A matematikai statisztika alapfogalmai

Kemény Sándor

Budapesti Műszaki és
Gazdaságtudományi Egyetem, Kémiai
és Környezeti Folyamatmérnöki
Tanszék

kemeny@mail.bme.hu

He uses statistics as a drunken man uses lamp-posts – for support rather than illumination.

Az ingadozás, bizonytalanság elkerülhetetlen

- a gyártott termékpéldányok különböznek
- az ismételt mérési eredmények nem azonosak
- ha egy tételből többször veszünk mintát, a talált selejtarány változik
- ha másik mintát veszünk a vízből, nem lesz teljesen azonos
- ha másik napon veszünk mintát, nem lesz ugyanolyan

A matematikai statisztika célja:
a mintából következtessünk a sokaságra

2 fő módszer: becslés és hipotézisvizsgálat

Sokaság és minta

a sokaság érdekel
a minta van a kezünkben

az egy év alatt gyártott darabok sokasága (mi a minta?)
a lehetséges mérési eredmények sokasága (mi a minta?)
a lehetséges gyártott darabok sokasága (mi a minta?)

1. példa

A páciensek kétféle gyógyszert kaptak, kisorsolva, hogy ki melyiket. Kettős vak vizsgálatot végeztek: az orvos és a páciens sem tudja, hogy ki melyik gyógyszert kapja.

Van-e a két gyógyszer között különbség a tekintetben, hogy egyforma arányban gyógyultak-e tőlük a betegek?

Gyógyszer típusa	Gyógyult	Nem gyógyult	Σ
A	23	7	30
B	18	13	31
Σ	41	20	61

$$\hat{\pi}_A = \frac{23}{30} = 0.7667 \qquad \hat{\pi}_B = \frac{18}{31} = 0.5806$$

„enumerative"

Statisztikai következtetés: a sokaság érdekel, de a minta van a kezünkben.

„inferential”

π_1 annak valószínűsége, hogy a beteg az A gyógyszertől meggyógyul
π_2 annak valószínűsége, hogy a beteg a B gyógyszertől meggyógyul

$$H_0 : \pi_1 = \pi_2 \qquad H_1 : \pi_1 \neq \pi_2$$

Az A és B gyógyszernél a gyógyulás relatív gyakorisága külön-külön binomiális eloszlást követ π_1 és π_1 paraméterrel

$$RR = \frac{\pi_1}{\pi_2}$$ kockázati arány (Risk Ratio)

$$1.24 < RR < 1.41$$

2. példa

Egy analitikai módszer torzítatlanságának vizsgálatára 5 ismételt mérést végeztek egy 3.25% ismert koncentrációjú munka-standarddel.

Az eredmények: 3.25, 3.27, 3.24, 3.26 és 3.24.

Elfogadva, hogy az adatok közelítőleg normális eloszlásúak, ellenőrizzük 5%-os szignifikanciaszinten a torzítatlanság hipotézisét!

$H_0 : \mu = \mu_0 = 3.25$ nullhipotézis

$H_1 : \mu \neq \mu_0 = 3.25$ ellenhipotézis

$\bar{x} =$

$s =$

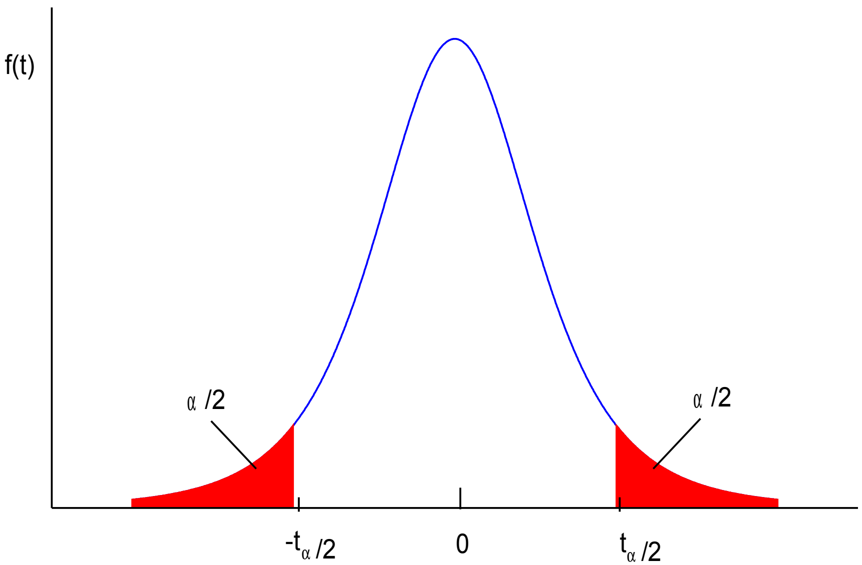
$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} =$$

