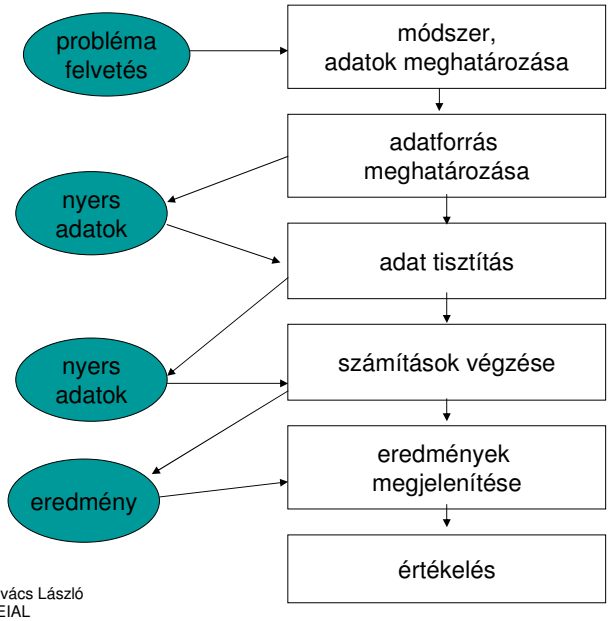
Adatelemzés és adatbányászat MSc

2. téma

Adatelemzési, statisztikai elemek áttekintése

Dr. Kovács László
ME GEIAL

Adatelemzés módszertana



Dr. Kovács László
ME GEIAL

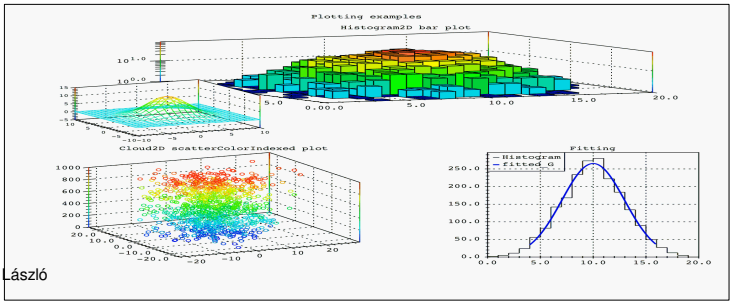
Adatelemzési módszerek

Elemzés célja:

Az adatsor tartalmának olyan formában történő reprezentálása, amely megkönnyíti a vizsgált problémakör szempontjából releváns részek felismerését

Módszerek:

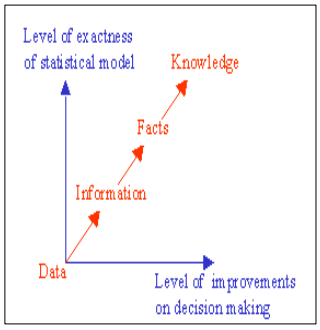
- lényeg kiemelés (összesítő adatok)
- formátum átalakítás (grafikus reprezentálás)



Adatelemzési módszerek

Adatelemzés nehézségei:

Forrásadatok meghatározása és kiemelése
Adatok konzisztencia ellenőrzése
Zajok kiküszöbölése
Módszerek használata
Paraméterek meghatározása
Nagy adatmennyiség kezelése, tömörítés
Megfelelő reprezentáció kiválasztása
Eredmények validálása



A megfelelő elemzés megfelelően releváns és kellően nagy adatmennyiségen nyugszik

Adatelemzési módszerek

Az elemzések rendszerint statisztikai alapokon nyugszanak

Módszeresség előnyei: - döntéshez jogalapot ad
- csökkenti a bizonytalanságot

Módszerek:

- várható érték, szórás
- eloszlások, sűrűség függvények
- interpolációk, extrapolációk
- regresszió analízis

A gyakorlatban több modellt is ki kell próbálni a probléma megoldásához

A cél: jó döntés meghozatala bizonytalan információk mellett

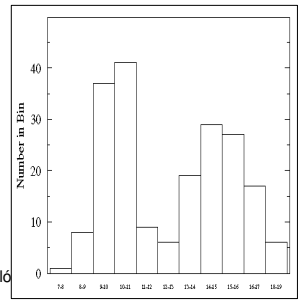
Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Statisztikai alapfogalmak ismétlése

