gyors STATISZTIKA

Biomatematika
2018/2019 II
félév
Tóth Gergely
Kémia 541-be.

http://tothgergely.web.elte.hu ← letölthető innen

Vegyes hallgatói háttér ⇒ alapfogalmak is megjelennek

(elmélet, angol - gyakorlati rész támogatására)

Valószínűségszámítás: események sokaságának tárgyalása valószínűségi alapon (determinisztikus tól összelett) (populáció)

Statisztika: minta tulajdonságai alapján mit mondhatunk a sokaságról

Várható érték és becslései

vsi-ben. $E(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx$

↑ val. sűrűségfv.

folytonos

diszkrét val.v.-ll. eloszlás

$E(x) = \sum_i x_i \cdot p(x_i)$

diszkrét ↑ pl kockára $p(x_i) = 1/6$

$E(x) = 3,5$

minta: N darab $y_1, y_2 \dots y_n$

átlag $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$

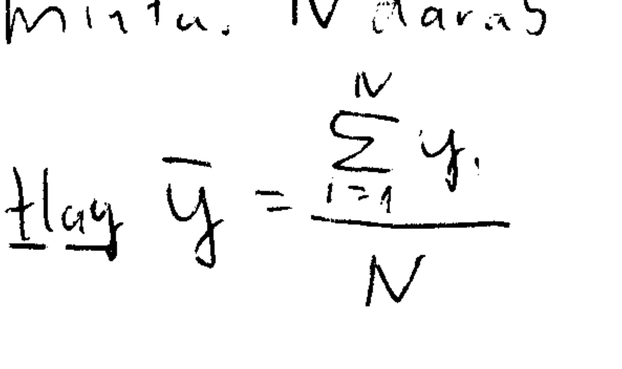
kilógó adatnál: median

1) sorbarendezés $y_1 \le y_2 \le y_3 \le y_n$

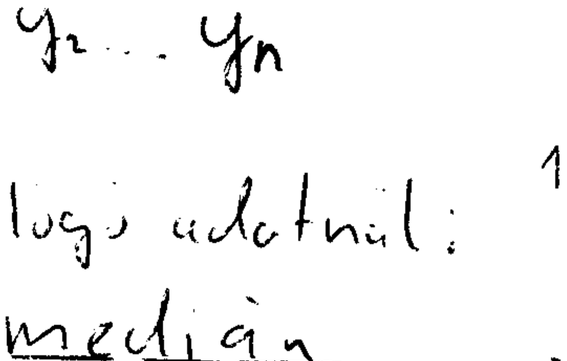
2) középső, vagy középsők átlaga

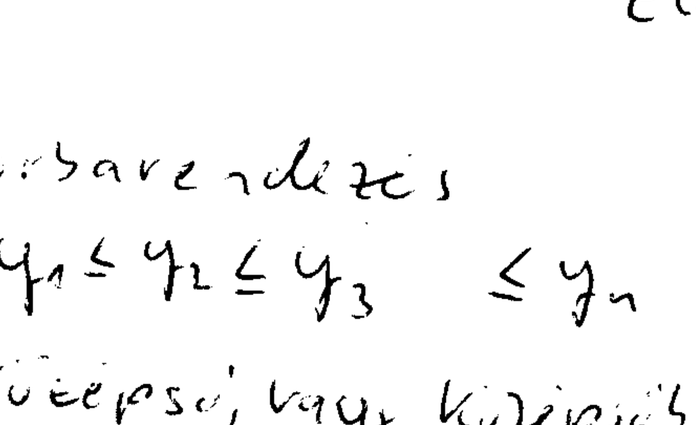
robosztus becslés: sokféle adatra jó, hibásakra is (nemérzékeny)

módusz: leggyakoribb



diszkrét val.v.-ll. eloszlása

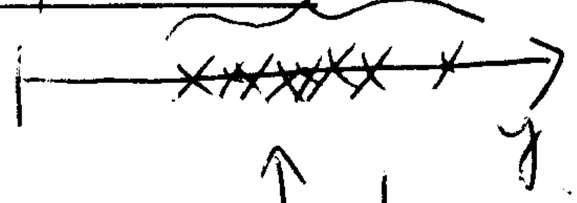




bimodális eset

f(x): valsűrűség fv-ny, integrálja 1

Terjedelem:



Variancia, szórásnégyzet: $\sigma^2 = \int_{-\infty}^{\infty} (x - E(x))^2 f(x)\,dx$

$$\sigma^2 = \sum (x_i - E(x))^2 p(x_i)$$

mintából:

estimated variance $s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$

becsült szórás $s = \sqrt{s^2}$ (standard deviation)

Variációs koefficiens (coefficient of variation)

