Tilastollinen ennustaminen ja otantateoria

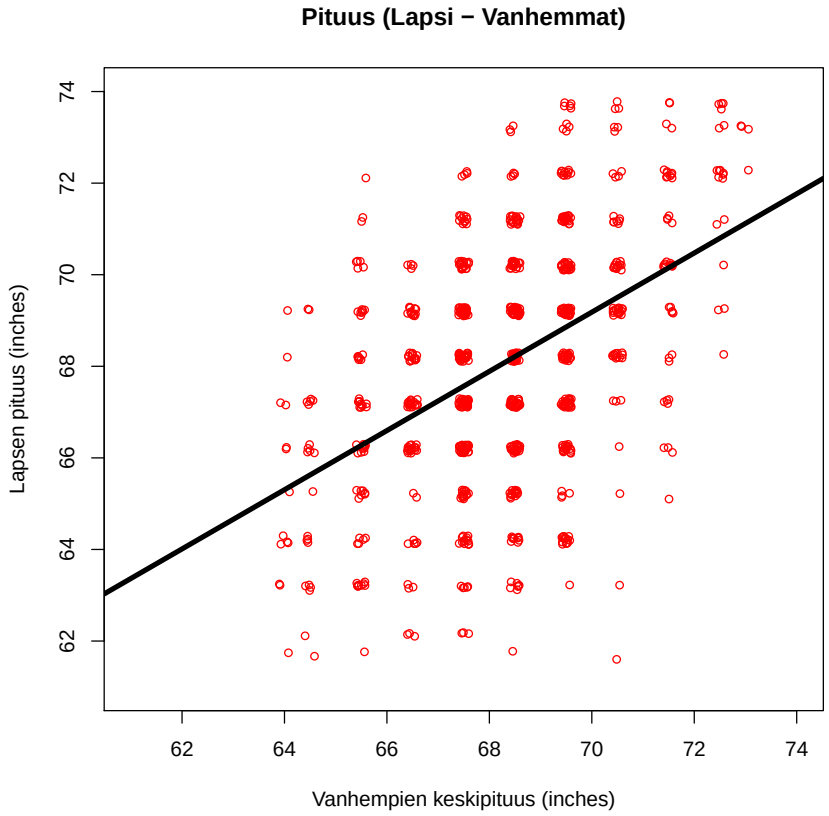
Ennustaminen tutuksi

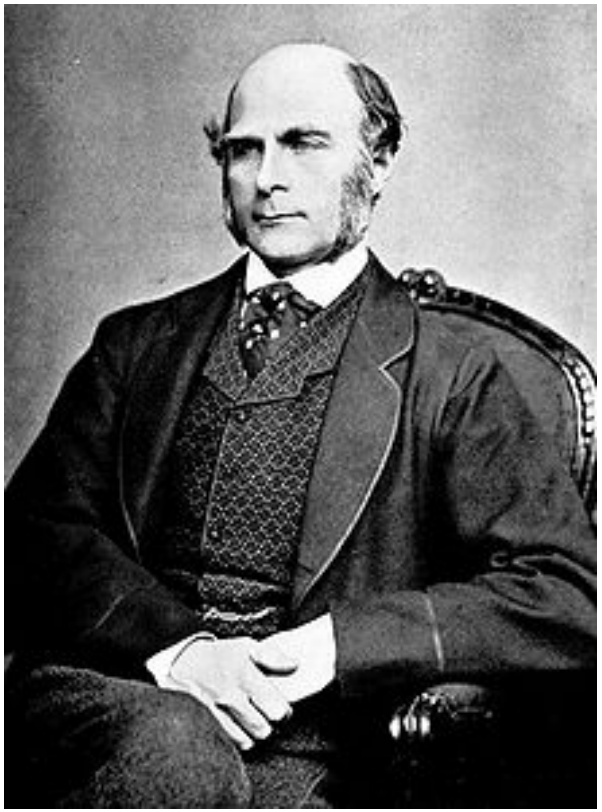
Tilastotiede ja ennustaminen

- Tilastotieteen tehtävänä on kehittää menetelmiä reaalimaailman *satunnaisilmiöiden* kuvaamiseen, selittämiseen ja **ennustamiseen**.
- Tilastollisessa mallinnuksessa pyritään satunnaisilmiöstä aikaisemmin havainnoitujen (toteutuneiden) arvojen $y_1, y_2, \ldots, y_n$ sisältämän informaation avulla luomaan satunnaismuuttujalle Y sellainen tilastollinen malli, jota hyväksikäyttämällä pystytään mahdollisimman hyvin selittämään ja **ennustamaan** satunnaismuuttujan Y (tulevaisuudessa tapahtuvaa) käyttäytymistä.

