ADATREDUKCIÓ I.

Középértékek

Adatredukció

1. Mi a középérték: azonos fajta számszerű adatok közös jellemzője.
2. Követelmények:
   a) Számított középérték: közbenső helyet foglaljanak el, azaz

   $$x_{min} \leq \text{középérték} \leq x_{max}$$

   b) Helyzeti középérték: tipikus értékek legyenek (gyakran forduljonak elő).
   c) Legyenek könnyen meghatározhatók.
   d) Legyenek egyértelműen definiálva.
3. A középérték **az azonos fajta adatok tömegének számszerű jellemzője.**

Középértékek

Számított középértékek

- Aritmetikai átlag: $\overline{X}$
- Harmonikus átlag: $\overline{X}_h$
- Geometriai átlag: $\overline{X}_g$
- Kvadratikus átlag: $\overline{X}_q$

Helyzeti középértékek

- Módusz ***Mo***
- Medián ***Me***

Számított középértékek

- Matematikai összefüggés alapján számíthatók ki:
  - **Számtani (Aritmetikai) átlag**
    - Egyszerű
    - Súlyozott
  - **Harmonikus átlag**
    - Egyszerű
    - Súlyozott
  - **Mértani (Geometriai) átlag**
    - Egyszerű
    - Súlyozott
  - Négyzetes (Kvadratikus) átlag
    - Egyszerű
    - Súlyozott

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$\bar{x}_h = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i}$$

$$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

$$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{\sum_{i=1}^{n} f_i}}$$

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

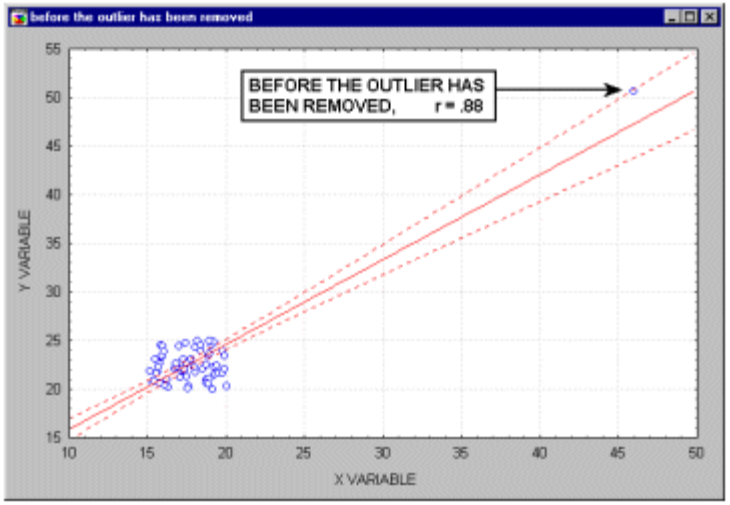
$$\bar{x}_g = \sqrt[\sum_{i=1}^{n} f_i]{\prod_{i=1}^{n} x_i^{f_i}}$$

