Tõenäosusteooria ja statistika II

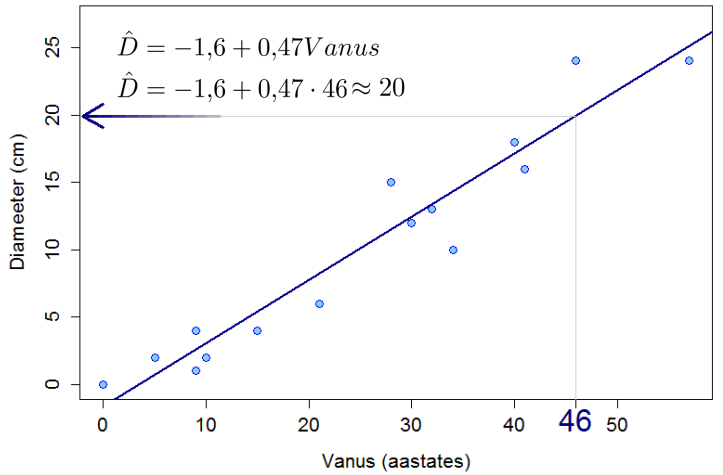
Loeng 16

Lihtne lineaarne regressioon
Tõepärasuhte test

Märt Möls

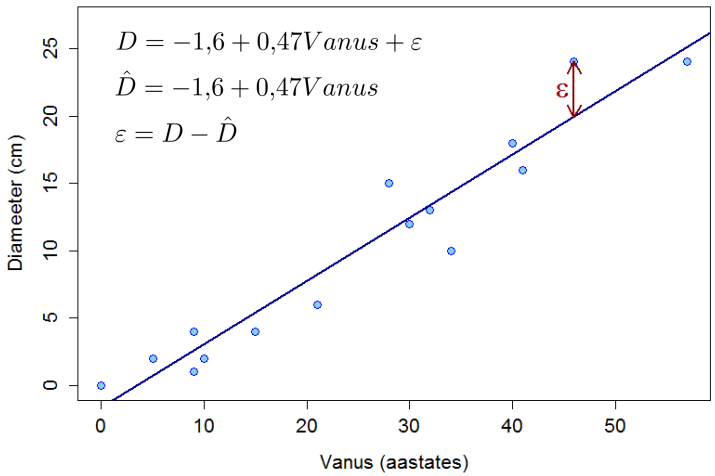
Prognoos

Remmelgas



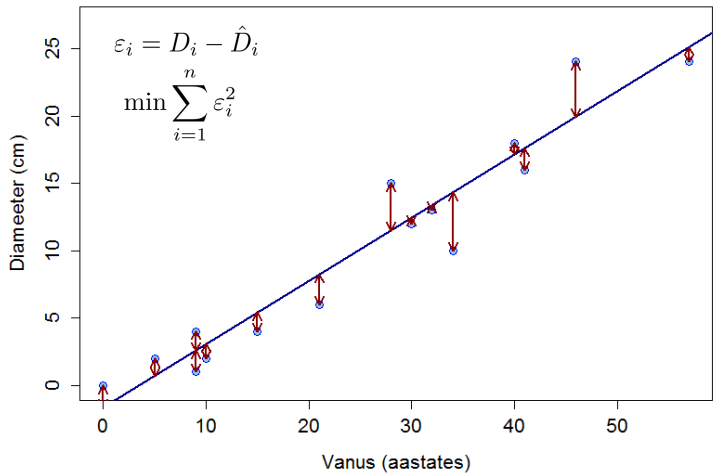
Prognoosiviga (jääk)

Remmelgas



Vähimruutude meetod

Remmelgas



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-1)$$

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-X_i)$$

$$\begin{cases} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i = 0 \end{cases}$$

$$\begin{cases} n\overline{Y} - n\beta_0 - n\beta_1 \overline{X} = 0 & | \cdot \overline{X} \\ \sum_{i=1}^{n} X_i Y_i - n\beta_0 \overline{X} - \beta_1 \sum_{i=1}^{n} X_i^2 = 0 \end{cases}$$

$$\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} - \beta_1 \left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \right) = 0$$

$$\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} - \beta_1 \left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \right) = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\hat{\mathrm{cov}}(X,Y)}{\hat{\sigma}_x^2}$$

