Tilastomenetelmien perusteet
MTTTA1
Luentorunko

Raija Leppälä

20. joulukuuta 2018

Sisältö

Luku 1

Johdanto

Opintojaksolla Tilastollisen päättelyn perusteet tutustuttiin todennäköisyysjakaumiin, otosjakaumiin, parametrien estimointiin sekä hypoteesien testaukseen. Tällä kurssilla tutustutaan varianssianalyysiin, regressioanalyysiin sekä χ^2-yhteensopivuustestiin ja χ^2-riippumattomuustestiin sekä hyvin lyhyesti epäparametrisiin testeihin.

Yksisuuntainen varianssianalyysi on yleistys kahden riippumattoman otoksen t-testistä. Regressioanalyysin avulla mallitetaan muuttujien välistä riippuvuutta. χ^2-yhteensopivuustestin avulla voidaan testata sitä, onko otos peräisin tietystä jakaumasta. χ^2-riippumattomuustesti testaa kahden muuttujan välistä riippumattomuutta perustana ristiintaulukko. Epäparametrisissä testeissä voidaan tinkiä jakaumaoletuksista, joita parametrisissä testeissä joudutaan tekemään.

Empiirisessä tutkimuksessa on käytössä satunnaisotos, jonka perusteella pyritään tekemään johtopäätelmiä populaatiosta. Yksinkertaisimmissa tilanteissa johtopäätelmien teko voidaan perustaa otoksesta laskettuun sopivaan testisuureeseen, jonka todennäköisyysjakauma nollahypoteesin vallitessa tunnetaan. Tilastollinen päättely sisältää aina tiettyä epävarmuutta, mutta sitä pyritään hallitsemaan juuri näiden otossuureiden todennäköisyysjakaumien avulla. Seuraavassa lyhyesti kertauksena opintojaksolla Tilastollisen päättelyn perusteet esillä olleita asioita.

1.1 Jatkuvista jakaumista

1.1.1 Normaalijakauma

Jatkuvan satunnaismuuttujan X, joka voi saada kaikki reaalilukuarvot, sanotaan noudattavan normaalijakaumaa parametrein μ ja σ^2 ($\sigma > 0$), jos sen tiheysfunktio on

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}[(x-\mu)/\sigma]^2}, \qquad -\infty < x < \infty$$

Merkitään $X \sim N(\mu, \sigma^2)$. Tällöin $E(X) = \mu$ ja $\mathrm{Var}(X) = \sigma^2$. Jos $X \sim N(0, 1)$, niin kyse on nk. standardoidusta normaalijakaumasta.

Usein merkitään $Z \sim N(0,1)$, $f(z) = \phi(z)$ ja $F(z) = \Phi(z)$. Standardoidun normaalijakauman kertymäfunktion arvot on taulukoitu. Näitä taulukoita voidaan käyttää hyväksi laskettaessa normaalijakaumaan liittyviä todennäköisyyksiä.

Jos $X \sim N(\mu, \sigma^2)$, niin $Z = (X - \mu)/\sigma \sim N(0,1)$.

Olkoon $Z \sim N(0,1)$. Määritellään z_α siten, että $P(Z \geq z_\alpha) = \alpha$. Samoin $z_{\alpha/2}$ siten, että $P(Z \geq z_{\alpha/2}) = \alpha/2$.

Esimerkki 1.1.1

$$\alpha = 0.1, \qquad z_\alpha = 1.28, \qquad z_{\alpha/2} = 1.65.$$

1.1.2 Studentin t-jakauma

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

Olkoon t_{df} Studentin t-jakaumaa noudattava satunnaismuuttuja. Määritellään $t_{\alpha;df}$ siten, että $P(t_{df} \geq t_{\alpha;df}) = \alpha$ ja $P(t_{df} \geq t_{\alpha/2;df}) = \alpha/2$. Näitä Studentin t-jakauman ylempiä fraktiileja eri vapausastein on taulukoitu.

Esimerkki 1.1.2

$$\begin{aligned} &\alpha = 0.1, && t_{\alpha;23} = 1.32, && t_{\alpha/2;23} = 1.714; \\ &\alpha = 0.01, && t_{\alpha;120} = 2.358, && t_{\alpha/2;120} = 2.617. \end{aligned}$$

1.2 Satunnaisotos, otossuure, otantajakauma

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

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

Satunnaisotoksesta muodostetut funktiot ovat satunnaismuuttujia, joita kutsutaan *otossuureiksi*. Otossuuren todennäköisyysjakaumaa kutsutaan *otanta-* tai *otosjakaumaksi*.

1.3 Estimointi

Estimointi on populaation tuntemattoman parametrin arviointia sopivan otossuureen avulla. Tätä otossuuretta kutsutaan *estimaattoriksi* ja sen arvoa *estimaatiksi*. Näin tehtäessä puhutaan *piste-estimoinnista*. Esimerkiksi voidaan estimoida populaation odotusarvoa otoskeskiarvolla, populaation varianssia otosvarianssilla.

Estimaattori on *harhaton*, jos sen odotusarvo on estimoitava parametri.

Esimerkki 1.3.1 Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos jakaumasta, jonka odotusarvo on μ ja varianssi σ^2. Tällöin $\overline{X}$ on μ:n harhaton estimaattori. Estimaattorin keskivirhe on $\sigma/\sqrt{n}$.

Väliestimoinnin yhteydessä ilmoitetaan *väli*, jolle arvellaan tuntemattoman parametrin kuuluvan. Tämä nk. *luottamusväli* muodostetaan vastaavan piste-estimaattorin ja piste-estimaattorin otantajakauman keskihajonnan eli *estimaattorin keskivirheen* avulla.

1.4 Tilastollinen testaus

Tilastollinen *hypoteesi* on väittämä populaatiosta, sen jakaumasta ja/tai jakauman parametrista. *Hypoteesin testaus* tarkoittaa väittämän tutkimista otoksen perusteella. *Testauksessa* määritellään sopiva otossuure, jota kutsutaan *testisuureeksi*, ja lasketaan otoksesta sille arvo, jonka perusteella väittämä hyväksytään tai hylätään. Väittämä laaditaan siten, että sen ollessa tosi testisuureen todennäköisyysjakauma tunnetaan. Havaitun otoksen perusteella lasketaan testisuurelle arvo, jonka avulla päätellään sopiiko saatu arvo testisuureen jakaumaan vai kuuluuko se harvinaisten arvojen joukkoon. Jos testisuureen arvo sopii väittämän jakaumaan hyväksytään väittämä. Jos laskettu testisuureen arvo voidaan katsoa kovin harvinaiseksi, niin väittämä hylätään ja hyväksytään nk. vaihtoehtoinen hypoteesi.

Hypoteesin testauksessa asetetaankin siis kaksi väittämää, joista jompi kumpi on välttämättä voimassa: Nollahypoteesi H_0, jonka ollessa tosi testisuuren jakauma tunnetaan sekä vaihtoehtoinen hypoteesi H_1.

Testauksen vaiheet:

1. Asetetaan H_0 ja H_1 siten, että jompi kumpi väittämä välttämättä voimassa.
2. Valitaan riskitaso (merkitsevyystaso) α eli oikean H_0:n hylkäämisen todennäköisyys.
3. Muodostetaan testisuureen otosjakauma, kun oletetaan H_0 todeksi.
4. Määrätään testisuureen harvinaisten arvojen joukko eli testin *kriittinen alue*, joka riippuu valitusta merkitsevyystasosta sekä vaihtoehtoisesta hypoteesista H_1.
5. Lasketaan otoksesta testisuureelle arvo.
6. Hylätään H_0, jos saatu arvo kuuluu kriittiselle alueelle, muulloin hyväksytään.

Testauksen yhteydessä informatiivista on myös ilmoittaa todennäköisyys, että H_0:n vallitessa saadaan havaittu tai sitä harvinaisempi arvo testisuureelle. Tämä todennäköisyys on *pienin riskitaso*, jolla H_0 voidaan hylätä. Tätä todennäköisyyttä merkitään p:llä ja puhutaan *p-arvosta*. Testimenettelyssä voidaan nyt laskea testisuureen arvoon liittyvä p-arvo ja hylätä H_0 mikäli p on pienempi

kuin valittu α. Testisuureita (ks. Tilastollisen päättelyn perusteet kaavakokoelma http://www.sis.uta.fi/tilasto/mtttp5/syksy2018/kaavat.pdf):

1) $H_0\colon \mu = \mu_0$

Oletetaan, että $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 on tunnettu. Tällöin H_0:n ollessa tosi

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

2) $H_0\colon \mu = \mu_0$

Oletetaan, että $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 on *tuntematon*. Tällöin H_0:n ollessa tosi

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$$

Esimerkki 1.4.1 Testataan hypoteesia, että populaation odotusarvo on 50. Viiden alkion otoksen perusteella otoskeskiarvoksi saadaan 65 ja keskihajonnaksi 11.6 Mikä on pienin riskitaso, jolla nollahypoteesi voidaan hylätä yksisuuntaisessa testissä?

$$\begin{aligned} &H_0\colon \mu = 50 && \overline{x} = 65 \\ &H_1\colon \mu > 50 && s = 11.6 \\ & && n = 5 \end{aligned}$$

Jos H_0 tosi, niin $t = \dfrac{\overline{X} - 50}{s/\sqrt{n}} \sim t(n-1)$.

$$t_{\text{hav.}} = \frac{65 - 50}{11.6/\sqrt{5}} = 2.89, \qquad t_{0.025;4} = 2.776, \qquad t_{0.01;4} = 3.747,$$
$$0.01 < p < 0.025.$$

3) $H_0\colon \pi = \pi_0$

Olkoon populaatiossa π % viallisia. Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos tästä populaatiosta. Jos H_0 on tosi, $p \sim N(\pi_0, \pi_0(100 - \pi_0)/n)$, likimain ja

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}} \sim N(0,1), \qquad \text{likimain,}$$

missä p on viallisten %-osuus otoksessa.

Esimerkki 1.4.2 Eräs puolue väittää, että suomalaisista 40 % kannattaa sitä. Väitteen tutkimiseksi teet kyselyn 5000 henkilölle, joista 1800 ilmoitti kannattavansa kyseistä puoluetta. Onko puolue arvioinut kannatuksensa oikein?

$$H_0 \colon \pi = 40\ \%,$$
$$H_1 \colon \pi < 40\ \%.$$

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}} \sim N(0,1), \qquad \text{likimain,}$$

kun H_0 tosi (eli $\pi_0 = 40\ \%$).

$$z_{\text{hav.}} = \frac{36 - 40}{\sqrt{40 \cdot 60/5000}} = -5.77, \qquad -z_{0.001} = -3.08.$$

Koska $z_{\text{hav.}} < -3.08$, niin H_0 hylätään 0,1 %:n riskitasolla ja päätellään, että puolue on arvioinut kannatuksensa liian suureksi.

4) $H_0 \colon \mu_1 = \mu_2$

Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta ja $Y_1, Y_2, \ldots, Y_m$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta, missä σ_1 ja σ_2 tunnettuja sekä satunnaisotokset toisistaan riippumattomia. Jos H_0 tosi, niin

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim N(0,1).$$

5) $H_0 \colon \mu_1 = \mu_2$

Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta ja $Y_1, Y_2, \ldots, Y_m$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta, missä σ_1 ja σ_2 tuntemattomia mutta yhtä suuria sekä satunnaisotokset toisistaan riippumattomia. Jos H_0 tosi, niin

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{1/n + 1/m}} \sim t(n + m - 2),$$

missä

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n + m - 2}.$$

Esimerkki 1.4.3 Psykologi on kehittänyt testin, joka koostuu muutamasta yksinkertaisesta käsin suoritettavista tehtävistä ja jonka tarkoitus on paljastaa mahdollinen lievä kehityshäiriö. Hän on poiminut satunnaisotoksen sekä normaaleista lapsista että kehityshäiriöisistä. Suoritusajat ovat:

Normaali	204	218	197	183	227	233	191	
Kehityshäiriö	243	228	261	202	343	242	220	239

Kelpaako testi tarkoitukseen?

$$\begin{array}{lll} H_0\colon \mu_N = \mu_K & \overline{x}_N = 207.57 & \overline{x}_K = 247.25 \\ H_1\colon \mu_N < \mu_K & s_N^2 = 18.87^2 & s_K^2 = 42.48^2 \\ & n_N = 7 & n_K = 8 \end{array}$$

Riippumattomien otosten t-testi odotusarvojen erotukselle.

$$t_{\text{hav.}} = \frac{207.57 - 247.25}{33.7\sqrt{\frac{1}{7} + \frac{1}{8}}} = -2.28, \qquad t_{0.01;13} = 2.65, \qquad t_{0.025;13} = 2.16.$$

H_0 voidaan hylätä esim. 2.5 %:n riskitasolla, mutta ei 1 %:n riskitasolla. Jos kiinnitetään 2.5 %:n riski, niin tehdään päätelmä, että testi kelpaa.

6) H_0: $\mu_1 = \mu_2$ (vastinparitilanne), H_0: $\mu_D = 0$

H_0:n ollessa tosi testisuure

$$t = \frac{\overline{D}}{s_D/\sqrt{n}} \sim t(n-1).$$

Esimerkki 1.4.4 Halutaan tutkia erään menetelmän vaikutusta ihmisen hengitystilavuuteen. Tehdään 5 alkion satunnaisotos populaatiosta ja mitataan koehenkilöiden hengitystilavuudet ennen menetelmän soveltamista sekä menetelmän soveltamisen jälkeen. Tulokset ohessa. Onko menetelmällä ollut vaikutusta?

Hengitystilavuus					
ennen	2750	2360	2950	2830	2260
jälkeen	2850	2380	2930	2860	2330

(Liski & Puntanen)

Lasketaan erotukset ja saadaan 100, 20, −20, 30, 70. Näistä keskiarvo 40 ja keskihajonta 46.37, joten

$$t_{\text{hav.}} = \frac{40}{46.37/\sqrt{5}} = 1.93.$$

$t_{0.10;4} = 1.533 < t_{\text{hav.}} < 2.132 = t_{0.05;4}$, joten yksisuuntaisessa testissä $0.05 < p < 0.1$. Päätellään ei vaikutusta.

Suurten otosten tapauksessa edellä esitettyjä testejä voidaan käyttää myös muidenkin kuin normaalijakaumien yhteydessä.