Tilastotiede tutuksi - Galton ja ennustaminen

– Sir Francis Galton (1822 – 1911) https://en.wikipedia.org/wiki/Francis_Galton

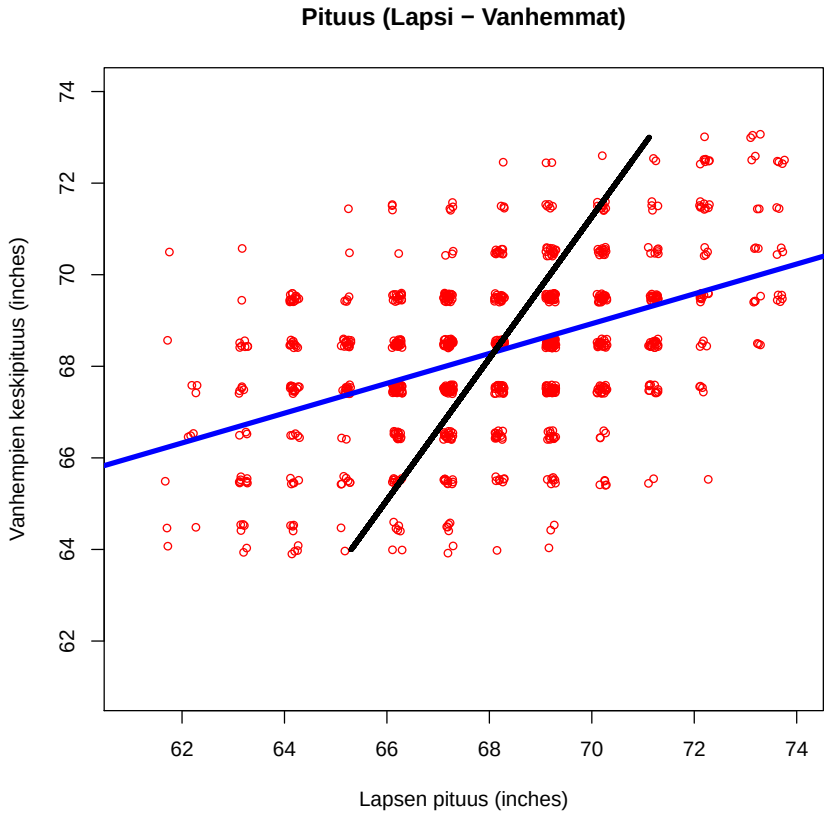


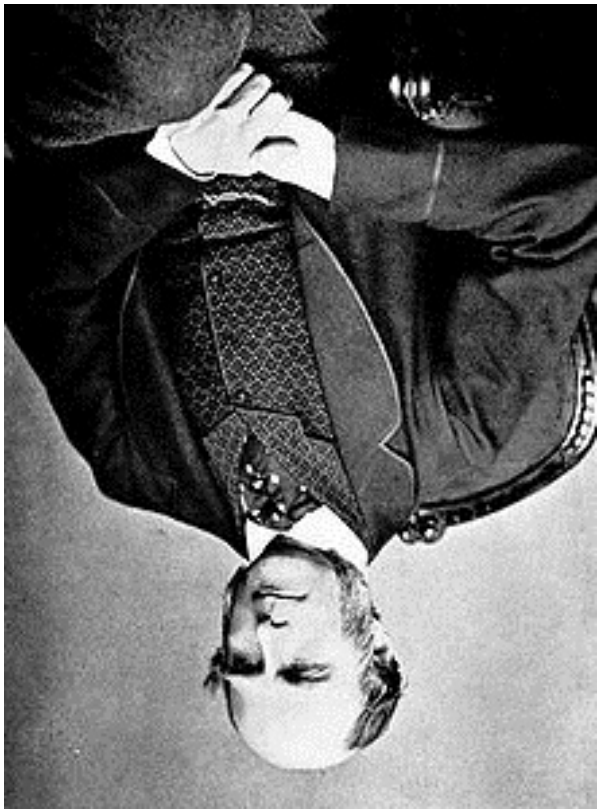


– Kuinka pitkäksi ennustaisit lapsen kasvavan jos vanhempien keskipituuden tiedetään olevan 72.5 tuumaa (184.2 cm)?

Tilastotiede tutuksi - Galton ja ennustaminen - osa 2

– Sir Francis Galton (1822 – 1911) https://en.wikipedia.org/wiki/Francis_Galton

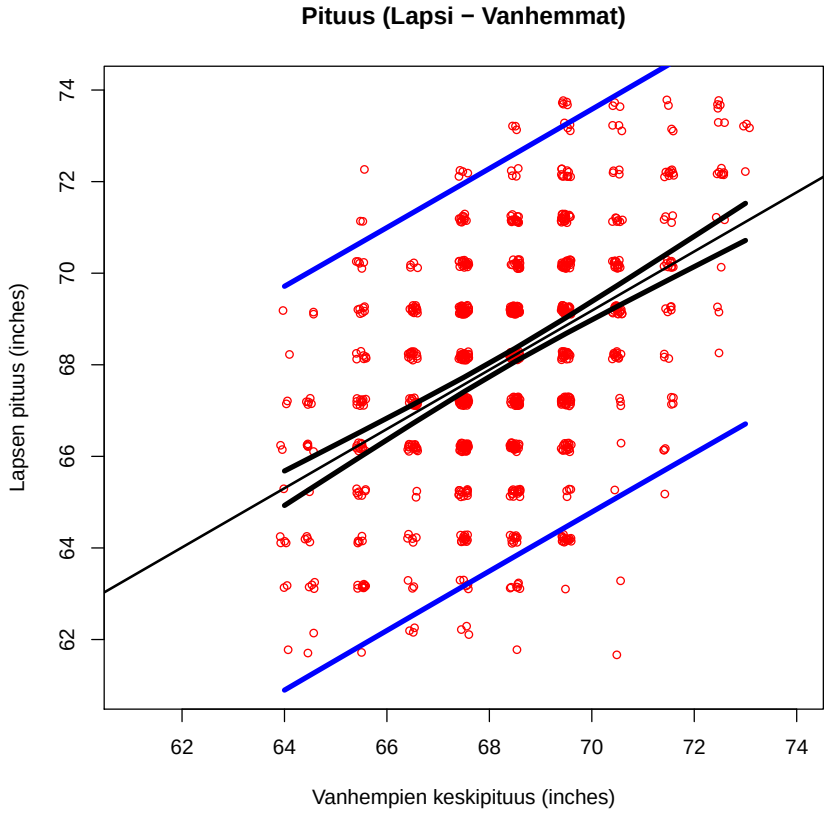


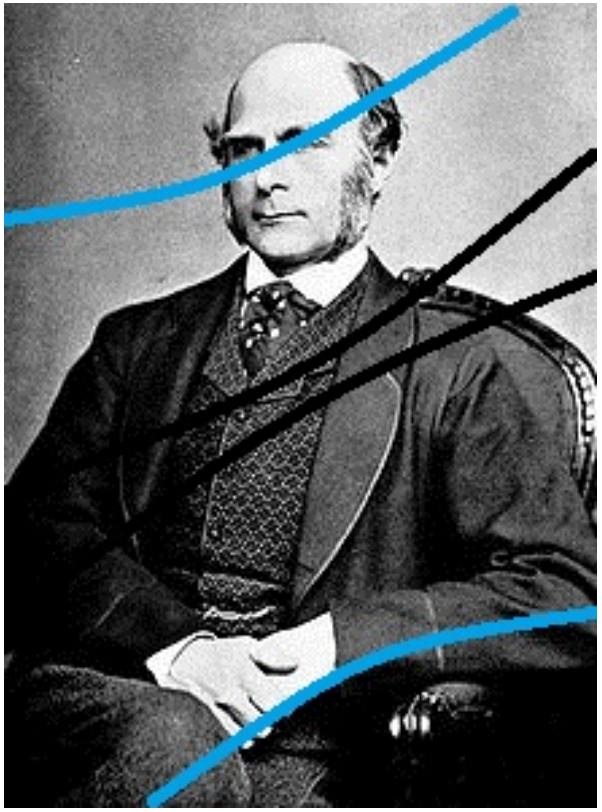


– Kuinka pitkäksi ennustaisit vanhempien keskipituuden olleen jos lapsen pituuden tiedetään olevan 70.8 tuumaa (179.8 cm)?

Tilastotiede tutuksi - Galton ja ennustaminen - osa 3

– Sir Francis Galton (1822 – 1911) https://en.wikipedia.org/wiki/Francis_Galton





– Jos vanhempien keskipituuden tiedetään olevan 72.5 tuumaa (184.2 cm), niin mille välille lapsen pituus 95 % luottamuksella toteutuu?

