerkki 5. Tarkastellaan miesten painoindeksejä. Luokitellaan painoindeksi luokkiin alle 25, 25–30, 30–35, yli 35. Kun uudelleen koodauksen yhteydessä on annettu selitteet koodeille *Normaalipainoinen*, *Lievä ylipaino*, *Merkittävä ylipaino*, *Sairaaloinen lihavuus* sekä uudelle muuttujalle selite *Lihavuus*, saadaan seuraava taulukko

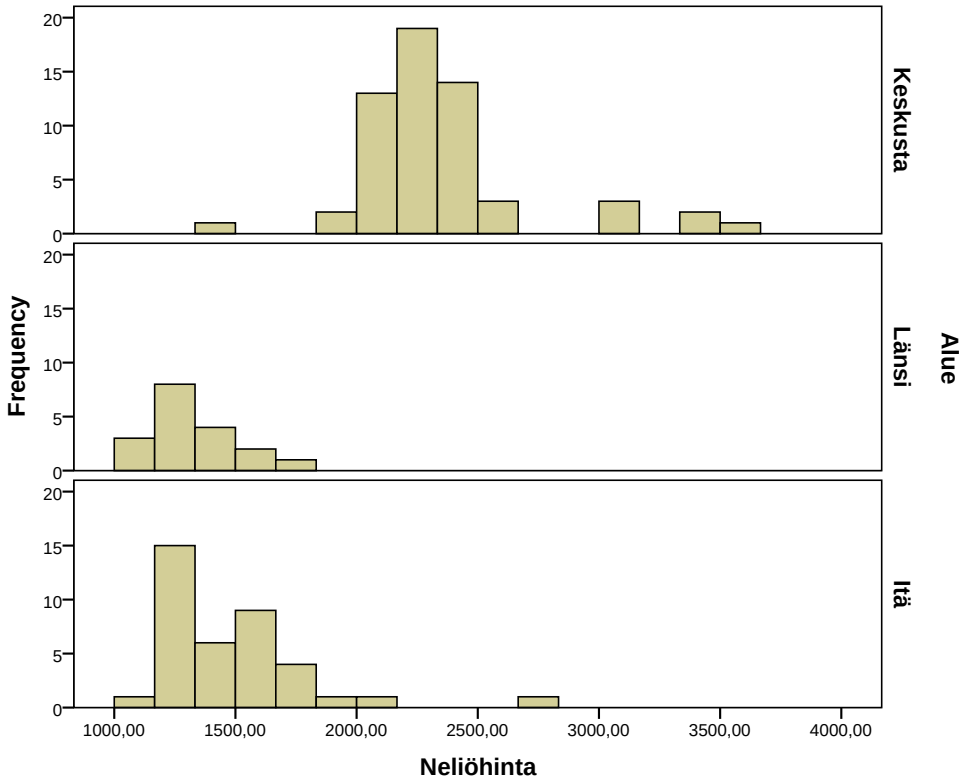
Lihavuus

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normaalipainoinen	125	49,6	49,8	49,8
	Lievä ylipaino	102	40,5	40,6	90,4
	Merkittävä lihavuus	21	8,3	8,4	98,8
	Sairaaloinen lihavuus	3	1,2	1,2	100,0
	Total	251	99,6	100,0	
Missing	System	1	,4		
Total		252	100,0		

missä on luokitellun painoindeksin mukaan miesten lukumäärät (Frequency) ja prosentuaaliset määrät (Valid Percent) sekä kumulatiiviset prosentit (Cumulative Percent). Huomataan, että aineistossa normaalipainoisia on vain puolet. ■

Jos aineistossa on puuttuvia tietoja, niin niiden lukumäärä näkyy frekvenssijakaumassa. Kun käyttää uudelleen koodausta, niin on syytä tarkistaa, että havaintoja ja puuttuvia tietoja on saman verran kuin alkuperäisessäkin muuttujassa.

Esimerkki 6. Esimerkin 4 jakaumasta huomattiin, että lähes puolet huoneistoista oli kaksioita. Tutkitaan nyt kaksioiden neliöhintoja. Lasketaan aluksi neliöhinta Transform ▸ Compute ▸ Target Variable on *Neliöhinta* ja Numeric Expression *HINTA*/*NELIOT*. Kun halutaan analysoida vain kaksioita, niin määritellään ennen analyysien tekemistä Data-valikossa Select Cases… if-ehto *Huoneisto* = 2. Verrataan kaksioiden neliöhinnan jakaumia alueittain. Annetaan frekvenssihistogrammin teon yhteydessä ryhmittelymuuttujaksi *Alue* (Panel by Rows: *Alue*). Saadaan samaan kuvaan kolme frekvenssihistogrammia



Huomataan selvät erot jakaumien sijainneissa sekä vaihteluväleissä. ■

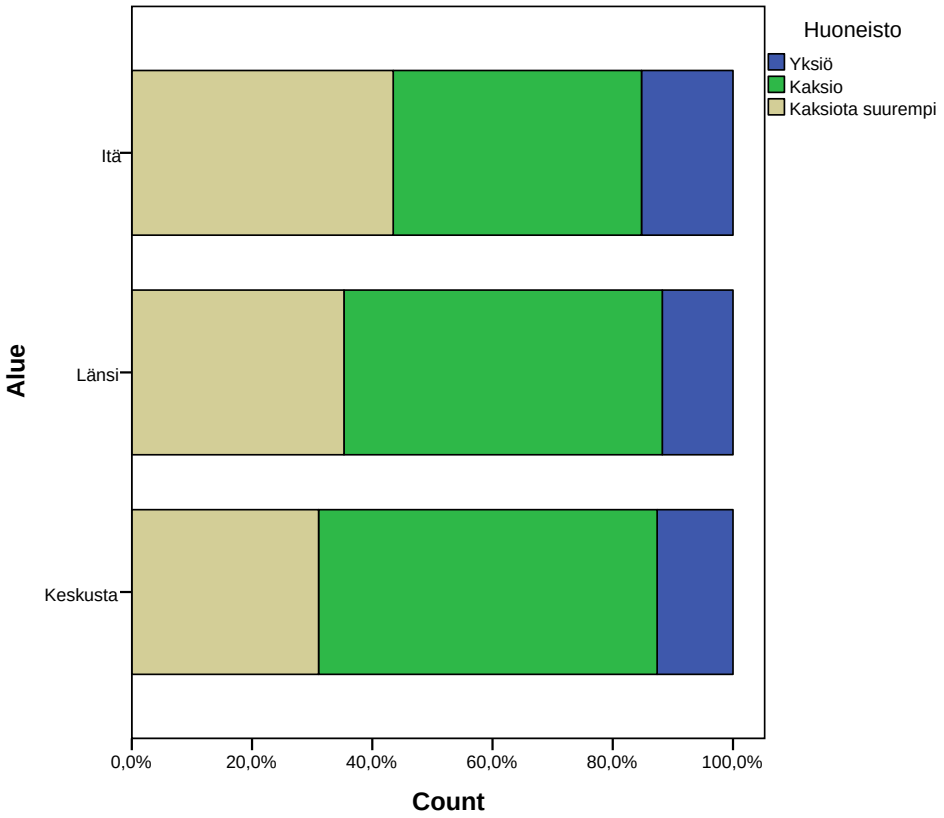
Esimerkki 7. Esimerkin 4 *Huoneisto*-muuttujan jakauma voidaan tehdä alueittain. Tällöin Split file ryhmittelymuuttujaksi annetaan *Alue*, sitten muodostetaan frekvenssijakauman muuttujasta *Huoneisto*, saadaan tulos

Huoneisto

Alue			Frequency	Percent	Valid Percent	Cumulative Percent
Keskusta	Valid	Yksiö	13	12,6	12,6	12,6
		Kaksio	58	56,3	56,3	68,9
		Kaksiota suurempi	32	31,1	31,1	100,0
		Total	103	100,0	100,0	
Länsi	Valid	Yksiö	4	11,8	11,8	11,8
		Kaksio	18	52,9	52,9	64,7
		Kaksiota suurempi	12	35,3	35,3	100,0
		Total	34	100,0	100,0	
Itä	Valid	Yksiö	14	15,2	15,2	15,2
		Kaksio	38	41,3	41,3	56,5
		Kaksiota suurempi	40	43,5	43,5	100,0
		Total	92	100,0	100,0	

Paremmin jakaumia voidaan kuitenkin verrata ristiintaulukoimalla muuttujat, (ks. esimerkki 12) tai käyttämällä tilanteeseen sopivaa graafista esitystä (ks. esimerkki 8). ■

Esimerkki 8. Tarkastellaan graafisesti esimerkin 4 *Huoneisto*-muuttujan jakaumia alueittain. Piirretään pylväsdiagrammit Graphs ► Legacy Dialogs ► Bar... Stacked, Category Axis: *Alue*, Define Stacks by: *Huoneisto*, käännetään ne vaakatasoon ja pyydetään prosenttijakaumat alueittain. Saadaan kuvaaja



Huomataan, että jakaumissa on jonkin verran alueellisia eroja. ■

4.2 Tunnuslukuja

Tunnusluvun avulla pyritään kuvaamaan muuttujan jakaumaa muuttujan arvoista lasketulla luvulla. Kuvataan esimerkiksi jakauman sijaintia sopivan keskiluvun avulla tai muuttujan arvojen vaihtelua hajontaluvun avulla. Muuttujan mitta-asteikko määrittää sen, mitä tunnuslukuja jakauman kuvaamiseen voidaan käyttää.