Adott $\{a_i\}$ mintahalmazra értelmezhető:

- empirikus átlag : $= \Sigma\, a_i /n$,
- medián :
- módusz :
- várható érték : a'
- szórásnégyzet : $\sigma^2 = (\Sigma\, (a_i - a')^2)/(n-1)$



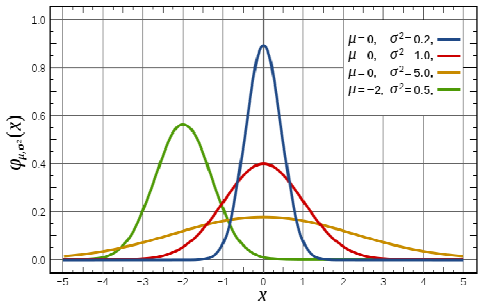
Sokszor több populáció együttesével (mixture) kell dolgozni

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

A statisztikai modellek alkalmazásának előfeltételei:

- zajok kiszűrése (outliers)
- homogén populációk vizsgálata
- véletlenszerűség ellenőrzése
- normál eloszlásra alakítás



várható érték

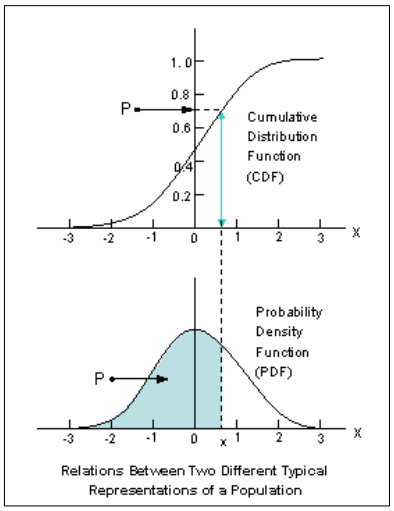
$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

szórásnégyzet

Normál eloszlás

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek



Valószínűségi változókat vezethetünk be

Centrális határeloszlás tétele: Nagy n-re a mért empirikus átlagok normális eloszlást mutatnak

A $x = (a - a')/\sigma$ változó N(0,1) normál eloszlású lesz

A N(0,1) eloszlás esetén az $|x| >\approx 2.8$ pontok „lehetetlen” eseményeknek tekintetők

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Zajok szűrése

Zajok kiszűrésének módszere: az elosztást N(0,1)-re hozva a 2.5-nél nagyobb abszolút értékek zajoknak tekinthetők

Példa: 46, 48, 38, 45, 47, 58, 44, 45, 43, 44

46	0,2	458	0,04	25,73333	0,039426
48	2,2	45,8	4,84	5,072803	0,433685
38	-7,8		60,84		-1,53761
45	-0,8		0,64		-0,1577
47	1,2		1,44		0,236556
58	12,2		148,84		2,404982
44	-1,8		3,24		-0,35483
45	-0,8		0,64		-0,1577
43	-2,8		7,84		-0,55196
44	-1,8		3,24		-0,35483
a	a-a’	a’	$(a-a')^2$	σ	x

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Véletlenszerűség ellenőrzése

Wald-Wolfowitz teszt: figyeli a sorozatok (runs) eloszlását (ne legyen se túl kevés, se túl sok sorozat)

Induló adatsor: mérési értékek
Lépések:
- a’ átlag kiszámítása
- sig(a-a’) –val helyettesítjük a-kat
- n^+, n^- (elemek db), R (sorozatok száma) meghatározása
- a” =1 + $2n^+n^-/(n^+ + n^-)$
- $\sigma^2 = (a"-1)(a"-2)/(n^+ + n^- -1)$.
- $z = (R-a") / \sigma$
- ha $|z| > z_0$ akkor nem véletlen a sorozat (~2.5)

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Véletlenszerűség ellenőrzése

Példa: 3, 5, 12, 7, 9, 8, 21, 17, 87, 22, 18, 24

Adatelemzési módszerek

Regresszió

Feladat: adott mérési pontra legjobban illeszkedő görbe megkeresése

Adottak: mérési pontok, függvényosztály (paraméteresen)