Tilastotiede tutuksi - Lapio ja äärellisen populaation otanta

– Puutarhalapioita valmistavalla yrityksellä on myyntipisteitä kuudessa eri maakunnassa. Yritys on vuoden alussa toimittanut maakuntien myyntipisteisin myytäväksi puutarhalapioita alla olevan taulukon mukaisesti. Lisäksi yritys on asettanut lapioille eri maakunnissa eri hinnan arvioimansa kilpailu- ja markkinatilanteen perusteella. Neljästä maakunnasta yritys on saanut tiedon, kuinka moni sen toimittamista lapioista on vuoden alusta lähtien myyty. Pirkanmaan ja Kanta-Hämeen osalta yritys ei kuitenkaan ole saanut tietoa myytyjen lapioiden lukumäärästä.

	Alue					
	Uusimaa	Varsinais-Suomi	Pohjanmaa	Satakunta	Pirkanmaa	Kanta-Häme
Toimitettu	120	75	50	35	75	35
Myyty	43	40	39	24	Avoin!	Avoin!
Hinta X	45	35	30	25	35	25

– Oletetaan, että todennäköisyys että yksittäinen puutarhalapio i tulee myydyksi
Y_i=myyty/ei myyty riippuu lapiolle asetetusta hinnasta x_i logistisen regressiomallin mukaan

$$\mathrm{E}(Y_i) = \theta(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}},$$
$$\mathrm{Var}(Y_i) = \theta(x_i)(1-\theta(x_i)).$$

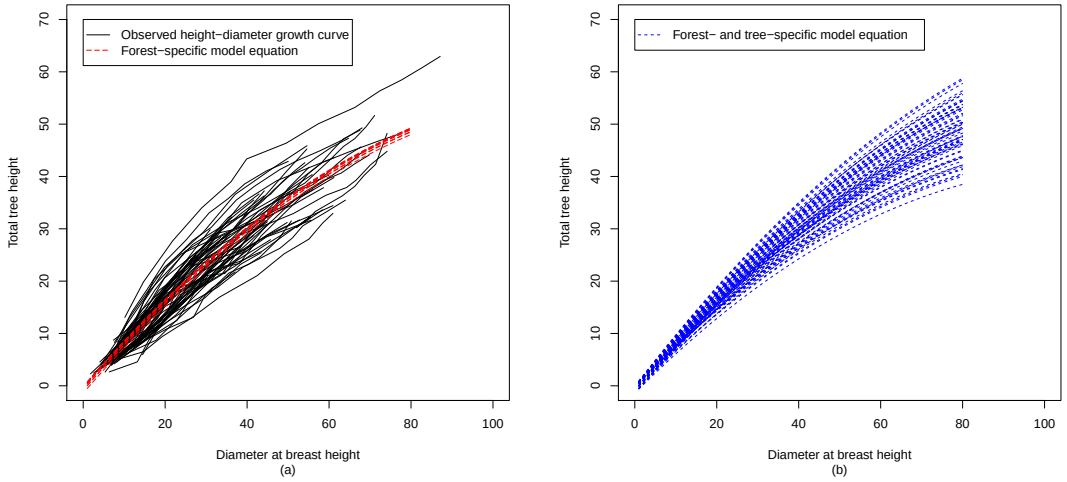
– Ennusta, kuinka monta puutarhalapiota yritys on myynnyt alkuvuoden aikana!

Tilastotiede tutuksi - Metsä ja ennustemallit

– In height-diameter relationship models, the height of the tree is the response variable with the expected value assumed to depend on the value of the diameter:

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \beta_2 x_{ijk}^2 + \gamma_{1j} + \gamma_{2j} x_{ijk} + \gamma_{3i} + \varepsilon_{ijk},$$

y_{ijk} is the height measured at time k for the j:th tree in the i:th forest,
x_{ijk} is the diameter measured at time k for the j:th tree in the i:th forest,
β_0, β_1 and β_2 are fixed unknown parameters,
γ_{1j} and γ_{2j} are tree-specific unknown random intercepts and slopes, respectively,
γ_{3i} are forest-specific unknown random intercepts, and, ε_{ijk} are unknown random error terms.



– Ennusta, mitä on puun j korkeus y_{ijk} ajanhetkellä k läpimitan x_{ijk} tilanteessa?

Merkinnät ennustamisessa

– Oletetaan, että satunnaismuuttuja Y_i noudattaa jakaumaa $Y_i \sim f_{Y_i}(y_i; \theta_1, \theta_2, \ldots, \theta_k)$, missä parametrien voidaan olettaa riippuvan selittävistä muuttujista funktioiden $g_1, g_2, \ldots, g_k$ kautta tyyliin

$$\begin{aligned}
\theta_1 &= g_1(x_{i1}, x_{i2}, \ldots, x_{ip}; \beta_{0_1}, \beta_{1_1}, \beta_{2_1}, \ldots, \beta_{r_1}),\\
\theta_2 &= g_2(x_{i1}, x_{i2}, \ldots, x_{ip}; \beta_{0_2}, \beta_{1_2}, \beta_{2_2}, \ldots, \beta_{r_2}),\\
&\vdots\\
\theta_k &= g_k(x_{i1}, x_{i2}, \ldots, x_{ip}; \beta_{0_k}, \beta_{1_k}, \beta_{2_k}, \ldots, \beta_{r_k}).
\end{aligned}$$

– Olkoon aineiston $y_1, y_2, \ldots, y_n$ toteutuneita arvoja satunnaismuuttujista $Y_1, Y_2, \ldots, Y_n$. Merkitään havaintoyksiköön i_* liittyvää vielä toteutumatonta satunnaismuuttujaa Y_{i_*}:llä.

– **Ennustusongelma:**

Mikä on sellainen funktio $h(y_1, y_2, \ldots, y_n; \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$, joka on "paras mahdollinen ennuste" satunnaismuuttujan Y_{i_*} (joskus tulevaisuudessa toteutuvalle) tuntemattomalle arvolle y_{i_*}?

Ennusteet järjestykseen

Paras ennuste ja paras lineaarinen ennuste:

– Jos yhteistiheysfunktio

$$Y_1, Y_2, \ldots, Y_n, Y_{i_*} \sim f_{Y_i}(y_1, y_2, \ldots, y_n, y_{i_*}; \theta_1, \theta_2, \ldots, \theta_k),$$

on täysin tiedossa, niin silloin tuntemattoman arvon y_{i_*} **paras ennuste** (engl. Best Predictor, BP) on ehdollinen odotusarvo

$$\text{BP}(y_{i_*}) = \text{E}(Y_{i_*}|y_1, y_2, \ldots, y_n). \tag{1}$$

– Jos satunnaismuuttujien $Y_1, Y_2, \ldots, Y_n, Y_{i_*} = \mathbf{y}, Y_{i_*}$, odotusarvot $\boldsymbol{\mu}$ ja kovarianssirakenne

$$\text{E}\begin{pmatrix}\mathbf{y}\\ Y_{i_*}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \mu_{i_*}\end{pmatrix}, \quad \text{Cov}\begin{pmatrix}\mathbf{y}\\ Y_{i_*}\end{pmatrix} = \begin{pmatrix}\mathbf{V} & \mathbf{w}\\ \mathbf{w}' & v_{i*}\end{pmatrix}$$

täysin tiedossa, niin silloin tuntemattoman arvon y_{i_*} **paras lineaarinen ennuste** (engl. Best Linear Predictor, BLP) on ehdollinen lineaarinen odotusarvo

$$\text{BLP}(y_{i_*}) = \mu_{i_*} + \mathbf{w}'\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}). \tag{2}$$

