Tilastotieteen peruskurssi c

Tilastollisia malleja

Henri Pesonen

Luku 1

Tilastollisia malleja

Tällä kurssilla tarkastelemme pääosin ns. selittävän tutkimuksen malleja, joissa on aina vähintään yksi selitettävä muuttuja, eli vaste, jonka arvoista ja arvojen vaihteluista olemme kiinnostuneita, ja selittävä muuttuja, jonka avulla vaihtelua selitetään. Jatkossa emme tarkastele muuttujien yhteisjakaumia, vaan olemme kiinnostuneista pelkästään selitettävän muuttujan jakaumasta. Jos selittäjä on tekijä, tutkimme selitettävän muuttujan vaihtelua tekijän määräämissä osissa ja jos selittäjä on vaste, tarkastelemme selitettävän muuttujan ehdollista jakaumaa selittäjän eri arvoilla. Selitysmalliksi voidaan ajatella esimerkiksi kahden populaation odotusarvojen vertailu, jossa tutkittiin vaikuttaako dikotominen (kaksiarvoinen) selittäjä (populaatio 1 tai 2) kvantitatiiviseen vasteeseen (havaintoaineisto).

1.1 Yksisuuntainen varianssianalyysi

Varianssianalyysi (analysis of variance, ANOVA) on kokoelma malleja ja menetelmiä, joissa pyritään jakamaan satunnaismuuttujan havaittu vaihtelu osiin vaihtelun lähteen mukaisesti. Käytännössä varianssianalyysillä voidaan testata ovatko useamman satunnaismuuttujan odotusarvot toisistaan poikkeavat, ja varianssianalyysiä voidaankin ajatella kahden riippumattoman otoksen t-testin laajennuksena useammalle ryhmälle. Nimestään huolimatta varianssianalyysillä **ei siis testata ryhmien varianssien eroja** vaan odotusarvojen eroja. Jos tilannetta mallitetaan siten, että yksi kategorinen muuttuja määrittelee vertailtavat ryhmät, on kyseessä yksisuuntainen varianssianalyysi (one-way ANOVA).

Jotta pystymme analysoimaan tilannetta, tarvitsemme aineistolle tilastollisen mallin. Vaikutusten malli havainnolle $y_{i,j}$ voidaan kirjoittaa muodossa

$$y_{i,j} = \mu + \beta_i + \epsilon_{i,j},$$

tai

$$y_{i,j} = \mu_i + \epsilon_{i,j},$$

jossa $y_{i,j}$ on i, j. havainto ilmaistuna yleisen odotusarvon μ ja $i.$ ryhmän odotusarvon β_i, tai $i.$ ryhmän odotusarvon μ_i, ja virhetermin $\epsilon_{i,j}$ avulla. Tällä kurssilla yksinkertaistuksen vuoksi emme merkitse erikseen satunnaismuuttujia isoilla kirjaimilla. Virhetermit oletetaan usein riippumattomiksi ja samoin jakautuneiksi (independent identically distiributed, iid) normaalijakautuneiksi satunnaismuuttujiksi, joille käytetään merkintää

$$\epsilon_{i,j} \overset{\text{iid}}{\sim} \mathsf{Normal}\left(0, \sigma^2\right).$$

Virhetermit ovat **ei-havaittavia** satunnaismuuttujia. Nyt siis havainnoille

$$\begin{aligned} \mathsf{E}\left(y_{i,j}\right) &= \mu_i \\ \mathsf{Var}\left(y_{i,j}\right) &= \sigma^2. \end{aligned}$$

Olemme tehneet siis **samavarianssisuusoletuksen**

$$\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2 > 0,$$

eli jokaisen ryhmän havaintojen hajonta on sama.

Havaintoaineisto koostuu k:sta toisistaan riippumattomasta ryhmästä

$$\begin{aligned} &(y_{1,1}, y_{1,2}, \ldots, y_{1,m_1}) \quad (\ m_1 \text{ kpl havaintoja }) \\ &\qquad \vdots \\ &(y_{k,1}, y_{1,2}, \ldots, y_{k,m_k}) \quad (\ m_k \text{ kpl havaintoja }), \end{aligned}$$

ja yhteensä $n = m_1 + \cdots + m_k$ havainnosta. Joskus on kätevää ajatella, että aineisto koostuu kahden tilastollisen muuttujan arvoista: havainnoista y_i sekä niihin liittyvistä selittävän muuttujan A arvoista a_i. Voitaisiin siis merkitä että saamme havainnot $(y_i, a_i), i = 1, \ldots, n$, jossa Y on kvantitatiivisen muuttujan vaste, ja A:n arvo $a_i \in \{1, \ldots, k\}$.

Esimerkki 1. Tarkastellaan kahdella eri valmistustavalla tuotetun muovin elastisuutta. Kerätään havaintoaineisto

Valmistustapa	1	2
	6.4	6.7
	7.2	8.0
	6.2	8.2
	8.6	8.1
	6.3	7.2.

Aineisto voidaan kirjoittaa muodossa $y_{1,1} = 6.4, y_{1,2} = 7.2, \ldots, y_{1,5} = 6.3, y_{2,1} = 6.7, y_{2,2} = 8.0, \ldots, y_{2,5} = 7.2$ tai muodossa $(y_1, a_1) = (6.4, 1), (y_2, a_2) = (7.2, 1), \ldots, (y_6, a_6) = (6.7, 2), \ldots, (y_{10}, a_{10}) = 7.2$

Estimoidaan ensin mallin kiinnostavia parametreja $\mu_1, \ldots, \mu_k, \sigma^2$. Jokaiselle ryhmälle voidaan laskea ryhmäkohtaiset otoskeskiarvot

$$\bar{y}_{i,\cdot} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{i,j}, \quad i = 1, \ldots, k,$$

sekä koko havaintoaineistolle otoskeskiarvo

$$\bar{y}_{\cdot,\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m_i} y_{i,j} = \frac{1}{n} \sum_{i=1}^{k} m_i \bar{y}_{i,\cdot}.$$

Voidaan näyttää että estimaattorit $\bar{y}_{1,\cdot}, \ldots, \bar{y}_{k,\cdot}$ ovat harhattomia ja tarkentuvia vastaaville ryhmille. Varianssin σ^2 harhaton ja tarkentuva estimaattori on jäännöskeskineliö (vastaa yhteisotosvarianssia)

$$s_E^2 = \frac{(m_1 - 1)s_1^2 + \cdots + (m_k - 1)s_k^2}{n - k}.$$

Lähdetään tarkastelemaan vaihtelun lähteitä vaikusten mallissa. Vaihtelua havaintoaineistoon tuo vaihtelu eri ryhmien i ja j, $i \neq j$, välillä, sekä vaihtelu jokaisen ryhmän l sisällä. Hajonnan tunnuslukuna koko havaintoaineistoille voidaan käyttää **kokonaisneliösummaa** (total sum of squares, SST)

$$\mathsf{SST} = \sum_{i=1}^{k} \sum_{j=1}^{m_i} (y_{i,j} - \bar{y}_{\cdot,\cdot})^2, \tag{1.1}$$

josta voitaisiin laskea mm. otosvarianssi jakamalla summa luvulla $n-1$. Varianssianalyysissä jaamme kokonaisneliösumman (1.1) osiin, joista kaikki paitsi yksi kohdistuu yhdelle riippumattomalle muuttujalle, sekä loppuosalle joka yhdistetään jäännösvirheeseen.

$$\mathsf{SST} = \begin{cases} \text{Ryhmään 1 liittyvä neliösumma} \\ \text{Ryhmään 2 liittyvä neliösumma} \\ \vdots \\ \text{Ryhmään k liittyvä neliösumma} \\ \text{Jäännösvirheen neliösumma} \end{cases}$$

Tarkastellaan ensin ryhmittelevään muuttujaan A liittyvää neliösummaa. Havainnon ij kohdalla tekijä A selittää eron $\bar{y}_{i,\cdot} - \bar{y}_{\cdot,\cdot}$, eli paljonko havaintoaineiston osat vaihtelevat toisistaan. Yhteensä riippumattomuusoletuksen perusteella saamme tekijälle A neliösumman

$$\mathsf{SSA} = \sum_{i=1}^{k} m_i(\bar{y}_{i,\cdot} - \bar{y}_{\cdot,\cdot})^2 = v_A s_A^2,$$

missä $v_A = k - 1$ on tekijän A vapausaste ja

$$s_A^2 = \frac{\sum_{i=1}^{k} m_i(\bar{y}_{i,\cdot} - \bar{y}_{\cdot,\cdot})^2}{k-1}$$

on A:n jäännöskeskineliö. Vaihtelua havaintoaineiston eri osien sisällä kuvaa jäännösneliösumma (sum of squares for error)

$$\mathsf{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{i,j} - \bar{y}_{i,\cdot})^2 = v_E s_E^2,$$

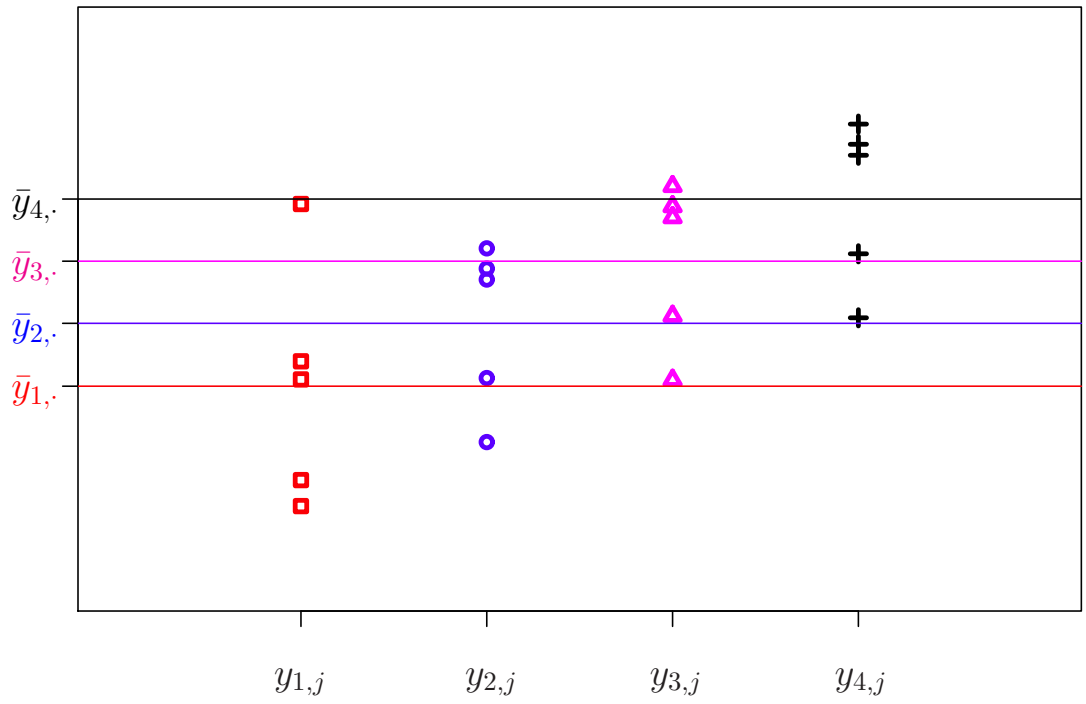
missä $v_E = n - k$ on jäännösten vapauste. Mekaanisella kaavanmuokkauksella voidaan näyttää, että kokonaisneliösumma voidaan ilmaista tekijään A liittyvän neliösumman sekä jäännösneliösumman avulla

$$\mathsf{SST} = \mathsf{SSA} + \mathsf{SSE}.$$

Vaihtelun tyyppejä on havainnollistettu Kuvassa 1.1.

Esimerkki 2. Tarkastellaan eri valmistustavoilla tuotetun muovin elastisuutta. Saadaan seuraava satunnainen otos muovien elastisuudesta.

Valmistustapa	1	2	3	4
Havainnot	6.4	6.7	7.2	7.7
	7.2	8.0	8.5	9.0
	6.2	8.2	8.7	9.2
	8.6	8.1	8.6	9.1
	6.3	7.2	7.7	8.2
$\sum_{j=1}^{5} y_{i,j}$	34.7	38.2	40.7	43.2
$\bar{y}_{i,\cdot}$	6.94	7.64	8.14	8.64



Kuva 1.1: SSA mittaa ryhmien odotusarvojen $\bar{y}_{i,\cdot}$ vaihtelua, eli vaakasuorien viivojen hajontaa, ja SSE mittaa havaintojen $y_{i,j}$ yhteenlaskettua vaihtelua ryhmittäin.

Aineistoa on havainnollistettu Kuvassa 1.1. Huomaa että kyseessä on simuloitu aineisto, jossa esimerkin vuoksi valmistapojen 3 ja 4 arvot ovat laskettu valmistustavan 2 havainnoista, siten että $y_{3,i} = y_{2,i} + 0.5$ ja $y_{4,i} = y_{2,i} + 1.0$. Tästä seuraa että ryhmien $2, 3$ ja 4 havaintojen ryhmäkohtainen hajonta on kaikilla sama. Tilanne ei siis edes simuloi satunnaisuutta, vaan tärkeintä on tarkastella neliösummien arvojen muuttumista, kun ryhmien otoskeskiarvot muuttuvat ja hajonnat pysyvät samoina.

(a) Otetaan huomioon ainoastaan valmistustavat 1 ja 2. Nyt

$$
\begin{aligned}
\bar{y}_{\cdot,\cdot} &= \frac{1}{10}\left(34.7 + 38.2\right) = 7.29 \\
\mathsf{SSA} &= 5 \cdot (6.94 - 7.29)^2 + 5 \cdot (7.64 - 7.29)^2 = 1.225 \\
\mathsf{SSE} &= \quad (6.4 - 6.94)^2 + (7.2 - 6.94)^2 + (6.2 - 6.94)^2 + (8.6 - 6.94)^2 + (6.3 - 6.94)^2 \\
&\quad + (6.7 - 7.64)^2 + (8.0 - 7.64)^2 + (8.2 - 7.64)^2 + (8.1 - 7.64)^2 + (7.2 - 7.64)^2 \\
&= 5.804 \\
\mathsf{SST} &= \quad (6.4 - 7.29)^2 + (7.2 - 7.29)^2 + (6.2 - 7.29)^2 + (8.6 - 7.29)^2 + (6.3 - 7.29)^2 \\
&\quad + (6.7 - 7.29)^2 + (8.0 - 7.29)^2 + (8.2 - 7.29)^2 + (8.1 - 7.29)^2 + (7.2 - 7.29)^2 \\
&= 7.029 \\
v_A &= 2 - 1 = 1 \\
v_E &= 5 + 5 - 2 = 8
\end{aligned}
$$

Voidaan tarkastaa, että $\mathsf{SST} = \mathsf{SSA} + \mathsf{SSE} = 1.225 + 5.804 = 7.029$.

(b) Otetaan huomioon ainoastaan valmistustavat 1 ja 3. Nyt

$$
\begin{aligned}
\bar{y}_{\cdot,\cdot} &= \frac{1}{10}\left(34.7 + 40.7\right) = 7.54 \\
\mathsf{SSA} &= 5 \cdot (6.94 - 7.54)^2 + 5 \cdot (8.14 - 7.54)^2 = 3.6 \\
\mathsf{SSE} &= \quad (6.4 - 6.94)^2 + (7.2 - 6.94)^2 + (6.2 - 6.94)^2 + (8.6 - 6.94)^2 + (6.3 - 6.94)^2 \\
&\quad + (7.2 - 8.14)^2 + (8.5 - 8.14)^2 + (8.7 - 8.14)^2 + (8.6 - 8.14)^2 + (7.7 - 8.14)^2 \\
&= 5.804 \\
\mathsf{SST} &= \quad (6.4 - 7.54)^2 + (7.2 - 7.54)^2 + (6.2 - 7.54)^2 + (8.6 - 7.54)^2 + (6.3 - 7.54)^2 \\
&\quad + (7.2 - 7.54)^2 + (8.5 - 7.54)^2 + (8.7 - 7.54)^2 + (8.6 - 7.54)^2 + (7.7 - 7.54)^2 \\
&= 9.404 \\
v_A &= 2 - 1 = 1 \\
v_E &= 5 + 5 - 2 = 8
\end{aligned}
$$

Voidaan tarkastaa, että $\mathsf{SST} = \mathsf{SSA} + \mathsf{SSE} = 3.6 + 5.804 = 9.404$. Huomaa että SSE on sama verrattuna (a)-kohdan tilanteeseen, sillä havaintoaineiston osien hajonta on sama. Lisäksi huomaa, että SSA on (b)-kohdassa suurempi, sillä ryhmien arvot poikkeavat keskimääräisesti toisistaan enemmän kuin (a)-kohdassa.

(c) Otetaan huomioon ainoastaan valmistustavat 1 ja 4. Nyt

$$\begin{aligned}
\bar{y}_{\cdot,\cdot} &= \frac{1}{10}(34.7 + 43.2) = 7.79 \\
\mathsf{SSA} &= 5 \cdot (6.94 - 7.79)^2 + 5 \cdot (8.64 - 7.79)^2 = 7.225 \\
\mathsf{SSE} &= \quad (6.4 - 6.94)^2 + (7.2 - 6.94)^2 + (6.2 - 6.94)^2 + (8.6 - 6.94)^2 + (6.3 - 6.94)^2 \\
&\quad + (7.7 - 8.64)^2 + (9.0 - 8.64)^2 + (9.2 - 8.64)^2 + (9.1 - 8.64)^2 + (8.2 - 8.64)^2 \\
&= 5.804 \\
\mathsf{SST} &= \quad (6.4 - 7.79)^2 + (7.2 - 7.79)^2 + (6.2 - 7.79)^2 + (8.6 - 7.79)^2 + (6.3 - 7.79)^2 \\
&\quad + (7.7 - 7.79)^2 + (9.0 - 7.79)^2 + (9.2 - 7.79)^2 + (9.1 - 7.79)^2 + (8.2 - 7.79)^2 \\
&= 13.029 \\
v_A &= 2 - 1 = 1 \\
v_E &= 5 + 5 - 2 = 8
\end{aligned}$$

Voidaan tarkastaa, että $\mathsf{SST} = \mathsf{SSA} + \mathsf{SSE} = 7.225 + 5.804 = 13.029$. Huomaa että SSE on sama verrattuna (a)-kohdan tilanteeseen, sillä havaintoaineiston osien hajonta on sama. Lisäksi huomaa, että SSA on (c)-kohdassa huomattavasti suurempi, sillä ryhmien arvot poikkeavat keskimääräisesti toisistaan huomattavasti enemmän kuin (a)-kohdassa.

(d) Otetaan huomioon kaikki valmistustavat. Nyt

$$\begin{aligned}
\bar{y}_{\cdot,\cdot} &= \frac{1}{20}(34.7 + 38.2 + 40.7 + 43.2) = 7.84 \\
\mathsf{SSA} &= 5 \cdot (6.94 - 7.84)^2 + 5 \cdot (7.64 - 7.84)^2 + 5 \cdot (8.14 - 7.84)^2 + 5 \cdot (8.64 - 7.84)^2 = 7.9 \\
\mathsf{SSE} &= \quad (6.4 - 6.94)^2 + (7.2 - 6.94)^2 + (6.2 - 6.94)^2 + (8.6 - 6.94)^2 + (6.3 - 6.94)^2 \\
&\quad + (6.7 - 7.64)^2 + (8.0 - 7.64)^2 + (8.2 - 7.64)^2 + (8.1 - 7.64)^2 + (7.2 - 7.64)^2 \\
&\quad + (7.2 - 8.14)^2 + (8.5 - 8.14)^2 + (8.7 - 8.14)^2 + (8.6 - 8.14)^2 + (7.7 - 8.14)^2 \\
&\quad + (7.7 - 8.64)^2 + (9.0 - 8.64)^2 + (9.2 - 8.64)^2 + (9.1 - 8.64)^2 + (8.2 - 8.64)^2 \\
&= 9.268 \\
\mathsf{SST} &= \quad (6.4 - 7.84)^2 + (7.2 - 7.84)^2 + (6.2 - 7.84)^2 + (8.6 - 7.84)^2 + (6.3 - 7.84)^2 \\
&\quad + (6.7 - 7.84)^2 + (8.0 - 7.84)^2 + (8.2 - 7.84)^2 + (8.1 - 7.84)^2 + (7.2 - 7.84)^2 \\
&\quad + (7.2 - 7.84)^2 + (8.5 - 7.84)^2 + (8.7 - 7.84)^2 + (8.6 - 7.84)^2 + (7.7 - 7.84)^2 \\
&\quad + (7.7 - 7.84)^2 + (9.0 - 7.84)^2 + (9.2 - 7.84)^2 + (9.1 - 7.84)^2 + (8.2 - 7.84)^2 \\
&= 17.168 \\
v_A &= 4 - 1 = 3 \\
v_E &= 5 + 5 + 5 + 5 - 4 = 16
\end{aligned}$$

Voidaan tarkastaa, että $\mathsf{SST} = \mathsf{SSA} + \mathsf{SSE} = 7.9 + 9.268 = 17.168$.

Tarkastellaan tekijän A vaikutusta havaintoaineiston osien odotusarvoihin $\mu_1, \ldots, \mu_k$. Tätä varten muodostetaan nolla- ja vastahypoteesi

H_0 : Kaikkien ryhmien todelliset odotusarvot ovat yhtäsuuria.

H_v : Joidenkin ryhmien odotusarvojen välillä on eroja.

