MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi
Luennot, osa II

G. Gripenberg

Aalto-yliopisto

11. helmikuuta 2015

Otos

- *Tavoitteena on saada tietoa satunnaismuuttujasta X.*
- *Teemme n koetta, esim. mittauksia, jotka antavat tuloksiksi $x_1, x_2, \ldots, x_n$ ja katsomme, että x_j on satunnaismuuttujan X_j saama arvo.*
- *Satunnaismuuttujat $X_1, \ldots, X_n$ muodostavat otoksen, jonka koko on n, ja luvut $x_1, x_2, \ldots, x_n$ muodostavat havaitun otoksen.*
- *Oletetaan (tavallisesti ja erikseen sanomatta) että satunnaismuuttujat $X_1, X_2, \ldots, X_n$ ovat riippumattomia ja että niillä on sama jakauma joka on sama kuin tutkittavan satunnaismuuttujan X jakauma.*

Mitta-asteikot

- *Laatueroasteikko: Luokkia ilman luonnallista järjestystä.*
- *Järjestysasteikko: Luokkia joilla on luonnollinen järjestys.*
- *Välimatka-asteikko: Numeerisia arvoja, erotukset merkityksellisiä, nolla mielivaltainen.*
- *Suhdeasteikko: Numeerisia arvoja, luonnollinen nollakohta.*

Huom!

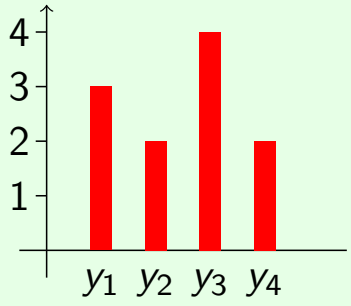
Oletus, että otoksen satunnaismuuttujat X_j ovat riippumattomia edellyttää, että käytämme "poimintaa takaisinpanolla" mutta tämä ehto harvoin täyttyy. Mutta käytännössä tulee esiin monta muuta paljon vakavampaa ongelmaa kun yritetään ottaa otos niin että oletukset ovat edes suurin piirtein voimassa ja tämä on tärkeä seikka, jota ei tässä käsitellä!

Havaintoarvojen jakauma ja sen kuvaaminen

Otoksen havaituista arvoista $x_1, x_2, \ldots, x_n$ voidaan muodostaa diskreetti todennäköisyysjakauma, ns. empiirinen jakauma siten, että $\Pr(H = x) = \frac{1}{n}|\{\, j : x_j = x \,\}|$ *(joka siis on tasainen diskreetti jakauma jos kaikki havaintoarvot poikkeavat toisistaan). Tätä jakaumaa voidaan odotusarvon, mediaanin, muiden kvanttiilien ym. lisäksi kuvata pylväsdiagrammilla tai histogrammilla tilanteesta riippuen.*

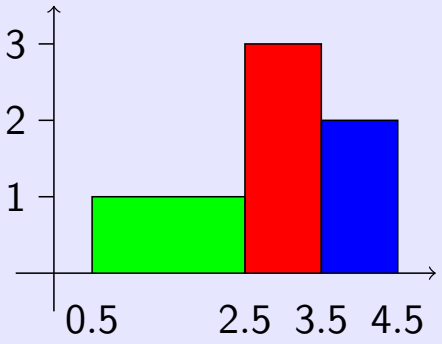
Pylväsdiagrammi

Jos otoksen havaintoarvojen mitta-asteikko on laatuero- tai järjestysasteikko ja/tai alkuperäinen satunnaismuuttuja on diskreetti voidaan havaitut otosarvot $x_1, x_2, \ldots, x_n$ esittää pylväsdiagrammmina missä jokaisen pylvään y_k korkeus on havaittu frekvenssi $f_k = |\{ j : x_j = y_k \}|$.



Histogrammi

Jos satunnaismuuttuja on jatkuva ja havaintoarvojen mitta-asteikko on välimatka- tai suhdeasteikko voidaan havaitut otosarvot $x_1, x_2, \ldots, x_n$ esittää histogrammina eli luokiteltuina frekvensseinä siten, että valitaan luokkarajat $a_0 < a_1 < \ldots < a_m$, lasketaan frekvenssit $f_k = |\{ j : a_{k-1} < x_j \leq a_k \}|$ ja nämä esitetään suorkaiteina, joiden pinta-alat ovat verrannollisia frekvensseihin.



Aritmeettinen keskiarvo

Jos X_j, $j = 1, \ldots, n$ on otos satunnaismuuttujasta X niin sen (aritmeettinen) keskiarvo on

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j,$$

ja

$$\mathrm{E}(\overline{X}) = \mathrm{E}(X) \quad \textit{ja} \quad \mathrm{Var}(\overline{X}) = \frac{1}{n} \mathrm{Var}(X),$$

koska odotusarvo on lineaarinen, riippumattomien satunnaismuuttujien summan varianssi on varianssien summa ja $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$.

Otosvarianssi

Jos X_j, $j = 1, \ldots, n$ on otos satunnaismuuttujasta X niin sen otosvarianssi on

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2,$$

ja (jos $\mathrm{Var}(X) < \infty$*)*

$$\mathrm{E}(S^2) = \mathrm{Var}(X),$$

joten otosvarianssi on varianssin harhaton estimaattori ja tämä on syy siihen, että nimittäjässä on $n - 1$ eikä n.

Huom

Jos $x_1, x_2, \ldots, x_n$ ovat havaittuja arvoja otoksesta satunnaismuuttujasta X niin näiden keskiarvo on $\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$ ja (havaittu) otosvarianssi on $s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2$.
Jos Matlab/Octavessa havainnot on vektorissa `x` *niin keskiarvo lasketaan komennolla* `mean(x)` *ja otosvarianssi komennolla* `var(x)`.

χ^2-jakauma

Jos satunnaismuuttujat $X_j \sim \mathrm{N}(0, 1)$, $j = 1, 2, \ldots, n$ ovat riippumattomia ja

$$C = \sum_{j=1}^{n} X_j^2$$

niin sanomme, että C on χ^2-jakautunut n:llä vapausasteella eli $C \sim \chi^2(n)$. Silloin

$$\mathrm{E}(C) = n \quad \textit{ja} \quad \mathrm{Var}(C) = 2n$$

ja C:llä on tiheysfunktio

$$f_C(t) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} t^{\frac{n}{2}-1} \mathrm{e}^{-\frac{t}{2}}, \quad t \geq 0.$$

och $f_C(x) = 0$ då $x < 0$.

Normaalijakauman otosvarianssi

Jos X_j, $j = 1, 2, \ldots, n$ *on otos* $\mathrm{N}(\mu, \sigma^2)$*-jakautuneesta satunnaismuuttujasta niin otosvarianssille pätee*

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1).$$

t-jakauma

Jos $Z \sim \mathrm{N}(0,1)$ *ja* $C \sim \chi^2(m)$ *ovat riippumattomia ja*

$$W = \frac{Z}{\sqrt{\frac{1}{m} C}}$$

niin sanomme, että W *on* t*-jakautunut* m*:llä vapausasteella eli* $W \sim t(m)$*.*
Silloin $\mathrm{E}(W) = 0$ *jos* $m > 1$ *ja* $\mathrm{Var}(W) = \frac{m}{m-2}$ *jos* $m > 2$ *ja* W*:llä on tiheysfunktio*

$$f(t) = \frac{1}{\sqrt{m\pi}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}}, \quad t \in \mathbb{R}.$$

Otos normaalijakaumasta

Jos X_j, $j = 1, 2, \ldots, n$ *on otos* $\mathrm{N}(\mu, \sigma^2)$*-jakautuneesta satunnaismuuttujasta niin*

$$\frac{\overline{X} - \mu}{\sqrt{\frac{1}{n} S^2}} \sim t(n-1).$$

Piste-estimaatti ja estimaattori

Oletamme, että tiedämme (tai uskomme) että X on satunnaismuuttuja, jolla on pistetodennäköisyysfunktio tai tiheysfunktio $f(x, \theta)$ missä parametri θ (joka myös voi olla vektori) on tuntematon. Miten voimme estimoida parametrin θ?

- *Otamme havaitun otoksen x_j, $j = 1, \ldots, n$.*
- *Laskemme estimaatin $\hat{\theta} = g(x_1, x_2, \ldots, x_n)$ missä g on jokin funktio.*
- *Huomaa, että $\hat{\theta}$ on luku tai vektori kun taas $\hat{\Theta} = g(X_1, X_2, \ldots, X_n)$ on satunnaismuuttuja.*
- *Joskus sanalla estimaattori tarkoitetaan funktiota g ja joskus satunnaismuuttujaa $\hat{\Theta}$.*

Väliestimaatti

Sen sijaan että parametrin estimaatiksi otetaan luku (tai vektori) voidaan myös laskea väli, jolla arvioidaan parametrin sijaitsevan.

Momenttimenetelmä, esimerkki

Satunnaismuuttujasta X on saatu havainnot 0.46, 0.20, 0.19, 0.09, 0.46 *ja* 0.16. *Meillä on syytä uskoa, että X on* $\mathrm{Exp}(\lambda)$*-jakautunut mutta parametriä λ emme tunne. Miten voimme laskea parametrille estimaatin? Koska tiedämme, että $\mathrm{E}(X) = \frac{1}{\lambda}$ niin on luonnolista käyttää momenttimenetelmää, eli laskemme havaintoarvojen keskiarvon ja saamme*

$$\overline{x} = \frac{1}{6}\sum_{j=1}^{6} = \frac{1}{6}(0.46 + 0.20 + 0.19 + 0.09 + 0.46 + 0.16) = 0.26$$

Koska $\lambda = \frac{1}{\mathrm{E}(X)}$ estimaatiksi tulee

$$\hat{\lambda} = \frac{1}{0.26} \approx 3.8.$$

Eksponenttijakauman tapauksessa voimme siis ottaa parametrin estimaattoriksi $\overline{X}^{-1}$ missä $\overline{X}$ on otoskeskiarvo. Estimaattori on siis satunnaismuuttuja kun taas estimaatti on havaintoarvoista laskettu luku. Tämä estimaattori ei ole harhaton koska $\mathrm{E}(\overline{X}^{-1}) > \lambda$ mutta kun n kasvaa se lähestyy oikeata arvoa eli $\lim_{n\to\infty} \Pr\left(\left|\lambda - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^{-1}\right| > \epsilon\right) = 0$ kaikilla $\epsilon > 0$.

Suurimman uskottavuuden menetelmä, esimerkki

Matkustat vieraaseen kaupunkiin ja lentokentällä näet kolme taksiautoa, joiden numerot ovat 57, 113 *ja* 758. *Montako taksiautoa on tässä kaupungissa?*

Oletamme, että kaupungissa on N *taksiautoa joiden numerot ovat* $1, 2, \ldots, N$ *ja että todennäköisyys että lentokentällä olevalla taksilla on numero* j *on* $\frac{1}{N}$ *kaikilla* $j = 1, 2, \ldots, N$.