Parhaat käytännön ennusteet:

– Jos yhteistiheysfunktio

$$Y_1, Y_2, \ldots, Y_n, Y_{i_*} \sim f_{Y_i}(y_1, y_2, \ldots, y_n, y_{i_*}; \theta_1, \theta_2, \ldots, \theta_k),$$

rakenne on tiedossa, mutta parametrit $\theta_1, \theta_2, \ldots, \theta_k$ tuntemattomia niin silloin tuntemattoman arvon y_{i_*} **suurimman uskottavuuden ennuste** (engl. Maximum Likelihood Predictor, MLP) on sellainen arvo $\hat{y}_{i_*}$, joka maksimoi yhteistiheysfunktion

$$\hat{y}_{i_*} = \text{MLP}(y_{i_*}) = \max_{\theta_1, \theta_2, \ldots, \theta_k, y_{i_*}} f_{Y_i}(y_1, y_2, \ldots, y_n, y_{i_*}; \theta_1, \theta_2, \ldots, \theta_k). \tag{3}$$

– Jos satunnaismuuttujien $Y_1, Y_2, \ldots, Y_n, Y_{i_*} = \mathbf{y}, Y_{i_*}$, odotusarvorakenne on lineaarinen ja kovarianssirakenne tunnettu

$$\text{E}\begin{pmatrix}\mathbf{y}\\ Y_{i_*}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \beta_0 + \beta_1 x_{i_*1} + \beta_2 x_{i_*2} + \cdots + \beta_r x_{i_*r}\end{pmatrix}, \quad \text{Cov}\begin{pmatrix}\mathbf{y}\\ Y_{i_*}\end{pmatrix} = \begin{pmatrix}\mathbf{V} & \mathbf{w}\\ \mathbf{w}' & v_{i*}\end{pmatrix}$$

niin silloin tuntemattoman arvon y_{i_*} **paras lineaarinen harhaton ennuste** (engl. Best Linear Unbiased Predictor, BLUP) on estimoitu ehdollinen lineaarinen odotusarvo

$$\text{BLUP}(y_{i_*}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i_*1} + \hat{\beta}_2 x_{i_*2} + \cdots + \hat{\beta}_r x_{i_*r} + \mathbf{w}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{4}$$

Ennustaminen ja äärellisen populaation otantateoria

Kokonaissumman ennusteongelma:

– Äärellisen populaation tilanteessa ajatellaan, että populaation jokaiselta havaintoyksiköltä i, ($i = 1, 2, \ldots, N$), on (ainakin teoriassa) mitattavissa tai havaittavissa muuttujan Y arvo Y_i:

$$\Omega = \{Y_1, Y_2, \ldots, Y_{N-1}, Y_N\}.$$

Mitä on kokonaissumman $Y_{S_\Omega} = Y_1 + Y_2 + \cdots + Y_{N-1} + Y_N$ tulevaisuudessa toteutuva arvo?

– Sisältäköön otos toteutuneita arvoja y_i "ensimmäiset" n kappaletta satunnaismuuttujista Y_i. Kokonaissumma tällöin jakautuu kahteen osaan

$$\begin{aligned} Y_{S_\Omega} &= y_1 + y_2 + \cdots + y_{n-1} + y_n + Y_{n+1} + Y_{n+2} + \cdots + Y_{N-1} + Y_N \\ &= \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} Y_i = Y_{S_\Delta} + Y_{S_\Psi}. \end{aligned} \tag{5}$$

– Kokonaissumman Y_{S_Ω} ennustaminen vastaa otoksesta poisjäänneiden havaintojen summan Y_{S_Ψ} ennustamista.

Lapioiden kokonaissumman ennustaminen

	Alue					
	Uusimaa	Varsinais-Suomi	Pohjanmaa	Satakunta	Pirkanmaa	Kanta-Häme
Toimitettu	120	75	50	35	75	35
Myyty	43	40	39	24	Avoin!	Avoin!
Hinta X	45	35	30	25	35	25

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	hinta	-,111	,018	36,626	1	,000	,895
	Constant	4,345	,708	37,620	1	,000	77,086

a. Variable(s) entered on step 1: hinta.

$$\theta(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}},$$

$$\hat{y}_{i_*} = \hat{\theta}(x_{i_*} = 35) = \frac{\exp(4.345 + (-0.111 * 35))}{1 + \exp(4.345 + (-0.111 * 35))} = 0.613,$$

$$\hat{y}_{i_*} = \hat{\theta}(x_{i_*} = 25) = \frac{\exp(4.345 + (-0.111 * 25))}{1 + \exp(4.345 + (-0.111 * 25))} = 0.827,$$

$$\hat{Y}_{S_\Omega} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}$$

$$= 43 + 40 + 39 + 24 + 75 * 0.613 + 35 * 0.827 = 220.92 \approx 221.$$

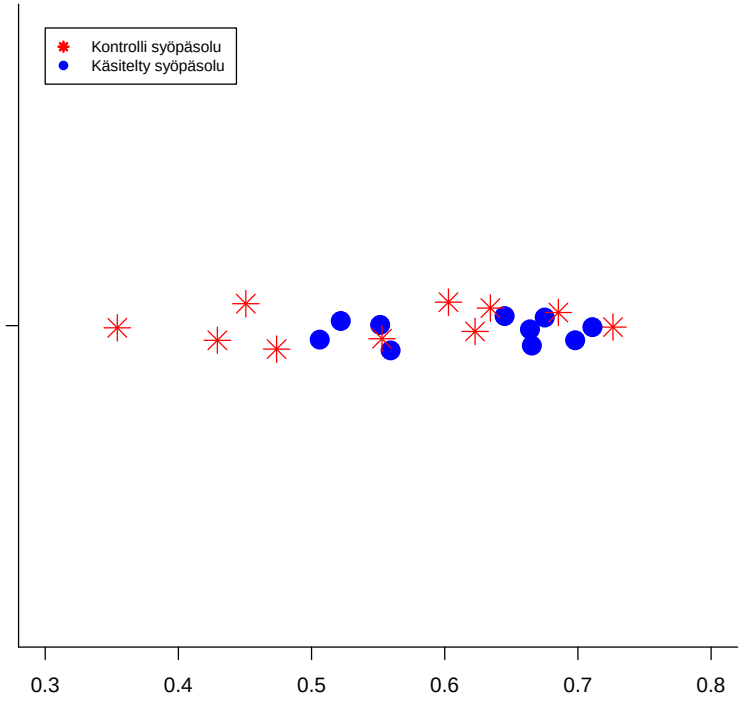
Ennustaminen tilastollisessa päättelyssä