1.5 SPSS-ohjeita

Luottamusvälit:

1) Luottamusväli populaation odotusarvolle

Analyze
Compare Means ▶ One-Sample T Test...

Muuttujan oltava vähintään intervalliasteikollinen.

2) Luottamusväli populaation odotusarvojen erotukselle riippumattomien otosten tilanteessa (ks. myös t-testi odotusarvojen yhtäsuuruudelle)

Analyze
Compare Means ▶ Independent-Samples T Test...

Riippuvan muuttujan oltava vähintään intervalliasteikollinen; selittävä muuttuja kahdessa luokassa

3) Luottamusväli populaation odotusarvojen erotukselle riippuvien otosten tilanteessa (vastinparitilanne) (ks. myös vastaava t-testi odotusarvojen yhtä suuruudelle)

Analyze
Compare Means ▶ Paired-Samples T Test...

Tarkasteltava ominaisuus mitattu vähintään intervalliasteikolla. ”Ennen” ja ”jälkeen” muuttujina havaintomatriisissa.

4) Luottamusvälit prosentuaalisille osuuksille: Ohjelmistolla lasketaan prosentuaaliset osuudet aineistossa esim. frekvenssijakauman tai ristiintaulukon avulla

Analyze
Descriptive Statistics ▶ Frequencies...
Crosstabs...

ja tämän jälkeen itse kyseinen luottamusväli.

Testisuureet:

1) H_0: $\mu = \mu_0$

Analyze
Compare Means ▶ One Sample T Test...

Muuttujan oltava vähintään intervalliasteikollinen.

2) H_0: $\pi = \pi_0$. Lasketaan vastaava %-osuus otoksesta ja sen avulla z-testisuurelle arvo. Prosenttiosuuden saa selville muodostamalla frekvenssijakauman muuttujasta. (Ei-parametrisistä testeistä löytyy mahdollisuus kyseisen testin suorittamiseen z-testisuurella, jolloin tulostuu vain p-arvon, tai käyttäen yhteensopivuustestiä.)

3) H_0: $\mu_1 = \mu_2$ (riippumattomat otokset)

Analyze
Compare Means ▶ Independent-Samples T Test...

Riippuvan muuttujan oltava vähintään intervalliasteikollinen; selittävä muuttuja kahdessa luokassa.

4) H_0: $\mu_1 = \mu_2$ (vastinparitilanne)

Analyze
Compare Means ▶ Paired-Samples T Test...

Vastinparien arvot havaintomatriisissa oltava eri muuttujissa!

Luku 2

Varianssianalyysi (ANOVA, Analysis of Variance)

2.1 Yksisuuntainen varianssianalyysi

Esimerkki 2.1.1 Tutkitaan golfpallojen lento-ominaisuuksia (mitataan lentomatkaa, Distance). Tutkittavana on kolmen erimerkkisen pallon (Brand A, B, C) ominaisuudet.

Lentomatkan ehdolliset keskiarvot ovat:

Brand A	251,28
Brand B	261,06
Brand C	269,95

Esimerkin otoskeskiarvot poikkeavat ryhmittäin toisistaan jonkin verran antaen viitteitä siitä, että populaatiossa odotusarvot saattaisivat olla eri suuret. Nyt voidaankin, samalla tavalla kuin kahden otoksen t-testissä, testata poikkeavatko odotusarvot toisistaan. Erona t-testiin on se, että kahden otoksen sijaan voi olla useampia otoksia (tässä 3).

Analysointimenetelmä on nimeltään yksisuuntainen varianssianalyysi ja tässä

$$H_0 \colon \mu_A = \mu_B = \mu_C,$$
$$H_1 \colon \text{kaikki odotusarvot eivät ole samoja.}$$

Testisuureena varianssianalyysissä on nk. F-testisuure, joka muodostetaan kahden neliösumman avulla ja jolla siis *testataan odotusarvojen yhtäsuuruutta*.

Perusoletuksen yksisuuntaisessa varianssianalyysissä (1-VA) on se, että meillä on I kappaletta toisistaan riippumattomia satunnaisotoksia normaalijakaumista, joiden varianssit ovat tuntemattomia, mutta yhtä suuria. Siis

$$\begin{aligned}
&Y_{11}, Y_{12}, \ldots, Y_{1n_1} && \text{satunnaisotos } N(\mu_1, \sigma^2)\text{:sta,} \\
&Y_{21}, Y_{22}, \ldots, Y_{2n_2} && \text{satunnaisotos } N(\mu_2, \sigma^2)\text{:sta,} \\
&\quad\vdots \\
&Y_{I1}, Y_{I2}, \ldots, Y_{In_I} && \text{satunnaisotos } N(\mu_I, \sigma^2)\text{:sta.}
\end{aligned}$$

Halutaan tutkia ovatko jakaumien odotusarvot yhtä suuret, jolloin

$$H_0\colon \mu_1 = \mu_2 = \ldots = \mu_I,$$
$$H_1\colon \text{kaikki odotusarvot eivät ole samoja.}$$

Joitain merkintöjä testisuureen määritystä varten:

$n = n_1 + n_2 + \ldots + n_I$,

$Y_{ij} = i.$ ryhmän $j.$ havainto,

$\overline{Y_i} = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} Y_{ij} = i.$ ryhmän keskiarvo,

$\overline{Y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} =$ yleiskeskiarvo eli kaikkien havaintojen keskiarvo.

Kokonaisneliösumma:

$$\begin{aligned} SST &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2 = \sum_{i=1}^{I} n_i(\overline{Y_i}-\overline{Y})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_i})^2 \\ &\overset{\text{merk.}}{=} SSB + SSW. \end{aligned}$$

Y:n kokonaisvaihtelua kuvaava SST voidaan jakaa kahteen osaan:

Kokonaisvaihtelu (SST)
= ryhmien välinen vaihtelu (SSB) + ryhmien sisäinen vaihtelu (SSW).

SSB:n yhteydessä puhutaan myös malliin liittyvästä neliösummasta ja SSW:n yhteydessä jäännösneliösummasta (SSE).

Merkitään vielä

$$MSB = \frac{SSB}{I-1} \qquad \text{ja} \qquad MSW = \frac{SSW}{n-I},$$

missä neliösummat on jaettu nk. vapausasteillaan, jolloin saadaan keskineliösummat.

Voidaan osoittaa, että MSW on σ^2:n harhaton estimaattori aina ja MSB on σ^2:n harhaton estimaattori, kun H_0 on tosi. Lisäksi H_0:n ollessa tosi $F = MSB/MSW$ noudattaa Fisherin F-jakaumaa vapausastein $I-1$ ja $n-I$. Merk.

$$F = \frac{MSB}{MSW} \sim F(I-1, n-I).$$

F-jakauma määritellään siis kaksin vapausastein. Olkoon $F_{df1,df2}$ Fisherin F-jakaumaa noudattava satunnaismuuttuja. Määritellään $F_{\alpha;df1,df2}$ siten, että $P(F_{df1,df2} \geq F_{\alpha;df1,df2}) = \alpha$.

Näitä F-arvoja on taulukoituna eri vapausastein muutamilla α:n arvoilla, ks. esim. http://www.sis.uta.fi/tilasto/mttta1/kevat2019/F_jakauma.pdf.

Edellä varianssianalyysin testauksen yhteydessä estimoidaan σ^2:sta kahdella tavalla. Jos H_0 ei ole tosi, niin MSB pyrkii yliestimoimaan varianssia, jonka seurauksena F-arvo tulee "liian suureksi". Nyt H_0 voidaan hylätä riskitasolla α, jos otoksen perusteella laskettu F:n arvo $F_{\text{havaittu}} > F_{\alpha;I-1,n-I}$.

Varianssianalyysi-nimitys on hieman harhaanjohtava. Varianssianalyysin yhteydessä testataan odotusarvojen yhtä suuruutta. Toki varianssienkin yhtäsuuruuden testaaminen voidaan (ja pitääkin) suorittaa, mutta se on oletusten paikkansa pitävyyden selvittämistä, eikä varsinaisesti riippuvuustarkastelujen tekemistä (ks. Esim. 2.1.4, SPSS-tulostus → Levenen testi sekä opintojakson Usein kysyttyä -sivu https://coursepages.uta.fi/mttta1/kevat-2019/usein-kysyttya/). Nimitys tullee testisuureesta, joka perustuu kahteen varianssin estimaattoriin.

Varianssianalyysin tulokset on tapana esittää taulukkona

vaihtelu	neliösummat (SS)	vapausasteet (df)	keskineliösummat (MS)	F-arvo	p-arvo
välinen	SSB	$I-1$	$MSB = \dfrac{SSB}{I-1}$	$F = \dfrac{MSB}{MSW}$	$P(F \geq F_{\text{hav.}})$
sisäinen (jäännös)	SSW	$n-I$	$MSW = \dfrac{SSW}{n-I}$	$\sim F(I-1, n-I)$ kun H_0 tosi	
kokonais	SST	$n-1$			

Esimerkki 2.1.2 Neliösummat sekä testaus esimerkin 2.1.1 tilanteessa.

Brand	lkm	keskiarvo	keskihajonta
A	10	251,28	5,977
B	10	261,06	3,866
C	10	269,95	4,501
	$n = 30$	$\overline{y} = 260{,}76$	

$$
\begin{aligned}
SS_1 &= (n_1-1)s_1^2 = (10-1)5{,}977^2 \approx 321{,}52\\
SS_2 &= (n_2-1)s_2^2 = (10-1)3{,}866^2 \approx 134{,}51\\
SS_3 &= (n_3-1)s_3^2 = (10-1)4{,}501^2 \approx 182{,}33\\
SSW &= SS_1 + SS_2 + SS_3 \approx 638{,}36\\
SSB &= 10(251{,}28-260{,}76)^2 + 10(261{,}06-260{,}76)^2\\
&\quad + 10(269{,}95-260{,}76)^2 \approx 1744{,}17
\end{aligned}
$$

$$
\begin{aligned}
MSB &= SSB/(I-1) = 1744{,}17/(3-1) \approx 872{,}08\\
MSW &= SSW/(n-I) = 638{,}36/(30-3) \approx 23{,}64\\
F &= MSB/MSW = 872{,}08/23{,}64 \approx 36{,}87\\
F_{0{,}01;2,27} &= 5{,}49.
\end{aligned}
$$

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	F-ratio	Prob $> F$
Model	2	1744.1647	872.082	36.8864	0.0000
Error	27	638.3450	23.642		
Total	29	2382.5097			

Esimerkki 2.1.3 Tutkitaan kolmen eri valmennusmenetelmän vaikutusta urheilusuoritukseen saatiin aineisto:

Menetelmä 1:	6	4	6	4
Menetelmä 2:	14	9	10	11
Menetelmä 3:	5	11	8	8

Onko valmennusmenetelmien vaikutuksilla merkitsevää eroa?

(Liski & Puntanen, Tilastotieteen peruskurssi II)

$$\overline{y}_1 = 5, \quad \overline{y}_2 = 11, \quad \overline{y}_3 = 8, \quad \overline{y} = 8,$$
$$n_1 = n_2 = n_3 = 4, \quad n = 12,$$
$$SST = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$$
$$= (6-8)^2 + \cdots + (8-8)^2 = 108,$$
$$SSB = \sum_{i=1}^{3} n_i(\overline{y}_i - \overline{y})^2$$
$$= 4(5-8)^2 + 4(11-8)^2 + 4(8-8)^2 = 72,$$
$$SSW = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$
$$= (6-5)^2 + (4-5)^2 + (6-5)^2 + (4-5)^2$$
$$+ (14-11)^2 + (9-11)^2 + (10-11)^2 + (11-11)^2$$
$$+ (5-8)^2 + (11-8)^2 + (8-8)^2 + (8-8)^2 = 36,$$
$$MSB = SSB/(I-1) = 72/(3-1) = 36,$$
$$MSW = SSW/(n-I) = 36/(12-3) = 4,$$
$$F = MSB/MSW = 36/4 = 9,$$
$$F_{0.01;2,9} = 8.02.$$

Tulos SPSS-ohjelmalla:

ANOVA

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	72.000	2	36.000	9.000	.007
Within Groups	36.000	9	4.000		
Total	108.000	11			

Esimerkki 2.1.4 Tutkitaan eri autotyyppien (A, B ja C) kulutusta. On saatu aineisto, jossa kulutusarvot (miles per gallon) ovat:

A-autot	B-autot	C-autot
22.2	24.6	22.7
19.9	23.1	21.9
20.3	22.0	23.3
21.4	23.5	24.1
21.2	23.6	22.1
21.0	22.1	23.4
20.3	23.5	

Vaikuttaako autotyyppi keskimääräiseen kulutukseen?

(Newbold (1995), Statistics for Business and Economics)

Tulos SPSS-ohjelmalla:

Descriptives

MILES

	n	Mean	Std. Deviation	Std. Error	95 % Confidence Interval for Mean		Min.	Max.
					Lower Bound	Upper Bound		
A	7	20.9000	.79162	.29921	20.1679	21.6321	19.90	22.20
B	7	23.2000	.90921	.34365	22.3591	24.0409	22.00	24.60
C	6	22.9167	.84004	.34294	22.0351	23.7982	21.90	24.10
Total	20	22.3100	1.33610	.29876	21.6847	22.9353	19.90	24.60

Test of Homogeneity of Variances

MILES

Levene Statistic	*df*1	*df*2	Sig.
.036	2	17	.965

ANOVA

MILES

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	21.670	2	10.835	15.038	.000
Within Groups	12.248	17	.720		
Total	33.918	19			

Jos yksisuuntaisessa varianssianalyysissä H_0 hylätään ja täten H_1 hyväksytään, niin usein halutaan lisäksi selvittää minkä ryhmien välillä odostusarvot poikkeavat toisistaan. Tämä voidaan tehdä parittaisten luottamusvälien avulla. Muodostetaan tavanomaiset luottamusvälit $(\mu_i - \mu_j)$:lle

$$\overline{X_i} - \overline{X_j} \pm t_{\alpha/2;n_i+n_j-2}\, s_{ij} \sqrt{1/n_i + 1/n_j},$$

missä

$$s_{ij}^2 = \frac{(n_i - 1)s_i^2 + (n_j - 1)s_j^2}{n_i + n_j - 2}$$

(luottamusväli odotusarvojen erotukselle, ks. MTTTP5).

