Változók függőségi viszonyainak vizsgálata

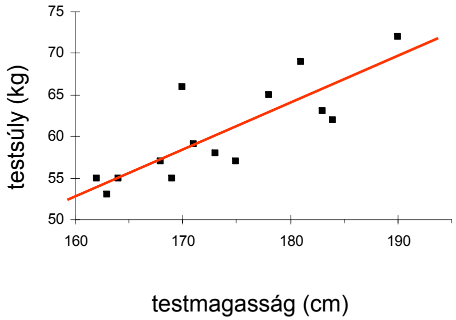
kategoriális változók

Asszociációs kapcsolat

	szemüveges	nem sz.
nő	1	3
férfi	5	3

számszerű változók

Korrelációs kapcsolat

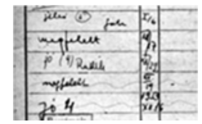


KAD 2011.11.08

Ismétlés: változók, mérési skálák típusai

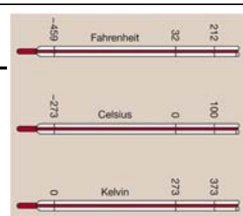
JELLEMZŐ

- kategoriális
  - nominális
  - ordinális
- számszerű
  - diszkrét
  - folytonos



jeles (5)
jó (4)
közepes (3)
elégséges (2)
elégtelen (1)

A számszerű változó skálatípusai

	diszkrét		folytonos	
intervallum-skála kivonás értelmezett, „nincs” 0 pont	naptári napok		hőmér-séklet °C-ban	
arány-skála hányados értelmezett, van 0 pont	fogak száma		hőmér-séklet K-ban	

Regresszió és korreláció

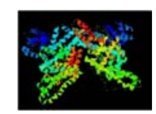
(visszatérés, hátrálás; visszafordulás) (viszony, összefüggés, kölcsönösség)

Gyakorlati megközelítés (pl.1)

mennyi fehérje van a vérplazmában?
(db, mol, g, …)

mekkora a vérplazma fehérjekoncentrációja?
(db/L, mol/L, g/L)

nm



1 db HSA molekula

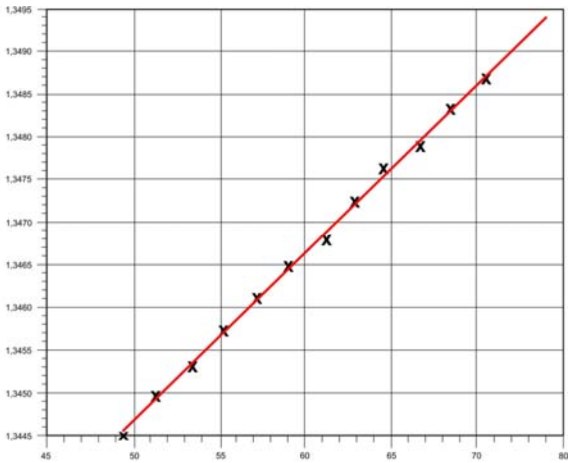
Nephrosis (súlyos vesebetegség) esetén értéke erősen lecsökken

direkt módszer: megszámolni egy adott térfogatban levő fehérje molekulák számát(?)

közvetett módszer:
keresni egy olyan (könnyen) mérhető fizikai mennyiséget, amely szigorúan monoton kapcsolatban van a megismerni kívánt mennyiséggel (legegyszerűbb ilyen függvény ...)

észrevétel:
a vérplazmában a fény lassabban halad, ha sok benne a fehérje (magas a fehérjekoncentráció), azaz nagyobb a törésmutatója (determinisztikus kapcsolat, de: mérési hiba mindig van)

plazma törésmutató



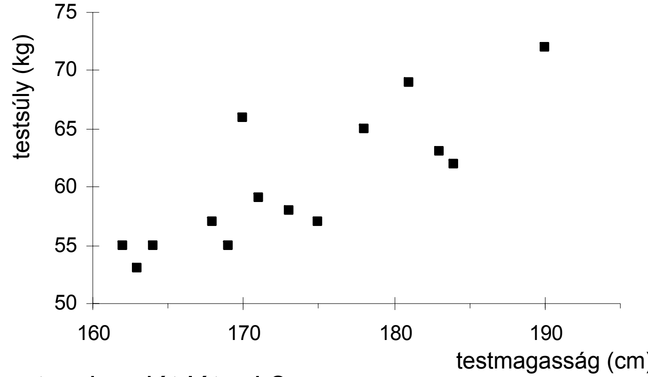
baktériumok szaporodása

a változások determinisztikus és sztochasztikus része mindig együtt fordul elő

plazma fehérje koncentráció (g/L)

(pl.2) E2 csoport (1994.09) tagjainak adatai (összetartozó értékpárok)

cm	kg
162	55
163	53
164	55
168	57
169	55
170	66
171	59
173	58
175	57
178	65
181	69
183	63
184	62
190	72

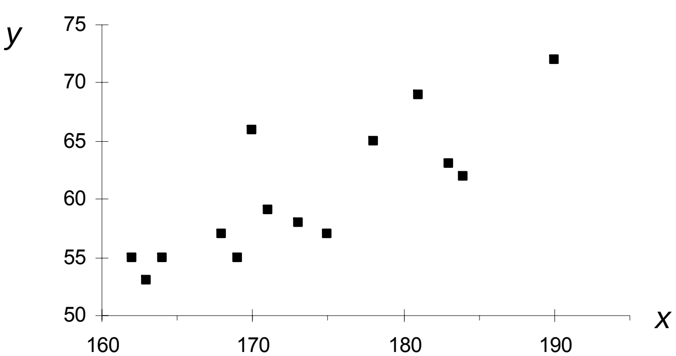


milyen tendenciát látunk?