Esimerkkiaineisto: 10 havaintoa luokittain

- Tarkastellaan, kuinka eturauhassyöpäsolut reagoivat tilanteessa missä niitä käsitellään käsittelyillä "trichostatin A" ja "demethylating agent 5-aza-2-deoxycytidine".
- Erityisesti tarkastellaan, kuinka hsa-miR-31 nimisen mikroRNA:n ekspressioarvo muuttuu käsittelyiden vaikutuksesta.
- Tutkimuksen alussa käytössä oli 10 eturauhassyöpäsoluja sisältävää soluviljelmää (kontrollit) ja 10 eturauhassyöpäsoluja sisältävää soluviljelmää, joille oltiin tehty käsittelyt "trichostatin A" ja "demethylating agent 5-aza-2-deoxycytidine".
- Tutkimusongelmana on testata, onko syöpäsolujen käsittelyillä vaikutusta hsa-miR-31 nimisen mikroRNA:n ekspressioarvoihin kontrolli syöpäsolujen arvoihin verrattuna. Tätä voidaan testata kahden riippumattoman otoksen t-testillä

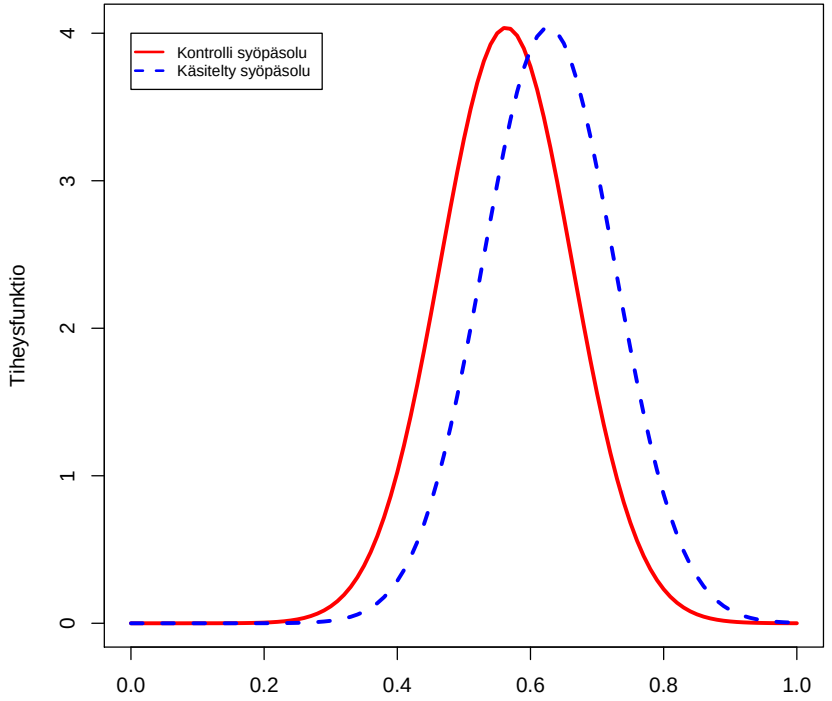
hsa-miR-31	**Ekspressioarvot**				
A=**Kontrolli syöpäsolu**	0.3541437	0.4292698	0.4506302	0.4737521	0.5528881
	0.6027864	0.6227665	0.6343017	0.6853181	0.7263142
B=**Käsitelty syöpäsolu**	0.5061544	0.5219287	0.5515087	0.5593351	0.6449594
	0.6640027	0.6654243	0.6750765	0.6979383	0.7108782

Havaitut ekspressioarvot: 10 havaintoa



$\hat{\mu}_A = 0.5532171$ ja $\hat{\mu}_B = 0.6197206$

Estimoidut tiheysfunktiot: 10 havaintoa



Kliininen ero: $\hat{\mu}_B - \hat{\mu}_A = 0.06650354$

Kahden otoksen t-testi

– Klassisesti t-testissä oletetaan, että ekspressioarvot noudattavat normaalijakaumaa.

– Merkitään kontrollisolujen ekspressioarvoja satunnaismuuttuja y_A:lla ja käsiteltyjen solujen ekspressioarvoja satunnaismuuttuja y_B:lla.

– Satunnaismuuttujien y_A ja y_B oletetaan noudattavat normaalijakaumia

$$y_A \sim N(\mu_A, \sigma^2), \qquad y_B \sim N(\mu_B, \sigma^2), \tag{6}$$

missä μ_A ja μ_B ovat tuntemattomia odotusarvoja, sekä σ^2 on yhteinen tuntematon varianssiparametri.

– Tutkimusongelma voidaan nähdä seuraavien hypoteesien testausongelmana:

$$H_0 : \mu_A = \mu_B, \tag{7a}$$
$$H_1 : \mu_A \neq \mu_B. \tag{7b}$$

– Edellä olevia hypoteeseja voidaan testata kahden riippumattoman otoksen t-testillä

$$t = \frac{\hat{\mu}_B - \hat{\mu}_A}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}, \tag{8}$$

missä

$$\hat{\mu}_A = \bar{y}_A = \frac{\sum_{i=1_A}^{n_A} y_i}{n_A}, \tag{9a}$$

$$\hat{\mu}_B = \bar{y}_B = \frac{\sum_{i=1_B}^{n_B} y_i}{n_B}, \tag{9b}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1_A}^{n_A} (y_i - \hat{\mu}_A)^2 + \sum_{j=1_B}^{n_B} (y_j - \hat{\mu}_B)^2}{n_A + n_B - 2}. \tag{9c}$$

– Testisuure t noudattaa Studentin t-jakaumaa vapausastein $n_A + n_B - 2$ kun H_0 hypoteesi on tosi.

– Testin havaittu p-arvo (merkitsevyystaso) on todennäköisyys

$$p_{hav} = 2 \times P(t > |t_{hav}|). \tag{10}$$

t-testin tulokset - 10 havaintoa

– Aineistosta saadaan nyt laskettua seuraavia otostunnuslukuja ja t-testin tuloksia.