Jos halutaan, että kaikki parittaiset luottamusvälit sisältävät todellisen erotuksen todennäköisyydellä, joka on vähintään $1 - \alpha$, niin voidaan käyttää esim. nk. Bonferronin luottamusväliä

$$\overline{X_i} - \overline{X_j} \pm t_{\alpha^*/2;n-I}\, s \sqrt{1/n_i + 1/n_j},$$

missä

$$s^2 = \frac{SSW}{n - I} = MSW, \qquad \alpha^* = \frac{2\alpha}{I(I-1)}.$$

Esimerkki 2.1.5 Monivertailu esimerkin 2.1.4 tilanteessa.

Multiple Comparisons

Dependent Variable: MILES

Bonferroni

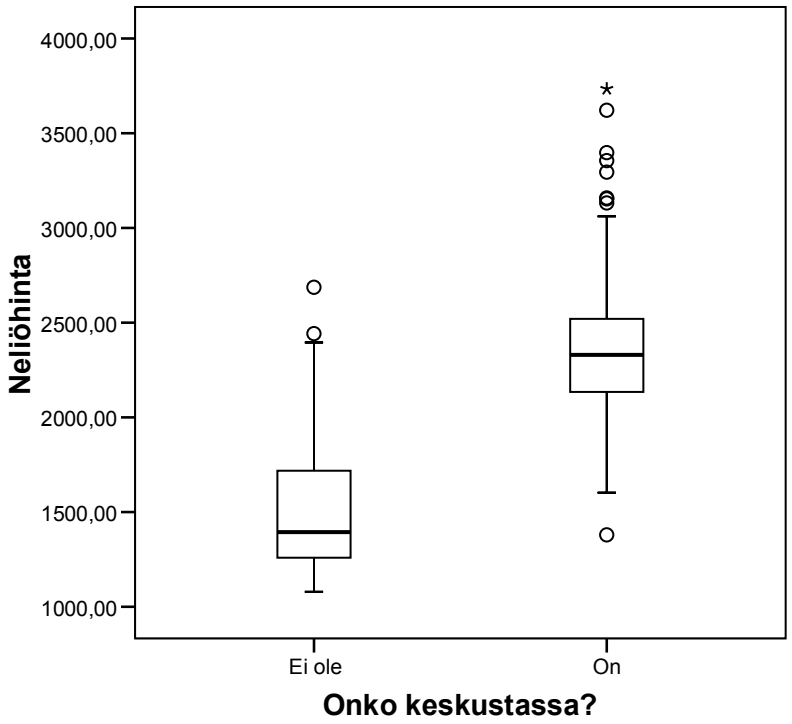
(*I*) AUTO	(*J*) AUTO	Mean Difference (*I* − *J*)	Std. Error	Sig.	95 % Confidence Interval for Mean	
					Lower Bound	Upper Bound
A	B	−2.3000*	.45371	.000	−3.5046	−1.0954
	C	−2.0167*	.47224	.002	−3.2705	−.7629
B	A	2.3000*	.45371	.000	1.0954	3.5048
	C	.2833	.47224	1.000	−.9705	1.5371
C	A	2.0167*	.47224	.002	.7629	3.2705
	B	−.2833	.47224	1.000	−1.5371	.9705

*The mean difference is significant at the .05 level.

Yksisuuntainen varianssianalyysi on kahden populaation tilanteessa identtinen riippumattomien otosten t-testin kanssa. Tällöin $t^2 = F$.

Esimerkki 2.1.6 Tampereella myynnissä olleita kerrostalohuoneistoja, jotka olivat esittelyssä 7. - 14.4.2006). Aineisto http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Asunnot_2006.sav.

a) Asuntojen neliöhinnat keskustassa ja ei-keskustassa (t-testi ja 1-VA).



Group Statistics

Neliöhinta

Onko keskustassa?	N	Mean	Std. Deviation	Std. Error Mean
Ei ole	126	1503.2538	325.34129	28.98371
On	103	2397.6072	408.02462	40.20386

Independent Samples Test

Neliöhinta

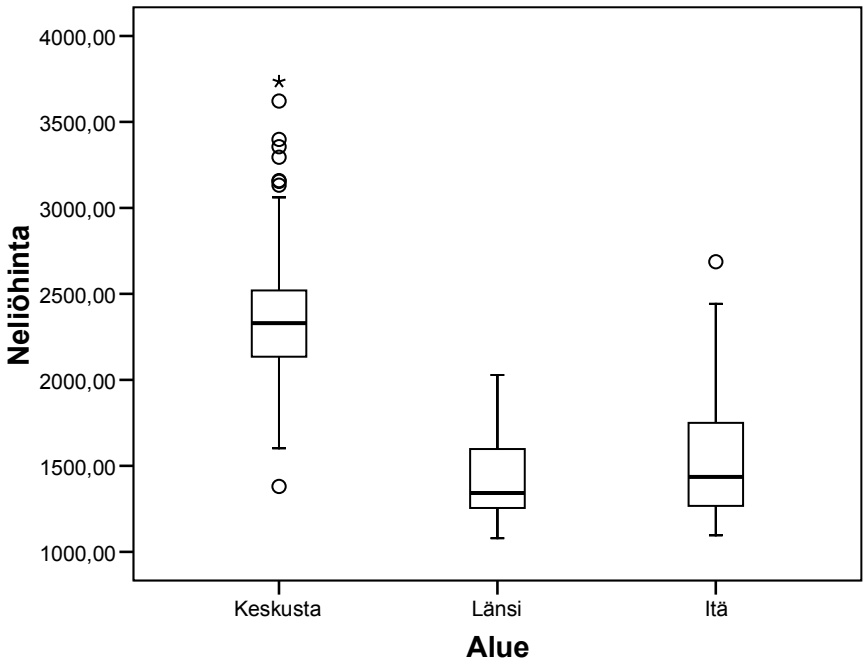
	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95 % Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	1.235	.268	−18.455	227	.000	−894.35342	48.46101	−989.84436	−798.86248
Equal variances not assumed			−18.045	193.029	.000	−894.35342	49.56214	−992.10630	−796.60054

ANOVA

Neliöhinta

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	45330513	1	45330512.598	340.591	.000
Within Groups	30212246	227	133093.595		
Total	75542759	228			

b) Asuntojen neliöhinnat keskusta/länsi/itä (1-VA)



Descriptives

Neliöhinta

	N	Mean	Standard Deviation	Standard Error	95 % Confidence Interval for Mean		Min.	Max.
					Lower Bound	Upper Bound		
Keskusta	103	2397.6072	408.02462	40.20386	2317.8630	2477.3514	1380.00	3734.43
Länsi	34	1414.2870	260.39544	44.65745	1323.4307	1505.1433	1079.75	2028.57
Itä	92	1536.1328	341.69439	35.62410	1465.3699	1606.8957	1096.55	2687.20
Total	229	1905,5176	575.61088	38.03744	1830.5677	1980.4674	1079.75	3734.43

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	45699081	2	22849540.291	173.035	.000
Within Groups	29843678	226	132051.673		
Total	75542759	228			

Varianssianalyysin käyttö edellyttää siis selitettävältä muuttujalta vähintään intervalliasteikollista mittausta (normaalijakaumaoletukset). Selittävälle muuttujalle ei aseteta mitta-asteikon suhteen vaatimuksia. Jos selittävä muuttuja on numeerinen on se tietysti ensin luokiteltava sopivasti.

Ks. varianssianalyysistä http://www.fsd.uta.fi/menetelmaopetus/varianssi/anova.html.

2.1.1 SPSS-ohjeita

Yksisuuntainen varianssianalyysi

Analyze
Compare Means ▶ One-Way ANOVA...

2.2 Kaksisuuntainen varianssianalyysi

Esimerkki 2.2.1 Eräällä kurssilla luennot esitettiin toiselle ryhmälle televisioituina ja toiselle ryhmälle tavalliseen tapaan. Osallistujille tehtiin testi sekä ennen

että jälkeen kurssin. Näiden testipistemäärien erotukset olivat:

Naiset		Miehet				
Tavallinen	TV	Tavallinen		TV	TV	TV
20.3	6.2	12.5	45.4	28.1	29.7	9.3
23.5	15.6	7.8	6.3	−7.8	39.1	1.5
4.7	25.0	21.9	18.8	17.1	9.4	4.7
21.9	4.7	−7.8	9.4	14.1	20.3	15.6
15.6	28.1	−3.1	−3.1	18.8	14.1	26.7
20.3	17.2	3.1		17.2	18.7	12.5
26.6	14.1	29.7		1.5	17.2	10.9
21.9	31.2	18.8		20.3	1.5	32.8
−9.4	12.6	28.1		4.7	25.0	−6.2
4.7	9.4	36.0		15.6	29.7	3.1
−1.6	17.2	4.7		34.4	25.0	28.1
25.0	23.4	−3.3		18.8	23.4	37.5
						20.3

Haluttaessa tutkia vaikuttaako opetustapa oppimiseen, voidaan käyttää t-testiä tai yksisuuntaista varianssianalyysiä. Samoin jos tutkitaan onko sukupuolella vaikutusta oppimiseen. Mielenkiintoisempaa lienee kuitenkin sen selvittäminen, miten opetustapa ja sukupuoli yhdessä vaikuttavat oppimiseen. Tällöin selitetään numeerista muuttujaa kahdella luokittelutason muuttujalla. Analysointi voidaan suorittaa *kaksisuuntaisella varianssianalyysillä*.

Usein varianssianalyysin yhteydessä selittäviä muuttujia kutsutaan faktoreiksi (A ja B) ja niiden luokkia tasoiksi. Faktorin B vaikutus selitettävään muuttujaan saattaa olla erilaista A:n eri tasoilla. Tällöin sanotaan, että A:lla ja B:llä on *yhdysvaikutusta eli interaktiota*.

Kaksisuuntaisen varianssianalyysin avulla pyritään selvittämään:

1. Onko A:lla B:stä riippumatonta vaikutusta selitettävään eli onko A:lla omavaikutusta?
2. Onko B:llä A:sta riippumatonta vaikutusta selitettävään eli onko B:llä omavaikutusta?
3. Onko A:lla ja B:llä yhdysvaikutusta?

Esim. Onko opetustavan vaikutus pistemäärään erilaista naisilla ja miehillä?

Kohtiin 1–3 liittyy jokaiseen oma F-testisuureensa. Mielenkiintoisin tutkittava on tietysti yhdysvaikutus.

F-testisuureet määritellään samaan tapaan kuin yksisuuntaisessa varianssianalyysissä neliösummien avulla.

Merkitään: SS_A on A:n omavaikutukseen liittyvä neliösumma, SS_B on B:n omavaikutukseen liittyvä neliösumma, SS_{AB} on A:n ja B:n yhdysvaikutukseen liittyvä

neliösumma ja *SSE* jäännösneliösumma. Näihin neliösummiin perustuen määritellään testisuureet.

Testaukset kaksisuuntaisessa varianssianalyysissä:

1. H_0: A:lla ei ole omavaikutusta,
   H_1: A:lla on omavaikutusta.
   Jos H_0 tosi, niin
   $$F_A = \frac{MS_A}{MSE} \sim F_{df_A, df_{SSE}},$$
   missä
   $$MS_A = \frac{SS_A}{df_A} \qquad \text{ja} \qquad MSE = \frac{SSE}{df_{SSE}}.$$
   (MS-neliösummat saadaan siis, kun jaetaan neliösummat vapausasteillaan, joiden määrittäminen kuten neliösummien laskukin jätetään ohjelmiston tehtäväksi!)
   Nyt H_0 hylätään riskitasolla α, jos otoksen perusteella laskettu F_A:n arvo $> F_{\alpha;df_A,df_{SSE}}$.
2. H_0: B:llä ei ole omavaikutusta,
   H_1: B:llä on omavaikutusta.
   Jos H_0 tosi, niin
   $$F_B = \frac{MS_B}{MSE} \sim F_{df_B, df_{SSE}},$$
   missä
   $$MS_B = \frac{SS_B}{df_B} \qquad \text{ja} \qquad MSE = \frac{SSE}{df_{SSE}}.$$
   Nyt H_0 hylätään riskitasolla α, jos otoksen perusteella laskettu F_B:n arvo $> F_{\alpha;df_B,df_{SSE}}$.
3. H_0: A:lla ja B:llä ei yhdysvaikutusta,
   H_1: A:lla ja B:llä on yhdysvaikutusta.
   Jos H_0 tosi, niin
   $$F_{AB} = \frac{MS_{AB}}{MSE} \sim F_{df_{AB}, df_{SSE}},$$
   missä
   $$MS_{AB} = \frac{SS_{AB}}{df_{AB}} \qquad \text{ja} \qquad MSE = \frac{SSE}{df_{SSE}}.$$
   Nyt H_0 hylätään riskitasolla α, jos otoksen perusteella laskettu F_{AB}:n arvo $> F_{\alpha;df_{AB},df_{SSE}}$.

Esimerkki 2.2.2 Esimerkin 2.2.1 tilanteessa kaksisuuntainen varianssianalyysi. Aineisto http://www.sis.uta.fi/tilasto/tiltp3/kevat2004/Aineistoja/OPETUS.SAV.

Ehdolliset keskiarvot

Report

PISTEET

Sukupuoli	Opetustapa	Mean	*N*	Std. Deviation
Nainen	Tavallinen	14,4583	12	11,82505
	TV	17,0583	12	8,44915
	Total	15,7583	24	10,13813
Mies	Tavallinen	13,2471	17	15,20671
	TV	17,1000	37	11,66660
	Total	15,8870	54	12,86560

Kaksisuuntainen varianssianalyysi

Tests of Between-Subjects Effects

Dependent Variable: PISTE

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	213,754*	3	71,251	,483	,695
Intercept	15155,879	1	15155,879	102,674	,000
sukupuol	5,417	1	5,417	,037	,849
opetust	164,901	1	164,901	1,117	,294
sukupuol · opetust	6,217	1	6,217	,042	,838
Error	10923,261	74	147,612		
Total	30726,030	78			
Corrected Total	11137,014	77			

*R squared = ,019 (Adjusted R Squared = −,021)

Sukupuolella ei omavaikutusta (p=0,849), opetustavalla ei omavaikutusta (p=0,294), ei yhdysvaikutusta (p=0,838).