Testataksemme havaintoaineiston poikkeaman merkitsevyyttä nollahypoteesista, muodostetaan **F-testisuure**, joka määritellään

$$F = \frac{\text{Ryhmien vaihtelu}}{\text{Ryhmien sisäinen vaihtelu}}.$$

Jos F:n havaittu F_{hav} arvo on suuri, silloin ryhmien välinen vaihtelu on suurta verrattuna ryhmien sisäiseen vaihteluun. Tämä on epätodennäköistä jos nollahypoteesi olisi voimassa, ja mitä suurempi on F_{hav}, niin sitä enemmän on evidenssiä nollahypoteesia vastaan. Kriittinen alue testille löytyy siis jakauman oikeasta hännästä. F-jakauman arvoja taulukoidaan yleensä kirjoittamalla esim. 0.05-yläkvantiileja eri vapausasteilla. Käytännössä siis testi suoritetaan tasolla 0.05 vertaamalla havaittua testisuureen arvoa taulukoituihin 0.05-kriittisiin arvoihin $f_{0.05}^{(v_A, v_E)}$, jotka ovat kriittisten alueiden alarajoja.

Nollahypoteesi jää voimaan, jos havaittu arvo on taulukoitua arvoa pienempi ja nollahypoteesi hylätään, jos havaittu arvo on taulukoitua suurempi. Matemaattisesti testisuure voidaan kirjoittaa muodossa

$$F = \frac{\mathsf{SSA}/v_A}{\mathsf{SSE}/v_E} = \frac{s_A^2}{s_E^2} \sim \mathsf{F}(v_A, v_E),$$

ja merkitsevyystaso nollahypoteesia vastaan on

$$\mathsf{p-arvo} = \mathsf{P}\left(F \geq F_{\text{hav}}\right).$$

Varianssianalyysin tulokset kootaan usein yhteen niin sanottuun ANOVA-taulukkoon 1.1, johon usein lasketaan myös lisäsarakkeeseen testin $\mathsf{p-arvo}$.

Tarkastellaan erityisesti kahden ryhmän tilannetta $k = 2$. Tässä tapauksessa voidaan näyttää mekaanisella kaavanmuokkauksella, että

$$\mathsf{SSA} = \sum_{i=1}^{2} m_i(\bar{y}_{i,\cdot} - \bar{y}_{\cdot,\cdot})^2 = \frac{(\bar{y}_{1,\cdot} - \bar{y}_{2,\cdot})^2}{\frac{1}{m_1} + \frac{1}{m_2}}.$$

Nyt F-testisuure on siis muotoa

$$F = \frac{\mathsf{SSA}/v_A}{s_E^2} = \frac{(\bar{y}_{1,\cdot} - \bar{y}_{2,\cdot})^2}{s_E^2\left(\frac{1}{m_1} + \frac{1}{m_2}\right)} \sim \mathsf{F}(1, v_E),$$

Vaihtelun lähde	SS	df	MS	F
Ryhmien välinen vaihtelu	SSA	v_A	s_A^2	s_A^2/s_E^2
Ryhmien sisäinen vaihtelu	SSE	v_B	s_E^2	
Kokonaisvaihtelu	SST	$n-1$		

Taulukko 1.1: ANOVA-taulukko

joka on tutun kahden riippumattoman otoksen t-testin testisuureen neliö. Tämä tarkoittaa sitä, että $\mathsf{t}(v_E)$-jakautuneen satunnaismuuttujan neliön jakauma on $\mathsf{F}(1, v_E)$-jakauma. Lisäksi nähdään että tämän kappaleen F-testi on ekvivalentti kaksitahoisen kahden riippumattoman otoksen t-testin kanssa.

Esimerkki 3. Tarkastellaan edellisen esimerkin havaintoaineistoa muovin elastisuudesta. Testataan tasolla 0.05 voidaanko hylätä nollahypoteesi, jonka mukaan vertailtavien ryhmien odotusarvot ovat samat.

(a) Tarkastellaan ainoastaan valmistustapoja 1 ja 2. Kriittinen arvo on 5.32. Nyt testisuureen havaittu arvo on

$$F_{\text{hav}} = \frac{\mathsf{SSA}/v_A}{\mathsf{SSE}/v_E} = \frac{1.225/1}{5.804/8} = 1.688491.$$

Koska havaittu testisuureen arvo on pienempi kuin kriittinen arvo, jää nollahypoteesi voimaan tasolla 0.05. Kootaan tulokset ANOVA-taulukkoon.

Vaihtelun lähde	SS	df	MS	F
Ryhmien välinen vaihtelu	1.225	1	1.225	1.688491
Ryhmien sisäinen vaihtelu	5.804	8	0.7255	
Kokonaisvaihtelu	7.029	9		

(b) Tarkastellaan ainoastaan valmistustapoja 1 ja 3. Nyt testisuureen havaittu arvo on

$$F_{\text{hav}} = \frac{\mathsf{SSA}/v_A}{\mathsf{SSE}/v_E} = \frac{3.6/1}{5.804/8} = 4.962095,$$

jota verrataan kriittiseen arvoon $f_{0.05}^{(1,8)} = 5.32$. Koska $F_{\text{hav}} < 5.32$, niin nollahypoteesi jää voimaan tasolla 0.05.

(c) Tarkastellaan ainoastaan valmistustapoja 1 ja 4. Nyt testisuureen havaittu arvo on

$$F_{\text{hav}} = \frac{\mathsf{SSA}/v_A}{\mathsf{SSE}/v_E} = \frac{7.225/1}{5.804/8} = 9.958649,$$

jota verrataan kriittiseen arvoon $f_{0.05}^{(1,8)} = 5.32$. Koska $F_{\text{hav}} > 5.32$, niin voidaan nollahypoteesi hylätä merkitsevyystasolla 0.05.

(d) Tarkastellaan kaikkia valmistustapoja. Nyt testisuureen havaittu arvo on

$$F_{\text{hav}} = \frac{\mathsf{SSA}/v_A}{\mathsf{SSE}/v_E} = \frac{7.9/3}{9.268/16} = 4.546108,$$

jota verrataan kriittiseen arvoon $f_{0.05}^{(3,16)} = 3.24$. Koska $F_{\text{hav}} > 3.24$, voidaan nollahypoteesi hylätä merkitsevyystasolla 0.05.

Vaihtelun lähde	SS	df	MS	F
Ryhmien välinen vaihtelu	7.9	3	2.6333	4.546108
Ryhmien sisäinen vaihtelu	9.268	16	0.5793	
Kokonaisvaihtelu	17.168	19		

1.2 Parivertailut

Testasimme kaikkien ryhmien odotusarvojen yhtäsuuruutta $\mu_1 = \cdots = \mu_k$ ja F-testimme antoi merkitsevän p-arvon nollahypoteesia vastaan. Tämän jälkeen meille jää tutkittavaksi, että mitä tämä tarkoittaa. Merkitsevä erohan voi olla minkälaista tahansa, esim. μ_1:n ja kaikkien muiden ryhmien odotusarvojen välillä tai μ_2:n ja kaikki muiden ryhmien odotusarvojen välillä. Vastahypoteesissa ei ole mitenkään määritelty eron tyyppiä vaan riittävää on, että jonkinnäköistä eroa löytyy. Mahdollisia eroja voidaan tarkastella muodostamalla ryhmien sijaintiparametreista $\mu_1, \ldots, \mu_k$ useita lineaarisia yhdistelmiä. Erityisesti usein tarkastellaan niin sanottuja **kontrasteja**. Parametrien $\mu_1 \ldots, \mu_k$ kontrasti on mikä tahansa niiden lineaarinen kombinaatio δ, joka on muotoa

$$\delta = \sum_{i=1}^{k} c_i \mu_i,$$

jossa kertoimien c_i täytyy toteuttaa ehdot

$$\sum_{i=1}^{k} c_i = 0, \quad \text{ja} \quad \sum_{i=1}^{k} |c_i| > 0.$$

Eli kertoimet summautuvat nollaksi ja vähintään kaksi kertoimista ovat erisuuria kuin nolla.

Esimerkki 4. Yksinkertaisimpia kontrasteja ovat sijaintiparametrien parittaiset vertailut, joille kurssilla käytetään omaa merkintää $\delta_{k,l}$ ryhmien k ja l vertailussa. Kertoimet ja kontrasti ovat

$$\begin{aligned} &c_k = 1 \\ &c_l = -1, k \neq l \\ &c_i = 0, \text{ kun } i \neq k \text{ ja } i \neq l \\ &\delta_{kl} = \mu_k - \mu_l. \end{aligned}$$

Jos $\delta_{k,l} > 0$, niin $\mu_k > \mu_l$.

Esimerkki 5. Verrataan uuden tarkkaavaisuushäiröihin kehitetyn hoitomenetelmän tehoa kolmeen vanhaan standardimenetelmään. Oletetaan että tarkkaavaisuushäiriöitä voidaan mitata standarditestin antamalla pistearvolla, jonka pienempi arvo vastaa huonompaa tarkkaavaisuutta. Uudella menetelmällä hoidettujen potilaiden pistejakauman odotusarvoa merkitään symbolilla μ_1, kun taas μ_2, μ_3 ja μ_4 vastaavat standardimenetelmillä hoidettujen potilaiden pistejakaumien odotusarvoja. Nyt valitaan kontrasti, jonka määrittelee $c_1 = 3, c_2 = -1, c_3 = -1, c_4 = -1$, eli

$$\delta = 3\mu_1 - \mu_2 - \mu_3 - \mu_4.$$

Jos $\delta > 0$ niin keskimäärin uusi hoitomenetelmä antaa parempia pisteitä kuin standardimenetelmät.

Vaikka monimutkaiset kontrastit ovatkin oiva työkalu havaintoaineiston analysoinnissa, niin tämän kurssin puitteissa keskitymme ainoastaan parivertailuihin.

Kontrastien lineaarisuuden vuoksi parametrien δ estimaatit saadaan laskettua helposti eri ryhmien odotusarvojen estimaattien $\bar{y}_{i,\cdot}$ avulla

$$\hat{\delta} = \sum_{i=1}^{k} c_i \bar{y}_{i,\cdot}.$$

Estimaatin $\hat{\delta}$ keskivirhe saadaan laskettua kaavalla

$$\mathsf{SE}(\hat{\delta}) = s_E \sqrt{\sum_{i=1}^{k} \frac{c_i^2}{m_i}}.$$

Huomaa, että kaikille kontrasteille käytetään keskihajonnan estimaattia s_E, joka on laskettu koko aineistosta. Koska havainnot $y_{i,j}$ olivat olettamuksemme mukaisesti normaalijakautuneita, niin myös kontrastien estimaattorit $\hat{\delta}$ ovat normaalijakautuneita

$$\hat{\delta} \sim \mathsf{Normal}\left(\sum_{i=1}^{k} c_i \mu_i, s_E^2 \sum_{i=1}^{k} \frac{c_i^2}{m_i}\right).$$

Tämän vuoksi normalisoidut kontrastit ovat meidän mallissamme t-jakautuneita satunnaismuuttujia

$$\frac{\hat{\delta} - \delta}{\mathsf{SE}(\hat{\delta})} \sim \mathsf{t}(v_E),$$

ja δ:lle saadaan tutusti $100(1-\alpha)\%$-luottamusväli

$$\hat{\delta} \pm t_{\alpha/2}^{(v_E)} \mathsf{SE}(\hat{\delta}).$$

Tämä on siis yhden kontrastin luottamusväli. Yleensä sen sijaan että tarkastellaan yksittäisten parien eroa, tarkastelemme usein kaikkia mahdollisia pareittaisia eroja. Näitä on yhteensä $k(k-1)/2$ kpl. Kun tarkastellaan samanaikaisesti useita pareittaisia eroja (tai yleisemmin useita kontrasteja), tämä täytyy ottaa huomioon luottamusvälejä muodostaessa. Tätä voidaan perustella siten, että kun muodostetaan 95%-luottamusväli yhdelle parivertailulle, niin 5% tapauksista luottamusväli ei sisällä ryhmien odotusarvojen todellista eroa. Eli kahdessakymmenessä 95%-luottamusvälin muodostuksessa noin yksi on sellainen ettei se sisällä todellista ryhmien odotusarvojen eroa. Jos suoritamme samanaikaisesti 20 parivertailua, niin on suurella todennäköisyydellä ainakin yhdessä tapauksessa sattuman vuoksi luottamusväli on pielessä. Tämän ongelman vuoksi käytämme erilaisia korjauksia samanaikaisessa tarkastelussa.

Tämän kurssin puitteissa tutustumme kahteen korjausmenetelmään: **Bonferroni**- ja**Tukeyn HSD**-menetelmiin (Honestly Significant Difference, HSD). Bonferroni-menetelmä perustuu siihen että jokaisen tarkasteltavan parittaisen eron (r kpl) tarkastelu suoritetaan luottamustasolla $1-\alpha/r$, jolloin yhteensä kaikkien parittaisten erojen joukkoon liittyy luottamustaso $1-\alpha$. Eli Bonferroni-korjauksella laskettu luottamusväli yhdelle parivertailulle on

$$\hat{\delta} \pm t_{\alpha/(2r)}^{(v_E)} \mathsf{SE}(\hat{\delta}),$$

joka on hyvin konservatiivinen, ja kasvattaa luottamusväliä huomattavasti. Tukeyn HSD- menetelmää käytetään kun havaintoaineisto on **tasapainoinen**, eli jokainen ryhmä sisältää yhtä monta havaintoa, ja haluamme testata kaikkia mahdollisia parittaisia eroja. Tukeyn menetelmä perustuu ns. studentized range-jakaumaan. Luottamusväliä muodostaessa siis yksinkertaisesti käytämme studentized range-jakauman yläkvantiilia t-jakauman yläkvantiilin sijaan. Nyt $100\%(1-\alpha)$-luottamusväli samanaikaisille tarkasteluille on muotoa

$$\hat{\delta} \pm \frac{q(\alpha, k, v_E)}{\sqrt{2}} \mathsf{SE}(\hat{\delta}),$$

jossa $q(\alpha, k, v_E)$ on studentized range-jakauman α-yläkvantiili parametrilla k sekä vapausastein v_E. Erilaisia korjausmenetelmiä samanaikaisiin vertailuihin on hyvin suuri määrä, joista moni on optimaalinen jollekin tietylle tilanteelle.

Esimerkki 6. Halutaan tutkia neljän eri opetusmetodin vaikutuksia lukiolaisten matematiikan opetuksessa. 24 opiskelijaa jaettiin 6 hengen ryhmiin, joissa jokaisessa käytettiin eri opetusmetodia lukukauden ajan. Lukukauden lopussa opiskelijoille järjestettiin tasokoe, josta kerättiin havaintoaineisto

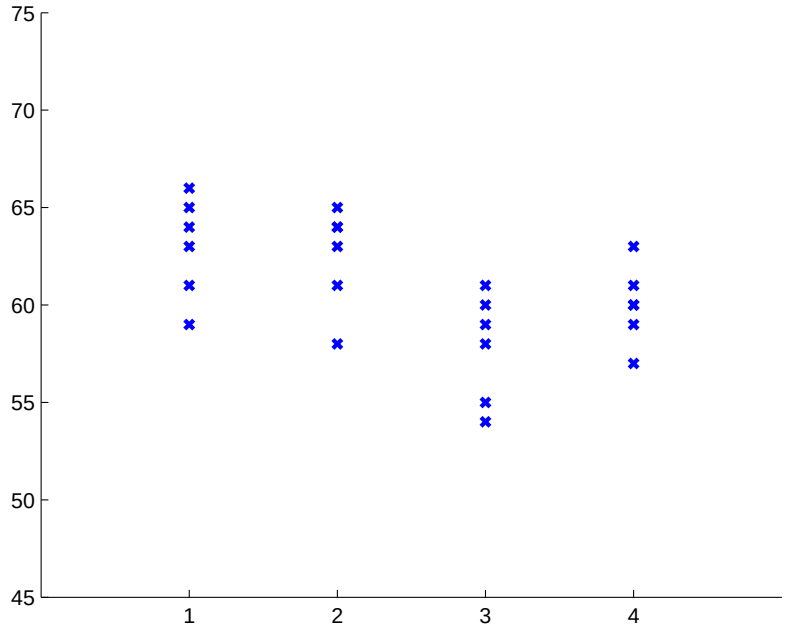
Opetusmetodi	**Pisteet tasokokeesta**					
1	59	63	65	61	64	66
2	58	61	64	63	65	64
3	54	60	55	58	59	61
4	60	59	57	60	61	63

Havainnollistetaan aineistoa ensin piirtämällä Kuva 1.2.

Olkoon μ_i opetusmetodilla i opetuksen saaneiden koululaisten pisteiden jakauman odotusarvo. Testataan hypoteesiparia

$$\mathsf{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$\mathsf{H}_v : \text{Kaikki odotusarvot } \mu_i \text{ eivät ole samoja.}$$

ANOVA-taulukko on muotoa



Kuva 1.2: Esimerkin 6 havaintoaineisto ryhmittäin.

Vaihtelun lähde	SS	df	MS	F
Ryhmien välinen vaihtelu	103	3	34.333	5.44
Ryhmien sisäinen vaihtelu	126.333	20	6.3167	
Kokonaisvaihtelu	229.333	23		

Luetaan $\mathsf{F}(v_1, v_2)$-jakauman 0.05-yläkvantiilitaulukosta kohdalta $v_1 = 3$ ja $v_2 = 20$ kriittinen arvo 3.10. Koska havaittu testisuureemme arvo $F_{\text{hav}} = 5.44$ kuuluu kriittiselle alueelle, niin eri ryhmien odotusarvojen välillä on merkitseviä eroja tasolla 0.05. Lähdetään tutkimaan havaintoaineistoa parivertailujen avulla. Koska havaintoaineisto on tasapainoinen, niin jokaisella parivertailulla on sama keskivirhe

$$\mathsf{SE}(\hat{\delta}_{i,j}) = s_E\sqrt{\frac{1}{6}+\frac{1}{6}} = \sqrt{6.3167}\sqrt{\frac{1}{3}} = 1.45.$$

Vapausasteita t-jakautuneella suureella on nyt $v_E = 20$, joten jokaisen (korjaamattoman) luottamusvälin yläkvantiili on $t^{(20)}_{0.025} = 2.086$.

Vertailukohde	Est.	alaraja (ei korjausta)	yläraja (ei korjausta)
$\mu_1 - \mu_2$	0.5	-2.53	3.53
$\mu_1 - \mu_3$	5.17	2.14	8.20
$\mu_1 - \mu_4$	3	-0.03	6.03
$\mu_2 - \mu_3$	4.67	1.64	7.70
$\mu_2 - \mu_4$	2.5	-0.53	5.53
$\mu_3 - \mu_4$	-2.17	-5.20	0.86

Muodostetaan myös korjatut luottamusvälit Bonferroni sekä Tukeyn HSD-korjauksilla. Bonferronin luottamusväli lasketaan kaavalla

$$\hat{\delta} \pm t^{(20)}_{0.05/(12)} \mathsf{SE}(\hat{\delta}),$$

ja koska $t^{(20)}_{0.05/(12)} = t^{(20)}_{0.0042}$:n arvoa ei löydy taulukosta, käytämme approksimaatiota $t^{(20)}_{0.005}$. Tukeyn HSD on muotoa

$$\hat{\delta} \pm \frac{q(0.05, 4, 20)}{\sqrt{2}} \mathsf{SE}(\hat{\delta}).$$

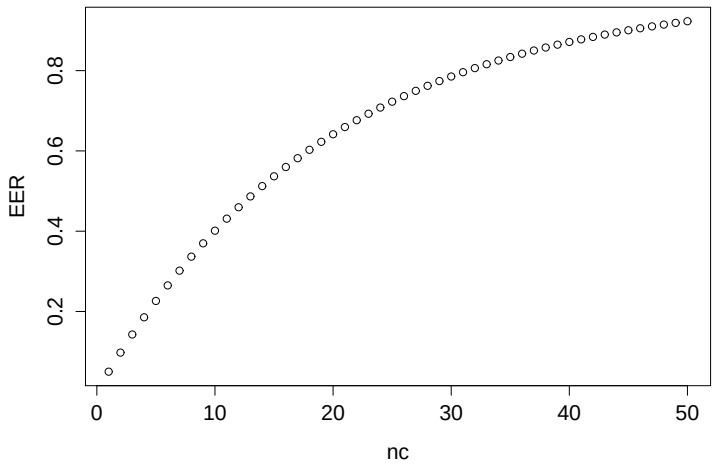
Saamme arvot

Vertailukohde	Est.	alaraja (Bonferroni)	yläraja (Bonferroni)	alaraja (Tukey HSD)	yläraja (Tukey HSD)
$\mu_1 - \mu_2$	0.5	-3.63	4.62	-3.56	4.56
$\mu_1 - \mu_3$	5.17	1.04	9.30	1.11	9.23
$\mu_1 - \mu_4$	3	-1.12	7.13	-1.06	7.06
$\mu_2 - \mu_3$	4.67	0.54	8.80	0.61	8.73
$\mu_2 - \mu_4$	2.5	-1.63	6.63	-1.56	6.56
$\mu_3 - \mu_4$	-2.17	-6.30	1.96	-6.23	1.89

Luottamusvälien laskemisen lisäksi tai sijaan, tehdään usein parivertailuja (yleisemmin vertailtaisiin useita erilaisia kontrasteja) tilastollisten testien avulla tietyn kahden ryhmän odotusarvojen erojen tutkimiseksi. Testisuureena käytetään kahden riippumattoman otoksen t-testin muuttujaa

$$T = \frac{\hat{\delta}_{i,j}}{\mathsf{SE}(\hat{\delta})}.$$

Samoin kuin luottamusvälien laskemisessa, usean tilastollisen testin tekeminen samanaikaisesti saattaa johtaa ongelmiin. Samanaikaisessa parivertailussa ongelmana on tyypin I virheen tekemisen kasvaminen. Tyypin I virhe tarkoittaa, että hylkäämme nollahypoteesin, vaikka se on tosi, eli tyypin I virhe on testin p − arvo. Jos tehdään useita vertailuja käyttäen tilastollisesti merkitsevän eron rajana p − arvoa 0.05, keskimäärin joka $20.$ vertailussa esiintyy tyypin I virhe, eli saamme merkitsevän tuloksen sattuman kautta. Tästä johtuen ei ole mielekästä tehdä samanaikaisesti suurta määrää parittaisia vertailuja. Tyypin I kokonaisvirheen kasvamisen ylärajaa havainnollistetaan Kuvassa 1.3. Tämä yläraja saavutettaisiin, jos jokainen parivertailu olisi tilastollisesti riippumaton muista parivertailuista. Näin ei kuitenkaan oikeasti ole, sillä esimerkiksi erojen $\mu_1 - \mu_2$ ja $\mu_1 - \mu_3$ tarkastelut ovat tilastollisesti riippuvia.