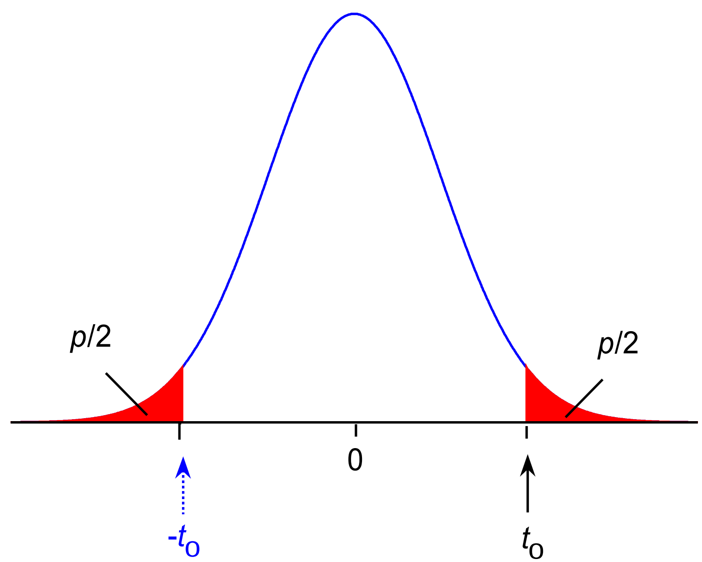
	Test of means against reference constant (value) (Test.sta)							
Variable	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
mert	3.25200(	0.01303ε	5	0.00583ı	3.25000(	0.34299ī	4	0.74886ε

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} =$$

táblázattal

software





$$t_{0.05/2}(4) = 2.776$$

Első- és másodfajú hiba

nullhipotézis	döntés: a H_0 hipotézist elfogadjuk	döntés: a H_0 hipotézist elutasítjuk
H_0 igaz	helyes döntés	elsőfajú hiba (α)
H_0 nem igaz	másodfajú hiba (β)	helyes döntés

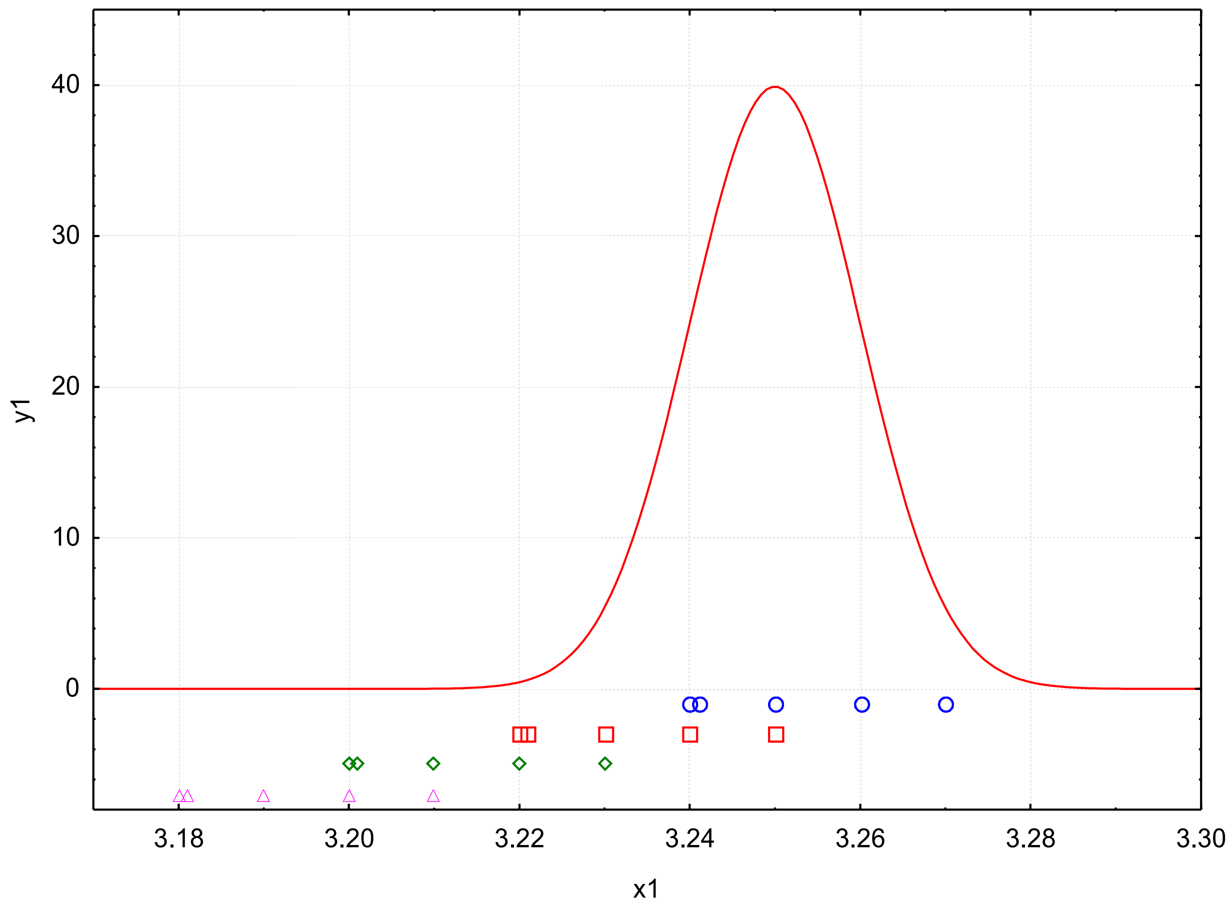
“fail to reject”