Jos haluamme käyttää momenttimenetelmää niin laskemme ensin joukossa $\{1, \ldots, N\}$ *tasaisesti jakautuneen satunnaismuuttujan odotusarvon ja se on* $\mathrm{E}(X) = \sum_{i=1}^{N} i \cdot \frac{1}{N} = \frac{N(N+1)}{2N} = \frac{N+1}{2}$, *josta seuraa, että* $N = 2\mathrm{E}(X) - 1$. *Sitten laskemme havaintojen keskiarvon* $\overline{x} = \frac{1}{3}(57 + 113 + 758) = 309.33$ *jolloin estimaatiksi saamme* $\hat{N} = 2 \cdot 309.33 - 1 \approx 618$ *mikä on selvästi liian pieni luku.*

Toinen mahdollisuus on käyttää suurimman uskottavuuden periaatetta: Jos taksiautoja on N *niin todennäköisyys, että näet auton numerolla* 57 *on* $\frac{1}{N}$, *samalla todennäköisyydellä näet auton numerolla* 113 *ja auton numerolla* 758, *olettaen, että* $N \geq 758$ *koska muuten todennäköisyys, että näet auton numerolla* 758 *on* 0.

Suurimman uskottavuuden menetelmä, esimerkki, jatk.

Näin ollen

$$\mathcal{L}(N) = \Pr(\text{"näet numerot 57, 113 ja 758"}) = \begin{cases} \frac{1}{N^3}, & N \geq 758, \\ 0, & N < 758. \end{cases}$$

Suurimman uskottavuuden menetelmän mukaisesti valitsemme estimaatin $\hat{N}$ *siten että uskottavuusfunktio* $\mathcal{L}(N)$ *on mahdollisimman iso, eli tässä tapauksessa* $\hat{N} = 758$.

Vastaava tulos pätee yleisestikin, eli jos $X_1, X_2, \ldots, X_k$ *on otos tasajakaumasta joukossa* $\{1, 2, \ldots, N\}$ *(tai jatkuvassa tapauksessa välillä* $[0, N]$*) niin* N*:n suurimman uskottavuuden estimaattori on*

$$\hat{N} = \max(X_1, X_2, \ldots, X_k).$$

Tämä on harhainen estimaattori koska selvästikin $\mathrm{E}(\hat{N}) < N$ *mutta mikä* $\mathrm{E}(\max(X_1, X_2, \ldots, X_k))$ *oikein on?*

Suurimman uskottavuuden menetelmä, esimerkki, jatk.

Nyt $\Pr(\max(X_1, X_2, \ldots, X_k) \leq m) = \Pr(X_j \leq m, j = 1, \ldots, k) = \left(\frac{m}{N}\right)^k$ *josta seuraa, että* $\Pr(\max(X_1, X_2, \ldots, X_k) = m) = \left(\frac{m}{N}\right)^k - \left(\frac{m-1}{N}\right)^k$ *ja odotusarvoksi tulee*

$$\mathrm{E}\big(\max(X_1, X_2, \ldots, X_k)\big) = \sum_{m=1}^{N} m \left(\left(\frac{m}{N}\right)^k - \left(\frac{m-1}{N}\right)^k \right)$$

josta voi päätellä, että

$$\frac{k}{k+1} N < \mathrm{E}(\max(X_1, X_2, \ldots, X_k)) < \frac{k}{k+1} N + 1.$$

Näin ollen parempi estimaattori N:lle voisi olla

$$\frac{k+1}{k} \max(X_1, X_2, \ldots, X_k),$$

joka on harhaton jatkuvassa tapauksessa. Näin ollen parempi estimaatti taksiautojen lukumäärälle olisi $\frac{4}{3} \cdot 758 \approx 1011$.

Momenttimenetelmä

Jos satunnaismuuttujan X *pistetodennäköisyys- tai tiheysfunktio* $f_X(t, \theta)$ *on sellainen, että* θ *voidaan esittää odotusarvon* $\mathrm{E}(X)$ *funktiona eli* $\theta = h(\mathrm{E}(X))$ *niin parametrin* θ *momenttiestimaattori on*

$$\hat{\Theta} = h\left(\frac{1}{n} \sum_{j=1}^{n} X_j \right).$$

Jos parametri, tai parametrit, voidaan esittää muodossa $h(\mathrm{E}(X), \mathrm{E}(X^2))$ *niin estimaattoriksi valitaan*

$\hat{\Theta} = h\left(\frac{1}{n} \sum_{j=1}^{n} X_j, \frac{1}{n} \sum_{j=1}^{n} X_j^2 \right)$.

Suurimman uskottavuuden periaate

Jos todennäköisyysjakauman pistetodennäköisyys- tai tiheysfunktio on $f(x,\theta)$ *niin parametrin* θ *suurimman uskottavuuden estimaatti* $\widehat{\theta}$ *valitaan siten, että*

$$L(\widehat{\theta}, x_1, x_2, \ldots, x_n) = \max_{\theta} L(\theta, x_1, x_2, x_n),$$

missä

$$L(\theta, x_1, x_2, x_n) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta)$$

on ns. uskottavuusfunktio ja x_j, $j = 1, \ldots, n$ *on havaittu otos satunnaismuuttujasta jonka pistetodennäköisyys- tai tiheysfunktio on* $f(x,\theta)$.
Diskreetissä tapauksessa $L(\theta, x_1, x_2, x_n)$ *on todennäköisyys, että saadaan havaittu otos* x_j, $j = 1, \ldots, n$ *kun parametrin arvo on* θ. *Jatkuvassa tapauksessa* $(2h)^n L(\theta, x_1, \ldots, x_n)$ *on pienillä positiivisilla h:n arvoilla suurin piirtein todennäköisyys, että saadaan havaittu otos* y_j, $j = 1, \ldots, n$ *siten, että* $|y_j - x_j| < h$ *kaikilla* j.

Eksponenttijakauman parametrin luottamusväli, esimerkki

Oletamme, että meillä on otos $\mathrm{Exp}(\lambda)$*-jakaumasta siten, että otoksen koko on* 50 *ja keskiarvo* 0.8. *Momenttimenetelmällä saamme silloin parametrille* λ *estimaatin* $\hat{\lambda} = \frac{1}{0.8} = 1.25$ *mutta tässä on kyse siitä, että miten voimme määrittää välin siten, että jos laskemme monella otoksella ja tällä samalla menetelmällä monta tällaista väliä, niin suurin piirtein esim.* 95% *tapauksista ovat sellaisia, että parametri kuuluu väliin, jonka olemme siinä tapauksessa laskeneet havaintoarvoista.*
Tätä varten tarvitsemme satunnaismuuttujan, jonka jakauma ainakin approksimatiivisesti on täysin tunnettu eikä siis riipu mistään tuntemattomista parametreistä. Keskeisen väliarvolauseen nojalla tällaiseksi jakaumaksi otetaan usein normaalijakauma $\mathrm{N}(0,1)$ *ja näin teemme nytkin. Unohdamme hetkeksi numeroarvot ja oletamme että meillä on otos* $X_1, X_2, \ldots, X_{50}$ *satunnaismuuttujasta* $X \sim \mathrm{Exp}(\lambda)$. *Otoskeskiarvon* $\overline{X} = \frac{1}{50}\sum_{j=1}^{n} X_j$ *odotusarvo on silloin* $\mathrm{E}(\overline{X}) = \mathrm{E}(X) = \frac{1}{\lambda}$ *ja varianssi* $\mathrm{Var}(\overline{X}) = \frac{1}{50}\mathrm{Var}(X) = \frac{1}{50} \cdot \frac{1}{\lambda^2}$.

Eksponenttijakauman parametrin luottamusväli, esimerkki, jatk.

Jos n on riittävän iso niin

$$\frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{50\lambda^2}}} \sim_a \mathrm{N}(0,1).$$

Jos $Z \sim \mathrm{N}(0,1)$ *niin*

$$\Pr\left(F^{-1}_{\mathrm{N}(0,1)}(0.025) \leq Z \leq F^{-1}_{\mathrm{N}(0,1)}(0.975)\right) = \Pr(-1.96 \leq Z \leq 1.96) = 0.95,$$

joten

$$\Pr\left(-1.96 \leq \frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{50\lambda^2}}} \leq 1.96\right) \approx 0.95.$$

Nyt

$$-1.96 \leq \frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{50\lambda^2}}} \leq 1.96 \quad \Leftrightarrow \quad \frac{1 - \frac{1.96}{\sqrt{50}}}{\overline{X}} \leq \lambda \leq \frac{1 + \frac{1.96}{\sqrt{50}}}{\overline{X}},$$

Eksponenttijakauman parametrin luottamusväli, esimerkki, jatk.

joten todennäköisyys että λ on satunnaismuuttujien $\frac{0.72}{\overline{X}}$ ja $\frac{1.28}{\overline{X}}$ välillä on myös 0.95. *Näin ollen eksponenttijakauman parametrin* 95 %*:n approksimatiivinen luottamusväli kun otoskoko on* 50 *on*

$$\left[\frac{0.72}{\overline{X}}, \frac{1.28}{\overline{X}}\right].$$

Tässä tapauksessa havaituksi luottamusväliksi tulee $[0.9, 1.6]$.

Eksponenttijakauman kohdalla ei ole erityisen hankalaa saada epäyhtälöt parametrille, mutta jos näin olisi ollut (kuten on asian laita esim Bernoulli-jakauman kohdalla) olisimme voineet varianssin lausekkeessa $\frac{1}{\lambda^2}$ käyttää λ:n estimaattoria $\overline{X}^{-1}$ jolloin luottamusväliksi olisi tullut

$$\left[\frac{1}{\overline{X} + \frac{1.96}{\sqrt{50}}\overline{X}}, \frac{1}{\overline{X} - \frac{1.96}{\sqrt{50}}\overline{X}}\right] = \left[\frac{0.78}{\overline{X}}, \frac{1.38}{\overline{X}}\right],$$

jolloin havaituksi luottamusväliksi tässä tapauksessa olisi tullut $[0.97, 1.73]$.

Luottamusväli

Todennäköisyysjakauman parametrin θ luottamusväli luottamustasolla $1-\alpha$ on väliestimaattori

$$I(X_1, X_2, \ldots, X_n) = [L(X_1, X_2, \ldots, X_n), U(X_1, X_2, \ldots, X_n)]$$

siten, että

$$\Pr\big(\theta \in I(X_1, X_2, \ldots, X_n)\big) = 1 - \alpha.$$

Usein sanaa luottamusväli käytetään myös välistä $I(x_1, x_2, \ldots, x_n)$, missä siis satunnaismuuttujat X_j on korvattu havaituilla arvoilla x_j, $j = 1, \ldots, n$.

Huom!