Keskilukuja ovat moodi, mediaani ja keskiarvo. Moodi on se muuttujan arvo tai luokka, joka esiintyy useimmin tai jossa on eniten havaintoja. Mediaani on sellainen muuttujan arvo, jota pienempiä ja suurempia arvoja on yhtä paljon. Mediaania voidaan käyttää, kun järjestyksellä on tulkinta eli muuttuja on vähintään järjestysasteikollinen. Aritmeettinen keskiarvo on sallittu kvantitatiivisten muuttujien yhteydessä.

Muuttujan arvot vaihtelevat tilastoyksiköstä toiseen. Vaihtelun voimakkuutta pyritään mittaamaan erilaisia tunnuslukuja käyttäen. Kvantitatiivisten muuttujien yhteydessä vaihtelua mitataan varianssin avulla. Varianssi mittaa kuinka tiiviisti muuttujien arvot ovat keskittyneet keskiarvon ympärille. Varianssin neliöjuuri on nimeltään keskihajonta, joka useimmiten ilmoitetaan vaihtelun mittarina.

Ala- ja yläkvartiili 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 % havaituista arvoista on pienempiä kuin yläkvartiili. Alakvartiili, mediaani ja yläkvartiili jakavat muuttujan arvot neljään havaintomääriltään yhtä suuriin osiin. Yhdessä näitä tunnuslukuja kutsutaan kvartiileiksi. Muuttujan arvot voidaan jakaa

viiteen, kuuteen, jne. havaintomääriltään yhtä suuriin osiin. Yleisesti näitä osiin jakavia tunnuslukuja kutsutaan fraktiileiksi.

Tunnuslukuja voidaan tarkastella ehdollisina. Ehdollisia keskiarvoja (tai mediaaneja) voidaan käyttää tutkittaessa riippuvuutta kahden muuttujan välillä. Ehdollisten keskiarvojen käyttö riippuvuuden tutkimisessa edellyttää tietysti sitä, että selitettävä muuttuja on kvantitatiivinen.

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 halutuista muuttujista mm. keskiarvo, keskihajonta, varianssi, pienin arvo, suurin arvo, vaihteluväli sekä tunnusluvut ehdollisina antamalla kohdassa Factor List ryhmittelymuuttuja

Compare Means ▸

Means... saadaan tunnusluvut ehdollisina antamalla ehtomuuttuja kohdassa Independent List

Esimerkki 9. Painoindeksin sekä huoneiden lukumäärän tunnuslukuja (pyydetty keskiarvo, mediaani, keskihajonta, pienin ja suurin arvo, alakvartiili, mediaani, yläkvartiili)

Statistics

Painoindeksi		
N	Valid	251
	Missing	1
Mean		25,4086
Median		25,1045
Std. Deviation		3,66131
Minimum		18,03
Maximum		48,96
Percentiles	25	23,0564
	50	25,1045
	75	27,3526

Statistics

Huoneita		
N	Valid	229
	Missing	0
Mean		2,31
Median		2,00
Std. Deviation		,823
Minimum		1
Maximum		6
Percentiles	25	2,00
	50	2,00
	75	3,00

■

Histogrammin teon yhteydessä saa muuttujan keskiarvon ja keskihajonnan automaattisesti (ks. esimerkki 2).

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

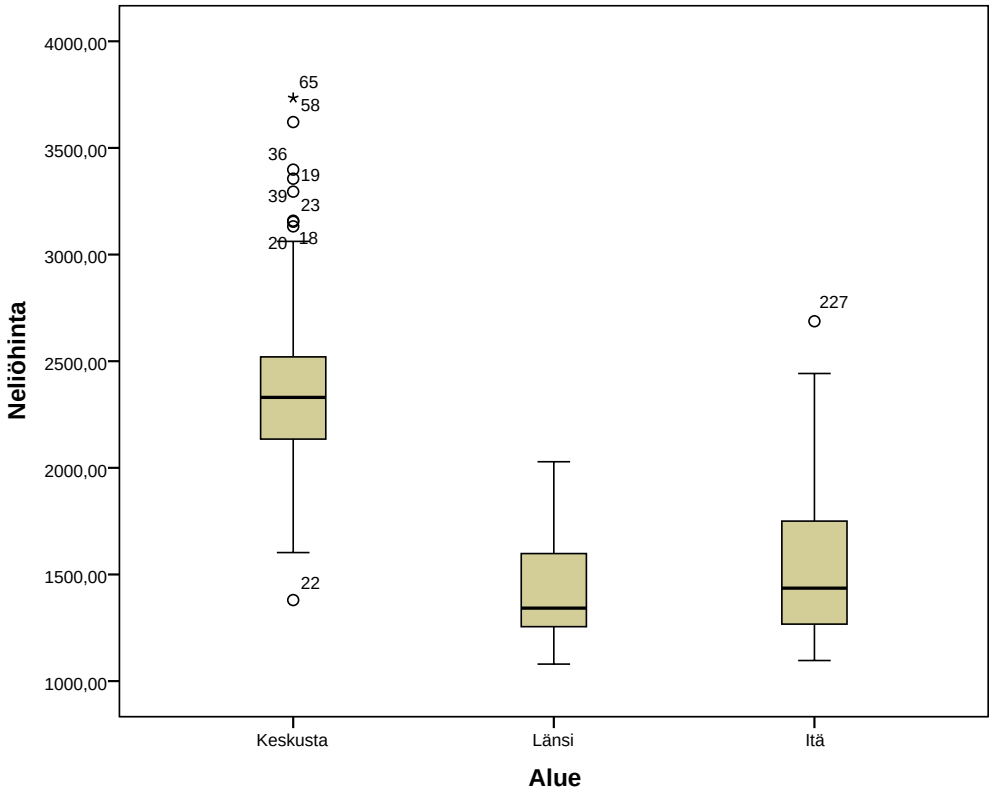
Graphs ▸ **Legacy Dialogs** ▸ **Boxplot...** antamalla Variable-kohtaan tutkittava muuttuja ja Category-kohtaan ryhmittelymuuttuja.

Esimerkki 10. Vaikuttaako sijainti asunnon neliöhintaan? Asunnon neliöhinta (laskettu esimerkissä 6) on selitettävä eli riippuva muuttuja (y) ja sijainti (*Alue*) selittävä eli riippumaton muuttuja (x). Pyritään selvittämään *Neliöhinta*-muuttujan arvojen vaihtelua sillä, millä alueella huoneisto sijaitsee. Eräs mahdollisuus riippuvuuden selvittämisessä on keskiarvojen vertailu ryhmittäin, ehdollisten keskiarvojen käyttö. Lasketaan *Neliöhinta*-muuttujasta keskiarvot alueittain sekä vertaillaan keskiarvoeroja. Jos ehdolliset keskiarvot poikkeavat toisistaan sanotaan, että *Alue*-muuttujalla voidaan selittää *Neliöhinta*-muuttujan vaihtelua. Sanotaan, että *Neliöhinta*-muuttuja riippuu *Alue*-muuttujasta. Jos ehdolliset keskiarvot ovat lähes samoja, niin riippuvuutta ei ole. Ehdolliset keskiarvot lasketaan valikosta Compare Means ▸ Means… antamalla Dependent List -muuttujaksi *Neliöhinta* ja Independent List -muuttujaksi *Alue*. Näin saadaan tulos