Kuva 1.3: Monivertailujen kokonaisvirheen esiintymisasteen (experimentwise error rate, EER) ylärajan kasvaminen vertailujen määrän (nc) funktiona (EER $\leq 1 - (1 - \alpha)^{\text{nc}}$), kun testi tehdään tasolla $\alpha = 0.05$.

Tyypin I virhettä voidaan kontrolloida kahdella vaihtoehtoisella tavalla. Ensimmäisessä asetetaan tutkimuskysymykset tarkkaan etukäteen ja tehdään vain tutkimuskysymysten mukaiset vertailut. Tässä tapauksessa meillä on vähemmän eri vertailuja ja näin myös vähemmän satunnaisuudesta aiheutuvia löydöksiä. Tätä tapaa voidaan pitää jossain mielessä luontevampana pienissä kokeellisissa tutkimuksissa, etenkin jos tutkimuskysymykset ovat selkeitä. Parittaisia samanaikaisia testauksia, ilman erityisiä korjausmenetelmiä kutsutaan joskus Tukeyn LSD-testiksi (least significant difference, LSD). Toisessa tavassa vertaillaan esimerkiksi kaikkia ryhmiä toi-

siin ryhmiin, mutta korjataan parivertailujen p − arvoja ottamalla huomioon vertailujen määrä. Bonferroni-menetelmässä valittujen parien erotuksien testaus suoritetaan t-testillä ja havaitut merkitsevyystasot, eli p − arvot kerrotaan testien kokonaismäärällä. Jos kaikkia parien erotuksia testataan, on Tukeyn HSD menetelmä optimaalinen. Tukeyn HSD-menetelmässä lasketaan havaittu t-testisuureen arvo, jonka itseisarvoa verrataan studentized-range jakauman α-yläkvantiiliin. Eri korjausmenetelmiä on hyvin monenlaisia, jotka monet toimivat erinomaisesti tietynlaisissa tilanteissa (Scheffen, Dunnettin, Duncanin, Sidakin, ym. menetelmät).

On tärkeää huomata että edellä mainitut menetelmät perustuvat oletettuun malliimme, jossa oletimme että havaintoaineisto on normaalijakautunutta, sekä kaikkien ryhmien havainnoilla on sama varianssi (**varianssien homogeenisuus**). Molempia voidaan tutkailla tilastollisilla testeillä, joiden teoriaan emme tämän kurssin puitteissa perehdy. Jos esimerkiksi Shapiro-Wilkin, tai Kolmogorov-Smirnovin testi hylkää hypoteesin normaaliudesta, tutkimme mahdollisuutta käyttää esimerkiksi epäparametrisia testejä. Ryhmävarianssien homogeenisuutta voidaan tutkia esimerkiksi Bartlettin tai Levenen testillä. Havaintoaineiston normaalijakautuneisuutta voitaisiin tutkia myös havaintojen jäännösten (residuaalien)

$$e_{i,j} = y_{i,j} - \bar{y}_{i,\cdot}$$

avulla. Jäännöksiä tarkasteltaessa teemme esimerkiksi Shapiro-Wilkin testin kaikille residuaaleille yhtäaikaisesti. Tämä on mahdollista koska jäännökset ovat havaintojen erotuksia ryhmän otoskeskiarvosta, niistä on siis tietyssä mielessä poistettu yksittäisen ryhmän vaikutus.

Jos ryhmillä on tilastollisesti merkitsevästi erisuuret varianssit, mutta aineistot ovat normaalijakautuneita, voimme ottaa käyttöön esimerkiksi robustin testin. Tässä tapauksessa voisimme käyttää Brown-Forsythen tai Welchin testiä. Jos nämä testit antavat merkitsevän eron ryhmien odotusarvojen välille, voidaan tarkastella aineistoa tarkemmin parivertailujen ja Tamhanen korjausten avulla.

Esimerkki 7. Niin sanottu Gunningin fog-indeksi mittaa englanninkielisen tekstin vaikealukuisuutta. Indeksi lasketaan kaavalla

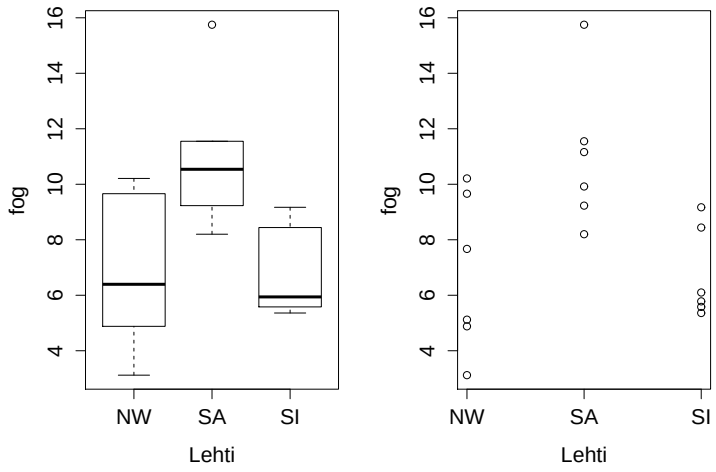
$$\begin{aligned}\text{fog} = 0.4 \cdot (&\text{keskimääräisten sanojen lukumäärä lauseessa} \\ &+ \text{yli } 3\text{-tavuisten sanojen prosenttiosuus}).\end{aligned}$$

Shuptrine ja McVicker (1981) tekivät tutkimuksen, jossa verrattiin kolmen lehden luettavuutta, valitsemalla satunnaisesti 6 mainosta kustakin lehdes-

tä ja määrittämällä niihin liittyvät fog-indeksit. Havaintoaineisto on

Lehti	**fog-indeksi**					
Newsweek	10.21	9.66	7.67	5.12	4.88	3.12
Scientific American	15.75	11.55	11.16	9.92	9.23	8.20
Sports Illustrated	9.17	8.44	6.10	5.78	5.58	5.36,

jota ollaan havainnollistettu Kuvassa 1.4. Muodostetaan hypoteesit



Kuva 1.4: Esimerkin 7 havaintoaineisto.

$\mathsf{H}_0 : \mu_1 = \mu_2 = \mu_3$
$\mathsf{H}_v :$ Lehtien fog-indeksit eivät ole samat

Tässä vaiheessa tarkasteltaisiin mallimme oletuksien paikkaansapitävyyttä. Esimerkksi R:llä voidaan laskea Shapiro-Wilkin testin merkitsevyystasot (`shapiro.test()`) ryhmien havainnoille. Koko aineistolle Shapiro-Wilkin $\mathsf{p-arvo} = 0.507$, lisäksi Shapiro-Wilk antaa ryhmien aineistoille $\mathsf{p-arvot}$ 0.534 (Newsweek), 0.375 (SA), 0.065 (SI), joten aineistoja voidaan pitää normaalijakautuneina. Varianssin homogeenisuutta tarkastellaan Bartlettin testillä (`bartlett.test()`), joka antaa merkitsevyystason 0.4831. Täten ryhmien varianssien erot eivät ole tilastollisesti merkitseviä esimerkiksi tasolla 0.05. Muodostetaan aineistolle ANOVA-taulukko.

Vaihtelun lähde	SS	df	MS	F
Ryhmien välinen vaihtelu	70.929	2	35.464	5.976
Ryhmien sisäinen vaihtelu	89.013	15	5.934	
Kokonaisvaihtelu	159.941	17		

F-jakauman 0.05-yläkvantiilitaulukosta voidaan vapausasteiden $v_A = 2$ ja $v_E = 15$ kohdalta lukea kriittinen piste $f_{0.05}^{(2,15)} = 3.68$ F-testille tasolla 0.05. Koska havaittu testisuureen arvo $F_{\text{hav}} > f_{0.05}^{(2,15)}$, niin voimme hylätä nollahypoteesin tasolla 0.05. Koska F-testin perusteella hylkäämme nollahypoteesin tasolla 0.05, tarkastelemme erojen tyyppejä parivertailujen avulla. Parivertailujen tulokset ovat

	Lehti i	**Lehti** j	$\hat{\delta}_{i,j}$	$\mathsf{SE}(\hat{\delta}_{i,j})$	95%**-CI ala**	95%**-CI ylä**	$\mathsf{p}-\mathsf{arvo}$
LSD:	NW	SA	-4.192	1.406	-7.189	-1.195	$\in (0.01, 0.02)$
	NW	SI	0.038	1.406	-2.959	3.035	$\in (0.5, 1)$
	SA	SI	4.230	1.406	1.233	7.227	$\in (0.005, 0.01)$
B.rroni:	NW	SA	-4.192	1.406	-7.851	-0.532	$\in (0.03, 0.06)$
	NW	SI	0.038	1.406	-3.621	3.698	≈ 1
	SA	SI	4.230	1.406	0.570	7.890	$\in (0.015, 0.03)$
HSD:	NW	SA	-4.192	1.406	-7.844	-0.5388625	$\in (0, 0.05)$
	NW	SI	0.038	1.406	-3.614	3.691	$\in (0.05, 1)$
	SA	SI	4.230	1.406	0.577	7.882	$\in (0, 0.05)$.

Riippumatta siitä mitä korjausmenetelmää käytetään (tai LSD:n tapauksessa ei käytetä), parien erotuksien ja keskivirheiden estimaatit ovat

$$\hat{\delta}_{i,j} = \bar{y}_{i,\cdot} - \bar{y}_{j,\cdot}$$
$$\mathsf{SE}(\hat{\delta}_{i,j}) = s_E\sqrt{\frac{1}{3}}.$$

Tukeyn LSD-testille parivertailujen luottamusvälit lasketaan tutulla tavalla

$$\hat{\delta}_{i,j} \pm t_{\alpha/2}^{(v_E)}\mathsf{SE}(\hat{\delta}),$$

ja Bonferroni-korjatulle testille

$$\hat{\delta}_{i,j} \pm t^{(v_E)}_{0.05/(2\cdot 3)}\mathsf{SE}(\hat{\delta}).$$

Koska $0.05/(2 \cdot 3) = 0.0083$, käytämme luottamusvälien ja $\mathsf{p}-$arvon laskemisessa 0.01-yläkvantiilia. Tukeyn HSD-testissä laskemme samanaikaiset luottamusvälit vertailuille kaavalla

$$\hat{\delta}_{i,j} \pm \frac{q[0.05, 3, 15]}{\sqrt{2}}\mathsf{SE}(\hat{\delta}).$$

Katsomme $\mathsf{p}-$arvon yksinkertaisesti luottamusvälin avulla. Näin pystymme selvittämään yksinkertaisesti onko $\mathsf{p}-$arvo suurempi vai pienempi kuin 0.05.

1.3 Kurskal-Wallisin H-testi

Samoin kuin voidaan ajatella että F-testi on t-testin yleistys useamman kuin kahden ryhmän sijaintiparametrien vertailuun, voidaan ajatella että **Kruskal-Wallisin H-testi** on Mann-Whitneyn U-testin yleistys useamman kuin kahden ryhmän vertailuun. Kuten U-testi, myös Kruskal-Wallis on siis järjestyslukuihin perustuva epäparametrinen testi, eikä vaadi oletusta normaalijakautuneisuudesta. Kruskal-Wallisin käyttö on suotavaa jos ryhmien havaintoaineistoja ei saada kaikkia samalla muunnoksella lähelle normaalijakaumaa. Oletetaan, että ryhmien havainnot ovat toisistaan riippumattomia ja ryhmien jakaumat ovat identtisiä muuten paitsi jakauman mediaanin suhteen. Testin tulkinta on yksikäsitteistä vain siinä tapauksessa kun jakaumat ovat samanmuotoisia.

Tarkastellaan hypoteesiparia

H_0 : Kaikki k ryhmää ovat mediaanin puolesta identtisiä

H_v : Kaikki ryhmät eivät ole mediaanin puolesta identtisiä.

Kuten teimme U-testissä, korvaamme nyt kaikki havainnot ryhmään katsomatta järjestysluvuilla $r_{i,j}$, jotka kattavat siis luvut $1, \ldots, n$. Jos havaintoaineistossa on sidoksia, eli yhtäsuuria lukuja, järjestysluvut korvataan sidoksien järjestyslukujen keskiarvolla. Eli jos $3., 4.$ ja $5.$ suurimmat luvut ovat sidoksia, jokainen niistä saa järjestysluvuksi $r_{i,j} = (3 + 4 + 5)/3 = 4$. Testisuure Kruskal-Wallisin testissä, jos havaintoaineistossa ei ole sidoksia, on muotoa

$$H = \left(\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{m_i}\right) - 3(n+1),$$

jossa

$$R_i = \sum_{j=1}^{m_i} r_{i,j}$$

on ryhmän i järjestyslukujen summa. Jos datassa on sidoksia, testisuuretta korjattaisiin jakamalla H luvulla

$$c = 1 - \frac{\sum_{i=1}^{G} t_i^3 - t_i}{n^3 - n},$$

jossa G on eri sidosryhmien määrä ja t_i on tiettyyn sidosryhmään liittyvien lukujen määrä ryhmässä i. Merkitään korjattua testisuuretta

$$H_c = \frac{H}{c}.$$

Jos sidoksia on havaintoaineistossa vähän, korjauksella ei ole kovinkaan suurta vaikutusta testisuureen arvoon.

Nyt jos jakaumat ovat hypoteesin H_0 mukaisia, testisuure H_c noudattaa likimain khin-neliö-jakaumaa vapausastein $k-1$, eli

$$H_c \sim \chi^2(k-1).$$

Approksimaatio on yleensä hyvin tarkka jos jokaisessa ryhmässä on havaintoja luokkaa $m_i \geq 5$.

Jos Kruskal-Wallis antaa tilastollisesti merkitsevän $\mathsf{p-arvon}$, niin voimme lähteä tekemään parivertailuja ryhmien välillä kuten aiemmin. Koska emme ole tehneet oletusta normaalijakaumasta, mutta sen sijaan olemme olettaneet että ryhmien jakaumat ovat samanmuotoisia, voimme käyttää Mann-Whitneyn U-testiä parivertailujen tekemiseen. Samoin kuten aiemmin, jos vertailuja on paljon, on hyvä korjata parivertailujen $\mathsf{p-arvot}$. Yksinkertaisin korjaus, jota voidaan käyttää tässä tapauksessa on Bonferronin korjaukset, eli kerromme lasketut merkitsevyystasot vertailujen lukumäärällä.

Jonckheere-Terpstranin testiä voitaisiin käyttää jos vastahypoteesi olisi muodostettu tarkemmin kuin yllä siten, että olettaisimme odotusarvojen välille monotonisen järjestyksen. Toisin sanoen testaisimme nollahypoteesiamme hypoteesia

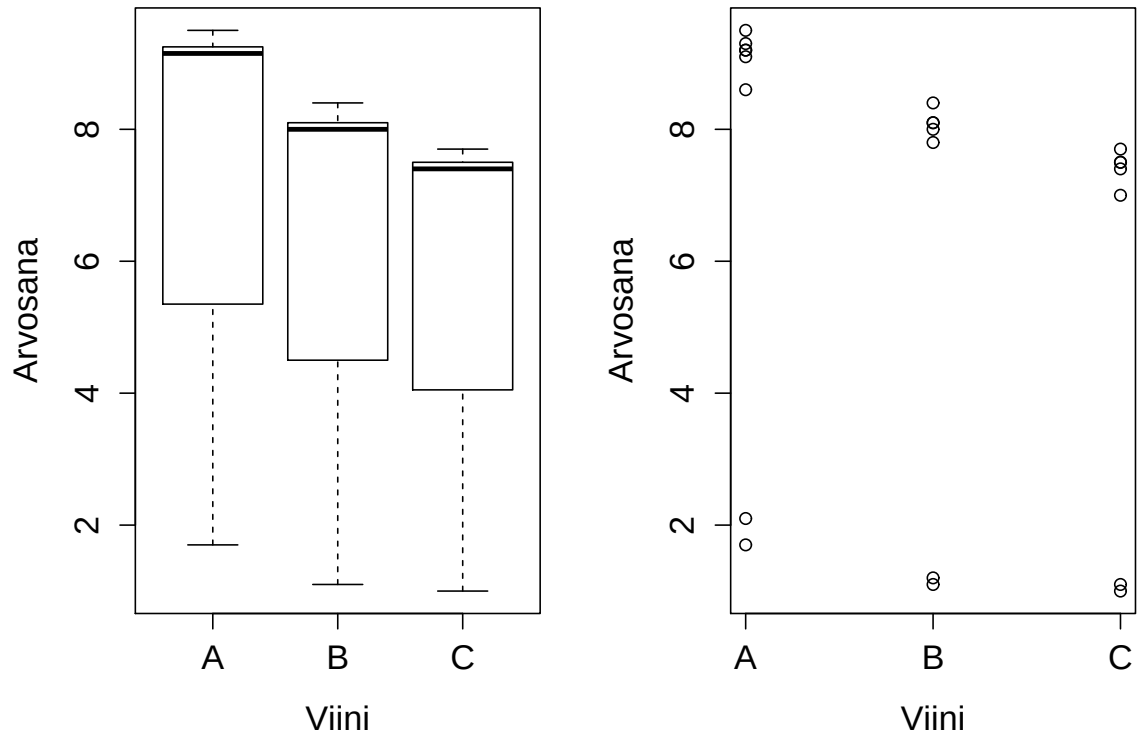
$$\begin{aligned} \mathsf{H}_v: &\quad \mu_1 \geq \mu_2 \geq \ldots \geq \mu_k \quad \text{tai} \\ \mathsf{H}_v: &\quad \mu_1 \leq \mu_2 \leq \ldots \leq \mu_k \end{aligned}$$

vastaan. Jonckheere-Terpstran testin yksityiskohtiin emme paneudu tämän kurssin puitteissa.

Esimerkki 8. Halutaan tutkia eroavatko kolme eri viiniä arvostelijoilta saatavien keskimääräisten pisteiden suhteen. Kullekin viinille on muodostettu oma asiantuntijaraati, eli kukin arvostelija on maistanut ja arvostellut vain yhtä viiniä. Keräämme arvostelijoilta havaintoaineiston

Viini	**Arvosana**							
A-luokan ranskalainen	1.70	2.10	8.60	9.10	9.20	9.20	9.30	9.50
Chilestä hyvää	1.10	1.20	7.80	8.00	8.10	8.10	8.40	
Erkin pikakivääri (Unkari)	1.00	1.10	7.00	7.40	7.50	7.50	7.70	

jota on havainnollistettu Kuvassa 1.5. Muista piirtää aina havaintoaineistosta kuva ennen analyysiä. Kuvasta näemme että havaintoaineisto ei näytä normaalijakautuneelta, mutta keskenään jakaumat ovat hyvin samannäköisiä.



Kuva 1.5: Esimerkin 8 havaintoaineisto. (A:A-luokan ranskalainen, B:Chilestä hyvää, C:Erkin pikakivääri)

Ilman realisista oletusta normaalijakautuneisuudesta (SPSS:llä tehty Kolmogorov-

Smirnovin normaalisuustesti hylkää kaikkien ryhmien normaalijakautuneisuushypoteesin merkitsevyystasolla $\alpha_i \leq 0.002$), emme käytä F-testiä viinien keskimääräisten pisteiden vertailuun, vaan käytämme nyt epäparametrista Kruskal-Wallisin H-testiä. Ensin laitamme havainnot suuruusjärjestykseen

Viini	**Arvosana**								R_i
A-luokan ranskalainen	1.70	2.10	8.60	9.10	9.20	9.20	9.30	9.50	
r_{1j}	5	6	17	18	19.5	19.5	21	22	128
Chilestä hyvää	1.10	1.20	7.80	8.00	8.10	8.10	8.40		
r_{2j}	2.5	4	12	13	14.5	14.5	16		76.5
Erkin pikakivääri (Unkari)	1.00	1.10	7.00	7.40	7.50	7.50	7.70		
r_{3j}	1	2.5	7	8	9.5	9.5	11		48.5.

Lasketaan H-testisuure

$$\begin{aligned} H &= \left(\frac{12}{22(22+1)} \sum_{i=1}^{3} \frac{R_i^2}{m_i} \right) - 3(22+1) \\ &= \frac{12}{506} \left(\frac{128^2}{8} + \frac{76.5^2}{7} + \frac{48.5^2}{7} \right) - 69 \\ &= 7.36533. \end{aligned}$$

Eri sidosryhmiä on $G = 4$ kappaletta (havainnot 1.10, 7.50, 8.10 ja 9.20) ja jokaisessa sidosryhmässä on $t_i = 2$ havaintoa. Näiden lukujen avulla voimme laskea korjauskertoimen

$$c = 1 - \frac{2^3 - 2 + 2^3 - 2 + 2^3 - 2 + 2^3 - 2}{22^3 - 22} = 0.9977414,$$

joka on pienellä sidoksien määrällä hyvin lähellä lukua yksi. Korjattu H-testisuure on

$$H_c = \frac{H}{c} = \frac{7.36533}{0.9977414} = 7.382003.$$

Vapausasteita tekijällä on nyt $3 - 1 = 2$kpl, joten testisuureen H jakauma noudattaa siis likimäin jakaumaa $\chi^2(2)$. Testisuure osuu $\chi^2(2)$-jakaumassa kvantiilien $q^{(2)}_{0.025}$ ja $q^{(2)}_{0.01}$ väliin. Tästä seuraa suoraan, että testin $\mathsf{p} - \mathsf{arvo}$ on välillä $(0.01, 0.025)$ Tämä on tarpeeksi pieni että voimme hylätä nollahypotesin tasolla 0.025.

Tarkastellaan parittaisia ryhmien eroja Mann-Whitneyn U-testin avulla. Muodostetaan parittaisille ryhmille järjestyslukutaulukot.