Tavallisesti luottamusväli valitaan **symmetriseksi** *siten, että*

$$\Pr(\theta < L(X_1, X_2, \ldots, X_n)) = \Pr(\theta > U(X_1, X_2, \ldots, X_n)) = \frac{1}{2}\alpha.$$

Usein joutdutaan tyytymään siihen, että luottamusväliin liittyvät ehdot ovat voimassa ainoastaan approksimatiivisesti.

Odotusarvon luottamusväli kun $X \sim \mathrm{N}(\mu, \sigma^2)$

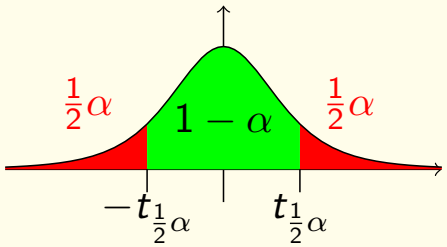
Jos $X_1, X_2, \ldots, X_n$ on otos, jonka keskiarvo on $\overline{X}$ ja otosvarianssi S^2, $\mathrm{N}(\mu, \sigma^2)$-jakautuneesta satunnaismuuttujasta niin

$$\left[\overline{X} - \sqrt{\frac{S^2}{n}} F^{-1}_{\mathrm{t}(n-1)}\left(1 - \tfrac{1}{2}\alpha\right), \overline{X} + \sqrt{\frac{S^2}{n}} F^{-1}_{\mathrm{t}(n-1)}\left(1 - \tfrac{1}{2}\alpha\right)\right],$$

on μ:n symmetrinen luottamusväli luottamustasolla $1-\alpha$.

Miksi?

Jos W on $t(n-1)$-jakautunut ja merkitään $t_{\frac{1}{2}\alpha} = F^{-1}_{\mathrm{t}(n-1)}\left(1 - \frac{1}{2}\alpha\right) = -F^{-1}_{\mathrm{t}(n-1)}\left(\frac{1}{2}\alpha\right)$ niin $\Pr(-t_{\frac{1}{2}\alpha} \le W \le t_{\frac{1}{2}\alpha}) = 1 - \alpha$. Jos nyt $W = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}}$ niin $-t_{\frac{1}{2}\alpha} \le W \le t_{\frac{1}{2}\alpha}$ jos ja vain jos $\overline{X} - \sqrt{\frac{S^2}{n}} t_{\frac{1}{2}\alpha} \le \mu \le \overline{X} + \sqrt{\frac{S^2}{n}} t_{\frac{1}{2}\alpha}$.



Todennäköisyyden p luottamusväli kun jakauma on Bernoulli(p)

Jos $X_1, X_2, \ldots, X_n$ *on otos* Bernoulli(p)*-jakautuneesta satunnaismuuttujasta ja otoskeskiarvo on* $\overline{X}$ *niin*

$$\left[\overline{X} - \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} F^{-1}_{\mathrm{N}(0,1)}\left(1-\tfrac{1}{2}\alpha\right), \overline{X} + \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} F^{-1}_{\mathrm{N}(0,1)}\left(1-\tfrac{1}{2}\alpha\right)\right]$$

on todennäköisyyden p **approksimatiivinen** *luottamusväli luottamustasolla* $1-\alpha$.

Miksi?

Jos $\widetilde{Z}$ *on approksimatiivisesti* $\mathrm{N}(0,1)$*-jakautunut ja merkitään* $z_{\frac{1}{2}\alpha} = F^{-1}_{\mathrm{N}(0,1)}\left(1-\frac{1}{2}\alpha\right)$ *niin* $\Pr(-z_{\frac{1}{2}\alpha} \leq \widetilde{Z} \leq z_{\frac{1}{2}\alpha}) \approx 1-\alpha$. *Nyt* $\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim_a \mathrm{N}(0,1)$ *mutta* p *korvataan nimittäjässä sen estimaattorilla* $\overline{X}$ *ja jos* $\tilde{Z} = \frac{\overline{X}-p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}}$ *niin* $-z_{\frac{1}{2}\alpha} \leq \widetilde{Z} \leq z_{\frac{1}{2}\alpha}$ *täsmälleen silloin kun*

$$\overline{X} - \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} z_{\frac{1}{2}\alpha} \leq p \leq \overline{X} + \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} z_{\frac{1}{2}\alpha}.$$

Huom!

Usein käytetty merkintä on

$$t_\alpha = t_\alpha(m) = F^{-1}_{\mathrm{t}(m)}(1-\alpha) = -F^{-1}_{\mathrm{t}(m)}(\alpha),$$

mikä siis tarkoittaa, että jos X *on* $\mathrm{t}(m)$ *jakautunut satunnaismuuttuja, niin*

$$\Pr(X \leq -t_\alpha) = \Pr(X \geq t_\alpha) = \alpha \quad \textit{ja} \quad \Pr(|X| \geq t_\alpha) = 2\alpha.$$

Vastaava merkintä $\mathrm{N}(0,1)$*-jakaumalle on* z_α.

Varianssin luottamusväli kun jakauma on N(μ, σ^2)

Jos $X_1, X_2, \ldots, X_n$ *on otos* $\mathrm{N}(\mu, \sigma^2)$*-jakautuneesta satunnaismuuttujasta ja otosvarianssi on* S^2 *niin*

$$\left[\frac{(n-1)S^2}{F^{-1}_{\chi^2(n-1)}\left(1-\frac{1}{2}\alpha\right)}, \frac{(n-1)S^2}{F^{-1}_{\chi^2(n-1)}\left(\frac{1}{2}\alpha\right)}\right]$$

on varianssin σ^2 *symmetrinen luottamusväli luottamustasolla* $1-\alpha$.

Hypoteesin testaus

- *Testataan onko syytä hylätä hypoteesi* H_0*, ns.* **nollahypoteesi** *sen perusteella, että saadut tulokset ovat hyvin epätodennäköisiä jos nollahypoteesi on voimassa vai onko kaikki vain seuraus sattumasta?*
- *Nollahypoteesi on tavallisesti jonkinlainen (vasta)väite, jonka kumoamiseksi tarvitaan perusteluita.*
- *Laskujen suorittamisen kannalta on oleellista, että nollahypoteesiksi valitaan riittävän yksiselitteinen väite, esim.* $\theta = \theta_0$ *eikä* $\theta \neq \theta_0$ *joka on liian epämääräinen. Usein riittää, että nollahypoteesillä on yksiselitteinen ääritapaus, esim.* $\theta = \theta_0$ *jos nollahypoteesi on* $\theta \leq \theta_0$*.*
- *Nollahypoteesiin voidaan lisätä muita oletuksia jakaumista, riippumattomuudesta jne., joilla kaikilla voi olla merkitystä tuloksen kannalta mutta joita ei välttämättä pyritä kumoamaan tässä testissä.*

Hypoteesin testaus, jatk.

- *Kun otos on otettu ja havaittu, lasketaan testimuuttujalle arvo ja tämä* **testimuuttuja** *valitaan siten, että jos nollahypoteesi pätee niin sen jakauma on (ainakin suunnilleen) jokin standardijakauma.*
- *Nollahypoteesin perusteella lasketaan todennäköisyys, ns. p-***arvo***, että tämä testimuuttuja saa arvon, joka poikkeaa nollahypoteesin perusteella odotetusta vähintään yhtä paljon kuin havaittu otos.*
- *Jos p-arvo on pienempi kuin annettu* **merkitsevyystaso** *hylätään nollahypoteesi.*
- *Merkitsevyystaso on siis todennäköisyys (usein likiarvo ja jos nollahypoteesi sisältää epäyhtälöitä, yläraja todennäköisyydelle), että nollahypoteesi hylätään vaikka se olisi voimassa.*

Esimerkki: Kolikonheitto

Haluamme selvittää onko todennäköisyys, että tulos on klaava on 0.5 *tai jotain muuta kun heitämme tiettyä kolikkoa. Tästä syystä heitämme tätä kolikkoa* 400 *kertaa ja saamme* 170 *klaavaa ja* 230 *kruunaa. Mitä johtopäätöksiä voimme näiden tulosten perusteella vetää?*
Tässä tapauksessa valitsemme nollahypoteesiksi $H_0 : p = 0.5$ *missä siis* $p = \Pr(T)$. *Nollahypoteesiksi emme voisi valita* $p \neq 0.5$ *koska sen nojalla emme voisi laskea mitään ja epäyhtälö* $p \leq 0.5$ *olisi ollut perusteltu jos olisimme epäilleet, että* $p > 0.5$ *jolloin saatujen tulosten perusteella olisimme voineet todeta ettei mitään puhu nollahypoteesia vastaan. Vastaavasti* $p \geq 0.5$ *olisi ollut perusteltu jos olisimme epäilleet tai toivoneet, että* p *olisi pienempi kuin* 0.5 *mutta se seikka että saimme vähemmän kuin* $0.5 \cdot 400$ *klaavaa olisi harhaanjohtava peruste valita nollahypoteesiksi* $p \geq 0.5$.
Seuraavaksi oletamme siis että nollahypoteesi $H_0 : p = 0.5$ *on voimassa ja että eri heittojen tulokset ovat toisistaan riippumattomia. Nyt todennäköisyys* $\binom{400}{170} \cdot 0.5^{170} \cdot (1 - 0.5)^{230} = 0.00044$ *että saamme täsmälleen* 170 *klaavaa on epäolennainen mutta sen sijaan todennäköisyys*

Esimerkki: Kolikonheitto, jatk.

että saamme korkeintaan 170 *klaavaa on olennainen ja se on*

$$\sum_{j=0}^{170} \binom{400}{j} \cdot 0.5^j \cdot (1 - 0.5)^{400-j} = 0.00156$$

Koska odotusarvo on 200 *ja nollahypoteesi on* $p = 0.5$ *niin* 230 *tai sitä suurempi lukumäärä klaavoja olisi yhtä poikkeava tulos joten ns. p-arvoksi tulee*

$p = \Pr(\text{"korkeintaan 170 tai vähintään 230 klaavaa"}) = 2 \cdot 0.00156 = 0.0031.$

Tästä voimme vetää johtopäätöksen, että on hyvin epätodennäköistä että olisimme saaneet näin vähän klaavoja jos klaavan todennäköisyys todella on 0.5 *ja tämän voi esittää myös siten, että koska* $p < 0.01$ *hylkäämme nollahypoteesin merkitsevyystasolla* 1%. *Mutta jos merkitsevyystasoksi olisi valittu* 0.1% *niin silloin emme hylkäisi nollahypoteesia.*

Esimerkki: Kolikonheitto, jatk.

Jos emme halua laskea binomijakaumalla voimme käyttää normaaliapproksimaatiota esim. siten että satunnaismuuttuja X_j saa arvon 1 jos j:nnessä heitossa tulee klaava, muuten 0 jolloin $\sum_{j=1}^{400} X_j$ on klaavojen lukumäärä ja keskiarvo $\overline{X}$ niiden osuus.

$$\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{400}}} \sim_a \mathrm{N}(0,1),$$

jos p on klaavan todennäköisyys.
Huomaa, että kun käytämme normaaliapproksimaatiota on yhdentekevää käytämmekö keskiarvoa Bernoulli-jakautuneista muuttujista tai summaa $\sum_{j=1}^{400} X_j$ (jolloin nimittäjässä on $\sqrt{p(1-p)400}$) joka on binomijakautunut satunnaismuuttuja.