Neliöhinta

Alue	Mean	N	Std. Deviation
Keskusta	2397,6072	103	408,02462
Länsi	1414,2870	34	260,39544
Itä	1536,1328	92	341,69439
Total	1905,5176	229	575,61088

jossa on ehdolliset keskiarvot (Mean) ja keskihajonnat (Std.Deviation). Näyttäisi siis siltä, että keskustan neliöhinnat ovat keskimäärin korkeampia kuin keskustan ulkopuolella (ks. testaus 6.2). Neliöhinnan jakaumissa esiintyvä vaihtelu on myös jonkin verran erilaista, keskustassa keskihajonta on suurin. Tämä näkyy hyvin myös laatikko-jana -kuviosta



joka on tehty valikosta Graphs ▸ Boxplot… antamalla Variable-kohtaan *Neliöhinta* ja Category-kohtaan *Alue*. Laatikko-jana -kuviossa keskimmäinen viiva on

neliöhinnan mediaanin kohdalla ja laatikon ylä- ja alareunat ylä- ja alakvartiileissa. Ylimmät ja alimmat viivat ovat suurimpien ja pieneimpien arvon kohdalla, ellei ryhmässä ole poikkeavia arvoja. Kuviosta nähdään, että keskustassa neliöhinnan jakauma on ylempänä kuin muilla alueilla ja siinä on myös enemmän vaihtelua. Keskustassa on myös muutamia kovin kalliita asuntoja. ■

5 Pisteparvi ja korrelaatiokerroin

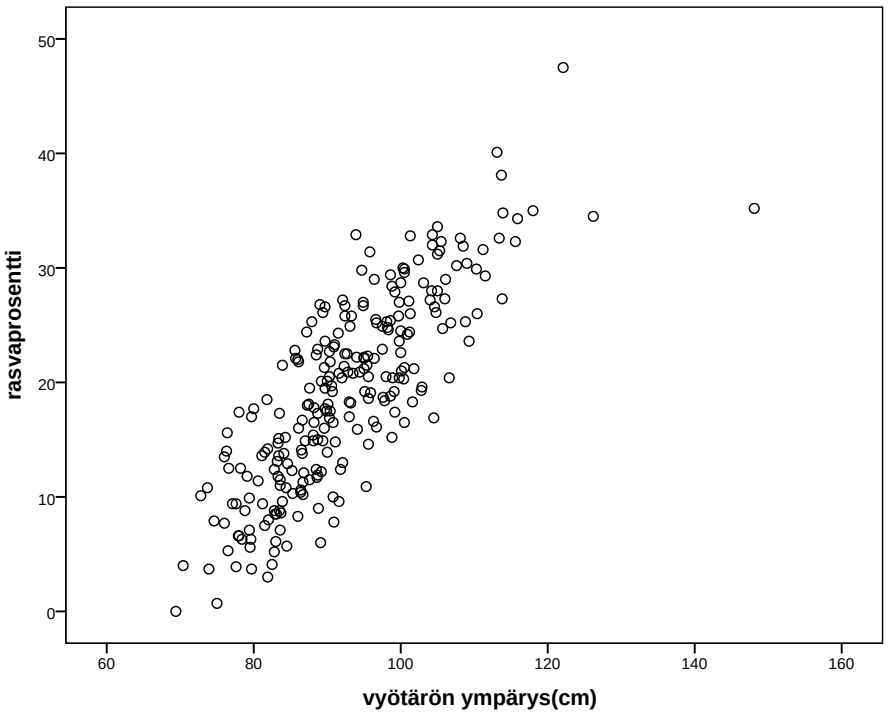
Kun tarkastellaan kahta muuttujaa samanaikaisesti, niin on kyse kaksiulotteisesta jakaumasta. Tällöin ollaan kiinnostuneita muuttujien välisestä riippuvuudesta. Kaksiulotteisen jakauman graafinen esitystapa on pisteparvi eli korrelaatiodiagrammi, joka antaa hyvän yleiskuvan mahdollisesta riippuvuudesta ja sen laadusta. Pisteparvi on järkevä piirtää, kun selitettävä muuttuja on kvantitatiivinen. Tulkinnassa on huomattava selittävän muuttujan mitta-asteikko. Pisteparvi saadaan valikosta

Graphs ▸ Legacy Dialogs ▸ Scatter/Dot... antamalla selitettävä y pystyakselille ja selittävä x vaaka-akselille.

Myös kaksiulotteisista jakaumista voidaan määritellä tunnuslukuja, jotka nyt mittaavat riippuvuuden voimakkuutta. Yksi tällainen tunnusluku on korrelaatiokerroin (r), joka mittaa kahden kvantitatiivisen muuttujan välisen suoranomaisen eli lineaarisen riippuvuuden voimakkuutta. Lineaarinen riippuvuus voi olla positiivista tai negatiivista, korrelaatiokerroin vastaavasti joko positiivinen tai negatiivinen. Korrelaatiokertoimen ollessa lähellä nollaa lineaarista riippuvuutta ei ole. Täydellinen lineaarinen riippuvuus on silloin, kun korrelaatiokerroin on itseisarvoltaan yksi. Korrelaatiokertoimen (korrelaatiomatriisin) voi laskea valikosta

Analyze ▸ Correlate ▸ Bivariate... antamalla halutut muuttujat.

Esimerkki 11. Rasvaprosentin ja vyötärön ympärysmitan riippuvuus.



Muuttujien välillä huomataan olevan voimakas suoranomainen riippuvuus. Siis rasvaprosentti riippuu lineaarisesti vyötärön ympärysmitasta (ks. tarkemmin 6.4).

Korrelaatiomatriisiksi saadaan

Correlations

		rasvaprosentti	vyötärön ympärys(cm)
rasvaprosentti	Pearson Correlation	1	,813**
	Sig. (2-tailed)		,000
	N	252	252
vyötärön ympärys(cm)	Pearson Correlation	,813**	1
	Sig. (2-tailed)	,000	
	N	252	252

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

jossa korrelaatiokerroin on 0,813 kertoen voimakkaasta positiivisesta lineaarisesta riippuvuudesta (ks. testaus 6.4). ■

6 Joitain yleisesti käytettyjä analysointimenetelmiä

Tilastollinen hypoteesi on väittämä populaatiosta, sen jakaumasta tai jakauman parametrista. Hypoteesin testaus tarkoittaa väittämän tutkimista otoksen perusteella. Väitteen paikkansa pitävyyttä tutkitaan otoksen (käytettävissä olevan aineiston) perusteella laskemalla tilanteeseen sopiva testisuure. 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 väitettä ei uskota. Mikä sitten on harvinaista? Testauksessa harvinaisiksi arvoiksi katsotaan sellaisten arvojen joukko, jonka todennäköisyys on melko pieni. Tämän todennäköisyyden kiinnittäminen määrittää testaukseen liittyvän riskitason. Testauksessa on tapana ilmoittaa nk. p-arvo, joka kertoo todennäköisyyden saada väitteen ollessa tosi otoksesta saatua arvoa harvinaisempi arvo. Tämä on pienin ristitaso, jolla asetettu väite voidaan hylätä. Jos siis testaukseen liittyvä p-arvo on pieni, sanotaan vaikka 0,01, 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, sen halutaan usein olevan suuruusluokkaa pienempi kuin 5 %, 2,5 %, 1 %.

Jos p-arvo on pienempi kuin 0,05, on tapana sanoa, että tulos on tilastollisesti melkein merkitsevä. Jos p-arvo on pienempi kuin 0,01, sanotaan tuloksen olevan tilastollisesti merkitsevän ja p-arvon ollessa pienempi kuin 0,001 tilastollisesti erittäin merkitsevän.

Hypoteesin testauksessa asetetaankin siis kaksi väittämää, joista toinen on välttämättä voimassa. Asetetaan nollahypoteesi H_0, jonka ollessa tosi, testisuuren

todennäköisyysjakauma tunnetaan, sekä vaihtoehtoinen hypoteesi H_1. Nollahypoteesi H_0 tulee aina asettaa käytetyn testin sanelemalla tavalla.

Seuraavaksi esiteltävissä menetelmissä pyritään selittämään yhtä muuttuja. Selittäviä muuttujia on yksi tai useampia. Analysointimenetelmän valintaan vaikuttaa muuttujien mittaustaso. Tässä esityksessä käydään läpi kolme perusanalyysiä.

6.1 Ristiintaulukko ja riippumattomuustesti

Kahden kvalitatiivisen muuttujan välinen riippuvuustarkastelu voidaan tehdä ristiintaulukon avulla vertailemalla selitettävän muuttujan prosenttijakaumia selittäjän luokissa. Riippuvuuden merkitsevyys voidaan testata. Testisuureena käytetään χ^2-riippumattomuustestisuuretta ja hypoteesit asetetaan

H_0: ei riippuvuutta

H_1: on riippuvuutta.

Testin käyttöön liittyy joitain oletuksia (ei kuitenkaan mitta-asteikkovaatimuksia). Tilanteissa, jossa ristiintaulukointi on tehty siten, että molemmilla muuttujilla on kaksi luokkaa (nelikenttä), testiä voidaan käyttää, jos $n > 40$. Jos nelikentässä $20 \leq n \leq 40$, niin nk. teoreettiset frekvenssit (frekvenssit, jos riippuvuutta ristiintaulukon perusteella ei olisi) eivät saa olla alle viiden. Muulloin kaikkien teoreettisten frekvenssien oltava suurempia kuin yksi sekä enintään 20 % saa olla alle viiden. Jos vaatimukset eivät täyty, muutetaan muuttujien luokituksia.

