A statisztikai adatelemzés módszerei és gyakori hibái

Varga József

DE OEC Nukleáris Medicina Intézet

Statisztika az orvosi irodalomban:

- >50%: csak %-számítás, átlag, szórás
- >80%: a fentieken kívül:
  t-próba, χ^2-próba, lineáris regresszió

- A statisztikai módszereket alkalmazó orvosi cikkek kb. fele rosszul használja azokat!
  (Circulation 61:1-7, 1980.)

Célkitűzés:

- Példák köré csoportosítva:
  - „Rutineljárások" buktatóinak bemutatása
  - Az egyes próbák alkalmazásának pontosítása
  - Tanulságok levonása
- A statisztikai próba elvének tisztázása
- Próba kiválasztása a:
  - kérdésfelvetéshez
  - rendelkezésre álló adatokhoz
- Néhány szó a kísérletek tervezéséről

Eszközök:

- Kiindulás a Microsoft Excelből
- Példák elérhető programokkal:
  - Excel
  - SPSS
  - StatistiX
  - PS
  - DstPlan

Ingyenesen letölthető segédprogramok

Mintaméret	DSTPLAN:	http://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=41
	PS:	http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize
	GPower3:	http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register
Excel	Box Charter:	http://peltiertech.com/Excel/Charts/BoxWhisker.html
	Better Histograms:	http://www.treeplan.com/better_down.htm

1. pl.: Két adatsor

A csop.	B csop.
499.91	500.61
513.58	513.86
509.78	512.40
503.94	501.87
509.63	511.66
504.38	508.06
495.68	499.98
494.12	496.26
495.66	496.44
500.91	501.90
506.41	506.29
494.72	496.24
510.48	510.82
501.82	505.30
496.49	498.34
505.65	506.33

(2*30 adat)

„Rutinból": 2 mintás t-próba

Kétmintás t-próba egyenlő szórásnégyzeteknél

	A csop.	*B csop.*
Várható érték	503.6	504.9
Variancia	61.3	58.3
Megfigyelések	30	30
Súlyozott variancia	59.8	
Feltételezett átlagos e	0	
df	58	
t érték	-0.643	
P(T<=t) egyszélű	0.261	
t kritikus egyszélű	1.672	
P(T<=t) kétszélű	**0.523**	
t kritikus kétszélű	2.002	

Párba állíthatók az adatok?

Kétmintás t-próba egyenlő szórásnégyzeteknél

	A csop.	*B csop.*
Várható érték	503.6	504.9
Variancia	61.3	58.3
Megfigyelések	30	30
Súlyozott variancia	59.8	
Feltételezett átlagos e	0	
df	58	
t érték	-0.643	
P(T<=t) egyszélű	0.261	
t kritikus egyszélű	1.672	
P(T<=t) kétszélű	**0.523**	
t kritikus kétszélű	2.002	

Kétmintás párosított t-próba a várható értékre

A csop.	*B csop.*
503.6	504.9
61.3	58.3
30	30
1.0	
0	
29	
-5.145	
0.0000	
1.699	
0.0000	
2.045	

Tanulság:

- Ha az adatok egyedenkénti ismételt mérésekből származnak (pl. kezelés előtt és után), párosított próbát kell végezni.

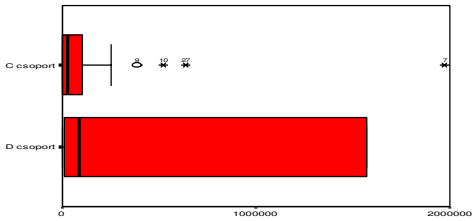
2. pl.: Ismét 2 adatsor …

C csop.	D csop.
6009	9194353
21699	457714
12288	1570493
38400	161
62862	44914
307	23868
1974844	945
251601	13727340
386505	15320464
523818	62831
104339	550644517
1136	638
2608	30213518
68741	4403455
30549	57234
89	524447
188	427643302
2559	676305
9770	3904
17497	17655
2466720	1430
52125	102120

Kétmintás t-próba egyenlő szórásnégyzeteknél

	C csop.	*D csop.*
Várható érték	5.34E+05	3.52E+07
Variancia	3.01E+12	1.55E+16
Megfigyelések	30	30
Súlyozott variancia	7.77E+15	
Feltételezett átlagos eltérés	0	
df	58	
t érték	-1.524	
P(T<=t) egyszélű	0.066	
t kritikus egyszélű	1.672	
P(T<=t) kétszélű	**0.133**	
t kritikus kétszélű	2.002	

Hát normális ez?



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
C csoport	.379	30	.000	.339	30	.000
D csoport	.463	30	.000	.310	30	.000

a. LilliefarsSignificanceCorrection

t-próba a logaritmusokra:

„Szemre" kevésbé különböznek

Kétmintás t-próba egyenlő szórásnégyzeteknél

	C csop.	*D csop.*
Várható érték	5.34E+05	3.52E+07
Variancia	3.01E+12	1.55E+16
Megfigyelések	30	30
Súlyozott variancia	7.77E+15	
Feltételezett átlagos eltérés	0	
df	58	
t érték	-1.524	
P(T<=t) egyszélű	0.066	
t kritikus egyszélű	1.672	
P(T<=t) kétszélű	**0.133**	
t kritikus kétszélű	2.002	

Kétmintás t-próba egyenlő szórásnégyzeteknél

	log(C)	*log(D)*
Várható érték	4.308	5.140
Variancia	1.779	3.135
Megfigyelések	30	30
Súlyozott variancia	2.45731	
Feltételezett átlagos eltérés	0	
df	58	
t érték	-2.056	
P(T<=t) egyszélű	0.022	
t kritikus egyszélű	1.672	
P(T<=t) kétszélű	**0.044**	
t kritikus kétszélű	2.002	

Szignifikáns!

Tanulság:

- A próbák alkalmazásának feltételei vannak
- *t*-próba végzése előtt ellenőrizni kell:
  - az eloszlások normalitását, és
  - hogy a szórások azonosak-e.
  - Jobbra ferde eloszlásoknál érdemes megpróbálni a logaritmikus transzformációt.
- A minták eloszlásáról képet kell kapnunk, mielőtt tovább dolgozunk velük.
- Tisztázni kell, hogyan is működnek a próbák.

Leíró statisztikák

- Hely („középérték")
  - Átlag
  - Medián
- Változékonyság
  - Tartomány
  - Szórás (=SD), szórásnégyzet (=variancia)
  - Negyedelő pontok közti (interkvartilis) tartomány
  - Csúcsok
  - Ferdeség

Populációs paraméterek és mintafüggvények

Populáció paramétere	**Jele**	**A becsléshez használt mintafüggvény**	**Jele**
Várható érték (expected value)	μ E(X)	Átlag (mean)	$\bar{x}$
Szórásnégyzet (population variance)	σ^2 V(X)	Korrigált tapasztalati szórásnégyzet (sample variance)	SD^2
Szórás	σ	Korrigált tapasztalati szórás (standard deviation)	SD
Eloszlás		Tapasztalati eloszlás	
Sűrűség-függvény		Gyakoriság-eloszlás (hisztogram)	

Leggyakoribbmintafüggvények

Átlag:

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

Szórás ("standard deviáció"):

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

Variációs együttható:

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

Pl: 4 adatcsoport (N, K, F, L)

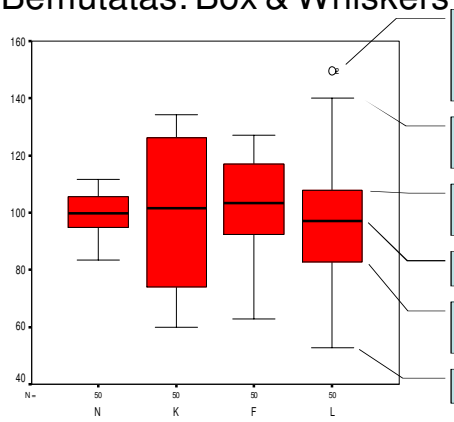
Középérték:

	N	K	F	L
Mean	99.9	100.1	103.4	95.6
5% Trimmed Mean	100.2	100.4	104.0	95.3
Median	99.6	101.6	103.3	96.8

Változékonyság:

	N	K	F	L
SE of Mean	1.0	3.8	2.1	3.1
Variance	51.3	725.7	230.9	487.1
Std. Deviation	7.2	26.9	15.2	22.1
Minimum	83.5	59.7	63.0	52.7
Maximum	111.6	134.1	127.2	149.4
Range	28.1	74.3	64.2	96.7
Interquartile Range	11.3	52.4	24.4	26.0

Bemutatás: Box & Whiskers



Kieső érték (a doboztól az 1,5-szeres hosszánál messzebb)

Legnagyobb érték (kieső elhagyva)

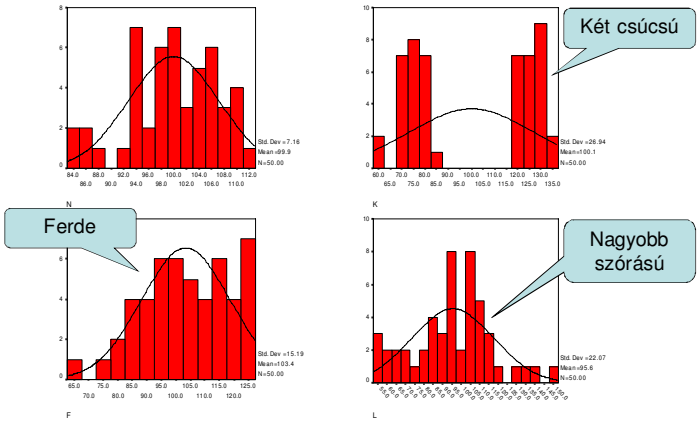
Felső negyedelő pont (kvartilis)

Medián

Alsó negyedelő pont (kvartilis)

Legkisebb érték

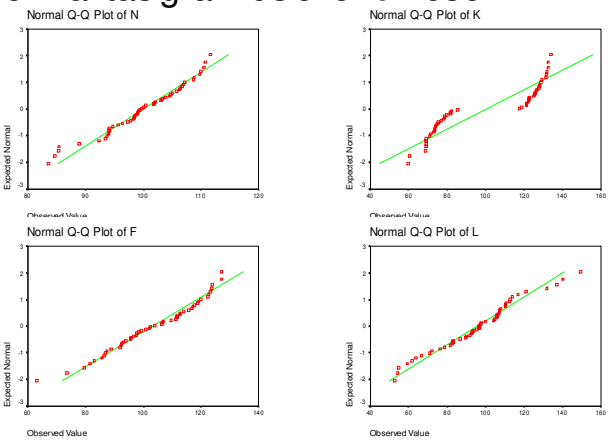
Bemutatás: Gyakorisági hisztogramok



Normalitás ellenőrzése

- Grafikus ábrázolással
  - „Box & whiskers": a negyedelő pontok és a teljes tartomány
  - Kiegyenesítés:
    - „Q-Q plot": a kvantilisek (osztóértékek) kapott és normális eloszlásnál várható értékei
- Statisztikai próbával
  - Kolmogorov-Szmirnov
    - a megfigyelt és várt eloszlás legnagyobb abszolút különbsége
  - Shapiro-Wilk
  - Ferdeség (skewness) és csúcsosság (kurtosis)
    - Várt érték normális eloszlásnál: mindkettőre 0
    - Összehasonlítás a becslés hibájának 2-szeresével

Normalitás grafikus ellenőrzése:



Normalitás-próbák

			N	K	F	L
Ferdeség	Skewness		-0.46	-0.04	-0.38	0.02
		SE	0.34	0.34	0.34	0.34
Csúcsosság (Lapultság)	Kurtosis		-0.24	-1.89	-0.42	0.03
		SE	0.66	0.66	0.66	0.66

Becslés és hibája

- Centrális határeloszlás-tétel:
  Ha a populáció tetszőleges eloszlású μ várható értékkel és σ szórással, akkor a belőle vett n elemű minta $\bar{x}$ átlaga, mint valószínűségi változó, aszimptotikusan (ha n→∞) NORMÁLIS eloszlású μ várható értékkel és σ^2/n szórásnégyzettel.
- Az átlag szórásának (standard error of the mean, SEM) becslése:

$$SEM \approx \frac{SD}{\sqrt{n}}$$

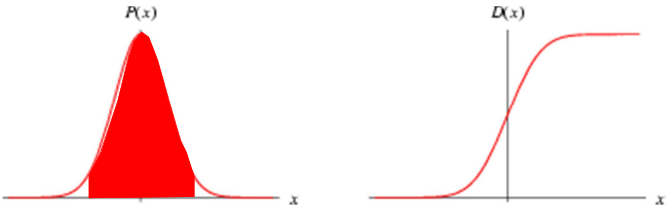
- De: a minta szórása (SD) nem követ χ^2-eloszlást

Összehasonlítás: A minta szórása és a paraméter-becslés hibája

- A minta szórása: leíráshoz
  - Az egyedek (ill. egyedi mérések) változékonyságát jellemzi
  - A mintaelemszámtól független
- A becslés hibája: becsléshez
  - Függ a felhasznált mintá(k) szórásától
  - A mintaelemszám növelésével csökkenthető
  - Ebből számolható a konfidencia-intervallum

Eloszlás- és sűrűség-függvény

Gaussian Distribution



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

$$D(x) \equiv \int_{-\infty}^{x} P(x)\,dx$$

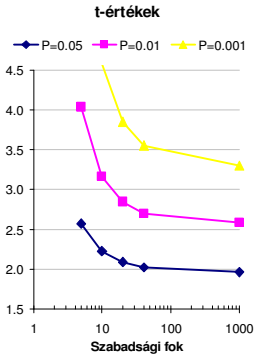
$$\int_{-\infty}^{\infty} P(x)\,dx = 1.$$

Referencia-tartomány és konfidencia-intervallum

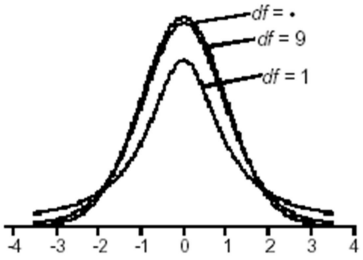
- Referencia-tartomány:
  - Az adatok eloszlásából számolható
  - Az az intervallum, melybe az adatok adott (általában 95) %-a esik
  - Normális eloszlás esetén kb. az *átlag ± 2*SD*
- Konfidencia-intervallum: $\mathrm{CI}(a,b) \equiv \int_b^a P(x)\,dx,$
  - A becslés hibájából (standard error, SE) számolható
  - Az az intervallum, melybe a populáció becsült paramétere adott (általában 95) %-os valószínűséggel esik a minta alapján
  - Az átlagra: minden eloszlás esetén kb. *becsült érték* ± 2*SE

SEM és konfidencia-intervallum: Normális vagy t-eloszlással?

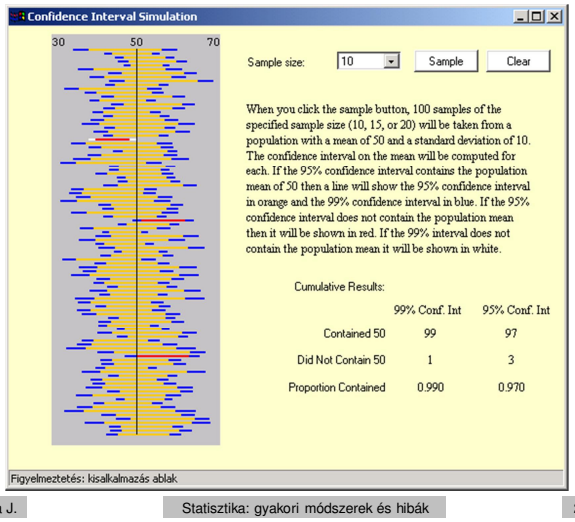
- Ha a populáció szórása (σ) ismert:
  - A becslés hibája: $SEM = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
  - Konfidencia-határok: $\pm 1.96 \cdot \sigma_{\bar{x}}$
- Ha a szórást a mintából számoljuk:
  - A becslés hibája: $sem = s_{\bar{x}} = \frac{SD}{\sqrt{n}}$
  - Konfidencia-határok: $\pm t_{n-1} \cdot s_{\bar{x}}$



t-eloszlás



Általában: >30 elemű mintából számolt átlag eloszlása már normálisnak tekinthető



Pl.: Poisson-eloszlás

Normális eloszlás:

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

Poisson-eloszlás:

$$f_{Poisson}(\mu,x) = \frac{\mu^x * e^{-\mu}}{x!}$$

,