Esimerkki: Kolikonheitto, jatk.

Koska nollahypoteesin mukaan $p = 0.5$ niin tämä testimuuttuja saa arvon

$$\frac{\frac{170}{400} - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{400}}} = -3.$$

Koska testimuuttujan itseisarvoltaan suuret arvot ovat poikkeavia niin p-arvo on

$$p = \Pr(|Z| \geq 3) = 2F_{\mathrm{N}(0,1)}(-3) = 0.0027.$$

Vaikka normaaliapproksimaatiolla avulla saatu p-arvo poikkeaa tarkasta arvosta suuruusluokka on aivan sama ja samoin johtopäätökset.

Normaalijakautunut satunnaismuuttuja, odotusarvon testaus

Jos X_j, $j = 1, 2, \ldots, n$ on otos satunnaismuuttujasta X joka on $\mathrm{N}(\mu, \sigma^2)$-jakautunut ja nollahypoteesi on $\mu = \mu_0$ (tai $\mu \leq \mu_0$ tai $\mu \geq \mu_0$) niin testimuuttujaksi valitaan

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim \mathrm{t}(n-1),$$

missä $\overline{X}$ on otoskeskiarvo ja S^2 otosvarianssi.

Huom!

Normaalijakaumaoletuksesta seuraa, että tämä ei ole approksimatiivinen jakauma joten pieni otoskoko n ei välttämättä ole ongelmallinen.

Esimerkki

Oliko maaliskuu 2014 poikkeuksellinen sademäärän suhteen?
Maaliskuussa 2014 sademäärät olivat joillakin mittausasemilla seuraavat:

	1	2	3	4	5	6	7	8	9	10
Sademäärä	33	27	30	22	28	28	24	31	34	22

Vastaavat keskiarvot vuosilta 1981–2010 olivat

	1	2	3	4	5	6	7	8	9	10
Keskiarvo	39	37	38	36	36	26	35	29	30	21

Nyt on järkevää laskea miten paljon vuoden 2014 keskiarvot poikkeavat pitkäaikaisista keskiarvoista ja erotukset ovat seuraavat

	1	2	3	4	5	6	7	8	9	10
Erotus	-6	-10	-8	-14	-8	2	-11	2	4	1

Esimerkki, jatk.

Koska kysytään oliko maaliskuu 2014 poikkeuksellinen sademäärän suhteen niin nollahypoteesiksi valitaan väite ettei se ollut. Nollahypoteesiksi ei voi valita että vuosi 2014 oli poikkeuksellinen koska sen perusteella ei voi laskea mitään eikä kysymykseen sisältynyt mitään siitä millä tavalla sademäärät olisivat olleet poikkeuksellisia joten mitään sellaista ei ole syytä ottaa nollahypoteesiinkaan.
Nollahypoteesi on siis, että erotus vuoden 2014 ja pitkäaikaisen keskiarvon välillä on $\mathrm{N}(\mu, \sigma^2)$*-jakautunut missä* $\mu = 0$ *ja että nämä erotukset eri paikkakunnilla ovat riippumattomia, mikä ehkä on kyseenalaista.*
Erotusten keskiarvo on -4.8 *ja otosvarianssi* 41.733*. Näin ollen testimuuttuja* $W = \frac{\overline{X}-0}{\sqrt{\frac{S^2}{10}}}$ *saa arvon* -2.3496*. Koska nollahypoteesin nojalla testimuuttuja* W *on* $t(10-1)$ *jakautunut niin p-arvoksi tulee*

$$p = \Pr(|W-0| \geq |-2.3496-0|) = \Pr(W \leq -2.3496 \text{ tai } W \geq 2.3496)$$
$$= F_{\mathrm{t}(9)}(-2.3496) + 1 - F_{\mathrm{t}(9)}(2.3496) = 2F_{\mathrm{t}(9)}(-2.3496) = 0.043333,$$

joten hylkäämme nollahypoteesin merkitsevyystasolla 0.05*.*

Esimerkki, jatk

Jos olisi kysytty oliko maaliskuussa 2014 sademäärä poikkeusellisen pieni niin nollahypoteesiksi olisimme valinneet väitteen ettei se ollut eli että erotusten jakauma on $\mathrm{N}(\mu, \sigma^2)$ *missä* $\mu \geq 0$*. Testimuuttuja olisi ollut täsmälleen sama mutta p-arvoksi olisi saatu*

$$p = \Pr(W \leq -2.3496) = F_{\mathrm{t}(9)}(-2.3496) = 0.021667.$$

Jos olisi kysytty oliko maaliskuussa 2014 sademäärä poikkeusellisen suuri niin nollahypoteesiksi olisimme taas valinneet väitteen ettei se ollut eli että erotusten jakauma on $\mathrm{N}(\mu, \sigma^2)$ *missä nyt* $\mu \leq 0$*. Testimuuttuja olisi ollut täsmälleen sama ja koska keskiarvo on negatiivinen tämä on täysin nollahypoteesin mukaista eikä sitä pidä hylätä. Tällaisissa tapauksissa ei otosvarianssiakaan tarvitse laskea, riittää että laskemme keskiarvon.*

Osuus tai todennäköisyys, normaaliapproksimaatio

Jos X_j, $j = 1, 2, \ldots, n$ *on otos satunnaismuuttujasta* X *joka on* Bernoulli(p)*-jakautunut ja nollahypoteesi on* $p = p_0$ *(tai* $p \leq p_0$ *tai* $p \geq p_0$*) niin testimuuttujaksi valitaan*

$$\frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim_a \mathrm{N}(0,1).$$

Aivan yhtä hyvin voidaan otoksesta laskea summa $Y = \sum_{j=1}^{n} X_j$ *joka on* Bin(n, p) *jakautunut ja testimuuttuja (joka siis ei muutu) kirjoitetaan silloin muodossa*

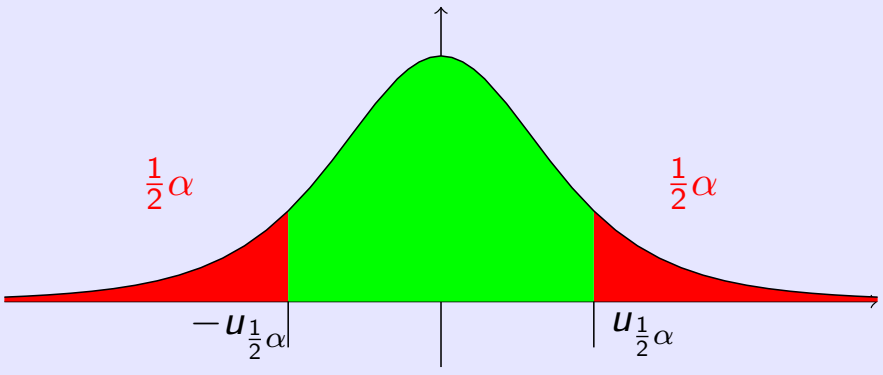
$$\frac{Y - np_0}{\sqrt{p_0(1-p_0)n}} \sim_a \mathrm{N}(0,1).$$

Huom!

Tässä tapauksessa jakauma on approksimatiivinen ja yleensä katsotaan, että approksimaatio on riittävän hyvä jos $\min(np_0, n(1-p_0)) \geq 10$.

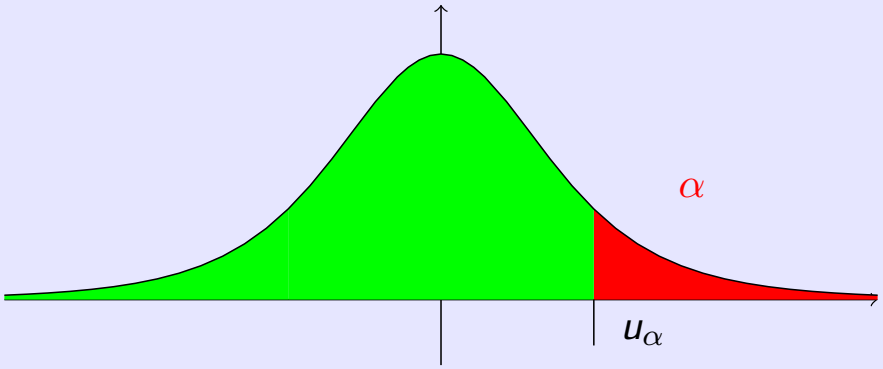
p-arvo, hylkäysalue, t(m)- tai N(0,1)-testimuuttuja

- *Oletetaan, että testimuuttuja* U *on* t(m)*- tai (approksimatiivisesti)* N(0,1)*-jakautunut (jolloin sen kertymäfunktio on* F_U*) ja se testissä saa arvon* u_*.
- *Jos vaihtoehto nollahypoteesille on kaksisuuntainen, eli nollahypoteesi on muotoa* $\mu = \mu_0$, $p = p_0$ *jne., eli jos tulokset ovat täsmälleen nollahypoteesin mukaisia ainoastaan kun testimuuttuja on* 0 *niin*
  - *p-arvo on* $\Pr(U \leq -|u_*| \text{ tai } U \geq |u_*|) = 2F_U(-|u_*|)$.
  - *Nollahypoteesi hylätään merkitsevyystasolla* α *jos* $p < \alpha$ *eli jos testimuuttujan arvo on hylkäysalueella* $(-\infty, -u_{\frac{1}{2}\alpha}) \cup (u_{\frac{1}{2}\alpha}, \infty)$ *missä* $u_{\frac{1}{2}\alpha} = -F_U^{-1}(\frac{1}{2}\alpha) = F_U^{-1}(1 - \frac{1}{2}\alpha)$.



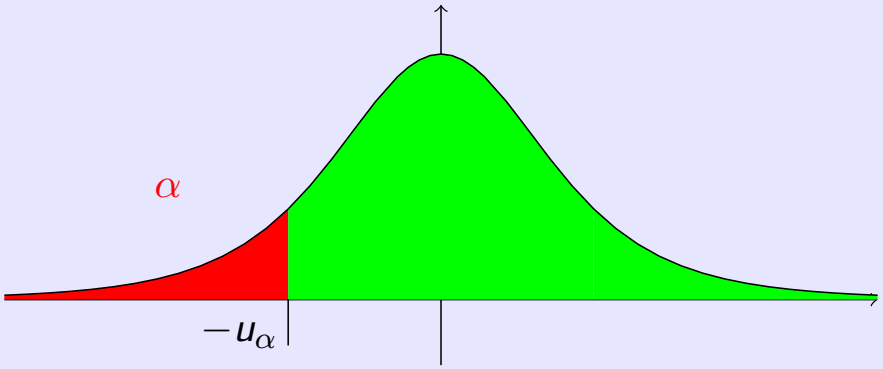
p-arvo, hylkäysalue, t(m)- tai N(0, 1)-testimuuttuja, jatk.