χ^2-testisuureen arvot ovat ei-negatiivisia, joten harvinaisten arvojen joukko muodostuu suurista arvoista.

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. Kvantitatiivinen muuttuja on siis ensin luokiteltava (Transform ▸ Recode into Different Variables ...).

Esimerkki 12. Tarkastellaan esimerkin 7 tilannetta. Halutaan selvittää, onko alueella vaikutusta huoneiston tyyppiin.

Nyt asetetaan

H_0: huoneiston tyypin ja alueen välillä ei riippuvuutta

H_1: huoneiston tyypin ja alueen välillä on riippuvuutta.

Suoritetaan edellä esitetyllä tavalla näiden muuttujien ristiintaulukointi. Saadaan ristiintaulukko ja testaukseen liittyvät tulokset

Huoneisto * Alue Crosstabulation

			Alue			
			Keskusta	Länsi	Itä	Total
Huoneisto	Yksiö	Count	13	4	14	31
		% within Alue	12,6%	11,8%	15,2%	13,5%
	Kaksio	Count	58	18	38	114
		% within Alue	56,3%	52,9%	41,3%	49,8%
	Kaksiota suurempi	Count	32	12	40	84
		% within Alue	31,1%	35,3%	43,5%	36,7%
Total		Count	103	34	92	229
		% within Alue	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	4,674[a]	4	,322
Likelihood Ratio	4,700	4	,319
Linear-by-Linear Association	1,040	1	,308
N of Valid Cases	229		

a. 1 cells (11,1%) have expected count less than 5. The minimum expected count is 4,60.

Aluksi huomataan, että *Huoneisto*-muuttujan prosentuaaliset jakaumat alueittain poikkeavat jonkin verran 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 4,60, alle viiden teoreettisia frekvenssejä 11,1 %). χ^2-riippumattomuustestisuureen arvo (Pearson Chi-Square) on 4,674 ja p-arvo on 0,322, joten H_0 hyväksytään. Tehdään johtopäätelmä, että muuttujien välillä ei ole riippuvuutta. ■

Esimerkki 13. Tarkastellaan ARVIO-aineistoa, jossa on eräältä kurssilta saatua opiskelijapalautetta. Halutaan selvittää, onko opintosuunnalla vaikutusta annettuun palautteeseen. Aineistossa on muuttuja *Opintojakson työläys*, joka mittaa kurssin työläyttä (muodostettu *kurssi*-muuttujasta yhdistämällä reunimmaiset luokat), sekä palautteen antajan opintosuunta (*Opiskelijan opintosuunta*). Nyt asetetaan

H_0: opintosuunnan ja työläyden välillä ei riippuvuutta

H_1: opintosuunnan ja työläyden välillä on riippuvuutta.

Saadaan ristiintaulukko ja testaukseen liittyvät tulokset

Opintojakson työläys * Opiskelijan opintosuunta Crosstabulation

			Opiskelijan opintosuunta		
			hallinto	taloust	Total
Opintojakson työläys	työläs	Count	13	16	29
		% within Opiskelijan opintosuunta	68,4%	34,8%	44,6%
	sopiva	Count	5	15	20
		% within Opiskelijan opintosuunta	26,3%	32,6%	30,8%
	vähätöinen	Count	1	15	16
		% within Opiskelijan opintosuunta	5,3%	32,6%	24,6%
Total		Count	19	46	65
		% within Opiskelijan opintosuunta	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymptotic Significance (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.

Huomataan, että *Opintojakson työläys* -muuttujan prosentuaaliset jakaumat opintosuunnittain poikkeavat paljonkin 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 on 4,68, alle viiden teoreettisia frekvenssejä on 16,7 %). χ^2-riippumattomuustestisuureen arvo (Pearson Chi-Square) on 7,668 ja p-arvo on 0,022. Jos päättely tehdään 5 %:n riskitasolla, niin H_0 hylätään ja päätellään riippuvuutta olevan. Jos valitaan pienempi riskitaso kuin 2,2 %, niin päätellään, että riippuvuutta ei ole. ■

Tilastollisten testin suorittaminen tapahtuu periaatteessa kaikissa tilanteissa edellä esitetyllä tavalla. Asetetaan testattava hypoteesi, lasketaan testisuureen arvo ja pienin riskitaso, jolla nollahypoteesi voidaan hylätä. Tämän p-arvon perusteella joko hyväksytään väittämä tai hylätään se. Eri tilanteissa nollahypoteesi, testisuure ja sen jakauma ovat erilaisia.

6.2 Odotusarvojen yhtäsuuruuden testaaminen *t*-testillä

Tutkittaessa kvantitatiivisen muuttujan riippuvuutta kvalitatiivisesta muuttujasta, jolla on kaksi luokkaa, voidaan käyttää riippumattomien otosten t-testiä kahden populaation odotusarvojen yhtäsuuruuden testaamiseksi. Tutkitaan siis populaatioiden keskiarvoja, joita voidaan arvioida otoskeskiarvojen avulla.

Hypoteesit asetetaan

H_0: populaatioiden odotusarvot ovat samoja

H_1: populaatioiden odotusarvot eivät ole yhtä suuria.

Vaihtoehtoinen hypoteesi voidaan asettaa myös yksisuuntaisena, jolloin H_1: toisen populaation odotusarvo on toista suurempi. Riippumattomien otosten t-testissä oletetaan, että käytössä on riippumattomat satunnaisotokset normaalijakaumista, joiden varianssit ovat yhtä suuret, mutta tuntemattomat. Testisuure, jota käytetään, noudattaa nollahypoteesin ollessa tosi nk. *Studentin t*-jakaumaa, joka määritellään nk. vapausastein. Tämä jakauma on symmetrinen origon suhteen. Siis harvinaisten arvojen joukko muodostuu kaksisuuntaisessa testissä itseisarvoltaan suurista arvoista.

Riippumattomien otosten t-testi saadaan valikosta

Analyze ▸ Compare Means ▸ Independent-Samples T Test... annetaan selitettävä (Test Variable) sekä selittävä, ryhmittely -muuttuja (Grouping Variable).

Tuloksena saadaan testisuureen lisäksi myös ryhmäkeskiarvot ja -varianssit sekä testisuure varianssien yhtäsuuruuden testaamiseksi.

Esimerkki 14. Onko keskustassa ja keskustan ulkopuolella olevien asuntojen keskimääräisissä neliöhinnoissa eroja?

Asetetaan

H_0: neliöhinnan odotusarvot samoja molemmissa populaatioissa

H_1: neliöhinnan odotusarvot eivät samoja molemmissa populaatioissa.

Tehdään tarkastelu esimerkin 10 aineistosta. Suoritetaan riippumattomien otosten t-testi edellä esitetyllä tavalla ja saadaan tulokset

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)
Neliöhinta	Equal variances assumed	1,235	,268	-18,455	227	,000
	Equal variances not assumed			-18,045	193,029	,000

Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		Lower	Upper
-894,35342	48,46101	-989,84436	-798,86248
-894,35342	49,56214	-992,10630	-796,60054

Tässä on siis riippumattomat otokset keskustan ulkopuolelta ja keskustasta. Otoskoot ovat 126 ja 103. Neliöhinnan keskiarvojen erotus on −894,35. Neliöhinnan

otosvarianssit ($325{,}34^2$ ja $408{,}02^2$) poikkeavat toisistaan. Nyt tuloksista löytyy testisuure (Levene's Test for Equality of Variances) hypoteesille H_0: Populaation varianssit voidaan olettaa samoiksi. Koska tähän liittyvä p-arvo on $0{,}268 > 0{,}05$, H_0 hyväksytään ja todetaan, että vaatimus varianssien yhtäsuuruudesta voidaan kuitenkin olettaa olevan täytetty. Jos näin ei olisi, niin t-testin tulokset luettaisiin toiselta riviltä. Normaalijakaumaoletus jätetään tässä testaamatta. Populaatioiden odotusarvojen yhtäsuuruuden testaamiseen liittyvä t-testisuureen arvo on $-18{,}455$ ja kaksisuuntaiseen testiin liittyvä p-arvo $< 0{,}001$. Jos riskitasoksi valitaan 0,001 %, niin nollahypoteesi hylätään ja tehdään päätelmä, että neliöhinnat ovat keskimäärin erisuuruiset.