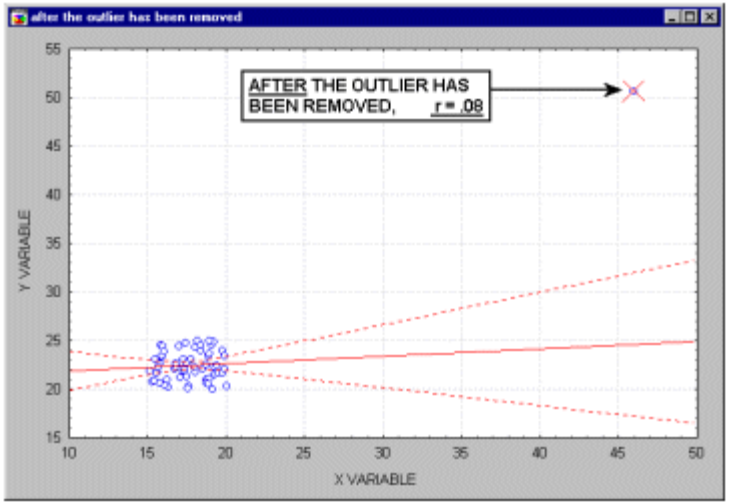
Helyzeti mutatók

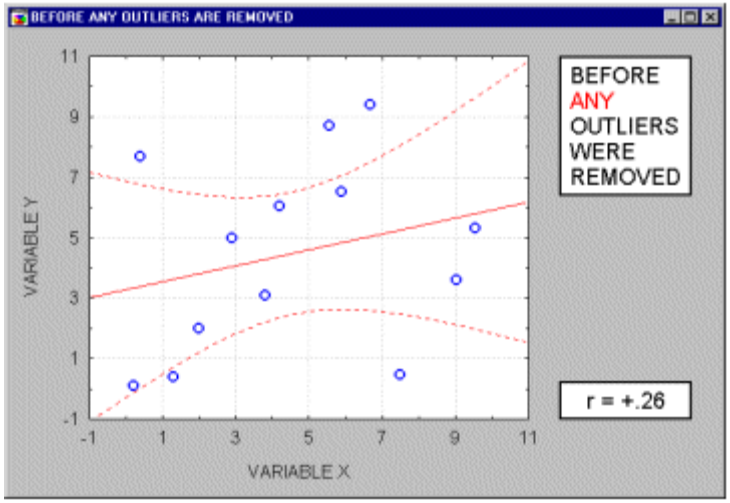
- Adatokat nagyságszerint rendezzük.
- Meghatározzuk a küszöb értéket és felosztjuk a tartományt a kívánt részre.
- Kvantilisek: az összes előforduló érték
  j/k (j=1,2,…,k-1) része kisebb és 1-j/k része nagyobb. Pl.

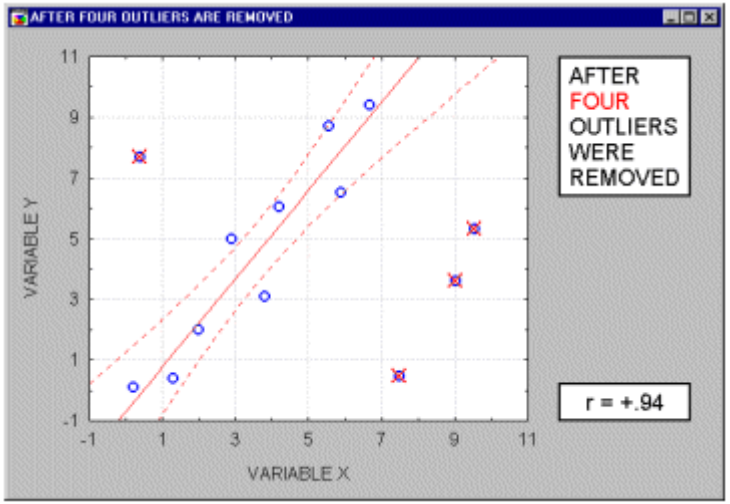
  k=2: Medián (Me)
  k=3: tercilis
  k=4: Qvartilis (Q1, Q2=Me, Q3)
  k=5: kvintilis
  k=10: decilis
  k=100: percentilis

Outlier


before the outlier has been removed
BEFORE THE OUTLIER HAS BEEN REMOVED, r = .88
Y VARIABLE
X VARIABLE


after the outlier has been removed
AFTER THE OUTLIER HAS BEEN REMOVED, r = .08
Y VARIABLE
X VARIABLE


BEFORE ANY OUTLIERS ARE REMOVED
BEFORE ANY OUTLIERS WERE REMOVED
r = +.26
VARIABLE Y
VARIABLE X


AFTER FOUR OUTLIERS ARE REMOVED
AFTER FOUR OUTLIERS WERE REMOVED
r = +.94
VARIABLE Y
VARIABLE X

Egyéb átlagok

- **Interquartile mean (IQM)** vagy **midmean**:
- **Nem érzékeny az outlier értékekre:**

$$x_{\mathrm{IQM}} = \frac{2}{n} \sum_{i=\frac{n}{4}+1}^{\frac{3n}{4}} x_i$$

Trimean vagy Tukey's trimean

- Kombinálja a medián és a midhinge előnyeit tekintettel az extrém értékekre:

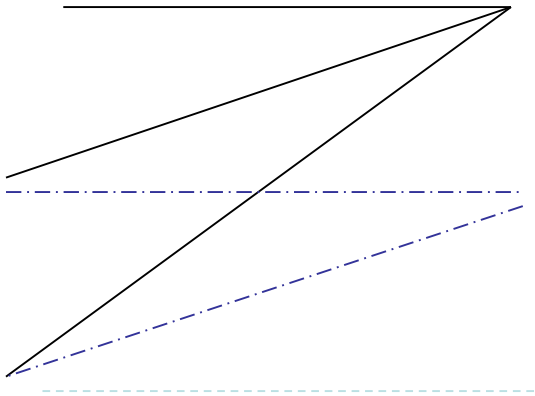
$$TM = \frac{Q_1 + 2Q_2 + Q_3}{4}$$

Az egyes adatfajtáknál milyen középértékeket alkalmazzunk?

Átlag

Medián

Módusz



Kvantitatív

Ordinális

Nominális

ADATREDUKCIÓ II.
Szóródás és mérése

A szóródás mérése

Szóródás: azonos fajta számszerű adatok ***különbözősége***

Mérése: az ismérvértékek

- valamilyen **középértéktől** vett vagy
- **egymás közötti**

különbségei alapján történik.

Szóródási mutatók

- A szóródás terjedelme
- Átlagos abszolút eltérés
- Szórásnégyzet, szórás, relatív szórás
- (Átlagos különbség)
- Koncentráció

A szóródás terjedelme

A legnagyobb és legkisebb ismérvérték különbsége

R vagy $T = X_{max} - X_{min}$

Interquartilis terjedelem:

$$IQT = Q_3 - Q_1$$

- A mutatószámok kifejezik, hogy mekkora értékközben ingadoznak az ismérvértékek.
- Gyakorlatban kevéssé használatos, mert csupán a két szélső értékre támaszkodik.

A szórásnégyzet (variancia) és szórás

Variancia
vagy:
szórásnégyzet

$$Var(x) = s^2 = \frac{\sum_{i=1}^{N} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{N} f_i}$$

Az egyes értékek számtani átlagtól vett **eltérés-négyzeteinek *átlaga:***

Korrigálatlan szórás:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$$

Korrigált szórás:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}$$

A szórás kiszámítható a négyzetes és a számtani átlag négyzeteinek különbségéből is:

$$s = \sqrt{\bar{x}_q^2 - \bar{x}^2}$$

Relatív szórás

$$V\% = \frac{s}{\bar{x}} * 100$$

Elvonatkoztat az ismérv-értékek

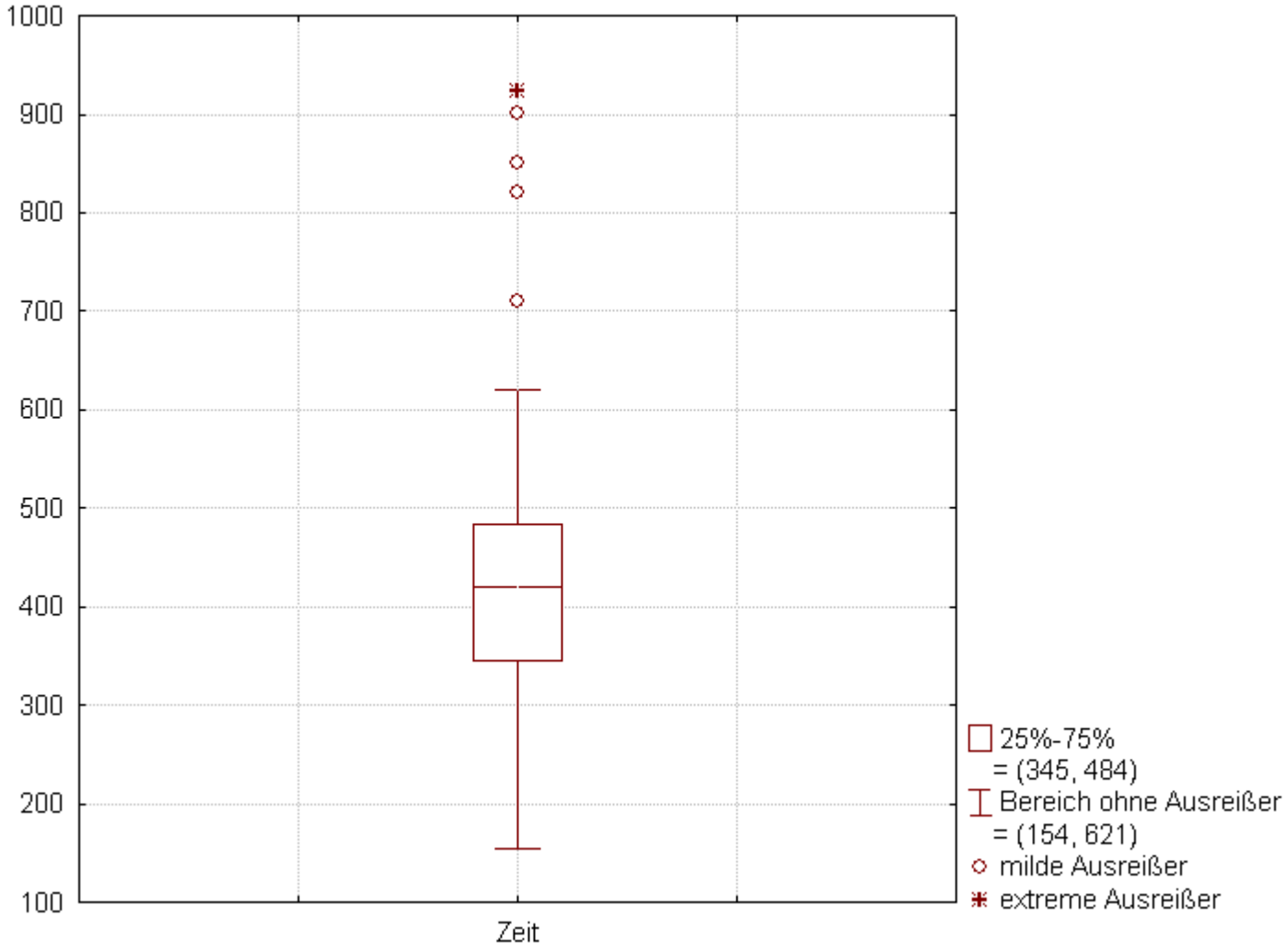
- nagyságrendjétől
- és mértékegységétől.

Azt mutatja meg, hogy a szórás hányad része (hány százaléka) az átlagnak.

Relatív szórás (variációs együttható, V)

- Az adatok szórását osztjuk az átlaggal, majd szorozzuk 100%-al
- Kicsi: a szórás, ha V<15%,
- Közepes: ha 15%<V<25%,
- Nagy: ha 25%<V<35%,
- Extrém (szélsőséges): ha V>35%

Box-and-whisker plot négy + nagyon távoli extrém értékkel: definiálva Q3 + 1.5(IQR) and Q3 + 3(IQR) alapján



Átlag szórása (Standard error, SEM)

- A mintaválasztás jóságát mutatja: a 0 közeli érték a jó érték, mert ekkor helyes a mintaválasztás (dimenziós érték!):

$$S_{\bar{x}} = \frac{S}{\sqrt{N}}$$

A szórás tulajdonságai

- Ha minden x értékhez ugyanazt a konstans számot hozzáadjuk (x+a), a szórás változatlan marad.
- Ha minden x értéket ugyanazzal a k konstans számmal megszorozzuk, (kx), a szórás is k-szorosára változik.
- Az eltérésnégyzet-összeg az ***átlagtól való eltérésekkel*** számolva a ***legkisebb***
- A **szórásnégyzet** felírható a **négyzetes átlag** és a **számtani átlag négyzetének** a különbségeként.
- A sokaságot jellemző **teljes szórásnégyzet** (variancia) megegyezik a rész-sokaságok **külső és belső szórásnégyzetének** összegével (ANOVA témakör):

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

Hiányzó értékek kezelése (Missing values)

- **Hiányzó érték:** nem regisztrált adat.
- **Hatása:** erőteljesen befolyásolhatják az elemzés eredményeit.
- **Többváltozós** módszereknél esetszám kiesés.

Hiányzó értékek jelölése

- 0 kód esetén a teendő
- kód használata: -99999
- Szoftver felé való közlés
- Hiányzó értékek kezelése:

  - üresen hagyjuk,

  - átlagot tesszük be: a *helyettesítés rombolja a változók eloszlásfüggvényét, konfidencia-intervallumát, megnöveli* az eloszlások csúcsosságát, a változók közötti lineáris kapcsolatokat is megváltoztatja, a korrelációs együttható közelebb kerül a 0-hoz.

MI (*multiple imputation)*

- Az MI célja, hogy a helyettesítésekkel együtt
- megtartsuk a változók eloszlását és a változók közötti asszociációkat.
- Szimuláción és legtöbbször Bayes-i alapokon álló technika, ahol a megfigyelt adatokból *m>1 verzióban modelleznek lehetséges adatokat a hiányzók helyére,* majd a végén egy algoritmus szerint kombinálják az eredményeket (a becsléseket és a szórásokat).

MI

- Általános szabályként olyan változók esetében használhatjuk az imputálást, ahol változónként maximum az adatok 30–40%-a hiányzik, de a teljes adatbázisban nincs több hiányzó, mint a teljes mátrix 10–15%-a.
- Ezek az arányok a szakirodalom szerint egyáltalán nem adnak okot aggodalomra a helyettesítés metódusát illetően.

Aszimmetria mérőszámai

Ferdeség mérése

- =FERDESÉG() – **SKEW()**

$$Ferdeség = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

  - A ferdeség az eloszlás középérték körüli aszimmetriájának mértékét jelzi. A pozitív ferdeség a pozitív értékek irányába nyúló aszimmetrikus eloszlást jelez, míg a negatív ferdeség a negatív értékek irányában torzított.
- =CSÚCSOSSÁG() – **KURT()**

$$Csúcsosság = -\frac{3(n-1)^2}{(n-1)(n-2)}$$

  - Egy adathalmaz csúcsosságát számítja ki. A függvény a normális eloszláshoz viszonyítva egy eloszlás csúcsosságát vagy laposságát adja meg. A pozitív értékek viszonylag csúcsos, a negatív értékek viszonylag lapos eloszlást jelentenek.

Aszimmetria

Az aszimmetria Pearson-féle **A-mutatószáma:**

$$A = \frac{\bar{x} - Mo}{\sigma}$$

Szimmetrikus eloszlás esetén: **A = 0**

Jobb oldali aszimmetria esetén: **A > 0**

Bal oldali aszimmetria esetén: **A < 0**

Az aszimmetria **F-mutatószáma**

$$F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$

Szimmetrikus eloszlás esetén: F **= 0**

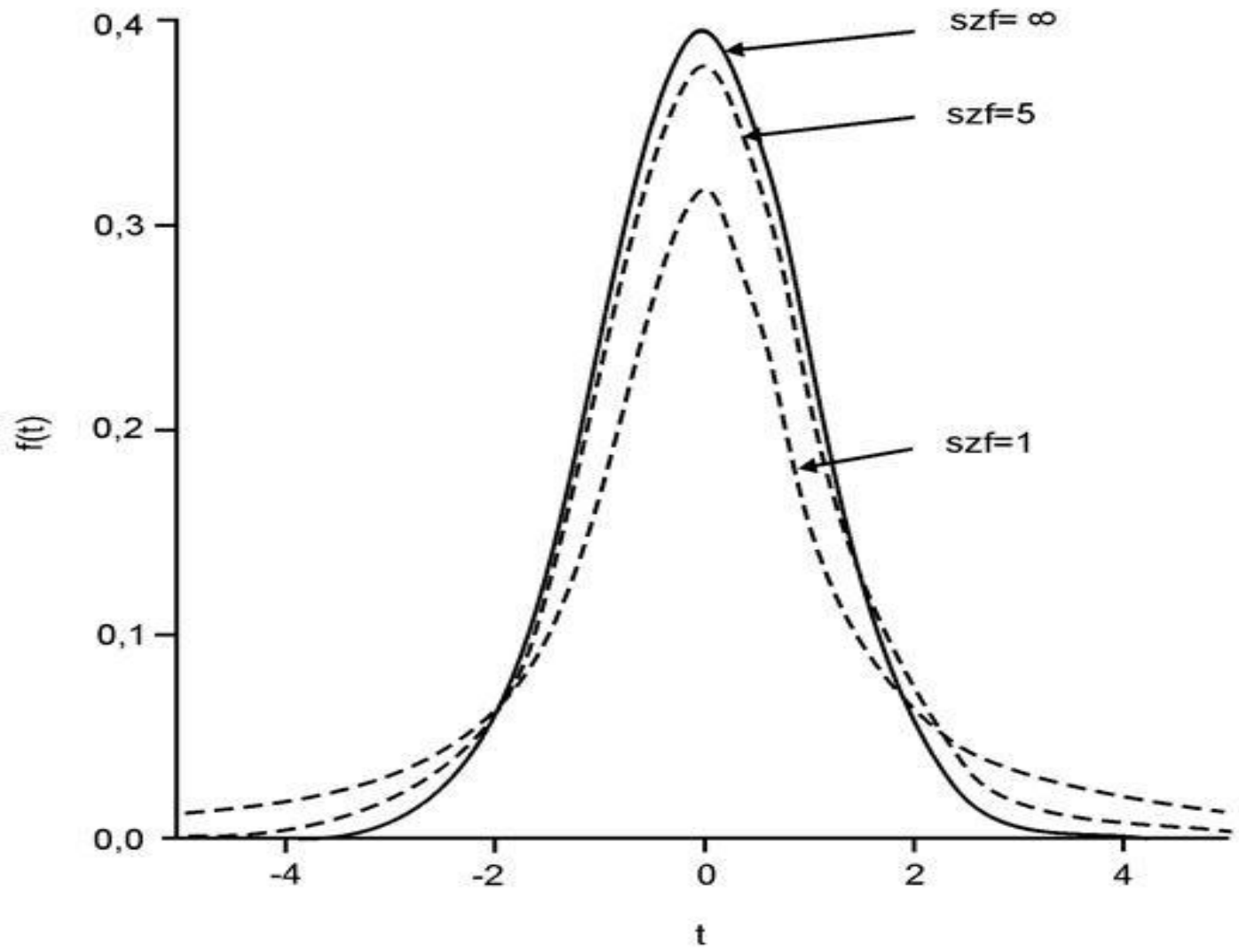
Jobb oldali aszimmetria esetén: F **> 0**

Bal oldali aszimmetria esetén: F **< 0**

Konfidenciaintervallum
(Confidence interval)

- Határozzuk meg $\bar{\chi}$ körül azt az intervallumot amibe előre meghatározott valószínűséggel esik a várható érték (μ).
- A várható értéket (μ) pontosan nem tudjuk, de $\bar{x}$ körül van: nagy (1-α) valószínűséggel a fenti intervallumban, és kicsi (α) valószínűséggel esik ezen kívülre.
- Ezt az intervallumot a *<u>várható érték becslésére szolgáló 100·(1- α)% konfidencia intervallumnak</u>* <u>nevezzük</u>.
- Leggyakrabban 90 v. 95%-os megbízhatósági szintet választunk (vagyis α = 0,1 ill. 0,05).

t-eloszlás



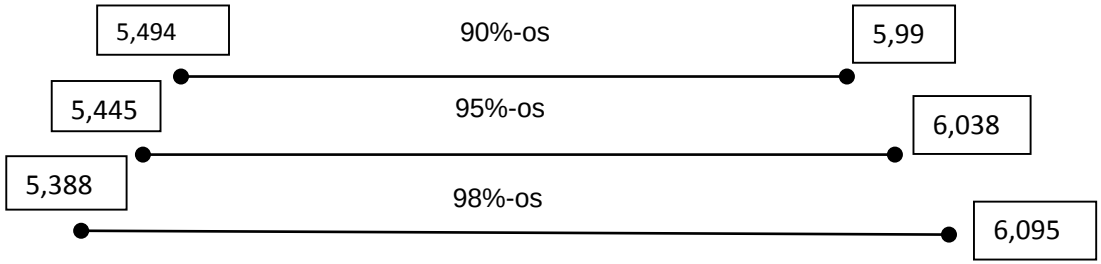
IV. táblázat. A t-(Student-) próba kritikus értékei 80% -os (90%-os) , 90% -os (95% -os), 95% -os (97.5% -os), 98% -os (99% -os), 99% -os (99.5% -os) , 99.9% -os (99.95% -os) kétoldali (egyoldali) szintre

szabadsági fok, f	Statisztikai biztonság					
	80%	90%	95%	98%	99%	99.9%
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460
40	1.3031	1.6839	2.0211	2.4233	2.7045	3.5510
50	1.2987	1.6759	2.0086	2.4033	2.6778	3.4960
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.4602
80	1.2922	1.6641	1.9901	2.3739	2.6387	3.4163
100	1.2901	1.6602	1.9840	2.3642	2.6259	3.3905
200	1.2858	1.6525	1.9719	2.3451	2.6006	3.3398
500	1.2832	1.6479	1.9647	2.3338	2.5857	3.3101
∞	1.2816	1.6449	1.9600	2.3263	2.5758	3.2905

N=96, df=95, α=0,05

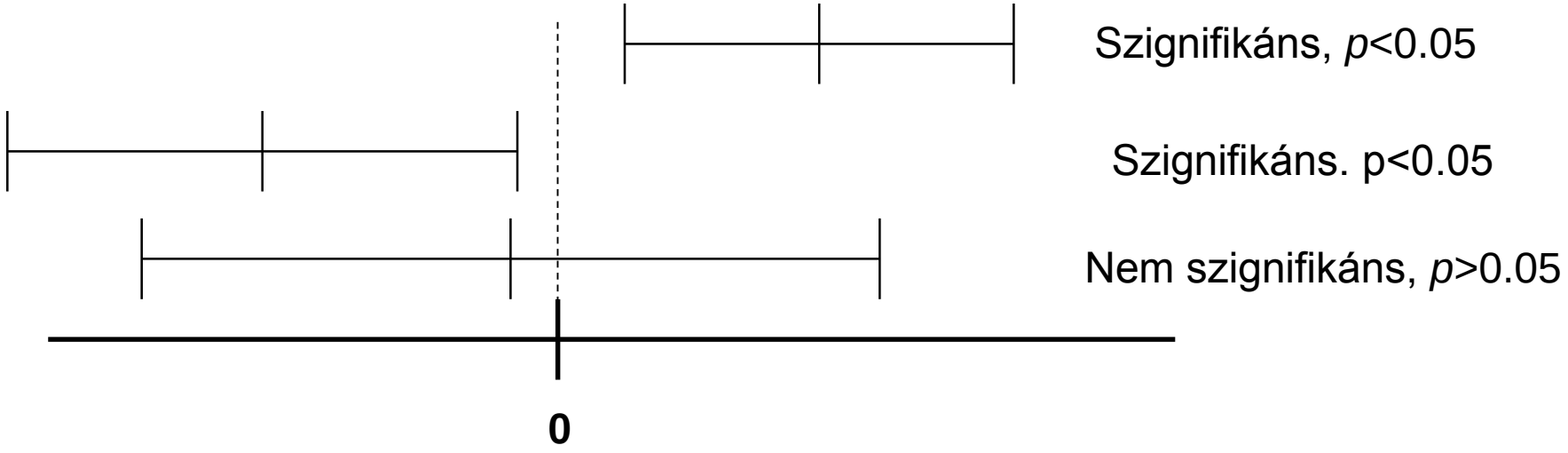
95%-os CI	Mean	5,742
	Standard Error	0,149
	$t_{kritikus}$	1,984
	$K=t_{kritikus}*SE$	0,297
	$L_{l=átlag} - K$	5,445
	$L_{l=átlag} + K$	6,038

CI intervallumok ábrázolása



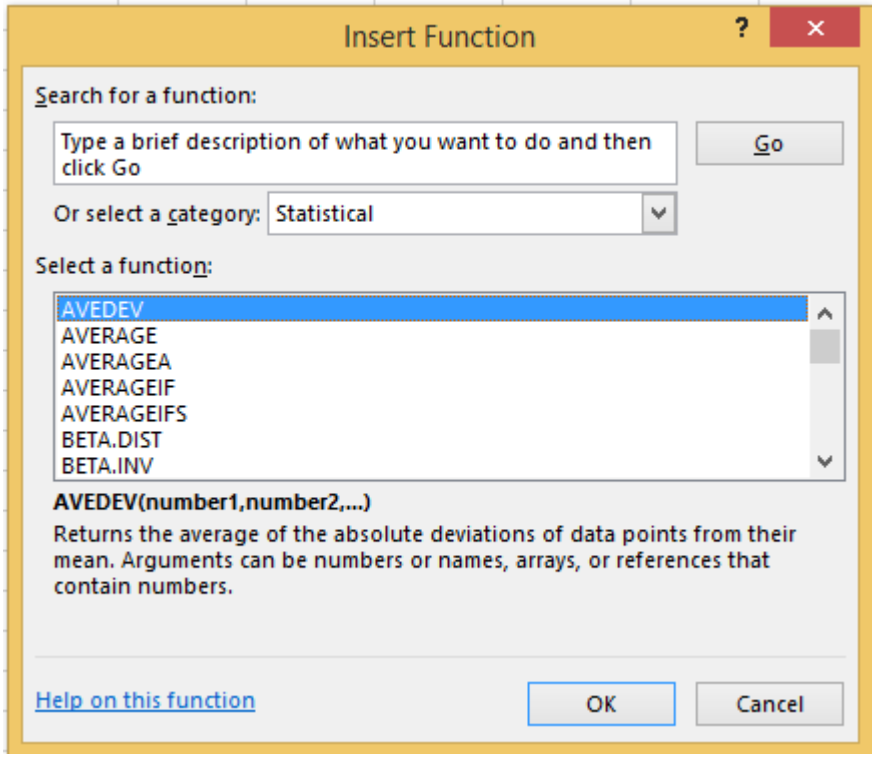
Szignifikancia vizsgálatok és a konfidenciaintervallum kapcsolata (H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$)

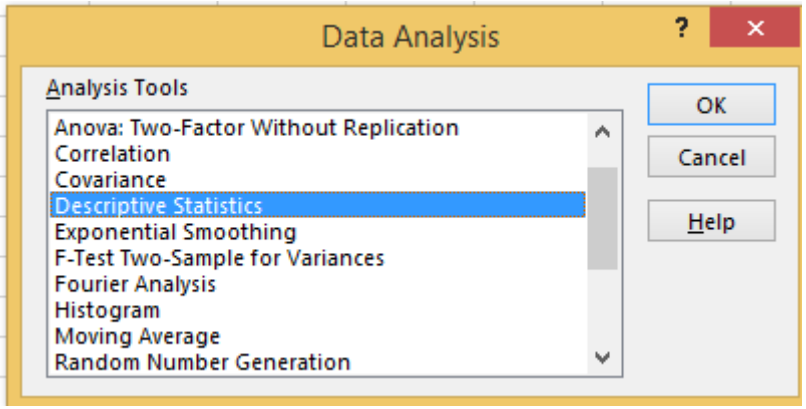
p-érték	szignifikancia	95% CI
p<0.05	szign. 5%-os szinten	pl. (4.5, 10.7) 0 nincs benne a konf. intervallumban
p ≥ 0.05	nem szign. 5%-os sz.	pl. (-1.72, 5.81) 0 benne van a konf. intervallumban



Megjegyzés: ha relatív kockázatot (RR) vagy esélyhányados (OR) vizsgálunk, akkor a konfidenciaintervallumban az **1**-et keressük, hogy az értéket tartalmazza-e.

Statisztikai függvények Excelben





Átlagra függvények az Excelben

- Számtani átlag: **=ÁTLAG( ) =AVERAGE()**
- Mértani átlag: **=MÉRTANI.KÖZÉP =GEOMEAN()**
- Harmonikus átlag: **=HARM.KÖZÉP() =HARMEAN()**
- Kvadratikus átlag:

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)}.$$

$$\sqrt{\frac{\sum x^2}{n}}$$

=SQRT(SUMSQ(A1:A10)/COUNTA(A1:A10))

Szórásfüggvények az Excelben

- =ÁTL.ELTÉRÉS – átlagos abszolút eltérés =AVEDEV()
- =SZÓRÁSP()
- =STDEV()- szórás
- =VAR() – variancia (szórásnégyzet)

Számláló - keresőfüggvények függvények

- =DARAB () =COUNT()
  - a megadott tartomány számmal kitöltött celláinak a számát adja
- =DARAB2() =COUNTA()
  - a megadott tartomány értékkel kitöltött celláinak (nem üres) a számát adja
- =DARABTELI () =COUNTIF ()
  - a megadott tartományban megszámolja, hogy hány darab cella felel meg a megadott kritériumnak
- =DARABÜRES () =COUNTBLANK ()
  - A megadott tartományban megszámolja hány db cella üres

Excel függvényei

- =MEDIÁN() =MEDIAN() : medián
- =MODE() : leggyakoribb érték
- =KVARTILIS() =QUARTILE()
- =PERCENTILIS() = PERCENTILE(): k-dik percentilis
- =SZÁZALÉKRANG() =PERCENTRANK(): egy értéknek egy adathalmazon vett százalékos rangját adja
- =MAX
- =MIN
- =KICSI() =SMALL(): egy adathalmaz k-dik legkisebb elemét adja értékül!
- =NAGY() =LARGE(): egy adathalmaz k-dik legnagyobb elemét adja értékül!
- =SORSZÁM() = RANK(): egy szám sorszámát adja, meg ha az adatokat sorba rendezzük

Kritikus-értéket számoló függvények

- Student's t-distribution

  a) the two-tailed value:
  =T.INV.2T(0.05,10) = 2.2281

  b) the left-tailed value:
  =T.INV(0.025,10) = -2.2281

- Normal distribution

  =NORM.S.INV(1-(0,05/2)) = 1,9600