Viini	**Arvosana**								R_i
A-luokan ranskalainen	1.70	2.10	8.60	9.10	9.20	9.20	9.30	9.50	
r_{1j}	3	4	10	11	12.5	12.5	14	15	82
Chilestä hyvää	1.10	1.20	7.80	8.00	8.10	8.10	8.40		
r_{2j}	1	2	5	6	7.5	7.5	9		38

Mahdolliset U-testisuureet ovat

$$U_1 = 82 - \frac{1}{2}8(8+1) = 46$$
$$U_2 = 38 - \frac{1}{2}7(7+1) = 10,$$

joista valitaan pienempi, eli $U_2 = 10$. Nyt $k_1 = 8$ ja $k_2 = 7$, joten kaksitahoisen testin $\mathsf{p-arvo} = 2 \cdot 0.020 = 0.040$. Eli voimme tasolla 0.04 sanoa että A-luokan ranskalaisen ja Chilestä hyvää viinien välillä on eroa. Jos käytämme konservatiivista Bonferronin korjausta, niin saamme korjatuksi $\mathsf{p-arvoksi}$ $3 \cdot 0.040 = 0.12$.

Viini	**Arvosana**								R_i
A-luokan ranskalainen	1.70	2.10	8.60	9.10	9.20	9.20	9.30	9.50	
r_{1j}	3	4	10	11	12.5	12.5	14	15	82
Erkin pikakivääri (Unkari)	1.00	1.10	7.00	7.40	7.50	7.50	7.70		
r_{3j}	1	2	5	6	7.5	7.5	9		38.

Mahdolliset U-testisuureet ovat

$$U_1 = 82 - \frac{1}{2}8(8+1) = 46$$
$$U_2 = 38 - \frac{1}{2}7(7+1) = 10,$$

joista valitaan pienempi, eli $U_2 = 10$. Nyt $k_1 = 8$ ja $k_2 = 7$, joten kaksitahoisen testin $\mathsf{p-arvo} = 2 \cdot 0.020 = 0.040$. Eli voimme tasolla 0.04 sanoa että A-luokan ranskalaisen ja Erkin pikakivääri viinien välillä on eroa. Jos käytämme konservatiivista Bonferronin korjausta, niin saamme korjatuksi $\mathsf{p-arvoksi}$ $3 \cdot 0.040 = 0.12$.

Viini	**Arvosana**							R_i
Chilestä hyvää	1.10	1.20	7.80	8.00	8.10	8.10	8.40	
r_{2j}	2.5	4	10	11	12.5	12.5	14	66.5
Erkin pikakivääri (Unkari)	1.00	1.10	7.00	7.40	7.50	7.50	7.70	
r_{3j}	1	2.5	5	6	7.5	7.5	9	38.5.

Mahdolliset U-testisuureet ovat

$$U_1 = 82 - \frac{1}{2}8(8+1) = 38.5$$
$$U_2 = 38 - \frac{1}{2}7(7+1) = 10.5,$$

joista valitaan pienempi, eli $U_2 = 10.5$. Koska luemme merkitsevyystason kokonaislukuja sisältävästä taulukosta, täytyy meidän pyöristää testisuureen arvo. Pyöristämme $U_2 \approx 10$, eli kohti vähemmän merkitsevää $\mathsf{p} - \mathsf{arvoa}$. Nyt $k_1 = 7$ ja $k_2 = 7$, joten kaksitahoisen testin $\mathsf{p} - \mathsf{arvo} = 2 \cdot 0.036 = 0.072$. Eli voimme tasolla 0.05 sanoa että Chilestä hyvää ja Erkin pikakivääri viinien välillä ei ole eroa. Jos käytämme konservatiivista Bonferronin korjausta, niin saamme korjatuksi $\mathsf{p} - \mathsf{arvoksi}$ $3 \cdot 0.040 = 0.216$.

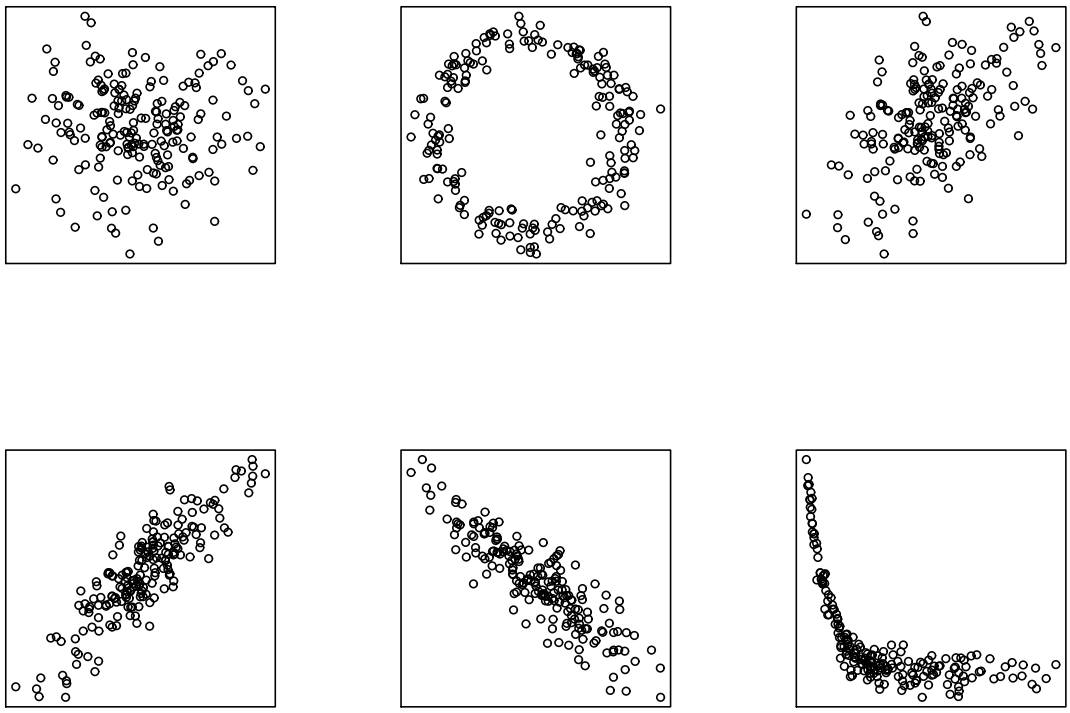
1.4 Korrelaatio

Kahden muuttujan yhteyksien tutkimiseksi (yleensä ei-kokeellisessa tutkimuksessa) usein tarkastellaan näiden muuttujien välistä korrelaatiota. Korrelaatiokerrointa käytettäessä voidaan selittää vaihtelua ja yhteyksiä muuttujien välillä, mutta usein ei ole perusteltua tehdä voimakkaita kausaali- eli syy-seurauspäätelmiä.

Oletetaan että kaksi numeerista satunnaismuuttujaa X ja Y ovat vasteita ja olemme keränneet n:n arvoparin havaintoaineiston

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Jos X:llä ja Y:llä on mahdollisia erisuuria arvoja vain vähän, niin voimme tutkia aineistoa kaksiulotteisten frekvenssitaulujen avulla. Kun otetaan havaintojen kvantitatiivisuus huomioon, tai erisuuria arvoja on hyvin paljon, kannattaa aineistoa tutkia sirontakuvioilla (scatter plot, scatter diagram).



Kuva 1.6: Joitakin sirontakuvion muotoja.

Sirontakuvio on pisteistä $(x_i, y_i), i = 1, \ldots, n$ muodostuva yleiskuva muuttujien yhteisjakaumasta. Mielenkiintoista kuvassa on sironnan määrä ja muoto. Erilaisia sirontakuvioita löytyy Kuvasta 1.6.

Havaintoaineiston analysointi on hyvä aloitettaa aineiston graafisesta tarkastelusta, mutta aineistoa on mahdollista luonnehtia myös korrelaatiokertoimien avulla. Tavallinen muuttujien yhteyttä kuvaava suure on **Pearsonin tulomomenttikorrelaatiokerroin**, joka on oikeastaan muuttujien lineaarisen yhteyden voimakkuuden ja suunnan mitta. Kaikki pisteet (x_i, y_i) ovat samalla suoralla, kun

$$(x_i - \bar{x})(y_j - \bar{y}) = (x_j - \bar{x})(y_i - \bar{y}), \quad i, j = 1, \ldots, n, \quad i \neq j.$$

Pisteiden kokonaisvaihtelua tästä oletuksesta voidaan mitata neliösummalla

$$q^2 = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \left[(x_i - \bar{x})(y_j - \bar{y}) - (x_j - \bar{x})(y_i - \bar{y}) \right]^2,$$

joka voidaan ilmaista otosvarianssien ja Pearsonin tulomomenttikorrelaa-

tiokertoimen r_P avulla

$$q^2 = (n-1)^2 s_X^2 s_Y^2 (1 - r_P^2).$$

Kerroin on muotoa

$$r_P = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2} \sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}},$$

jossa

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

on (X, Y):n otoskovarianssi ja s_X ja s_Y ovat muuttujien otosvarianssit. Käytännössä tulomomenttikorrelaatiokerroin mittaa kuinka lähelle suoraa viivaa sirontakuvion pisteet tulevat. Pearsonin tulomomenttikorrelaatiokertoimen voidaan osoittaa saavan arvoja

$$-1 \leq r_P \leq 1,$$

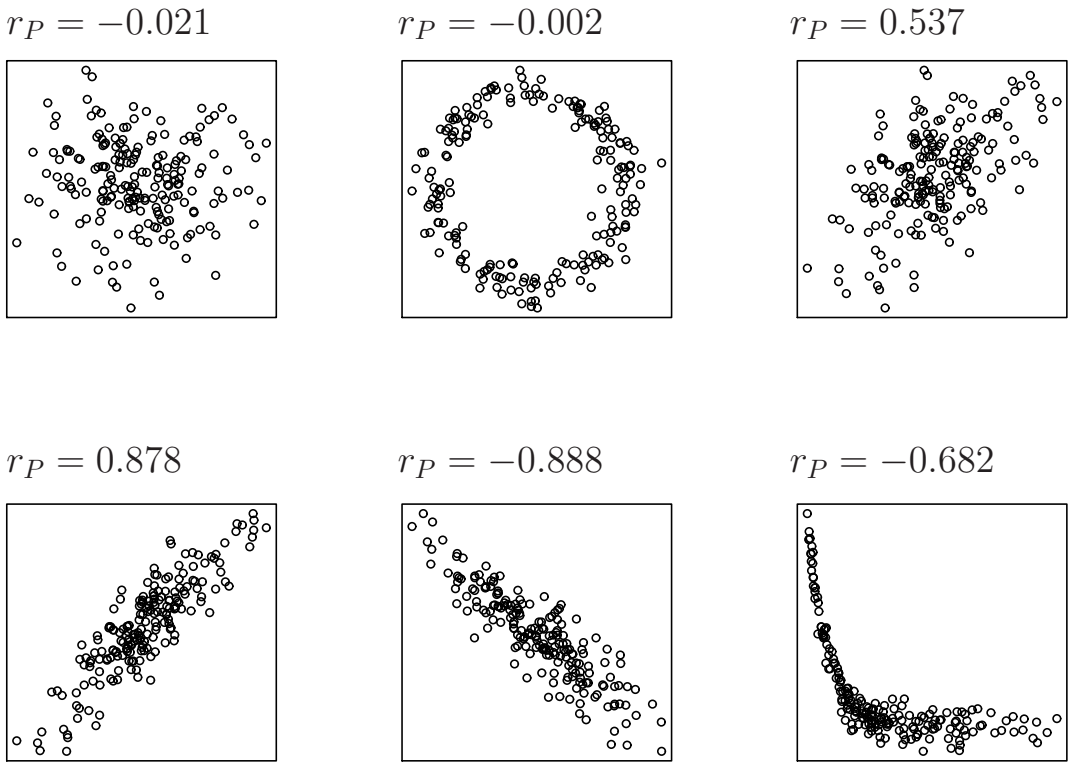
siten että mitä lähempänä kerroin on nollaa, sitä vähemmän muuttujien välillä on korrelaatiota. Jos kerroin on pienempi kuin 0, on korrelaatio negatiivista ja jos kerroin on suurempi kuin 0, on korrelaatio positiivista. Eri korrelaatiokertoimen arvoja on laskettu Kuvan 1.7 havaintoaineistoille. Yleisesti eri korrelaatiokertoimille r suuretta r^2 kutsutaan selitysasteeksi ja lukua $100\% r^2$ selitysprosentiksi.

On tärkeää huomata, että on olemassa muitakin korrelaatiokertoimia kuin Pearsonin. Jos havaintoaineisto ei selvästikään noudata normaalijakaumaa, tai muuttujat ovat järjestysasteikollisia, ja mielenkiintoinen riippuvuuden muoto on monotonisesti positiivinen tai negatiivinen riippuvuus (jos X kasvaa niin Y kasvaa, tai jos X kasvaa niin Y pienenee), niin voimme käyttää havaintojen järjestyslukuja havaintojen arvojen sijaan ja laskea **Spearmanin järjestyskorrelaatiokertoimen**. Spearmanin järjestyskorrelaatiokerrointa r_S laskiessa järjestämme havainnot x_i ja y_j ja saamme järjestyslukuparit (x_i', y_i'). Jälleen sidokset korvataan vastaavien järjestyslukujen keskiarvolla. Kerroin on samaa muotoa kuin Pearsonin kerroin

$$r_S = \frac{\sum_{i=1}^n (x_i' - \bar{x'})(y_i' - \bar{y'})}{\sqrt{\sum_{j=1}^n (x_j' - \bar{x'})^2} \sqrt{\sum_{k=1}^n (y_k' - \bar{y'})^2}},$$

joka supistuu muotoon

$$r_S = 1 - \frac{6 \sum_{i=1}^n (x_i' - y_i')^2}{n(n^2 - 1)},$$



Kuva 1.7: Pearsonin tulomomenttikorrelaatiokertoimia.

jos aineistossa ei ole sidoksia. Spearmanin korrelaatiokertoimia on havainnollistettu Kuvassa 1.8

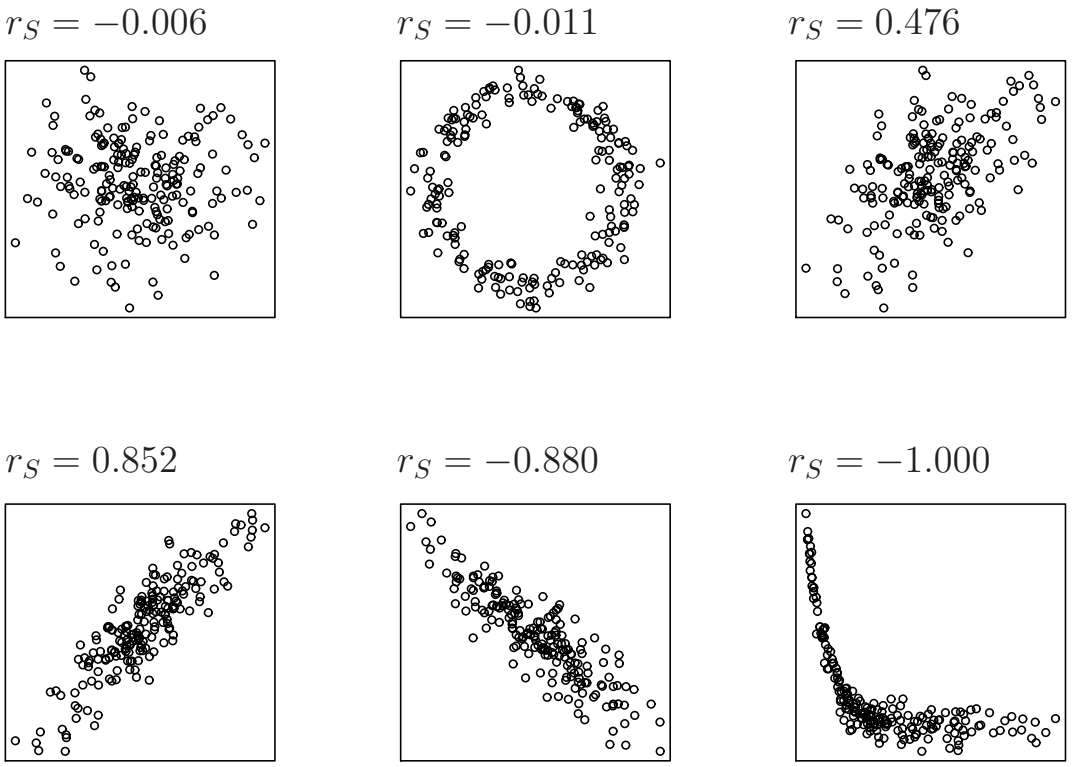
Pearsonin otoskorrelaatiokertoimen r_{XY} voidaan ajatella olevan korrelaatiokertoimen

$$\rho = \mathsf{Corr}(X, Y) = \frac{\mathsf{E}\left((X - \mathsf{E}\left(X\right))(Y - \mathsf{E}\left(Y\right))\right)}{\sqrt{\mathsf{Var}\left(X\right)}\sqrt{\mathsf{Var}\left(Y\right)}},$$

estimaatti. Korrelaatiokertoimen testauksessa lähtökohta on muuttujien X ja Y yhteisjakauman normaalisuus. Kaksiulotteista normaalijakaumaa on havainnollistettu Kuvassa 1.9. Kuvassa 1.10 on piirretty tasa-arvokäyriä ($f(x, y) = c$) kaksiulotteisten normaalijakaumille tiheysfunktioille $f(x, y)$ eri korrelaatiokertoimilla, sekä vastaavista jakaumista simuloitujen otosten sirontakuvia.

Normaalijakaumaoletuksen ollessa voimassa, tai havaintoaineiston ollessa suuri, voimme käyttää korrelaation merkitsevyyden, eli hypoteesien

$$\mathsf{H}_0 : \rho = 0 \quad \mathsf{H}_v : \ \rho \neq 0$$



Kuva 1.8: Pearsonin tulomomenttikorrelaatiokertoimia.

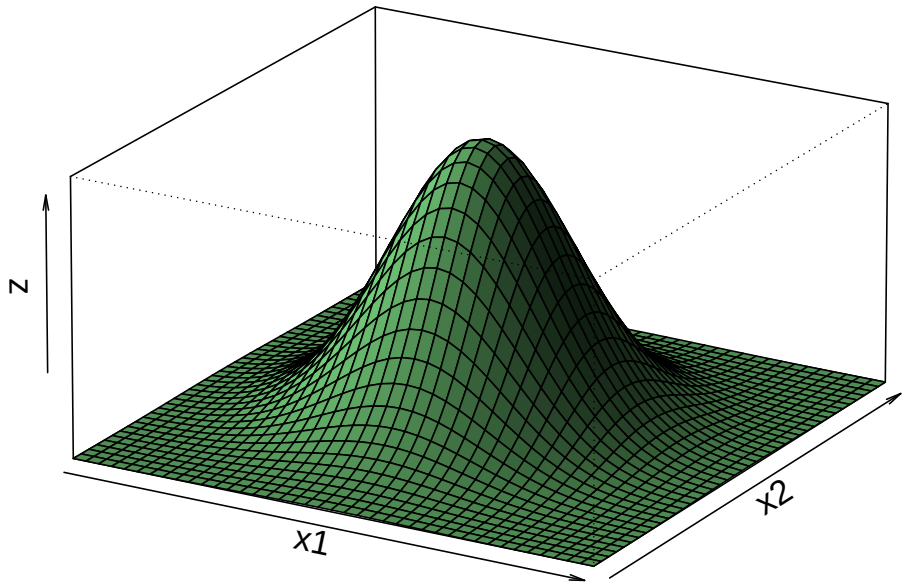
testaukseen suuretta

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}. \tag{1.2}$$

Nollahypoteesin ollessa voimassa testisuure noudattaa jakaumaa t$(n-2)$. Korrelaatiokertoimena testisuureen laskemisessa voidaan käyttää joko Pearson tai Spearmanin kerrointa. Huomaa, että testi kertoo vain poikkeaako kerroin tilastollisesti merkitsevästi nollasta. Testi ota kantaa kertoimen suuruuteen! Toisin sanoen, korrelaatio voi olla vahvaa, mutta tilastollisesti eimerkitsevää, tai korrelaatio voi olla heikkoa, mutta tilastollisesti merkitsevää. Joskus testaukseen käytetään Fisherin muunnosta muuttujasta (1.2), mutta siihen emme paneudu tämän kurssin puitteissa.