Tulostuksesta löytyy myös 95 % luottamusväli odotusarvojen erotukselle. Testin sijaan voidaan käyttää tätä luottamusväliä johtopäätelmän tekemisessä. Jos luottamusväli sisältää nollan niin populaation odotusarvojen erotus voidaan arvioida olevan nolla (eli keskustan ulkopuolella ja keskustassa neliöhintojen odotusarvot samoja). Tässä luottamusväli, jolle populaatioiden odotusarvojen erotuksen arvellaan kuuluvan, on $(-989{,}84, -798{,}86)$. ■

Esimerkki 15. Onko keskustassa ja keskustan ulkopuolella olevien kaksioiden keskimääräisissä neliömäärissä eroja?

Asetetaan

H_0: neliömäärien odotusarvot samoja molemmissa populaatioissa

H_1: neliömäärien odotusarvot eivät samoja molemmissa populaatioissa.

Tarkastellaan edellisen esimerkin aineistoa. Valiteen analysoitavaksi vain kaksiot (Data ► Select Cases ...) ja suoritetaan riippumattomien otosten t-testi. Saadaan tulokset

Group Statistics

	Onko keskustassa?	N	Mean	Std. Deviation	Std. Error Mean
Neliöt	Ei ole	56	55,05	6,132	,819
	On	58	53,39	6,326	,831

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Neliöt	Equal variances assumed	,026	,873	1,428	112	,156
	Equal variances not assumed			1,429	111,998	,156

Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		Lower	Upper
1,667	1,168	-,646	3,981
1,667	1,167	-,645	3,979

Populaation varianssit voidaan olettaa olevan samoja, koska varianssien yhtäsuuruuden testaamiseen liittyvä p-arvo on 0,873 > 0,05. Asetettu H_0 hyväksytään 5 %:n riskitasolla, koska t-testiin liittyvä p-arvo on 0,156. Jos testaus tehtäisiin yksisuuntaisena, niin p-arvo olisi $0{,}156/2 = 0{,}078$.

Jos halutaan tehdä päättely luottamusvälin avulla, niin todetaan nollan kuuluvan odotusarvojen erotuksen 95 %:n luottamusvälille $(-0{,}646, 3{,}981)$ ja päätellään odotusarvojen olevan yhtä suuret. ■

6.3 Varianssianalyysi

Tutkittaessa kvantitatiivisen muuttujan riippuvuutta kvalitatiivisesta muuttujasta, jolla on useampi kuin kaksi luokkaa, voidaan käyttää yksisuuntaista varianssianalyysiä populaatioiden odotusarvojen yhtäsuuruuden testaamiseksi. Tämä on siis yleistys edellä esitetylle riippumattomien otoksien t-testille. Nytkin testattavana hypoteesina on

H_0: populaation odotusarvot ovat samoja

H_1: kaikki odotusarvot eivät ole samoja.

Testin käyttöön liittyy samat oletukset kuin t-testissäkin. On tehty riippumattomat otokset normaalijakaumista, joiden varianssit yhtä suuret mutta tuntemattomat. Jos otoksia on kaksi, voi tehdä joko t-testin tai suorittaa varianssianalyysin.

Varianssianalyysissä käytetään nk. F-testisuuretta odotusarvojen yhtäsuuruuden testaamiseksi. Tämän testisuureen arvot ovat ei-negatiivisia, joten harvinaisten arvojen joukko muodostuu suurista arvoista.

Varianssianalyysi suoritetaan valikosta

Analyze ▸ Compare Means ▸ One-Way ANOVA... annetaan selitettävä, riippuva (Dependent) muuttuja sekä selittävä (Factor) muuttuja.

Tuloksena saadaan testisuureen lisäksi pyydettäessä (Options...) myös ehdolliset keskiarvot ja varianssit sekä testisuure varianssien yhtäsuuruuden testaamiseksi. Jos saadaan tulos, että odotusarvot eivät kaikki ole yhtäsuuria, voidaan myös tehdä monivertailuja ryhmittäin (Post Hoc...). Nimitys yksisuuntainen varianssianalyysi tulee siitä, että on yksi selittäjä. Nimitys varianssianalyysi on hieman harhaanjohtava, koska analyysissä ei testata varianssien yhtäsuuruutta (paitsi oletusten tutkimisessa) vaan odotusarvojen yhtäsuuruutta.

Esimerkki 16. Halutaan tutkia, vaikuttaako alue keskimääräiseen neliöhintaan. Nyt

H_0: neliöhinnan odotusarvot ovat samoja kaikilla alueilla

H_1: neliöhinnan odotusarvot kaikki eivät yhtä suuria.

Tehdään varianssianalyysi edellä esitetyllä tavalla esimerkin 10 tilanteesta. Saadaan tulokset

Neliöhinta

	N	Mean	Std. Deviation
Keskusta	103	2397,6072	408,02462
Länsi	34	1414,2870	260,39544
Itä	92	1536,1328	341,69439
Total	229	1905,5176	575,61088

Test of Homogeneity of Variances

Neliöhinta

Levene Statistic	df1	df2	Sig.
1,929	2	226	,148

ANOVA

Neliöhinta

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	45699080,58	2	22849540,29	173,035	,000
Within Groups	29843678,05	226	132051,673		
Total	75542758,64	228			

Multiple Comparisons

Dependent Variable: Neliöhinta

Bonferroni

(I) Alue	(J) Alue	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Keskusta	Länsi	983,32022*	71,87439	,000	809,9641	1156,6764
	Itä	861,47438*	52,12868	,000	735,7435	987,2052
Länsi	Keskusta	-983,32022*	71,87439	,000	-1156,6764	-809,9641
	Itä	-121,84584	72,93296	,289	-297,7552	54,0635
Itä	Keskusta	-861,47438*	52,12868	,000	-987,2052	-735,7435
	Länsi	121,84584	72,93296	,289	-54,0635	297,7552

*. The mean difference is significant at the 0.05 level.

Ensin huomataan, että ehdolliset otoskeskiarvot näyttäisivät poikkeavan toisistaan ainakin verrattaessa keskustan keskiarvoja muihin. Kun testataan varianssien yhtäsuuruutta (kolmessa populaatiossa, tarkastellaan neliöhintaa kolmella alueella), voidaan olettaa niiden olevat yhtä suuret, koska $p = 0{,}148$. Testattaessa odotusarvojen yhtäsuuruutta saadaan $F = 173{,}035$ ja $p < 0{,}001$, joten H_0 hylätään ja tehdään johtopäätelmä, että kaikilla alueilla neliöhinnat eivät ole keskimäärin samoja. Missä sitten on eroja? Alueittain vertailu (Multiple Comparisons) kertoo, että eroja on keskustan ja muiden alueiden välillä ($p < 0{,}001$), mutta ei idän ja lännen välillä ($p = 0{,}289$). ■

Esimerkki 17. Halutaan tutkia, vaikuttaako huoneistotyyppi keskimääräiseen neliöhintaan. Koska esimerkissä 10 todettiin keskimääräisissä neliöhinnoissa olevan eroja keskustassa ja keskustan ulkopuolella, niin tehdään varianssinalyysi näissä ryhmissä erikseen (Data ► Split File ...).

Nyt

H_0: neliöhinnan odotusarvot ovat samoja huoneistotyypeittäin

H_1: neliöhinnan odotusarvot kaikki eivät yhtä suuria.

Saadaan tulokset

Neliöhinta

Onko keskustassa?		N	Mean
Ei ole	Yksiö	18	1876,4512
	Kaksio	56	1436,4447
	Kaksiota suurempi	52	1446,0183
	Total	126	1503,2538
On	Yksiö	13	2768,6696
	Kaksio	58	2339,6051
	Kaksiota suurempi	32	2351,9919
	Total	103	2397,6072

Test of Homogeneity of Variances

Neliöhinta

Onko keskustassa?	Levene Statistic	df1	df2	Sig.
Ei ole	,214	2	123	,808
On	1,010	2	100	,368

ANOVA

Neliöhinta

Onko keskustassa?		Sum of Squares	df	Mean Square	F	Sig.
Ei ole	Between Groups	2927273,954	2	1463636,977	17,472	,000
	Within Groups	10303595,08	123	83769,066		
	Total	13230869,03	125			
On	Between Groups	2051644,770	2	1025822,385	6,871	,002
	Within Groups	14929732,24	100	149297,322		
	Total	16981377,01	102			

Multiple Comparisons

Dependent Variable: Neliöhinta
Bonferroni

Onko keskustassa?	(I) Huoneisto	(J) Huoneisto	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Ei ole	Yksiö	Kaksio	440,00653*	78,42011	,000	249,6681	630,3449
		Kaksiota suurempi	430,43288*	79,15037	,000	238,3220	622,5438
	Kaksio	Yksiö	-440,00653*	78,42011	,000	-630,3449	-249,6681
		Kaksiota suurempi	-9,57364	55,73885	1,000	-144,8609	125,7137
	Kaksiota suurempi	Yksiö	-430,43288*	79,15037	,000	-622,5438	-238,3220
		Kaksio	9,57364	55,73885	1,000	-125,7137	144,8609
On	Yksiö	Kaksio	429,06445*	118,56855	,001	140,3596	717,7693
		Kaksiota suurempi	416,67765*	127,08246	,004	107,2422	726,1131
	Kaksio	Yksiö	-429,06445*	118,56855	,001	-717,7693	-140,3596
		Kaksiota suurempi	-12,38680	85,08603	1,000	-219,5644	194,7908
	Kaksiota suurempi	Yksiö	-416,67765*	127,08246	,004	-726,1131	-107,2422
		Kaksio	12,38680	85,08603	1,000	-194,7908	219,5644

*. The mean difference is significant at the 0.05 level.