Feladat: a mérési pontokra legjobban illeszkedő paraméterek meghatározása

Optimalizálási feladat:

Célfüggvény: illeszkedési hiba: eltérések négyzetösszege

Optimalizálási módszerek:

- Derivált zérushelye
- Gradiens módszer
- Sztohasztikus keresés

$$\{(x_i, y_i)\}$$

$$\{f(\overline{p}_i, x)\}$$

$$E(\overline{p}_i) = \sum_i (f(\overline{p}_i, x_i) - y_i)^2$$

Adatelemzési módszerek

Lineáris regresszió

többváltozós lineáris regresszió: a mérési pontokat legjobban közelítő függvény meghatározása

egy függő változó feltételes várható érték becslésére szolgál

$$E(y|x_1\ x_2..) = F(x_1,x_2,..,\alpha_1,\alpha_2,..)$$
$$y = F(x_1,x_2,..,\alpha_1,\alpha_2,..) + \varepsilon$$

lineáris regresszió : a paraméterekben lineáris az F függvény

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 + \varepsilon$$
$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 + \alpha_4 x_1^2 + \alpha_5 x_2^2 + \varepsilon$$

a feltétel szerint ε egy 0 várható értékű, azonos paraméterű normál eloszlású

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Lineáris regresszió

a paraméterek várható értékének meghatározása a legkisebb négyzetek elvével történik
elemi esetre:

$$y_i = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 + \varepsilon_i$$
$$\varepsilon_i = y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3)$$
$$E(\varepsilon_i) = 0$$
$$\Sigma\, \varepsilon_i^2 \rightarrow \text{minimális}$$

a szélsőérték szükséges feltétele a deriváltak zérus értéke

$$\partial_{a1} \Sigma_i (y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3))^2 = 0$$
$$\partial_{a2} \Sigma_i (y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3))^2 = 0$$
$$\partial_{a3} \Sigma_i (y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3))^2 = 0$$

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Lineáris regresszió

egy változós esetre:

$$\partial_{a1} \Sigma_i (y_i - (a_1 x_i + a_2))^2 = 0$$
$$\partial_{a2} \Sigma_i (y_i - (a_1 x_i + a_2))^2 = 0$$

$$\partial_{a1} \Sigma_i (y_i^2 + a_1^2 x_i^2 + a_2^2 + 2 a_1 a_2 x_i - 2 y_i a_1 x_i - 2 y_i a_2) = 0$$

$$\partial_{a1} \Sigma_i (a_1^2 x_i^2 + 2 a_1 a_2 x_i - 2 y_i a_1 x_i + a_2^2 - 2 y_i a_2 + y_i^2) = 0$$

$$\partial_{a2} \Sigma_i (a_2^2 + 2 a_1 a_2 x_i - 2 y_i a_2 + a_1^2 x_i^2 - 2 y_i a_1 x_i + y_i^2) = 0$$

$$a_1 \Sigma_i x_i^2 + a_2 \Sigma_i x_i - \Sigma_i y_i x_i = 0$$

$$a_2 n + a_1 \Sigma_i x_i - \Sigma_i y_i = 0$$

$$a_1 = (n \Sigma_i x_i y_i - \Sigma_i x_i \Sigma_i y_i) / (n \Sigma_i x_i^2 - \Sigma_i x_i \Sigma_i x_i)$$

$$a_2 = (\Sigma_i y_i - a_1 \Sigma_i x_i) / n$$

Adatelemzési módszerek

Lineáris regresszió

Minta {(2.,4.),(4.,6.2), (6.,4.5)} és f(a,b,x) = ax+b

Statisztikai próbák

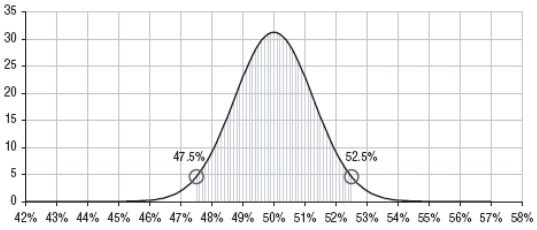
A statisztikusok is ideális világból indulnak ki.