Yksisuuntainen varianssianalyysi, selittäjänä opetustapa

ANOVA

PISTE

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	203,417	1	203,417	1,414	,238
Within Groups	10933,597	76	143,863		
Total	11137,014	77			

Yksisuuntainen varianssianalyysi, selittäjänä sukupuoli

ANOVA

PISTE

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	,275	1	,275	,002	,966
Within Groups	11136,739	76	146,536		
Total	11137,014	77			

Ks. varianssianalyysistä http://www.fsd.uta.fi/menetelmaopetus/varianssi/anova.html.

2.2.1 SPSS-ohjeita

Kaksisuuntainen varianssianalyysi

Analyze

General Linear Model ▶ Univariate...

Luku 3

χ^2-yhteensopivuus- ja riippumattomuustestit

Tutustutaan ensin jakaumaan, jota noudattavaa testisuuretta yhteensopivuus- ja riippumattomuustestien yhteydessä tullaan käyttämään.

Olkoon $Z_1, Z_2, \ldots, Z_k$ riippumattomia satunnaismuuttujia siten, että kukin $Z_i \sim N(0,1)$. Tällöin $Z_1^2 + Z_2^2 + \cdots + Z_k^2$ noudattaa nk. χ^2-jakaumaa vapausastein k. Merkitään χ_k^2. Voidaan osoittaa, että $E(\chi_k^2) = k$ ja $\mathrm{Var}(\chi_k^2) = 2k$.

χ^2-jakauman jakauman tiheysfunktion muoto määräytyy vapausasteiden perusteella (ks. http://onlinestatbook.com/2/chi_square/distribution.html). Huomataan siis, että χ^2-jakauma ei ole symmetrinen ja että χ^2-jakautunut satunnaismuuttuja saa arvokseen ei-negatiivisia reaalilukuja.

Olkoon χ^2-jakaumaa vapausastein k noudattava satunnaismuuttuja χ_k^2.

Määritellään luku $\chi_{\alpha;k}^2$ siten että $P(\chi_k^2 \geq \chi_{\alpha;k}^2) = \alpha$. Näitä arvoja on taulukoitu muutamilla α:n arvoilla ja eri vapausastein, ks. http://www.sis.uta.fi/tilasto/mttta1/kevat2019/chi.pdf.

3.1 χ^2-yhteensopivuustesti

χ^2-yhteensopivuustestin avulla voidaan testata sitä, onko satunnaisotos peräisin tietystä jakaumasta. Tässä siis hypoteesina väite jakaumasta, ei ainostaan jostain sen parametrista, kuten tähän asti esillä olleissa hypoteeseissa on ollut.

Esimerkki 3.1.1 H_0: Otos peräisin diskr. tasajakaumasta; H_1: otos ei ole peräisin ko. jakaumasta. H_0: Otos peräisin normaalijakaumasta; H_1: otos ei ole peräisin normaalijakaumasta.

Olkoon n alkion satunnaisotoksen muuttuja luokiteltu siten, että luokkien lukumäärä on k. Olkoon lisäksi näiden luokkien frekvenssit $f_1, f_2, \ldots, f_k$. Testataan sitä, onko havaitut frekvenssit sopusoinnussa nk. teoreettisten eli odotettujen frekvenssien $e_1, e_2, \ldots, e_k$ kanssa. Teoreettisen frekvenssit määrätään sen perusteella, mistä jakaumasta ajattelemme otoksen olevan peräisin.

Nyt H_0: Otos peräisin tietyllä tavalla jakautuneesta populaatiosta. Jos H_0 tosi, niin

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

ja H_0 hylätään riskitasolla α, jos otoksen perusteella laskettu χ^2:n arvo

$$\chi^2_{\text{havaittu}} > \chi^2_{\alpha;k-1}.$$

χ^2-yhteensopivuustestiä voidaan käyttää, jos kaikki teoreettiset frekvenssit ovat > 1 ja enintään 20 % < 5.

Esimerkki 3.1.2 Yhtiö tietää aikaisempien vuosien perusteella, että talven loputtua 80 % sen asiakkaista on maksanut laskunsa ajoissa, 10 % kuukauden myöhässä, 6 % 2 kuukautta myöhässä ja 4 % enemmän kuin kaksi kuukautta myöhässä. Viimeisimmän talven loputtua tehdään 400 lähetetyn laskun satunnaisotos, jossa ajallaan maksaneita on 287, 49 kuukauden myöhässä, 30 kaksi kuukautta myöhässä ja 34 enemmän kuin kaksi kuukautta myöhässä. Onko tämän perusteella epäiltävissä, että asiakkaiden laskujen maksutavoissa on muutosta aiempiin vuosiin? (Newbold (1995), Statistics for Business and Economics)

H_0: ei muutosta. Lasketaan χ^2-yhteensopivuustestisuure.

f_i	e_i
287	$0{,}8 \times 400 = 320$
49	$0{,}1 \times 400 = 40$
30	$0{,}06 \times 400 = 24$
34	$0{,}04 \times 400 = 16$
400	

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{4} \frac{(f_i - e_i)^2}{e_i} \\
&= \frac{(287-320)^2}{320} + \frac{(49-40)^2}{40} + \frac{(30-24)^2}{24} + \frac{(34-16)^2}{16} \\
&\approx 27{,}58 > \chi^2_{0{,}005;3} = 12{,}84
\end{aligned}$$

Voidaan siis päätellä, että on tapahtunut muutosta.

Yhteensopivuustestin yhteydessä joudutaan teoreettisia frekvenssejä laskettaessa usein estimoimaan jakauman parametrit. Tällöin käytetyn testisuureen jakauman vapausasteet vähenevät estimoitujen parametrien määrällä.

Esimerkki 3.1.3 Jos halutaan tutkia onko otos peräisin normaalijakaumasta, on aluksi estimoitavana kaksi parametria (odotusarvo ja varianssi). Tehtiin 1000

alkion satunnaisotos ja saatiin otoskeskiarvoksi 50 ja keskihajonnaksi 10. Muodostetaan luokiteltu jakauma otoksen perusteella siten, että yksi luokka on $(40, 50)$. Mikä on tämän luokan teoreettinen frekvenssi?

$$H_0\text{: otos peräisin } N(50, 100)\text{:sta.}$$

Luokan $(40, 50)$ teoreettinen frekvenssi saadaan laskemalla H_0:n mukaisessa tilanteessa vastaava todennäköisyys

$$P(40 \leq X \leq 50) = \Phi\left(\frac{50-50}{10}\right) - \Phi\left(\frac{40-50}{10}\right) = \cdots = 0{,}3413,$$

joten $e_i = 0{,}3413 \times 1000 \approx 341{,}3$.

Esimerkki 3.1.4 Eräällä tilastotieteen kurssilla ilmoittautumisen yhteydessä "heitettiin noppaa" siten, että lomakkeessa oli kysymys: *"Kuvittele heittäväsi noppaa. Heittosi tulos on __"*

Silmäluvun jakaumaksi saatiin:

silmäluku	frekvenssi	%
1	8	6.6
2	5	4.1
3	17	13.9
4	27	22.1
5	26	21.3
6	39	32.0
	122	

Testataan tapahtuiko heittäminen satunnaisesti.

$$H_0\text{: otos peräisin Tasd}(1, 6)\text{:sta.}$$

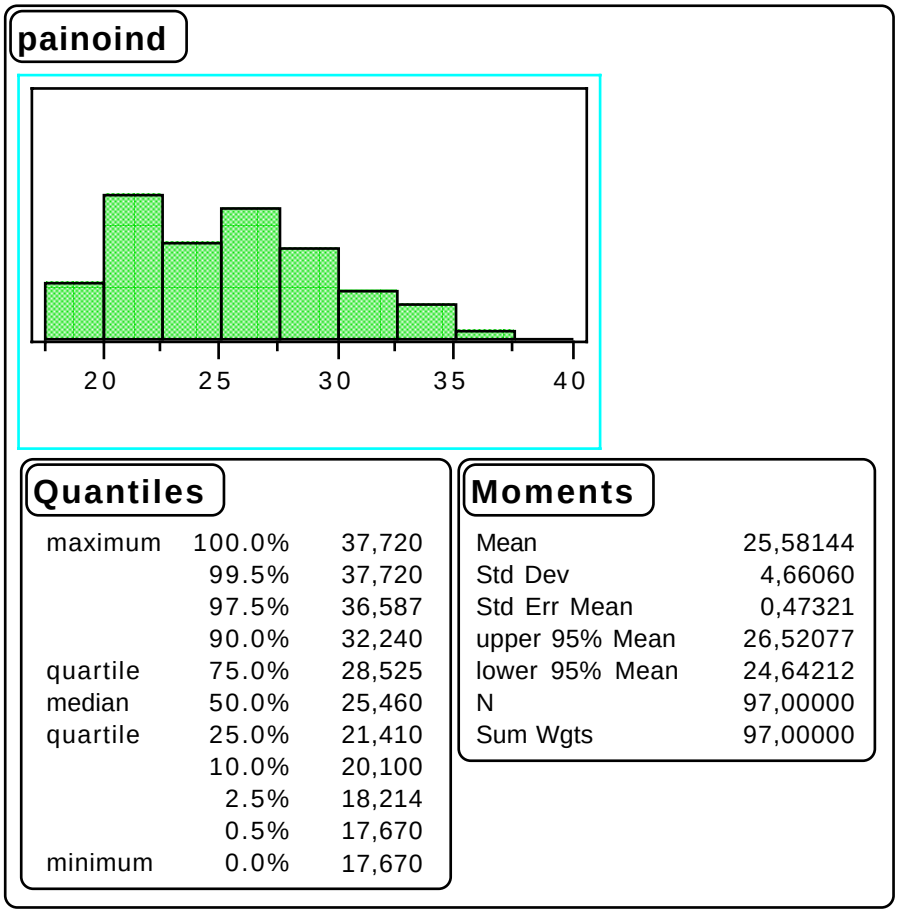
Jos H_0 on tosi, niin kaikkia silmälukuja tulisi olla saman verran eli $122/6 = 20{,}3$.

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{6} \frac{(f_i - e_i)^2}{e_i} = \frac{(8-20{,}3)^2}{20{,}3} + \cdots + \frac{(39-20{,}3)^2}{20{,}3} \\ &\approx 40{,}6 > \chi^2_{0{,}005;5} = 16{,}75,\end{aligned}$$

joten nopanheitto ei ole tapahtunut satunnaisesti.

Esimerkki 3.1.5 Onko painoindeksi normaalisti jakautunut?

$$H_0\text{: otos peräisin } N(25{,}58; 4{,}66^2)\text{:sta.}$$



Kuva 3.1: Esimerkin 3.1.5 painoindeksin jakauma ja tunnuslukuja.

Lasketaan χ^2-yhteensopivuustestisuure.

Painoindeksi	frekvenssi	odotettu frekv.
alle 20,1	9	11,5
20,1–21,4	15	6,3
21,4–25,5	26	30,0
25,5–28,5	23	23,6
28,5–32,2	15	18,1
yli 32,2	9	7,5
	97	97,0

Esimerkiksi 1. luokan teoreettinen frekvenssi saadaan laskemalla H_0:n mukaisessa tilanteessa vastaava todennäköisyys

$$P(X \leq 20{,}1) = \Phi\left(\frac{20{,}1 - 25{,}58}{4{,}66}\right) = 1 - \Phi(1{,}18) = 0{,}119,$$

joten $e_1 = 0{,}119 \times 97 \approx 11{,}5$

$$\chi^2 = \sum_{i=1}^{6} \frac{(f_i - e_i)^2}{e_i} = \frac{(9-11{,}5)^2}{11{,}5} + \cdots + \frac{(9-7{,}5)^2}{7{,}5} \approx 13{,}94.$$

Koska on estimoitu 2 parametria (odotusarvo ja varianssi), niin vapausasteet ovat $6-2-1=3$. Koska $\chi^2_{0.005;3} = 12{,}84$ ja $\chi^2_{0.001;3} = 16{,}27$ niin $0{,}001 < p < 0{,}005$. Päättelemme, että otos ei ole peräisin normaalijakaumasta.

3.1.1 SPSS-ohjeita

χ^2 -yhteensopivuustesti

Analyze
Nonparametric Tests ▶ Chi-Square...

3.2 χ^2-riippumattomuustesti

Ristiintaulukoiden perusteella voitiin tutkia muuttujien välistä riippuvuutta vertailemalla selitettävän muuttujan ehdollisia prosenttijakauma. Tällöin

$$H_0\colon X \text{ ja } Y \text{ ovat riippumattomia,}$$
$$H_1\colon X \text{ ja } Y \text{ ovat riippuvia.}$$

χ^2-riippumattomuustestillä voidaan testata asetettua nollahypoteesia käyttäen perustana ristiintaulukkoa.

Olkoon muodostettu ristiintaulukko

		x				
		1	2	...	J	
y	1	f_{11}	f_{12}	...	f_{1J}	$f_{1\cdot}$
	2	f_{21}	f_{22}	...	f_{2J}	$f_{2\cdot}$
	⋮	⋮	⋮		⋮	⋮
	I	f_{I1}	f_{I2}	...	f_{IJ}	$f_{I\cdot}$
		$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot J}$	n

Jos H_0 on tosi, niin

$$\frac{e_{ij}}{f_{\cdot j}} = \frac{f_{i\cdot}}{n} \qquad \text{eli} \qquad e_{ij} = \frac{f_{i\cdot} f_{\cdot j}}{n}.$$

Lisäksi kun H_0 on tosi, niin

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

Nyt H_0 hylätään riskitasolla α, jos otoksen perusteella laskettu χ^2:n arvo

$$\chi^2_{\text{havaittu}} > \chi^2_{\alpha;(I-1)(J-1)}.$$

Jos molemmat muuttuja on luokiteltu kahteen luokkaan (kyse nelikentästä), niin testisuure voidaan laskea

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

Riippumattomuustestin yhteydessä ei tarvita siis populaatioon liittyviä jakaumaoletuksia, kuten esimerkiksi varianssianalyysin yhteydessä tehtiin. Riippumattomuustestiä voidaan siis käyttää jo luokitteluasteikollisten muuttujien yhteydessä.