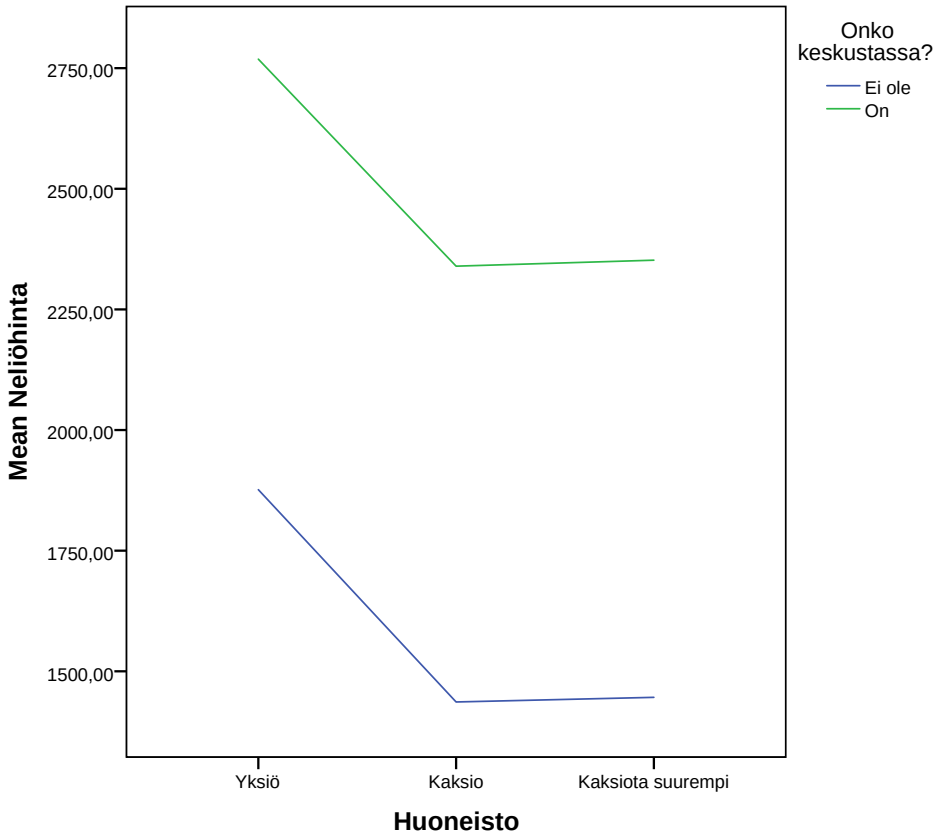
Ensin huomataan, että sijainnista riippumatta aineistossa yksiöiden neliöhinnan keskiarvot näyttäisivät poikkeavan muista. Kun testataan huoneistotyypeittäin varianssien yhtäsuuruutta, voidaan olettaa niiden olevat yhtä suuret sekä keskustassa ($p = 0{,}368$) että keskustan ulkopuolella ($p = 0{,}808$). Testattaessa odotusarvojen yhtäsuuruutta saadaan keskustassa $F = 6{,}871$ ja $p = 0{,}002$ ja keskustan ulkopuolella $F = 17{,}472$ ja $p < 0{,}001$, joten molemmissa ryhmissä H_0 hylätään ja tehdään johtopäätelmät, että huoneistopyypeittäin neliöhinnat eivät ole keskimäärin samoja. Missä sitten on eroja? Huoneistotyypeittäin vertailu (Multiple Comparisons) kertoo, että yksiöt eroavat muista, mutta kaksioiden ja sitä suurempien välillä ei keskimääräisissä neliöhinnoissa ole eroja. Sama päättely tehdään sekä keskustassa että keskustan ulkopuolella olevien asuntojen osalta. ■

Jos halutaan selittää kvantitatiivista muuttujaa kahdella kvalitatiivisella muuttujalla samanaikaisesti, voidaan joissain tilanteissa käyttää kaksisuuntaista varianssianalyysiä. Analyysi saadaan tehtyä valikosta

Analyze ► General Linear Model ► Univariate... annetaan selitettävä, riippuva (Dependent) muuttuja sekä selittävät (Fixed Factors) muuttujat.

Kaksisuuntaisessa varianssianalyysissä voidaan tutkia molempien selittäjien omavaikutusta sekä yhdysvaikutusta. Jokaiseen saadaan omat F-testit.

Esimerkki 18. Tehdään kaksisuuntainen varianssianalyysi huoneistotyypin ja sijainnin samanaikaisesta vaikutuksesta keskimääräiseen neliöhintaan. Esimerkin 17 ryhmäkeskiarvoista nähdään, että ne käyttäytyvät hyvin samalla tavalla huoneistotyypeittäin keskustassa ja keskustan ulkopuolella. Keskiarvojen vertailu voidaan tehdä graafisesti Graphs ► Legacy Dialogs ► Line... Multiple pyytäen keskiarvot *Neliöhinnasta* ja antamalla ryhmittelymuuttujat Category Axis *Huoneisto*, Define Lines by *Onko keskustassa?* Näin saadaan kuvaaja



Etukäteen voisi ryhmäkeskiarvojen perusteella arvioida, että yhdysvaikutusta ei olisi havaittavissa. Tehdään kaksisuuntainen varianssianalyysi ja saadaan tulos

Tests of Between-Subjects Effects

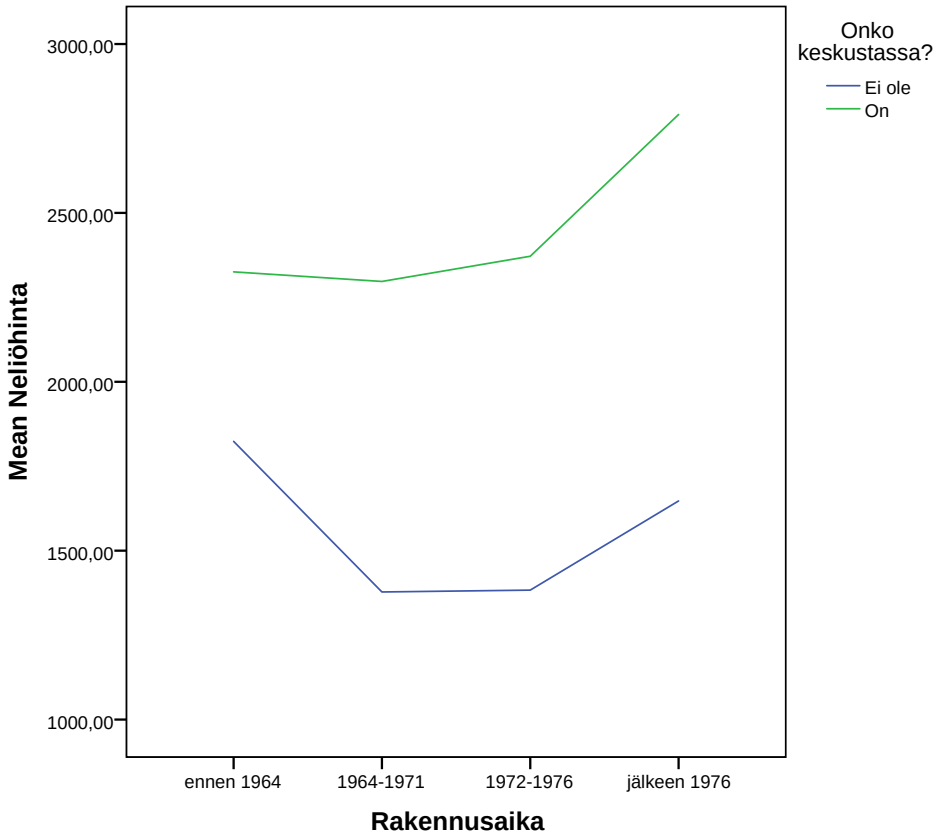
Dependent Variable: Neliöhinta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	50309431,3[a]	5	10061886,26	88,922	,000
Intercept	684718896,0	1	684718896,0	6051,216	,000
Huoneisto	4834593,872	2	2417296,936	21,363	,000
Onko keskustassa?	33464982,52	1	33464982,52	295,747	,000
Huoneisto * Onko keskustassa?	1047,606	2	523,803	,005	,995
Error	25233327,31	223	113153,934		
Total	907041110,3	229			
Corrected Total	75542758,64	228			

a. R Squared = ,666 (Adjusted R Squared = ,658)

Huomataan, että yhdysvaikutukseen liittyvä $p = 0{,}995$ ($F = 0{,}005$). Päätellään, että yhdysvaikutusta ei ole. ■

Esimerkki 19. Vaikuttaako rakennusaika samalla tavalla neliöhintaan keskustassa ja muualla? Ryhmäkeskiarvoista nähdään, että ne käyttäytyvät eri tavalla rakennusajankohdan (luokiteltu *Rakennusvuosi*-muuttuja) mukaan keskustassa ja keskustan ulkopuolella.