- *Jos vaihtoehto nollahypoteesille on yksisuuntainen ja nollahypoteesi on muotoa $\mu \leq \mu_0$, $p \leq p_0$ jne., eli jos tulokset ovat täsmälleen nollahypoteesin mukaisia kun testimuuttuja on ≤ 0 niin*
  - *p-arvo on* $\Pr(U \geq u_*) = 1 - F_U(u_*)$.
  - *Nollahypoteesi hylätään merkitsevyystasolla α jos $p < \alpha$ eli jos testimuuttujan arvo on hylkäysalueella (u_α, ∞) missä $u_\alpha = -F_U^{-1}(\alpha) = F_U^{-1}(1 - \alpha)$.*



p-arvo, hylkäysalue, t(m)- tai N(0, 1)-testimuuttuja, jatk.

- *Jos vaihtoehto nollahypoteesille on yksisuuntainen ja nollahypoteesi on muotoa $\mu \geq \mu_0$, $p \geq p_0$ jne., eli jos tulokset ovat täsmälleen nollahypoteesin mukaisia jos testimuuttuja on ≥ 0 niin*
  - *p-arvo on* $\Pr(U \leq u_*) = F_U(u_*)$.
  - *Nollahypoteesi hylätään merkitsevyystasolla α jos $p < \alpha$ eli jos testimuuttujan arvo on hylkäysalueella $(-\infty, -u_\alpha)$ missä $u_\alpha = -F_U^{-1}(\alpha) = F_U^{-1}(1 - \alpha)$.*



Normaalijakauma, kaksi otosta, sama varianssi, odotusarvojen vertailu

Jos X_j, $j = 1, 2, \ldots, n_x$ *ja* Y_j, $j = 1, 2, \ldots, n_y$ *ovat (riippumattomia) otoksia satunnaismuuttujista* X *ja* Y *missä* $X \sim \mathrm{N}(\mu_x, \sigma^2)$ *ja* $Y \sim \mathrm{N}(\mu_y, \sigma^2)$ *ja nollahypoteesi on* $\mu_x = \mu_y$ *(tai* $\mu_x \leq \mu_y$ *tai* $\mu_x \geq \mu_y$*) niin testimuuttujaksi valitaan*

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2} \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim \mathrm{t}(n_x + n_y - 2).$$

Esimerkki

Vuosina 1660-1740 syntyi Pariisissa 377 649 *tyttöä ja* 393 535 *poikaa ja samana aikana syntyi Lontoossa* 698 900 *tyttöä ja* 737 687 *poikaa. Onko tyttöjen osuuksissa eroja?*

Olkoon X_j *satunnaismuuttuja, jonka arvo on* 1 *jos lapsi* j *Pariisissa on tyttö ja* 0 *jos hän on poika. Vastaavasti* Y_j *on satunnaismuuttuja, jonka arvo on* 1 *jos lapsi* j *Lontoossa on tyttö ja* 0 *jos hän on poika. Lisäksi oletamme, että kaikki nämä satunnaismuutujat ovat riippumattomia ja että* $\Pr(X_j = 1) = p_P$ *ja* $\Pr(Y_j = 1) = p_L$*. Nollahypoteesi on tässä tapauksessa* $H_o : p_P = p_L$*.*

Nollahypoteesin perusteella emme tiedä mikä $p_P = p_L$ *on mutta voimme laskea estimaatin* $\hat{p}$ *tälle todennäköisyydelle toeteamalla, että kaiken kaikkiaan syntyi* 2 207 771 *lasta ja näistä yhteensä* 1 076 549 *oli tyttöjä joten* $\hat{p} = \frac{1\,076\,549}{2\,207\,771} \approx 0.48762$*. Tiedämme myös keskiarvot havaituista muuttujista ja ne ovat* $\overline{x} = 0.4897$ *ja* $\overline{y} = 0.4865$*.*

Satunnaismuuttujan $\overline{X}$ *varianssi on noin* $\frac{\widehat{P}(1-\widehat{P})}{n_P}$ *missä* $n_P = 771184$ *on Pariisissa syntyneiden lasten lukumäärä.*

Esimerkki, jatk.

Samoin satunnaismuuttujan $\overline{Y}$ *varianssi on noin* $\frac{\widehat{P}(1-\widehat{P})}{n_L}$ *missä* $n_L = 771184$ *on Lontoossa syntyneiden lasten lukumäärä.*
Näin ollen satunnaismuuttujan $\overline{X} - \overline{Y}$ *varianssi on noin* $\frac{\widehat{P}(1-\widehat{P})}{n_P} + \frac{\widehat{P}(1-\widehat{P})}{n_L}$ *joten testimuuttuja*

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{P}(1-\widehat{P})(\frac{1}{n_P} + \frac{1}{n_L})}}$$

on suurin piirtein $\mathrm{N}(0,1)$*-jakautunut.*
Tässä tapauksessa testimuuttujan arvoksi tulee

$$z = \frac{0.48970 - 0.48650}{\sqrt{0.48762 \cdot (1 - 0.48762) \cdot \left(\frac{1}{771184} + \frac{1}{1436587}\right)}} = 4.5350.$$

Nyt tämän testin p-arvo on

$$p \approx \Pr(|Z| \geq 4.535) = 2 \cdot F_{\mathrm{N}(0,1)}(-4.5350) = 0.00000576,$$

eli voimme hyvin perustein hylätä nollahypoteesin.

Kahden osuuden tai todennäköisyyden vertailu

Jos X_j, $j = 1, 2, \ldots, n_x$ *ja* Y_j, $j = 1, \ldots, n_y$ *ovat (riippumattomia) otoksia satunnaismuuttujista* X *ja* Y *missä* $X \sim \mathrm{Bernoulli}(p_x)$ *ja* $Y \sim \mathrm{Bernoulli}(p_y)$ *ja nollahypoteesi on* $p_x = p_y$ *(tai* $p_x \leq p_y$ *tai* $p_x \geq p_y$*) niin testimuuttujaksi valitaan*

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{P}(1-\widehat{P})(\frac{1}{n_x} + \frac{1}{n_y})}} \sim_a \mathrm{N}(0,1)$$

missä

$$\widehat{P} = \frac{n_x \overline{X} + n_y \overline{Y}}{n_x + n_y}.$$

Yhteensopivuus

Jos X_j, $j = 1, \ldots, n$ on otos satunnaismuuttujasta X jonka arvojoukko on $\cup_{k=1}^{m} A_k$ missä joukot A_k ovat pistevieraita ja nollahypoteesi on $\mathrm{H}_0 : \Pr(X \in A_k) = p_k$, $k = 1, \ldots, m$ niin testimuuttujaksi valitaan

$$C = \sum_{k=1}^{m} \frac{(O_k - np_k)^2}{np_k} \sim_a \chi^2(m-1),$$

missä O_k on joukon $\{ j : X_j \in A_k \}$ alkioden lukumäärä.

p-arvo, hylkäysalue, χ^2-testimuuttuja

- *Oletetaan että testimuuttuja C on (approksimatiivisesti) $\chi^2(k)$-jakautunut ja saa arvon c_*.*
- *Jos vaihtoehto nollahypoteesille on yksisuuntainen ja pienet testimuuttujan arvot ovat sopusoinnussa nollahypoteesin kanssa niin*
  - *p-arvo on $\Pr(C \geq c_*) = 1 - F_{\chi^2(k)}(c_*)$.*
  - *Nollahypoteesi hylätään merkitsevyystasolla α jos $p < \alpha$ eli jos testimuuttujan arvo on hylkäysalueella $(F^{-1}_{\chi^2(k)}(1-\alpha), \infty)$.*
- *Jos vaihtoehto nollahypoteesille on yksisuuntainen ja suuret testimuuttujan arvot ovat sopusoinnussa nollahypoteesin kanssa niin*
  - *p-arvo on $\Pr(C \leq c_*) = F_{\chi^2(k)}(c_*)$.*
  - *Nollahypoteesi hylätään merkitsevyystasolla α jos $p < \alpha$ eli jos testimuuttujan arvo on hylkäysalueella $(0, F^{-1}_{\chi^2(k)}(\alpha))$.*
- *Jos vaihtoehto nollahypoteesille on kaksisuuntainen niin*
  - *p-arvo on $2\min(F_{\chi^2(k)}(c_*), 1 - F_{\chi^2(k)}(c_*))$.*
  - *Nollahypoteesi hylätään merkitsevyystasolla α jos $p < \alpha$ eli jos testimuuttujan arvo on hylkäysalueella $(0, F^{-1}_{\chi^2(k)}(\frac{1}{2}\alpha)) \cup (F^{-1}_{\chi^2(k)}(1 - \frac{1}{2}\alpha), \infty)$.*

Esimerkki: Kolikonheitto toisella tavalla

Kuten aikaisemmassa esimerkissä olemme heittäneet kolikkoa 400 *kertaa ja saaneet* 170 *klaavaa ja* 230 *kruunaa. Otamme taas nollhypoteesiksi* $H_0 : p = 0.5$ *missä* $p = \Pr(T)$ *eli klaavan todennäköisyys (ja tässä tapauksessa muotoa* $p \geq 0.5$ *oleva nollahypoteesi ei ole mahdollinen). Kirjoitamme havaitut lukumäärät taulukkoon*

T	H
170	230

ja laskemme χ^2*-yhteensopivuustestin testimuuttujan* $C = \sum_{k=1}^{m} \frac{(O_k - np_k)^2}{np_k}$ *arvon joka on*

$$c = \frac{(170 - 400 \cdot 0.5)^2}{400 \cdot 0.5} + \frac{(230 - 400 \cdot 0.5)^2}{400 \cdot 0.5} = \frac{30^2}{200} + \frac{30^2}{200} = 9.$$

Nyt C *on suunnilleen* $\chi^2(2-1)$*-jakautunut ja ainostaan* C*:n suuret arvot ovat ristiriidassa nollahypoteesin kanssa joten testin p-arvo on*

Esimerkki: Kolikonheitto toisella tavalla, jatk.

$$p = \Pr(C \geq 9) = 1 - F_{\chi^2(1)}(9) = 0.0027.$$

Mistä nyt johtuu, että saimme täsmälleen saman tuloksen kuin aikaisemmin kun käytimme normaaliapproksimaatiota?
Jos $Y \sim \mathrm{Bin}(n, p)$ *niin*

$$\frac{(Y - np)^2}{np} + \frac{((n - Y) - n(1-p))^2}{n(1-p)} = \frac{(Y - np)^2}{np} + \frac{(-Y + np)^2}{n(1-p)}$$
$$= \frac{(Y - np)^2}{n}\left(\frac{1}{p} + \frac{1}{1-p}\right) = \frac{(Y - np)^2}{np(1-p)} = \left(\frac{Y - np}{\sqrt{np(1-p)}}\right)^2,$$

eli χ^2*-testin testimuuttuja on binomijakauman normaaliapproksimaation neliö ja* $\chi^2(1)$ *jakautunut satunnasimuuttuja on määritelmän mukaan* $\mathrm{N}(0,1)$*-jakautuneen satunnaismuuttujan neliö.*
Jos χ^2*-testissä luokkien lukumäärä* m *on suurempi kuin* 2 *niin on huomattavasti vaikeampaa osoittaa, että* $C \sim_a \chi^2(m-1)$.