Null-hipotézis elve: a mérési eltérés csak a véletlen műve

A vizsgálat tárgya:
- milyen paraméterű az ideális eloszlás?
- mennyire teljesül a null-hipotézis?

A mérési adatokon próbákat lehet végrehajtani a hipotézis ellenőrzésére, a hipotézis konfidencia szintjének megállapítására

Az elemzés megadja, hogy milyen konfidencia értékkel fog a paraméter egy megadott konfidencia intervallumba esni.



Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-eloszlás

Feltétel: legyenek X_i független normál eloszlású változók, (μ,σ) paraméterekkel.

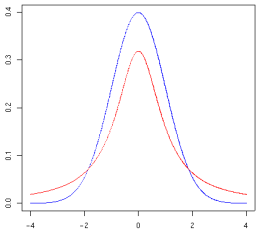
Ekkor a minta átlaga (n: mintaszám): $\overline{X} = \frac{\sum X_i}{n}$

minta szórásnégyzete: $S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$

Ekkor normál(0,1) eloszlású: $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$

Illetve Chi-négyzet eloszlású: $\frac{(n-1)S^2}{\sigma^2}$

Emiatt T Student eloszlású lesz: $\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$



Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Egyváltozós eset.

- null hipotézis: az eloszlás várható értéke: μ
- feladat: a tapasztalati eloszlás illeszkedik-e?
- vizsgált eloszlás: $t = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}}$
- függetlenségi tényező: $n - 1$

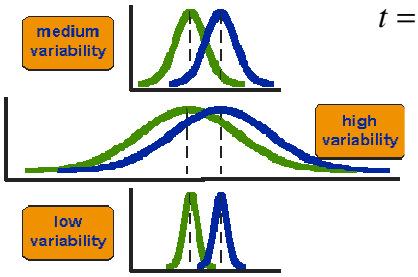
Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Kétváltozós eset.

A feladat adott kontroll és mérési eloszlás mellett eldönteni, hogy a mérési eloszlás mennyire illeszkedik a kontrollra



$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{(n_1 - 1)S^2{}_1 + (n_2 - 1)S^2{}_2}{n_1 + n_2 - 2} \cdot \dfrac{n_1 + n_2}{n_1 n_2}}}$$

Függetlenségi tényező: $n_1 + n_2 - 2$

Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

T-tábla használata:

- az oszlop jelöli a konfidenciát
- a sor jelöli a függetlenségi értéket
- a cella adja meg az előírt maximum t értékek

(ha a tábla érték nagyobb mint a számított, akkor megtartjuk a hipotézist)

FD	0.1	0.05	0.01
5	2.02	2.57	4.03
6	1.94	2.45	3.71
7	1.89	2.37	3.50
9	1.83	2.26	2.68
20	1.72	2.09	2.85
30	1.70	2.04	2.75

5% a kockázat, hogy úgy vetjük el a hipotézist, hogy mégis igaz

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Adott az alábbi adatsor: 483, 502, 498, 496, 502, 483, 494, 491, 505, 486.

Kérdés: tekintető-e 5%-os kockázat mellett a eloszlás m=500 várható értékűnek?

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

X=494, S = 8.05, μ=500, n=10, df=9

t=2.36, tablazat:2.26

megoldás: nem fogadható el a hipotézis

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Adott az alábbi adatsor, két eltérő helyen élő egyedhalmaz súlyértékei:

X1: 52; 57; 62; 55; 64; 57; 56; 55

X2: 41; 34; 33; 36; 40; 25; 31; 37; 34; 30; 38.

Kérdés: tekintető-e azonosnak a két eloszlás 5%-os kockázat mellett?