Näin voisi olla odotettavissa yhdysvaikutusta. Tehdään kaksisuuntainen varianssianalyysi ja saadaan tulos

Tests of Between-Subjects Effects

Dependent Variable: Neliöhinta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	51473967,7[a]	7	7353423,957	67,519	,000
Intercept	699593758,9	1	699593758,9	6423,680	,000
Onko keskustassa?	34440038,88	1	34440038,88	316,229	,000
Rakennusaika	4400240,537	3	1466746,846	13,468	,000
Onko keskustassa? * Rakennusaika	2298538,654	3	766179,551	7,035	,000
Error	24068790,94	221	108908,556		
Total	907041110,3	229			
Corrected Total	75542758,64	228			

a. R Squared = ,681 (Adjusted R Squared = ,671)

Huomataan, että yhdysvaikutukseen liittyvä $p < 0{,}001$ ($F = 7{,}035$), joten päätellään yhdysvaikutusta olevan. Myös molempien muuttujien omavaikutukset ovat merkitseviä ($F = 316{,}229$, $p < 0{,}001$, $F = 13{,}468$, $p < 0{,}001$). ■

6.4 Regressioanalyysi

Regressioanalyysillä tutkitaan muuttujan y riippuvuutta muuttujista x_1, x_2, ..., x_k. Pyritään löytämään malli, joka kertoisi y:n riippuvuuden selittäjistä. Kaikkien muuttujien oletetaan olevan kvantitatiivisia. Tosin joissain tilanteissa selittäjissä voi

olla dikotomisia muuttujia, mikä on sitten huomioitava mallin tulkinnassa tietyllä tavalla.

Regressioanalyysin yhteydessä ajatellaan selitettävän muuttujan y riippuvuuden muuttujista $x_1, x_2, \ldots, x_k$ olevan muotoa

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

missä Y on satunnaismuuttuja (response) selitettävä muuttuja, havaittavissa oleva; $x_1, x_2, \ldots, x_k$ ovat selittäviä, ei-satunnaisia, havaittuja, kontrolloitavissa olevia; ε on satunnaismuuttuja, virhetermi (ei havaittavissa oleva), $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ ovat mallin tuntemattomat parametrit, jotka aineiston perusteella ovat estimoitavissa (arvioitavissa). Jos $k = 1$, on kyse yhden selittäjän regressiomallista, jos $k = 2$ kahden selittäjän, jne. Vakiokerroin β_0 voi tarvittaessa puuttua mallista.

Tavanomainen yhden selittäjän regressioanalyysi sopii käytettäväksi lineaarisesti riippuvien muuttujien yhteydessä. Tällöin pisteparveen voidaan sovittaa suora, jonka ympärille pisteiden ajatellaan ryhmittyneen. Tällöin y:n riippuvuus muuttujasta x on muotoa

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

missä β_0 ja β_1 ovat mallin parametrit sekä ε satunnaisvirhe. Tässä yhden selittäjän regressiomallissa ajatellaan siis satunnaismuuttujan Y:n muodostuvan x:n avulla selitettävästä osasta $\beta_0 + \beta_1 x$ sekä satunnaisvaihtelusta ε. Jos oletetaan, että on tehty n havaintoa muuttujan x eri arvoilla, niin malli voidaan kirjoittaa muodossa

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

Regressiomallissa oletetaan, että jokainen $\varepsilon_i \sim N(0, \sigma^2)$ ja ε_i:t ovat riippumattomia.

Mallin estimointi sisältää parametrien β_0 ja β_1 estimoinnin. Tässä siis estimoidaan suora, jonka ajatellaan kuvaavan y:n riippuvuutta x:stä. Estimoitu malli (suora) on $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Yhden selittäjän regressioanalyysissä siis estimoidaan pisteparveen sovitettava suora, suoran vakiokerroin sekä kulmakerroin. Ajatellaan, että y:n arvot määräytyvät keskimäärin x:n arvoista estimoidun suoran yhtälön mukaisesti. Mitä paremmin pisteet ovat keskittyneet suoran ympärille, sitä voimakkaampaa on riippuvuus. Korrelaatiokerroin mittaa tätä lineaarisen riippuvuuden voimakkuutta. Otoksesta laskettua korrelaatiokerrointa käyttäen voidaankin testata, onko populaatiossa kahden muuttujan välinen korrelaatiokerroin nolla. Tällöin

H_0: populaatiossa muuttujien välinen korrelaatiokerroin on nolla

H_1: populaatiossa muuttujien välinen korrelaatiokerroin ei ole on nolla.

Tässä käytetään testisuuretta, joka noudattaa Studentin t-jakaumaa. Kun SPSS:llä lasketaan korrelaatiomatriisi (ks. luku 5), niin saadaan samalla tähän t-testiin liittyvä p-arvo.

Esimerkki 20. Esimerkissä 11 otoskorrelaatiokerroin rasvaprosentin ja vyötärön ympärysmitan välillä on 0,813. Kun testataan hypoteesia H_0: rasvaprosentti ja vyötärön ympärysmitta eivät riipu lineaarisesti toisistaan, se hylätään, koska $p < 0{,}001$. Lineaarista riippuvuutta siis on ja se voidaan mallittaa suorittamalla regressioanalyysi (ks. esimerkki 21). ■

Regressioanalyysissä estimoinnin lisäksi suoritetaan erilaisia mallin uskottavuuden ja hyvyyden tarkasteluja. Ensimmäisenä on selvitettävä, voidaanko estimoitujen parametrien perusteella päätellä, että mallin parametrit ovat nollasta poikkeavia.

Testataan yhden selittäjän mallissa aluksi sitä, onko x merkitsevä selittäjä. Tällöin testattavana hypoteesina on

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0,$$

johon saadaan t-testisuure. Jos x on todettu merkitseväksi selittäjäksi, niin seuraavaksi voidaan tutkia, onko vakiokertoimen β_0 syytä olla mallissa.

Tällöin

$$H_0\colon \beta_0 = 0$$
$$H_1\colon \beta_0 \neq 0.$$

Tähänkin saadaan t-testisuure.

Lisäksi saadaan laskettua mallin selityskerroin R^2. Ilmoittamalla $100R^2$, voidaan puhua mallin selitysasteesta. Yhden selittäjän regressiomallissa $100R^2 = 100r^2$ kertoen kuinka monta prosenttia y:n vaihtelusta kyseisellä yhden selittäjän mallilla voidaan x:n avulla selittää. Selityskertoimella on tämä tulkinta kuitenkin vain silloin, kun mallissa on vakiokerroin.

Regressioanalyysin suoritus tapahtuu valikosta

Analyze ▸ Regression ▸ Linear… annetaan selitettävä, riippuva (Dependent) muuttuja sekä selittävä(t), (riippumattomat, Independent(s)) muuttuja(t).