Esimerkki 9. Seuraavassa asetelmassa on eräästä otoksesta 18 arvoparia (x, y):

x :	2.5	1.7	2.1	0.2	2.4	2.8	1.2	3.7	0.9	1.4	0.4	1.6	3.5	0.7	2.7	1.3	3.2	3.3
y :	2.2	1.5	2.1	0.4	1.3	2.4	1.4	3.8	1.0	1.7	1.1	0.9	3.3	0.6	1.6	0.7	2.8	2.5



Kuva 1.9: Kaksiulotteinen normaalijakauma

Jota on havainnollistettu Kuvassa 1.11

Kun pienestä aineistosta lasketaan käsin kovarianssin ja korrelaatiokertoimien arvot, kannattaa ensin laskea arvot

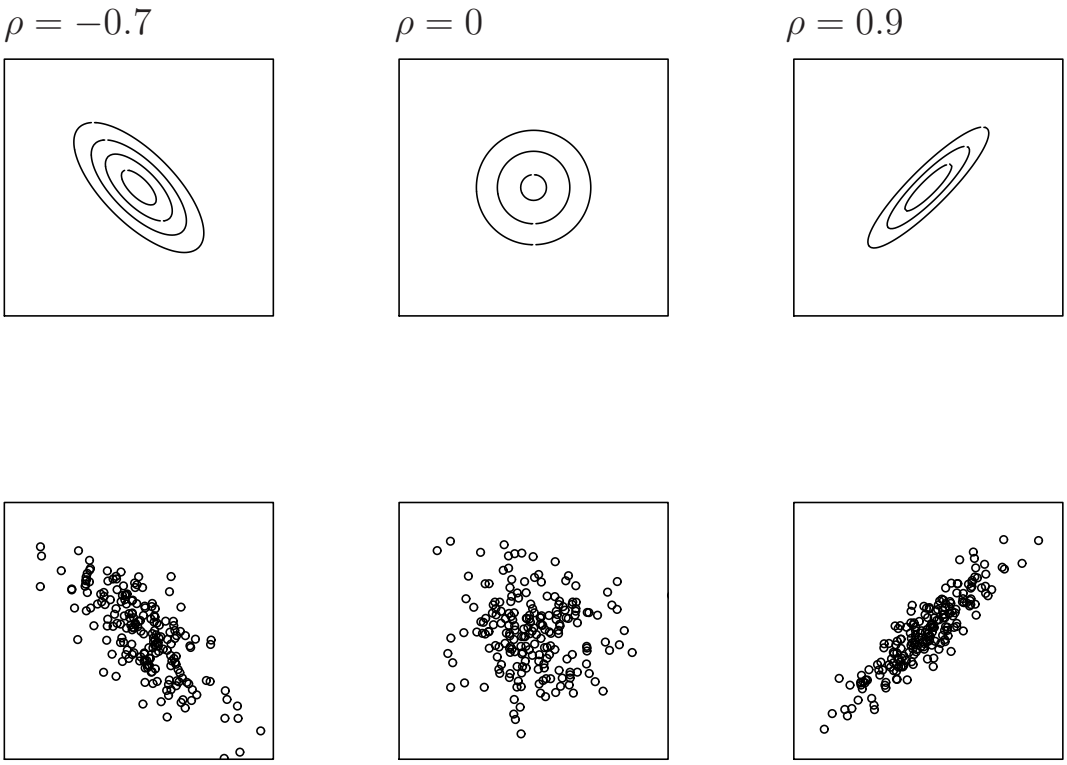
$$\sum_{i=1}^{18} x_i = 35.6, \quad \sum_{i=1}^{18} y_i = 31.3$$
$$\sum_{i=1}^{18} x_i^2 = 90.66, \quad \sum_{i=1}^{18} y_i^2 = 69.81, \quad \sum_{i=1}^{18} x_i y_i = 77.69,$$

joiden avulla voidaan laskea otoskovarianssi

$$s_{XY} = \frac{1}{18-1}\left[\sum_{i=1}^{18} x_i y_i - \frac{1}{18}\sum_{j=1}^{18} x_j \sum_{k=1}^{18} y_k\right] = 0.929$$

ja

$$s_X^2 = 1.191, \quad s_Y^2 = 0.905$$
$$r_P = 0.894.$$



Kuva 1.10: Kaksiulotteisten normaalijakaumien tasa-arvokäyriä sekä niistä simuloituja satunnaismuuttujia.

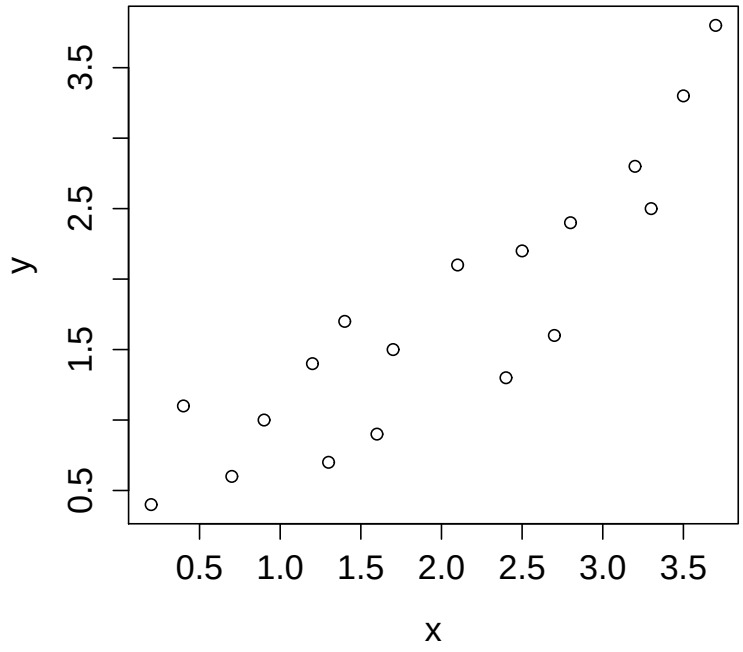
Lineaarisen riippuvuuden selitysprosentti on $100r_P^2 = 80$.

Lasketaan Spearmanin korrelaatiokerroin. Siirrytään $x-$ ja $y-$arvoista järjestyslukuihin x'_i ja y'_i

x :	2.5	1.7	2.1	0.2	2.4	2.8	1.2	3.7	0.9	1.4	0.4	1.6	3.5	0.7	2.7	1.3	3.2	3.3
y :	2.2	1.5	2.1	0.4	1.3	2.4	1.4	3.8	1.0	1.7	1.1	0.9	3.3	0.6	1.6	0.7	2.8	2.5
x' :	12	9	10	1	11	14	5	18	4	7	2	8	17	3	13	6	15	16
y' :	13	9	12	1	7	14	8	18	5	11	6	4	17	2	10	3	16	15
$x' - y'$:	-1	0	-2	0	4	0	-3	0	-1	-4	-4	4	0	1	3	3	-1	1
$(x' - y')^2$:	1	0	4	0	16	0	9	0	1	16	16	16	0	1	9	9	1	1

Nyt järjestettyjen lukujen avulla voidaan laskea

$$\sum_{i=1}^{18} d_i^2 = 100,$$



Kuva 1.11: Esimerkin 9 havaintoaineisto.

ja edelleen

$$r_S = 1 - \frac{6 \cdot 100}{18(18^2 - 1)} = 0.897.$$

Testaamme vielä Pearsonin korrelaatiokertoimen poikkeavuutta nollasta korrelaatiotestin avulla. Muodostamme hypoteesit

$$\mathsf{H}_0 : \rho = 0, \quad \mathsf{H}_v : \rho \neq 0,$$

joita testaamme tasolla 0.01. Havaittu testisuureen arvo on

$$t_{\text{hav}} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.894}{\sqrt{\frac{1-0.894^2}{18-2}}} = 7.98095.$$

Testisuureemme noudattaa likimääräisesti $\mathsf{t}(16)$-jakaumaa, jolle $0.01/2$-yläkvantiili on $t^{(16)}_{0.005} = 2.921$. Koska testisuureen havaittuarvo osuu kriittiselle alueelle, niin voimme hylätä nollahypoteesin tasolla 0.01. Voimme myös katsoa t-jakauman yläkvantiilitaulukosta rajat p-arvolle. Koska

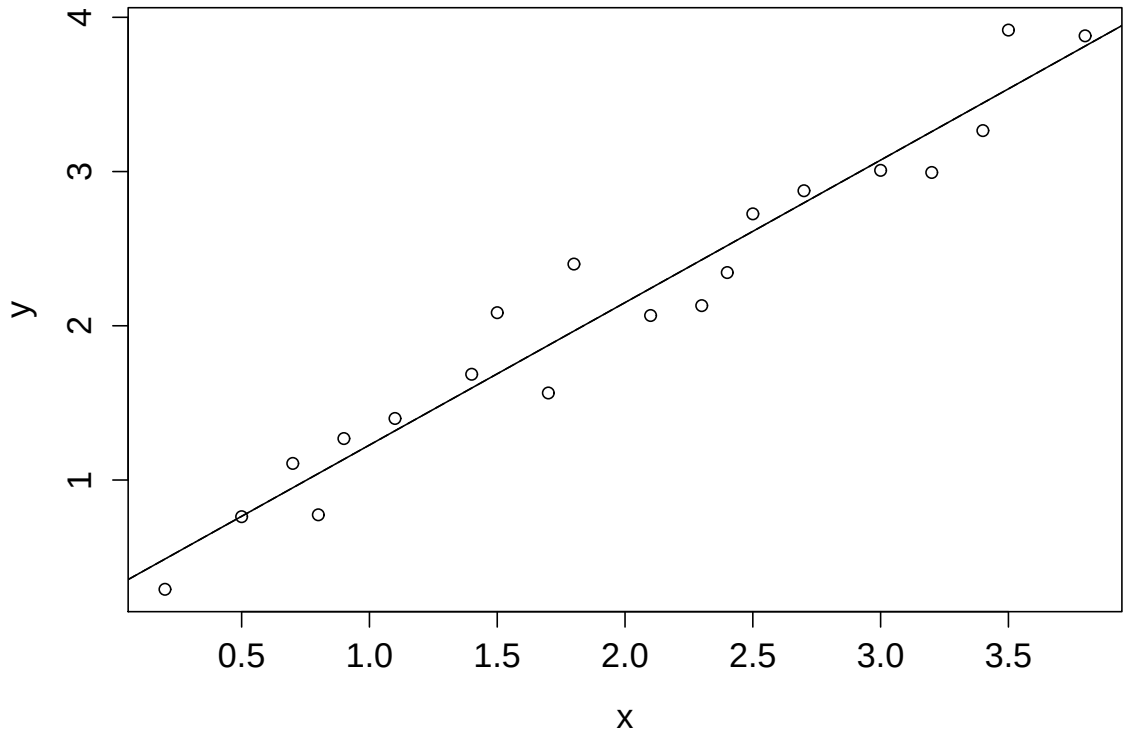
$$7.98095 = t_{\text{hav}} > t^{(16)}_{0.0005},$$

niin merkitsevyystasolle pätee

$$\mathsf{p-arvo} < 2 \cdot 0.0005 = 0.001.$$

1.5 Yksinkertainen lineaarinen regressioanalyysi

Oletetaan että tarkastellaan kahta kvantitatiivista muuttujaa X ja Y, joista ollaan kerätty havaintoaineisto. Haluamme tarkastella muuttuvatko vasteen Y:n arvot säännönmukaisesti, kun X:n arvot muuttuvat, eli haluamme selittää Y:n X:n avulla. Selittävä muuttuja X voi olla vaste tai tekijä. Tilanne eroaa nyt korrelaation tutkimisesta siten, että kun korrelaatiossa oltiin kiinnostuttu vain siitä osuvatko havainnot kuinka hyvin samalle suoralle, niin nyt olemme kiinnostuneita esimerkiksi kyseisen suoran ominaisuuksista. Havainnollistetaan tilannetta Kuvalla 1.12, jossa havaintoaineistoon ollaan sovitettu suora, jonka kulmakerroin kertoo kuinka paljon Y muuttuu kun X muuttuu ja jonka vakiotermi kertoo minkä arvon Y saisi kun $X = 0$.



Kuva 1.12: Havaintoaineiston sirontakuvio ja sovitettu yksinkertainen regressiosuora.

Y:n muutosta X:n suhteen tarkastellaan usein parametrisesti, niin sanotun **regressiofunktion** avulla. Ensin valitsemme jonkin parametrisen funktion, jonka muoto mahdollisesti kuvaa havaintoaineistoa. Tämän jälkeen

havaintoaineiston avulla regressiofunktion parametrien arvoja estimoidaan, siten että estimoiduilla parametrien arvoilla käyrä sopii mahdollisimman hyvin havaintoaineistoon. Tätä kutsutaan käyrän sovittamiseksi aineistoon. Yleisiä regressiofunktioita ovat yksinkertainen lineaarinen regressiofunktio

$$g(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x$$

ja polynomiregressiofunktio

$$g(x; \beta_0, \beta_1, \ldots, \beta_m) = \beta_0 + \beta_1 x + \cdots + \beta_m x^m,$$

jotka ovat molemmat lineaarisia funktioita tuntemattomien parametrien β_i suhteen. Lisäksi usein joudutaan käsittelemään hankalampia epälineaarisia regressiofunktioita, kuten

$$g(x; \beta_0, \beta_1, \beta_2) = \frac{1}{\beta_0 + \beta_1 \exp(-\beta_2 x)}$$
$$g(x; \beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 \exp(\beta_1 x) + \beta_2 \exp(\beta_3 x).$$

Regressiokäyrät sovitetaan laskemalla parametreille $\beta_0, \ldots, \beta_m$ estimaatit $\hat{\beta}_0, \ldots, \hat{\beta}_m$. Estimaattien avulla voidaan laskea **sovite** $\hat{y}$ millä tahansa sopivalla x:n arvolla

$$\hat{y} = g(x; \hat{\beta}_0, \ldots, \hat{\beta}_m).$$

Erityisesti voidaan laskea mallin $g(\cdot)$ mukainen sovite havainnolle (x_i, y_i),

$$\hat{y}_i = g(x_i; \hat{\beta}_0, \ldots, \hat{\beta}_m),$$

jota voidaan verrata havaittuun arvoon y_i laskemalla jäännös

$$e_i = y_i - \hat{y}_i$$

havaitun arvon sekä sovitetun arvon $\hat{y}_i$ välillä. Tarkastellaan tarkemmin yksinkertaista lineaarista regressiofunktiota. Mallin yhtälö havainnolle on

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

jossa siis $g(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x$, ja ϵ_i on ei-havaittava virhetermi. Oletamme, että $\mathsf{E}(\epsilon_i) = 0$ ja $\mathsf{Var}(\epsilon_i) = \sigma^2$. Lisäksi oletamme että virhetermit ovat korreloitumattomia $\mathsf{Cov}(\epsilon_i, \epsilon_j) = 0$, kun $i \neq j$. Miten haetaan estimaatit parametreille havaintojen perusteella? Erilaisia menetelmiä on hyvin monia, mutta yksinkertainen tapa on hakea $\hat{\beta}_i$:t siten, että jäännösten e_i neliösumma

$$q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

minimoituu.

Välttämätön ehto funktion $q(\cdot,\cdot)$ ääriarvokohdille on, että sen derivaatat muuttujien $\hat{\beta}_0$:n ja $\hat{\beta}_1$:n suhteen ovat nollia. Eli haetaan derivaattojen nollakohdat

$$\begin{aligned} 0 &= \frac{\mathrm{d}q(\hat{\beta}_0,\hat{\beta}_1)}{\mathrm{d}\hat{\beta}_0} = 2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]\cdot(-1) \\ \Leftrightarrow \quad 0 &= \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \end{aligned}$$

ja

$$\begin{aligned} 0 &= \frac{\mathrm{d}q(\hat{\beta}_0,\hat{\beta}_1)}{\mathrm{d}\hat{\beta}_1} = 2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]\cdot(-x_i) \\ \Leftrightarrow \quad 0 &= \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2. \end{aligned}$$

Näistä yhtälöistä voimme ratkaista niin sanotut **pienimmän neliösumman estimaatit**

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{j=1}^{n} x_j \sum_{k=1}^{n} y_k}{\sum_{l=1}^{n} x_l^2 - \frac{1}{n}\left(\sum_{m=1}^{n} x_m\right)^2} = \frac{s_{XY}}{s_X^2} \\ \hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\hat{\beta}_1 \sum_{j=1}^{n} x_j = \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

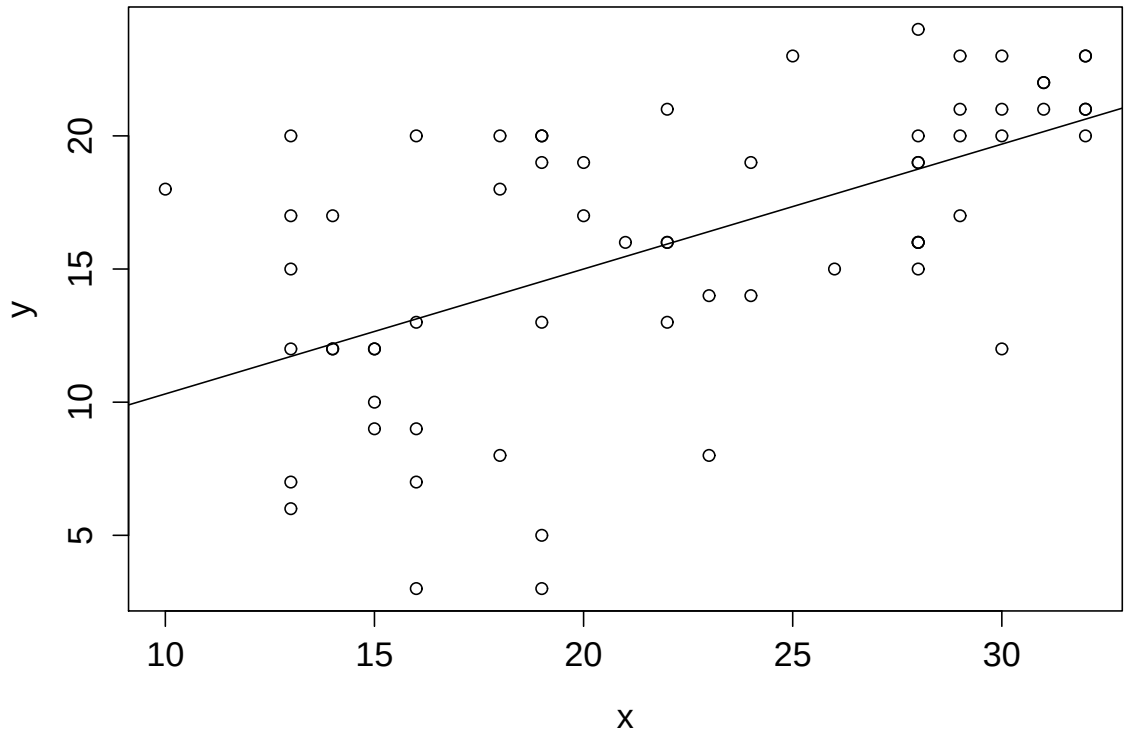
Pienimmän neliösumman menetelmä on mielekäs käsittelemässämme tapauksessa, jossa ei-havaittavat virhetermit ϵ_i oletettiin nollakeskisiksi, samavarianssisiksi ja keskenään korreloitumattomiksi. Tässä tapauksessa estimaattorit $\hat{\beta}_0$ ja $\hat{\beta}_1$ ovat harhattomia. Regressiosuoran sovittamista on havainnollistettu Kuvassa 1.13.

Havaintojen varianssin σ^2 harhaton estimaattori saadaan jakamalla jäännösneliösumma

$$\mathsf{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{j=1}^{n}(x_j - \bar{x})^2$$

jäännösten vapausasteilla $v_E = n - 2$

$$s_E^2 = \frac{\mathsf{SSE}}{n-2}.$$



Kuva 1.13: Lukuvuoden 2012-13 Tilastotieteen peruskurssi b:n harjoituspisteet sekä tenttitulokset 18.12.2012 järjestetystä tentistä. Kuvaan lisäksi piirretty pienimmän neliösumman estimaatti-suora havaintoaineistolle. Sovitetulle suoralle $\hat{\beta}_0 = 5.62$ ja $\hat{\beta}_1 = 0.48$.

Estimaattoreiden $\hat{\beta}_0$ ja $\hat{\beta}_1$ varianssit ovat

$$\mathsf{Var}\left(\hat{\beta}_1\right) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\mathsf{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right),$$

joten estimaattoreiden keskivirheet ovat

$$\mathsf{SE}(\hat{\beta}_1) = s_E \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$
$$\mathsf{SE}(\hat{\beta}_0) = s_E \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Oletetaan mallillemme riippumattomat normaalijakautuneet virheet ϵ_i. Nyt parametrien $(1-\alpha)100\%$-luottamusvälit voidaan laskea t-jakauman

avulla kuten aiemmin

$$\hat{\beta}_0 \pm t_{\alpha/2}^{(v_E)} \mathsf{SE}(\hat{\beta}_0)$$
$$\hat{\beta}_1 \pm t_{\alpha/2}^{(v_E)} \mathsf{SE}(\hat{\beta}_1).$$

Voimme myös hakea luottamusvälin vasteen odotusarvolle mielivaltaisessa selittävän muuttujan pisteessä x_u. Odotusarvon $\mathsf{E}(Y \mid X = x_u) = \beta_0 + \beta_1 x_u$ estimaatti on

$$\hat{y}_u = \hat{\beta}_0 + \hat{\beta}_1 x_u,$$

jonka keskivirhe on

$$\mathsf{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_u) = s_E \sqrt{\frac{1}{n} + \frac{(x_u - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}. \tag{1.3}$$

Normaalijakaumaoletuksen ollessa voimassa, $(1-\alpha)100\%$-luottamusväli $\mathsf{E}(Y \mid X = x_u)$:lle saadaan keskivirheen avulla

$$\hat{\beta}_0 + \hat{\beta}_1 x_u \pm t_{\alpha/2}^{(v_E)} \mathsf{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_u). \tag{1.4}$$

Edellä kuvatut (1.3) sekä (1.4) liittyvät nimenomaan parametrien estimoinnin tarkkuuteen. Usein kuitenkin olemme kiinnostuneita miten vaste käyttäytyy kun keräämme lisää dataa. Eli mikä on *tulevan* havainnon jakauma? Nyt voimme ajatella että $y = \beta_0 + \beta_1 x$ on lineaarinen **prediktori** ja estimoitu regressiosuora $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ on lineaarisen prediktorin estimaatti. Kaksi eri merkitystä näkyvät laskettaessa keskivirhettä. Muuttujan Y selittävän muuttujan arvoon x liittyvä tulevan arvon ennuste on $\hat{\beta}_0 + \hat{\beta}_1 x$ ja arvon $(1-\alpha)100\%$-luottamusväli on

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{\alpha/2}^{(v_E)} s_E \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Ennustetun arvon luottamusväli on pidempi kuin estimointiin liittyvä luottamusväli, sillä ennustetun muuttujan luottamusvälissä otetaan huomioon tulevan arvon ei-havaittava virhe ϵ_i.