Kuitenkin, jotta χ^2-riippumattomuustestiä voidaan käyttää, on 1) $df > 1$, kaikkien teoreettisten frekvenssien oltava > 1 sekä enintään 20 % saa olla < 5; 2) $df = 1$, jos $n > 40$ testin käyttö sallittu, jos $20 \leq n \leq 40$ kaikkien teoreettisten frekvenssien oltava ≥ 5.

Jos edellä esitetyt vaatimukset eivät ole täytetty, voidaan koettaa luokituksia muuttumalla saada oletukset kuntoon.

Esimerkki 3.2.1 Erään kurssin arviointiin liittyvässä kyselyssä oli mm. seuraavat kysymykset:

A. Taustani
Pääaineeni on (ympyröi numero)

1. matematiikka tai tilastotiede
2. kansantaloustiede
3. tietojenkäsittelyoppi
4. jokin muu.

B. Kurssin arviointi
Tämä kurssi on mielestäni

työläs	1	2	3	4	5	vähätöinen
vaikea	1	2	3	4	5	helppo

Odotin kurssin olevan

työläämpi	1	2	3	4	5	vähätöisempi
vaikeampi	1	2	3	4	5	helpompi

Vastausten perusteella saatiin seuraavat ristiintaulukot:

		pääaine			
		kans.	mat. & til.	tko	
kurssin vaikeus	1–2	23	15	13	51
	3	6	15	10	31
	4–5	1	9	1	11
		30	39	24	93

		pääaine			
		kans.	mat. & til.	tko	
odotettu vaikeus	1–2	3	6	7	16
	3	18	26	9	53
	4–5	8	7	8	23
		29	39	24	92

		pääaine		
		kans.	mat. & til.	tko
odotettu vaikeus	1–2	10.34 %	15.38 %	29.17 %
	3	62.07 %	66.67 %	37.50 %
	4–5	27.59 %	17.95 %	33.33 %

Odotetut frekvenssit toiseen ristiintaulukkoon liittyen ovat:

		pääaine		
		kans.	mat. & til.	tko
odotettu vaikeus	1–2	5.04	6.78	4.17
	3	16.71	22.47	13.83
	4–5	7.25	9.75	6.00

Test	ChiSquare	Prob > ChiSq
Pearson	6.692	0.1531

Onko pääaineella vaikutusta siihen, kuinka vaikeana piti opintojaksoa?

	kans.	mat. & til.	tko
vaikea (1–2)	23 (16,5)	15 (21,4)	13 (13,2)
sopiva (3)	6 (10,0)	15 (13,0)	10 (8,0)
helppo (4–5)	1 (3,5)	9 (4,6)	1 (2,8)

Koska odotetuista frekvensseistä 33 % on alle 5, eivät testin oletukset ole voimassa. Muodostetaan uusi ristiintaulukko:

	kans.	mat. & til.	tko
vaikea (1–2)	23 (16,5)	15 (21,4)	13 (13,2)
sopiva tai helppo (3–5)	7 (13,5)	24 (17,6)	11 (10,8)

Lasketaan χ^2-riippumattomuustestisuure.

$$\chi^2 = \frac{(23-16{,}5)^2}{16{,}5} + \cdots + \frac{(11-10{,}8)^2}{10{,}8} \approx 9{,}94 > \chi^2_{0{,}01;2} = 9{,}21$$

H_0: ei riippuvuutta, hylätään 1 %:n riskitasolla (mutta ei 0,5 %). Voidaan päätellä, että eri koulutusohjelmien opiskelijoiden mielipiteet kurssin vaikeudesta ovat erilaiset. Kansantaloustieteilijöistä 76,7 % piti kurssia vaikeana, kun taas vastaava luku matematiikan ja tilastotieteen koulutusohjelmassa oli 38,5 %.

Esimerkki 3.2.2 Erään tilastotieteen tentin tulos pääaineittain. (odotetut frekvenssit suluissa).

	kans.	mat. & til.	tko	Yht.
Hylätty	13 (14,8) 33.33 %	22 (22,0) 37.93 %	14 (12,2) 43.75 %	49
Hyväksytty	26 (24,2) 66.67 %	36 (36,0) 62.07 %	18 (19,8) 56.25 %	80
Yhteensä	39	58	32	129

Lasketaan χ^2-riippumattomuustestisuure.

$$\chi^2 = \frac{(13-14{,}8)^2}{14{,}8} + \cdots + \frac{(18-19{,}88)^2}{19{,}8} \approx 0{,}81 < \chi^2_{0{,}05;2} = 5{,}99$$

H_0: ei riippuvuutta, hyväksytään.

Esimerkki 3.2.3

	Miehet	Naiset	Yhteensä
Hylätty	34	15	49
Hyväksytty	59	23	82
Yhteensä	93	38	131

$$\chi^2 = \frac{(34 \cdot 23 - 59 \cdot 15)^2 \cdot 131}{93 \cdot 38 \cdot 49 \cdot 82} \approx 0{,}09787,$$

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

3.2.1 SPSS-ohjeita

Ristiintaulukot ja χ^2

Analyze
Descriptive Statistics ▶ Crosstabs...
→ Statistics... Chi-square

Luku 4

Regressioanalyysi

Regressioanalyysillä tutkitaan jonkin muuttujan y riippuvuutta joukosta muita muuttujia $x_1, x_2, \ldots, x_k$.

Regressioanalyysin yhteydessä y:n riippuvuuden muuttujista $x_1, x_2, \ldots, x_k$ ajatellaan 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, satunnaisvirhe (ei havaittavissa oleva); $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ ovat mallin tuntemattomat parametrit, jotka aineiston perusteella ovat estimoitavissa.

4.1 Yksi selittävä muuttuja

Esimerkki 4.1.1 Nuoresta metsiköstä, jossa oli samanikäisiä puita, poimittiin arpomalla 10 puuta. Näistä puista mitattiin kuutiomäärät (y) ja poikkileikkauspinta-alat (x).

puu	pinta-ala (dm^2)	tilavuus (m^3)
1	2.59	0.161
2	3.89	0.273
3	4.60	0.309
4	5.22	0.338
5	5.75	0.398
6	5.89	0.401
7	6.30	0.426
8	7.03	0.459
9	8.28	0.549
10	9.63	0.633

Pisteparvesta (ks. Esim. 4.1.3) huomataan, että riippuvuus näyttää hyvin lineaariselta.

Tämän esimerkin tilanteessa pisteparveen voidaan sovittaa suora, jonka ympärille pisteiden ajatellaan ryhmittyneen. Tällöin y:n riippuvuuden x:stä ajatellaan olevan muotoa

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon,$$

missä β_0 ja β_1 ovat mallin parametrit sekä ε satunnaisvirhe. Mallissa ajatellaan siis satunnaismuuttujan Y:n muodostuvan x:n avulla selitettävästä osasta $\beta_0+\beta_1 x_1$ sekä satunnaisvaihtelusta ε. Regressioanalyysissä halutaankin estimoida β_0 ja β_1 havaitun aineiston perusteella. Näin tehtäessä siis (ed. malliin liittyen) estimoidaan suora, jonka ajatellaan kuvaavan y:n riippuvuutta x:stä.

Jos oletetaan, että edellä esitetystä yhden selittäjän regressiomallista on tehty havaintoja n kertaan selittävien muuttujien eri arvoilla, niin malli voidan kirjoittaa muodossa

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

Lisäksi regressiomallissa oletetaan, että $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$ sekä ε_i:t toisistaan riippumattomiksi. Tästä seuraa, että edellä esitetyn mallin tilanteessa $Y_i \sim N(\beta_0 + \beta_1 x_i,\, \sigma^2)$. Tämä tarkoittaa siis sitä, että jokaista x:n arvoa kohti on olemassa Y:n todennäköisyysjakauma. Havainnot ovat otoksia näistä jakaumista.

Esimerkki 4.1.2 Regressiomalli graafisesti. Ks. http://www.sis.uta.fi/tilasto/tiltp3/kevat2004/esim_4_1_2.pdf.

Havaintojen $y_1, y_2, \ldots, y_n$ ja $x_1, x_2, \ldots, x_n$ perusteella mallin parametrit voidaan estimoida (käyttäen kriteerinä sitä, että sovitettava suora on keskimäärin mahdollisimman lähellä kaikkia pisteitä, kyse pienimmän neliösumman estimoinnista, PNS-estimointi) seuraavalla tavalla:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}.$$

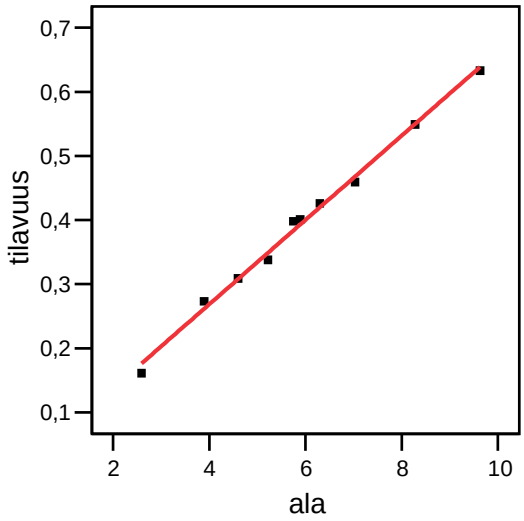
Näin saadaan regressiosuora (estimoitu)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, 2, \ldots, n,$$

missä $\hat{\beta}_0$ on estimoitu β_0 eli estimoitu vakiokerroin ja $\hat{\beta}_1$ on estimoitu β_1 eli estimoitu x:n regressiokerroin.

Estimoidusta mallista voidaan laskea estimoidut y:n arvot ja verrata niitä havaittuihin. Laskemalla erotukset $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$ saadaan residuaalit. PNS-estimoinnissa määrätään estimoidut mallin parametrit niin, että neliösumma $\sum e_i^2$ on mahdollisimman pieni.

Esimerkki 4.1.3 Esimerkin 4.1.1 aineistosta estimointitulokset.



Linear Fit:

Summary of Fit

R Square	0,994589
Root Mean Square Error	0,010588
Mean of Response	0,3947
Observations (or Sum Wgts)	10

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	F-ratio	Prob > F
Model	1	0,16484919	0,164849	1470,377	0,0000
Error	8	0,00089691	0,000112		
C Total	9	0,16574610			

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob > $\|t\|$
Intercept	0,0058446	0,01068	0,55	0,5991
ala	0,0657072	0,00171	38,35	0,0000

Esimerkki 4.1.4 Olkoon $y =$ satomäärä, $x =$ lannoitemäärä. Estimodaan regressiosuora oheisesta aineistosta. Lisäksi on laskettu neliösummia ja F-testisuure, jotka esitellään myöhemmin, s. 38.

(Liski & Puntanen, Tilastotieteen peruskurssi II)

x_i	y_i	$x_i y_i$	x_i^2	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
100	40	4000	10000	39,64	0,36
200	45	9000	40000	46,43	−1,43
300	50	15000	90000	53,21	−3,21
400	65	26000	160000	60,00	5,00
500	70	35000	250000	66,79	3,21
600	70	42000	360000	73,57	−3,57
700	80	56000	490000	80,36	−0,36
2800	420	187000	1400000		
$\overline{x} = 400$	$\overline{y} = 60$				

$$\hat{\beta}_1 = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/7}{\sum x_i^2 - (\sum x_i)^2/7} = \frac{187000 - 2800 \times 420/7}{1400000 - 2800^2/7} \approx 0{,}06786$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 420/7 - 0{,}06786 \times 2800/7 \approx 32{,}857$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 32{,}857 + 0{,}06786 x_i, \qquad i = 1, \ldots, 7$$

$$e_i = y_i - \hat{y}_i = y_i - (32{,}857 + 0{,}06786 x_i), \qquad i = 1, \ldots, 7$$

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \approx 60{,}7$$

$$SST = \sum (y_i - \bar{y})^2 = 1350{,}0$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \approx 1289{,}286$$

$$R^2 = SSR/SST = 0{,}955$$

$$MSR = SSR/1, \qquad MSE = SSE/(7-2) = 12{,}143$$

$$F = MSR/MSE = 106{,}176 > F_{0{,}01;1,5} = 16{,}26.$$

On osoitettavissa, että

$$E(\hat{\beta}_1) = \beta_1 \qquad \text{ja} \qquad E(\hat{\beta}_0) = \beta_0.$$

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 aluksi sitä onko x merkitsevä selittäjä. Tällöin testattavana hypoteesina on

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

Jos H_0 on tosi, niin

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-2},$$

missä

$$s(\hat{\beta}_1) = \sqrt{\frac{MSE}{SS_x}}$$

on β_1:n estimoitu hajonta. Nyt H_0 hylätään riskitasolla α, jos aineiston perusteella laskettu $|t_{\text{hav.}}| > t_{\alpha/2;n-2}$.

Jos x on todettu merkitseväksi selittäjäksi, niin seuraavaksi tutkitaan, onko vakiokertoimen β_0 syytä olla mallissa.

Tällöin

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

Jos H_0 on tosi, niin

$$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \sim t_{n-2},$$

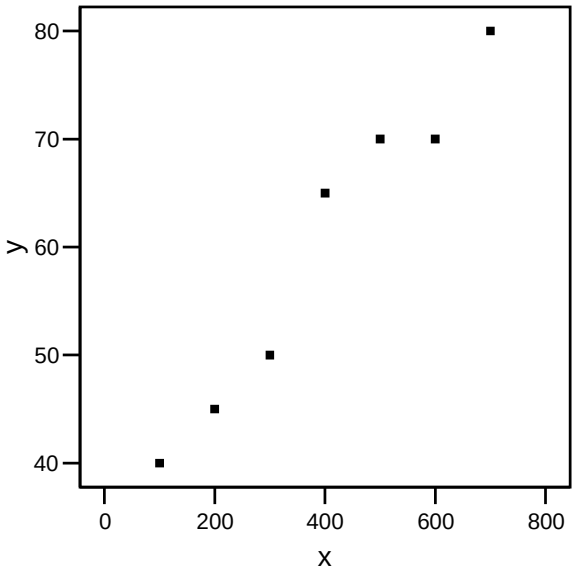
missä

$$s(\hat{\beta}_0) = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)}$$

on β_0:n estimoitu hajonta. Nyt H_0 hylätään riskitasolla α, jos aineiston perusteella laskettu $|t_{\text{hav.}}| > t_{\alpha/2;n-2}$.