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{(n_1-1)S^2{}_1 + (n_2-1)S^2{}_2}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1 n_2}}}$$

X1=57.25, X2=34.45
n1=8, n2=11
$S1^2$=15.36, $S2^2$=21.87
t=11.12, táblázat=2.11

megoldás: nem fogadható el a hipotézis

Dr. Kovács László
ME GEIAL

Adatelemzések SQL háttere

ROLLUP tag:
A részletező adatok mellett a magasabb szintű összegek megadásra kerülnek

```
SELECT deptno, job, count(*), sum(sal)
FROM emp
GROUP BY ROLLUP(deptno,job);
```

DEPTNO	JOB	COUNT(*)	SUM(SAL)
10	CLERK	1	1300
10	MANAGER	1	2450
10	PRESIDENT	1	5000
10		**3**	**8750**
20	ANALYST	2	6000
20	CLERK	2	1900
20	MANAGER	1	2975
20		**5**	**10875**

Dr. Kovács László
ME GEIAL

Adatelemzések SQL háttere

CUBE tag:
A részletező adatok mellett az összes tetszőleges szintű összegek is megadásra kerülnek

```
ELECT   deptno,   job,   count(*),   sum(sal)
FROM   emp
GROUP BY   CUBE(deptno,job);
```

DEPTNO	JOB	COUNT(*)	SUM(SAL)
10	CLERK	1	1300
10	MANAGER	1	2450
10	PRESIDENT	1	5000
10		**3**	**8750**
20	CLERK	2	1900
20	MANAGER	1	2975
20		**5**	**10875**
	CLERK	**4**	**4150**
	MANAGER	**3**	**8275**
	PRESIDENT	**1**	**5000**
		14	**29025**

Adatelemzések SQL háttere

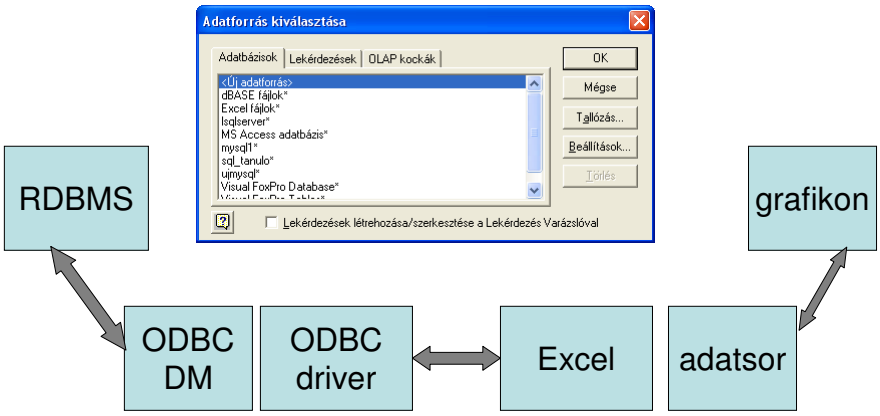
A számtásokat több segédopció támogatja (CASE, al-SELECT,..)

```
SELECT
  COUNT(*) as numvalues,
  MAX(freqnull) as freqnull,
  MIN(minval) as minval,
  SUM(CASE WHEN state = minval THEN freq ELSE 0 END) as numminvals,
  MAX(maxval) as maxval,
  SUM(CASE WHEN state = maxval THEN freq ELSE 0 END) as nummaxvals,
  SUM(CASE WHEN freq = maxfreq THEN 1 ELSE 0 END) as nummodes,
FROM
   (SELECT state, COUNT(*) as freq
    FROM orders
    GROUP BY state) osum CROSS JOIN
       (SELECT MIN(freq) as minfreq, MAX(freq) as maxfreq,
        MIN(state) as minval, MAX(state) as maxval,
        SUM(CASE WHEN state IS NULL THEN freq ELSE 0 END) as freqnull
         FROM (SELECT state, COUNT(*) as freq
         FROM orders
         GROUP BY state)
   osum)
summary
```

Adatok importálása Excelbe

Adatforrás megadása (ODBC)