Esimerkki 10. Jatketaan Esimerkin 9 käsittelyä sovittamalla aineistoon regressiosuora $y = \beta_0 + \beta_1 x$. Olemme laskeneet arvot

$$\sum_{i=1}^{18} x_i = 35.6, \quad \sum_{i=1}^{18} y_i = 31.3$$
$$\sum_{i=1}^{18} x_i^2 = 90.66, \quad \sum_{i=1}^{18} y_i^2 = 69.81, \quad \sum_{i=1}^{18} x_i y_i = 77.69,$$

joiden avulla voidaan laskea otoskeskiarvot, otosvarianssit sekä otoskovarianssit

$$\begin{aligned}\bar{x} &= \frac{1}{18}35.6 = 1.978, \quad \bar{y} = \frac{1}{18}31.3 = 1.739\\ s_X^2 &= 1.191, \quad s_Y^2 = 0.905\\ s_{XY} &= \frac{1}{18-1}\left[\sum_{i=1}^{18} x_i y_i - \frac{1}{18}\sum_{j=1}^{18} x_j \sum_{k=1}^{18} y_k\right] = 0.929.\end{aligned}$$

Näiden suureiden avulla voimme laskea regressiosuoran parametrien estimaatit

$$\begin{aligned}\hat{\beta}_1 &= \frac{s_{XY}}{s_X^2} = 0.780\\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 0.196,\end{aligned}$$

ja edelleen odotusarvon $\mathsf{E}\,(Y \mid X = x)$ estimaatin

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.196 + 0.780x.$$

Varianssin σ^2 estimaatti on jäännöskeskineliö

$$s_E^2 = \frac{\mathsf{SSE}}{18-2} = \frac{\sum_{i=1}^{18}(y_i - 1.739)^2 - 0.780^2 \sum_{j=1}^{18}(x_j - 1.978)^2}{16} = 0.191.$$

Parametrien keskivirheet ovat

$$\begin{aligned}\mathsf{SE}(\hat{\beta}_1) &= \sqrt{0.191}\sqrt{\frac{1}{\sum_{i=1}^{18}(x_i - 1.978)^2}} = 0.097\\ \mathsf{SE}(\hat{\beta}_0) &= \sqrt{0.191}\sqrt{\frac{1}{18} + \frac{1.978^2}{\sum_{i=1}^{18}(x_i - 1.978)^2}} = 0.218,\end{aligned}$$

ja sovitetun vasteen keskivirhe on

$$\mathsf{SE}(\hat{\beta}_0 + \hat{\beta}_1 x) = \sqrt{0.191}\sqrt{\frac{1}{18} + \frac{(x - 1.978)^2}{\sum_{i=1}^{18}(x_i - 1.978)^2}},$$

Tulevan ennustetun Y:n arvon keskivirhe liittyen arvoon $X = x_p$ on

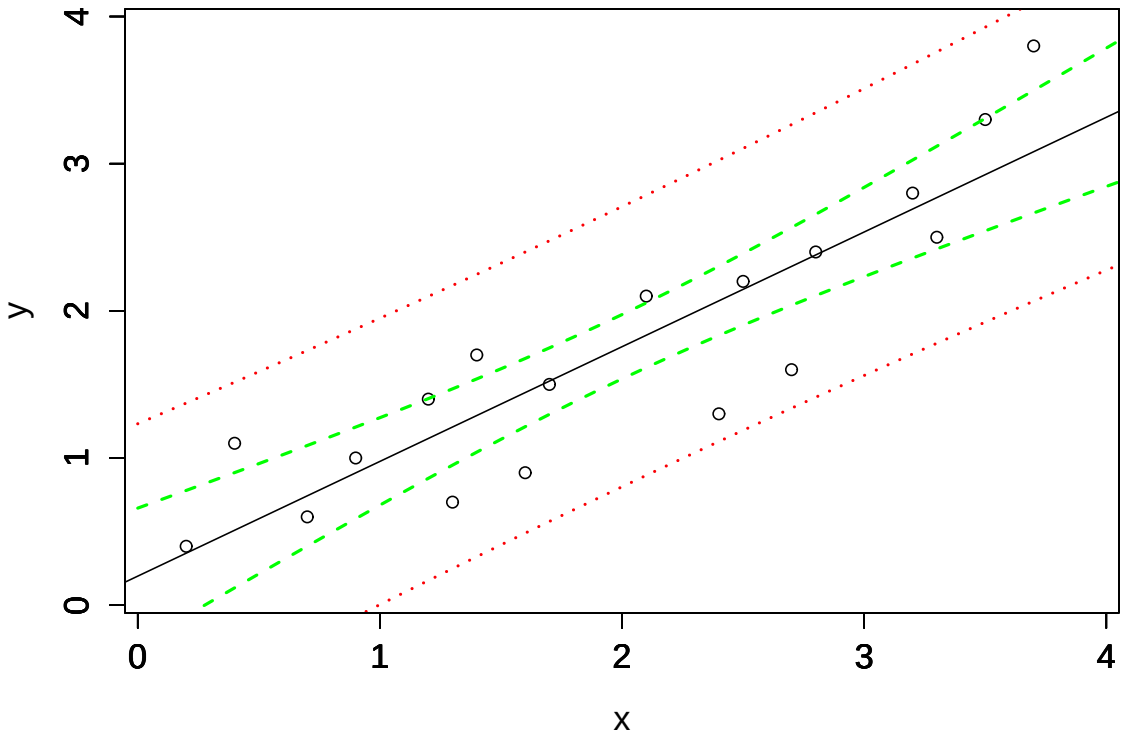
$$\mathsf{SE}(\hat{\beta}_0 + \hat{\beta}_1 x_p) = \sqrt{0.191}\sqrt{1 + \frac{1}{18} + \frac{(x_p - 1.978)^2}{\sum_{i=1}^{18}(x_i - 1.978)^2}}.$$

Koska $t_{0.025}^{(16)} = 2.120$ niin 95%-luottamusvälit parametreille, sovitetulle vas-

teelle sekä ennustetulle vasteelle ovat

$$\begin{aligned}
\beta_0: &\quad 0.196 \pm 2.120 \cdot 0.218 \\
\beta_1: &\quad 0.780 \pm 2.120 \cdot 0.097 \\
\beta_0 + \beta_1 x: &\quad 0.196 + 0.780x \pm 2.120 \cdot \sqrt{0.191}\sqrt{\frac{1}{18} + \frac{(x - 1.978)^2}{\sum_{i=1}^{18}(x_i - 1.978)^2}} \\
\beta_0 + \beta_1 x_p: &\quad 0.196 + 0.780x_p \pm 2.120 \cdot \sqrt{0.191}\sqrt{1 + \frac{1}{18} + \frac{(x_p - 1.978)^2}{\sum_{i=1}^{18}(x_i - 1.978)^2}},
\end{aligned}$$

joista kaksi viimeistä on piirretty sovitetun regressiosuoran ohella Kuvaan 1.14.



Kuva 1.14: Esimerkin 10 aineisto, sovitettu regressiosuora sekä 95%-luottamusvälit sovitetulle sekä ennustetulle vasteelle.

Regressio-mallin sopivuutta voidaan tarkastella tilastollisen merkitsevyystestauksen avulla. Regressiosuoran tapauksessa voimme testata kumpaakin parametria β_0 ja β_1, mutta usein mielenkiinto kohdistuu regressiosuo-

ran kulmakertoimeen. Normaalijakaumamallissa hypoteesin

$$\mathsf{H}_0 : \beta_1 = b_1$$

testisuureena voidaan käyttää t-testisuuretta

$$t = \frac{\hat{\beta}_1 - b_1}{\mathsf{SE}(\hat{\beta}_1)} \sim \mathsf{t}(v_E).$$

Vakiotermille testi on samaa muotoa, eli

$$\mathsf{H}_0 : \beta_0 = b_0,$$

jota testataan t-testisuureella

$$t = \frac{\hat{\beta}_0 - b_0}{\mathsf{SE}(\hat{\beta}_0)} \sim \mathsf{t}(v_E).$$

Yleisimmin testataan hypoteeseja

$$\begin{aligned} \mathsf{H}_0 &: \beta_1 = 0 \\ \mathsf{H}_v &: \beta_1 \neq 0, \end{aligned}$$

jolloin testiä kutsutaan selittäjän X merkitsevyyden testiksi. Yksinkertaisen regressiosuoran tapauksessa, eli tapauksessa jossa on vain yksi selittäjä, testiä kutsutaan **mallin merkitsevyyden testiksi**. Mallin merkitsevyys voidaan ajatella siten, että paljonko kulmakerroin sekä selittävän muuttujan arvo selittää vasteen arvosta, vai onko vasteen arvo sama riippumatta selittävän muuttujan arvosta, kuten havainnollistettu Kuvassa 1.15

Varianssianalyysi ja regressioanalyysi ovat läheistä sukua toisilleen. Oikeastaan kyseessä on samankaltainen lineaarinen malli sovellettuna eri tyyppisille muuttujille. Tästä kertoo sekin, että regressioanalyysinkin taustalla pyörivät erilaiset neliösummat ja regressioanalyysissäkin tarkastellaan ANOVA-taulukoita.

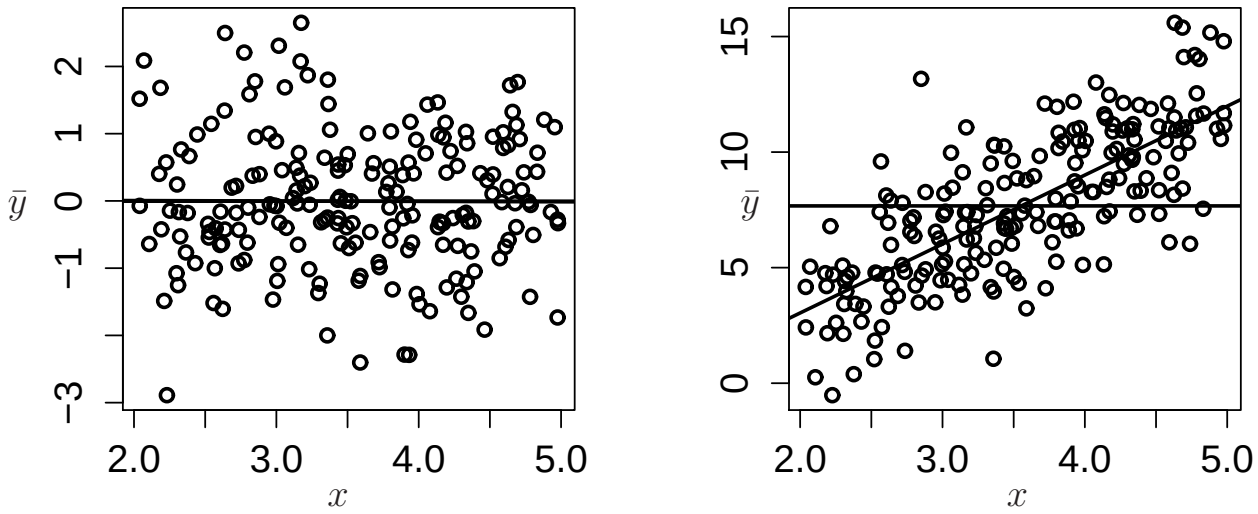
Regressiosuoran jäännösneliösumma

$$\mathsf{SSE} = \mathsf{SST} - \mathsf{SSA}$$

voidaan jakaa mallin kokonaisvaihteluun SST sekä mallin selittämään vaihteluun SSA

$$\mathsf{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\mathsf{SSA} = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2.$$



Kuva 1.15: Selittääkö kulmakerroin vastetta y, vai onko vaste riippumaton selittävän muuttujan x arvoista.

Kaksipuolinen mallinmerkitsevyyden testaus voidaan suorittaa F-testin avulla samoin kuin varianssianalyysissa. Testisuure on tällöin

$$F = \frac{\hat{\beta}_1^2}{\mathsf{SE}(\hat{\beta}_1)^2} = \frac{\mathsf{SSA}/v_A}{s_E^2} = \frac{s_A^2}{s_E^2} \sim \mathsf{F}(v_A, v_E) \quad (\mathsf{H}_0\text{:n ollessa voimassa}) ,$$

jossa SSA, s_A^2 ja $v_A = 1$ ovat hypoteesiin liittyvä neliösumma, keskineliö ja vapausaste. Korrelaatiokertoimen neliö, eli selitysaste, on eräs mallin hyvyyden kriteereistä

$$r^2 = 1 - \frac{\mathsf{SSE}}{\mathsf{SST}}.$$

Esimerkki 11. Jatketaan edelleen Esimerkin 9 käsittelyä. Testataan hypoteesia

$$\begin{aligned} \mathsf{H}_0 &: \beta_1 = 0 \\ \mathsf{H}_v &: \beta_1 \neq 0 \end{aligned}$$

Vaihtelun lähde	SS	df	MS	F
Tekijä	SSA	v_A	s_A^2	s_A^2/s_E^2
Jäännös	SSE	v_E	s_E^2	
Kokonais	SST	$n-1$		

Taulukko 1.2: Regressiomallin merkitsevyyden testaukseen liittyvä ANOVA-taulukko

tasolla 0.05. Muodostetaan ensin neliösummat

$$\begin{aligned}
\mathsf{SSE} &= \sum_{i=1}^{18}(y_i - 1.739)^2 - 0.780^2 \sum_{j=1}^{18}(x_j - 1.978)^2 = 3.062 \\
\mathsf{SSA} &= 0.780^2 \sum_{j=1}^{18}(x_j - 1.978)^2 = 12.321 \\
\mathsf{SST} &= \sum_{i=1}^{18}(y_i - 1.739)^2 = 15.383,
\end{aligned}$$

ja täydennetään ANOVA-taulukko regressiomallille.

Vaihtelun lähde	SS	df	MS	F
Tekijä	12.321	1	12.321	64.5
Jäännös	3.062	16	0.191	
Kokonais	15.383	17		

Nyt $\mathsf{F}(1, 16)$-jakauman taulukosta voidaan lukea kriittinen piste hypoteesillemme,

$$f_{0.05}^{(1,16)} = 4.49 \leq 64.5 = F_{\text{hav}},$$

joten voimme hylätä nollahypoteesin tasolla 0.05.

Jos testataan poikkeavatko vakiotermi β_0 sekä kulmakerroin β_1 nollasta

t-testillä, niin saamme havaitut t-testisuureet

$$t_{\text{hav,0}} = \frac{\hat{\beta}_0}{\mathsf{SE}(\hat{\beta}_0)} = \frac{0.196}{0.218} = 0.899$$
$$t_{\text{hav,1}} = \frac{\hat{\beta}_1}{\mathsf{SE}(\hat{\beta}_1)} = \frac{0.780}{0.097} = 8.041$$

Studentin t-jakauman yläkvantiilitaulukosta luemme, että voimme hylätä hypoteesin $\mathsf{H}_0 : \beta_1 = 0$ kaksipuolisella vastahypoteesilla merkitsevyystasolla $\mathsf{p} - \mathsf{arvo} < 0.001$. Lisäksi näemme, että hypoteesi $\mathsf{H}_0 : \beta_0 = 0$ kaksipuolisella vastahypoteesilla jää voimaan tasolla $0.2 < \mathsf{p} - \mathsf{arvo} < 0.5$.

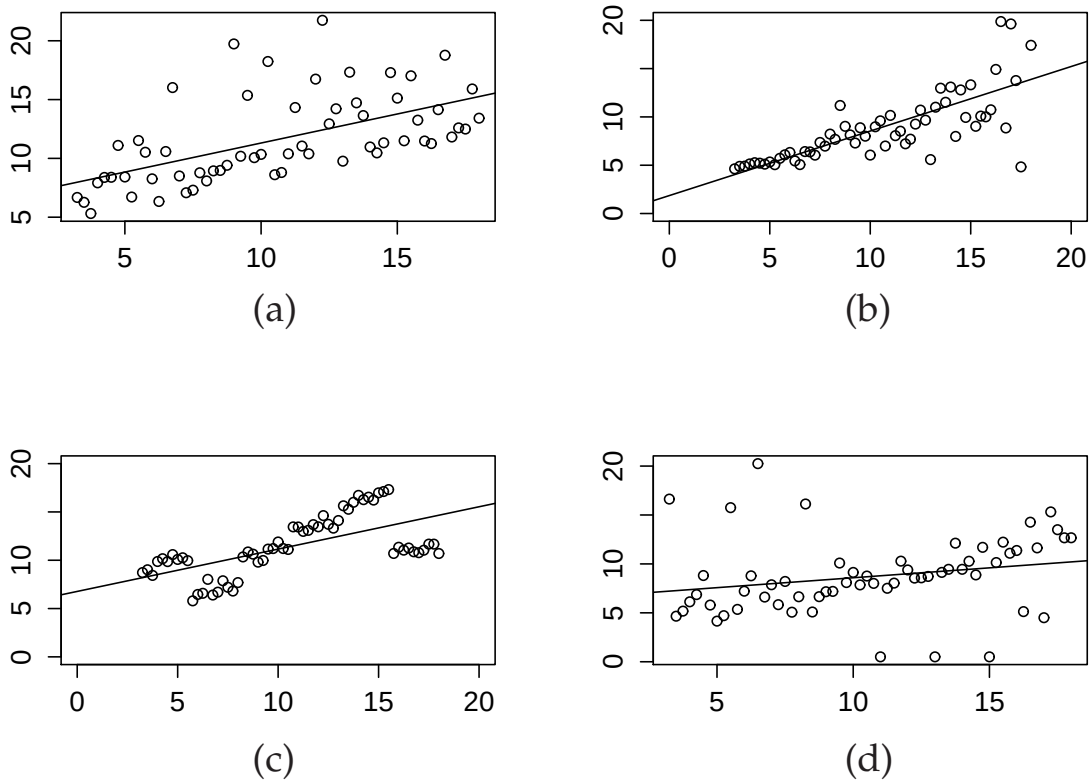
1.5.1 Diagnostinen tarkastelu

Regressioanalyysissä menetellään kuten ennenkin, ensin tarkastellaan mallin oletusten voimassaoloa, jonka jälkeen lähdetään analysoimaan varsinaisia tuloksia. Tärkeää on tarkastella onko merkkejä seuraavista:

- Virheiden jakaumaa ei voida pitää normaalina. (Kuva 1.16 (a))
- Virheiden varianssit ovat eri suuria. (Kuva 1.16 (b))
- Lineaarinen prediktori ei ole hyvä (eli perustilanteessa yhteys ei ole lineaarinen). Siinä on jotain turhia tai huonoja osia, siitä puuttuu jokin oleellinen osa, tai tarvitaan jotain selittäjän, selitettävän tai perusteellisia mallin yhtälön muodon muunnoksia.
- Virheet eivät ole riippumattomia. (Kuva 1.16 (c))
- Osa havainnoista on jollain tavalla poikkeavia siten, että niiden takia ei tavallisesti käytetty analyysitapa olekaan hyvä. (Kuva 1.16 (d))

Tarkastelu tapahtuu pääasiassa jäännösten $e_i = y_i - \hat{y}_i$ avulla. Usein tarkastellaan erilaisia modifikaatioita jäännöksistä. Näitä ovat mm.:

- Standardoitu jäännös, jossa jäännökset jaetaan s_E:llä.
- Studentoitu jäännös, jossa jäännökset jaetaan keskivirheellään, eli keskihajonnan $\sqrt{\mathsf{Var}\,(e_i)} = \sigma^2(1 - h_{i,i})$ estimaatilla $\mathsf{SE}(e_i) = s_E\sqrt{1 - h_{i,i}}$, jossa $h_{i,i}$ on i. havainnon **vipuarvo**, joka kuvaa kuinka suuri vaikutus havainnolla on regressiosuoraan. Käytännössä vipupiste kuvaa kuinka kaukana vastaavan selittäjän arvo on keskimääräisestä. Studentoiduilla jäännöksillä on yhtäsuuret varianssit.



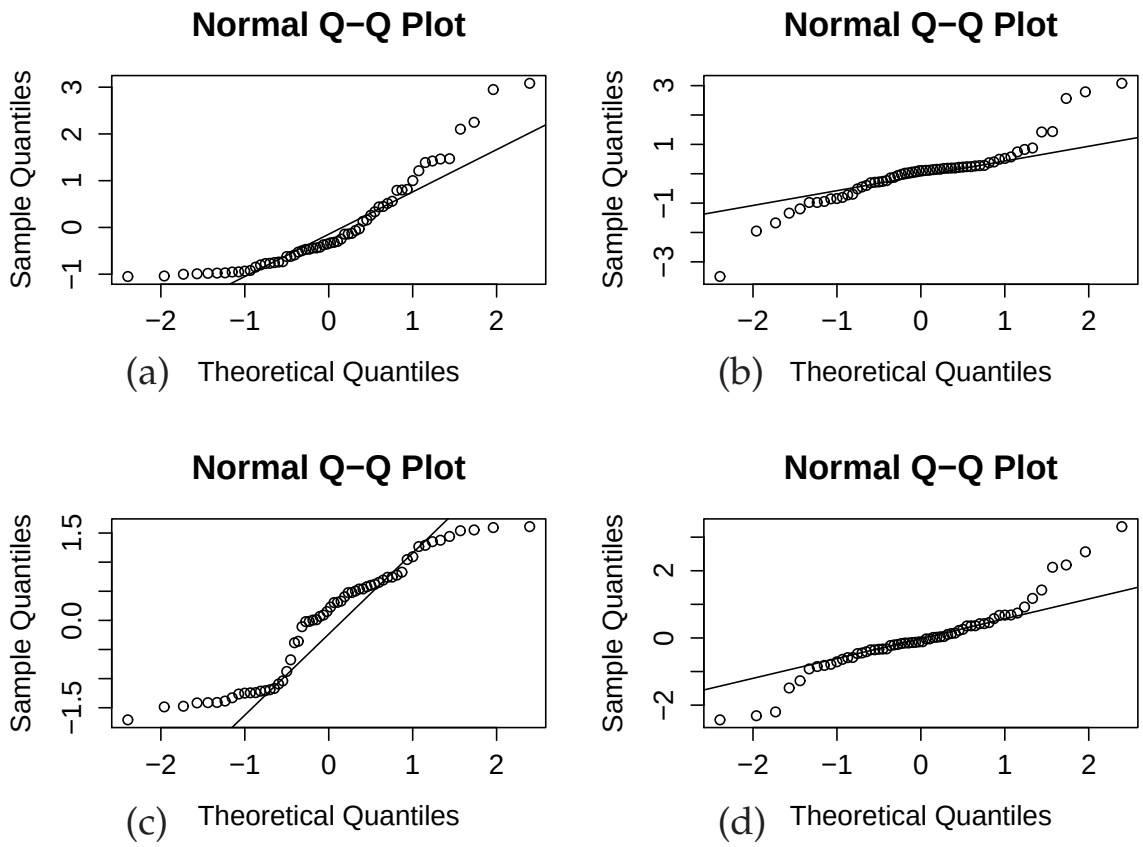
Kuva 1.16: Poikkeamia mallin ei-havaittavan virheen oletuksista.

- Ennustettu jäännös, jossa jäännös $e_i = y_i - \hat{y}_i$ lasketaan sovitetun arvon $\hat{y}_i$ avulla, joka ollaan laskettu jättämällä pois havainto (x_i, y_i). Näin saadaan paremmin esille poikkeavat havainnot.

- Studendoitu ennustettu jäännös, jossa $i.$ jäännös sekä jäännöksen varianssin estimaatti lasketaan ilman $i.$ havaintoa. Jakamalla jäännökset niiden keskihajonnan estimaateilla, muunnetut jäännökset noudattavat t-jakaumaa.

Joskus kirjallisuudessa Studentoiduja jäännöksiä kutsutaan standardoituiksi residuaaleiksi, ja Studentoituja ennustettuja jäännöksiä Studentoiduiksi jäännöksiksi.

Jäännöksiä voidaan tarkastella graafisesti esimerkiksi Studentoitujen residuaalien normaalikvantiilikuvion avulla (Kuva 1.17), jonka antaa yleiskuvan aineiston sopivuudesta, mutta yleensä tämän perusteella ei pystytä erittelemään mikä on pielessä jos aineisto ei näytä normaalilta. Toinen graafinen lähestymistapa on käyttää sirontakuvioita, joissa on esimerkiksi jäännökset selittävien muuttujien (Kuva 1.18) tai sovitearvojen (Kuva 1.19)



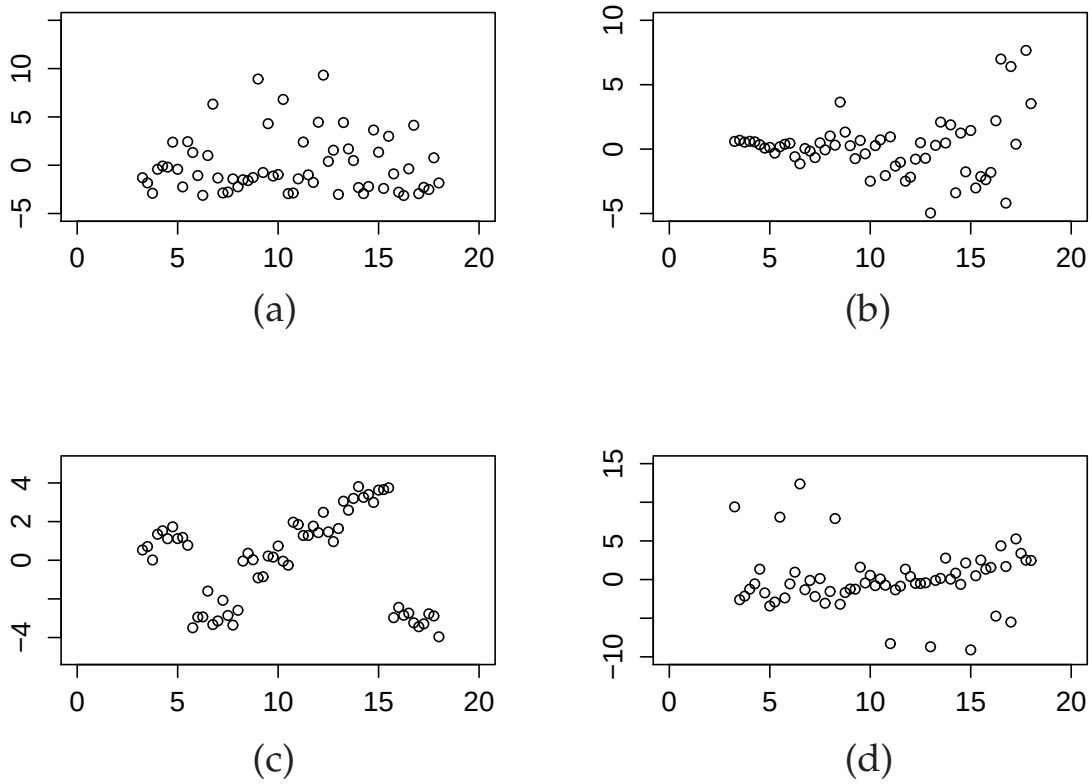
Kuva 1.17: Normaalikvantiilikuvio studentoiduista residuaaleista.

avulla ilmaistuna.

Jos diagnostinen tarkastelu paljastaa, että on hyvin mahdollista etteivät oletetun mallin oletukset täyty, niin on olemassa useita tapoja lähteä parantamaan tilannetta. Poikkeavat havainnot voidaan yrittää poistaa havaintoaineistosta, tai käyttää robusteja menetelmiä pienimmän neliösumman menetelmän sijaan. Selitettävän muuttujan muunnoksilla voidaan yrittää muokata mallia oletusten mukaiseksi. Malliin voidaan lisätä korkeamman asteen termejä (polynomiregressio), tai voimme yrittää lisätä malliin muita muuttujia, joiden mahdollinen yhteisvaikutus selittäisi vastetta. Jos yksinkertaisesti lisäämme selittäjiä malliin, on kyseessä useamman selittäjän lineaarinen regressioanalyysi. Jos selittäjiä on kaksi kvantitatiivista muuttujaa (X_1, X_2), näiden suhteen lineaarinen regressiofunktio on

$$g(x_1, x_2; \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

jonka geometrinen kuvaaja on taso, kuten Kuvassa 1.20. Käytännössä mallin merkitsevyyden testaus tapahtuu F-testillä, kuten regressiosuoran ta-



Kuva 1.18: Regressiosuoran sovituksesta saadut residuaalit vs. selittäjien arvot.

pauksessa. Testi on muotoa

$$\mathsf{H}_0 \ : \beta_1 = \beta_2 = 0$$

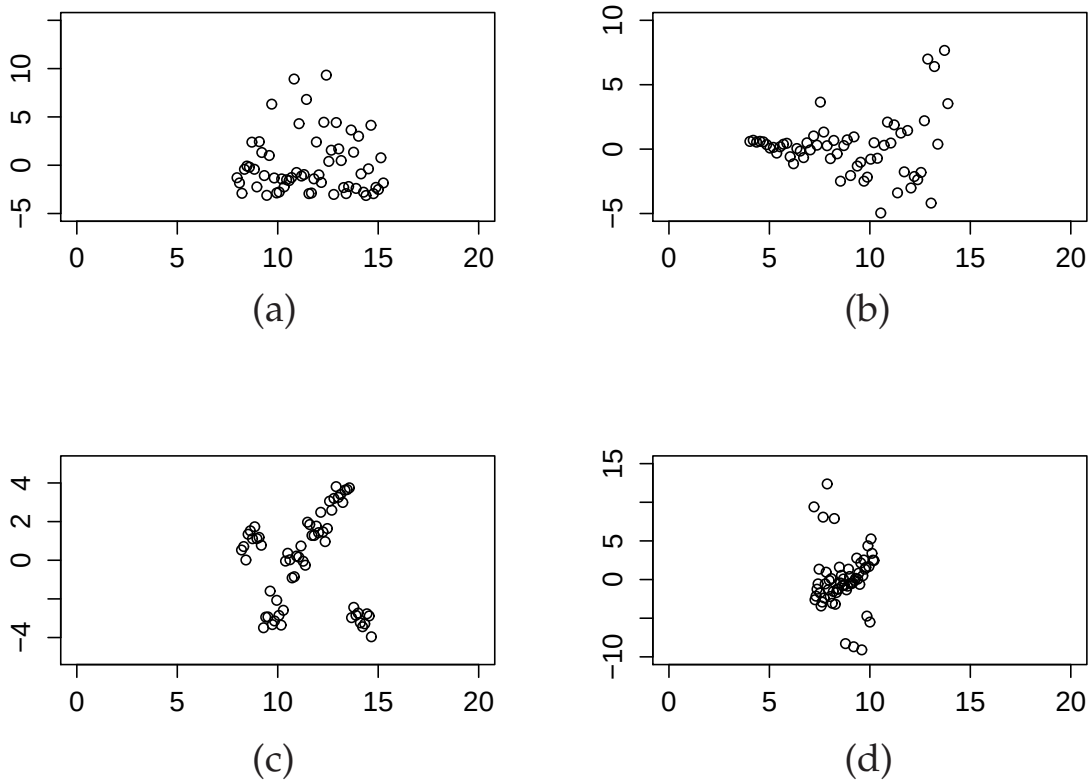
$$\mathsf{H}_v \ : \text{ Vähintään toinen parametreista } \beta_1, \beta_2 \text{ ei ole nolla.}$$

Eli testaamme sopiiko sovitettu taso merkitsevästä paremmin havaintoaineistoon kuin xy-tason suuntainen perustaso, jota on havainnollistettu Kuvassa 1.22.

Kaikki käsittelemämme lineaariset mallit kuuluvat niin sanottujen **yleisten lineaaristen mallien** joukkoon. Yleinen lineaarinen malli on muotoa

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon,$$

jossa selittäjiä on k kappaletta ja selittäjät saavat olla joko kategorisia (kuten varianssianalyysissä) tai numeerisia (kuten yksinkertaisessa lineaarisessa regressiossa). On tärkeää muistaa aineistoa analysoidessamme, että malli jota sovitamme on vain yksi mahdollinen malli tilanteelle. Erilaisten mallien sopivuutta havaintoaineistoon voidaan kuitenkin tarkastella. Mallien vertailuun voidaan periaatteessa käyttää mallien selitysastetta, mutta



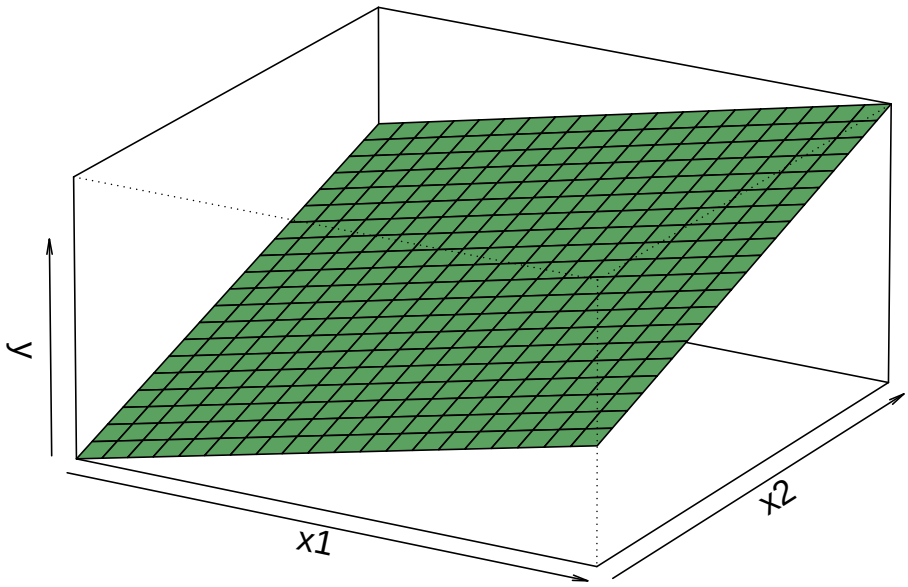
Kuva 1.19: Regressiosuoran sovituksesta saadut residuaalit vs. sovitetut arvot

jos malleissa on eri määrä selittäjiä, ongelmana on että selittävien muuttujien lisääminen malliin ei koskaan huononna selitysastetta. Tästä johtuen tavallinen selitysaste ei sovi sellaisten mallien vertailuun, joissa on eri määrä selittäjiä. Sen sijaan pitää käytää korjattua (adjusted) selitysastetta, jossa lasketaan neliösummien sijaan keskineliöiden suhde

$$R^2_{\text{adj}} = 1 - \frac{\frac{\text{SSE}}{n-k-1}}{\frac{\text{SST}}{n-1}} = 1 - \frac{s_E^2}{s_T^2},$$

jossa otetaan huomioon parametrien määrä. Käytännössä korjatussa selitysasteessa rankaistaan mallia selittäjien määrästä k.

Parametreja koskevia testejä voidaan tehdä kerroinkohtaisesti t-testeillä, tai testata F-testillä useamma selittäjän samanaikaista merkitsevyyttä. Kun malleissa on vähintään kaksi selittäjää, usein halutaan selvittää mikä selittäjistä on 'paras', ja mikä on selittäjien 'paremmuusjärjestys'? Kvantitatiivinen hyvyyden tarkastelu ei ole mielekästä, mutta paremmuusjärjestykseen voidaan todeta yksittäisten kerrointen t-testien merkitsevyystasojen avulla,



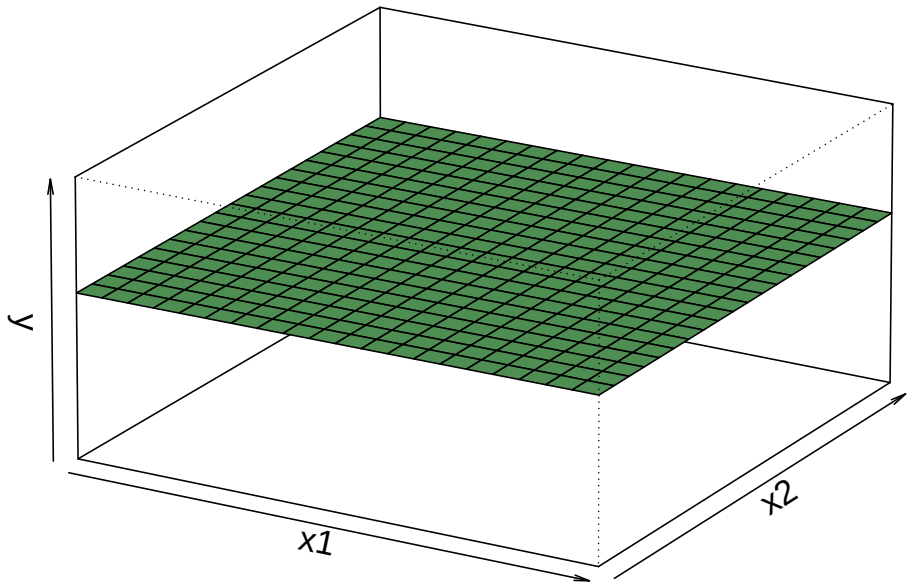
Kuva 1.20: Kahden selittäjän tapauksessa yksinkertainen lineaarinen malli on regressiotaso.

jossa testin suuri merkitsevyystaso tarkoittaa hyvää selittäjää.

1.5.2 Logistinen regressio

Usein tilastollisissa tutkimuksissa joudutaan käsittelemään tapauksia, joissa tarkastellaan kaksiarvoisen vasteen yhteyttä kvantitatiivisiin selittäjiin. Kaksiarvoinen vaste voidaan liittää monenlaisiin tapahtumiin, esim. onnistuminen/epäonnistuminen, läsnä/poissa, elossa/kuollut, raja ylittyy/ei ylity. Tapauksissa joissa muuttuja on kaksiarvoinen, on riittävää tarkastella todennäköisyyttä, että muuttuja saa toisen arvoistaan. Mallinnamme siis todennäköisyyttä rajoittamattoman vasteen sijaan. Emme voi suoraan sovittaa järkevästi regressiosuoraa vasteeseen jo siitäkin syystä, että on tarpeellista ottaa huomioon todennäköisyyden mahdolliset arvot $[0, 1]$.

Esimerkki 12. Tarkastellaan hyönteismyrkyn konsentraation vaikutusta erään kovakuoriaslajikkeen hävittämiseksi. Kun kovakuoriaiseen käytetään tietyn vahvuista myrkkyä, havaitaan joko vaikutus (kovakuoriainen kuolee)



Kuva 1.21: Nollahypoteesin mukainen perustaso.

tai ei havaita vaikutusta (kovakuoriainen ei kuole). Tilanne on Kuvan 1.23 mukainen, mitä pienempi myrkyn konsentraatio on, sitä pienempi todennäköisyys on, että hyönteinen kuolee.

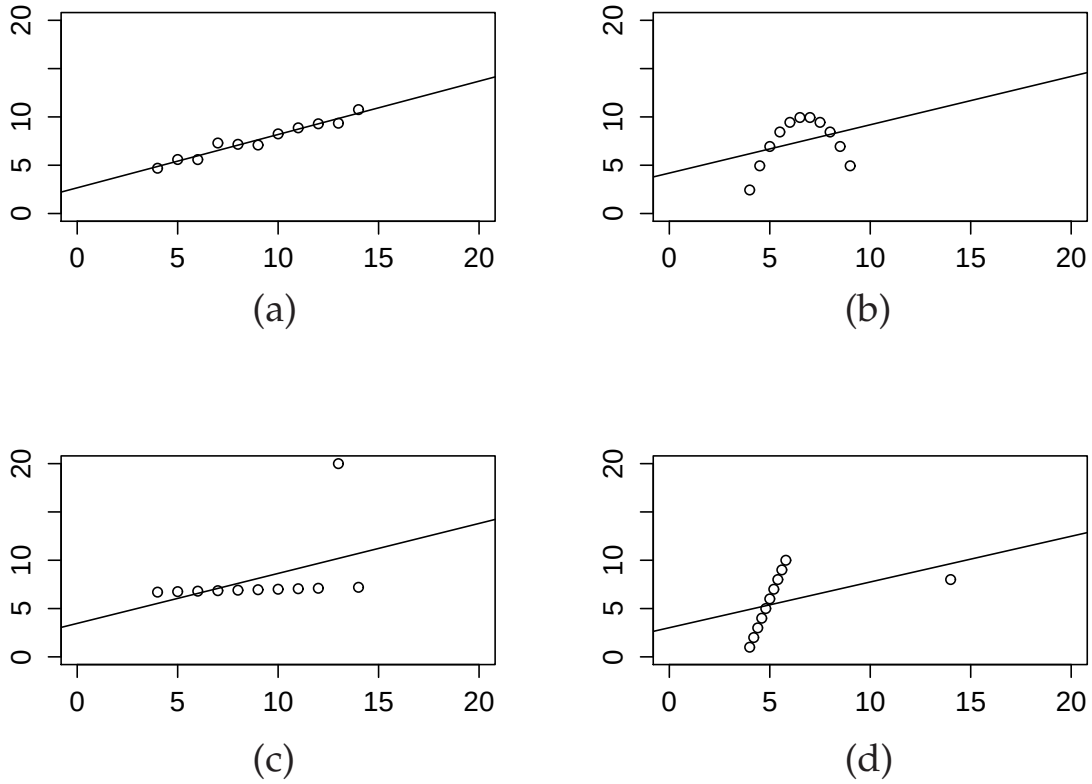
Oletetaan, että D on pienin annos, joka aiheuttaa jonkinlaisen reaktion satunnaisesti valitussa oliossa. Nyt d:n suuruinen annos aiheuttaa reaktion, jos satunnaisesti valitun olion kynnysarvo on korkeintaan d. Merkitään

$$p(d) = \mathsf{P}\,(D \leq d) = F(\beta_0 + \beta_1 d),$$

jossa F on kertymäfunktio (Kuvan 1.23 mukainen), ja $p(d)$ on todennäköisyys, että olio saa reaktion saadessaan annoksen d. Luvut $\beta_0 \in \mathbb{R}$ ja $\beta_1 > 0$ ovat tuntemattomia sijaintiin ja skaalaan liittyviä parametreja. Logistisessa regressiossa valitaan mallintavaksi jakaumaksi standardoitu logistinen jakauma

$$F(z) = \frac{1}{1+\exp(-z)} = \frac{\exp(z)}{1+\exp(z)} = 1 - \frac{1}{1+\exp(z)}.$$

Huomaa, että toisin kuin esimerkiksi normaalijakauman kertymäfunktio, logistisen jakauman kertymäfunktio voidaan laskea tarkasti, ilman integroin-



Kuva 1.22: Regressiosuoran sijoittaminen erilaisiin aineistoihin tapauksissa: (a) sopiva malli, (b) polynomi-malli, (c) poikkeava havainto, (d) vaikuttava havainto.

tia tai taulukoituja arvoja. Nyt voimme ratkaista yhtälöstä $F(z_i) = p_i = p(d_i)$

$$z_i = \mathsf{log}\left(\frac{p_i}{1-p_i}\right) = \mathsf{logit}(p_i),$$

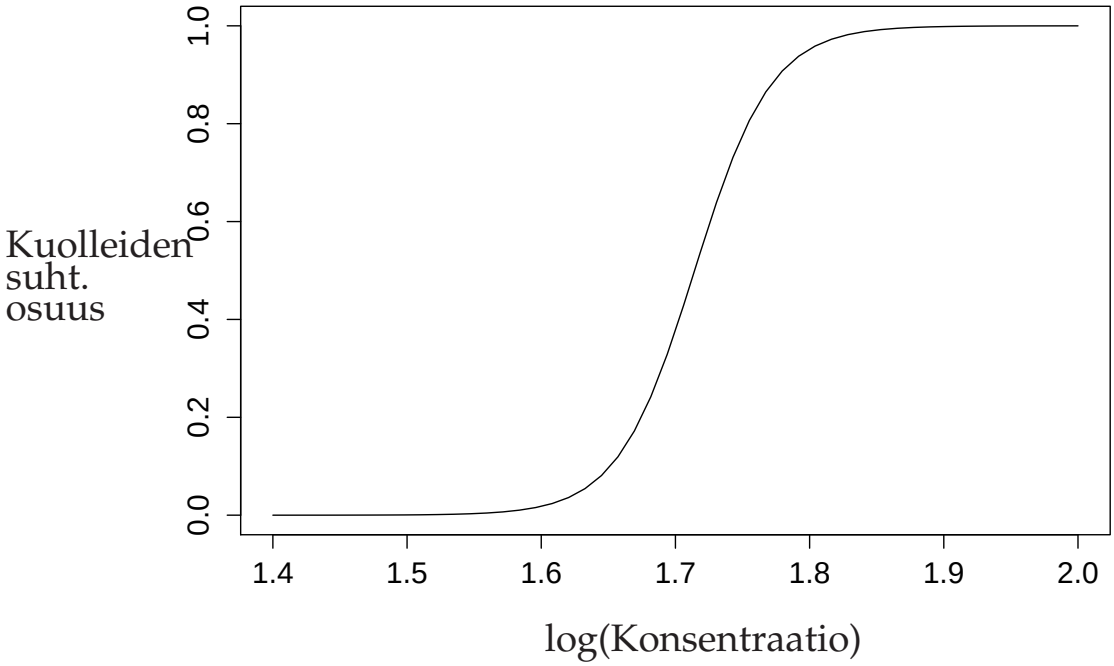
jossa $\mathsf{logit}(\cdot)$ on **logit-funktio**. Lukua

$$\frac{p_i}{1-p_i} = \exp(z_i),$$

jossa p_i on suotuisan tapahtuman todennäköisyys, kutsutaan **vedonlyöntisuhteeksi**. Yhteensä logistinen regressiomalli on yksinkertaisessa yhden selittävän muuttujan tapauksessa muotoa

$$\mathsf{logit}\,(p_i) = \beta_0 + \beta_1 d_i, \quad i = 1, \ldots, n.$$

Oletetaan, että n_i subjektia saa annoksen d_i, ja että y_i kappaletta näistä saa reaktion. Nyt Y_i on binomijakautunut satunnaismuuttuja parametrein



Kuva 1.23: Kuolleiden suhteellinen osuus myrkkyannoksen funktiona.

n_i ja p_i, jossa logistisen mallin mukaisesti

$$p_i = 1 - \frac{1}{1 + \exp(\beta_0 + \beta_1 d_i)}. \tag{1.5}$$

Oletetaan, että Y_i:t ovat riippumattomia, eli käytännössä samoille subjekteille ei anneta erikokoisia annoksia. Riippumattomuuden nojalla yhteistodennäköisyysjakauma annoksille on

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}. \tag{1.6}$$

Käytännössä parametrien β_0 ja β_1 estimaatit lasketaan siten, että niille yritetään arvot jotka maksimoivat funktion 1.6. Yleisesti ottaen menetelmää kutsutaan suurimman uskottavuuden estimoinniksi. Maksimoitava funktion voidaan kirjoittaa muodossa

$$\begin{aligned} l(\beta_0, \beta_1) &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{p_i}{1 - p_i}\right) + n_i \log(1 - p_i) \right] \\ &= \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_i d_i) + n_i \log(1 - p_i) \right]. \end{aligned}$$

Maksimiarvot löydetään ratkaisemalla yhtälöryhmä

$$\frac{\mathrm{d}l(\beta_0, \beta_1)}{\mathrm{d}\beta_0} = 0$$
$$\frac{\mathrm{d}l(\beta_0, \beta_1)}{\mathrm{d}\beta_1} = 0.$$

Yleisesti yhtälöryhmä pitää ratkaista numeerisesti käyttäen optimointimenetelmiä.

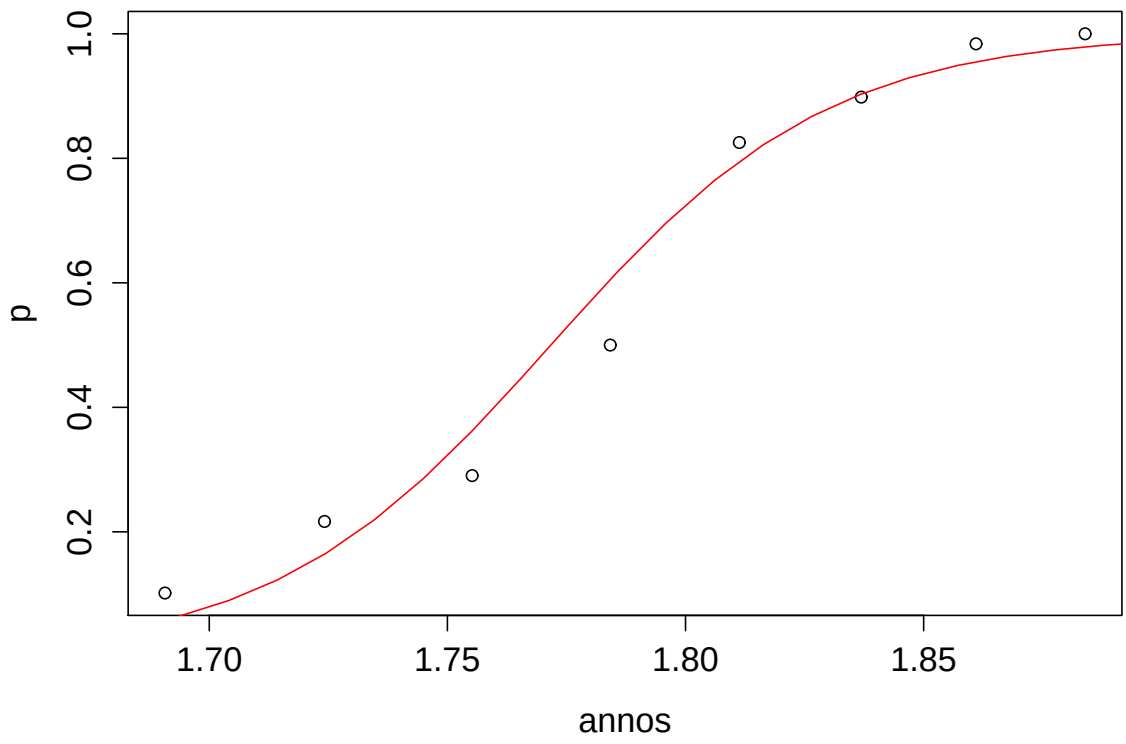
Esimerkki 13. Tarkastellaan tarkemmin Esimerkkiä 12

Annos	Kovakuoriaisia	Kuolleita
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Sovitetaan logistinen regressiomalli aineistoon käyttäen R:ää. Aineisto syötetään muodossa jossa jokaiselle kovakuoriaiselle annettu myrkkyannos ja reaktio (kuollut/ei kuollut) on oma kokeensa. Käytetään komentoa `glm()`. Saadaan seuraavat tulokset

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
c1            34.270      2.912   11.77   <2e-16 ***
```

Havainnollistetaan alkuperäistä aineistoa ja sovittettua käytää kuvassa 1.24



Kuva 1.24: Vaikutuksen todennäköisyys myrkkyannoksen funktiona.

Lähteet

[1] T. W. Anderson, *An Introduction to the Statistical Analysis of Data*, Houghton Mifflin Company, 1978

[2] R. B. Ash, *Basic Probability Theory*, Dover Publications, Inc., Mineola, New York

[3] M. Grönroos, *Johdatus tilastotieteeseen: Kuvailu, mallit ja päättely*, Oy Finn Lectura Ab, 2003, Helsinki.

[4] G. Casella, R. L. Berger, *Statistical Inference*, 2ed.

[5] K. Ruohonen, Tilastomatematiikka, Luentomoniste, TTY

[6] G. J. Kerns, *Introduction to Probability and Statistics Using R*, 2010.

[7] I. Mellin, *Todennäköisyyslaskenta: Todennäköisyys ja sen laskusäännöt*, Luentomoniste, TKK.

Luku 2

χ^2-testit

2.1 Riippumattomuustesti

Tarkastellaan tilannetta, jossa tutkittaviin tilastoyksiköihin liittyy kaksi tilastollista muuttujaa X ja Y, jotka ovat kategorisia tai diskreettejä numeerisia muuttujia. Molemmilla tilastollisilla muuttujilla on äärellinen määrä mahdollisia arvoja, ja molemmat muuttujista ovat vasteita. Nyt yhteisfrekvenssijakauma ja reunafrekvenssijakaumat voidaan esittää kontingenssitaulukon, eli ristiintaulukon muodossa. Olkoon X:n mahdolliset arvot $\{x_1, \ldots, x_r\}$ ja Y:n mahdolliset arvot $\{y_1, \ldots, y_s\}$. Nyt olkoon satunnaismuuttujien pistetodennäköisyysfunktiot

$$\begin{aligned} \mathsf{P}(X = x_i) &= p_i \\ \mathsf{P}(Y = y_j) &= q_j. \end{aligned}$$

Satunnaismuuttujat X ja Y ovat toisistaan riippumattomat jos ja vain jos

$$\mathsf{P}(X = x_i, Y = y_j) = p_{i,j} = \mathsf{P}(X = x_i)\,\mathsf{P}(Y = y_j) = p_i q_j.$$

Haluamme testata ovatko tilastolliset muuttujat X ja Y toisistaan riippumattomat käyttäen havaintoaineistoa, joka voidaan ilmaista yhteisfrekvenssijakauman avulla. Yhteisfrekvenssijakauma on kontingens-

sitaulukon avulla ilmaistuna

		Y				
		y_1	$\dots$	y_j	$\dots$	y_s	
X	x_1	$f_{1,1}$	$\dots$	$f_{1,j}$	$\dots$	$f_{1,s}$	$f_{1,\cdot}$
		$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	$f_{i,1}$	$\dots$	$f_{i,j}$	$\dots$	$f_{i,s}$	$f_{i,\cdot}$
		$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	$f_{r,1}$	$\dots$	$f_{r,j}$	$\dots$	$f_{r,s}$	$f_{r,\cdot}$
		$f_{\cdot,1}$	$\dots$	$f_{\cdot,j}$	$\dots$	$f_{\cdot,s}$	n

Kontingenssitaulukossa $f_{i,j}$ on niiden havaintojen lukumäärä, jossa tilastolliset muuttujat saavat arvot x_i ja y_j. Tilastollisten muuttujien X ja Y (reuna)frekvenssijakaumat muodostuvat frekvensseistä $f_{i,\cdot}$ ja $f_{\cdot,j}$, ja ilmoittavat montako kertaa x_i esiintyy aineistossa ($f_{i,\cdot}$) ja y_j esiintyy aineistossa ($f_{\cdot,j}$). Havaintojen kokonaismäärä on n.

Tarkastellaan hypoteesiparia

$$\begin{aligned} \mathsf{H}_0 &: p_{i,j} = p_i q_j, \quad \forall i, j \\ \mathsf{H}_v &: \exists\, i, j \; p_{i,j} \neq p_i q_j. \end{aligned}$$

Nyt odotettu frekvenssi $\mathsf{E}\,(F_{i,j})$ tapahtumalle $(X = x_i, Y = y_j)$ on binomijakauman (multinomijakauman) mukaisesti $\mathsf{E}\,(F_{i,j}) = np_{i,j}$, joka taas voidaan nollahypoteesin ollessa voimassa ilmaista np_iq_j. Todennäköisyyksiä p_i ja q_j estimoidaan suhteellisten frekvenssien avulla

$$\begin{aligned} \hat{p}_i &= \frac{f_{i,\cdot}}{n} \\ \hat{q}_j &= \frac{f_{\cdot,j}}{n}, \end{aligned}$$

jolloin voimme estimoida frekvenssin odotusarvoa

$$e_{i,j} = n\hat{p}_i\hat{q}_j = \frac{f_{i,\cdot}f_{\cdot,j}}{n}.$$

Standardoidut jäännökset määritellään

$$d_{i,j} = \frac{f_{i,j} - e_{i,j}}{\sqrt{e_{i,j}}}.$$

Riippumattomuuden testaamiseen voidaan käyttään standardoitujen

jäännösten neliösummaa

$$h = \sum_{i=1}^{r} \sum_{j=1}^{s} d_{i,j}^2.$$

Vastaavalla satunnaismuuttujalla

$$H = \sum_{i=1}^{r} \sum_{j=1}^{s} D_{i,j}^2$$

on nollahypoteesin ollessa voimassa Pearsonin approksimaation mukaisesti $\chi^2(v)$, jossa vapausasteet $v = (r-1)(s-1)$. Kriittinen alue testille löytyy jakauman $\chi^2(v)$ oikeasta hännästä, sillä poikkeamat riippumattomuusoletuksesta kasvattavat ylläolevaa neliösummaa. Joskus kirjallisuudessa mainitaan, että kaikkien odotettujen frekvenssien $f_{i,\cdot} f_{\cdot,j}/n$ pitäisi olla vähintään 5, jotta χ^2-approksimaatio olisi hyvä.

Esimerkki 14. Tarkastellaan tilannetta, jossa kolmelta eri linjalta valmistuu tuotteita (L_i) ja tuotteissa on kahden tyyppisiä virheitä (V_i). Halutaan tarkastella onko virhetyyppi ja linja toisistaan riippumattomia. Kerätään havaintoaineisto linjastoilla tapahtuneista erilaisten virheiden lukumäärästä

		Linja		
		L_1	L_2	L_3	
Virhetyyppi	V_1	15	21	45	81
	V_2	26	31	34	91
		41	52	79	172

Testataan hypoteesiparia

$$\mathsf{H}_0 : p_{i,j} = p_i q_j, \quad \forall i, j$$
$$\mathsf{H}_v : \exists\, i, j\ p_{i,j} \neq p_i q_j.$$

tasolla 0.05. Testisuureella

$$H = \sum_{i=1}^{r} \sum_{j=1}^{s} D_{i,j}^2$$

on nollahypoteesin ollessa voimassa Pearsonin approksimaation mukaisesti $\chi^2(2)$, joten kriittinen arvo testille tasolla 0.05 on $q_{0.05}^{(2)} = 5.991$.

Havaittu testisuureen arvo on

$$\begin{aligned} h_{\text{hav}} &= \frac{\left(15 - \frac{81\cdot 41}{172}\right)^2}{\frac{81\cdot 41}{172}} + \frac{\left(21 - \frac{81\cdot 52}{172}\right)^2}{\frac{81\cdot 52}{172}} + \frac{\left(45 - \frac{81\cdot 79}{172}\right)^2}{\frac{81\cdot 79}{172}} \\ &\quad + \frac{\left(26 - \frac{91\cdot 41}{172}\right)^2}{\frac{91\cdot 41}{172}} + \frac{\left(31 - \frac{91\cdot 52}{172}\right)^2}{\frac{91\cdot 52}{172}} + \frac{\left(34 - \frac{91\cdot 79}{172}\right)^2}{\frac{91\cdot 79}{172}} \\ &= 5.844, \end{aligned}$$

joka ei osu kriittiselle alueelle (kuitenkin hyvin lähelle kriittistä arvoa). Nyt tasolla 0.05 nollahypoteesi jää voimaan, eli virhetyyppi sekä linjasto ovat toisistaan riippumattomia.

2.2 Homogeenisuustesti

Tarkastellaan nyt tilannetta, jossa Y on edelleen vaste, mutta X on tekijä. Tekijä-vaste tilannetta tarkasteltaessa ei tarkastella riippuvuutta vaan muuttujien assiosiaatiota. Olkoon X:n mahdolliset arvot $\{x_1, \ldots, x_r\}$ ja koska x on tekijä, niin x:n jakauma on ennalta määrätty. Merkitään $f_{i,\cdot} = n_i$, jossa n_i on x_i:n frekvenssi. Nyt Y:n mahdolliset arvot ovat $\{y_1, \ldots, y_s\}$ ja voimme ajatella, että havaintoaineisto koostuu r kappaleesta Y:n frekvenssijakaumia. Olemme kiinnostuneita nyt ovatko nämä r kpl Y:n jakaumia samanlaisia, eli ovatko jakaumat **homogeenisia**.

		Y				
		y_1	$\cdots$	y_j	$\cdots$	y_s	
X	x_1	$f_{1,1}$	$\cdots$	$f_{1,j}$	$\cdots$	$f_{1,s}$	n_1
		$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	$f_{i,1}$	$\cdots$	$f_{i,j}$	$\cdots$	$f_{i,s}$	n_i
		$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	$f_{r,1}$	$\cdots$	$f_{r,j}$	$\cdots$	$f_{r,s}$	n_r
		$f_{\cdot,1}$	$\cdots$	$f_{\cdot,j}$	$\cdots$	$f_{\cdot,s}$	n

Homogeenisuuden testaaminen tapahtuu täsmälleen samoin kuin riippumattomuuden. Hypoteesipari on nyt (X on tässä muotoilussa tekijänä)

$$\begin{aligned} &\mathsf{H}_0 : p_{1,j} = p_{2,j} = \cdots = p_{r,j}, j = 1, \ldots, s \\ &\mathsf{H}_v : \exists\, i, j, k \text{ s.e. } p_{i,k} \neq p_{j,k}. \end{aligned}$$

Oletetaan, että nollahypoteesi on voimassa. Merkitään nyt nollahypoteesin mukaisia todennäköisyyksiä

$$p_j = p_{1,j} = p_{2,j} = \cdots = p_{r,j}.$$

Nyt frekvenssin $F_{i,j}$ odotusarvo on

$$\mathsf{E}\,(F_{i,j}) = n_i p_j.$$

Tuntematonta todennäköisyyttä estimoidaan suhteellisten frekvenssien avulla

$$\hat{p}_j = \frac{f_{\cdot,j}}{n},$$

jolloin saamme frekvenssin odotusarvoa $\mathsf{E}\,(F_{i,j})$ estimoivat odotetut frekvenssit

$$e_{i,j} = n_i \hat{p}_j = \frac{n_i f_{\cdot,j}}{n}.$$

Huomaa, että muoto on odotetuille jäännöksille täsmälleen sama kuin riippumattomuuden testauksessa. Nollahypoteesin ollessa voimassa testisuureella

$$H = \sum_{i=1}^{r}\sum_{j=1}^{s} D_{i,j}^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \left(\frac{F_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}\right)^2$$

on likimääräisesti $\chi^2(v)$ jakauma, jossa $v = (r-1)(s-1)$.

Esimerkki 15. Pesuaineiden markkinatutkimuksessa todettiin satunnaisesti valittujen kotitalouksien edustajien mielipiteen parhaasta pesuaineesta jakautuvan kolmella paikkakunnalla seuraavasti:

	Pesuaine			
	A	B	C	Otoskoko
Turku	232	108	60	400
Tampere	260	139	101	500
Lahti	197	106	97	400
	689	353	258	1300

Testataan onko mieltymykset pesuaineista samanlaisia eri paikkakunnilla. Testataan tasolla 0.05 hypoteesiparia

$$\mathsf{H}_0 : p_{1,j} = p_{2,j} = p_{3,j}, j = 1, 2, 3$$
$$\mathsf{H}_v : \exists\, i, j, k \text{ s.e. } p_{i,k} \neq p_{j,k}.$$

Nyt testisuureella on jakauma

$$H = \sum_{i=1}^{r} \sum_{j=1}^{s} D_{i,j}^2 \sim \chi^2(4),$$

ja kriittinen arvo testille on $q_{0.05}^{(4)} = 9.488$. Lasketaan odotetut frekvenssit ja taulukoidaan ne

	Pesuaine		
	A	B	C
Turku	$\frac{400 \cdot 689}{1300}$	$\frac{400 \cdot 353}{1300}$	$\frac{400 \cdot 258}{1300}$
Tampere	$\frac{500 \cdot 689}{1300}$	$\frac{500 \cdot 353}{1300}$	$\frac{500 \cdot 258}{1300}$
Lahti	$\frac{400 \cdot 689}{1300}$	$\frac{400 \cdot 353}{1300}$	$\frac{400 \cdot 258}{1300}$

Havaittu testisuureen arvo on

$$\begin{aligned} h_{\text{hav}} = & \frac{\left(232 - \frac{400 \cdot 689}{1300}\right)^2}{\frac{400 \cdot 689}{1300}} + \frac{\left(108 - \frac{400 \cdot 353}{1300}\right)^2}{\frac{400 \cdot 353}{1300}} + \frac{\left(60 - \frac{400 \cdot 258}{1300}\right)^2}{\frac{400 \cdot 258}{1300}} \\ & + \frac{\left(260 - \frac{500 \cdot 689}{1300}\right)^2}{\frac{500 \cdot 689}{1300}} + \frac{\left(139 - \frac{500 \cdot 353}{1300}\right)^2}{\frac{500 \cdot 353}{1300}} + \frac{\left(101 - \frac{500 \cdot 258}{1300}\right)^2}{\frac{500 \cdot 258}{1300}} \\ & + \frac{\left(197 - \frac{400 \cdot 689}{1300}\right)^2}{\frac{400 \cdot 689}{1300}} + \frac{\left(106 - \frac{400 \cdot 353}{1300}\right)^2}{\frac{400 \cdot 353}{1300}} + \frac{\left(97 - \frac{400 \cdot 258}{1300}\right)^2}{\frac{400 \cdot 258}{1300}} \\ & = 11.85963, \end{aligned}$$

joka kuuluu kriittiselle alueelle. Nollahypoteesi voidaan siis hylätä tasolla 0.05. Havaintoaineiston perusteella siis eri kaupungeissa preferoidaan eri suhteessa pesuaineita A, B ja C.

Lähteet

[1] T. W. Anderson, *An Introduction to the Statistical Analysis of Data*, Houghton Mifflin Company, 1978

[2] R. B. Ash, *Basic Probability Theory*, Dover Publications, Inc., Mineola, New York

[3] M. Grönroos, *Johdatus tilastotieteeseen: Kuvailu, mallit ja päättely*, Oy Finn Lectura Ab, 2003, Helsinki.

[4] G. Casella, R. L. Berger, *Statistical Inference*, 2ed.

[5] K. Ruohonen, Tilastomatematiikka, Luentomoniste, TTY

[6] G. J. Kerns, *Introduction to Probability and Statistics Using R*, 2010.

[7] I. Mellin, *Todennäköisyyslaskenta: Todennäköisyys ja sen laskusäännöt*, Luentomoniste, TKK.