Esimerkki, yhteensopivuus, χ^2-testi

Haluamme testaamalla selvittää onko satunnaismuuttujan $\frac{X}{Y}$, *missä* $X \sim \mathrm{N}(0,1)$ *ja* $Y \sim \mathrm{N}(0,1)$ *ovat riippumattomia, tiheysfunktio* $f(t) = \frac{1}{\pi}\frac{1}{1+t^2}$ *(jolloin siis kyse olisi ns. Cauchy-jakaumasta).*
Jos f *on tiheysfunktio niin kertymäfunktio on* $F(t) = \frac{1}{\pi}\arctan(t) + \frac{1}{2}$ *ja jos* U *on satunnaismuuttuja jonka kertymäfunktio on* F *niin*

$$\Pr(U \in (a_{k-1}, a_k]) = p_k = \frac{1}{8}, \quad k = 1, 2, \ldots 8$$

jos

$$a_0 = -\infty, \quad a_k = F^{-1}(\tfrac{k}{8}), \quad k = 1, 2, \ldots, 7, \quad a_8 = \infty,$$

eli

$$[a_0, a_1, \ldots, a_8] = [-\infty, -2.41421, -1, -0.41421, 0, 0.41421, 1, 2.41421, \infty].$$

Esimerkki, yhteensopivuus, χ^2-testi, jatk.

Jos nyt otamme otokset $x_1, \ldots, x_{100}$ *satunnaismuuttujasta* X *ja* $y_1, \ldots, y_{100}$ *satunnasimuuttujasta* Y *ja laskemme moniko luvuista* $\frac{x_j}{y_j}$ *kuuluu väliin* $(a_{k-1}, a_k]$ *niin saamme seuraavat tulokset:*

A_k	$(-\infty, -2.41421]$	$(2.41421, -1]$	$(-1, -0.41421]$	$(-0.41421, 0]$
O_K	7	12	12	11
A_k	$(0, 0.41421]$	$(0.41421, 1]$	$(1, 2.41421]$	$(2.41421, \infty]$
O_K	18	18	13	9

Tässä tapauksessa valitsimme välit siten, että $\Pr(U \in (a_{k-1}, a_k] = \frac{1}{8}$ *joten* $np_k = 100 \cdot \frac{1}{8} = 12.5$ *kaikilla* $k = 1, \ldots, 8$. χ^2*-testin testimuuttuja* $C = \sum_{k=1}^{m} \frac{(O_k - np_k)^2}{np_k}$ *saa arvon*

$$c = \frac{(7-12.5)^2}{12.5} + \frac{(12-12.5)^2}{12.5} + \frac{(12-12.5)^2}{12.5} + \frac{(11-12.5)^2}{12.5} + \frac{(18-12.5)^2}{12.5} + \frac{(18-12.5)^2}{12.5} + \frac{(13-12.5)^2}{12.5} + \frac{(9-12.5)^2}{12.5} = 8.48.$$

Esimerkki, yhteensopivuus, χ^2-testi, jatk.

Nyt C on suunnilleen $\chi^2(8-1)$ jakautunut ja ainostaan C:n suuret arvot ovat ristirridassa nollahypoteesin kanssa joten testin p-arvo on

$$p = \Pr(C \geq 8.48) = 1 - F_{\chi^2(7)}(8.48) = 0.29.$$

Näin ollen meidän ei pidä hylätä nollahypoteesia.
Huomaa, että hyvin pieni testiarvo, jonka seurauksena p-arvo olisi suunnilleen 1, ei merkitse välttämättä vahvaa tukea nollahypoteesille, vaan pikemmin herättää epäilyksen, että numerot olisi peukaloitu paremman yhteensopivuuden saavuttamiseksi.

Normaalijakautunut satunnaismuuttuja, varianssin testaus

Jos X_j, $j = 1, 2, \ldots, n$ on otos satunnaismuuttujasta X joka on $\mathrm{N}(\mu, \sigma^2)$-jakautunut ja nollahypoteesi on $\sigma^2 = \sigma_0^2$ (tai $\sigma^2 \leq \sigma_0^2$ tai $\sigma^2 \geq \sigma_0^2$ niin testimuuttujaksi valitaan

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1),$$

missä S^2 otosvarianssi.

Kahden odotusarvon vertailu, yleinen tapaus

Tietystä prosessista kerättiin mittausdataa tuotteen laadun varmistamiseksi ja prosessiin tehtiin muutoksia jotta mitatun suureen eli laadun varianssi pienenisi. Tässä onnistuttiin mutta arveltiin että samanaikaisesti myös laatu eli mitattu suure kasvoi. Tämän asian selvittämiseksi tehtiin mittauksia ennen muutoksia ja niiden jälkeen:

	Otoskoko	*Keskiarvo*	*Otosvarianssi*
Ennen	220	4.50	0.08
Jälkeen	250	4.56	0.04

Tässä meillä on otokset $X_1, X_2, \ldots, X_{220}$ *ja* $Y_1, Y_2, \ldots, Y_{250}$ *ja oletamme että kaikki nämä satunnaismuuttujat ovat riippumattomia, satunnaismuuttujilla* X_j *on sama jakauma ja samoin satunaismuuttujilla* Y_j *on sama jakauma. Sen sijaan meidän ei tarvitse tässä tapauksessa olettaa että ne olisivat normaalijakautuneita (ja koska varianssit eivät ole samoja emme voisi käyttää* t*-jakautunutta satunnaismuuttujaa) mutta kuitenkin sellaisia, että keskiarvot* $\overline{X}$ *ja* $\overline{Y}$ *ovat suunnilleen normaalijakautuneita keskeisen raja-arvolauseen nojalla.*

Kahden odotusarvon vertailu, yleinen tapaus, jatk.

Silloin myös

$$\overline{X} - \overline{Y} \sim_a N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{220} + \frac{\sigma_Y^2}{250}\right).$$

Tässä tapauksessa valitsemme nollahypoteesiksi $\mu_X \geq \mu_Y$ *vastaväitteenä arveluille, että laatu parani eli* $\mu_Y > \mu_X$*. Emme tiedä mitä* σ_X^2 *ja* σ_Y^2 *ovat mutta voimme estimoida ne otosvariansseilla* S_X^2 *ja* S_Y^2 *joten testimuuttujaksi otamme*

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{220} + \frac{S_Y^2}{250}}} \sim_a \mathrm{N}(0,1).$$

Testimuuttujan arvoksi tulee -2.622 *ja koska testimuuttujan suuret positiiviset arvot ovat sopusoinnussa nollahypoteesin kanssa niin p-arvoksi tulee*

$$p = \Pr(Z \leq -2.622) \approx F_{\mathrm{N}(0,1)}(-2.622) = 0.0044,$$

eli hylkäämme nollahypoteesin merkitsevyystasolla 0.01*.*

Mikä on p-arvon jakauma jos nollahypoteesi pätee?

p-arvo on yleensä luku mutta jos emme laske testimuuttujan arvolla vaan otamme testimuuttuja satunnaismuuttujana, niin silloin p-arvokin on satunnaismuuttuja. Lisäksi oletamme tässä, että meillä on testimuuttuja joka on jatkuva (ei siis esim. binomijakauman tapauksia). Mahdollisista approksimaatioista seuraa, että alla oleva tuloskin on approksimatiivinen. Tarkastelemme ensin tapausta missä nollahypoteesi on muotoa $\theta \leq \theta_0$ eli ainoastaan suuret positiiviset testimuuttujan arvot johtavat nollahypoteesin hylkäämiseen. Jos U on testimuuttuja niin p-arvo on

$$P = 1 - F_U(U).$$

Koska $P = 1 - F_U(Y) \leq t$ täsmälleen silloin kun $F_U(U) \geq 1 - t$ eli $U \geq F_U^{-1}(1-t)$ niin P:n kertymäfunktio on

$$\begin{aligned} F_P(t) = \Pr(P \leq t) = \Pr\Big(U \geq F_U^{-1}(1-t)\Big) &= 1 - \Pr\big(U \leq F_U^{-1}(1-t)\big) \\ &= 1 - F_U\Big(F_U^{-1}(1-t)\Big) = 1 - (1-t) = t, \end{aligned}$$

eli kyseessä on tasajakauma.

Mikä on p-arvon jakauma jos nollahypoteesi pätee? jatk.

Sama tulos saadaan jos nollahypoteesi on muotoa $\theta \geq \theta_0$.
Jos nollahypoteesi on muotoa $\theta = \theta_0$ ja testimuuttuja on U niin p-arvo on

$$P = 2\min(F_U(U), 1 - F_U(U)).$$

Nyt $F_u(U) \leq \frac{1}{2}t$ täsmällen silloin kun $U \leq F_U^{-1}(\frac{1}{2}t)$ ja $1 - F_u(U) \leq \frac{1}{2}t$ täsmällen silloin kun $U \geq F_U^{-1}(1 - \frac{1}{2}t)$ joten

$$\begin{aligned} F_P(t) = \Pr(P \leq t) &= \Pr\Big(P \leq t,\, F_U(U) \leq \tfrac{1}{2}\Big) + \Pr\Big(P \leq t,\, F_U(u) > \tfrac{1}{2}\Big) \\ &= \Pr\Big(F_U(U) \leq \tfrac{1}{2}t\Big) + \Pr\Big(1 - F_U(u) \leq \tfrac{1}{2}t\Big) \\ &= \Pr\Big(U \leq F_U^{-1}(\tfrac{1}{2}t)\Big) + \Pr\Big(U \geq F_U^{-1}(1 - \tfrac{1}{2}t)\Big) \\ &= F_U\Big(F_U^{-1}(\tfrac{1}{2}t)\Big) + 1 - F_U\Big(F_U^{-1}(1 - \frac{1}{2}t)\Big) = \tfrac{1}{2}t + 1 - (1 - \tfrac{1}{2}t) = t, \end{aligned}$$

joten tässäkin tapauksessa kyse on tasajakaumasta.