ezen előadásfél hosszú címe:
Ugyanabban a csoportban felvett többféle kvantitatív változó közötti kapcsolat elemzése. Korreláció, lineáris regresszió, a korrelációs koefficiens fogalma

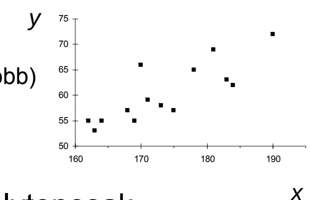
A **korrelációszámítás** két véletlen számszerű változó **szimmetrikus** kapcsolatával foglalkozik

akkor beszélünk korrelációs kapcsolatról az *x* és *y* véletlen változók között, ha vagy kis *x* értékekhez kis *y* értékek, nagy *x* értékekhez nagy *y* értékek (**pozitív** kapcsolat), vagy pedig kis *x* értékekhez nagy *y* értékek és nagy *x* értékekhez kis *y* értékek (**negatív** kapcsolat) tartoznak



itt: pozitív korreláció

Regressziós megközelítés



függvénykapcsolatot keresünk egy (vagy több) **független változó** (*x*) és egy **függő változó** (*y*) között

feltételezések: x és y számszerűek és folytonosak,
y valószínűségi változó (értékét nem csak a magyarázó változók, hanem a véletlen is befolyásolja)

A regressziós modell rögzíti a **függvény típusát**:

lineáris $y = (ax + b) + h$ (*a*: meredekség, *b*: tengelymetszet)

polinomiális $y = a + b_1x + b_2x^2 + \ldots + b_nx^n + h$

exponenciális $y = ab^x h$

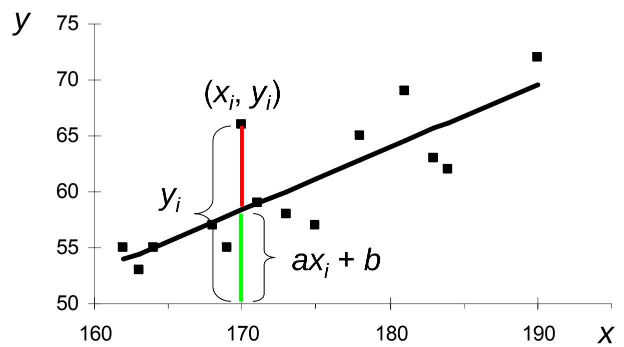
hatványfüggvényes $y = ax^b h$

és azt, hogy **hogyan hat a véletlen** a függő változóra:
Lineáris és polinomiális esetben a független változók értékétől független, **additív** ($+ h$) hibával, exponenciális és hatvány esetben **multiplikatív** ($\cdot h$) hibával.

A legegyszerűbb regressziós modell a **lineáris regresszió**

lineáris függvény: $y = (ax + b) + h$

$h_i = y_i - (ax_i + b)$ Ha a pont (x_i, y_i) az egyenes fölött van. (Hogyan írható fel, ha alatta van?)

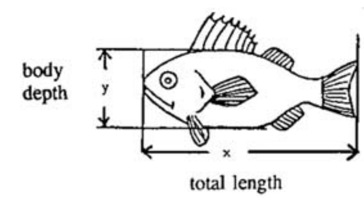


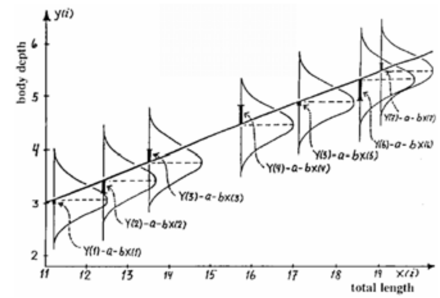
	x_i	y_i
1	162	55
2	163	53
3	164	55
4	168	57
5	169	55
6	170	66
7	171	59
8	173	58
9	175	57
10	178	65
11	181	69
12	183	63
13	184	62
14	190	72

Legjobb egyenes: hibák négyzetösszege a lehető legkisebb (**legkisebb négyzetek** módszere)

Az alkalmazhatóság feltételei

1. *x* és *y* között lineáris a kapcsolat.
2. A mintán belüli megfigyelési pontok egymástól függetlenek.
3. Minden rögzített x értékre az y értékek eloszlása normális.
4. Az y értékek eloszlása minden x értékre ugyanazzal a varianciával rendelkezik.
5. Az x értékeket „hiba nélkül” lehet mérni.





http://www.fao.org/docrep/w5449e/w5449e04.htm

a (négyzetes) **hibafüggvény**:

$$Q_{\mathrm{h}}(\ldots) = \sum_{i=1}^{n} \left[y_i - (ax_i + b)\right]^2$$

mi(k) a független változó(k)?

a és *b*

milyen a függvénykapcsolat *a*-ra és *b*-re nézve?

mindegyik változóban négyzetes a kapcsolat

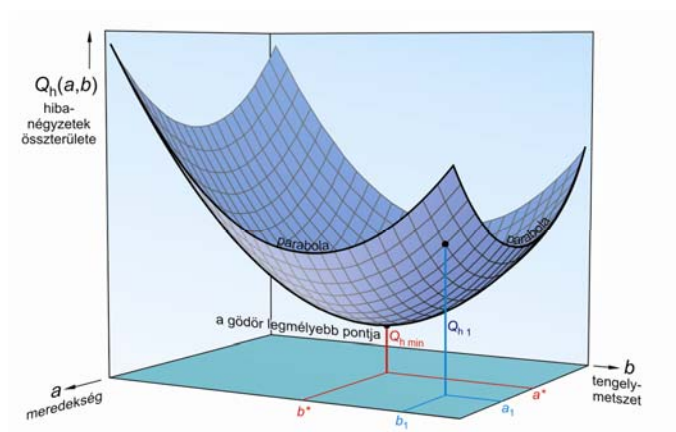
milyen függvénnyel ábrázolhatók?

különböző tágasságú parabolákkal

minimummal vagy maximummal rendelkeznek?

grafikonjuk minimummal rendelkező parabola

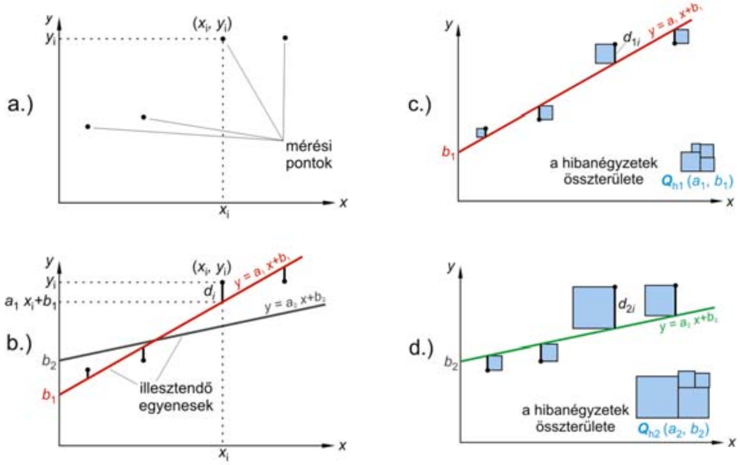
$$Q_{\mathrm{h}}(a, b) = \sum_{i=1}^{n} \left[y_i - (ax_i + b)\right]^2 = \min.$$



Gyak. jegyzet 2a fej. 14. ábra

A mérési pontokra legjobban illeszkedő egyenes ($y = ax + b$) keresése

a: meredekség
b: tengelymetszet



Gyak. jegyzet 2a fej. 13. ábra

Lineáris regresszió

$$Q_h(a, b) = \sum_{i=1}^{n} \left[y_i - (a x_i + b) \right]^2$$ hibafüggvény minimalizása

megoldási lehetőségek:

1. teljes négyzetté kiegészítés

   pl. $y = x^2 - 6x + 14 = (x-3)^2 + 5$, minimum $x = 3$-nál

2. differenciálszámítás

   differenciálhányados: az érintő iránytangense

   szélsőérték keresés: ahol a görbének minimuma (vagy maximuma) van, ott az érintő iránytangense zérus

   a szerinti és *b* szerinti differenciálhányadosok zérusok,
   2 egyenlet, 2 ismeretlen (2 ismeretlenes lineáris egyenletrendszer)

a „legjobb” meredekség:

($y = ax + b$)

$$\boxed{a^* = \frac{Q_{xy}}{Q_{xx}}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{vagy} \qquad a^* = \frac{s_{xy}^2}{s_x^2}$$

a „legjobb” tengelymetszet:

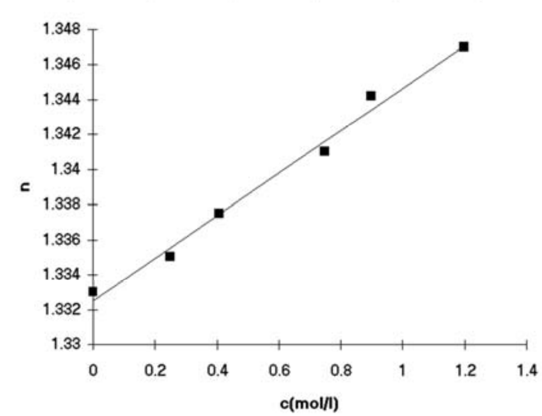
$$\boxed{b^* = \bar{y} - a^* \cdot \bar{x}} = \frac{\sum_{i=1}^{n} y_i}{n} - a^* \frac{\sum_{i=1}^{n} x_i}{n}$$

ahol $s_{xy}^2 = \dfrac{Q_{xy}}{n-1}$: kovariancia

Példa: refraktometria

(ismeretlen koncentráció meghatározása kalibrációs egyenes segítségével)

	A	B	C
1–7	ismeretlen sorszáma	koncentráció	törésmutató
8	i	c (mol/l)	n
9		0	1.333
10		0.25	1.3350
11		0.41	1.3375
12		0.75	1.3410
13		0.90	1.3442
14		1.20	1.3470
15	4	0.19	1.3348
16	10	1.13	1.3462
17	n=ac+n0		
18	n0=	1.3325	
19	a=	0.0121	
20	r=	0.995	



Milyen a pontok illeszkedése a regressziós egyeneshez?

ehhez a **korrelációszámítás** nyújt segítséget
(két véletlen változó szimmetrikus kapcsolatával foglalkozik)

a változók közötti kapcsolat erősségét vizsgálja (van erős és gyenge korreláció)

korrelációs együttható
(Pearson-féle)

$$r = \frac{Q_{xy}}{\sqrt{Q_{xx} \cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}$$

a számláló megegyezik a regressziós egyenes meredekségének számlálójával (a nevező mindkét esetben pozitív)

$$a^* = \frac{Q_{xy}}{Q_{xx}}$$

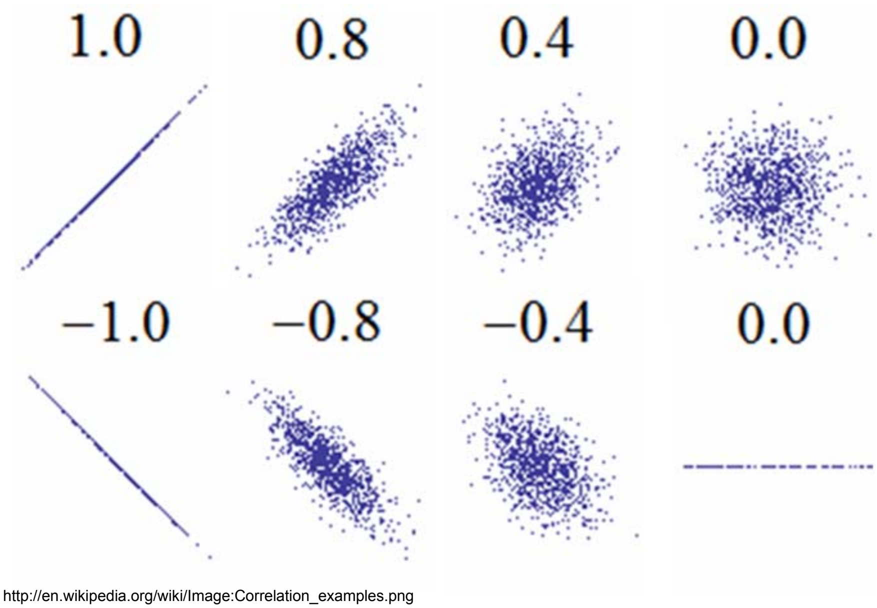
pozitív meredekség: $r > 0$ (pozitív korreláció)

negatív meredekség: $r < 0$ (negatív korreláció)

$-1 \le r \le 1$ $0 \le r^2 \le 1$

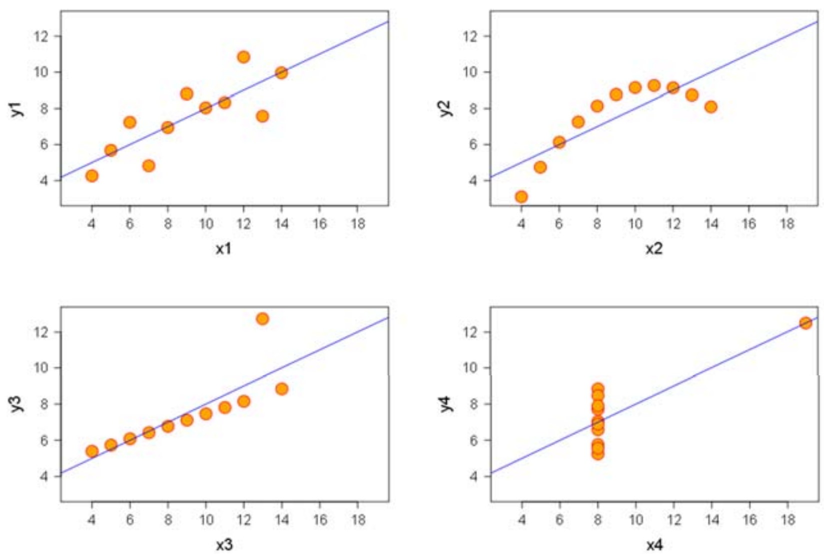
meghatározottsági együttható
(mennyire lehet az egyikből a másikat előre jelezni)

Példák korrelációs együtthatókra

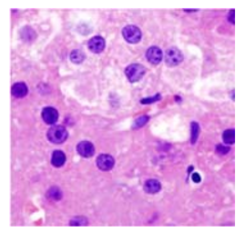


http://en.wikipedia.org/wiki/Image:Correlation_examples.png

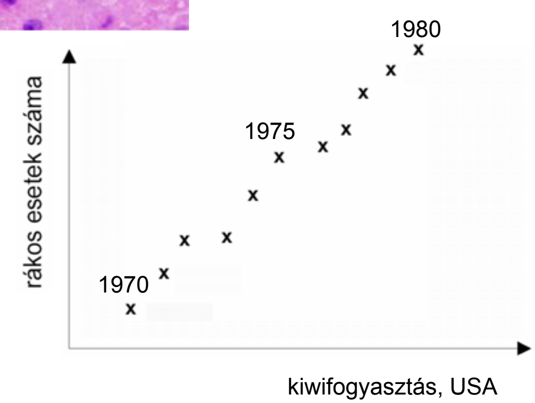
Extrém példa: *r* = 0.816, *y* = 3 + 0.5*x* (Anscombe)



http://en.wikipedia.org/wiki/Anscombe%27s_quartet



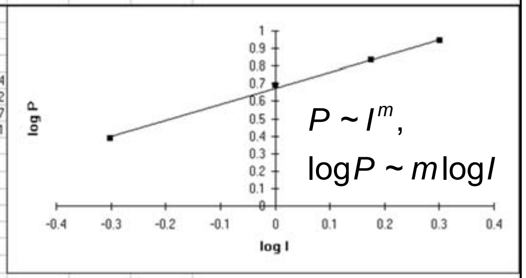
A korreláció jelenléte **nem** (feltétlenül) jelent **oksági kapcsolatot**



idő, mint rejtett változó

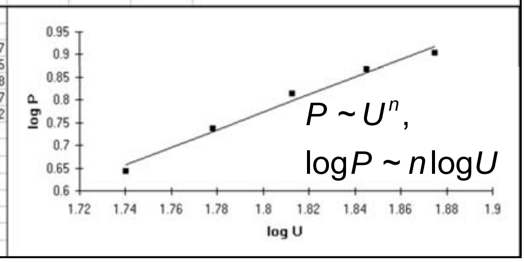
Példa: **Hatványfüggvényes regresszió visszavezetése lineáris regresszióra.** Röntgencső sugárteljesítményének mérése

	A	B	C	D	E
1	D1	97-10-16			
2					
3	I	P	logI	logP	
4	mA	r.e.			
5	0	0			
6	0.5	2.45	-0.30103	0.389166	0.396834
7	1	4.9	0	0.690196	0.673612
8	1.5	6.8	0.176091	0.832509	0.835517
9	2	8.8	0.30103	0.944483	0.950391
10					
11					
12	m=	0.919438	0.673612		
13		0.030392	0.007001		
14		0.99782	0.013744		
15		915.2208	2		
16		0.172873	0.000378		
17	r=	0.998909			
18					
19	U	P	logU	logP	
20	kV	r.e.			
21	55	4.4	1.740363	0.643453	0.656607
22	60	5.45	1.778151	0.736397	0.729975
23	65	6.5	1.812913	0.812913	0.797468
24	70	7.35	1.845098	0.866287	0.859957
25	75	8	1.875061	0.90309	0.918132
26					
27					
28	n=	1.941564	-2.72242		
29		0.145415	0.263339		
30		0.98345	0.015483		
31		178.2719	3		
32		0.042735	0.000719		
33	r=	0.991691			



$P \sim I^m$,

$\log P \sim m \log I$



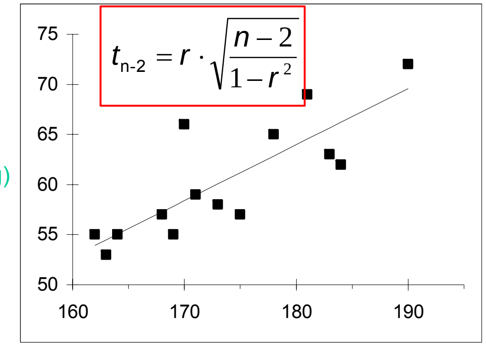
$P \sim U^n$,

$\log P \sim n \log U$

Korrelációs t-próba

Van-e kapcsolat a két mennyiség között?

cm	kg		E2
162	55	53.93	1994.09
163	53	54.49	
164	55	55.05	
168	57	57.28	
169	55	57.84	
170	66	58.39	
171	59	58.95	
173	58	60.07	
175	57	61.19	
178	65	62.86	
181	69	64.54	
183	63	65.65	
184	62	66.21	
190	72	69.56	
	a=	0.558	-36.5096 =b
		0.113	19.66358
r=	0.819	0.67	3.492297
n=	14	24.36	12
t=	4.935	297.1	146.3537



$$t_{n-2} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

m (kg)

h (cm)

H_0: nincs kapcsolat

Határozzuk meg pontosabban!

$$|t| = 4.935 > t_{12,\,krit(0,05)} = 2.18 \Rightarrow H_0 \text{ hamis } (p<=0.05)$$

Asszociációs kapcsolat. Khi-négyzet teszt (1)

Példa 1

	szemüveges	nem sz.	összesen
nő	28	75	103
férfi	48	49	97
	76	124	200

?

Kapcsolatvizsgálat kategorikus változók között. Khi-négyzet teszt

gyakorisági táblázat (kontingencia táblázat): két változó közös gyakoriságának táblázatos ábrázolása

X (pl. nem) és *Y* (szemüvegesség)

	szemüveges	nem sz.	összesen
nő	a=28	b=75	103
férfi	c=48	d=49	97
	76	124	200

kérdés: különbözik-e egy rögzített tulajdonság gyakorisága a két csoportban?

A nullhipotézis felállítása

H_0: nem és szemüvegesség egymástól függetlenek (nincs különbség a csoportokban)

$\frac{a}{b}=\frac{c}{d}$, vagy $\frac{a}{c}=\frac{b}{d}$

mekkora lenne a **várt gyakoriság** (expected frequency) a bal felső (*a*) cellában, ha a nullhipotézis igaz?

a nők száma:
a + *b* = 103

a szemüveges személyek száma:
a + *c* = 76

a nők aránya a mintában:
p(nő) = (*a* + *b*)/*n*= 103/200

a szemüvegesek aránya a mintában :
p(szemüveges) = (*a* + *c*)/*n*= 76/200

	szemü-veges	nem sz.	össze-sen
nő	a=28	b=75	103
fér-fi	c=48	d=49	97
	76	124	200

a megfigyelt (observed) gyakoriságok táblázata

Várt gyakoriságok. feltevés: H_0 igaz ⇨ a nem és a szemüvegesség független tulajdonságok

$$\text{várt gyakoriság a bal felső cellában}: \frac{a+b}{n}\cdot\frac{a+c}{n}\cdot n=\frac{(a+b)\cdot(a+c)}{n}$$

$$\text{várt gyakoriság a jobb felső cellában}: \frac{a+b}{n}\cdot\frac{b+d}{n}\cdot n=\frac{(a+b)\cdot(b+d)}{n}$$

$$\text{várt gyakoriság a bal alsó cellában}: \frac{c+d}{n}\cdot\frac{a+c}{n}\cdot n=\frac{(c+d)\cdot(a+c)}{n}$$

$$\text{várt gyakoriság a jobb alsó cellában}: \frac{c+d}{n}\cdot\frac{b+d}{n}\cdot n=\frac{(c+d)\cdot(b+d)}{n}$$

	sz.	nem	össz.
n	a=28	b=75	103
f	c=48	d=49	97
	76	124	200

megfigyelt (observed) kontingencia táblázat

	sz.	nem	össz.
n	103*76/200	103*124/200	103
f	97*76/200	97*124/200	97
	76	124	200

várt (expected) kontingencia táblázat

A várt gyakoriságok a megfigyelt gyakoriságokból

	sz.	nem	össz.
n	a=28	b=75	103
f	c=48	d=49	97
	76	124	200

megfigyelt (observed) kontingencia táblázat

	sz.	nem	össz.
n	103*76/200	103*124/200	103
f	97*76/200	97*124/200	97
	76	124	200

várt (expected) kontingencia táblázat

$$(\text{várt gyakoriság})=\frac{(\text{oszlopösszeg})\cdot(\text{sorösszeg})}{(\text{a minta elemszáma})}$$

Próbastatisztika

Ha a nullhipotézis igaz:
A megfigyelt és a várt gyakoriságokat tartalmazó kontingencia táblázatok megfelelő celláiban levő értékek nagyjából egyformák. A következő próbastatisztika (súlyozott négyzetes közép) **khi-négyzet eloszlású**:

$$\boxed{\chi^2=\sum_i\frac{(O_i-E_i)^2}{E_i},}$$

ahol
O_i a megfigyelt (observed)
E_i a(z el)várt gyariságok
az i-dik cellában.
Szabadsági fok: (sorok száma –1)*(oszlopok száma –1)

pl. 2*2 (négymezős-) táblázat: 1

A teszt végrehajthatóságának feltételei

n (a minta elemszáma) elegendően nagy:

a *várt gyakoriságokat* tartalmazó kontingencia táblázatban minden cellatartalomnak 1-nél nagyobbnak kell lenni

a *várt gyakoriságokat* tartalmazó kontingencia táblázatban azoknak a celláknak a száma, amelyekben a cellatartalom 1 és 5 közötti csak a cellák 20%-a lehet

(pl. négymeztős táblázat: minden cellában a cellatartalomnak 5-nél nagyobbnak kell lenni)

Speciális eset: négymezős táblázat (gyakorlati jegyzet 2.b.29)

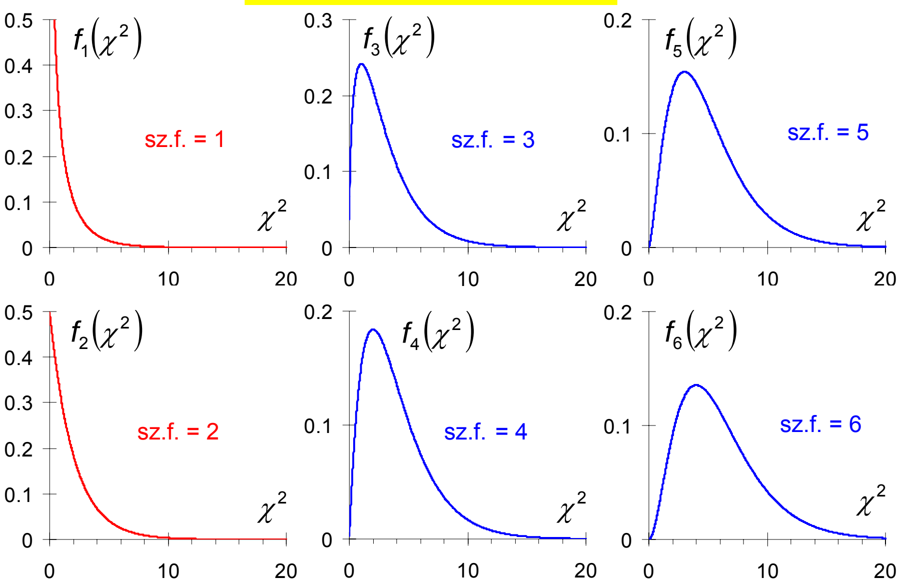
	a vizsgált tulajdonság		összesen
	megvan	nincs meg	
A csoport	***a***	***b***	*a+b*
B csoport	***c***	***d***	*c+d*
összesen	*a+c*	*b+d*	*n*

$$\chi^2_{\mathrm{M}} = \frac{n\cdot(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

a végrehajthatóság felétele:

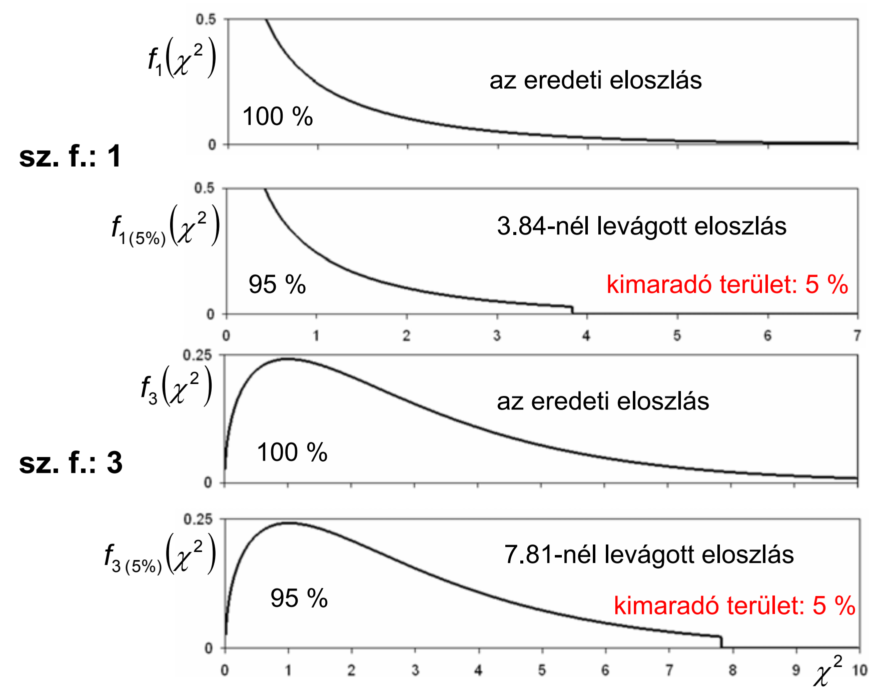
a két legkisebb részösszeg szorzata legyen nagyobb, mint 5*n*

Khi-négyzet eloszlások



módusz = 0, ha sz.f.=1 vagy 2

módusz = (sz.f.-2), ha sz.f. >2



szabadsági fok	0,99	0,975	0,95	0,05	0,025	0,01	0,001
	p (valószínűség)						
1	0,0000157	0,0000982	0,000393	3,84	5,02	6,63	10,83
2	0,0201	0,0506	0,103	5,99	7,88	9,21	13,82
3	0,115	0,216	0,352	7,81	9,35	11,34	16,27
4	0,297	0,484	0,711	9,49	11,14	13,28	18,47
5	0,554	0,831	1,15	11,07	12,83	15,09	20,51
6	0,872	1,24	1,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	14,07	16,01	18,47	24,32
8	1,65	2,18	2,73	15,51	17,53	20,09	26,13
9	2,09	2,70	3,33	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	18,31	20,48	23,21	29,59
11	3,05	3,61	4,57	19,68	21,92	24,72	31,26
12	3,57	4,40	5,23	21,03	23,34	26,22	32,91
13	4,11	5,01	5,89	22,36	24,74	27,69	34,53
14	4,66	5,63	6,57	23,68	26,12	29,14	36,12

Példa 1

A teszt alkalmazhatóságának feltétele:
a két legkisebb részösszeg szorzata legyen nagyobb, mint 5n

	szemü-veges	nem sz.	össze-sen
nő	a=28	b=75	103
fér-fi	c=48	d=49	97
	76	124	200

76*97 = 7372 > 5*200 =1000

a khi-négyzet teszt használható

$$\chi^2_M = \frac{200\cdot(28\cdot 49 - 48\cdot 75)^2}{76\cdot 124\cdot 103\cdot 97} = 10.54$$

10.54 > χ^2_{krit} =3,84 H_0 hamis

van kapcsolat a nem és a szemüvegesség (szemüvegviselési hajlandóság!) között

szabadsági fok	0,99	0,975	0,95	0,05	0,025	0,01	0,001
	p (valószínűség)						
1	0,0000157	0,0000982	0,000393	3,84	5,02	6,63	10,83

$$\chi^2_M = \frac{200\cdot(28\cdot 49 - 48\cdot 75)^2}{76\cdot 124\cdot 103\cdot 97} = 10.54$$

10.54 > χ^2_{krit} =3.84 H_0 hamis

10.54 > χ^2_{krit} =6.63 H_0 hamis

elvetjük a nullhipotézist, szignifikancia szint: <0.01

	A	B	C	D
1	megfigyelt gyakoriságok			
2		szemüveges	nem sz.	
3	nő	28	75	=SUM(B3:C3)
4	férfi	48	49	=SUM(B4:C4)
5		=SUM(B3:B4)	=SUM(C3:C4)	=SUM(D3:D4)
6	várt gyakoriságok			
7		szemüveges	nem sz.	
8	nő	=$D3*B$5/$D$5	=$D3*C$5/$D$5	=SUM(B8:C8)
9	férfi	=$D4*B$5/$D$5	=$D4*C$5/$D$5	=SUM(B9:C9)
10		=SUM(B8:B9)	=SUM(C8:C9)	=SUM(D8:D9)
11			szignifikanciaszint:	=CHITEST(B3:C4,B8:C9)
12			khi-négyzet érték:	=CHIINV(D11,1)

	A	B	C	D
1	megfigyelt gyakoriságok			
2		szemüveges	nem sz.	
3	nő	28	75	103
4	férfi	48	49	97
5		76	124	200
6	várt gyakoriságok			
7		szemüveges	nem sz.	
8	nő	39.14	63.86	103
9	férfi	36.86	60.14	97
10		76	124	200
11			szignifikanciaszint:	0.0012
12			khi-négyzet érték:	10.544

számolás Excel-lel

angol		*magyar*
SUM	=	SZUM
CHITEST	=	KHI.PRÓBA
CHIDIST	=	KHI.ELOSZLÁS
CHIINV	=	INVERZ.KHI

példa 2

	sz.	nem sz.	össz-szes
nő	1	3	4
férfi	5	3	8
	6	6	12

?

4*6 = 24 < 5*12 =60

a khi-négyzet teszt nem használható

(helyette: Fisher egzakt teszt)

a minta elemszámának növelése

	sz.	nem	össz.
nő	1	3	4
férfi	5	3	8
	6	6	12

12 → 200

	sz.	nem	össz.
nő	28	75	103
férfi	48	49	97
	76	124	200

nők

$$\frac{n_{sz}}{n_{nem}} = \frac{1}{3} = 0.33 \qquad \frac{n_{sz}}{n_{nem}} = \frac{28}{75} = 0.37$$

férfiak

$$\frac{n_{sz}}{n_{nem}} = \frac{5}{3} = 1.67 \qquad \frac{n_{sz}}{n_{nem}} = \frac{48}{49} = 0.98$$

sejtésünk van, de nem tudjuk igazolni

n növelésével (12 → 200): a sejtés igazolható lesz

Példa 3 (biofizika jegyzet 102. példa). Nem artériás típusú ischaemiás opticus neuropathia sikeres műtéti korrekciójáról jelent meg 1989-ben egy közlemény. Minthogy e betegségben korábban semmiféle hatásos kezelési módszer nem volt ismert, ezt a műtétet sok helyen alkalmazni kezdték. Rövidesen eredménytelen beavatkozásokról is megjelentek beszámolók, ezért számbavették 25 klinikai centrum 244 ilyen betegét, akik közül 119 főn elvégezték a műtétet, 125 betegen nem. A felmérés eredménye:

megfigyelt gyakoriságok

	műtött	nem m.	össz.
javult	39	53	92
változatlan	52	56	108
romlott	28	16	44
összes	119	125	244

várt gyakoriságok

	műtött	nem m.	össz.
javult	45	47	92
változatlan	53	55	108
romlott	21	23	44
összes	119	125	244

khi^2 = $(39–44.87)^2/44.87+(53–47.13)^2/47.13 +(52–52.67)^2/52.67+(56–55.33)^2/55.33 +(28–21.46)^2/21.46+(16–22.54)^2/22.54 = 5.407$

Mivel $5.407 < 5.991 = \chi^2_{\text{krit, sz.f.=2}}$, ezért nem vethetjük el a nullhipotézist. Azaz a mintánk alapján nincs okunk feltételezni különbséget a két módszer (műtét ill. nem műtét) hatásossága között.

egydimenziós kontingencia táblázatokkal kapcsolatos kérdés: a megfigyelt értékek illeszkednek-e egy feltételezett eloszláshoz?

Illeszkedésvizsgálat. Khi-négyzet teszt (2)

tiszta illeszkedésvizsgálat

(a gyakoriságokat ismert valószínűségekből kapott gyakoriságokkal hasonlítjuk össze)

- egyenletes eloszlásra történő i.v.
- egyéb ismert paraméterű eloszlásra történő i.v.

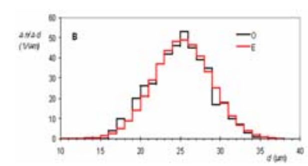
kockafeldobás eredménye

1	2	3	4	5	6
21	14	14	19	16	16

becsléses illeszkedésvizsgálat

(az eloszlás típusa alapján a megfigyelt gyakoriságokból becsüljük az eloszlás paramétereit)

- normalitás-vizsgálat
- egyéb becsült paraméteres i.v.



Egyenletes eloszlásra történő illeszkedésvizsgálat

A megfigyelt gyakoriságokat tartalmazó kontingencia táblázatot (bekeretezett rész, *O*) kibővítjük a várt gyakoriságokat tartalmazó segéd-kontingencia táblázattal (*E*). Feltételezzük, hogy a kocka nem cinkelt (H_0), ezért a 6 lehetséges esemény egyforma gyakoriságú: 100/6 = 16.7

a kockafeldobás eredménye

	1	2	3	4	5	6	össz.
O	21	14	14	19	16	16	100
E	16.7	16.7	16.7	16.7	16.7	16.7	100

khi^2= $(21–16.7)^2/16.7 + (14–16.7)^2/16.7 + (14–16.7)^2/16.7 + (19–16.7)^2/16.7 + (16–16.7)^2/16.7 + (16–16.7)^2/16.7 =$ $= 2.36 < 11.07 = \chi^2_{krit,\ sz.f.=5}$, a nullhipotézist megtartjuk.
A kocka nem cinkelt.

Normalitásvizsgálat

diszkrét számszerű változó esetén (ha nem diszkrét, akkor azzá tesszük)

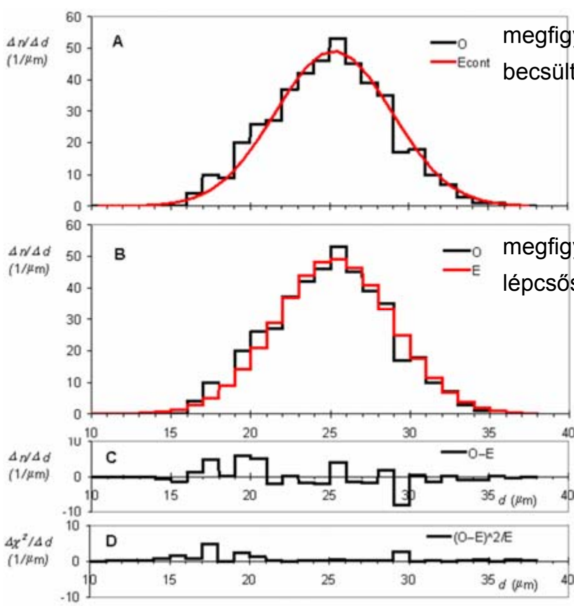
béka vörösvérsejtek hosszabbik átmérője (450 mérési adat)

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	össz.
O	4	10	9	20	26	27	37	42	46	53	45	39	35	17	18	10	7	3	1	1	450

nullhipotézis: a minta **normális eloszlású** populációból származik
az eloszlás elméleti értékeit a **tapasztalati értékekkel becsüljük**
μ becslése a számtani közép, avg = 24.76 μm
σ becslése az adatok tapasztalati szórása *s* = 3.65 μm
görbe alatti terület (1 ill. *n*): a sűrűségfv-t meg kell szorozni *n* = 450-nel

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	össz.
E	2.78	5.17	8.91	14.3	21.2	29.1	37.2	44.1	48.5	49.4	46.8	41	33.4	25.2	17.7	11.5	6.93	3.87	2.01	0.97	450

szabadsági fokok száma : *n*–*b*–1, *n* az osztályok száma, *b* az eloszlás paramétereinek száma (21 – 2 –1 = 18). A maximálisan méltányolható első fajta hibának vegyünk 5 %-ot! 18 szabadsági fok esetén ehhez a szignifikancia szinthez tartozó χ^2_a érték: 28.87.
khi^2 = $(4–2.78)^2/2.78 + (10–5.17)^2/5.17 + ... = 12.9$. Mivel 12.9 < 28.87, nincs okunk a nullhipotézis elvetésére.
Táblázatkezelővel: megkapjuk azt a szignifikancia szintet, amely mellett "elvethetnénk" a nullhipotézist, ez pedig 0.842 = 84.2 %.
A béka vörösvérsejtjeinek hosszabbik átmérője normális eloszlású.



megfigyelt gyakoriságok
becsült eloszlás sűrűségfüggvény

megfigyelt gyakoriságok
lépcsőssé tett becsült eloszlás ss. fv.

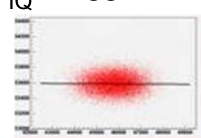
megfigyelt és várt függvények különbségei

khi-négyzet érték szemléletes ábrázolása (a görbe alatti terület)

Függőségi viszonyok lehetőségei

függetlenség

IQ



testmagasság

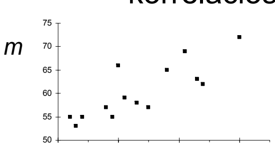
függőség

sztochasztikus viszony

determinisztikus viszony

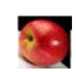
korrelációs | vegyes | asszociációs

m



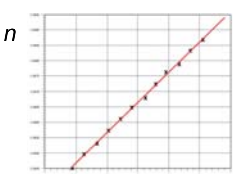
testmagasság

íz



színezettség

n



koncentráció

számszerű | ordinális | nominális | számszerű