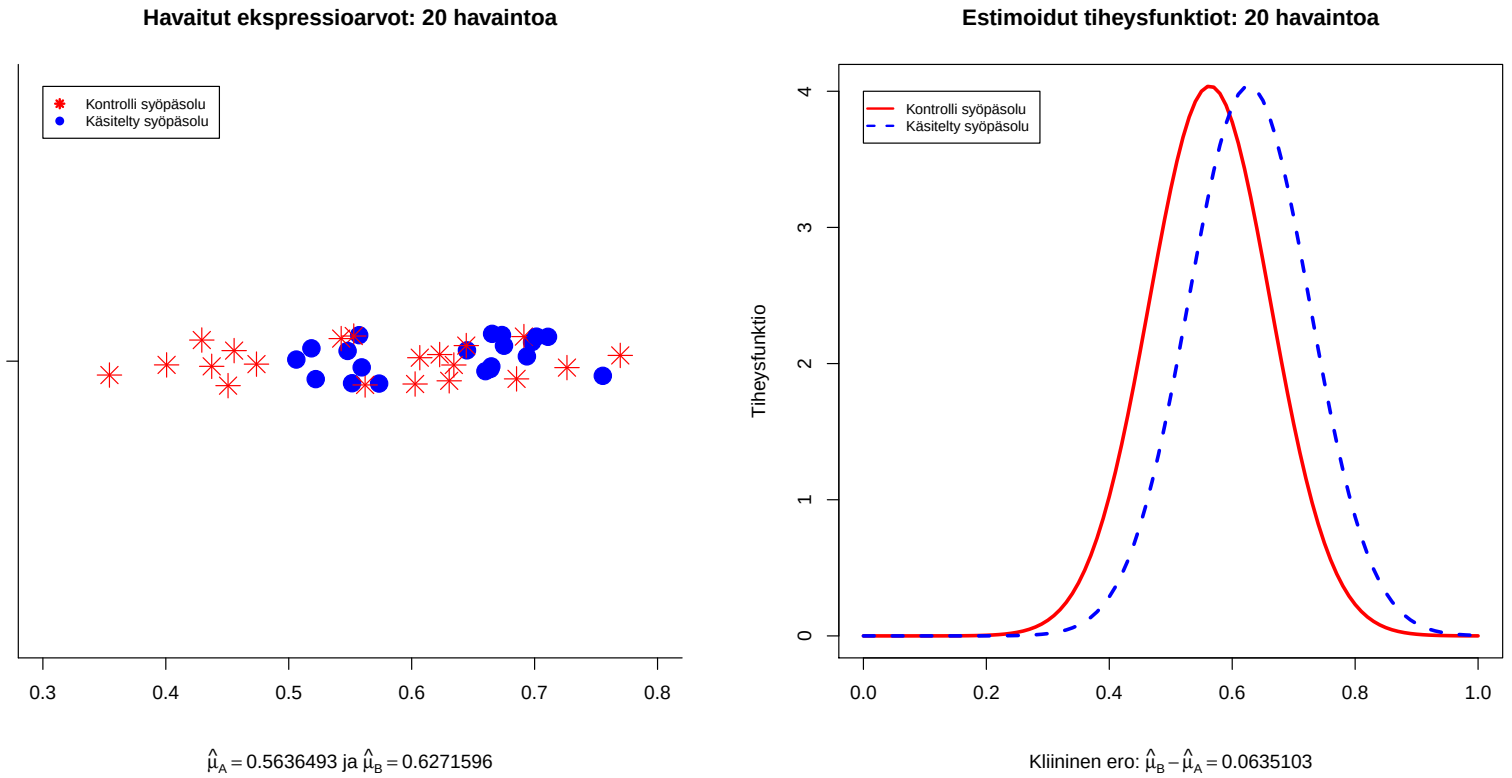
10 havaintoa - hsa-miR-31	**Keskiarvo**	**Varianssi**	t**-arvo**	p**-arvo**
A=**Kontrolli syöpäsolu**	0.5532171	0.01479047		
B=**Käsitelty syöpäsolu**	0.6197206	0.00588205		
Erotus	0.0665035	0.01033626	1.4627	0.1608

– Taulukon perusteella voidaan tehdä päättely, että yllä oleva H_0 hypoteesi jää voimaan kun päättely tehdään 5 % riskitasolla ja siten kontrolli eturauhassyöpäsolujen ja käsiteltyjen eturauhassyöpäsolujen keskimääräisissä ekspressiotasoissa ei ole t-testin perusteella eroa hsa-miR-31 mikroRNA:n suhteen.

Esimerkkiaineisto: 20 havaintoa luokittain

– Tutkimuksen edetessä aineistoa pystyttiin laajentamaan siten, että hsa-miR-31 mikroRNA:sta saatiin 10 uutta ekspressioarvoa kontrolli syöpäsoluviljelmistä ja 10 uutta ekspressioarvoa syöpäsoluviljelmistä, joille oltiin tehty käsittelyt "trichostatin A" ja "demethylating agent 5-aza-2-deoxycytidine".

hsa-miR-31	**Ekspressioarvot**				
A=**Kontrolli syöpäsolu**	0.3541437	0.4006350	0.4292698	0.4373371	0.4506302
	0.4555651	0.4737521	0.5426510	0.5528881	0.5621282
	0.6027864	0.6066067	0.6227665	0.6305202	0.6343017
	0.6444047	0.6853181	0.6913047	0.7263142	0.7696627
B=**Käsitelty syöpäsolu**	0.5061544	0.5184477	0.5219287	0.5478595	0.5515087
	0.5572843	0.5593351	0.5735056	0.6449594	0.6600103
	0.6640027	0.6647708	0.6654243	0.6734123	0.6750765
	0.6937213	0.6979383	0.7014992	0.7108782	0.7554756



– Laajennetun aineiston tilanteessa voidaan edelleen testata hypoteeseja

$$H_0 : \mu_A = \mu_B, \tag{11a}$$
$$H_1 : \mu_A \neq \mu_B, \tag{11b}$$

kahden riippumattoman otoksen t-testin avulla.

Tulosten vertailua havaintojen lukumäärän suhteen

– Laajennetusta aineistosta saadaan laskettua seuraavia t-testin tuloksia.

20 havaintoa - hsa-miR-31	**Keskiarvo**	**Varianssi**	t**-arvo**	p**-arvo**
A=**Kontrolli syöpäsolu**	0.5636493	0.01370696		
B=**Käsitelty syöpäsolu**	0.6271596	0.005811309		
Erotus	0.0635103	0.009759137	2.033	0.04908

– Voidaan tehdä päättely, että H_0 hypoteesi hylätään 5 % riskitasolla.

– Vertailtaessa 10 havainnon ja 20 havainnon aineistojen tuloksia saadaan seuraava taulukko.

hsa-miR-31	**Erotus**	**Varianssi**	t**-arvo**	p**-arvo**
10 havaintoa	0.0665035	0.01033626	1.4627	0.1608
20 havaintoa	0.0635103	0.009759137	2.033	0.04908

– Suurin tekijä päättelyn eroavuuteen on otoskoon kasvaminen.

Otoskoon vaikutus päättelyyn ja tilastollinen ennustaminen

– Otoskoiden n_A ja n_B suuruudella on merkittävä vaikutus tehtävään päättelyyn testattaessa odotusarvojen μ_A ja μ_B yhtäsuuruutta kahden riippumattoman otoksen t-testin avulla.