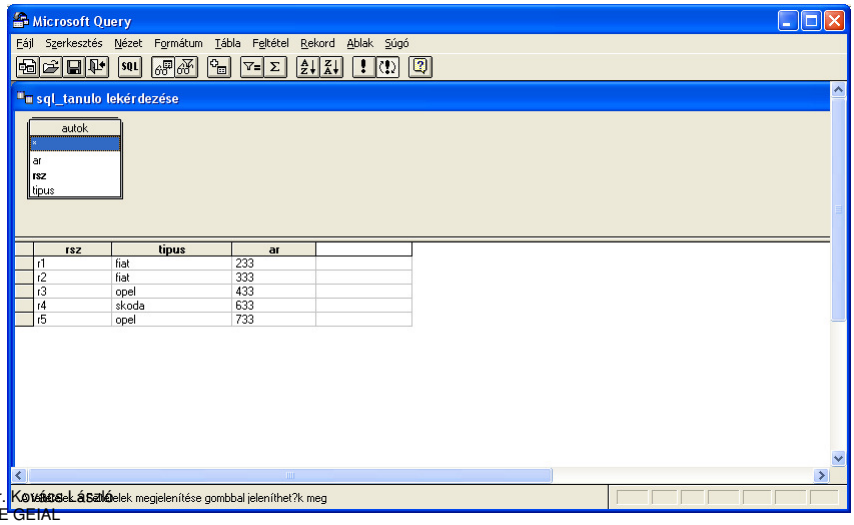
Adatok → Külső adatok importálása → Adatbázis lekérdezés



Adatok importálása Excelbe

Bejelentkezes

QBE felulet vagy SQL



Adatok importálása Excelbe

Grafikon felepitese

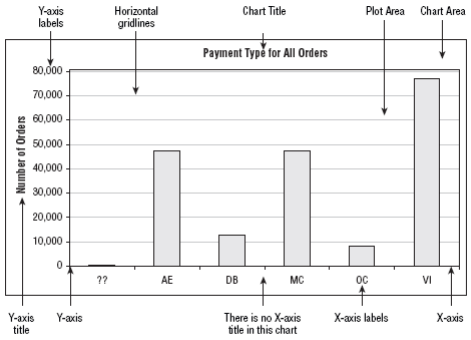


Figure 2-3: An Excel chart consists of many different parts.

Diagram Excelben

Parameterek

Forrasadat
- adattartomany
- adatsor

Diagram beallitasai
- cim
- tengely
- racsvonal
- jelmagyarazat
- feliratok

Mintak
- oszlop
- vonal
- korcikk
- …

Diagram terulet formazas
- mintazat
  - terulet
  - szegely
- betutipus

Tengely formazasa

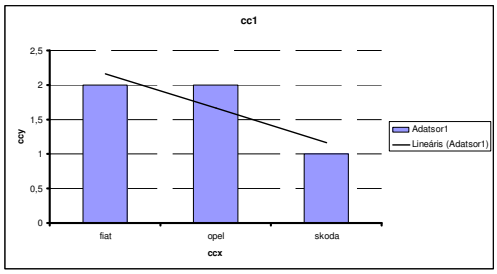
Racsok formazasa

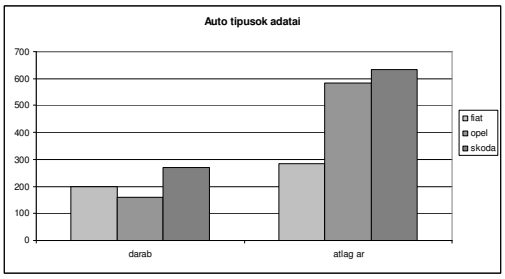
Adatsor formazasa

Feliratok formazasa

Trendvonal felvetele

Diagram Excelben

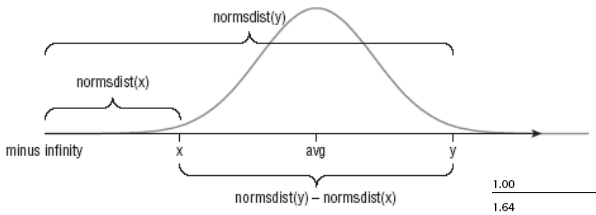




Dr. Kovács László
ME GEIAL

Statisztikai próbák

A hipotézis vizsgálat az ismert normál eloszláshoz kapcsolódik
Excel-ben a normsdist() függvény adja vissza az eloszlás értékét



1.00	68.269%
1.64	89.899%
1.96	95.000%
2.00	95.450%
2.50	98.758%
3.00	99.730%
3.29	99.900%
3.89	99.990%
4.00	99.994%
4.42	99.999%
5.00	100.000%

Kondidencia = ABS(normsdist(z)-normsdist(-z))

Dr. Kovács László
ME GEIAL