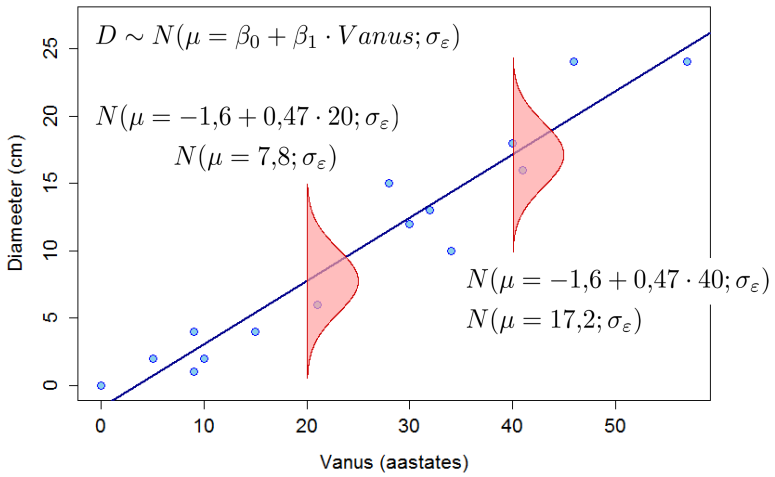
$$\hat{\mathrm{cov}}(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

$$\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$\begin{cases} n\overline{Y} - n\beta_0 - n\beta_1 \overline{X} = 0 \\ \sum_{i=1}^{n} X_i Y_i - n\beta_0 \overline{X} - \beta_1 \sum_{i=1}^{n} X_i^2 = 0 \end{cases}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Suurima Tõepära Meetod (ML) Remmelgas



Y_i tihedus (*i.* vaatluse panus tõepärasse):

$$\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} (Y_i - \beta_0 - \beta_1 X_i)^2 \right)$$

Tõepära (sõltumatud vaatlused):

$$L = (2\pi\sigma_\varepsilon^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right)$$

Log-tõepära:

$$l = -\frac{n}{2} \log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Maksimiseerime log-tõepära tundmatute parameetrite järgi:

$$\sigma_\varepsilon^2; \beta_0; \beta_1$$

$$l = -\frac{n}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{\partial l}{\partial \beta_0} = -\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n} 2\cdot(Y_i - \beta_0 - \beta_1 X_i)\cdot(-1)$$

$$\frac{\partial l}{\partial \beta_1} = -\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n} 2\cdot(Y_i - \beta_0 - \beta_1 X_i)\cdot(-X_i)$$

$$\frac{\partial l}{\partial \sigma_\varepsilon^2} = -\frac{n}{2}\cdot\frac{1}{2\pi\sigma_\varepsilon^2}\cdot 2\pi + \frac{1}{2(\sigma_\varepsilon^2)^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\begin{cases} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0 \\ -n/\sigma_\varepsilon^2 + \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2/\sigma_\varepsilon^4 = 0 \end{cases}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\hat{\mathrm{cov}}(X,Y)}{\hat{\sigma}_x^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$$

$$-n/\sigma_\varepsilon^2 + \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2/\sigma_\varepsilon^4 = 0$$

$$n\sigma_\varepsilon^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

$$\begin{cases} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = 0 \\ -n/\sigma_\varepsilon^2 + \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2/\sigma_\varepsilon^4 = 0 \end{cases}$$

Milliste omadustega on saadud hinnangud?

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

Milliste omadustega on saadud hinnangud?

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})\mathrm{E}\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(\mathrm{E}\left(Y_i\right) - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Milliste omadustega on saadud hinnangud?

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})\mathrm{E}\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\mathrm{E}\left(Y_i\right) - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Milliste omadustega on saadud hinnangud?