Jos on todettu x merkitseväksi selittäjäksi, mutta edellä H_0 on tullut hyväksytyksi, niin silloin uutena mallina onkin $Y = \beta x + \varepsilon$, joka voidaan estimoida. Tässä tapauksessa $\hat{\beta} = \sum x_i y_i / \sum x_i^2$.

Esimerkki 4.1.5 Sadon (y) riippuvuus lannoitemäärästä (x), aineisto esimerkissä 4.1.4.



Pisteparven perusteella lineaarista riippuvuutta. Suoritetaan regressioanalyysi selittäen satoa lannoitemäärällä ja saadaan seuraavat tulokset:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977*	.955	.946	3.48466

*Predictors: (Constant), x

Selitysprosentti $100 \cdot (R \text{ Square}) = 95{,}5\ \%$. Yhden selittäjän tilanteessa sama kuin $100r^2$, määritellään myöhemmin.

ANOVA*

Model		Sum of Squares	*df*	Mean Square	F	Sig.
1	Regression	1289.286	1	1289.286	106.176	.000†
	Residual	60.714	5	12.143		
	Total	1350.000	6			

*Dependent Variable: y
†Predictors: (Constant), x

Taulukossa neliösummat ja niiden vapausasteet, keskineliösummat ja F-testisuure (H_0: $\beta_1 = 0$). Ks. laskukaavat kaavakokoelmassa http://www.sis.uta.fi/tilasto/mttta1/kevat2019/kaavat.pdf, kaavat (3.8), (3.12), (3.13) ja (3.15), myös s. 38.

Coefficients*

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	32.857	2.945		11.157	.000
	x	6.786E−02	.007	.977	10.304	.000

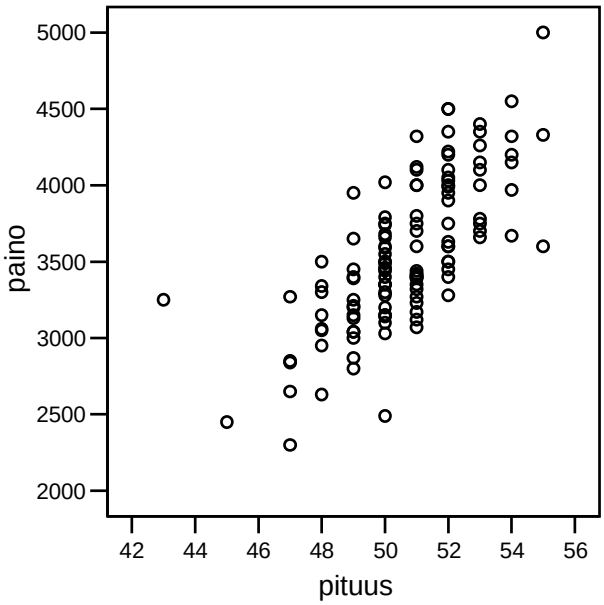
*Dependent Variable: y

Estimoitu lisäaineet regressiokerroin 0,06786 ja vakiokerroin 32,857. Lisäaine on merkittävä selittäjä, koska testattaessa hypoteesia H_0: $\beta_1 = 0$ päädytään sen hylkäämiseen (joko F-testin tai t-testin ($t = 10{,}304$) perusteella).

x	y	estimoitu y	residuaalit
100	40	39.64	.36
200	45	46.43	−1.43
300	50	53.21	−3.21
400	65	60.00	5.00
500	70	66.79	3.21
600	70	73.57	−3.57
700	80	80.36	−.36

Esimerkki 4.1.6 Esimerkkejä regressioanalyysistä.

a) Lapsen syntymäpainon riippuvuus pituudesta.



Paino näyttäisi riippuvan lineaarisesti pituudesta $r = 0.72$. Ks. korrelaatiokertoimen testaus kaavakokoelmassa http://www.sis.uta.fi/tilasto/mttta1/kevat2019/kaavat.pdf, kaava (1.4), myös s. 39.

Correlations

		PAINO	PITUUS
PAINO	Pearson Correlation	1	.720*
	Sig. (2-tailed)	.	.000
	N	120	120
PITUUS	Pearson Correlation	.720*	1
	Sig. (2-tailed)	.000	.
	N	120	120

*Correlation is significant at the 0.01 level

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.720*	.518	.514	339.132

*Predictors: (Constant), PITUUS

Selitysprosentti $100 \cdot (R \text{ square}) = 51{,}8\ \%$. Yhden selittäjän tilanteessa sama kuin $100r^2 = 100 \cdot 0{,}72^2$.

ANOVA*

Model		Sum of Squares	*df*	Mean Square	*F*	Sig.
1	Regression	14573162	1	14573162.20	126.711	.000†
	Residual	13571250	118	115010.596		
	Total	28144413	119			

*Dependent Variable: PAINO
†Predictors: (Constant), PITUUS

Coefficients*

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
		B	Std. Error	Beta		
1	(Constant)	−5211.574	779.297		−6.688	.000
	PITUUS	173.142	15.381	.720	11.257	.000

*Dependent Variable: PAINO

Estimoitu pituuden regressiokerroin 173,142 ja vakiokerroin −5211,574. Pituus on merkittävä selittäjä, koska testattaessa hypoteesia H_0: $\beta_1 = 0$ päädytään sen hylkäämiseen (joko F-testin (F = 126,711) tai t-testin (t = 11,257) perusteella. Yhden sentin lisäys pituudessa kohottaa painoa keskimäärin 173,142 g.

b) Veden pehmeysarvon riippuvuus lisäaineesta.

Model Summary

Model	*R*	*R* Square	Adjusted *R* Square	Std. Error of the Estimate
1	.909*	.826	.801	.33794

*Predictors: (Constant), Lisäaine

ANOVA*

Model		Sum of Squares	*df*	Mean Square	*F*	Sig.
1	Regression	3.796	1	3.796	33.241	.001†
	Residual	.799	7	.114		
	Total	4.596	8			

*Dependent Variable: Veden pehmeys
†Predictors: (Constant), Lisäaine

Coefficients*

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.118	.246		33.014	.000
	Lisäaine	.354	.061	.909	5.765	.001

*Dependent Variable: Veden pehmeys

Esimerkki 4.1.7 Mittayksikön vaikutus estimointitulokseen.

Malli $Y = \beta_0 + \beta_1 x + \varepsilon$.

$$\hat{\beta}_1 = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{SP_{xy}}{SS_x} = r_{xy}\frac{s_y}{s_x}.$$

Jos yhden selittäjän regressioanalyysissä tehdään muunnokset $z = ax + b$ ja $w = cy + d$, niin $r_{zw} = r_{xy}$, jos $ac > 0$ ja $r_{zw} = -r_{ry}$, jos $ac < 0$, $s_z = |a|s_x$ ja $s_w = |c|s_y$. Regressiokerroin on siis riippuvainen muuttujien mittayksiköistä. Kokeile esim. SAIDIT-aineistossa muuttamalla esimerkin 4.1.6 a) mittayksiköt kiloiksi ja metreiksi.

Samalla tavalla kuin varianssianalyysin yhteydessä regressioanalyysissäkin voidaan jakaa y:n vaihtelu (SST) kahteen osaan, joista toinen osa kertoo selitetyn vaihtelun (SSR) ja toinen selittämättä jäävän vaihtelun (SSE). On osoitettavissa, että $SST = SSR + SSE$, missä

$$\text{kokonaisneliösumma } SST = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

$$\text{regressioneliösumma } SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2,$$

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

Koska $SST = SSR + SSE$, niin

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST},$$

josta edelleen

$$\frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Merkitään $R^2 = SSR/SST$, jota kutsutaan mallin selityskertoimeksi, koska siinä verrataan mallin avulla selitettyä vaihtelua kokonaisvaihteluun. Ilmoittamalla $100R^2$, voidaan puhua mallin selitysasteesta. Jos $SSE = 0$, niin $R^2 = 1$. $R = r_{y,\hat{y}}$ on nimeltään yhteyskorrelaatiokerroin.

Yhden selittäjän mallissa, kun *vakio on* mukana, $R^2 = (r_{xy})^2$, jolloin $100(r_{xy})^2$ kertoo kuinka monta prosenttia y:n vaihtelusta voidaan x:n avulla selittää kyseisellä mallilla.

Esimerkki 4.1.8 Neliösummat ja selitysasteet edellisissä esimerkeissä.

Olkoon populaatiossa kahden muuttujan X ja Y välinen korrelaatiokerroin ρ. Siis $\rho = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y)$. Sitä voidaan estimoida otoksesta lasketulla otoskorrelaatiokertoimella r ja testata seuraavasti:

$$H_0\colon \rho = 0,$$
$$H_1\colon \rho \neq 0.$$

Jos H_0 on tosi, niin

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t_{n-2}.$$

Nyt H_0 hylätään riskitasolla α, jos aineiston perusteella laskettu $|t_{\text{hav.}}| > t_{\alpha/2;n-2}$.

Voidaan osoittaa, että yhden selittäjän mallissa vakion ollessa mukana $\hat{\beta}_1 = r_{xy}s_y/s_x$.

Esimerkki 4.1.9 Esimerkin 4.1.4 tilanteessa korrelatiokertoimen testaus.

$$H_0\colon \rho = 0;$$
$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t_{n-2}, \qquad \text{kun } H_0 \text{ tosi.}$$

Lasketaan aluksi korrelaatiokerroin ja sitten testisuure.

$$\begin{aligned} r &= SP_{xy}/\sqrt{SS_x SS_y} = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/7}{\sqrt{\left(\sum x_i^2 - (\sum x_i)^2/7\right)\left(\sum y_i^2 - (\sum y_i)^2/7\right)}} \\ &= \frac{187000 - 2800 \times 420/7}{\sqrt{(1400000 - 2800^2/7)(26550 - 420^2/7)}} \approx 0{,}977 \\ t &= \frac{0{,}977}{\sqrt{(1-0{,}977^2)/(7-2)}} \approx 10{,}304 > t_{0{,}005;5} = 4{,}032. \end{aligned}$$

4.2 Useampi selittävä muuttuja

Yleisesti regressioanalyysissä y:n riippuvuuden muuttujista $x_1, x_2, \ldots, x_k$ ajatellaan 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, jäännöstermi (ei havaittavissa oleva); β_0, β_1, β_2, …, β_k ovat mallin tuntemattomat parametrit, jotka aineiston perusteella ovat estimoitavissa. Lisäksi regressiomallissa oletetaan, että $\varepsilon_i \sim N(0,\sigma^2)$, $i = 1, 2, \ldots, n$ sekä ε_i:t toisistaan riippumattomiksi.

Parametrien estimointiin ei voida esittää samantapaisia lausekkeita kuin yhden selittäjän mallissa. Tyydytäänkin tässä vai toteamaan, että parametrit voidaan estimoida samojen periaatteiden mukaisesti kuin yhden selittäjän mallissa ja annetaan tarvittaessa tilastollisen ohjelmiston suorittaa estimointi sekä tarvittavien testisuureiden lasku.

Esimerkki 4.2.1 Olkoon

$$\begin{aligned} y &= \text{keuhkojen tilavuus (ml)}, \\ x_1 &= \text{ikä vuosina}, \\ x_2 &= \text{pituus (tuumina)}, \\ x_3 &= \text{poltettuja savukkeita / päivä (askeina)}. \end{aligned}$$

Olkoon eräästä 50 havainnon aineistosta estimoitu kolmen selittäjän regressiomalli ja saatu $\hat{y} = -39x_1 + 98.4x_2 - 180x_3$

Estimoi keuhkojen tilavuus, jos ikä on 20 vuotta, pituus 71 ja savukkeita kuluu aski päivässä.

Useamman selittäjän mallissa kertoimet voidaan tulkita siten, että yksittäinen kerroin kertoo keskimääräisen muutoksen y:ssä, kun kyseinen muuttuja kasvaa yhden yksikön muiden selittäjien pysyessä muuttumattomina.

Tässä iältään ja pituudeltaan samanlaisten henkilöiden kohdalla yhden paketin polttaminen päivässä vähentää keuhkojen tilavuutta keskimäärin 180 ml. Viiden vuoden iän lisäys pienentää keuhkojen tilavuutta keskimäärin $-39 \cdot 5$ ml $= 195$ ml, kun muut selittäjät pidetään vakiona.

Seuraavaksi käydään läpi useamman selittäjän malliin liittyvät testaukset. Ensinnäkin voidaan testata yksittäisiä kertoimia eli tutkia onko tarkasteltava muuttuja syytä lisätä malliin mukaan muiden selittäjien ollessa jo mallissa. Tällöin

$$\begin{aligned} H_0 &: \beta_i = 0, \\ H_1 &: \beta_i \neq 0. \end{aligned}$$

Jos H_0 on tosi, niin

$$t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \sim t_{n-k-1},$$

missä $s(\hat{\beta}_i)$ on $\hat{\beta}_i$:n estimoitu hajonta ja k on mallissa olevien selittäjien lukumäärä. Jos malliin ei kuulu vakiokerrointa, niin edellä testisuureen vapausasteet $n - k$. Lisäksi voidaan testata kaikkien selittäjien yhteisvaikutusta eli tutkia sitä saadaanko y:n vaihtelua selitettyä siten, että otetaan kaikki tarkasteltavat selittäjät yhtaikaa malliin. Regressiokertoimien yhteistestaus (kun vakiokerroin on mallissa mukana) voidaan muotoilla

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_k = 0, \\ H_1 &: \text{ainakin jokin } \beta_i \neq 0. \end{aligned}$$

Jos H_0 on tosi, niin

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} \sim F_{k,n-k-1}.$$

Neliösummat SST, SSR ja SSE määritellään kuten yhden selittäjänkin tilanteessa. MS-neliösummat saadaan kun vastaava neliösumma jaetaan nk. vapausasteillaan. On osoitettavissa, että $E(MSE) = \sigma^2$. Siis MSE on regressiomallissa olevan ε-satunnaistermin varianssin harhaton estimaattori. Mallin selitysvoimakkuudesta, kuten yhden selittäjänkin mallissa, kertoo vakiotermin ollessa mukana $R^2 = SSR/SST = 1 - SSE/SST$ (mallin selityskerroin) ja $100R^2$ on mallin selitysprosentti. Jos parametrien määrä on suuri suhteessa havaintomäärään tai jos halutaan vertailla malleja, jotka tehty eri aineistoista, voidaan käyttää

$$R^2(\text{adjusted}) = 1 - \frac{MSE}{SST/(n-1)}.$$

Yhteistestausta ei voida korvata peräkkäisillä t-testeillä. Yhden selittäjän tilanteessa t-testillä ja F-testillä testataan samaa hypoteesia H_0: $\beta_1 = 0$ ja tällöin $t^2 = F$.

