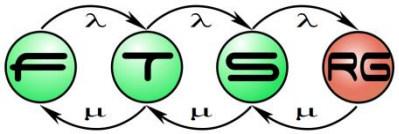
Adatelemzés és mérések

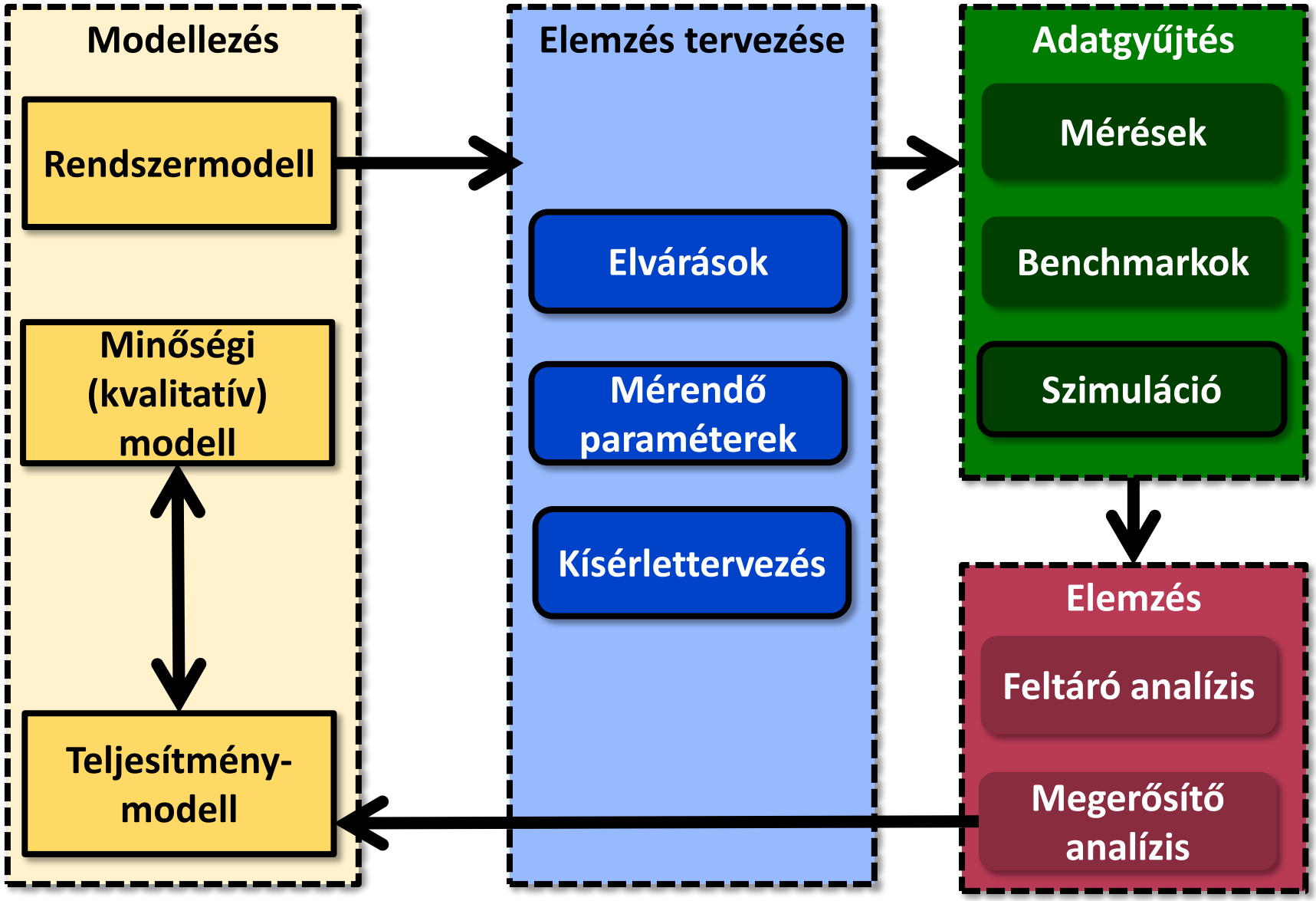
Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



Miről lesz ma szó?

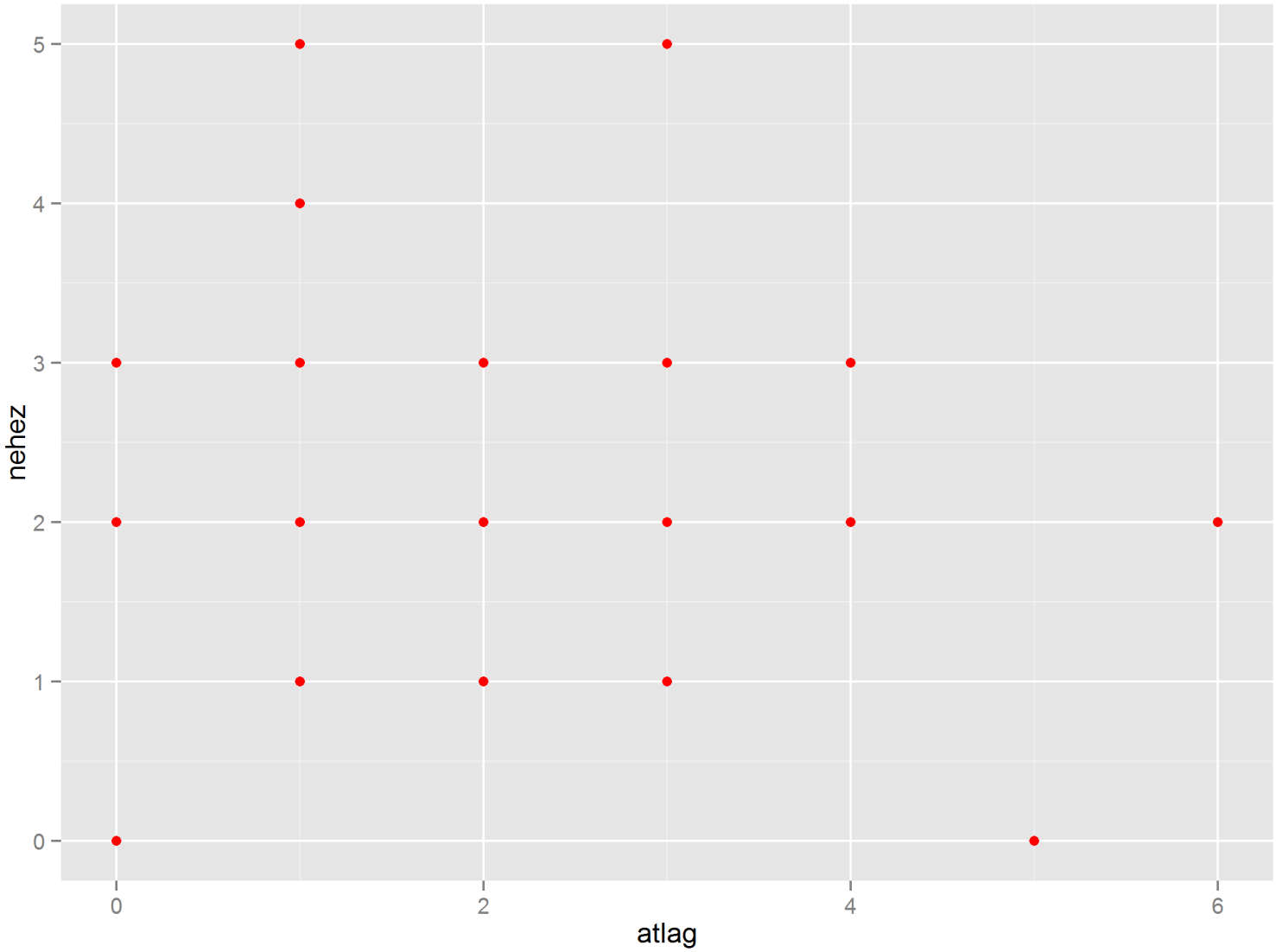
- Mérési adatok → összefüggések/előrejelzés?
  - Regresszió, átlagolás
  - Következető elemzés (alapok, célja)
- Benchmarking
  - Mérés helyett…
- Kísérlettervezés
  - Mennyit mérjünk? Mennyire higyjük el az eredményt?

Rendszermodelltől a teljesítménymodellig

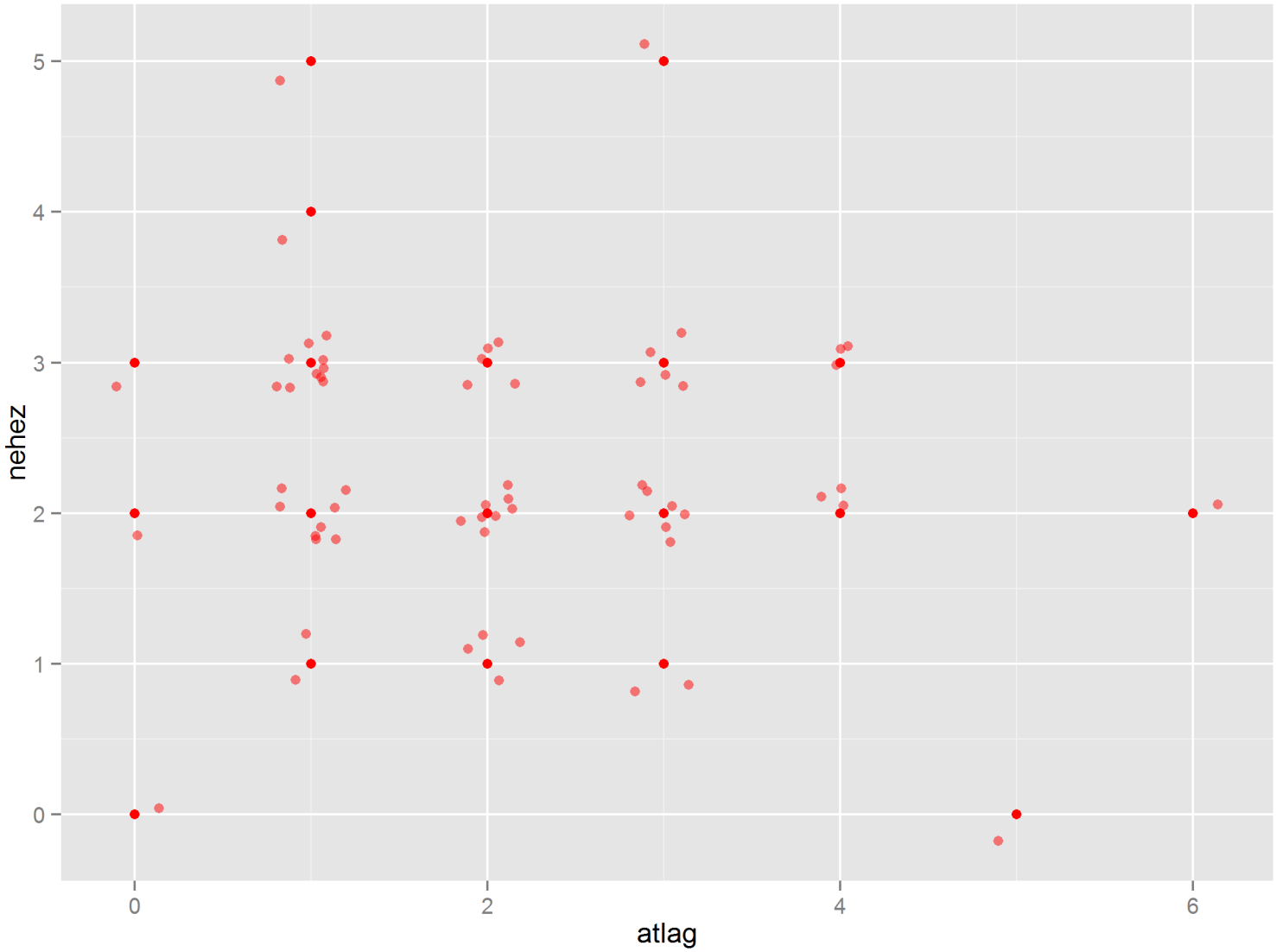


REGRESSZIÓS MÓDSZEREK

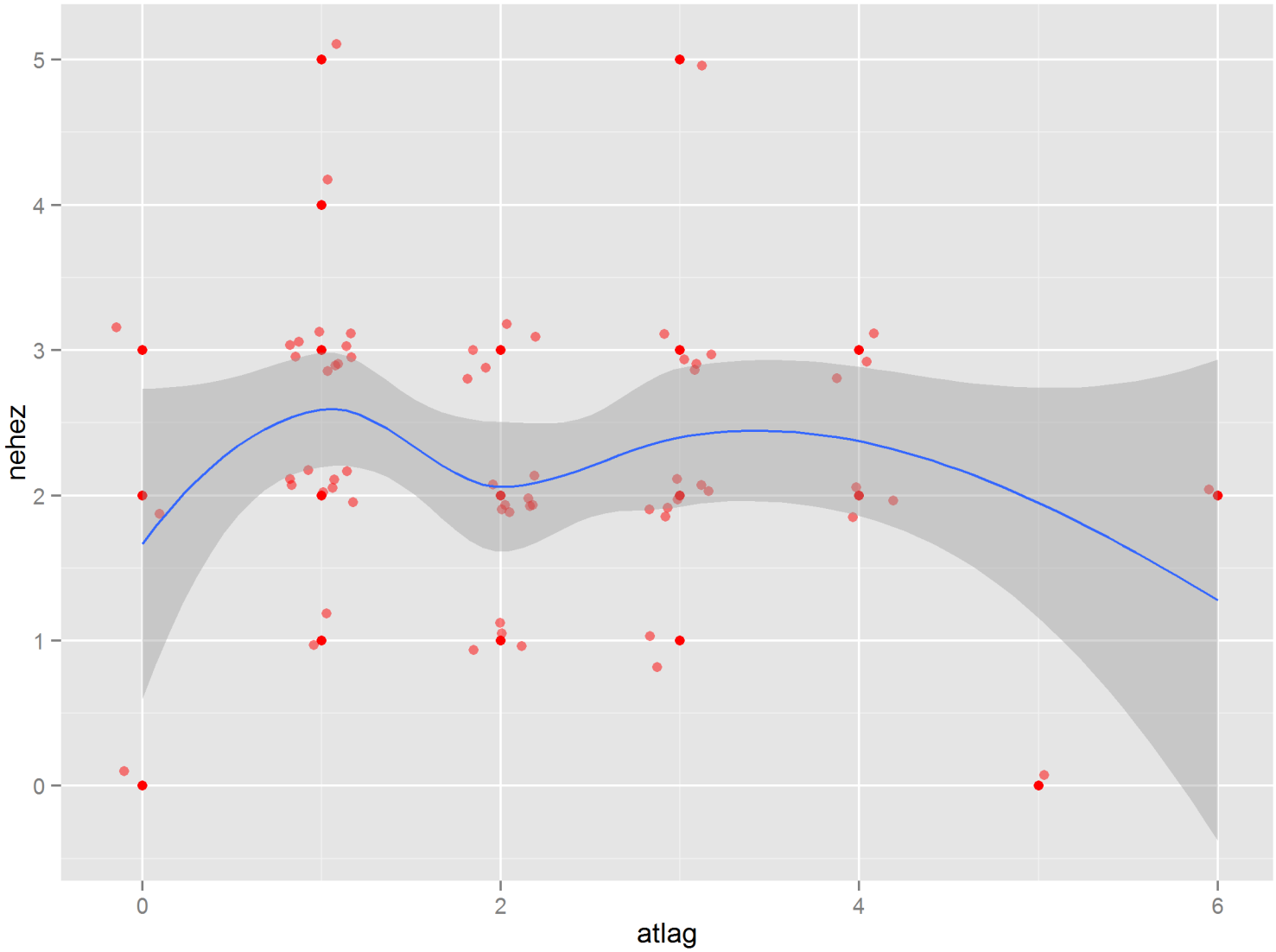
Hol volt, hol nem volt...