p értelmezése

Amit szeretnénk: $P(\mathrm{H}_0|\text{adatok}) = ?$

Ami kiszámítható: $P(\text{adatok}|\mathrm{H}_0) = ?$

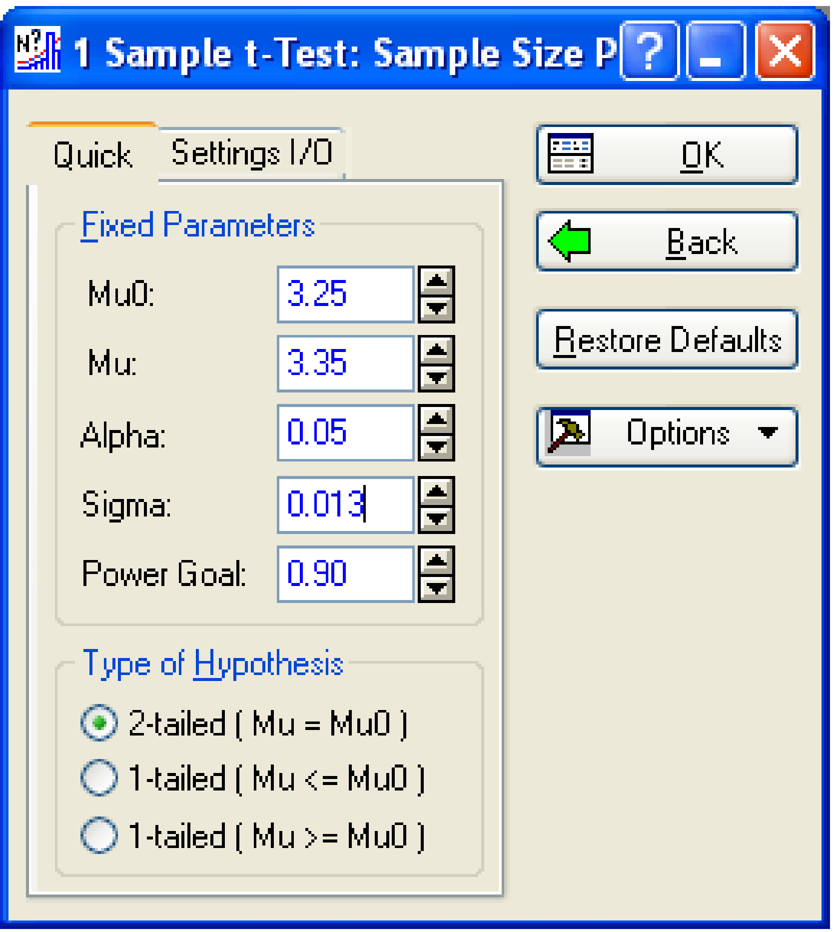
p annak valószínűsége, hogy a kapott vagy még szélsőségesebb adatok adódjanak, ha H_0 igaz.



40
30
20
10
0
y1
3.18
3.20
3.22
3.24
3.26
3.28
3.30
x1

Milyen biztonsággal tudnánk kimutatni 15 méréssel, ha 0.1 nagyságú atorzítás?

Statistics>Power Analysis>Sample size>One Mean, t-test



	Sample Size Calculat One Mean, t-Test H0: Mu = Mu0
	Value
Null Hypothesized Mean (Mu0	3.250(
True Population Mean (Mu)	3.350(
Population S.D. (Sigma)	0.013(
Standardized Effect (Es)	7.692:
Type I Error Rate (Alpha)	0.050(
Power Goal	0.900(
Actual Power for Required N	0.999!
Required Sample Size (N)	3.000(

$$H_0 : \mu \leq \mu_0 \qquad\qquad H_1 : \mu > \mu_0$$

$$H'_0 : \mu \geq \mu_0 \qquad\qquad H'_1 : \mu < \mu_0$$

Ha elutasítjuk H_0-t (vagy H\`$_0$-t), azt látjuk bizonyítva… (és csak $p \leq \alpha$ a valószínűsége annak, hogy tévedünk).

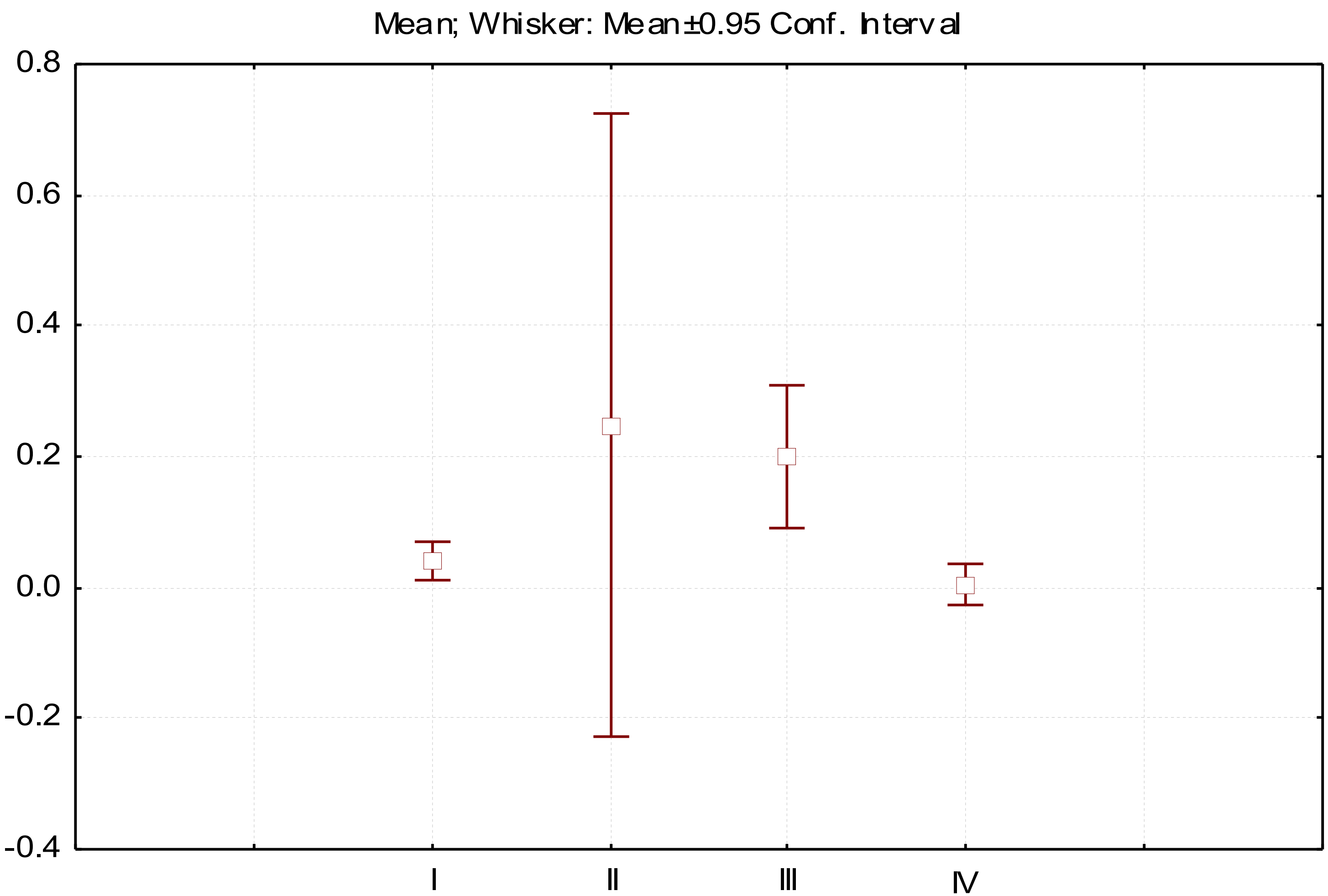
Ha sem H_0-t, sem H\`$_0$-t nem tudtuk elutasítani, nem jutunk döntésre, szürke zóna.

Ezért mondjuk, hogy az ellenhipotézisben legyen, amit bizonyítani akarunk.

Ha a másodfajú hiba β valószínűségével is foglalkoznánk, lehetne a nullhipotézisben, amit bizonyítani akarunk $\rightarrow n$ mintaelemszám.

A konfidencia-intervallum többet mond, mint a *p* érték!

J. H. Steiger, R.T. Fouladi: Noncentrality Interval Estimation and the Evaluation of Statistical Models, Chapter 9 in: L.L. Harlow, S.A. Mulaik, J.H. Steiger: What if there were no significance tests? Mahwah, NJ: Erlbaum (1997)



Kétmintás *t*-próba

3. példa

Két cipőtalp-anyag kopását hasonlítjuk össze, 10-10 fiú lábán, a használat során.

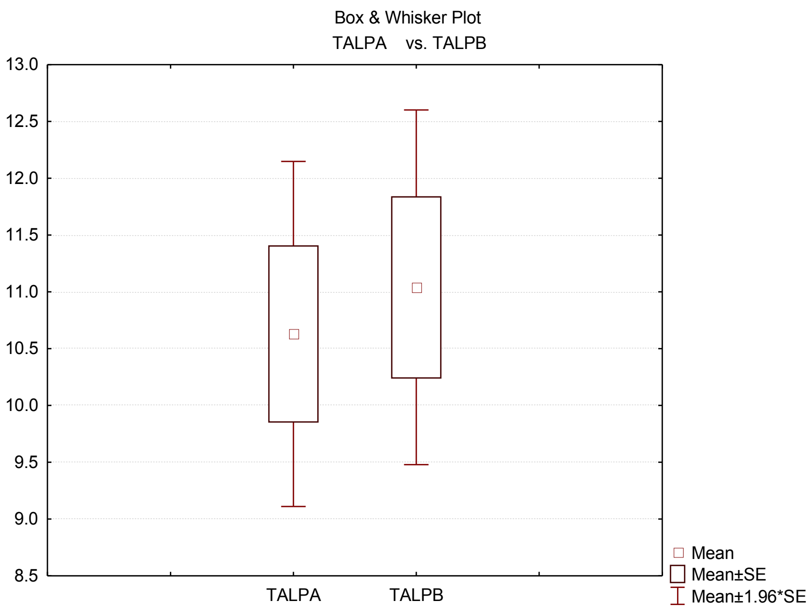
	n	átlag	szórásnégyzet
A	10	10.61	6.063
B	10	11.04	6.343