Samaan tapaan kuin varianssianalyysinkin yhteydessä regressioanalyysin tulokset on tapana ilmoittaa taulukkona, josta löytyy estimoidut kertoimet, niiden estimoidut hajonnat, t-testisuureet, neliösummat sekä F-testisuure.

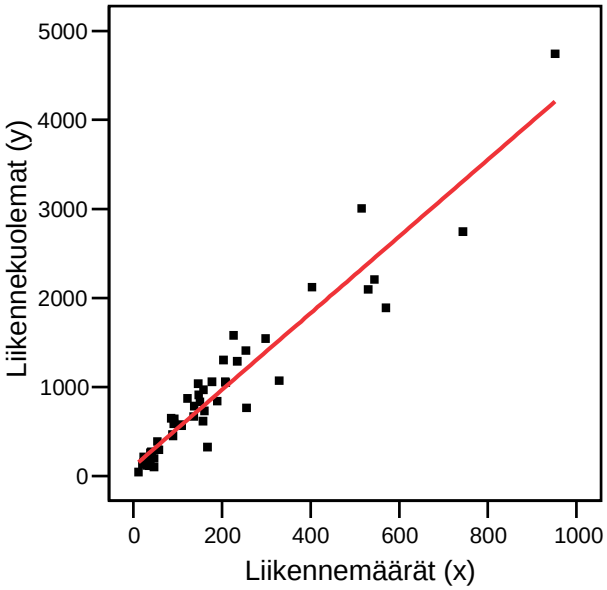
Esimerkki 4.2.2 Ks. https://www12.uta.fi/kirjasto/pdf/pdfkirjat/leppala.pdf.

Ks. regressioanalyysistä http://www.fsd.uta.fi/menetelmaopetus/regressio/analyysi.html.

4.3 Selittävien muuttujien valinnasta ja mallin oletuksista

Mallin valinta ei aina ole kovin helppoa. Pyritään valitsemaan niin monta 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. Monesti joudutaan tekemään erilaisia muunnoksia muuttujille ennen varsinaista mallin rakentamista. Taloudellisissa aineistoissa logaritmointi on usein tuiki tarpeellinen. On tilanteita, joissa selittäjät voivat olla esim. x, x^2, x^3, ...

Esimerkki 4.3.1 Regressioanalyysi, logaritmointi ja residuaalitarkasteluja. Aineisto Draper & Smith, Applied Regression Analysis (1981), s. 191, myös https://coursepages.uta.fi/mttta1/kevat-2019/esimerkkiaineistoja/.



Linear Fit:

Summary of Fit

R Square	0,913743
R Square Adj	0,911946
Root Mean Square Error	263,896
Mean of Response	926,76
Observations (or Sum Wgts)	50

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	F-ratio	Prob > F
Model	1	35410709	35410709	508,4744	0,0000
Error	48	3342772	69641,08		
C Total	49	38753481			

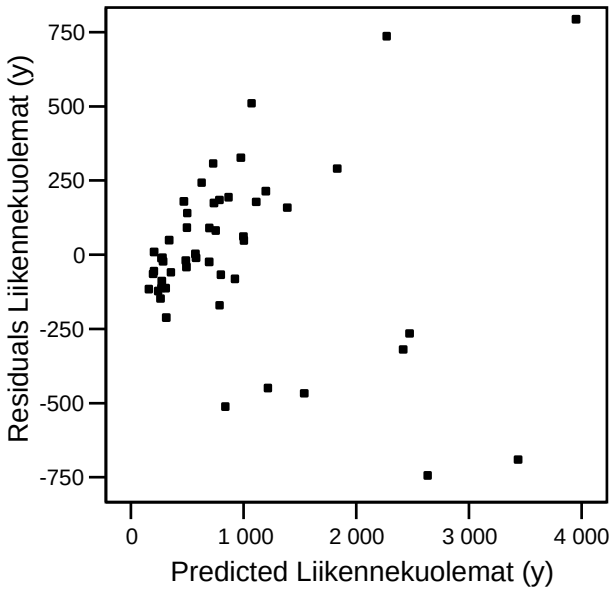
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob > $\|t\|$
Intercept	107,02891	52,09934	2,05	0,0454
Liikennemäärä (x)	4,306215	0,190968	22,55	0,0000

Tässä malli näyttää ihan hyvältä, jos tarkastellaan asiaa parametrien testauksen perusteella. Liikennemäärän kerroin on tilastollisesti merkitsevä ja selitysprosenttikin korkea 91 %.

Mallin parametrien testauksen lisäksi mallin sopivuutta tutkitaan myös residuaalien avulla. Tällöin tutkitaan mallin riittävyyttä ja oletusten voimassa olemista. Mallissa $Y = \beta_0 + \beta_1 x + \varepsilon$ tehdään oletukset, että $\varepsilon_i \sim N(0, \sigma^2)$ sekä ε_i:t toisistaan riippumattomia. Tehdään siis normaalijakaumaoletus, riippumattomuusoletus sekä vakiovarianssisuusoletus ε:sta. Jos malli oikea, niin residuaalien, jotka ovat ε:n

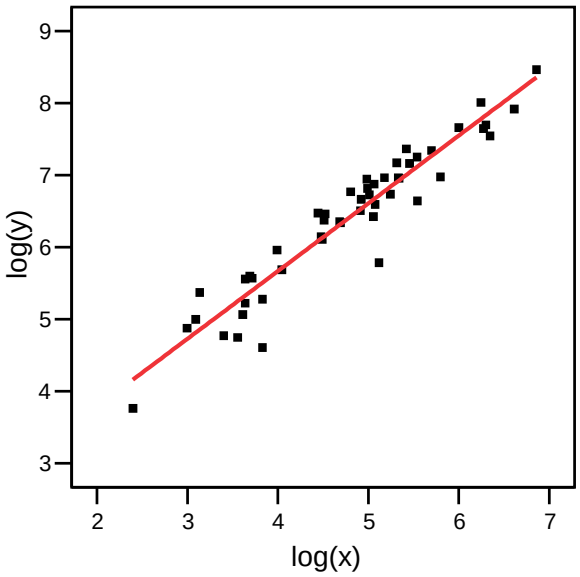
estimaatteja, tulisi käyttäytyä ε:n oletusten mukaisesti. Käyttäytymistä voidaan tutkia esim. piirtämällä pisteparvi residuaaleista ja estimoiduista y:n arvoista. Tässä esimerkissä pisteparvi näyttää hajaantuvan y:n estimoitujen arvojen kasvaessa.



Hajaantuminen on merkki siitä, että ei voida olettaa jokaisella ε_i:lla olevan samaa varianssia. Sama asia näkyy kyllä jo alkuperäisessä pisteparvessa, joka ihan selvästi hajoaa x:n kasvaessa.

Nyt voidaan menetellä siten, että logaritmoidaan molemmat muuttujat ja suoritetaan regressioanalyysi logaritmoiduilla arvoilla.

Näin saadaan seuraavat tulokset:



Linear Fit:

Summary of Fit

R Square	0,909514
R Square Adj	0,907629
Root Mean Square Error	0,307561
Mean of Response	6,397732
Observations (or Sum Wgts)	50

Analysis of Variance

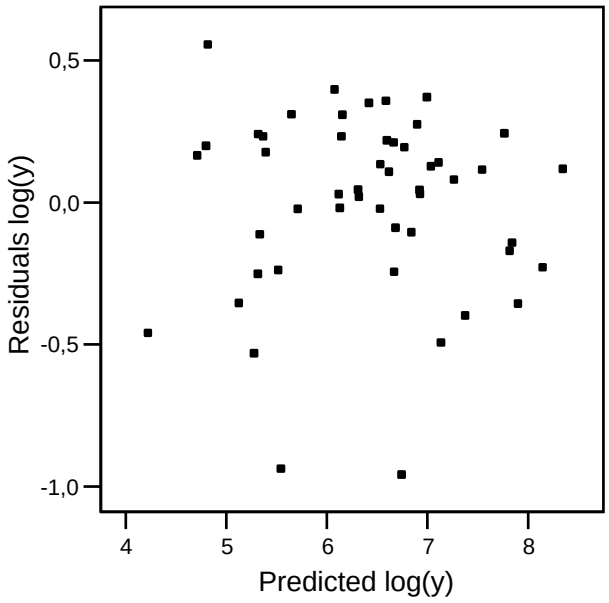
Source	*df*	Sum of Squares	Mean Square	F-ratio	Prob $> F$
Model	1	45,638716	45,6387	482,4705	0,0000
Error	48	4,540502	0,0946		
C Total	49	50,179218			

Parameter Estimates

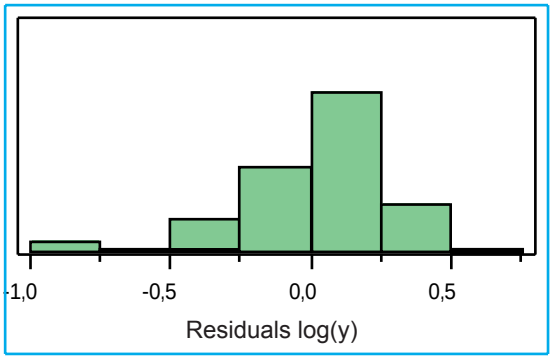
Term	Estimate	Std Error	t Ratio	Prob $> \|t\|$
Intercept	1,9036552	0,209172	9,10	0,0000
$\log(x)$	0,9410386	0,042842	21,97	0,0000

Tuloksista nähdään, että mallin parametrit ovat merkitseviä ja selitysprosenttikin 91.

Tässä mallissa residuaalit (alla) käyttäytyvät eri tavalla kuin edellä. Voidaan ajatella, että pisteparvi on x–akselin suuntainen nauha, joka kertoisi oletusten voimassa olemisesta sekä mallin riittävyydestä. Jos pisteparvessa olisi havaittavissa jotain muuta kuin nauhanomaista käyttäytymistä, niin se kertoisi, että tehdyt oletuksen malliin liittyen eivät pidä paikkaansa.



Katsotaan vielä residuaalien jakauma, joka pitäisi olla normaalinen.



Moments	
Mean	−0.00000
Std Dev	0.30441
Std Err Mean	0.04305
upper 95% Mean	0.08651
lower 95% Mean	−0.08651
N	50.00000
Sum Wgts	50.00000

Jos tässä testataan normaalisuutta, niin päädytään kyllä tulokseen, että otos ei ole peräisin normaalijakaumasta!

Esimerkki 4.3.2 Autoregressio. Aineisto Newbold, P., Statistics for Business and Economics. Prentice Hall, 1995 s. 588.

Twenty-eight quarterly observations from the United Kingdom on quantity of money in million pounds (y), income in million pounds (x_1) and the local authority interest rate (x_2) (aineisto myös https://coursepages.uta.fi/mttta1/kevat-2019/esimerkkiaineistoja/)

t	y_t	y_{t-1}	x_1	x_2	t	y_t	y_{t-1}	x_1	x_2
1	17602,5	•	14744	0,0805	15	17965,1	17734,9	15950	0,0582
2	17746,9	17602,5	14516	0,0828	16	18651,9	17965,1	15957	0,0482
3	17769	17746,9	14815	0,0781	17	19352,7	18651,9	16031	0,048
4	17909,1	17769	14900	0,0738	18	20444,1	19352,7	16295	0,0513
5	17855	17909,1	14829	0,0798	19	20835,3	20444,1	16151	0,0762
6	17470,8	17855	14900	0,0914	20	21827,4	20835,3	16803	0,0791
7	17352,6	17470,8	14980	0,0957	21	22375,2	21827,4	17528	0,1009
8	17481,2	17352,6	15085	0,0922	22	23217	22375,2	17301	0,091
9	17240,2	17481,2	14973	0,091	23	24011,6	23217	17503	0,1173
10	17467,7	17240,2	15359	0,0813	24	24975,2	24011,6	17455	0,1411
11	17619,8	17467,7	15362	0,0754	25	24736,3	24975,2	16620	0,1566
12	17683,1	17619,8	15540	0,0718	26	23407,3	24736,3	17779	0,1333
13	17954,9	17683,1	15404	0,0753	27	23560,7	23407,3	18040	0,1313
14	17734,9	17954,9	15649	0,0666	28	23421,2	23560,7	17827	0,1263

Estimoidaan malli $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 y_{t-1} + \varepsilon$.

Response: y

Summary of Fit	
R Square	0,975982
R Square Adj	0,972849
Root Mean Square Error	455,7342
Mean of Response	19928,38
Observations (or Sum Wgts)	27

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob $>\|t\|$
Intercept	−2297,819	1875,241	−1,23	0,2328
x_1	0,1573723	0,226106	0,70	0,4934
x_2	−14136,30	6351,172	−2,23	0,0361
y_{t-1}	1,0634212	0,126345	8,42	0,0000

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	F Ratio	Prob $> F$
Model	3	194109881	64703294	311,5323	0,0000
Error	23	4776955	207693,7		
C Total	26	198886836			

Nyt x_1 näyttää olevan tarpeeton ($t = 0{,}70$ ja $p = 0{,}4934$), joten jätetään tämä selittäjä pois mallista ja estimoidaan uusi malli $Y = \beta_0 + \beta_1 x_2 + \beta_2 y_{t-1} + \varepsilon$, joka tuottaa 97,5 %:n selitysasteen ja mallin kertoimet ovat merkittäviä vakiokerrointa lukuun ottamatta (p–arvot 0,1574; 0,0035; 0,000). Lisäksi H_0: $\beta_1 = \beta_2 = 0$, hylätään ($F = 477{,}3098$; $p = 0{,}000$). Malli on siis näiltä osin kaikin puolin kunnossa.

Response: y

Summary of Fit

R Square	0,975476
R Square Adj	0,973432
Root Mean Square Error	450,8126
Mean of Response	19928,38
Observations (or Sum Wgts)	27

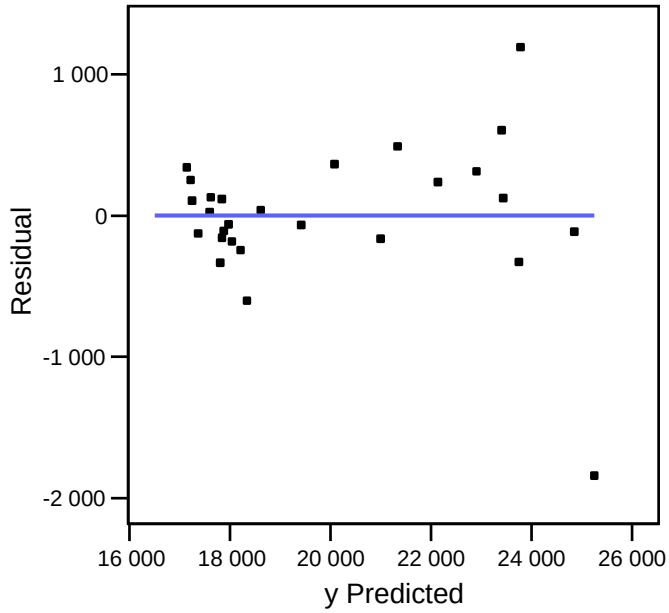
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob $>\|t\|$
Intercept	−1106,681	758,3441	−1,46	0,1574
x_2	−16666,87	5151,272	−3,24	0,0035
y_{t-1}	1,1426647	0,054183	21,09	0,0000

Analysis of Variance

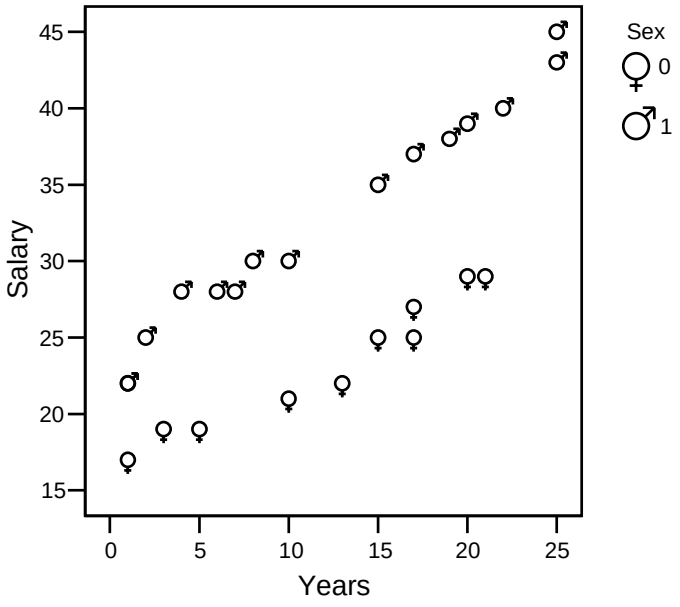
Source	*df*	Sum of Squares	Mean Square	F Ratio	Prob $> F$
Model	3	194009268	97004634	477,3098	0,0000
Error	24	4877568	203232		
C Total	26	198886836			

Tutkitaan vielä residuaalien käyttäytymistä.



Pisteparvi antaa kyllä viitteitä suuntaan, että vakiovarianssisuusoletus ei olisi ehkä voimassa. Toisaalta havaintoja on kovin vähän, joten pidemmän aikasarjan käyttö voisi olla jatkotoimenpiteenä aiheellinen.

Esimerkki 4.3.3 Dummy -muuttuja selittäjänä regressioanalyysissä. Palkan riippuvuutta palveluvuosista ja sukupuolesta. Aineisto: Younger (1985), A First Course in Linear Regression.



Salary	Years	Sex*	Salary	Years	Sex*
35	15	1	28	6	1
27	17	0	29	20	0
45	25	1	19	3	0
22	13	0	29	21	0
25	2	1	38	19	1
30	10	1	19	5	0
37	17	1	22	1	1
25	17	0	39	20	1
17	1	0	40	22	1
28	4	1	21	10	0
43	25	1	28	7	1
25	15	0	30	8	1
22	1	1			

*1 = mies

Palkka näyttää siis riippuvan paitsi palveluvuosista niin myös sukupuolesta. Voitaisiin tehdä yhden selittäjän regressioanalyysit miehillä ja naisilla erikseen. Yksi tapa olisi myös estimoida kahden selittäjän malli $\text{Salary} = \beta_0 + \beta_1 \cdot \text{Years} + \beta_2 \cdot \text{Sex} + \varepsilon$, jolloin saadaan estimoiduksi kaksi samansuuntaista suoraa

$$\begin{aligned} E(\text{Salary}) &= \beta_0 + \beta_1 \cdot \text{Years} && \text{(Naiset, Sex} = 0), \\ E(\text{Salary}) &= \beta_0 + \beta_1 \cdot \text{Years} + \beta_2 && \text{(Miehet, Sex} = 1). \end{aligned}$$

Estimointitulokset:

Response: Salary

Summary of Fit

R Square	0,970068
R Square Adj	0,967347
Root Mean Square Error	1,412216
Mean of Response	28,92
Observations (or Sum Wgts)	25

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob $>\lvert t\rvert$
Intercept	13,970213	0,627120	22,28	0,0000
Years	0,7647366	0,036088	21,19	0,0000
Sex	9,4176491	0,576540	16,33	0,0000

Analysis of Variance

Source	*df*	Sum of Squares	Mean Square	F Ratio	Prob $> F$
Model	3	1421,9642	710,982	356,4975	0,0000
Error	22	43,8758	1,994		
C Total	24	1465,8400			

Estimoinnin tulos:

Naiset: $\text{Salary(estimoitu)} = 13{,}970213 + 0{,}7647366 \cdot \text{Years},$
Miehet: $\text{Salary(estimoitu)} = 13{,}970213 + 0{,}7647366 \cdot \text{Years} + 9{,}4176491.$

Testaukset tehdään tavanomaiseen tapaan. Selitysprosentti 97.

4.4 Varianssianalyysimalli

Oletukset yksisuuntaisessa varianssianalyysissä:

$$\begin{aligned} &Y_{11}, Y_{12}, \ldots, Y_{1n_1} \quad \text{satunnaisotos } N(\mu_1, \sigma^2)\text{:sta,} \\ &Y_{21}, Y_{22}, \ldots, Y_{2n_2} \quad \text{satunnaisotos } N(\mu_2, \sigma^2)\text{:sta,} \\ &\quad \vdots \\ &Y_{I1}, Y_{I2}, \ldots, Y_{In_I} \quad \text{satunnaisotos } N(\mu_I, \sigma^2)\text{:sta.} \end{aligned}$$

Halutaan tutkia ovatko jakaumien odotusarvot yhtä suuret, jolloin

$$\begin{aligned} &H_0\colon \mu_1 = \mu_2 = \cdots = \mu_I, \\ &H_1\colon \text{kaikki odotusarvot eivät ole samoja.} \end{aligned}$$

Oletuksista seuraa, että varianssianalyysi voidaan ajatella mallina

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \text{missä } \varepsilon_{ij} \sim N(0, \sigma^2).$$

μ_1, μ_2, …, μ_I ovat mallin parametrit. Vaihtoehtoisesti myös $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$.

4.5 SPSS-ohjeita

Regressioanalyysi

Analyze
Regression ▶ Linear...

Korrelaatiokerroin

Analyze
Correlate ▶ Bivariate... Pearson

Luku 5

Epäparametrisista menetelmistä

Useimmissa tähän asti esillä olleissa testeissä ja menetelmissä on tehty oletuksia populaation jakaumasta. Oletetaan esimerkiksi, että jakauma on peräisin normaalijakaumasta. Testejä/menetelmiä, jotka perustuvat johonkin jakaumaoletukseen, kutsutaan parametrisiksi testeiksi/menetelmiksi.

On kehitetty menetelmiä, joissa jakaumaoletuksia ei tarvitse tehdä. Näihin liittyviä testejä kutsutaan epäparametrisiksi tai ei-parametrisiksi testeiksi. Käytännössä epäparametrisen menetelmän käyttö ei aseta muuttujalle korkeita mittaasteikkovaatimuksia. Useat epäparametriset menetelmät perustuvat järjestysasteikolliseen mittaukseen ja testaus kombinatoriikkaan. Tähän asti esillä olleita epäparametrisiä testejä ovat olleet χ^2-testit. Epäparametrisissä menetelmissä voidaan tutkia vaikkapa jakauman sijaintia tai vertailla kahden jakauman sijainteja luopumalla normaalijakaumaoletuksesta.

Tässä yhteydessä otetaan esimerkinomaisesti yksi epäparametrinen testi, merkkitesti. Oletetaan, että meillä on kaksi toisistaan riippuvaa otosta ja halutaan tutkia jakaumien sijainteja (vertaa vastinparien t-testi). Ei tehdä jakaumaoletuksia, jolloin meillä voi olla järjestysasteikollista mittausta.

Esimerkki 5.0.1 Halutaan vertailla kahta viinimerkkiä. Kahdeksan henkilöä maistaa molempia merkkejä ja arvioi makua järjestysasteikollisella mittarilla. Saadaan tulokset:

Henkilö	Viini 1	Viini 2	Merkki (+, jos ”Viini 1” > ”Viini 2”)
1	6	8	−
2	4	9	−
3	5	4	+
4	8	7	+
5	3	9	−
6	6	9	−
7	7	7	0 (jätetään pois)
8	5	9	−

Jos viinien laaduissa ei ole eroja, pitäisi plus- ja miinus-merkkejä olla saman verran. Merkitään X = plus-merkkien lkm.

$$H_0\colon \text{Viinien laaduissa ei eroja;}$$
$$H_1\colon \text{Viini 2 parempi.}$$

Jos H_0 on tosi, niin

$$\begin{aligned} P(\text{1. merkki on } +) &= 0.5, \\ P(\text{2. merkki on } +) &= 0.5, \\ &\vdots \\ P(\text{7. merkki on } +) &= 0.5. \end{aligned}$$

Lasketaan nyt todennäköisyys sille, että saadaan X:n arvo, joka on havaittu tai sitä pienempi H_0:n ollessa tosi. Tehdään johtopäätelmä lasketun todennäköisyyden perusteella. Jos tämä todennäköisyys $\leq \alpha$, niin hylätään H_0 riskitasolla α.

Jos H_0 on tosi, niin $X \sim \text{Bin}(7, 0.5)$ ja asetetut hypoteesit voidaan kirjata

$$\begin{aligned} H_0&\colon p = 0.5, \\ H_1&\colon p < 0.5. \end{aligned}$$

Nyt H_0:n ollessa tosi, $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \cdots = 0.2266$. (Newbold (1995), Statistics for Business and Economics)

Esimerkissä 5.0.2 olleen merkkitestin yhteydessä voidaan käyttää normaalijakaumaa approksimoimaan X:n jakaumaa, kun otoskoko suuri. Tällöin $X \sim N(np, np(1-p))$, likimain.

Esimerkki 5.0.2 Sata satunnaisesti valittua lasta vertaili jäätelömerkkejä A ja B. 56 lasta piti jäätelöä A parempana, 4 lasta piti molempia jäätelöitä yhtä hyvinä. Määritellään X = jäätelöä A parempana pitävien lkm.

$$H_0\colon \text{Jäätelöt samanlaisia.}$$

Jos H_0 tosi niin, $X \sim N\left(\frac{96}{2}, \frac{96}{4}\right)$, likimain ja

$$Z = \frac{X - 96/2}{\sqrt{96/4}} \sim N(0, 1),$$

johon testaus voidaan perustaa.

(Newbold (1995), Statistics for Business and Economics)

Muutamia epäparametrisiä testejä:

Wilcoxonin testi Merkkitestin tilanteisiin sopiva, kun mittaustaso sellainen, että voidaan vertailla erotuksien suuruutta.

Mann-Witneyn testi Kahden riippumattoman otoksen t-testin epäparametrinen vastine (normaalijakaumaoletus ei voimassa).

Kruskal-Wallis Epäparametrinen vastine yksisuuntaiselle varianssianalyysille (normaalijakaumaoletusta ei tehdä, selitettävä muuttuja voi olla järjestysasteikollinen).

Welch tai Browm-Forsythe Yksisuuntaisessa varianssianalyysissä, kun oletus varianssien yhtäsuuruudesta ei voimassa.

5.1 SPSS-ohjeita

Epäparametriset testit

Analyze
Nonparametric Tests ▶

Kirjallisuutta

Kirjallisuusluettelo, jota on käytetty tukena tämän luentorungon kirjoittamisessa.

Agresti, A. & Finlay, B., Statistical Methods for the Social Sciences. Prentice Hall, 1997.

Anderson, T. W. & Sclove, S. L., Introductory Statistical Analysis. Houghton Mifflin Company, 1974.

Clarke, G. M. & Cooke, D., A Basic course in Statistics. Arnold, 1998.

Devore, J. & Peck, R., Statistics, The Exploration and Analysis of Data. West Publishing Company, 1986.

Helenius, H., Tilastollisten menetelmien perustiedot. Statcon Oy, 1992.

Karjalainen, L. & Ruuskanen, A. Tilastomatematiikka. Pii-kirjat, 1994.

Liski, E. & Puntanen, S., Tilastotieteen peruskurssi I & II. Tampereen yliopisto.

Manninen P., Tilastotiedettä yhteiskuntatieteilijöille. Gaudeamus, 1978.

Mattila, S., Tilastotiede 1 & 2 , Gaudeamus.

Mellin, I., Johdatus tilastotieteeseen, 1. kirja, tilastotieteen johdantokurssi, Helsingin yliopisto.

Mellin, I., Johdatus tilastotieteeseen, 2. kirja, tilastotieteen jatkokurssi, Helsingin yliopisto.

Moore, D., The Basic Practice of Statistics, Freeman, 1997.

Moore, D., Introduction to the Practice of Statistics, 3rd ed., Freeman, 1998.

Newbold, P., Statistics for Business and Economics. Prentice Hall, 1995.

Ott, L. & Mendenhall, W., Understanding Statistics. Duxbury Press, 1985.

Siegel, A., Statistics and Data Analysis An Introduction. John Wiley & Sons, 1988.

Vasama, P.-M. & Vartia, Y., Johdatus tilastotieteeseen 1 & 2, Gaudeamus.