Szétszórjuk



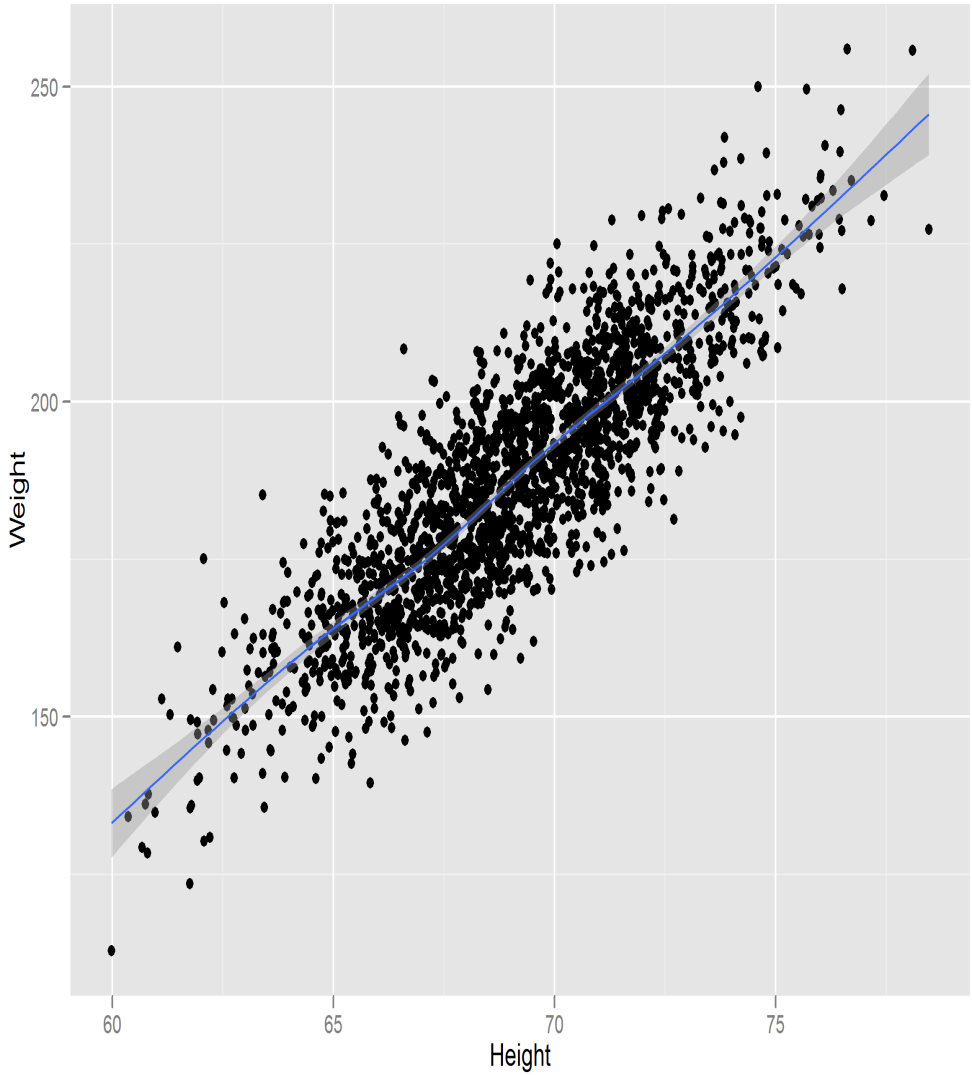
És megpróbáljuk közelíteni...



Regresszió

f függvény,

- bemenet:
  az attribútumok értéke,
- kimenet: megfigyelések legjobb közelítése
- „ökölszabály”
- Példa:
  testtömeg/magasság együttes eloszlás valójában egyenesre illeszthető,



Regressziós módszerek

- Alapelv:

$$Y_t = f(\bullet) + \varepsilon_t$$

Véletlen változó

Közelítés

Hiba

Jósolt esemény

$$Y = f(X_1, X_2, \ldots, X_n)$$

Megfigyelhető változók

- Átlagos hiba (mean error)

Becsült érték

$$ME = \frac{\sum_{t=1}^{n} (Y_t - F_t)}{n}$$

Mért érték

Lineáris regresszió

- Egyszerű lin. függvény illesztése az adatokra
  - nem vár alapvető változást a rendszer viselkedésében

$$Y = a + bX$$

- Legkisebb négyzetek módszere
  - keressük azokat az *a,b* paramétereket (itt: a: eltolás, b: meredekség), amelyekre

$$SSE = \sum_{t=1}^{n} \varepsilon_t^{\,2} = \sum_{t=1}^{n} \left(Y_t - F_t\right)^2$$ minimális (Sum of Squared Errors)

- cél: $$\sum_{t=1}^{n} \left(Y_t - F_t\right)^2 = \sum_{t=1}^{n} \left[Y_t - \left(a + bX_t\right)\right]^2$$

Levezetés (parc. deriválás)

$$\frac{d\sum_{t=1}^{n}\left[Y_t-\left(a+bX_t\right)\right]^2}{da}=\sum_{t=1}^{n}(-2)\left[Y_t-\left(a+bX_t\right)\right]=0$$

$$na=\sum_{t=1}^{n}\left(Y_t-bX_t\right)$$

$$a=\overline{Y}-b\overline{X}$$

Xi, Yi a mért értékpárok (pl. idő, terhelés)

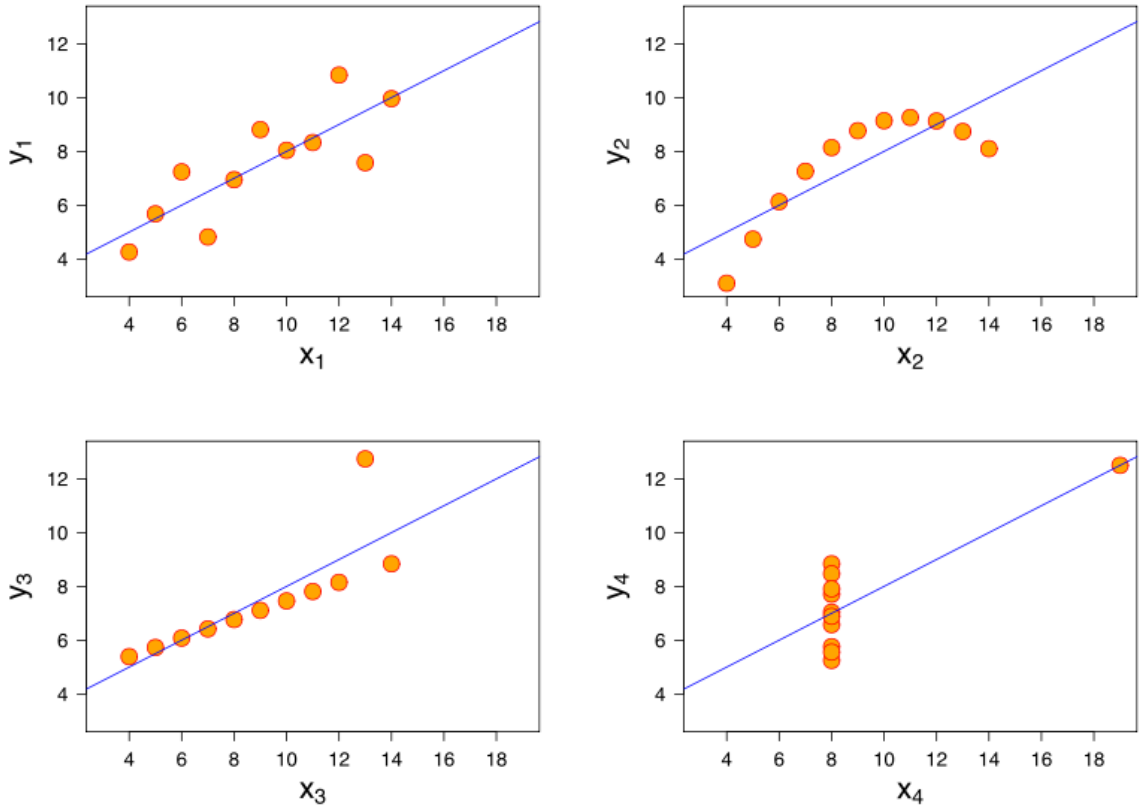
$$\frac{d\sum_{t=1}^{n}\left[Y_t-\left(a+bX_t\right)\right]^2}{db}=\sum_{t=1}^{n}X_t\left[Y_t-\left(a+bX_t\right)\right]=0$$

$$\sum_{t=1}^{n}X_t\left[Y_t-\frac{1}{n}\sum_{t=1}^{n}\left(Y_t-bX_t\right)-bX_t\right]=\sum_{t=1}^{n}X_tY_t-\frac{1}{n}\left(\sum_{t=1}^{n}X_t\right)\left(\sum_{t=1}^{n}Y_t\right)+\frac{1}{n}b\left(\sum_{t=1}^{n}X_t\right)\left(\sum_{t=1}^{n}X_t\right)-b\sum_{t=1}^{n}X_t^2=0$$

$$b=\frac{n\sum_{t=1}^{n}X_tY_t-\left(\sum_{t=1}^{n}X_t\right)\left(\sum_{t=1}^{n}Y_t\right)}{n\sum_{t=1}^{n}X_t^2-\left(\sum_{t=1}^{n}X_t\right)^2}$$

Lineáris regresszió

- Legjobban illeszkedő egyenes
- $min(\sum_{i=1}^{n}|Y_i - \hat{\mu}(x_i)|^2)$, ahol $\hat{\mu}(x) = ax + b$
- **DE:**
  Anscombe's quartet
  - **Minőségileg különböző adatok**
  - **Azonos regressziós egyenes**



Lineáris regresszió (folyt.)

- Korrelációs együttható (négyzete)
  - a változó becsült és tényleges értékének kapcsolata
  - 0 és 1 közti érték
  - 0: nincs kapcsolat
  - 1: függvényszerű kapcsolat

$$R^2 = \frac{\sum_{t=1}^{n}\left(F_t - \overline{Y}\right)^2}{\sum_{t=1}^{n}\left(Y_t - \overline{Y}\right)^2}$$

- Példa: E-mail szolgáltatás, 8 hétig mérjük a csúcsterhelést.

Hét	1	2	3	4	5	6	7	8
Max. terhelés (email/perc)	420	410	437	467	448	460	507	514

Hogyan közelíthető a terhelés változása?

Mekkora a korrelációs együttható?

Lin. regresszió példa

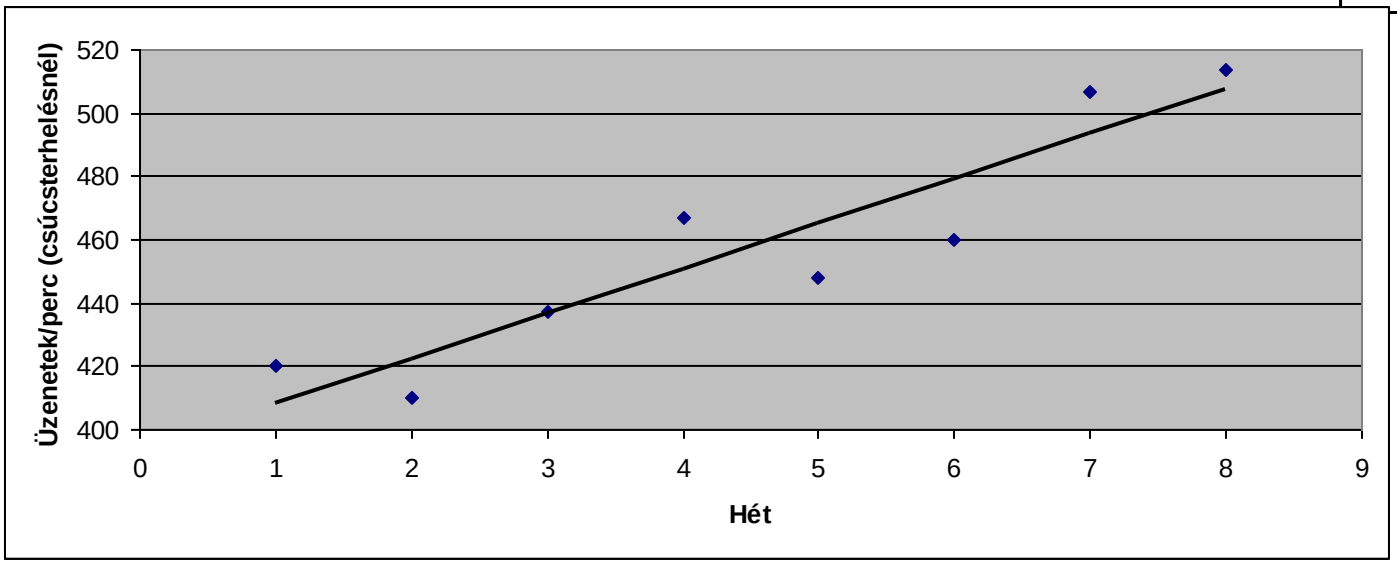
A legkisebb négyzetek módszerével

$Y=393.98+14.20X$

Korrelációs együttható:

$R^2=0.855$

Mért	Jósolt terhelés
420	408.18
410	422.38
437	436.58
467	450.78
448	464.98
460	479.18
507	493.38
514	507.58
	521.78



Két változó kapcsolatának vizsgálata

- Tfh. lineáris kapcsolat van az egyszerre bejelentkezett felhasználók száma és az elküldött emailek közt. (pl logok alapján)

Bejelentkezett felh. átlagos száma (1 óra alatt)	2450	2765	2241	2860	3011	2907	3209
Átl. terhelés (kimenő+bejövő emailek/óra)	19257	20488	18152	21450	21077	20639	22142

- Lineáris regressziós közelítés a legkisebb négyzetek módszerével:

ÜzenetekSzáma = f(*BejelentkezettFelhasználók*)

Y=9480.48 + 3.95X, R^2=0.937 erős kapcsolat

Nemlineáris módszerek

- Exponenciális közelítés:
  - jól illik a Web forgalom növekedéséhez
- Átalakítjuk a függvényt:

$$Y_t = a \times b^t$$

$$\log Y_t = \log a + t \log b$$

$$\log Yt = Y', \log a = a', \log b = b'$$

$$Y' = a' + b't$$

- Legkisebb négyzetek módszere használható
- Pl. adottak a legnagyobb mért terhelés értékei

Mennyi a várható legnagyobb terhelés az év végén?

Hónap	1	2	3	4	5	6	7	8	9	10
Max. kérések/sec (Y_t)	1035	1100	1160	1250	1350	1555	1770	1950	2210	2630
ln (Y_t)	6.942	7.003	7.056	7.13	7.207	7.349	7.478	7.575	7.7	7.874

Exp. terhelés példa

- Becslőfüggvény: $Y_t = a \times e^{bt}$
- Legkisebb négyzetek módszere a lineáris függvényre

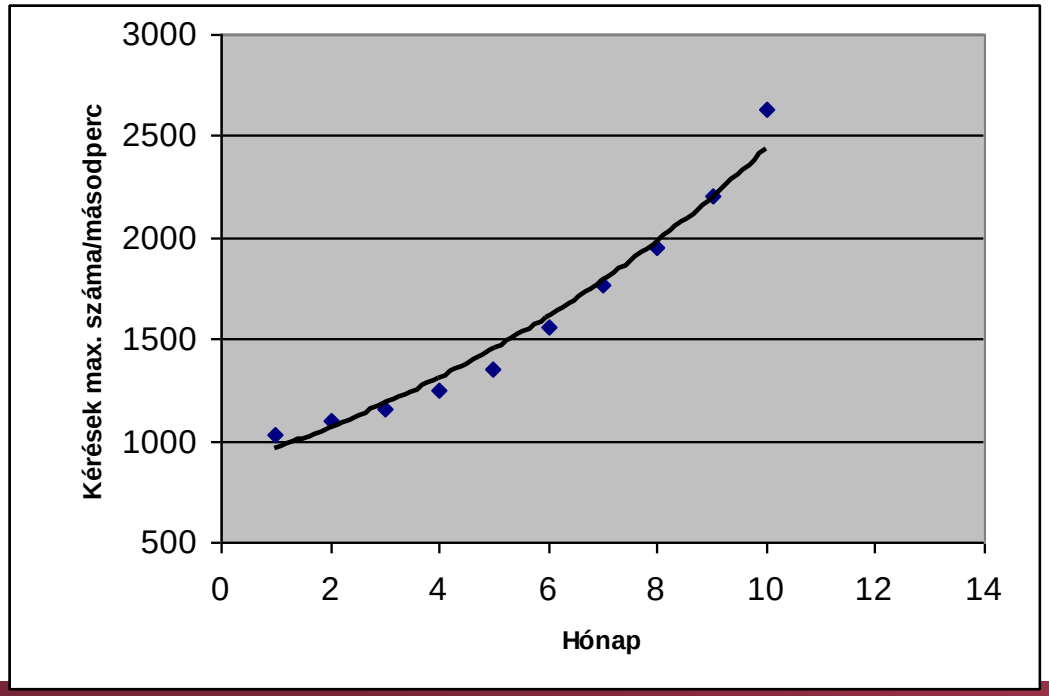
$$Y' = a' + b't, a' = 6.717, b' = 0.110, a = e^{a'}$$

- Eredmény:

$$Y_t = 826.33 \times e^{0.11t}$$

- 12. hónap:

Y_t=3093.3



Mozgó átlagok módszere

- Rövid távú előrejelzésre jó
- Egyszerre egy értéket ad meg
- A becsült érték az utolsó *n* érték átlaga

$$F_{t+1} = \frac{\sum_{i=t}^{t-n+1} Y_i}{n}$$

ahol Y_t a t. időpontban mért érték

F_{t+1} a becsült érték

n tipikusan 3 és 10 között van

(becslés hibája ne legyen túl nagy)

Exponenciális csúszóablak

- Egy értéket ad meg, az előző méréseket átlagolva
- Későbbi mérés nagyobb súllyal
- Súlyozza a mérési hibát
- Rövid távú előrejelzésre

$$F_{t+1} = F_t + \alpha\left(Y_t - F_t\right)$$

ahol F_t :a t. időpontra becsült érték

Y_t : t. időpontban mért érték

$Y_t - F_t$: mérési hiba a t. periódusban

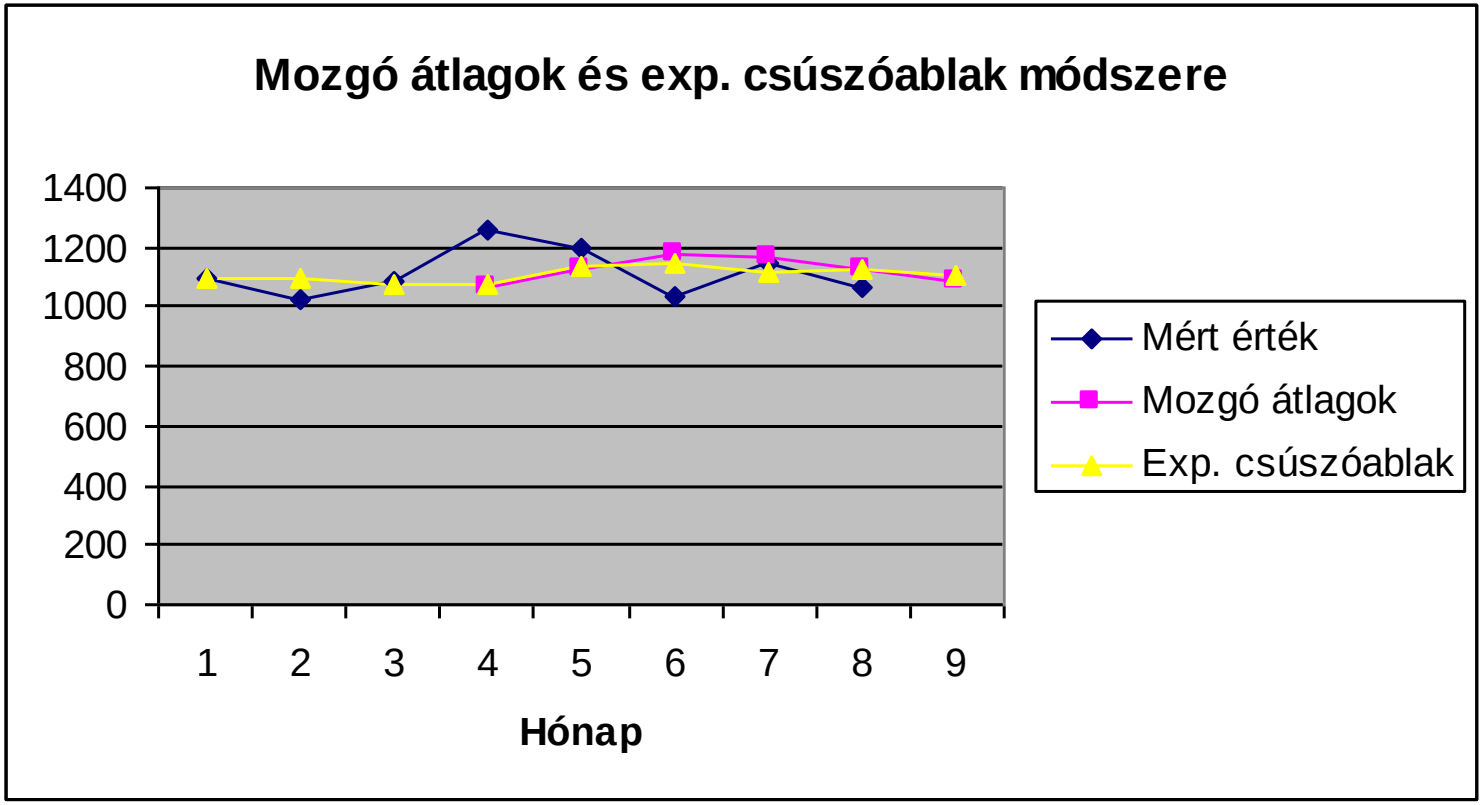
α : súlyozás ($0 \leq \alpha \leq 1$, gyakorlatban $0.05 \leq \alpha \leq 0.3$)

A két módszer összehasonlítása

- Adott sávszélesség igények, a két módszerel becsüljük a következő értéket

Hónap	Sávszélesség igény	Mozgó átlagok módszere (n=3)	Exp. csúszóablak (α = 0.3)
1	1100		1100.00
2	1020		1100.00
3	1090		1076.00
4	1255	1070.0000	1080.20
5	1195	1121.6667	1132.64
6	1039	1180.0000	1151.35
7	1145	1163.0000	1117.64
8	1066	1126.3333	1125.85
9		1083.3333	1107.90

A két módszer összehasonlítása



BENCHMARKING

Why benchmark?

Benchmarking

- Célok: szoftver/hardver eszközök teljesítményének összehasonlítása
  - Döntéstámogatás
    - melyiket előnyösebb megvenni/telepíteni
    - mekkora terhelésre elég a meglévő rendszer
  - Teljesítménytesztelés
    - kell-e még teljesítményen javítani és hol (fejlesztésnél)
    - optimális-e egy konkrét beállítás
    - van-e egy beállításnak teljesítményre gyakorolt hatása

Elvárások

- Ismételhetőség
  - Repeatability
- Reprodukálhatóság
  - Reproducilibity
- Relevancia
- Szabványok/megállapodások betartása
- Általánosított felhasználói eset
  - Átlag felhasználó számára értelmezhető legyen az eredmény

Benchmark terhelési modellek

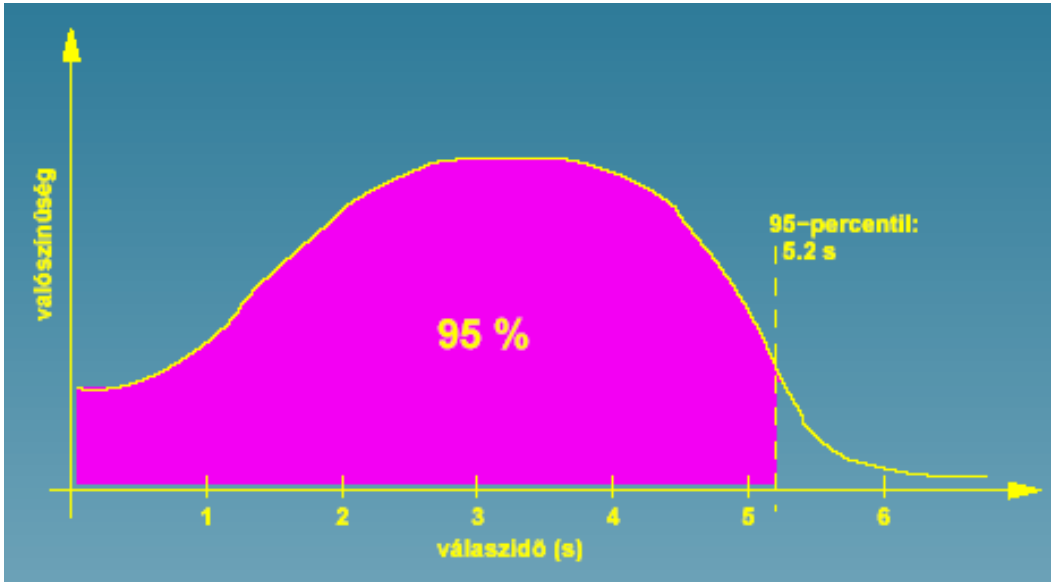
- Tudományos/műszaki rendszerek
  - nagy mennyiségű adat feldolgozása (number crunching)
  - párhuzamos módszerek
- Tranzakciókezelés (OLTP)
  - kliens-szerver környezet
  - sok gyors, párhuzamos tranzakció
- Batch jellegű adatfeldolgozás
  - riport készítés nagy mennyiségű adatból
- Döntéstámogatás
  - kevés, bonyolult lekérdezés
  - ad hoc műveletek
  - sok adat (pl. OLAP)
- Virtualizáció

Mérendő paraméterek (Metrikák)

- Futási idő
  - kezdet, vég?
  - eloszlás
  - CPU, I/O, hálózat,…
- Tranzakciósebesség
  - rendszer reakcióideje
  - akár egymásba ágyazott tranzakciók
- Áteresztőképesség
  - ***feldolgozott*** adatmennyiség / futási idő
  - terhelés függvényében

Mérendő paraméterek (2)

- Válaszidő
  - terhelés függvényében
    - felhasználók
    - tranzakciók száma, stb.
- X-Percentil
  - Egy adott halmaz X százaléka ez alatt az érték alatt van



Benchmark elvégzése

- Relevancia biztosítása
  - Tényleg azt az alkalmazást mérjük, amit kell
  - Terhelésgenerálás jellege közelítse a valódi terhelést
  - Minimalizáljuk a zavaró tényezőket

SPEC benchmarkok

- http://www.spec.org/benchmarks.html
  - Standard Performance Evaluation Corp.
- Erőforrás és alkalmazás szintű benchmarkok
  - CPU
  - Alkalmazások
  - Levelező szerverek
  - Web szerverek stb.
- Benchmark: megrendelhető szolgáltatás

SPEC CPU2006

- CPU intenzív
- CINT2006
  - Számításigényes egész számos
- CFP2006
  - Lebegőpontos

- Példa:

cpu2006.html

CINT2006 és CFP2006 terhelésgenerátorok

- CINT2006 :

400.perlbench	C	Programming Language
401.bzip2	C	Compression
403.gcc	C	C Compiler
429.mcf	C	Combinatorial Optimization
445.gobmk	C	Artificial Intelligence
456.hmmer	C	Search Gene Sequence
458.sjeng	C	Artificial Intelligence
462.libquantum	C	Physics / Quantum Computing
464.h264ref	C	Video Compression
471.omnetpp	C++	Discrete Event Simulation
473.astar	C++	Path-finding Algorithms
483.xalancbmk	C++	XML Processing

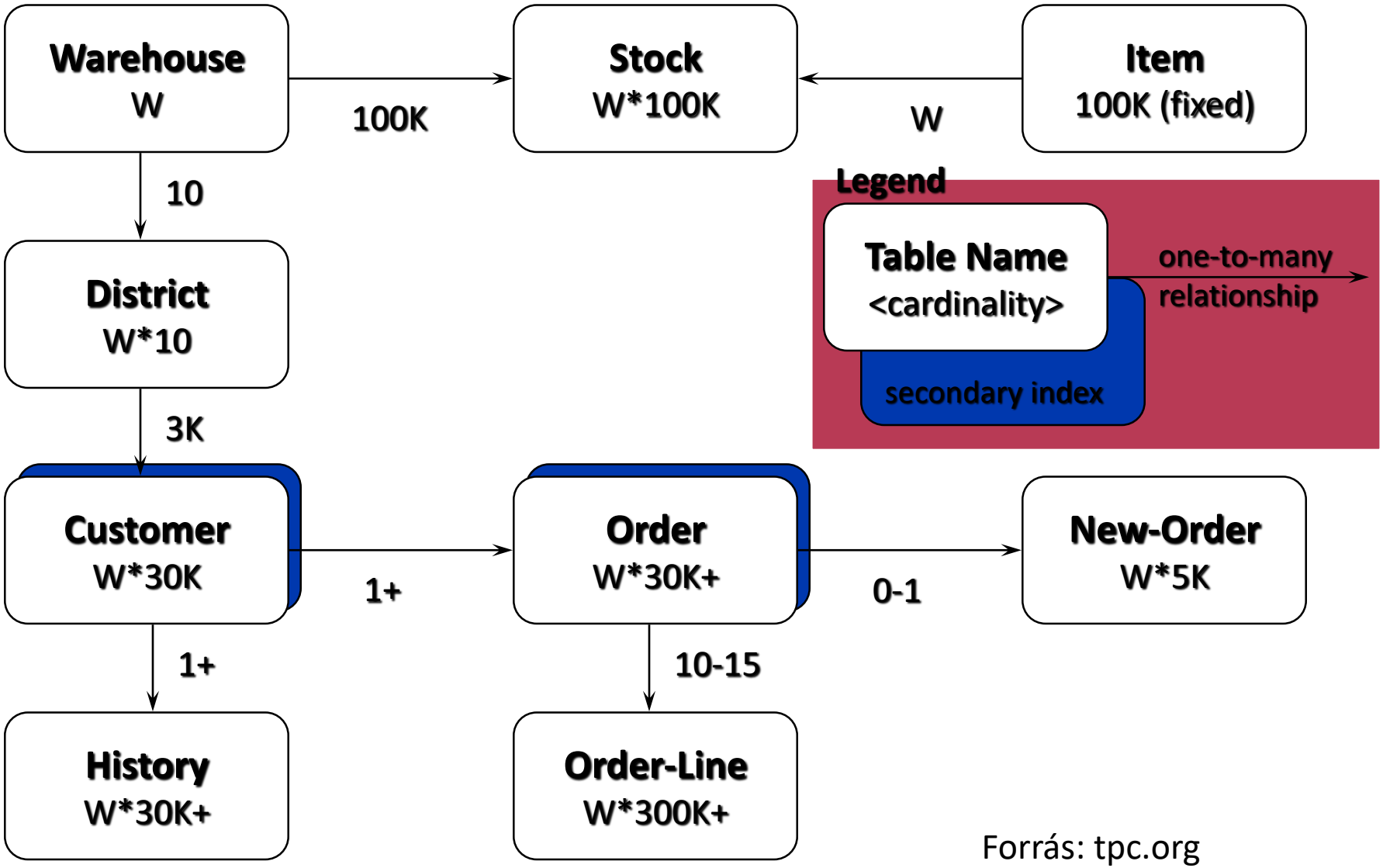
- CFP2006:

SPEC CFP2006 Benchmark Descriptions.htm

A TPC benchmark

- Adatbáziskezelő rendszerek mérése
  - RDBMS+OS+HW
- Mérési környezet
  - Mintaadatbázis: Ügyfelek és megrendelések
  - 5 fajta tranzakció (lekérdezés/módosítás) vegyesen
  - Felső korlát a futási időre
  - Valós körülmények: ACID tranzakciók, felhasználói gondolkodási idők
- Mért adatok
  - Áteresztőképesség (tpmC)
  - „Hatékonyság” ($/tpmC)

TPC-C séma



Forrás: tpc.org

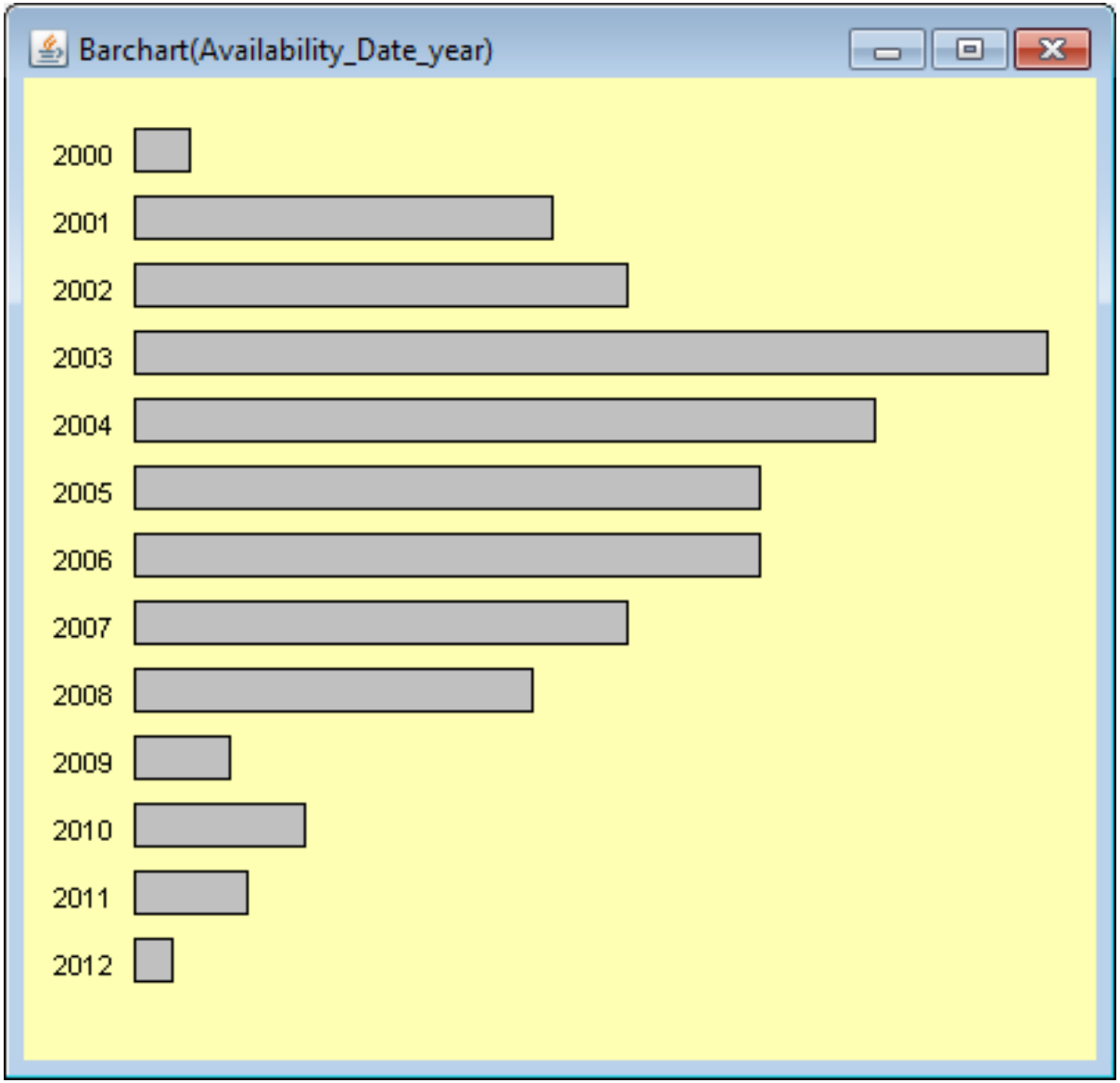
Mielőtt elemeznénk: tisztítunk

- A kiinduló adathalmazunk:

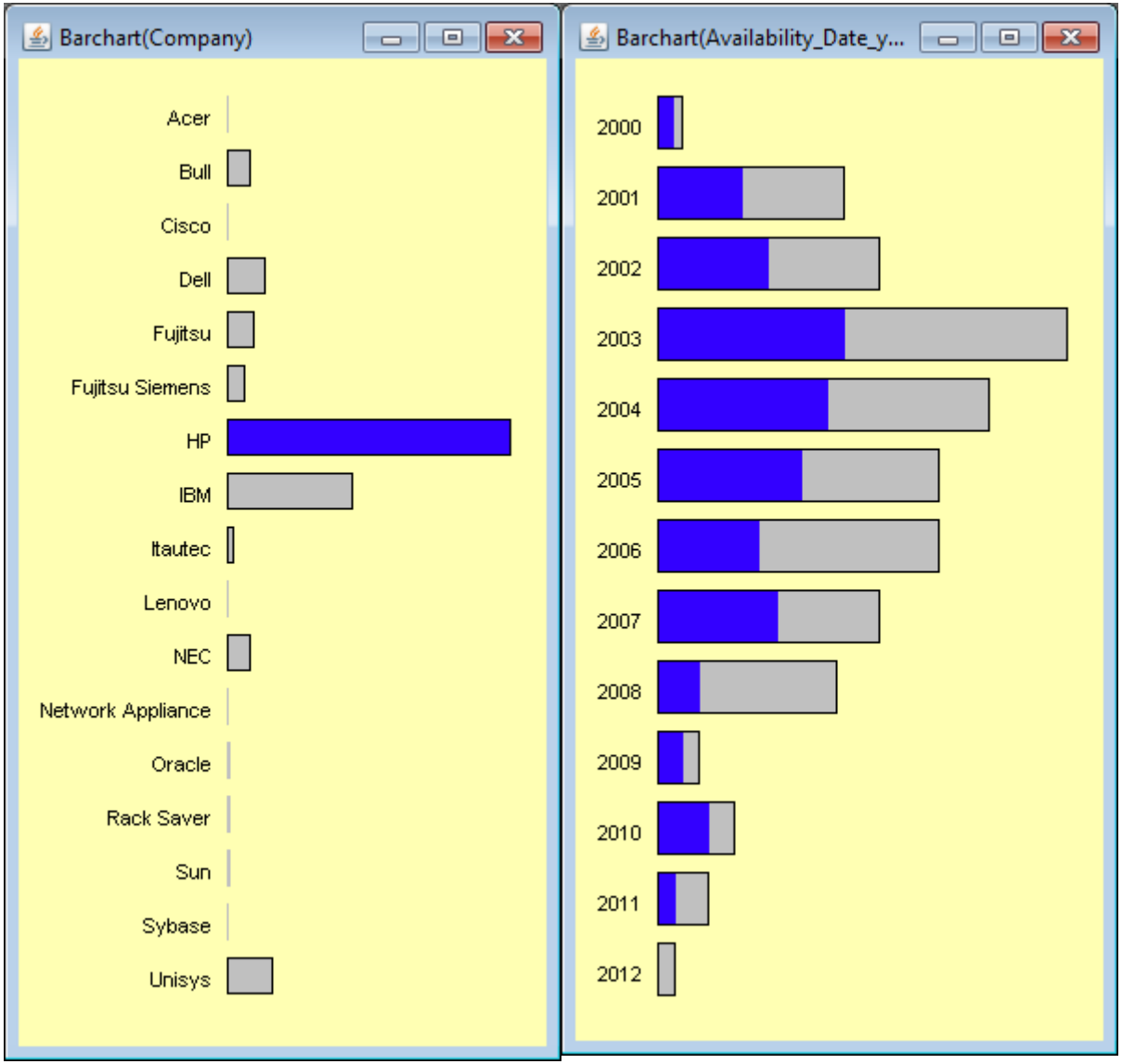
	A	B	C	D	E	F	G	H	I	J	K
1	TPC-C BENCHMARK RESULTS										
2	These results are valid as of date 6/12/2012 10:04:24 PM										
3											
4			TPC-C Results - Revision 5.X								
5											
6	Company	System	Spec. Revision	tpmC	Price/Perf	Total Sys. Cost	Currency	Database Software	Operating System	TP Monitor	Server CPU Type
7	Acer	Altos R710	5.5	66543	12.42	826507.55	AUD	Microsoft SQL Server	Microsoft Windows Serv	Microsoft CO	Intel Xeon - 3.6 GHz
8	Bull	Bull Escal	5.9	6085166	2.81	17127928	USD	IBM DB2 9.5	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 5.0
9	Bull	Bull Escal	5.9	629159	2.49	1566664	USD	IBM DB2 9.5 Enterpri	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.2
10	Bull	Bull Escal	5.8	1616162	3.54	5716286	USD	IBM DB2 9.1	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.7
11	Bull	Bull Escal	5.8	404462	3.51	1417121	USD	Oracle Database 10g	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.7

- Felesleges adatok:
  - Sorok (pl. a kezdő sorok, és az állomány végén lévő sorok, amelyek nem kapcsolódnak az eredményekhez)
  - Oszlopok (pl. Server CPU Type nekünk most nem kell)
  - Eltérő valutákban megadott költségek

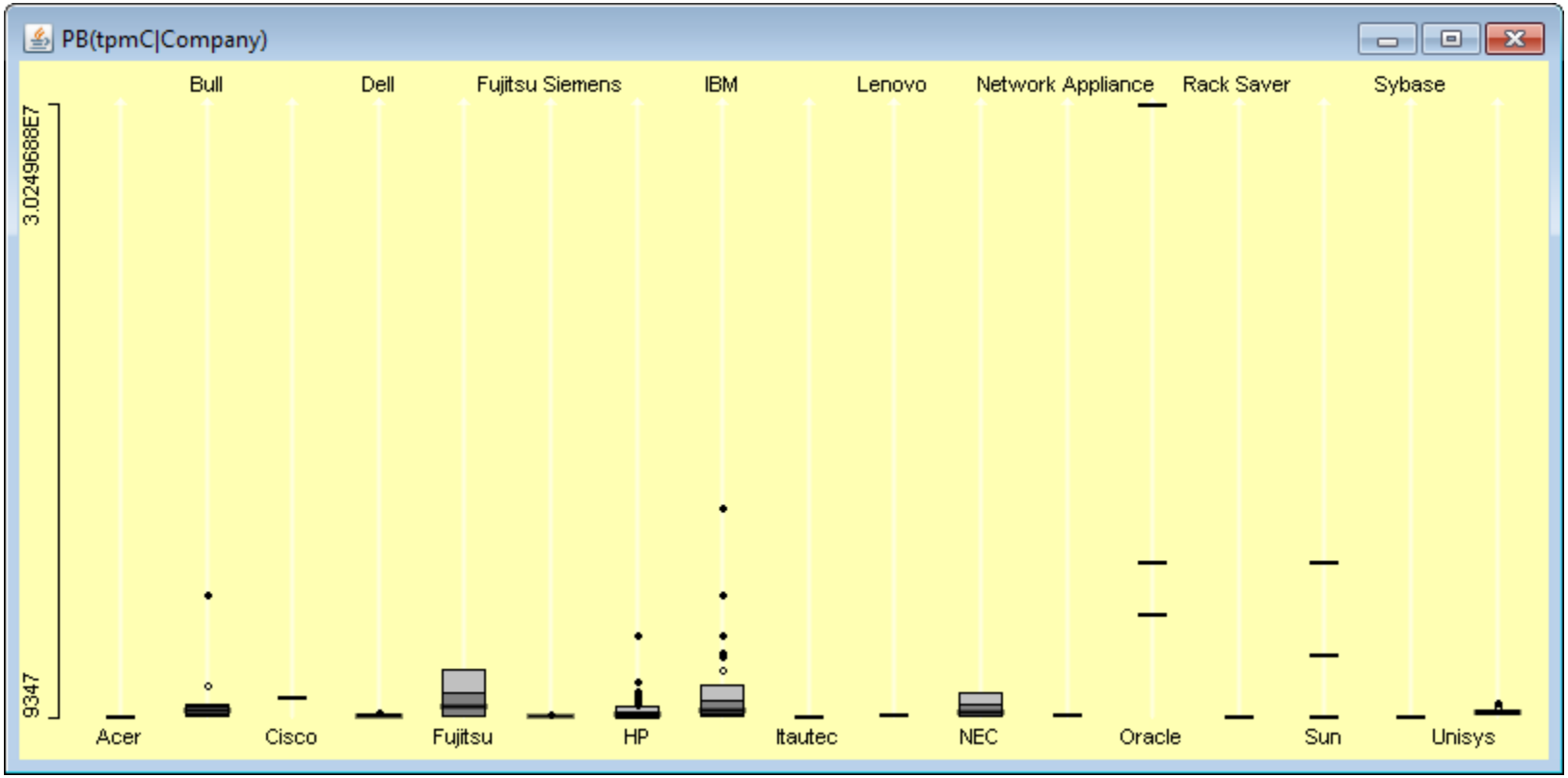
A benchmark mely évek eredményeit tartalmazza?



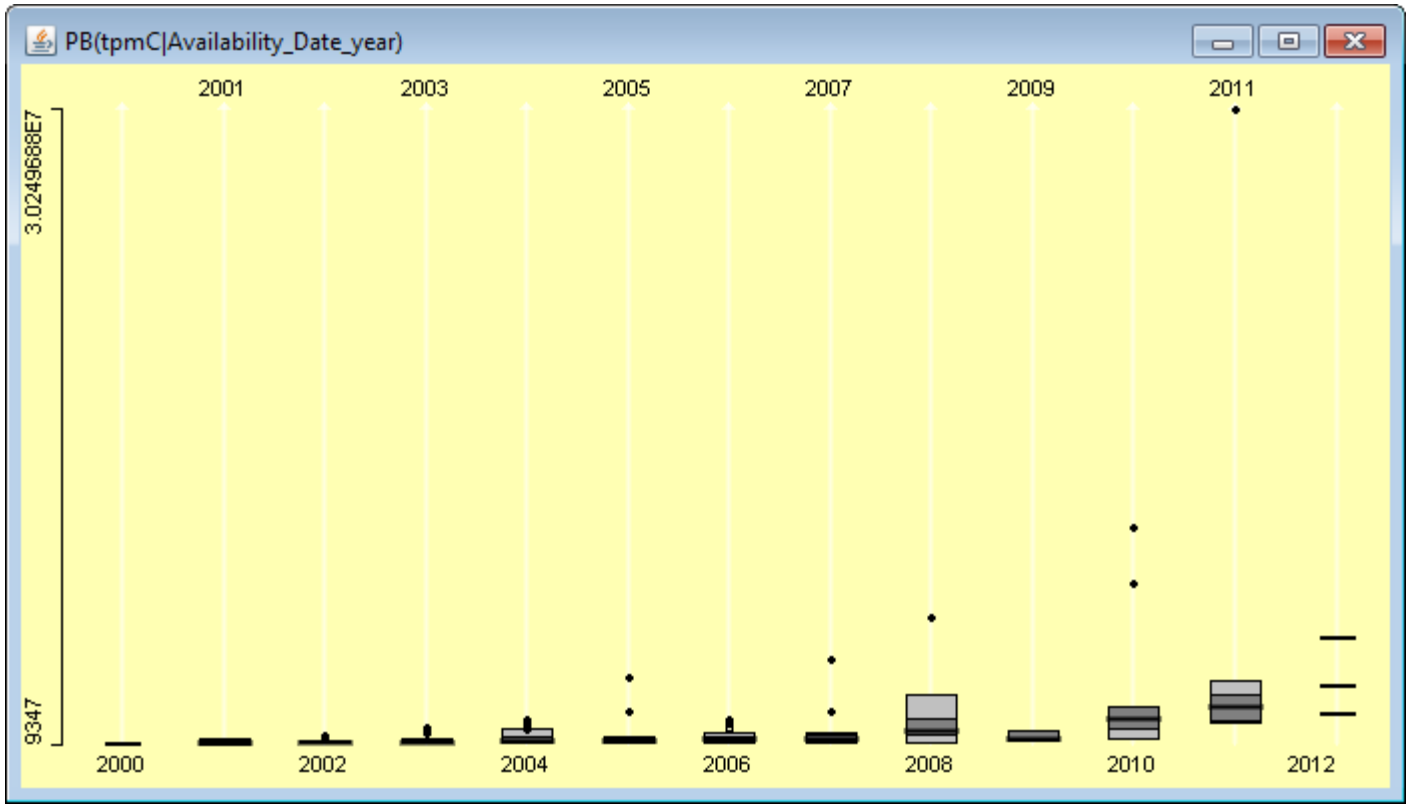
Mikor voltak aktívak a beszállítók?



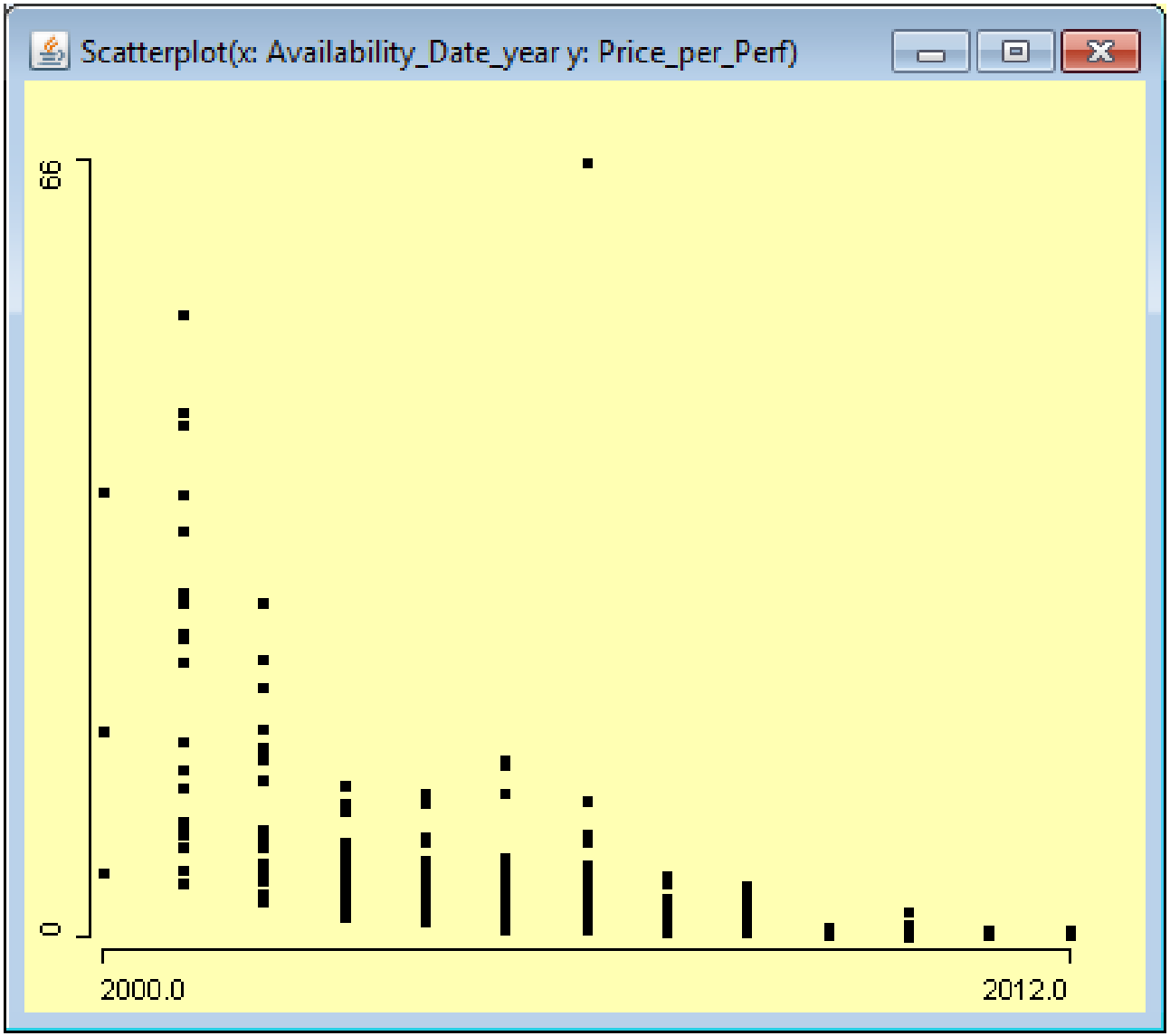
Hogyan alakulnak a teljesítmény mérőszámok?



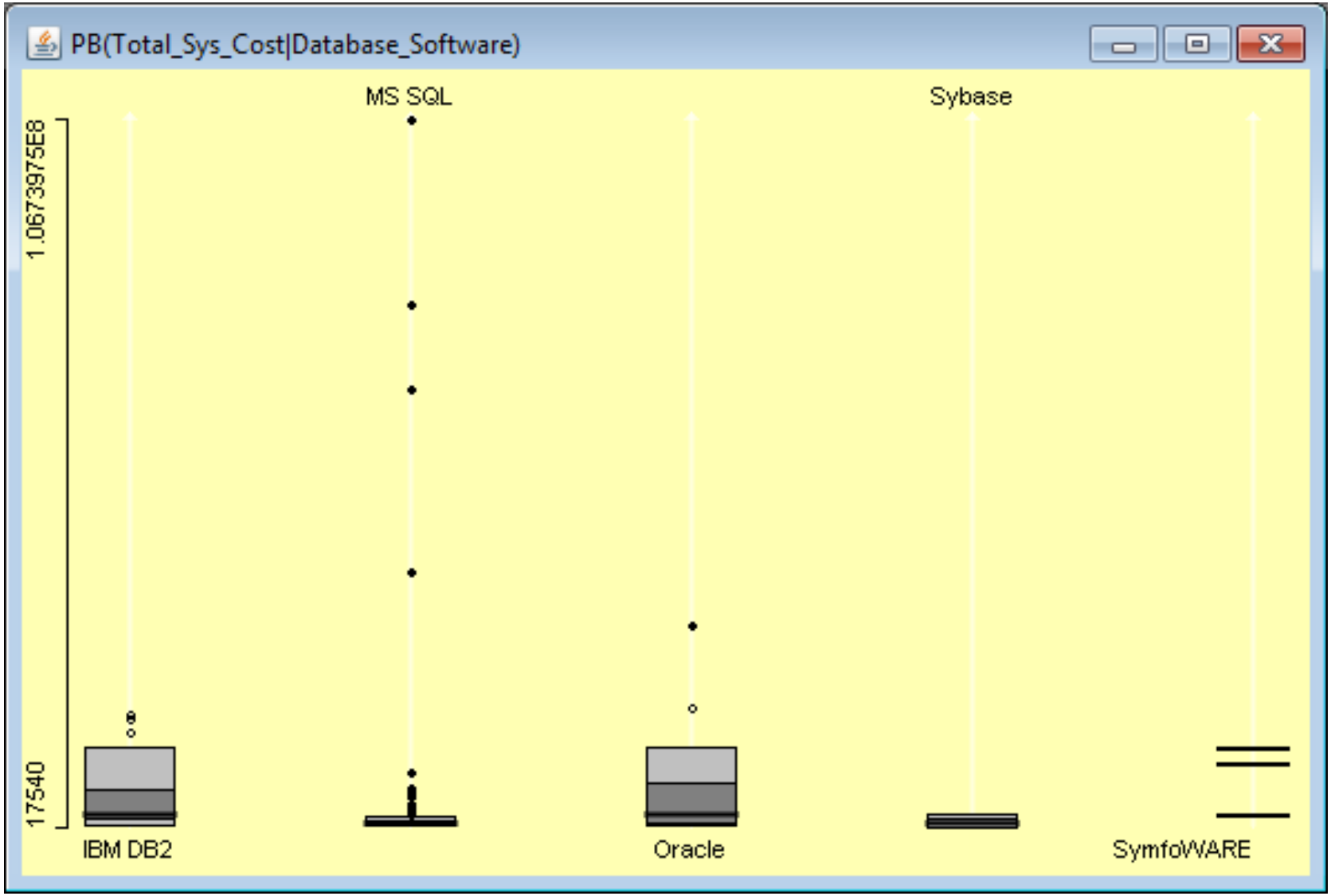
Időben a teljesítmény hogyan változott?



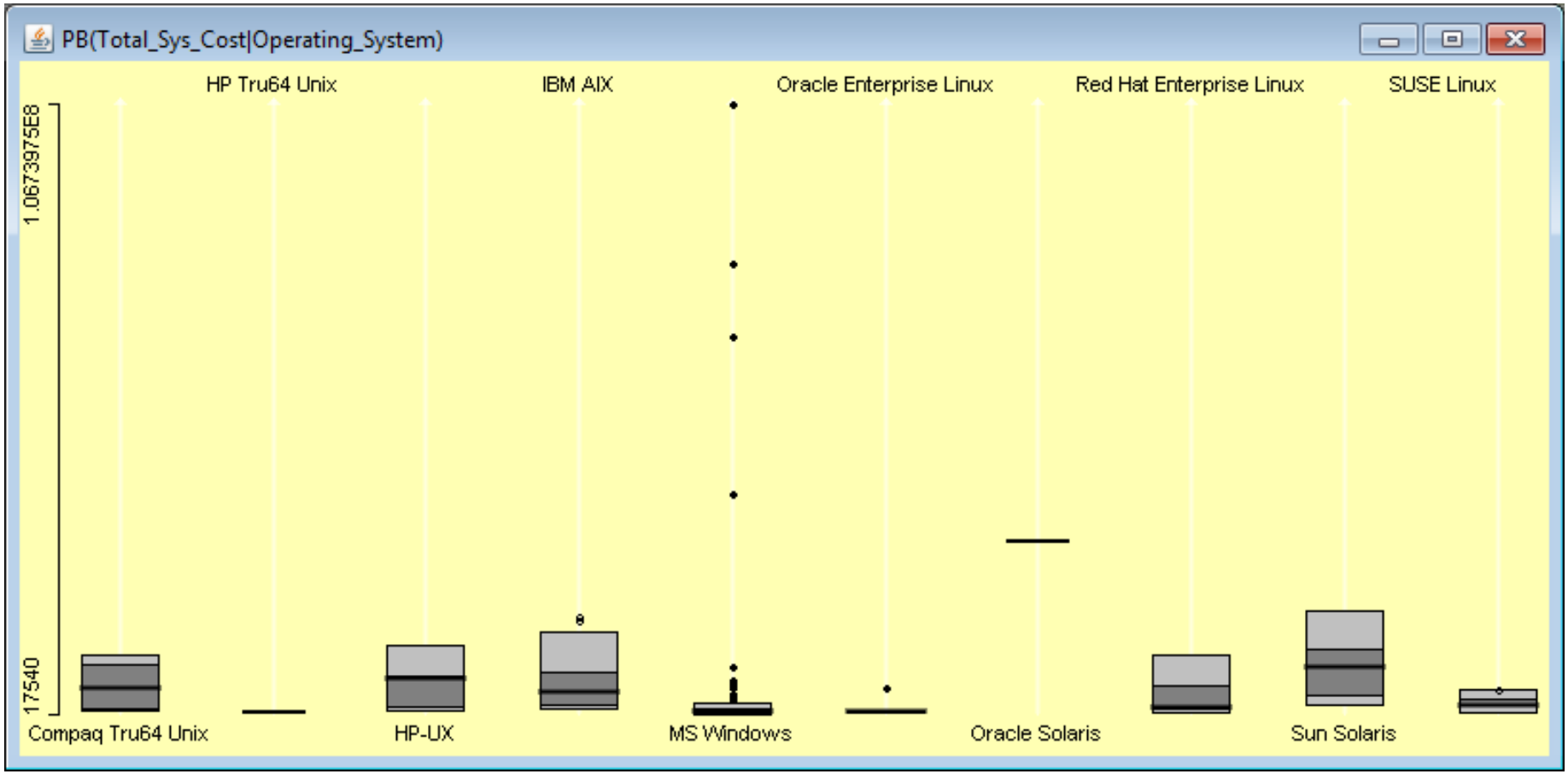
Hogyan alakultak mindeközben az árak?



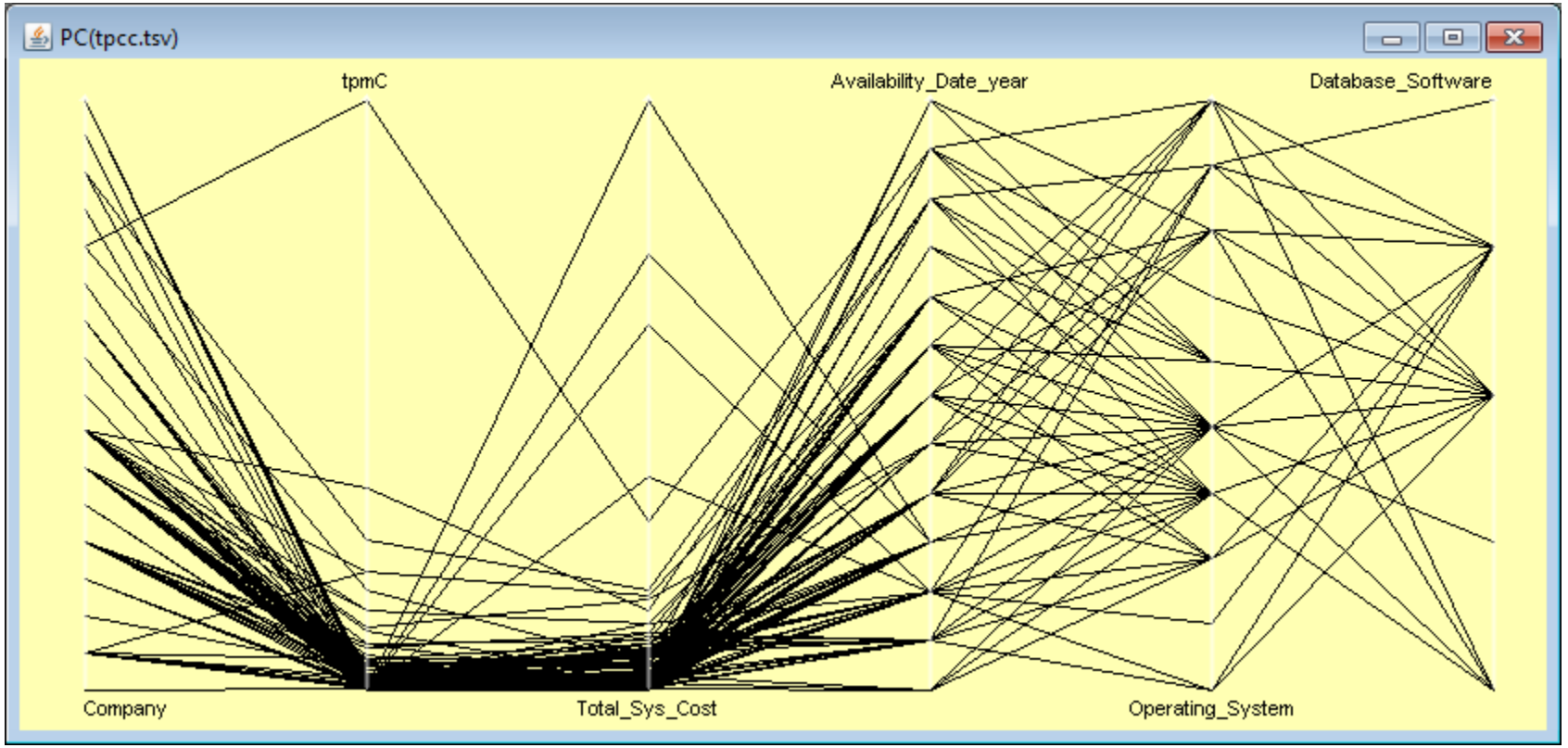
Milyen adatbázis-kezelő SW-t válasszunk?



Milyen OS-t válasszunk?



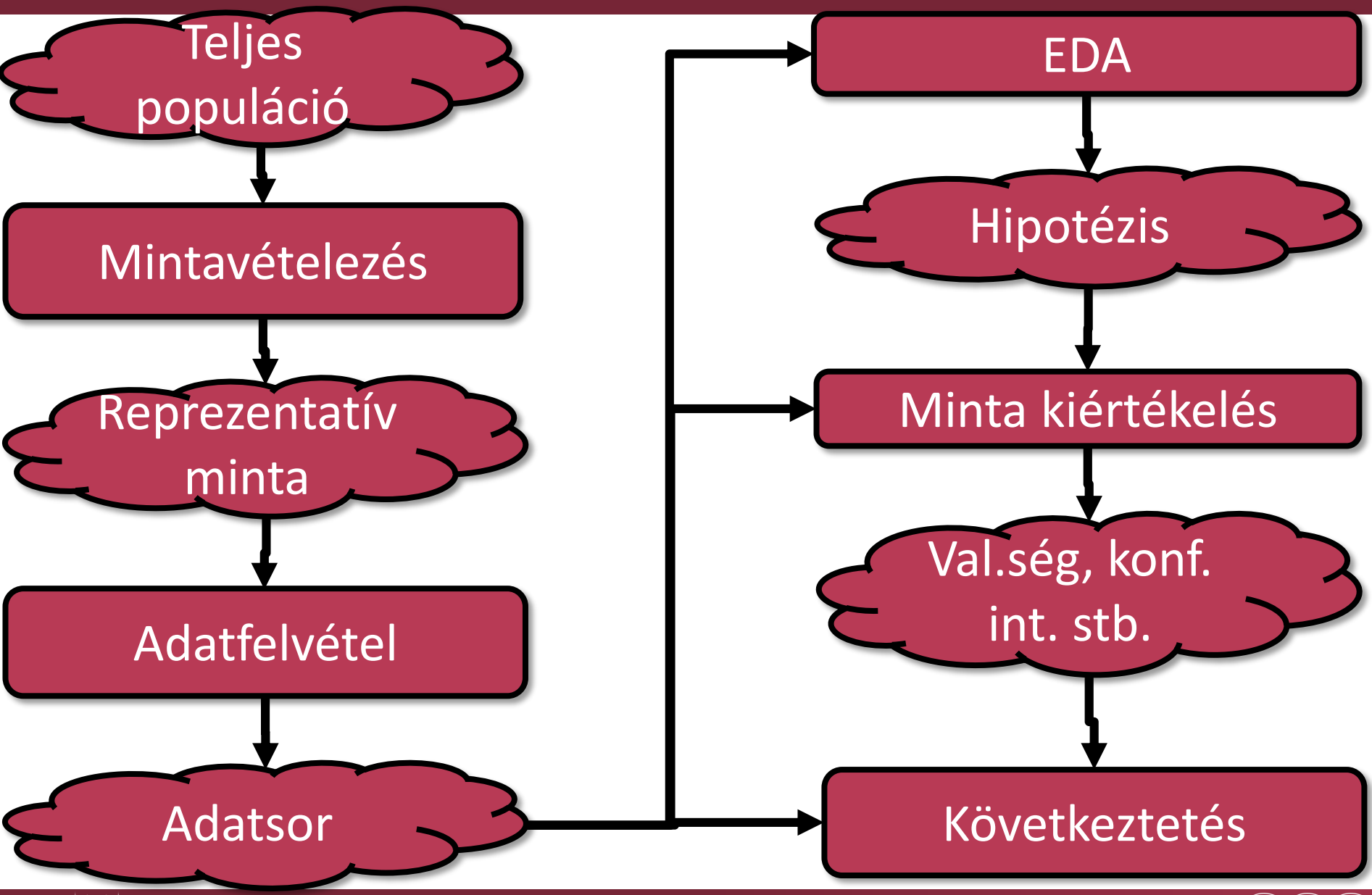
A „big picture”



KÖVETKEZTETŐ STATISZTIKA

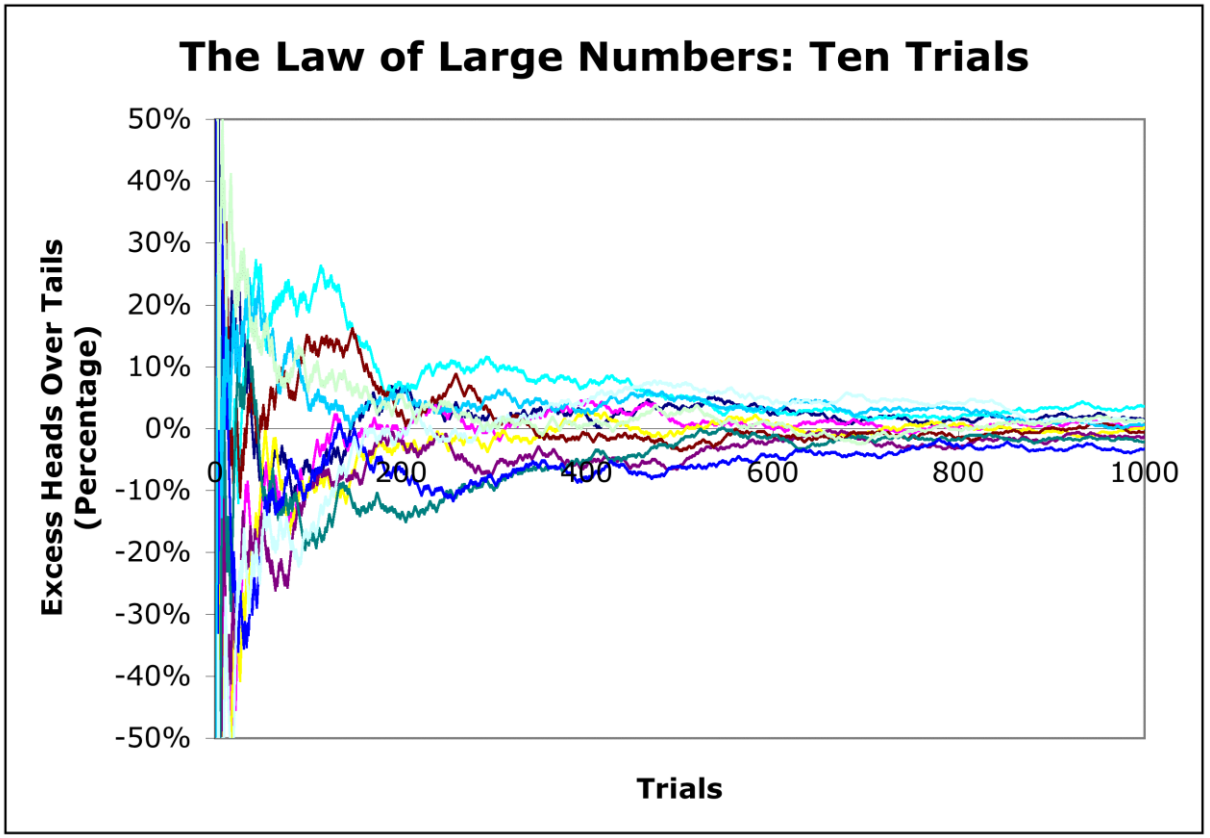
Következtető statisztika

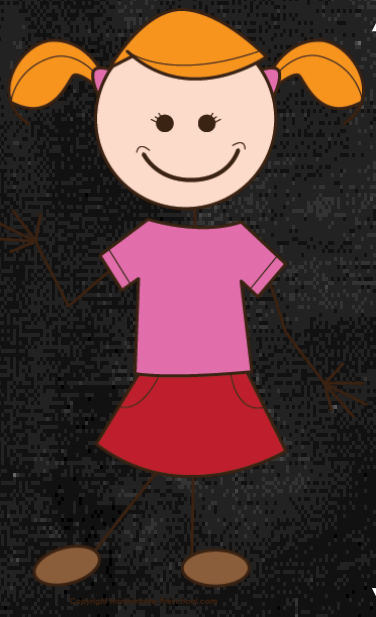
Következtető statisztika



Ökölszabályok

- LLN (Law of Large Numbers)
  - Ha a kísérletek száma tart a végtelenhez, az előfordulási gyakoriság az elméleti valószínűséghez konvergál





?

Magyarországi kamaszlányok

Békés

Heves

Vas

$\overline{x}_{Békés}$

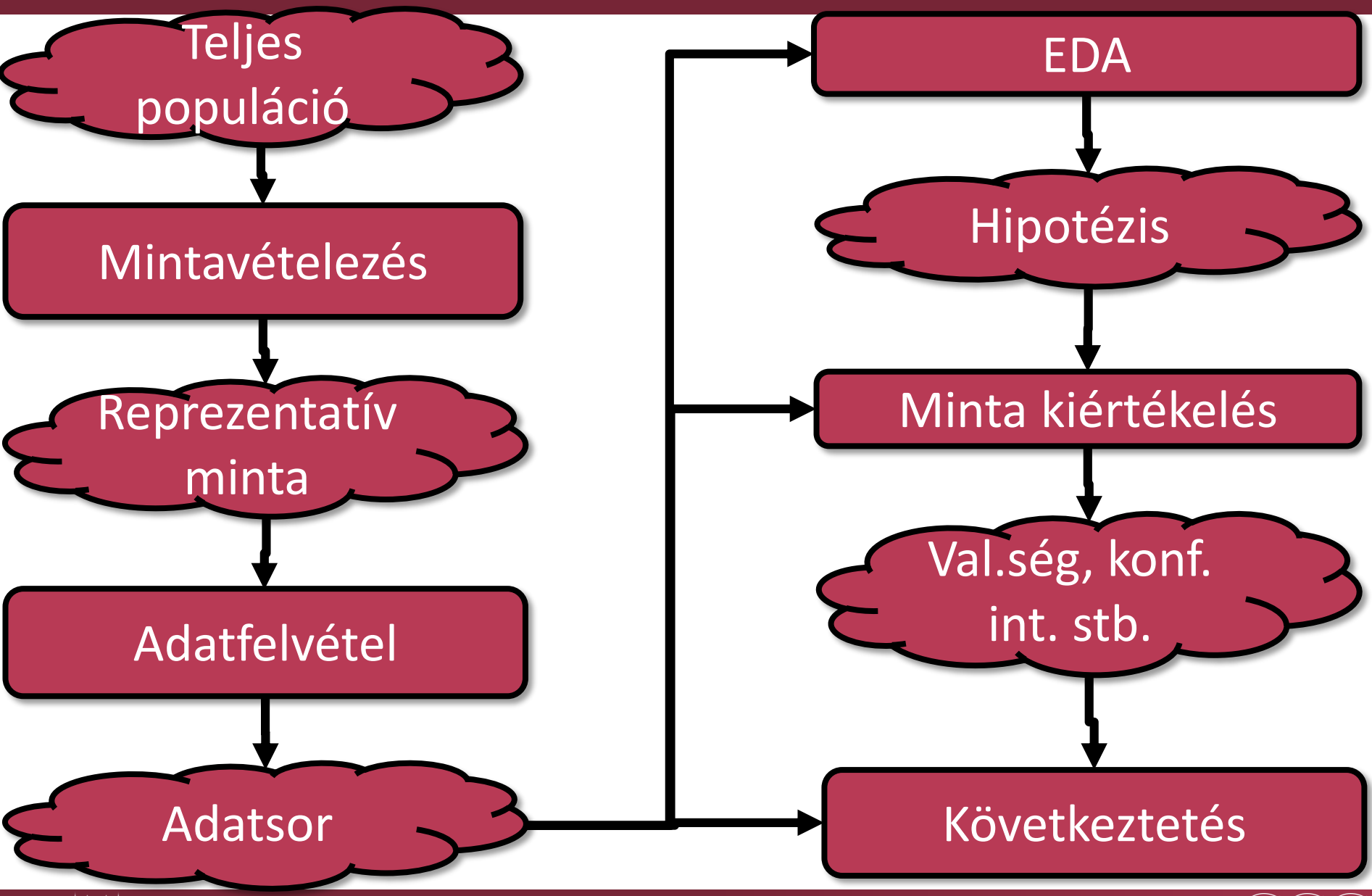
$\overline{x}_{Heves}$

$\overline{x}_{Vas}$

$$\mu = \frac{X_1 + X_2 + \ \ldots + X_N}{N}$$

$$\mu \approx mean(\overline{x})$$

Következtető statisztika



Következtetés

- Döntési bemenet
  - Valami küszöbérték
- Adatsor típusa
  - Megfigyelési tanulmány (observational study)
  - Kísérlet (experiment)

Különbség: a *köztes változók* eliminálása

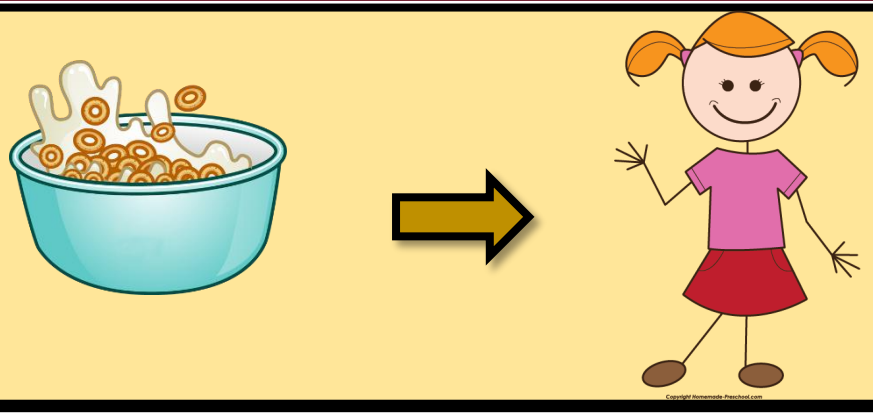
Esettanulmány

„Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast.„

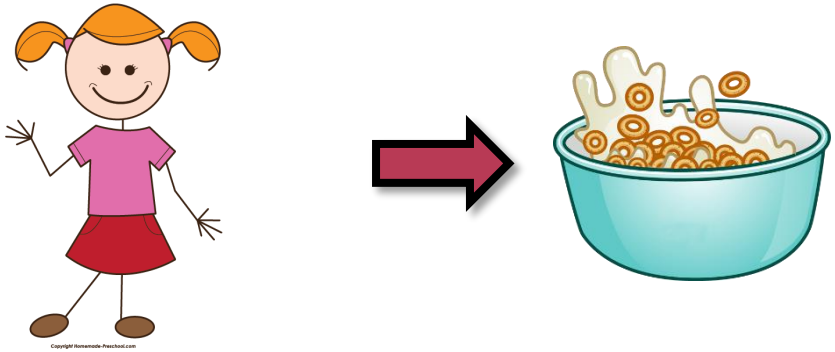
Forrás: http://usatoday30.usatoday.com/news/health/2005-09-08-cereal-slimming_x.htm

Esettanulmány

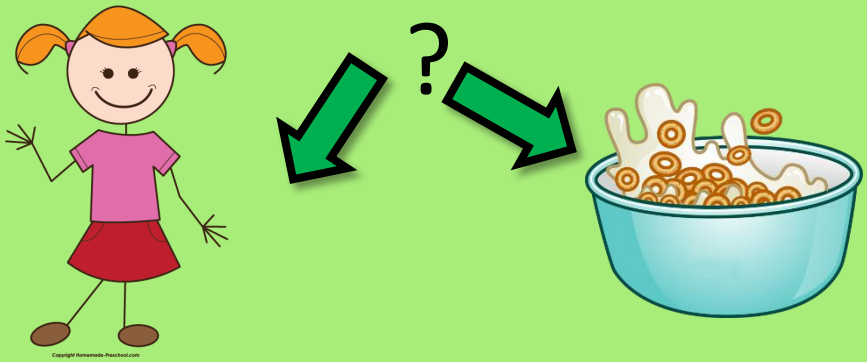
1. „Breakfast, cereal keep girls slim”



2. „Being slim causes girls to eat breakfast„



3. „A confounding variable is responsible for both”



Forrás: http://usatoday30.usatoday.com/news/health/2005-09-08-cereal-slimming_x.htm

Következtetés

- Döntési bemenet
  - Valami küszöbérték
- Adatsor típusa
  - Megfigyelési tanulmány (observational study)
    - A köztes változók kiléte bizonytalan
    - Csak korreláció, kauzális következtetések nem
  - Kísérlet (experiment)
    - A köztes változókat kiszűrtük (mintavételezés!)
    - Kauzális következtetések is

KÍSÉRLETTERVEZÉS ALAPFOGALMAK

Kísérlettervezés

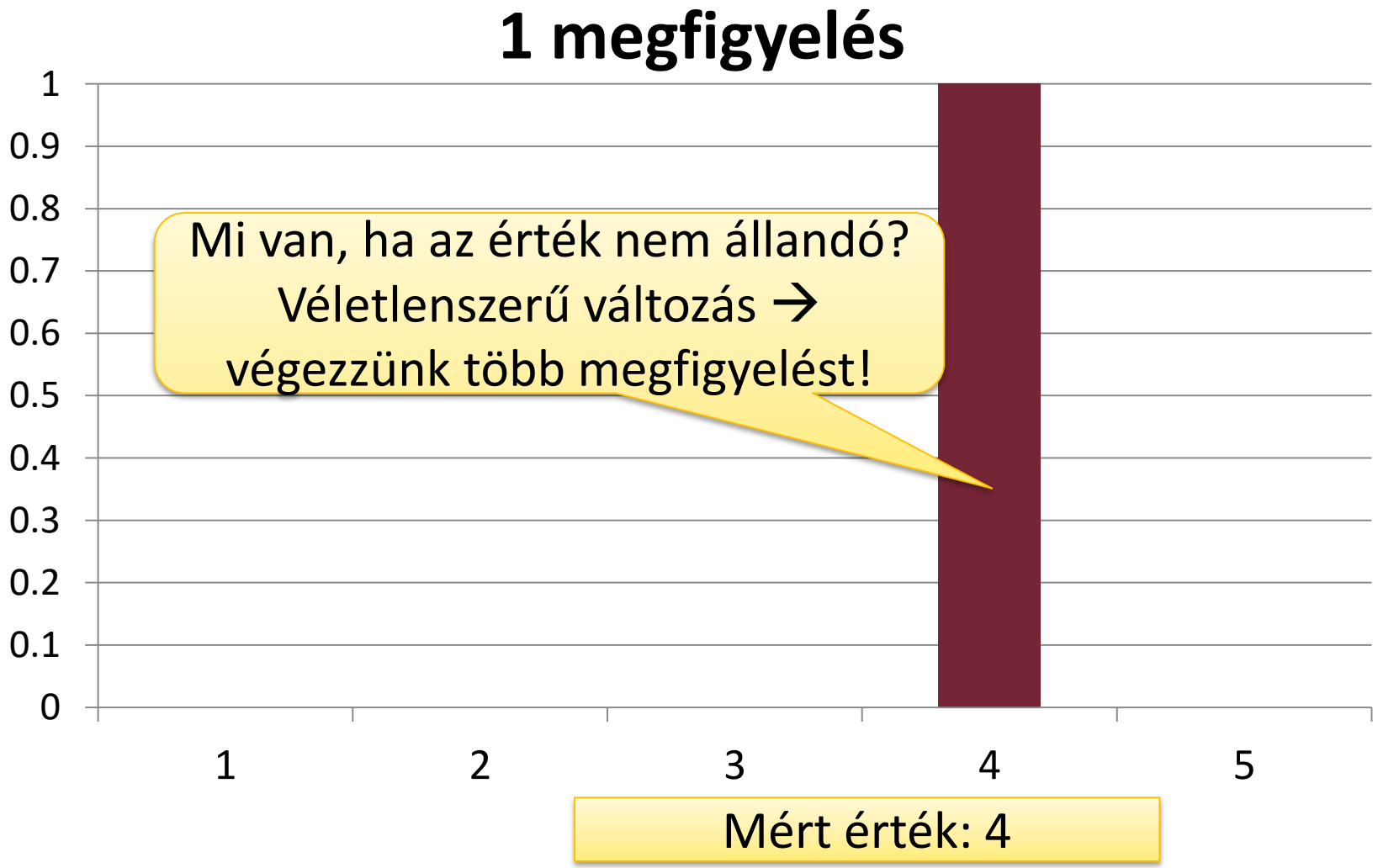
- Cél: a modell paraméterezése a valóság alapján
  - Vagy absztrakt modell a konkrét modell alapján
- Információt **kísérlet**ek révén szerzünk
  - Pontosan mit szeretnénk tudni?
  - Ehhez milyen megfigyelést, hányszor kell elvégezni?
  - A kapott eredményekből mire lehet következtetni?
- (statisztikai) **kísérlettervezés** (Design of Experiment, DOE)
  - hatékony eljárás a kísérletek tervezésére és elemzésére
  - valós és objektív konklúziók levonásához
- Kísérletterv: még a kísérlet elvégzése előtt

Kísérlettervezés

- Mire jó?
  - Alternatívák közötti választás
  - Érzékeny paraméterek, kulcsfaktorok
  - Megfelelő célérték, változékonyság csökkentése
  - Robosztussá tétel
- Fontos:
  - Világos cél, egyértelmű eredmények
  - Kis méret, alacsony költség
  - Valós viszonyok

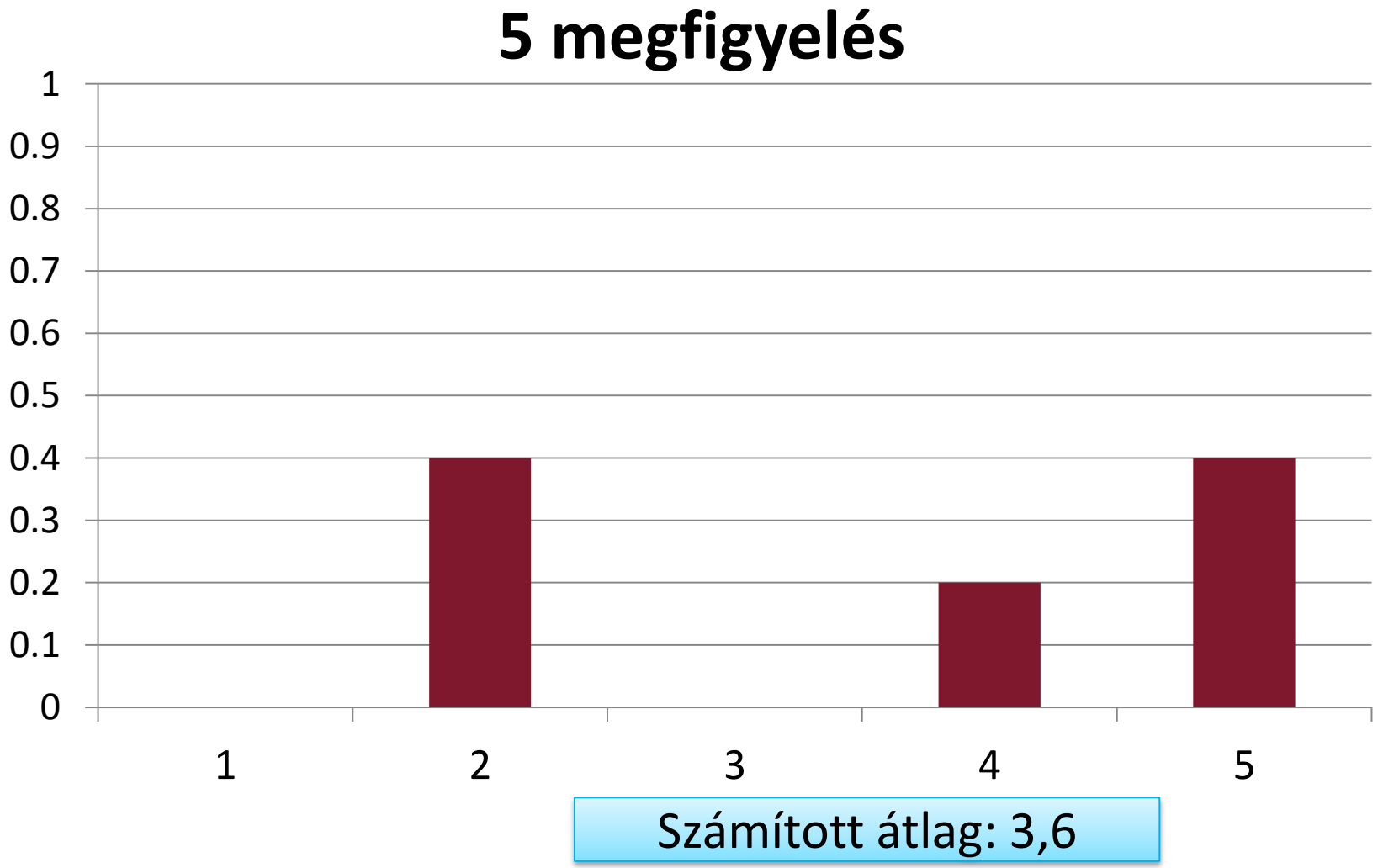
Számszerű jellemzők mérése

- Mérjük meg egy jellemző értékét! (1..5)



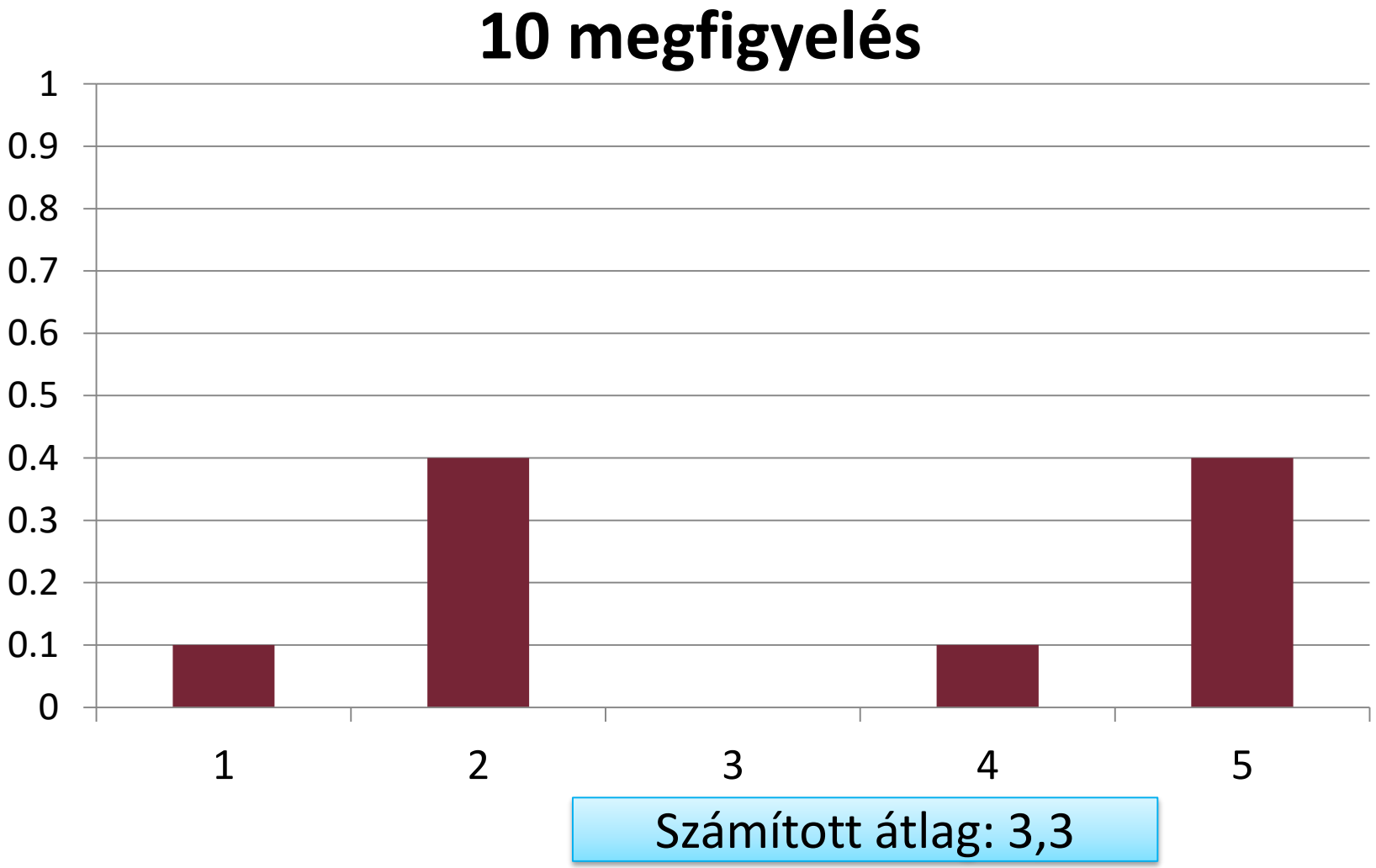
Számszerű jellemzők mérése

- Mérjük meg egy jellemző értékét! (1..5)



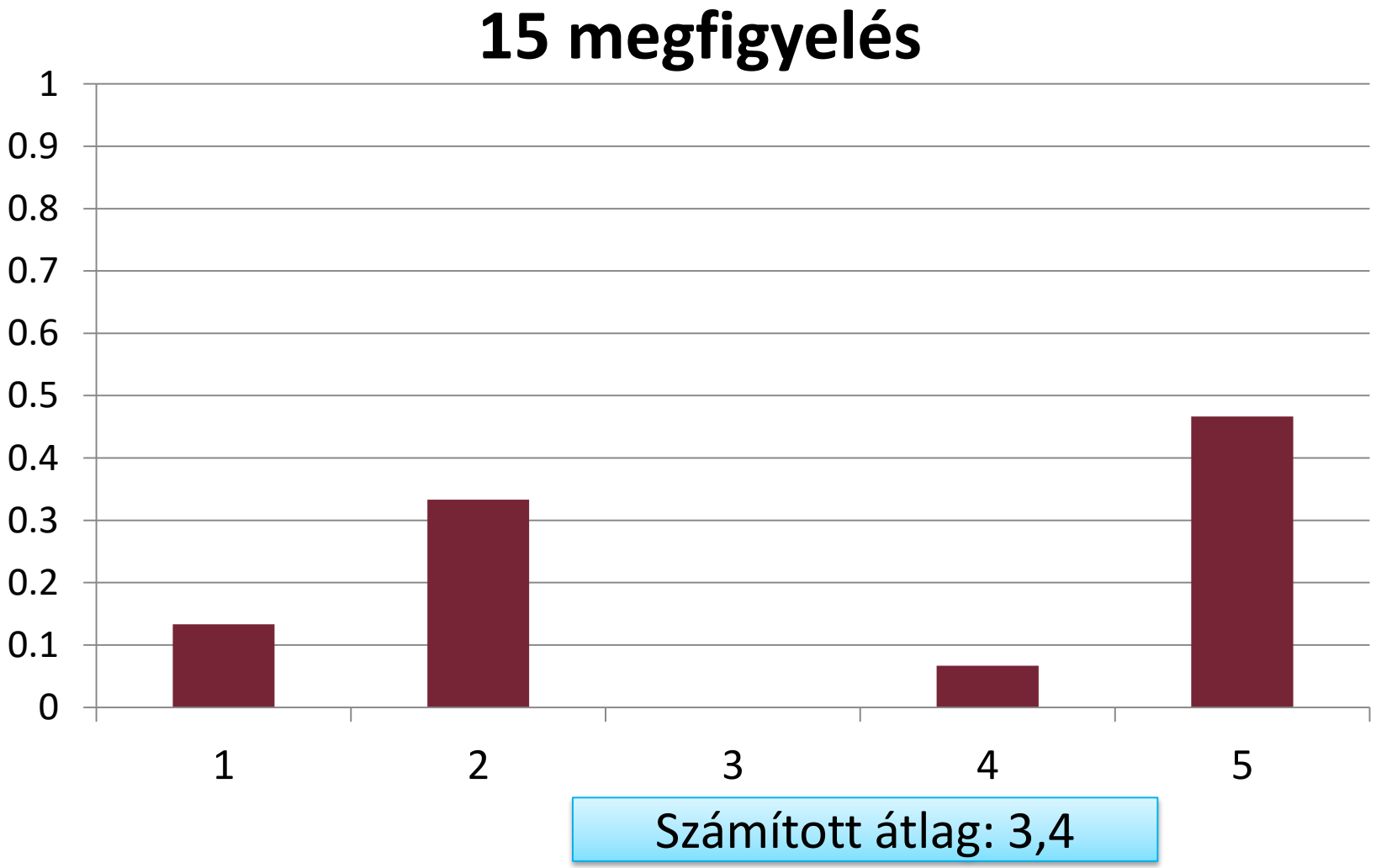
Számszerű jellemzők mérése

- Mérjük meg egy jellemző értékét! (1..5)



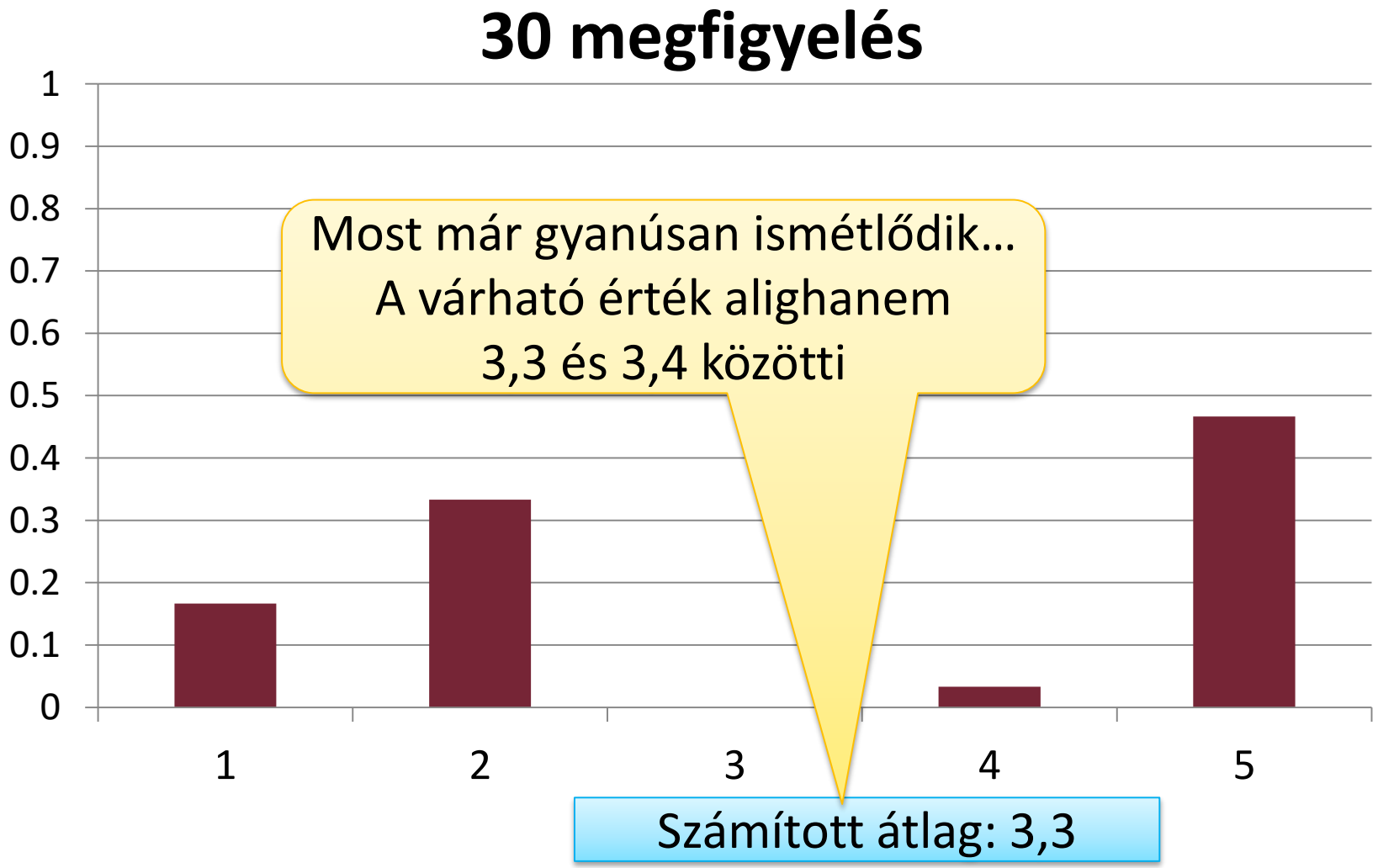
Számszerű jellemzők mérése

- Mérjük meg egy jellemző értékét! (1..5)



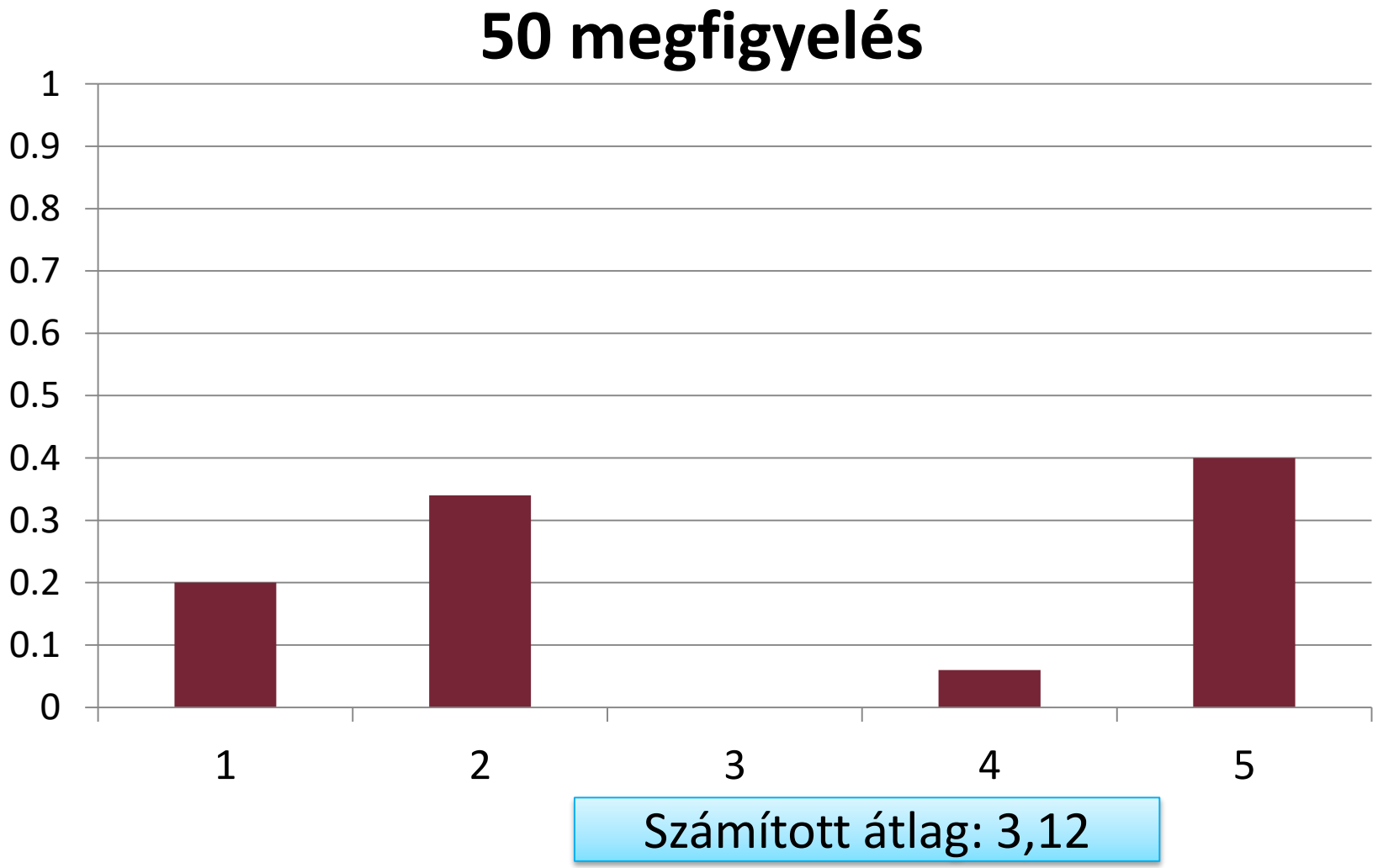
Számszerű jellemzők mérése

- Mérjük meg egy jellemző értékét! (1..5)



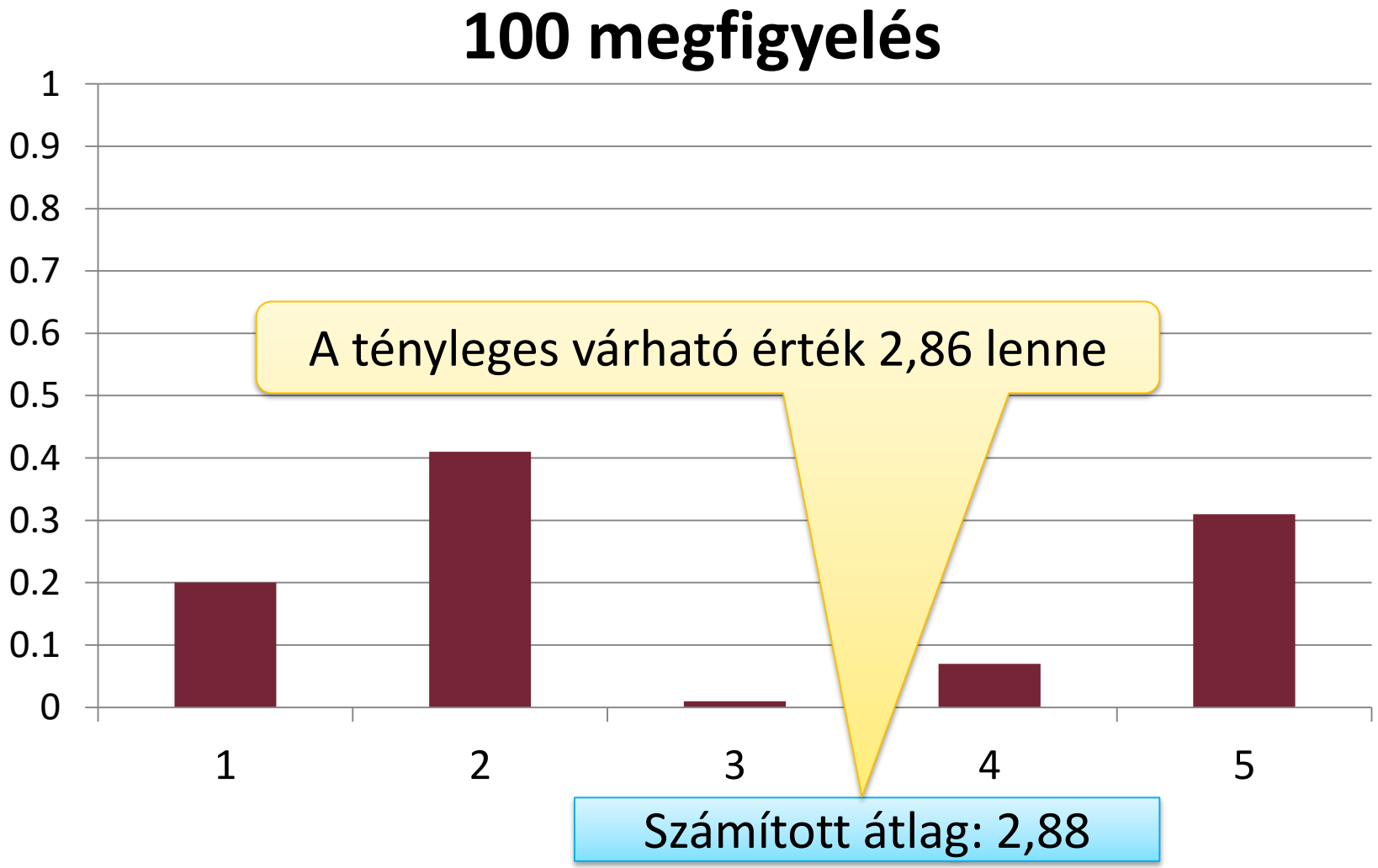
Számszerű jellemzők mérése

- Mérjük meg egy jellemző értékét! (1..5)



Számszerű jellemzők mérése

- Mérjük meg egy jellemző értékét! (1..5)



Ismétlés: tapasztalati átlag, szórás

- Valószínűségi változó: E (vizsgálandó jelenség)
  - Várható érték: $\mu = \mathrm{E}(X)$ átlagos viselkedés)
  - Szórás: $\sigma = \sqrt{\mathrm{E}(X-\mu)^2}$ (eltérések mértéke)
- Mintavétel: x_1, x_2, … x_t (mérések, megfigyelések)
  - Tapasztalati átlag: $\bar{x} = \frac{x_1 + x_2 + \ldots x_t}{t}$
  - Szórásra **nem jó**: $\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots(x_t-\bar{x})^2}{t}}$
  - Korrigált tapasztalati szórás: $\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots(x_t-\bar{x})^2}{t-1}}$

Tapasztalati átlag

- Módszer: számtani átlag képzése
  - Ismételt megfigyelések
  - Egymástól függetlenül
  - Azonos feltételek mellett
- Kérdések
  - Hány megfigyelést kell végezni?
  - A tapasztalati átlag mennyire jellemzi a valódi várható értéket?
- Először tisztázandó:
  - *A tapasztalati átlag eloszlása*

Tapasztalati átlag eloszlása

- Kísérlet = megfigyelések sorozata
- Megfigyelések sorozatának **tapasztalati átlaga**:
  - Egy jellemzőt ***t*** db független megfigyeléssel mérve,
  - majd a mért értékeket átlagolva kapott eredmény
- **Centrális határeloszlás tétele**ből következik:
  - Tetszőleges eloszlású jellemző
  - (de legyen *véges* ***m*** várható értékű és **s** szórású)
  - tapasztalati átlaga **t**→∞ esetén közelítőleg
  - **normális eloszlású**,
  - μ = ***m*** várható értékkel és $\sigma = {}^{\mathbf{s}}/_{\sqrt{t}}$ szórással

Tapasztalati átlag eloszlása

- Kísérlet = megfigyelések sorozata
- Megfigyelések sorozatának **tapasztalati átlaga**:
  - Egy jellemzőt ***t*** db független megfigyeléssel mérve,
  - [illegible]ött eredmény
- **C**[illegible]övetkezik:
  - [illegible]
  - (de legyen *véges* ***m*** várható [illegible]kű és **s** szórású)
  - tapasztalati átlaga **t**→∞ esetén közelítőleg
  - **normális eloszlású**,
  - μ = ***m*** várható értékkel és $\sigma = {}^{\mathbf{s}}/_{\sqrt{t}}$ szórással

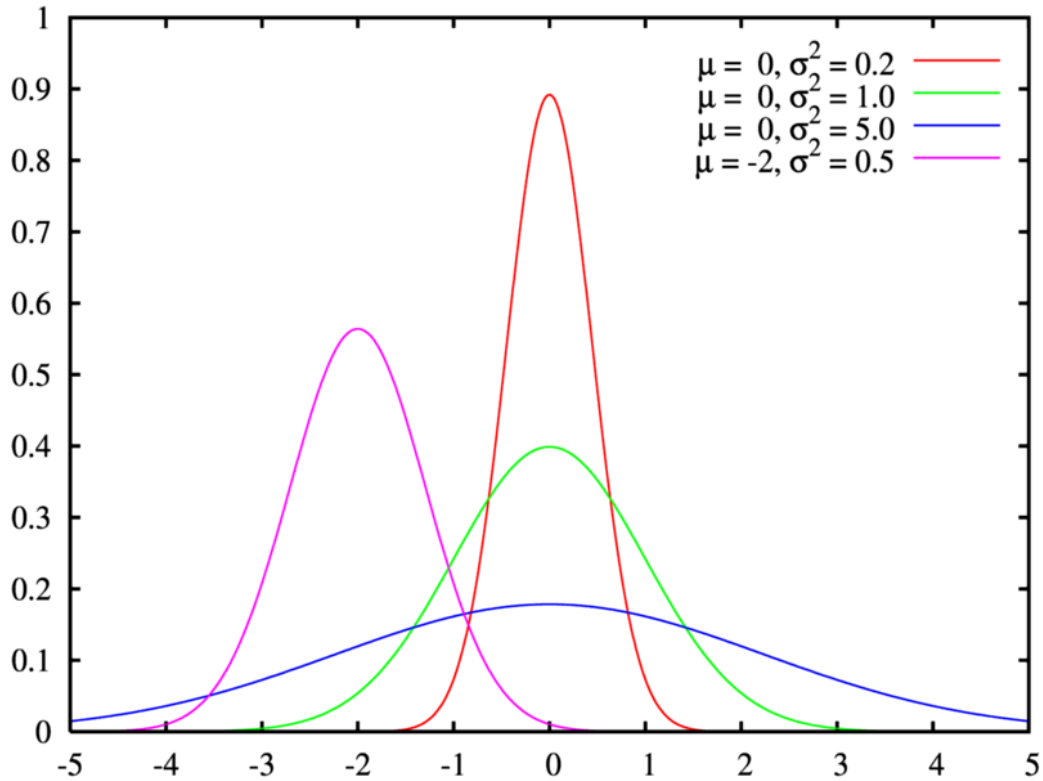
Ökölszabály:
ismert szórásnál t>30,
ismeretlen szórásnál t>100
után kezd elfogadható lenni a közelítés

A normális (Gauss) eloszlás

- Valószínűségsűrűség-függvénye: (nem kérdezzük)

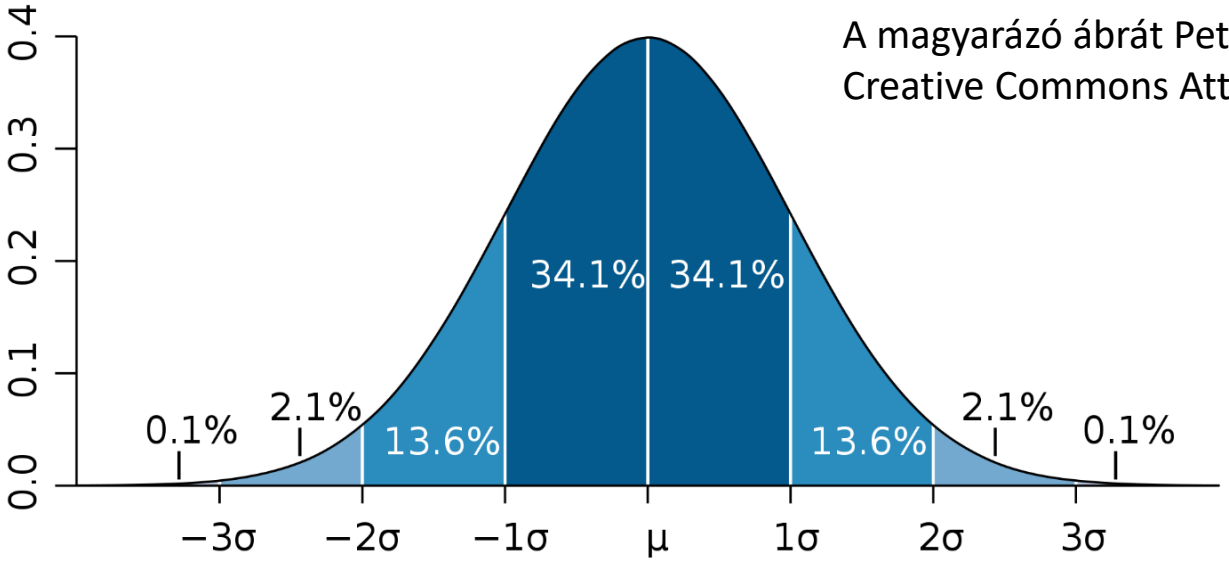
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Paraméterek
  - Várható értéke μ
    - → **m** a mi esetünkben
  - Szórása σ
    - → $^{s}/_{\sqrt{t}}$ esetünkben



A normális (Gauss) eloszlás

- A várható érték körül koncentrálódik



A magyarázó ábrát Petter Strandmark tette közzé Creative Commons Attribution 2.5 Generic licenc alatt

- A normális eloszlású változó…
  - az esetek **68%**-ában legfeljebb **1σ** messze kerül μ-től
  - az esetek **95%**-ában legfeljebb **2σ** messze kerül μ-től
  - az esetek **99,7%**-ában legfeljebb **3σ** messze kerül μ-től
  - …

Konfidenciaintervallumok

- Ha tehát a tetszőleges eloszlású, s szórású vizsgált jellemzőről t db (>30) megfigyelést végzünk
- A tapasztalati átlagáról…
  - **68%** biztonsággal kijelenthető, hogy legfeljebb $s/\sqrt{t}$ pontatlansággal becsli m értékét
  - **95%** biztonsággal $2s/\sqrt{t}$ sugarú intervallumba esik
  - **99,7%** biztonsággal $3s/\sqrt{t}$ sugarú intervallumba esik
- És t növelésével gyökösen szűkül az intervallum

Konfidenciaintervallumok

- Ha tehát a tetszőleges eloszlású, s szórású vizsgált jellemzőről t db (>30) megfigyelés végzünk
- A ta[illegible]gáról…
  - **68%** biztonsággal kijel[illegible] legfeljebb $s/\sqrt{t}$ pontatl[illegible]
  - **95%** biztonsággal $2s/\sqrt{t}$ sugarú intervallumba esik
  - **99,7%** biztonsággal $3s/\sqrt{t}$ sugarú intervallumba esik
- És t növelésével gyökösen szűkül az intervallum

Konfidenciaszint

Egyedi megfigyelés szórása

Konfidenciaintervallum sugara (félszélessége)

Kísérlettervezés példa

- A várható értékre 30 megfigyelés
  - Tapasztalati átlag: 2,3 s (jó-e ez? kell még mérni?)
  - Tapasztalati szórás: $\boldsymbol{s}$ = 1,1 s
- Cél
  - 99,7%-os konfidenciaintervallum 0,6 s széles legyen
- Kísérlettervezés
  - Elvárt sugár (félszélesség) = $3\sigma = {}^{3\boldsymbol{s}}/_{\sqrt{\boldsymbol{t}}} < 0{,}3$ s
    - (ez a σ az átlag szórása, nem az eredeti mért jellemzőé!)
  - Ezért t = 121 megfigyelés kell legalább
- Hol a csalás?

Korrekció

- Többnyire a tényleges eloszlás paraméterei *a priori* ismeretlenek (különben minek mérnénk?)
- Így nem használható fel a tényleges ***s*** szórás
- Csak a tapasztalati szórás használható → Gauss/normális helyett Student t-eloszlás
  - (más konfidenciaintervallumok)
- $t \rightarrow \infty$ esetén Student → normális
- Ökölszabály: $t>100$ esetén használható a Gauss