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})\mathrm{E}\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\mathrm{E}\left(Y_i\right) - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i - (\beta_0 + \beta_1 \overline{X}))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Milliste omadustega on saadud hinnangud?

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})\mathrm{E}\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\mathrm{E}\left(Y_i\right) - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\cancel{\beta_0} + \beta_1 X_i - (\cancel{\beta_0} + \beta_1 \overline{X}))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Milliste omadustega on saadud hinnangud?

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})\mathrm{E}\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\mathrm{E}\left(Y_i\right) - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i - \mathrm{E}\left(\overline{Y}\right))}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_1 X_i - \beta_1 \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$= \frac{\beta_1 \sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X}))}{\sum_{i=1}^n (X_i - \overline{X})^2} = \beta_1$$

Nihketa hinnang!

Milliste omadustega on saadud hinnangud?

$$\begin{aligned}\mathrm{E}(\hat{\beta}_0) = E(\overline{Y} - \hat{\beta}_1\overline{X}) &= E(\overline{Y}) - E(\hat{\beta}_1)\overline{X} \\ &= E(\overline{Y}) - \beta_1\overline{X} \\ &= \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} \\ &= \beta_0\end{aligned}$$

Jälle nihketa!

Milliste omadustega on saadud hinnangud?

$$\begin{aligned}\mathrm{D}(\hat{\beta}_1) &= \mathrm{D}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \\ &= \mathrm{D}\left(\sum_{i=1}^n \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}(Y_i - \overline{Y})\right) \\ &= \mathrm{D}\left(\sum_{i=1}^n \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} Y_i\right) \\ &= \sum_{i=1}^n \frac{(X_i - \overline{X})^2}{(\sum_{i=1}^n (X_i - \overline{X})^2)^2}\sigma^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\end{aligned}$$

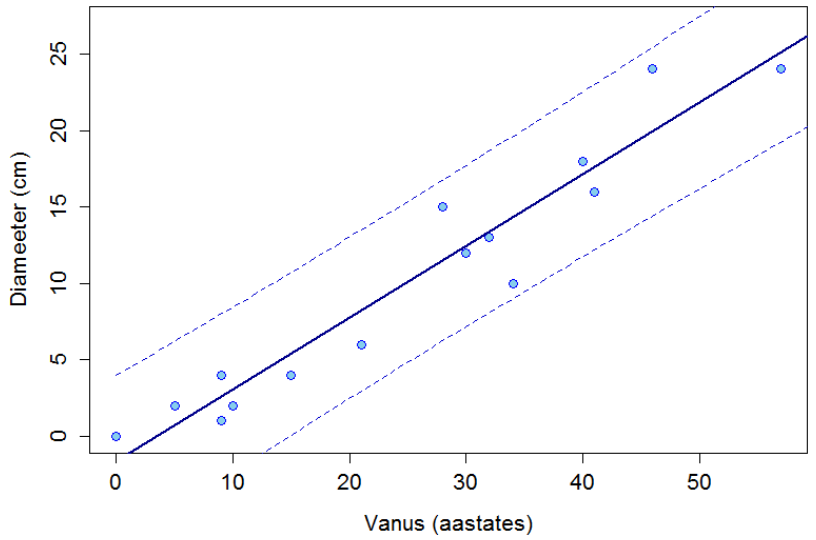
$$\hat{\beta}_1 \sim N\left(\beta_1; \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1; \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

$$\sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2/\sigma^2 \sim \chi^2(n-2)$$

Mudel uuritavale tunnusele – võimaldab leida prognoosiintervalli



Tõepärasuhte test

	parameetrid	tõepära	tõepärafunktsiooni maksimum
H_0: võime kasutada lihtsamat mudelit (vähem hinnatavaid parameetreid)	θ_0	$L(\theta_0)$	$L_0 := L(\hat{\theta}_0)$
H_1: peame kasutama suuremat mudelit – rohkem hinnatavaid parameetreid	θ	$L(\theta)$	$L := L(\hat{\theta})$

$$-2 \ln \frac{L_0}{L} \overset{H_0}{\underset{n \to \infty}{\sim}} {\chi_{df}}^2$$