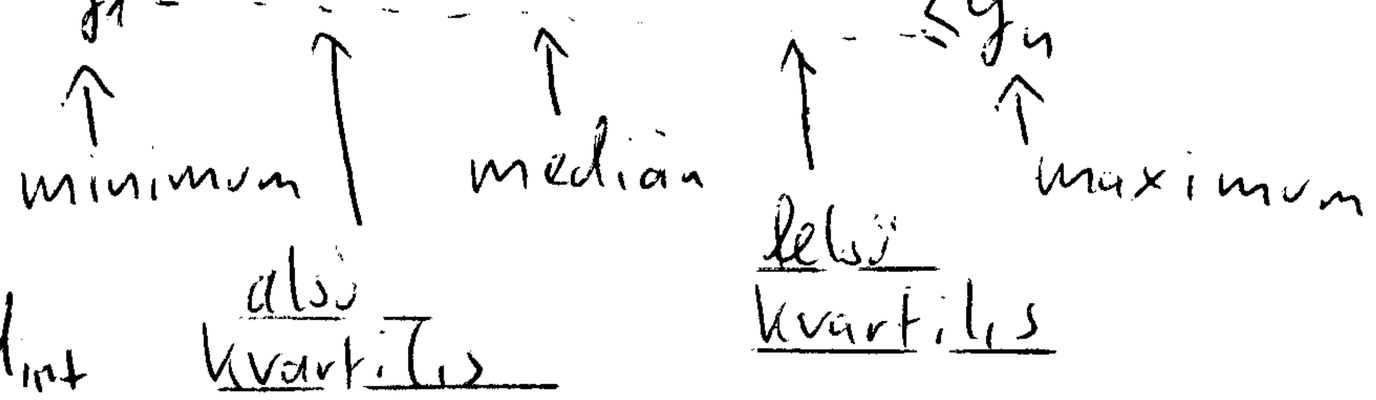
$c.o.f = \frac{s}{\bar{y}}$ más ha K vagy °C, 0 K-nél nincs gond!

Sorbarendezett mintánál

tetszőleges % = percentilis

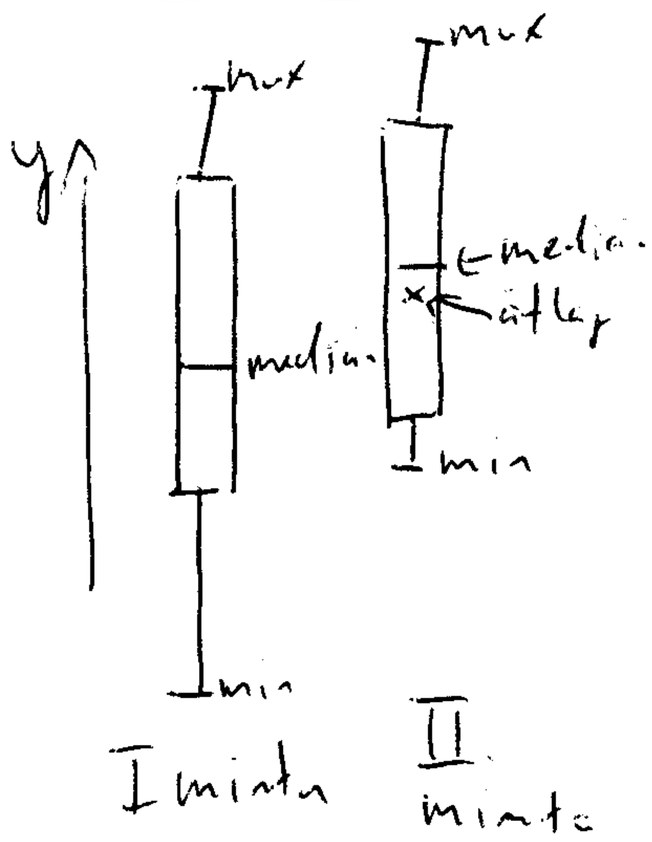
$y_{max} - y_{min}$ = range (terjedelem)

$y_{3/4} - y_{1/4}$ = interkvartilis távolság $= d_{int}$

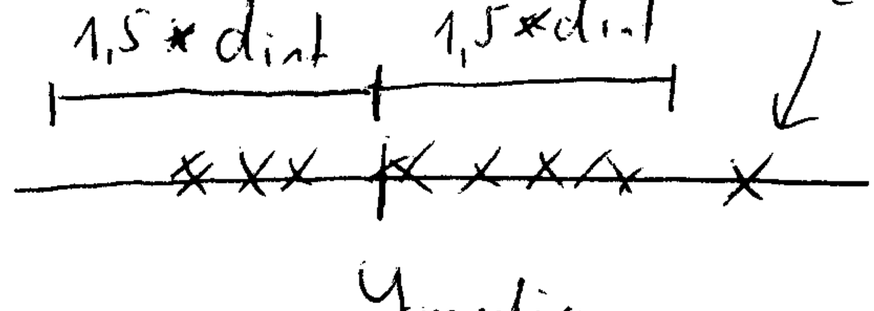


mire jó:

boxplot:



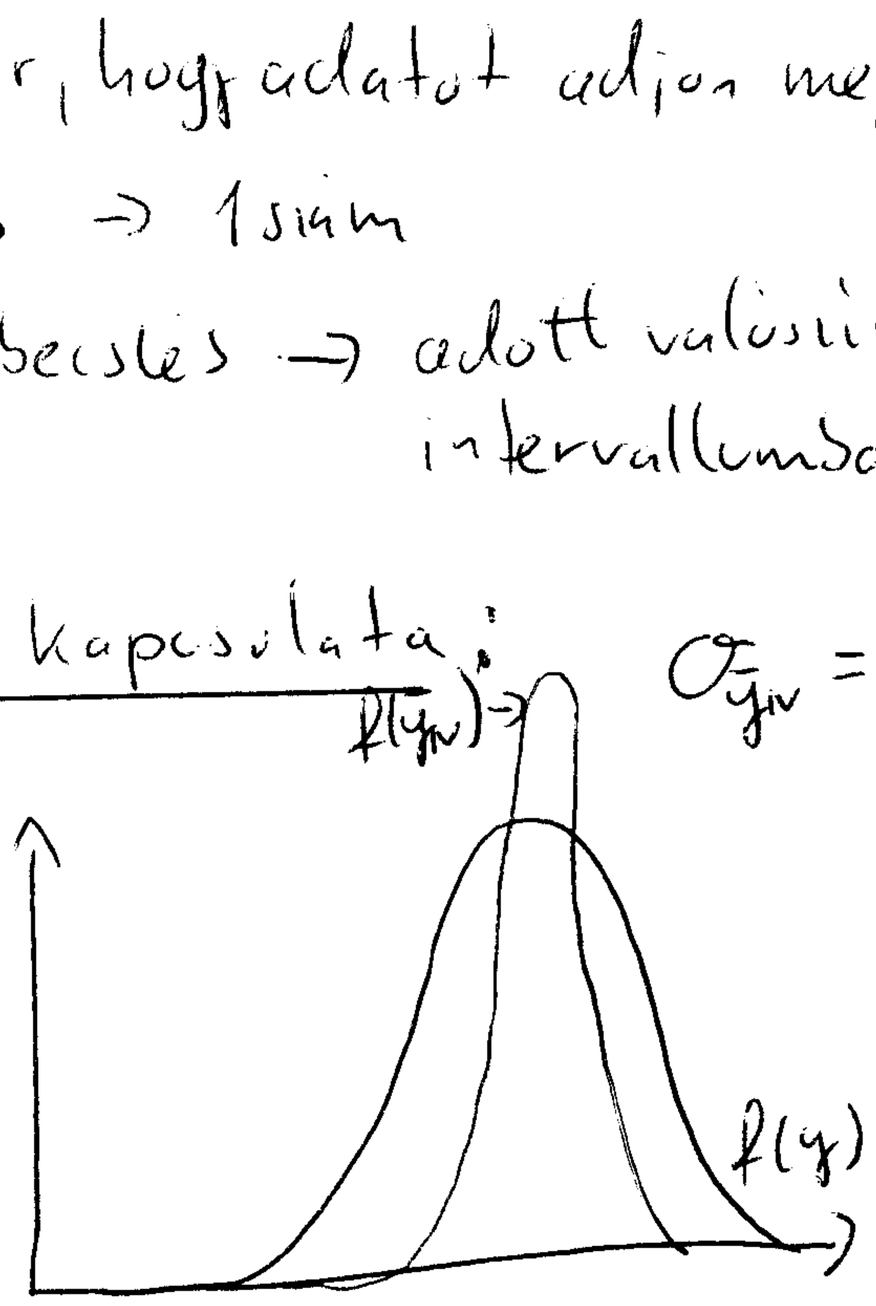
kilógó adat eldobása:



analitikus mér, hogy adatot adjon meg (becsüljön)

- pontbecslés → 1 szám
- intervallumbecslés → adott valószínűséggel ebben az intervallumban van (pl. 95%, 99%)

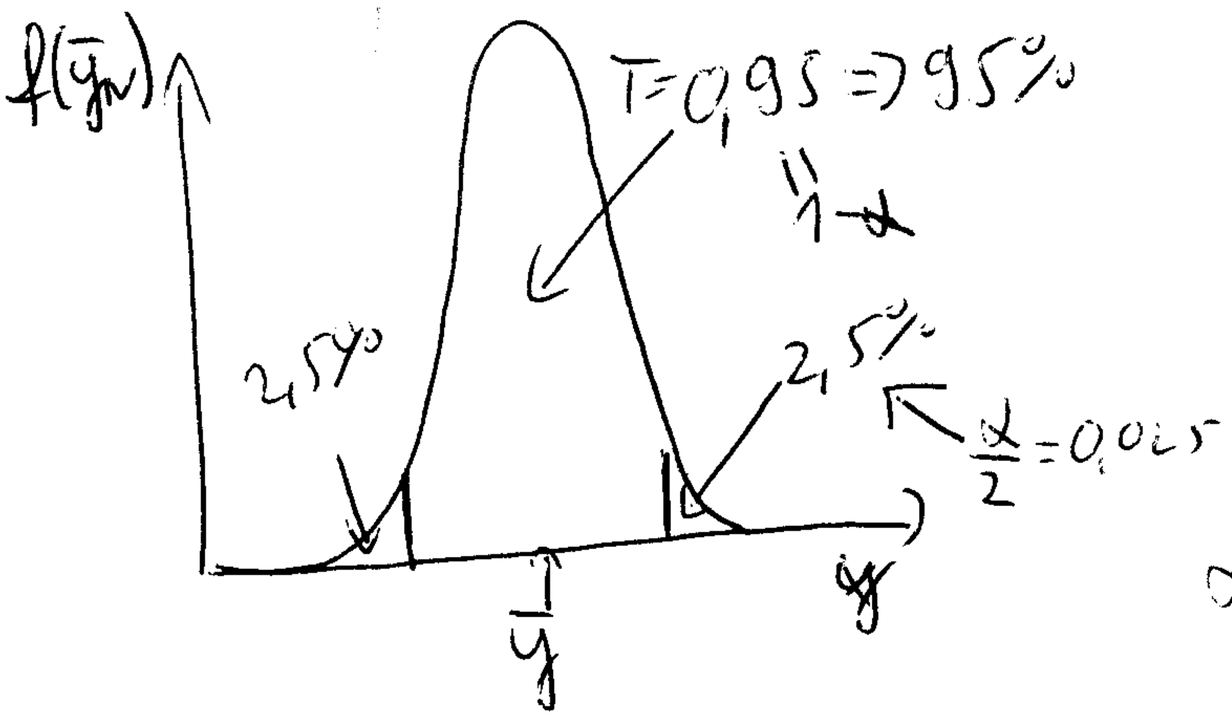
σ_y és $\sigma_{\bar{y}_N}$ kapcsolata:



$$\sigma_{\bar{y}_N} = \frac{\sigma_y}{\sqrt{N}}$$

⇒ több mérés, szűkül az intervallum

Konfidencia intervallum várható értékre: $\bar{y} \pm ?$

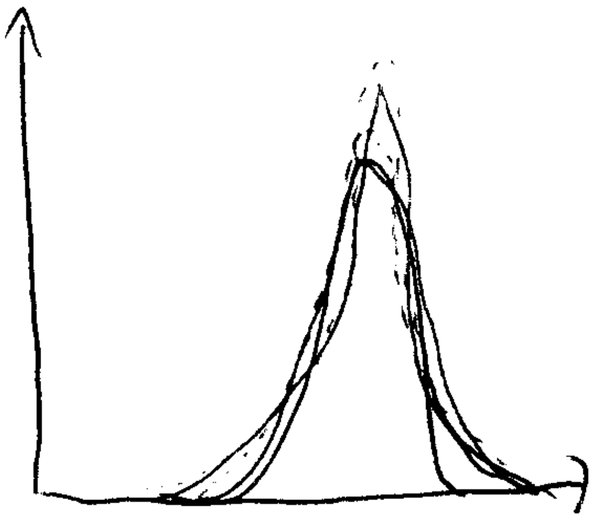


szimmetrikusra, ha normál eloszlású és $N \geq 30$

$$\bar{y} \pm \frac{1{,}96 \cdot s}{\sqrt{N}}$$

α szignifikanciaszint

kis mintánál gond, hogy s és $\bar{y}$ is ugyanabból van becsülve, normál eloszlás → student eloszlás (t-eloszlás)



$$\bar{y} \pm \frac{|t_\nu^{-1}(\frac{\alpha}{2})| \cdot s}{\sqrt{N}}$$

ν szabadsági fok

$\nu = N - 1$

$\bar{y} \pm \sim \neq y(s)$

↑ ez a konfidencia intervallum

↑ ide sosem szokás írni

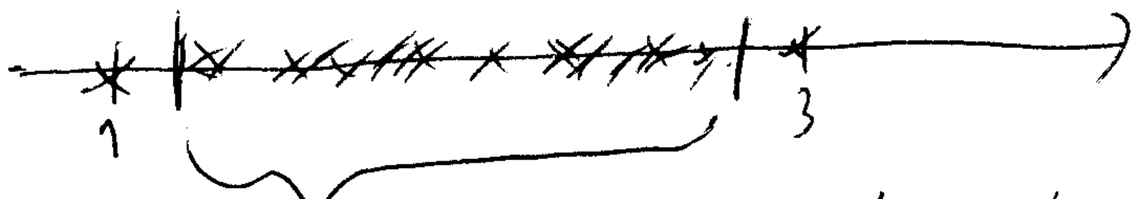
Konfidencia intervallum (nagyon) nem normális eloszlásra.

pl. bootstrap módszer: minta ⇒ populáció, visszatevéses újra mintavételezés:

$y_1=1,\ y_2=2,\ y_3=3$ ⇒ 27 új minta

111		27db
222		
333		
123	6	
122	·3	
322		
211		
311		
133		
233		

ezek $\bar{y}_{0,13}$ ←



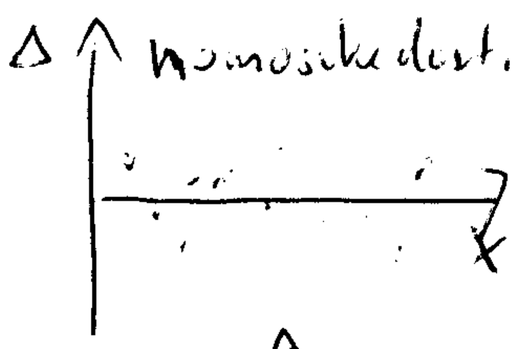
95% adja az intervallumot

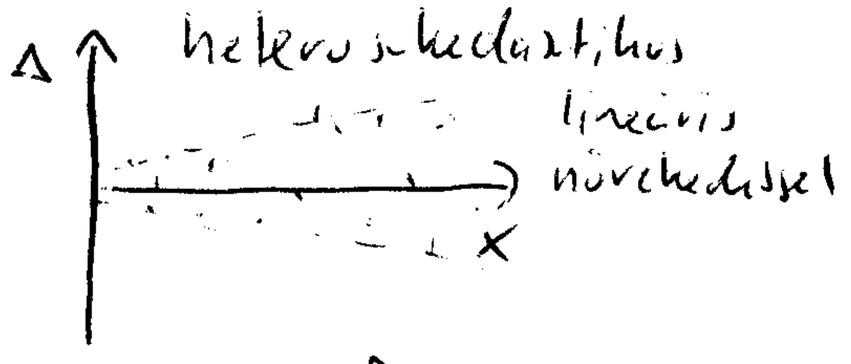
Statisztika: döntéseket hozhatunk

statisztikai tesztek (próbák, hipotézis vizsgálatok)

Pl: Validálás során torzítatlanság vizsgálható.

x pontosan bemért – y meghatározott koncentráció $\Delta = y - x$





átlagos torzítás vizsgálata

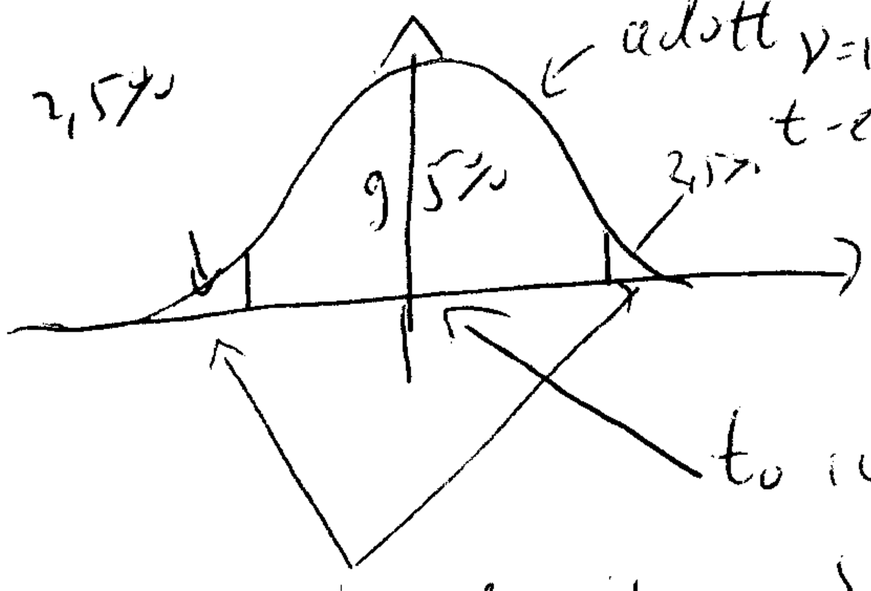
$H_0: \bar{E}(y-x) = 0 = E_0$

átlagos visszamérés módszere

$H_0: E\left(\frac{y}{x}\right) = 1 = E_0$

módszer: egymintás t-próba

$$t_0 = \frac{\bar{y} - E_0}{s/\sqrt{N}}$$



EXCEL:

T.ELOSZLÁS(t₀, N-1, 1)

↓

kivonás 1-ből

↙ ↘

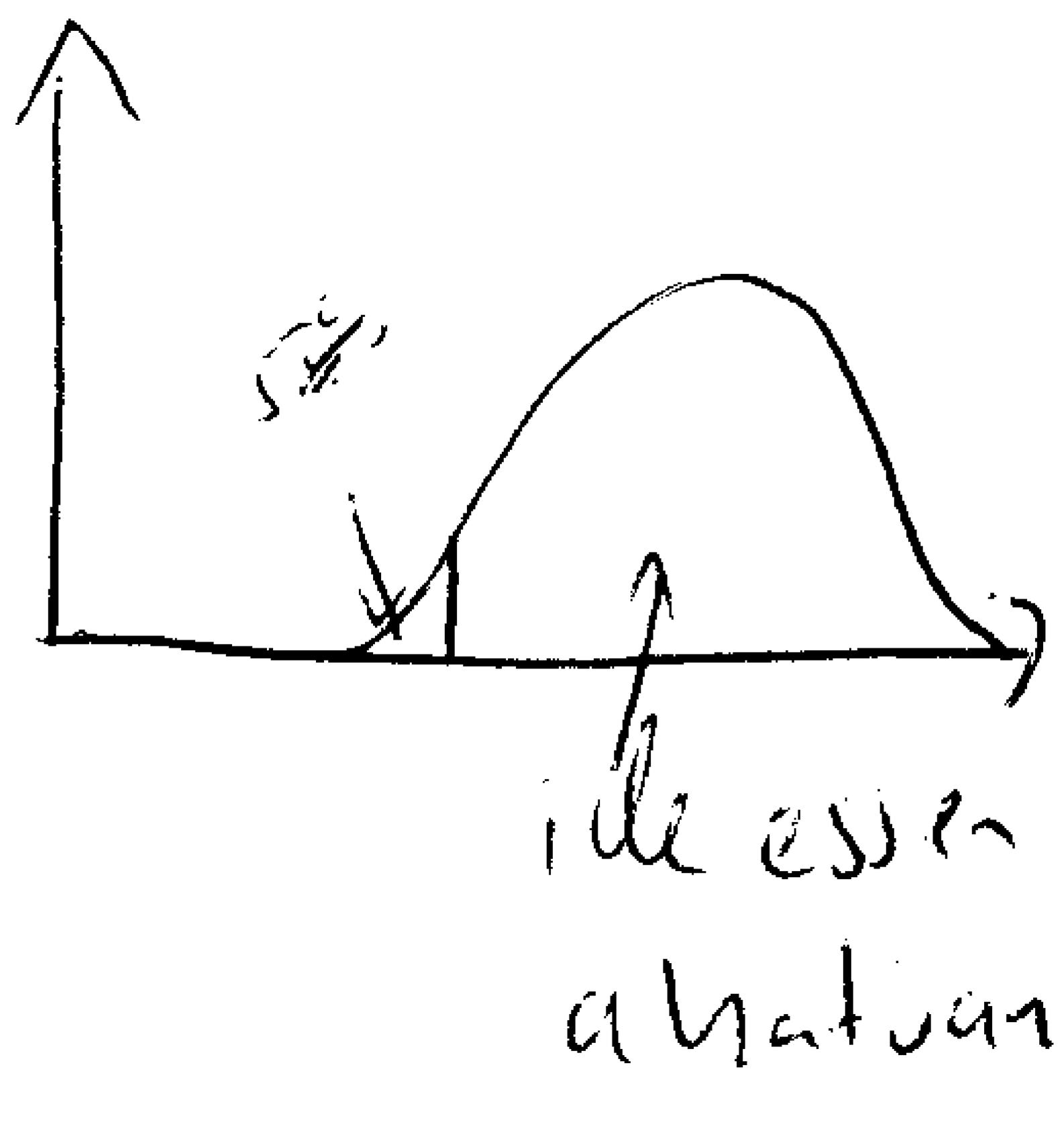
ha s helyett σ

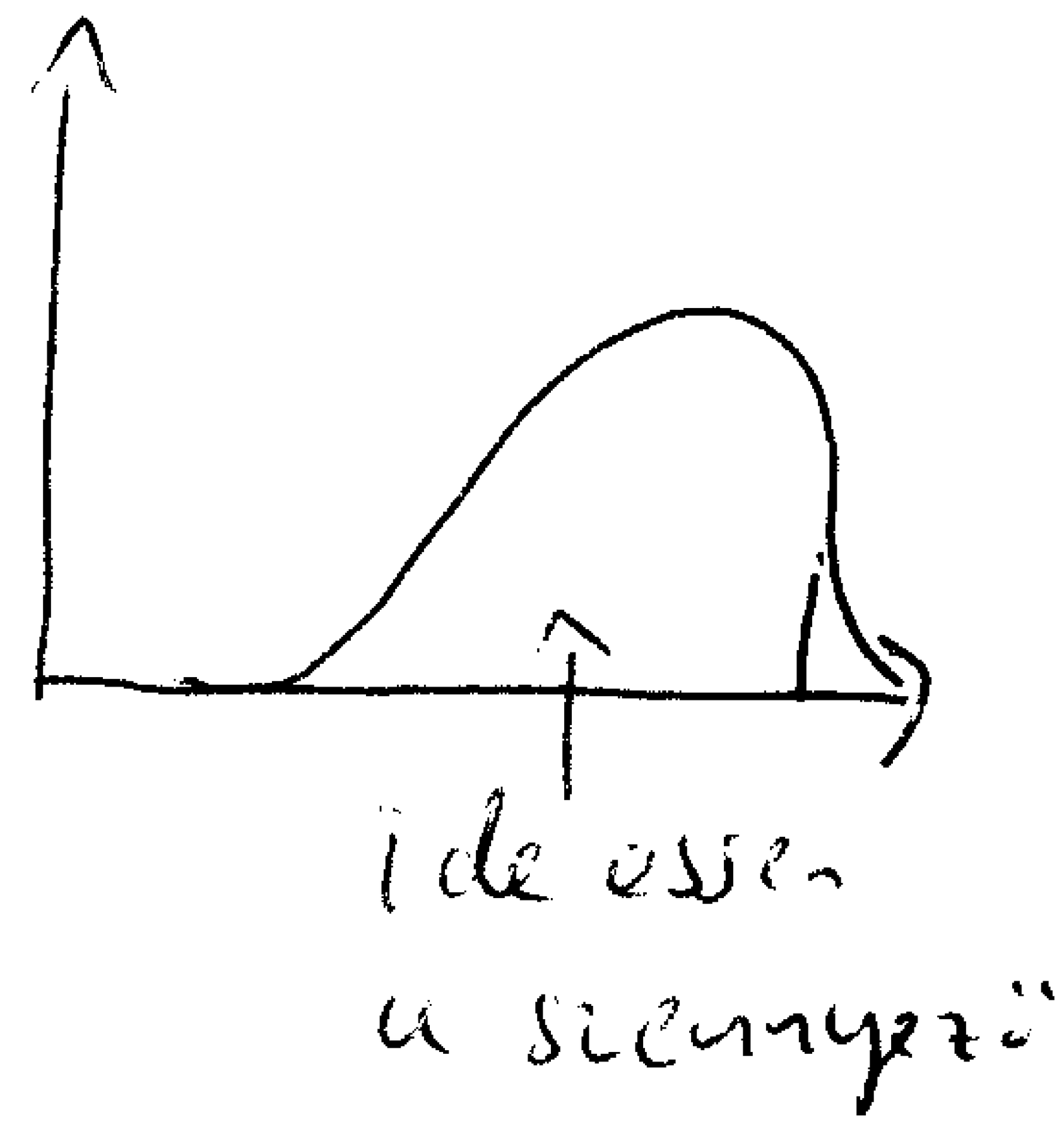
↓

u vagy z próba

annak, normál eloszlás

lehet egyoldali eset





A hagyományos statisztikai próbák H_0-nak kedveznek (a gyártónak!)

Validáció: két labor összehasonlítása

Kétmintás t-próba $H_0: E_1 = E_2$ $\quad H_a: E_1 \neq E_2$

mivel s_1 sokszor más, mint s_2, ezért többnyire az használjuk, hogy nem egyező varianciájuak. ~~közös~~ (variancia becslés mindkettőből)

labor 1 labor 2

⇒ EXCEL P-értéket kapunk pl. adatelemzés (van párosított is)

$0{,}05 \leq$ p-value H_0 ✓

$0{,}05 >$ p-value H_a

Regresszió: egyváltozós → egyenesillesztés $x \to y$

többváltozós → sík, hipersík illesztése $\underline{x} \to y$

N x_i, y_i pár. $y_i = a x_i + b + \varepsilon_i$

ε_i homoszkedasztikus → nem kell súlyozni a pontokat (w_i)

$w_i = 1/\sigma_i^2$

legkisebb négyzetek módszere:

OK, ha $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$

$E(\varepsilon_i) = 0$

normál eloszlású

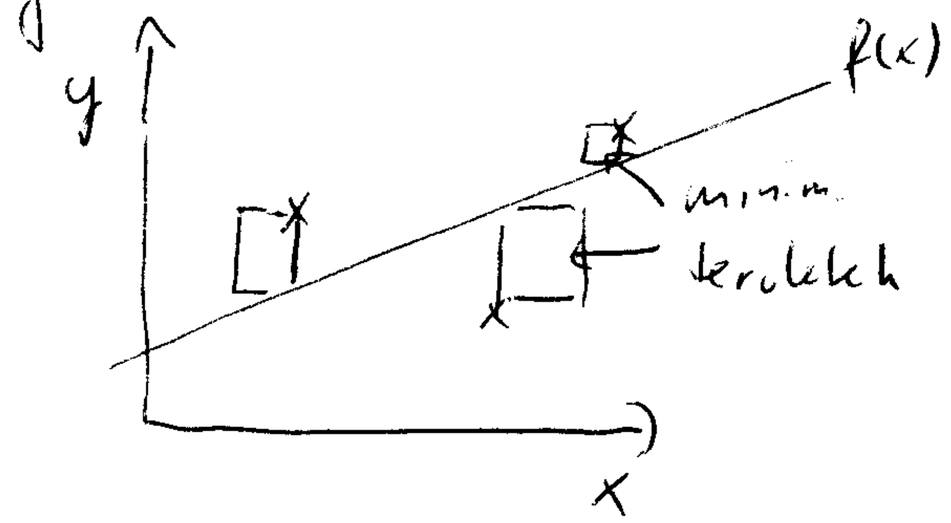
maximum likelihood becslés?

$$Q = \sum_{i=1}^{N} (y_i - a x_i - b)^2$$

$$\min Q \Rightarrow \frac{\partial Q}{\partial a} = 0, \quad \frac{\partial Q}{\partial b} = 0 \Rightarrow$$

$$a = \frac{\sum (x_i y_i - \bar{x}\bar{y})}{\sum (x_i x_i - \bar{x}\bar{x})} \qquad b = \bar{y} - a\bar{x}$$

grafikusan LKN



eredmény:

$\hat{y}_i$ számolt, y_i mért

$\bar{\hat{y}}_i = \bar{y}_i$

$a \pm$ $b \pm$ pl. 95% intervallumnal

↑ kivonva a 0, kell tengelymetszet!

<u>reziduális:</u> $r_i = y_i - \hat{y}_i$ $\quad s_r^2 = \sum r_i^2 / (N-2) \Rightarrow s_a^2, s_b^2$ essél

ábrázolásra, főleg standardizálva segít → kilógó értékek

<u>Regresszió variancia analízise:</u>

		négyzetösszeg
modell (regr.)	p	$MSS = \sum(\hat{y}_i - \bar{y})^2$
maradék (rezid.)	$N-p-1$	$RSS = \sum(\hat{y}_i - y_i)^2$
teljes (total)	$N-1$	$TSS = \sum(y_i - \bar{y})^2$

$(TSS = MSS + RSS)$

p = meredekségek száma, tengelymetsz. nincs benne

variancia

$MSS/p = s_M^2$

$RSS/(N-p-1) = s_e^2$

F_0 $\quad s_M^2 / s_r^2$ $\quad$ p-value

<u>R^2 - determinációs együttható</u>

$R^2 = 1 - \frac{RSS}{TSS}$ $\quad R^2 \in [0,1]$ (itt)

$\vec{y}, \vec{\hat{y}}$ <u>vektorok</u> egy N-dim térben, bezárt szögük cos²-ja

ha különböző N-ű modellek → R^2_{adj} ← lehet negatív

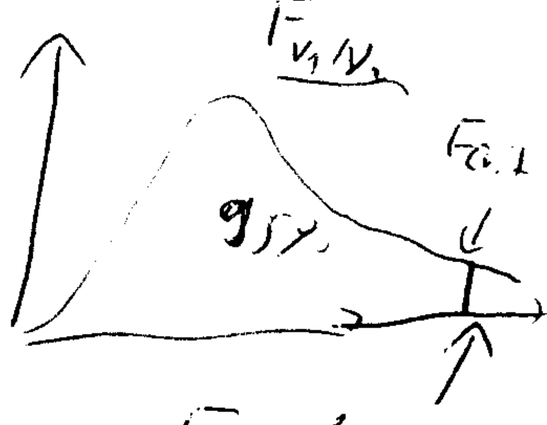
↓

de etcsak átskálázás $[-(N,n), 1]$

<u>F-eloszlás kell</u>

variancia ujjgyakorlat

a sokaságból



ha nem hányser nélküli regresszió, pl $b \equiv 0$ egyéb más modell.

$R^2 \in (-\infty, 1]$, $\bar{y}_0 \neq \bar{\hat{y}}$, $TSS \neq MSS + RSS$

RMSE - root mean square error

$$RMSE = \sqrt{\frac{\Sigma(\hat{y}_i - y_i)^2}{N}}$$

számítás, átlagos hiba (átlag négyzetes hiba)

nem statisztikai, de jó

$Q^2_1(R^2_{CV})$ $\hat{y}_{ji}$ ← olyan modellben számolt érték, ahol ez pont nem volt benne ☆

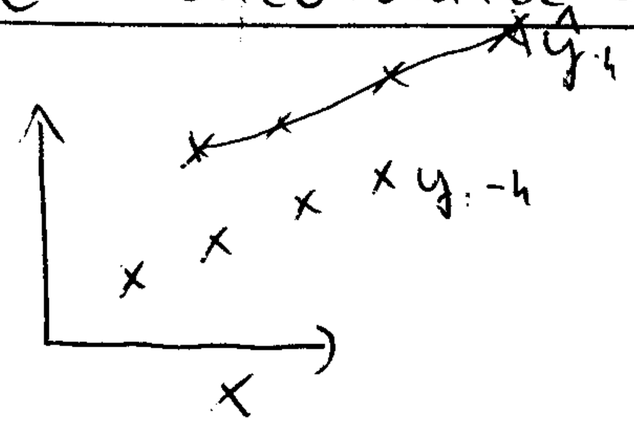
↑ cross-validation

test-tanító (test/training) felosztás

pl. 80% tanító → ~~erre~~ modell elkészítése 20%-on

$\approx R^2_{test}$ jellegű, $RMSE_{test}$

CCC Concordance Correlation Coefficient (újabb)



Csak az irányt azonos $\Rightarrow R^2 = 1$

vektor ≠ irányít. H jellemzi

ezek együttese kellene.

$CCC \in [-1, 1]$ minél korrelációs eh csak irányít. H jellemzőre, nem vektorokra

Feladat: illeszkedés jósága: R F, R^2, RMSE, CCC

robusztusság: Q^2 (belső validáció)

predikció test-training felosztás (külső validáció)

Példa:

$\underline{y} = (2,1; 2; 2,2; 2; 2,3; 2,1; 2,1)$ EXCEL

átlag	2,11428	(átlag)
medián	2,1	(medián)
módusz	2,1	(módusz)
tap. szórás	0,10694	(szórás)

$2,11428 \pm 0,09887 \Rightarrow \underline{2,11 \pm 0,10}$ ✓

$\Rightarrow 2,114 \pm 0,099$

egymintás t-~~elosítás~~ próba ~~$E_0 = 2,5$~~-re $E_0 = 2,25$-re

$t_0 = \dfrac{2,5 - \bar{y}}{s/\sqrt{N}} \rightarrow$ T.ELOSZL(t_0; $\nu = N-1$; 1)

↑ eloszlás fv.

$\Downarrow$

3,358 $\Rightarrow$ 0,9923

$\Downarrow$ ide esik $H_0: E \neq E_0$

$0,025 \leq \quad \leq 0,975$