Vizsgáljuk meg 0.05-os szinten, van-e különbség az átlagos kopásban!

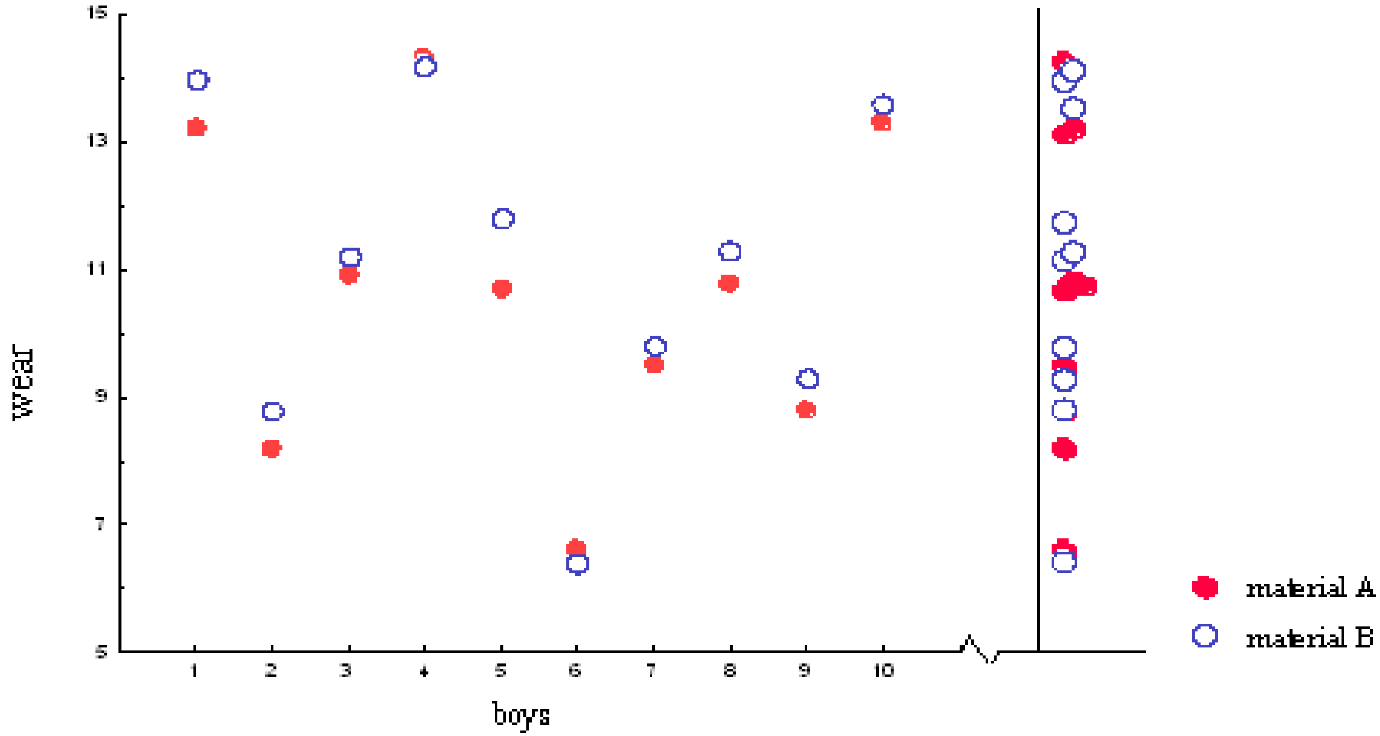
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$

Feltételezhetjük, hogy a két sokaság varianciája megegyezik? (*F*-próba!)