Korrelaatio

Satunnaismuuttujien X ja Y välinen korrelaatiokerroin on

$$\rho_{XY} = \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathrm{E}\Big(\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

ja jos (X_j, Y_j), $j = 1, \ldots, n$ on otos satunnaismuuttujasta (X, Y) niin **otoskorrelaatiokerroin** *on*

$$R_{xy} = \frac{\sum_{j=1}^n (X_j - \overline{X})(Y_j - \overline{Y})}{\sqrt{\sum_{j=1}^n (X_j - \overline{X})^2 \sum_{j=1}^n (Y_j - \overline{Y})^2}} = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}},$$

missä

$$S_{xy} = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})(Y_j - \overline{Y}),$$

ja

$$S_x^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2, \quad S_y^2 = \frac{1}{n-1}\sum_{j=1}^n (Y_j - \overline{Y})^2.$$

Otoskorrelaatokertoimen jakauma

- *Jos (X_j, Y_j), $i = 1, \ldots, n$ on otos normaalijakautuneesta satunnaismuuttujasta (X, Y) jonka korrelaatiokerroin on $\rho_{xy} = 0$ (ja $\sigma_x^2 > 0$, $\sigma_y^2 > 0$) niin pätee*

$$\frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}} \sim \mathrm{t}(n-2).$$

- *Jos (X_j, Y_j), $i = 1, \ldots, n$ on otos normaalijakautuneesta satunnaismuuttujasta (X, Y) ja $-1 < \rho_{xy} < 1$ (ja $\sigma_x^2 > 0$, $\sigma_y^2 > 0$) niin pätee*

$$\tfrac{1}{2}\ln\left(\frac{1+R_{xy}}{1-R_{xy}}\right) \sim_a \mathrm{N}\left(\tfrac{1}{2}\ln\left(\frac{1+\rho_{XY}}{1-\rho_{XY}}\right), \frac{1}{n-3}\right)$$

Pienimmän neliösumman menetelmä kun $y \approx b_0 + b_1 x$

Jos oletetaan, että muuttujien x ja y välinen yhteys voidaan esittää muodossa $y \approx b_0 + b_1 x$, pisteet (x_j, y_j), $j = 1, \ldots, n$ on annettu ja halutaan määrittää b_0 ja b_1 siten, että $\sum_{j=1}^n (y_j - b_0 - b_1 x_j)^2$ on mahdollisimman pieni niin voi olla hyödyllistä ensin laskea keskiarvot $\overline{x} = \frac{1}{n}\sum_{j=1}^n x_j$ ja $\overline{y} = \frac{1}{n}\sum_{j=1}^n y_j$ ja sitten minimoida neliösummaa

$$f(\tilde{b}_0, b_1) = \sum_{j=1}^n \Big((y_j - \overline{y}) - \tilde{b}_0 - b_1 (x_j - \overline{x}) \Big)^2 .$$

Koska $\sum_{j=1}^n (x_j - \overline{x}) = \sum_{j=1}^n (y_j - \overline{y}) = 0$ niin $\frac{\partial}{\partial \tilde{b}_0} f(\tilde{b}_0, b_1) = 2n\tilde{b}_0$ joten optimaalisuusehto $\frac{\partial}{\partial \tilde{b}_0} f(\tilde{b}_0, b_1) = 0$ antaa tulokseksi $\tilde{b}_0 = 0$. Nyt $\frac{\partial}{\partial \tilde{b}_1} f(0, b_1) = -2\sum_{j=1}^n \big((y_j - \overline{y}) - b_1 (x_j - \overline{x}) \big)(x_j - \overline{x})$ ja yhtälöllä $\frac{\partial}{\partial \tilde{b}_1} f(0, b_1) = 0$ on ratkaisu

$$b_1 = \frac{\sum_{j=1}^n (x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^n (x_j - \overline{x})^2} = \frac{s_{xy}}{s_x^2}.$$

Pienimmän neliösumman menetelmä, jatk.

Parametri b_0 lausekkeessa $y = b_0 + b_1 x$ on silloin

$$b_0 = \overline{y} - b_1 \overline{x}.$$

Huom

Edellä olevissa laskuissa ei esiintynyt satunnaismuuttujia, mutta voidaan hyvin ajatella, että muuttujien x ja y välillä vallitsee riippuvuus, joka on muotoa $y = \beta_0 + \beta_1 x$ mutta kun esim mitataan y-muuttujan arvoja esiintyy satunnaisia virheitä, josta seuraa että mitatut arvot ovatkin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n$$

missä ε_j ovat satunnaismuuttujia. Lausekkeen $\sum_{j=1}^n (y_j - b_0 - b_1 x_j)^2$ (eikä jonkin toisen lausekkeen) minimointi on järkevää nimenomaan jos oletetaan, ettei x_j-arvoissa ole virheitä ja kaikki poikkeamat suorasta johtuvat virheellisistä y_j-arvoista.

Esimerkki

Meillä on seuraavat havainnot:

x	1.0	1.9	2.7	3.2	3.8	4.7	5.1	5.5
y	-0.8	-0.4	-0.0	0.9	1.2	1.3	1.7	2.1

Ensin voimme laskea keskiarvot ja ne ovat

$$\overline{x} = 3.4875,$$
$$\overline{y} = 0.75.$$

Sitten meidän pitää laskea x:n otosvarianssi ja muuttujien x ja y otoskovarianssi ja saamme

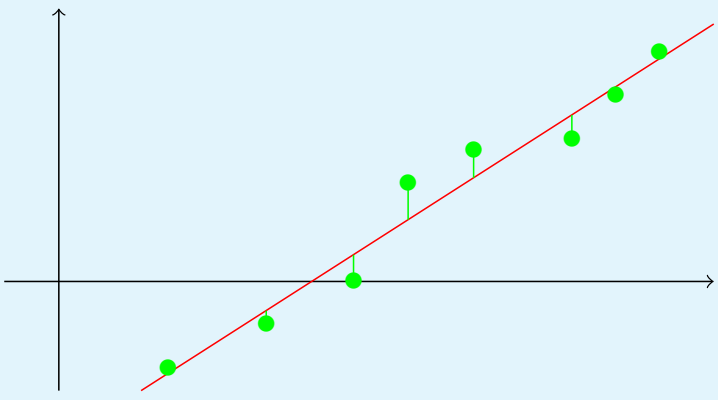
$$s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2 = 2.5184,$$
$$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}) = 1.6121.$$

Esimerkki, jatk.

Näin ollen

$$b_1 = \frac{s_{xy}}{s_x^2} = 0.64015,$$
$$b_0 = \overline{y} - b_1\overline{x} = -1.4825.$$

Pisteet ja suora näyttävät seuraavanlaisilta:



Regressio

- *Oletetaan, että jos x on annettu niin satunnaismuutuja Y on*
$$Y = \beta_0 + \beta_1 x + \varepsilon$$
missä ε on satunnaismuuttuja (eri satunnaismuuttuja eri x:n arvoilla).
- *Otos on tällaisessa tapauksessa muotoa (x_j, Y_j), $j = 1, \ldots, n$ missä $\varepsilon_j = Y_j - \beta_0 - \beta_1 x_j$ ovat riippumattomia satunnaismuuttujia joilla on sama jakauma, joka tavallisesti oletetaan olevan $\mathrm{N}(0, \sigma^2)$.*
- *Pienimmän neliösumman menetelmällä (joka on järkevä täsmälleen silloin kun $\varepsilon \sim \mathrm{N}(0, \sigma^2)$) saadaan seuraavat estimaattorit kertoimille β_1, β_0 ja* **jäännösvarianssille** *σ^2:*
$$B_1 = \frac{S_{xy}}{s_x^2},$$
$$B_0 = \overline{Y} - B_1 \overline{x},$$
$$S^2 = \frac{1}{n-2} \sum_{j=1}^{n} (Y_j - B_0 - B_1 x_j)^2.$$

Regressio, testimuuttujia

- *Oletetaan, että $\varepsilon_j \sim \mathrm{N}(0, \sigma^2)$, $j = 1, \ldots, n$ ovat riippumattomia ja $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$, $j = 1, \ldots, n$. Silloin*
$$B_0 \sim \mathrm{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)\right),$$
$$B_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right),$$
$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2).$$
- *Testimuuttujina voidaan käyttää*
$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)}} \sim \mathrm{t}(n-2),$$
$$W_1 = \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(n-2).$$

Estimaattorien väliset yhteydet

Edellä olevista määritelmistä seuraa myös (ja samat kaavat pätevät estimaateillekin), että

$$S^2 = \frac{n-1}{n-2} S_y^2 (1 - R_{xy}^2),$$

$$R_{xy} = B_1 \sqrt{\frac{S_x^2}{S_y^2}},$$

ja

$$\frac{B_1}{\sqrt{\frac{S^2}{(n-1)S_x^2}}} = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}}.$$

Tästä tuloksesta seuraa myös että nollahypoteesien $\beta_1 = 0$ ja $\rho_{xy} = 0$ testaukset antavat samat tulokset (kun oletuksena on normaalijakauma). Luku r_{xy}^2 eli satunnaismuuttujan R_{xy}^2 havaittu arvo sanotaan olevan regressiomallin **selitysaste**.

Estimaattorien väliset yhteydet, miksi?

Koska $B_1 = \frac{S_{xy}}{s_x^2}$, $B_0 = \overline{Y} - B_1\overline{x}$, $S^2 = \frac{1}{n-2}\sum_{j=1}^n (Y_j - B_0 - B_1 x_j)^2$ *ja* $S_{xy} = R_{xy}\sqrt{s_x^2 S_y^2}$ *niin*

$$(n-2)S^2 = \sum_{j=1}^n \left(B_0 + B_1 x_j - y_j\right)^2 = \sum_{j=1}^n \left(B_1(x_j - \overline{x}) - (y_j - \overline{y})\right)^2$$

$$= B_1^2 \sum_{j=1}^n (x_j - \overline{x})^2 - 2B_1 \sum_{j=1}^n (x_j - \overline{x})(y_j - \overline{y}) + \sum_{j=1}^n (y_j - \overline{y})^2$$

$$= (n-1)\left(B_1^2 S_x^2 - 2B_1 S_{xy} + S_y^2\right) = (n-1)\left(\frac{S_{xy}^2 S_x^2}{S_x^4} - 2\frac{S_{xy}^2}{S_x^2} + S_y^2\right)$$

$$= (n-1)\left(S_y^2 - R_{xy}^2 S_y^2\right) = (n-1)S_y^2(1 - R_{xy}^2),$$

joten

$$S^2 = \frac{n-1}{n-2} S_y^2 (1 - R_{xy}^2).$$

Estimaattorien väliset yhteydet, miksi? jatk.

Koska määritelmän mukaan

$$S_y^2 = \frac{S_{xy}^2}{s_x^2 R_{xy}^2} = \frac{B_1^2 s_x^2}{R_{xy}^2},$$

niin

$$S^2 = \frac{n-1}{n-2}\frac{B_1^2 s_x^2}{R_{xy}^2}(1 - R_{xy}^2).$$

Tästä seuraa, että

$$\frac{B_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} = \frac{B_1}{\sqrt{\frac{(n-1)B_1^2(1-R_{xy}^2)s_x^2}{(n-2)R_{xy}^2(n-1)s_x^2}}} = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}},$$

koska B_1 ja R_{xy} ovat samanmerkkiset.