– Kun on voimassa $\epsilon = |\mu_B - \mu_A| > 0$, niin on olemassa otoskoot n_A ja n_B, että t-testin perusteella H_0 hypoteesi hylätään valitulla riskitasolla α ehdolla, että testin tehokkuus on vähintään β.

Teoreema 1. Kaikille $\epsilon = |\mu_B - \mu_A| > 0$ on olemassa ϵ:n arvosta riippuvat otoskoot n_{A_ϵ} ja n_{B_ϵ} siten, että otossuureelle

$$t_\epsilon = \frac{\hat{\mu}_{B_\epsilon} - \hat{\mu}_{A_\epsilon}}{\sqrt{\hat{\sigma}^2_\epsilon \left(\frac{1}{n_{A_\epsilon}} + \frac{1}{n_{B_\epsilon}}\right)}}, \tag{12}$$

on voimassa todennäköisyys

$$P(t_\epsilon > q_{\alpha/2}) = \beta, \tag{13}$$

missä $q_{\alpha/2}$ on Studentin t-jakauman $100(1-\alpha/2)$ prosentin kvantiili vapausasteiden ollessa $df = n_{A_\epsilon} + n_{B_\epsilon} - 2$.

- Edellä olevasta teoreemasta seuraa ainakin kaksi huomattavaa käytännön ongelmaa.
  1. H_0 hypoteesi voidaan hylätä otoskoon vaikutuksesta tilanteessa, jossa kliininen ero $\epsilon = |\mu_B - \mu_A|$ on merkityksettömän pieni satunnaisilmiön kokonaisvaihtelun suhteen.
  2. Suurten otoskoiden tilanteessa (big data) hypoteesin testaus päättelyn välineenä on suhteellisen rajallinen, koska kaikista eroista tulee tilastollisesti merkitseviä.
- Tilastollinen ennustaminen ja erityisesti tilastollinen luottamusväliennustaminen mahdollistaa t-testin tyylisen kahden otoksen vertailun ilman, että otoskoolla voidaan kokonaan vaikuttaa tehtävään päättelyyn.
- Lisäksi tilastollinen ennustamisen kautta saadaan sellainen otoskoosta riippumaton minimi kliininen ero $\epsilon = |\mu_B - \mu_A|$, mitä pienemmillä eroilla ei voida kahden eri populaatio arvojen katsoa eroavan merkitsevästi toisistaan.

Hypoteesin testauksesta luottamusvälin kautta ennustepäättelyyn

– Tilastollinen hypoteesin testaus on yleisesti yhteydessä parametrien luottamusväliestimoimiseen.

– Testattaessa hypoteeseja

$$H_0 : \mu_A = \mu_B, \tag{14a}$$
$$H_1 : \mu_A \neq \mu_B, \tag{14b}$$

kahden riippumattoman otoksen t-testin avulla, H_0 hypoteesi hylätään 100α prosentin riskitasolla jos ja vain jos odotusarvojen erotuksen $\mu_B - \mu_A$ $100(1-\alpha)$ prosentin t-tyylinen luottamusväliestimaatti

$$\left[\hat{\mu}_B - \hat{\mu}_A - q_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}, \hat{\mu}_B - \hat{\mu}_A + q_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}\right] \tag{15}$$

ei sisällä arvoa nolla. Mikäli yllä oleva luottamusväliestimaatti sisältää arvon nolla, H_0 hypoteesi jää voimaan.

Satunnaismuuttujien erotuksen tarkastelu

- Kahden riippumattoman otoksen t-testissä ja yllä olevassa luottamusväliestimaatissa tarkastellaan odotusarvojen erotuksen $\mu_B - \mu_A$ eroavuutta nollasta.
- Todellisuudessa olisi hyödyllisempää vertailla satunnaismuuttujien erotuksen $y_B - y_A$ suuruutta.
- Mikäli satunnaismuuttujien erotus $y_B - y_A$ saa lähtökohtaisesti nollasta poikkeavia arvoja, tarkoittaa se sitä, että satunnaismuuttujien y_A ja y_B jakaumien sijainnissa on kokonaisuudessaan eroa.
- Esimerkkiaineiston tilanteessa jakaumien sijaintien ero tarkoittaisi, että käsittelyillä olisi vaikutusta syöpäsolujen ekspressioarvoihin.
- Satunnaismuuttujien y_A ja y_B erotuksen $y_B - y_A$ suuruutta puolestaan voidaan arvioida tilastollisen ennustamisen ja erityisen tilastollisen luottamusväliennustamisen avulla.

BLUP ennusteen ominaisuuksia

– Erotuksen $y_B - y_A$ paras lineaarinen harhaton ennuste on

$$\hat{y}_B - \hat{y}_A = \text{BLUP}(y_B - y_A) = \hat{\mu}_B - \hat{\mu}_A = \bar{y}_B - \bar{y}_A. \tag{16}$$

– Ennustevirhe $e = (y_B - y_A) - (\hat{y}_B - \hat{y}_A)$ noudattaa normaalijakaumaa

$$(y_B - y_A) - (\hat{y}_B - \hat{y}_A) \sim N\left(0, \sigma^2\left(2 + \frac{1}{n_A} + \frac{1}{n_B}\right)\right). \tag{17}$$

– Kun varianssi σ^2 korvataan estimaatilla $\hat{\sigma}^2$, niin suhde

$$t = \frac{(y_B - y_A) - (\hat{y}_B - \hat{y}_A)}{\sqrt{\hat{\sigma}^2\left(2 + \frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{18}$$

noudattaa Studentin t-jakaumaa vapausastein $df = n_A + n_B - 2$.

Tilastollinen luottamusväliennuste

– Kaavaa (18) voidaan käyttää pivotaalina, jonka avulla saadaan satunnaismuuttujien erotukselle $y_B - y_A$ muodostettua $100(1-\alpha)$ prosentin luottamusväliennuste

$$\left[\hat{\mu}_B - \hat{\mu}_A - q_{\alpha/2}\sqrt{\hat{\sigma}^2\left(2+\frac{1}{n_A}+\frac{1}{n_B}\right)}, \hat{\mu}_B - \hat{\mu}_A + q_{\alpha/2}\sqrt{\hat{\sigma}^2\left(2+\frac{1}{n_A}+\frac{1}{n_B}\right)}\right] \tag{19}$$

– Mikäli luottamusväliennuste sisältää arvon nolla, satunnaismuuttujat y_A ja y_B voivat saada yhtä suuria arvoja ja tehdään päättely, että satunnaismuuttujien y_A ja y_B jakaumat sijaitsevat kokonaisuudessaan lähekkäin.

– Jos taas luottamusväliennuste ei sisällä arvoa nolla, satunnaismuuttujat y_A ja y_B eivät $100(1-\alpha)$ luottamuksella saa yhtä suuria arvoja, eli satunnaismuuttujien y_A ja y_B jakaumat sijaitsevat kokonaisuudessaan toisistaan erillään.

Kliinisen eron suuruus

– Asymptoottisesti ennustevirheen varianssille on voimassa

$$\lim_{n_A, n_B \to \infty} \mathrm{Var}(e) = 2\sigma^2. \tag{20}$$

– Erotuksen $y_B - y_A$ asymptoottinen luottamusväliestimaatti on muotoa

$$\left[\hat{\mu}_B - \hat{\mu}_A - z_{\alpha/2}\sqrt{2\hat{\sigma}^2}, \hat{\mu}_B - \hat{\mu}_A + z_{\alpha/2}\sqrt{2\hat{\sigma}^2}\right], \tag{21}$$

missä $z_{\alpha/2}$ on standardoidun normaalijakauman $100(1-\alpha/2)$ prosentin kvantiili.

– Asymptoottisesti kliiniselle erolle $\epsilon = |\mu_B - \mu_A|$ tulee olla voimassa

$$\epsilon = |\mu_B - \mu_A| > z_{\alpha/2}\sqrt{2\sigma^2} \tag{22}$$

ennen kuin satunnaismuuttujien y_A ja y_B jakaumat sijaitsevat merkitsevästi toisistaan erillään.

Ennustepäättely ekspressioaineistossa

– Tarkastellaan luottamusväliennustamisen avulla, onko käsittelyillä vaikutusta syöpäsolujen ekspressioarvojen jakauman sijaintiin hsa-miR-31 mikroRNA:n suhteen.

– 95 % luottamusväliennusteeksi $y_B - y_A$:lle saadaan

$$\left[\hat{\mu}_B - \hat{\mu}_A - q_{\alpha/2}\sqrt{\hat{\sigma}^2\left(2 + \frac{1}{n_A} + \frac{1}{n_B}\right)}, \hat{\mu}_B - \hat{\mu}_A + q_{\alpha/2}\sqrt{\hat{\sigma}^2\left(2 + \frac{1}{n_A} + \frac{1}{n_B}\right)}\right]$$

$$= \left[0.0635103 \pm 2.024394 * \sqrt{0.009759137\left(2 + \frac{1}{20} + \frac{1}{20}\right)}\right]$$

$$= [-0.2262977, 0.3533183]\,. \qquad (23)$$

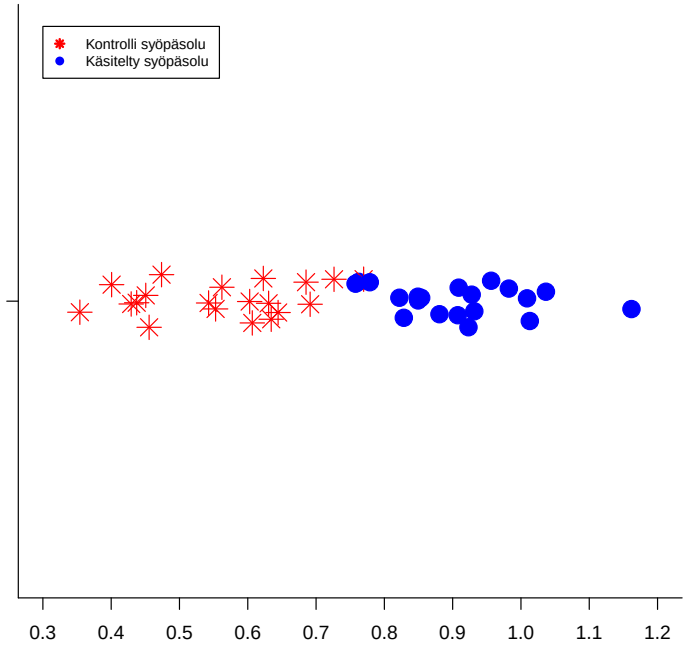
– Koska erotuksen $y_B - y_A$ 95 % luottamusväli sisältää arvon nolla, voidaan tehdä päättely, että 5 % riskitasolla tehdyillä käsittelyillä ei ole vaikutusta syöpäsolujen ekspressioarvojen jakauman sijaintiin hsa-miR-31 mikroRNA:n suhteen.

Ennustemerkitsevä kliininen ero

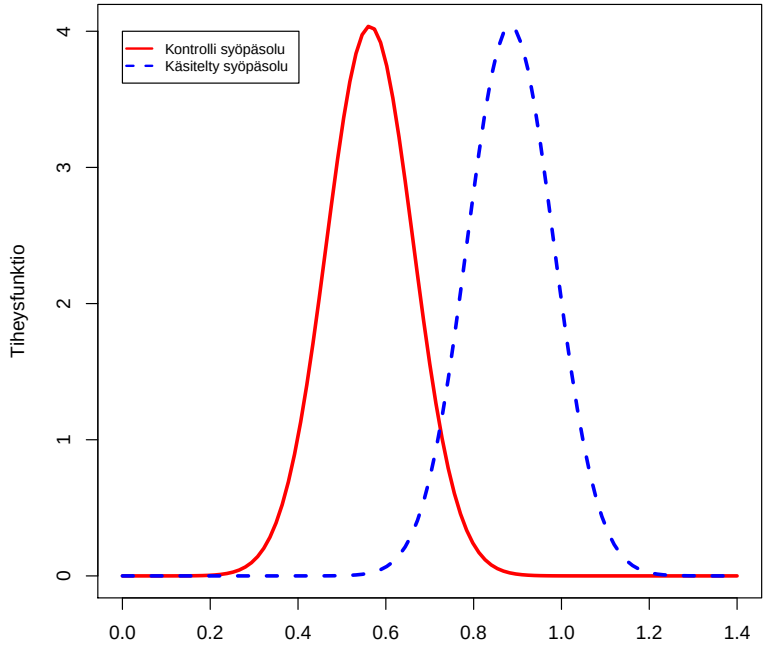
- Mielenkiintoista on vielä tutkia, milloin ekspressioarvojen jakaumat olisivat sijainniltaan niin paljon eroavat, että luottamusväliennustamisen perusteella tehtäisiin 5 % riskitasolla päättely, että käsittelyillä on vaikutusta syöpäsolujen ekspressioarvojen jakauman sijaintiin.
- Asymptoottisesti kliiniselle erolle ϵ tulee olla voimassa

$$\epsilon = |\mu_B - \mu_A| > 1.96 * \sqrt{2 * 0.009759137} = 0.2738273,$$

  missä $\sigma^2 = 0.009759137$.
- Käytännössä kun kliininen ero asetetaan tasolle $\epsilon = 0.32$, päästään 20 havainnon ekspressioaineistossa ennustemerkitseviin tuloksiin 80 prosentin tehokkuudella.



$\hat{\mu}_A = 0.5636493$ ja $\hat{\mu}_B = 0.8836493$



Kliininen ero: $\hat{\mu}_B - \hat{\mu}_A = 0.32$