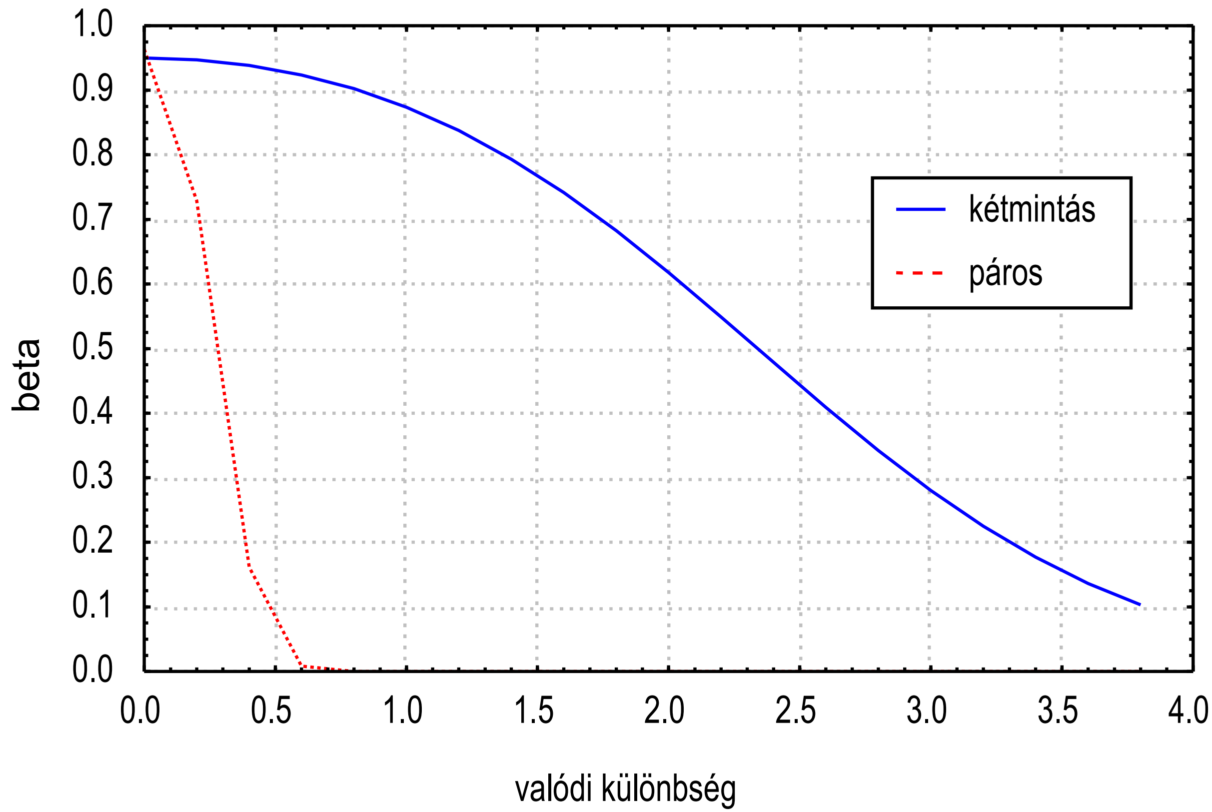
	T-test for Independent Samples (Fiucipo.sta) Note: Variables were treated as independent samples										
Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
TALPA vs. TALPB	10.6300	11.0400	-0.36891	18	0.71649	10	10	2.45132	2.51846	1.05552	0.93715



Páros t-próba



OC-görbe a fiúcipő-példához



VARIANCIAANALÍZIS (ANOVA)

Nevével ellentétben nem szórások, hanem átlagok összehasonlítására szolgál

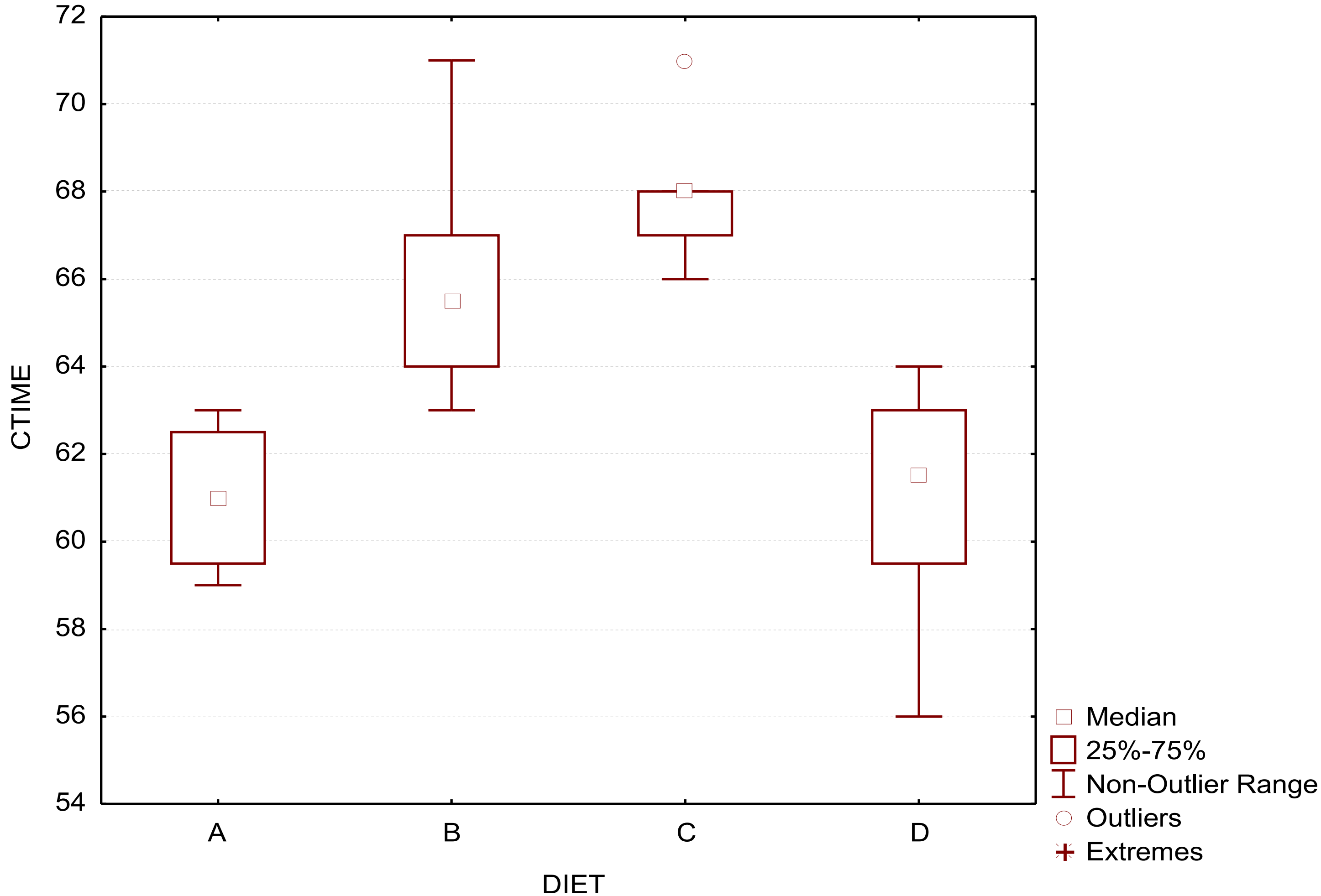
29. példa

(Box-Hunter-Hunter: Statistics for Experimenters, J. Wiley, 1978, p. 165)
Véralvadási idő (sec) négyféle diéta esetén

veralv.sta

Diéta			
A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$


72
70
68
66
64
62
60
58
56
54
CTIME
A
B
C
D
DIET
Median
25%-75%
Non-Outlier Range
Outliers
Extremes

Minitab>Stat>ANOVA>One-Way

Results for: blood.MTW

One-way ANOVA: CTIME versus DIET

```
Source  DF      SS     MS      F      P
DIET     3  228.00  76.00  13.57  0.000
Error   20  112.00   5.60
Total   23  340.00

S = 2.366   R-Sq = 67.06%   R-Sq(adj) = 62.12%

                               Individual 95% CIs For Mean Based on
                               Pooled StDev
Level  N    Mean  StDev  -----+---------+---------+---------+----
A      4  61.000  1.826  (-------*--------)
B      6  66.000  2.828                   (------*------)
C      6  68.000  1.673                          (------*-----)
D      8  61.000  2.619     (----*-----)
                         -----+---------+---------+---------+----
                            60.0      63.0      66.0      69.0
```

between

within

$$\hat{\sigma}_y^2 = s_R^2 = \frac{\sum_i s_i^2 (p_i - 1)}{\sum_i p_i - r}$$

döntés?

$$\sigma_{\bar{y}_i} = \sigma_y \Big/ \sqrt{p_i}$$

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

VARIANCIAANALÍZIS (ANOVA) véletlen faktorok esetén

Variancia-komponens-elemzés

A Gage R&R study mint ANOVA

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{k(ij)}$$

ismétlési hiba

alkatrész operátor kölcsönhatás

operá-tor:	A :			B :			C :		
minta	1.mérés	2.mérés	3.mérés	1.mérés	2.mérés	3.mérés	1.mérés	2.mérés	3.mérés
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									

A különböző alkatrészekre kapott mérési eredmények ingadozásának varianciája:

$$\sigma^2_{teljes} = \sigma^2_{alkatrész} + \sigma^2_{mérés}$$

A mérésnek tulajdonítható ingadozás:

$$\sigma^2_{mérés} = \sigma^2_{reprod} + \sigma^2_{ism}$$

A reprodukálhatósági variancia:

$$\sigma^2_{reprod} = \sigma^2_{kezelő} + \sigma^2_{alkatrész*kezelő}$$

Gage R&R for micro

Gage R&R Study - ANOVA Method

Two-Way ANOVA Table With Interaction

Source	DF	SS	MS	F	P
part	9	0.59893	0.066548	7.2628	0.000
operator	2	0.46126	0.230630	25.1700	0.000
part * operator	18	0.16493	0.009163	13.4987	0.000
Repeatability	60	0.04073	0.000679		
Total	89	1.26585			

Alpha to remove interaction term = 0.25

Source	VarComp
Total Gage R&R	0.0108891
Repeatability	0.0006788
Reproducibility	0.0102103
operator	0.0073822
operator*part	0.0028280
Part-To-Part	0.0063761
Total Variation	0.0172652

Gage R&R

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	0.0108891	63.07
Repeatability	0.0006788	3.93
Reproducibility	0.0102103	59.14
operator	0.0073822	42.76
operator*part	0.0028280	16.38
Part-To-Part	0.0063761	36.93
Total Variation	0.0172652	100.00

$$\frac{P}{T} = \frac{6\hat{\sigma}_{R\&R}}{USL - LSL}$$

Process tolerance = 1.52

Source	StdDev (SD)	Study Var (6 * SD)	%Study Var (%SV)	%Tolerance (SV/Toler)
Total Gage R&R	0.104351	0.626104	79.42	41.19
Repeatability	0.026054	0.156322	19.83	10.28
Reproducibility	0.101046	0.606275	76.90	39.89
operator	0.085920	0.515519	65.39	33.92
operator*part	0.053179	0.319076	40.47	20.99
Part-To-Part	0.079851	0.479104	60.77	31.52
Total Variation	0.131397	0.788382	100.00	51.87

$$\frac{P}{V} = \frac{6\hat{\sigma}_{R\&R}}{6\hat{\sigma}_{total}}$$

Number of Distinct Categories = 1

AZ ADATOK ÁBRÁZOLÁSA

Yogi Berra: " You can observe a lot by watching "

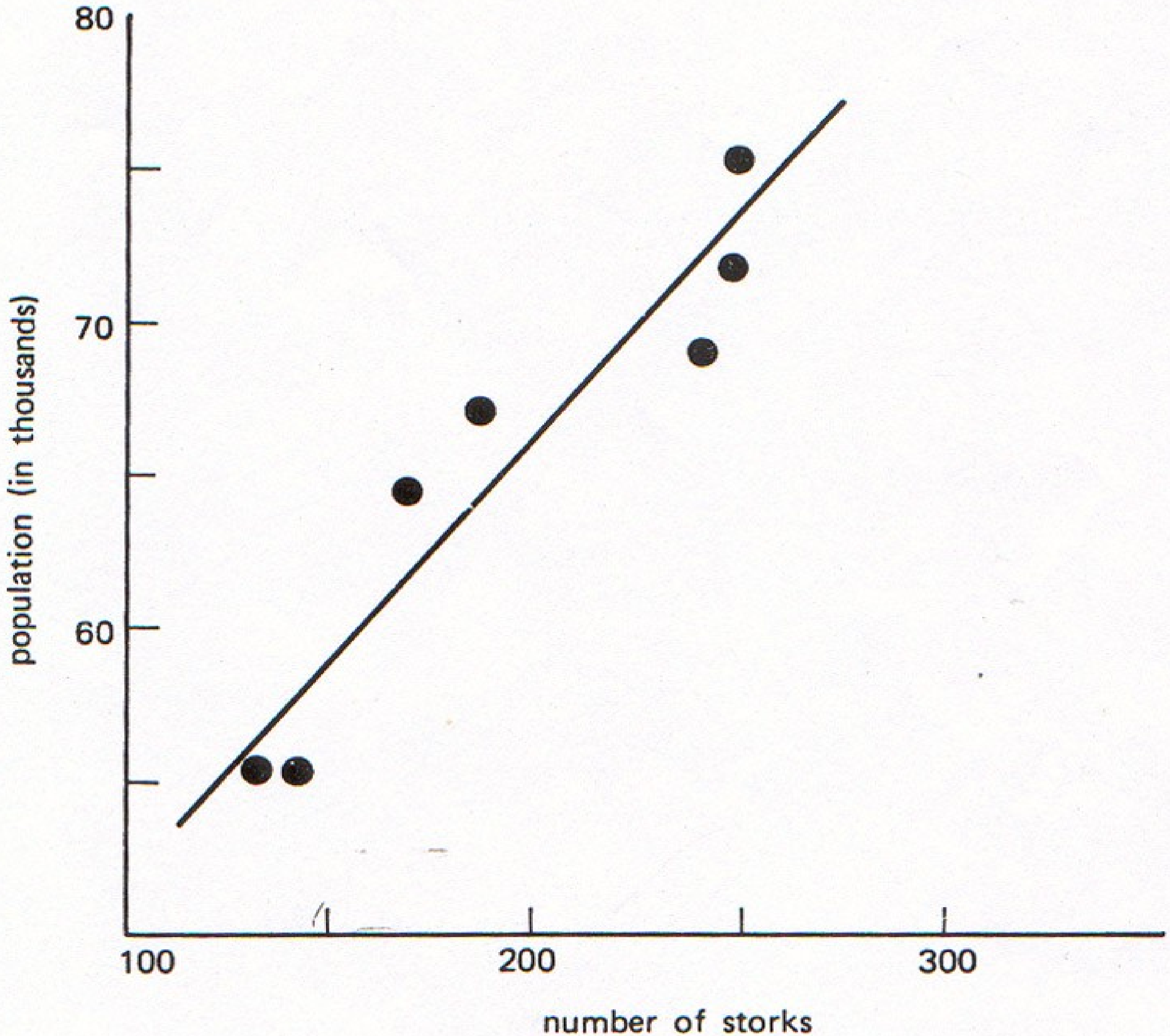


FIGURE 1.4. A plot of the population of Oldenburg at the end of each year against the number of storks observed in that year, 1930–1936.