Esimerkki 21. Tarkastellaan rasvaprosentin riippuvuutta vyötärön ympärysmitasta. Esimerkissä 11 olevasta pisteparvesta nähdään, että pisteparveen voidaan sovittaa suora. Kun suoritetaan regressioanalyysi, saadaan tulokset

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,813[a]	,662	,660	4,877

a. Predictors: (Constant), vyötärön ympärys(cm)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11631,527	1	11631,527	488,928	,000[b]
	Residual	5947,463	250	23,790		
	Total	17578,990	251			

a. Dependent Variable: rasvaprosentti

b. Predictors: (Constant), vyötärön ympärys(cm)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-39,280	2,660		-14,765	,000
	vyötärön ympärys(cm)	,631	,029	,813	22,112	,000

a. Dependent Variable: rasvaprosentti

Tarkasteltava regressiomalli on nyt

$$Rasvaprosentti = \beta_0 + \beta_1 Vyötärön\ ympärys(cm) + \varepsilon.$$

Kun malli on estimoitu, saadaan suora

$$\widehat{Rasvaprosentti} = -39{,}280 + 0{,}631\ Vyötärön\ ympärys(cm).$$

Esimerkiksi vyötärön ympärysmitan ollessa 90 cm rasvaprosentti on keskimäärin 17,51. Siis yhden senttimetrin lisäys nostaa rasvaprosentti keskimäärin 0,631. Mallissa molemmat kertoimet ovat merkitseviä ($t = 22{,}112$, $p < 0{,}001$; $t = -14{,}765$, $p < 0{,}001$). Rasvaprosenttia voidaan siis selittää vyötärön ympärysmitalla esitetyn mallin mukaisesti. Lisäksi saadaan selitysprosentiksi 66,2. Vyötärön ympärys siis selittää rasvaprosentista vaihtelusta 66,2 %. ■

Useamman selittäjän malliin liittyvät yksittäisten kertoimien testaukset t-testien avulla. Tällöin tutkitaan sitä, lisääkö kyseisen selittäjän tuonti malliin muiden jo siellä ollessa mallin selitysastetta riittävästi. Tällöin

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$.

Lisäksi testataan F-testillä kaikkien selittäjien yhteisvaikutusta eli sitä saadaanko y:n vaihtelua selitettyä siten, että otetaan kaikki tarkasteltavat selittäjät samanaikaisesti malliin mukaan. Tämä regressiokertoimien yhteistestaus (kun vakiokerroin on mallissa mukana) voidaan muotoilla

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: ainakin jokin $\beta_i \neq 0$.

Mallin valinta ei aina ole kovin helppoa. Pyritään valitsemaan niin monta merkitsevää selittäjää, että selitysaste on mahdollisimman hyvä. On kuitenkin pidettävä mielessä se, että mallin on oltava käyttötarkoitukseensa sopiva ja tulkittavissa oleva. Vaikka on olemassa erilaisia automaattisia mallinvalintamenettelyjä, on niitä syytä käyttää hyvin harkiten.

Testauksien lisäksi mallin sopivuuden ja hyvyyden tarkasteluihin liittyvät oletusten, jotka liittyvät satunnaisvirheeseen ε, voimassaolon tutkimiset. Niitä ei kuitenkaan käsitellä tässä yhteydessä.

Esimerkki 22. Lisätään esimerkissä 21 toiseksi selittäjäksi henkilön paino. Saadaan tulokset

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,848[a]	,719	,717	4,456

a. Predictors: (Constant), Paino_kg, vyötärön ympärys(cm)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12635,745	2	6317,872	318,242	,000[b]
	Residual	4943,245	249	19,852		
	Total	17578,990	251			

a. Dependent Variable: rasvaprosentti

b. Predictors: (Constant), Paino_kg, vyötärön ympärys(cm)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-45,952	2,605		-17,640	,000
	vyötärön ympärys(cm)	,990	,057	1,275	17,447	,000
	Paino_kg	-,326	,046	-,520	-7,112	,000

a. Dependent Variable: rasvaprosentti

Tarkasteltava regressiomalli on nyt

$$Rasvaprosentti = \beta_0 + \beta_1 Vyötärön\ ympärys(cm) + \beta_1 Paino\_kg + \varepsilon.$$

Estimointituloksen perusteella nähdään, että yksittäisten kertoimien testauksen yhteydessä kaikki nollahypoteesit hylätään eli mallin kertoimet ovat merkitseviä ($t = -17{,}640$, $p < 0{,}001$; $t = 17{,}640$, $p < 0{,}001$; $t = -7{,}112$, $p < 0{,}001$) ja selitysprosentti 71,9. Samoin yhteistestauksessa nollahypoteesi hylätään ($F = 318.242$, $p < 0{,}001$). Malli on siis kaikin puolin kunnossa ja rasvaprosenttia voidaan nyt estimoida

$$\widehat{Rasvaprosentti} = -45{,}952 + 0{,}990\, Vyötärön\ ympärys(cm) - 0{,}326\, Paino\_kg.$$

Esimerkiksi miehillä, joiden vyötärön ympärysmitta on 98 cm ja paino 95 kg, keskimäärin rasvaprosentti on 20,10. ■

7 Lopuksi

Oppaassa tarkasteltiin empiirisen tutkimuksen eri työvaiheita ja toteutusta SPSS-ohjelmistolla. Seuraavassa on lyhyesti yhteenveto tutkimuksen työvaiheista.

Kun havaintoaineisto on hankittu, muokataan se analysointia varten havaintomatriisimuotoon. Muuttujien mitta-asteikot on syytä selvittää, jotta analyysit tulee oikein valituksi. Havaintomatriisi talletetaan tietokoneelle joko käytettävällä tilastolaskentaohjelmistolla tai siten, että tämä ohjelmisto pystyy sen lukemaan. Tietojen taltioinnin oikeellisuus on syytä tarkistaa. Yleiskuvan saamiseksi aineistosta analysointi aloitetaan muuttujien jakaumien muodostamisella sekä tarpeellisten tunnuslukujen laskulla. Käytetään tarpeen mukaan tilanteeseen sopivia graafisia esityksiä. Jakaumien teon yhteydessä voidaan löytää tallennusvirheitä.

Seuraavaksi on vuorossa varsinainen analysointi. Valitaan kuhunkin tilanteeseen käyttökelpoinen menetelmä ja suoritetaan analyysi ja tulkitaan tulokset. Jokaiseen analysointivaiheeseen kuuluu siis johtopäätelmien teko. Esimerkiksi aineiston

kuvailun yhteydessä voidaan kiinnittää huomio jakauman muotoon. Riippuvuustarkastelujen yhteydessä tehdään johtopäätelmiä riippuvuussuhteista perustaen päätelmien teko analysoinnissa saatuihin tuloksiin.

Tilastollisen tutkimuksen keskeisen vaiheen muodostaakin näiden tutkimustulosten esittäminen sellaisessa kirjallisessa asussa, että lukija, jolle tutkimustulokset on tarkoitettu, saa sen sisältämän informaation mahdollisimman helposti, havainnollisesti ja yksikäsitteisessä muodossa. Raportointi on syytä jäsennellä selkeästi alaotsikointia ja kappalejakoa käyttäen. Kuviot ja taulukot laaditaan yleisten sopimusten mukaisesti, ne numeroidaan ja otsikoidaan. Kuvioiden ja taulukoiden on muodostettava sellaisia itsenäisiä kokonaisuuksia, että lukija voi muuhun tekstiin turvautumatta ymmärtää niissä esitetyn asian. (ks. http://www.fsd.uta.fi/menetelmaopetus/raportointi/numerotulokset.html)

Tässä oppaassa käytettiin vain murto-osaa tarjolla olevista menetelmistä. Menetelmät, joita esiteltiin, ovat ehkä kaikkein tavanomaisimpia ja useimmiten kaikilla tilastotieteen perusopintotasoisilla kursseilla esitettyjä. Lopuksi vielä yhteenveto näiden tilastollisten analyysien suorittamisesta SPSS-ohjelmistolla.

Analyze ▸

- **Descriptive Statistics** ▸ frekvenssijakaumat, tunnusluvut, ristiintaulukot
- **Compare Means** ▸ *t*-testi, yksisuuntainen varianssianalyysi
- **General Linear Model** ▸ kaksisuuntainen varianssianalyysi
- **Correlate** ▸ korrelaatiomatriisi
- **Regression** ▸ regressioanalyysi

Graphs ▸ **Legacy Dialogs** ▸

- **Bar...** pylväsdiagrammit
- **Pie...** piirakat
- **Boxplot...** laatikko-jana -kuviot
- **Scatter...** pisteparvet
- **Histogram...** frekvenssihistogrammit

Raportin analyyseissä käytetyt aineistot

Rasvaprosentti-aineistoa
http://www.sis.uta.fi/tilasto/tiltp_aineistoja/rasvaprosentti.sav,
muuttujien kuvaukset
http://www.sis.uta.fi/tilasto/tiltp_aineistoja/rasvaprosentti.PDF,
lähde
https://ww2.amstat.org/publications/jse/v4n1/datasets.johnson.html

Asunnot_2006-aineisto
http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Asunnot_2006.sav
Tampereella ETUOVI.com-palvelussa myynnissä olleita kerrostalohuoneistoja, jotka olivat esittelyssä 7.–14.4.2006. Mukana on vuosina 1950–2000 rakennetut talot, aineiston kerännyt Mika Hannula.

ARVIO-aineisto
http://www.sis.uta.fi/tilasto/tiltp_aineistoja/arvio.sav,
kyselylomake
http://www.sis.uta.fi/tilasto/tiltp3/kevat2003/Aineistoja/arviointi_lomake.pdf