* Kui fikseeritud parameetrid ei asu parameetri võimalike väärtuste ruumi piiril
* suure valimi korral

Tõepärasuhte test - eeldused

$$Y \sim N(\mu; \sigma^2)$$

Võime kasutada testimaks hüpoteese

$$H_0 : \mu = 0 \qquad H_0 : \sigma^2 = 10$$

Ei saa kasutada testimaks hüpoteesi

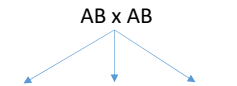
$$H_0 : \sigma^2 = 0$$

$$-2 \ln \frac{L_0}{L} \overset{H_0}{\underset{n \to \infty}{\sim}} {\chi_{df}}^2$$

* Kui fikseeritud parameetrid ei asu parameetri võimalike väärtuste ruumi piiril
* suure valimi korral

Tõepärasuhte test - näide

H_0: Teise põlvkonna hübriidide jaoks kehtib Mendeli seadus:



Järglane:	AA	AB	BB
Tõenäosus:	0,25	0,5	0,25

$$\hat{\theta}_0 = \begin{pmatrix} 0.25 \\ 0.5 \\ 0.25 \end{pmatrix}$$

$$\hat{\theta} = \begin{pmatrix} n_{AA}/n \\ n_{AB}/n \\ n_{BB}/n \end{pmatrix}$$

$$L_0 = 0,25^{n_{AA}} \cdot 0,5^{n_{AB}} \cdot 0,25^{n_{BB}}$$

$$L = (n_{AA}/n)^{n_{AA}} \cdot (n_{AB}/n)^{n_{AB}} \cdot (n_{BB}/n)^{n_{BB}}$$

$$L_0/L = (0,25n/n_{AA})^{n_{AA}} (0,5n/n_{AB})^{n_{AB}} (0,25n/n_{BB})^{n_{BB}}$$

$$-2 \ln(L_0/L) = 2 \sum_{i=1}^{k} O_i \ln(O_i/E_i)$$

Tuntakse ka G-testi nime all, sageli soovitatakse hii-ruut testi asemel.

$$-2 \ln \frac{L_0}{L} \overset{H_0}{\underset{n \to \infty}{\sim}} {\chi_{df}}^2$$

Tõepärasuhte test – alternatiivsed esitused

$$-2 \ln \frac{L_0}{L} \overset{H_0}{\sim} {\chi_{df}}^2$$

$$-2(l_0 - l) \overset{H_0}{\sim} \chi^2_{df}$$

$$2(l - l_0) \overset{H_0}{\sim} \chi^2_{df}$$

Tõepärasuhte test – tõestuse idee

Vaatame, kuidas saame tõestada, et tõepärasuhte testi teststatistik on asümptootiliselt hii-ruut jaotusega siis, kui tarvis on hinnata vaid ühte parameetrit (ja nullhüpoteesi kehtides polegi hinnatavaid parameetreid)

Arendame log-tõepära $l(\theta_0)$ Taylori ritta punkti $\hat{\theta}$ ümbruses:

$$l(\theta_0) \approx l(\hat{\theta}) + \underbrace{l'(\hat{\theta})}_{=0}(\theta_0 - \hat{\theta}) + \frac{l''(\hat{\theta})}{2}(\theta_0 - \hat{\theta})^2$$

$$l(\theta_0) - l(\hat{\theta}) \approx \frac{l''(\hat{\theta})}{2}(\theta_0 - \hat{\theta})^2$$

$$-2(l(\theta_0) - l(\hat{\theta})) \approx -\underbrace{\frac{l''(\hat{\theta})}{\sigma(\hat{\theta})^2}}_{\to 1} \underbrace{\left(\frac{\theta_0 - \hat{\theta}}{\sigma(\hat{\theta})}\right)^2}_{\text{kui H0 siis } \sim N(0;1)} \overset{H_0}{\sim} \chi^2_{df=1}$$