Liikenne-esimerkki

Tilastokeskuksen mukaan liikenteessä kuolleiden lukumäärät vuosina 2000–2013 olivat

2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
375	379	336	380	344	279	272	292	255	248

Tässä tapauksessa on edullista ottaa x-muuttujaksi vuosiluku josta vähennetään 2014 *jolloin taulukko näyttää tällaiselta:*

x	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
y	375	379	336	380	344	279	272	292	255	248

Tästä otoksesta voimme laskea seuraavat estimaatit:

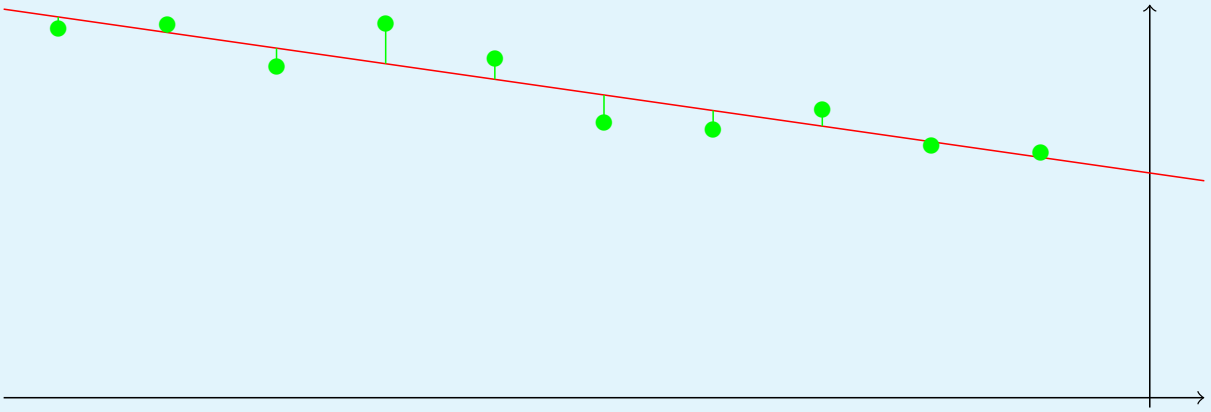
$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	s_{xy}
-5.5	316	9.1667	2772.8889	-145.5556

Liikenne-esimerkki, regressiosuora

Nyt saamme regressiomallin $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ *parametrien estimaateiksi*

$$b_1 = \frac{s_{xy}}{s_x^2} = -15.879,$$
$$b_0 = \overline{y} - b_1 \overline{x} = 228.67,$$
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = -0.91297.$$

Suora ja datapisteet näyttävät seuraavanlaisilta:



Liikenne-esimerkki, β_1:n testaus

Nyt voimme laskea jäännösvarianssin estimaatin joko suoraan alkuperäisestä datasta kaavalla

$$s^2 = \frac{1}{10-2} \sum_{j=1}^{10} (y_j - b_0 - b_1 x_j)^2,$$

mutta yleensä on helpompaa laskea kaavalla

$$s^2 = \frac{n-1}{n-2} s_y^2 (1 - r_{xy}^2) = \frac{9}{8} \cdot 2772.8889 \cdot \left(1 - (-0.91297)^2\right) = 519.35.$$

Nyt voimme testata nollahypoteesia $\beta_1 = 0$ *jolloin testimuuttujana on*

$$W_1 = \frac{B_1 - 0}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(10-2),$$

ja tämä testimuuttuja saa arvon

$$w_1 = \frac{-15.879}{\sqrt{\frac{519.35}{9 \cdot 9.1667}}} = -6.3287.$$

Liikenne-esimerkki, β_1:n testaus, jatk.

Koska nollahypoteesi oli $\beta_1 = 0$ *eikä esimerkiksi* $\beta_1 \geq 0$ *mikä voisi olla hyvin perusteltu, niin p-arvoksi tulee*

$$p = 2F_{\mathrm{t}(8)}(-6.3287) = 0.000226,$$

Liikenne-esimerkki, β_0:n testaus

Jos haluamme testata hypoteesia, $\beta_0 \leq 200$ *niin käytämme testimuuttujana*

$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)}} \sim \mathrm{t}(n-2).$$

Kun sijoitamme edellä lasketut arvot tähän kaavaan, niin saamme testimuuttujan arvoksi

Liikenne-esimerkki, β_0:n testaus, jatk.

$$w_0 = \frac{228.67 - 200}{\sqrt{519.35\left(\frac{1}{10} + \frac{(-5.5)^2}{(10-1)9.1667}\right)}} = 1.8416.$$

Koska nollahypoteesi oli $\beta_0 \leq 200$ *niin ainoastaan testimuuttujan isot positiiviset arvo ovat ristiriidassa nollahypoteesin kanssa, eli sen vaihtoehto on yksisuuntainen, joten p-arvo on*

$$p = 1 - F_{\mathrm{t}(8)}(1.8416) = 0.0514,$$

ja emme hylkää nollahypoteesia edes merkitsevyystasolla 0.05*.*

Liikennne-esimerkki, parametrien luottamusvälit

Parametrien β_0 ja β_1 luottamusvälit määritellään ja lasketaan samalla tavalla kuin normaalijakautuneen satunnaismuuttujan luottamsusväli, eli jos meidän pitää määrittää 99%:n luottamusväli parametrille β_1 niin ensin toteamme, että koska

$$W_1 = \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(n-2)$$

ja $F^{-1}_{\mathrm{t}(8)}(0.995) = -F^{-1}_{\mathrm{t}(8)}(0.005) = 3.3554$ *niin*

$$\Pr\left(-3.3554 \leq \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \leq 3.3554\right) = 1 - 0.005 - 0.005 = 0.99.$$

Koska $-3.3554 \leq \dfrac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \leq 3.3554$ *täsmälleen silloin kun*

Liikennne-esimerkki, parametrien luottamusvälit, jatk.

$$B_1 - 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}} \leq \beta_1 \leq B_1 + 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}} \text{ niin}$$

$$\Pr\left(\beta_1 \in \left[B_1 - 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}}, B_1 + 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}}\right]\right) = 0.99.$$

Kun sijoitamme aikaisemmin lasketut estimaatit tähän, niin saamme 99%:n luottamusväliksi

$$\left[-15.879 - 3.3554\sqrt{\frac{519.35}{9 \cdot 9.1667}}, -15.8791 + 3.3554\sqrt{\frac{519.35}{9 \cdot 9.1667}}\right] = \left[-24.295, -7.4628\right].$$

Normaalijakaumien ehdolliset jakaumat, selitysaste

Jos (X, Y) *on normaalijakautunut niin*

$$(Y|X = x) \sim \mathrm{N}\left(\mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho_{XY}^2)\sigma_Y^2\right),$$

eli

$$\mathrm{E}(Y|X = x) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \beta_0 + \beta_1 x,$$

missä

$$\beta_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2},$$
$$\beta_0 = \mu_Y - \beta_1\mu_X.$$

Pienimmän neliösumman menetelmällä saamme aivan vastaavasti parametrien β_0 *ja* β_1 *estimaateiksi*

$$b_1 = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2},$$
$$b_0 = \overline{y} - b_1\overline{x}.$$

Normaalijakaumien ehdolliset jakaumat, selitysaste, jatk.

Jos (X, Y) *on normaalijakautunut niin voimme siis kirjoittaa*

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

missä

$$\varepsilon \sim \mathrm{N}(0, (1 - \rho_{XY}^2)\sigma_Y^2).$$

Tässä siis jäännösvarianssi $(1 - \rho_{XY}^2)\sigma_Y^2$ *on se osa* Y*:n varianssista, jota ei voida selittää riipuvuudella muuttujasta* x *ja se osuus jolla* Y*:n varianssi pienenee on*

$$\frac{\rho_{XY}^2\sigma_Y^2}{\sigma_Y^2} = \rho_{XY}^2.$$

Analogisesti tämän kanssa sanomme, että mallin $Y_j = b_0 + b_1 x_j$ **selitysaste** *on* r_{xy}^2.

Interpolointi tai ekstrapolointi

Jos on tehty mittauksia ja saatu tulokset (x_j, y_j), $j = 1, \ldots, n$ niin usein halutaan myös tietää mikä olisi y:n arvo jos $x = x_0$. Eräs usein järkevä tapa vastata tähän kysymykseen on olettaa, että $y \approx b_0 + b_1 x$, määrittää b_0 ja b_1 ja sitten laskea $b_0 + b_1 x_0$. Yksinkertainen tapa suorittaa tämä lasku on, että korvataan luvut x_j, $j = 1, \ldots, n$ luvuilla $x_j - x_0$ ja sitten normaaliin tapaan lasketaan b_0:n estimaatti. Tämän lisäksi voidaan testata parametria β_0 koskevia hypoteesejä regressiomallissa $Y = \beta_0 + \beta_1 x + \varepsilon$.

Liikenne-esimerkki, parametrien jakauma

Jos oletetaan, että $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ missä satunnaismuuttujat ε_j ovat riippumattomia ja $\mathrm{N}(0, \sigma^2)$-jakautuneita, niin voidaan osoittaa, että B_0 ja B_1 ovat normaalijakautuneita odotusarvoilla β_0 ja β_1 ja variansseilla $\frac{\sigma^2}{(n-1)s_x^2}$ ja $\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2})$.
Toinen tapa saada käsitys siitä miten luotettavia lasketut estimaatit ovat on seuraava: Oletamme että kaskiulotteisen satunnaismuuttujan (X, Y) jakauma on havaittu empiirinen jakauma, eli

$$\Pr((X, Y) = (x_j, y_j)) = \frac{1}{10}, \quad j = 1, 2, \ldots, 10,$$

missä pisteet (x_j, y_j) tulevat havaitusta otoksesta.
Sitten generoimme tästä jakaumasta otoksia ja laskemme niistä estimaatit jolloin voimme saada käsityksen näiden estimaattien jakaumista. Tässä on kaksi virhelähdettä: Emme tunne alkuperäisen satunnaismuuttujan todellista jakaumaa vaan käytämme approksimaatiota ja emme laske kertymäfunktiota tai tiheysfunktiota tarkasti vaan luotamme simulointiin.

Liikenne-esimerkki, parametrien jakauma, jatk.

Voimme laskea esimerkiksi seuraavalla tavalla:

```
x=(2004:2013)-2014;
y=[375 379 336 380 344 279 272 292 255 248];
n=size(x,2);
m=2000;
b1=zeros(1,m);b0=b1;r=b1;s2=b1;
for j=1:m
  jj=floor(n*rand(1,n)+1); xx=x(jj); yy=y(jj);
  c=cov([x',y']);b1(j)=c(1,2)/var(xx);
  b0(j)=mean(yy)-b1(j)*mean(xx);
  r(j)=b1(j)*sqrt(var(xx)/var(yy));
  s2(j)=((n-1)/(n-2))*var(yy)*(1-r(j)*r(j));
end
```


Liikenne-esimerkki, parametrien jakauma, jatk.

Tällä tavalla simuloidut estimaattien b_0, b_1 s^2 ja r_{xy} jakaumat näyttävät tällaisilta: