Szent István Egyetem Állatorvos-tudományi Kar
Biomatematikai és Számítástechnikai Tanszék

14. előadás

Regresszió- és korrelációanalízis.

Nempraméteres próbák.

Fodor János
Biomatematika 2001

Copyright © 2001 jfodor@univet.hu
Last Revision Date: December 11, 2001 Version 1.10

Table of Contents

1 Lineáris regresszió **4**

2 A hibatag **10**

3 Az α és β paraméterek becslése **11**

3.1 ε szórásának becslése 18

4 A paraméterek konfidencia-intervallumai **19**

5 A korrelációs és determinációs együttható **20**

6 Nemparaméteres próbák 34

6.1 Illeszkedésvizsgálat 35

6.2 Normalitásvizsgálat 46

6.3 Homogenitásvizsgálat 50

1. Lineáris regresszió

Eddig egyszerre csak egy valószínűségi változót vizsgáltunk. Most két valószínűségi változó közti kapcsolatról, összefüggésről lesz szó.

Példa.

Hogyan függ a szarvasmarhák övmérete a testsúlyuktól?

10 elemű mintánkat az alábbi táblázat tartalmazza.

testsúly	övméret
440	180
442	182
450	182
458	183
460	184
470	184
475	185
480	186
490	187
495	188

Általában, minél nagyobb a testsúly, annál nagyobb az övméret. Legyen X a testsúlya, Y pedig az övmérete ugyanannak az állatnak. Feltesszük, hogy a kapcsolat **lineáris** a két változó között, amely a következő egyenlettel írható le:

$$Y = \beta \cdot X + \alpha + \varepsilon,$$

ahol α és β két ismeretlen paraméter. Ezek mutatják meg, hogy milyen a kapcsolat a két változó között. A regresszióanalízis segítségével becsüljük meg a két paraméter értékét.

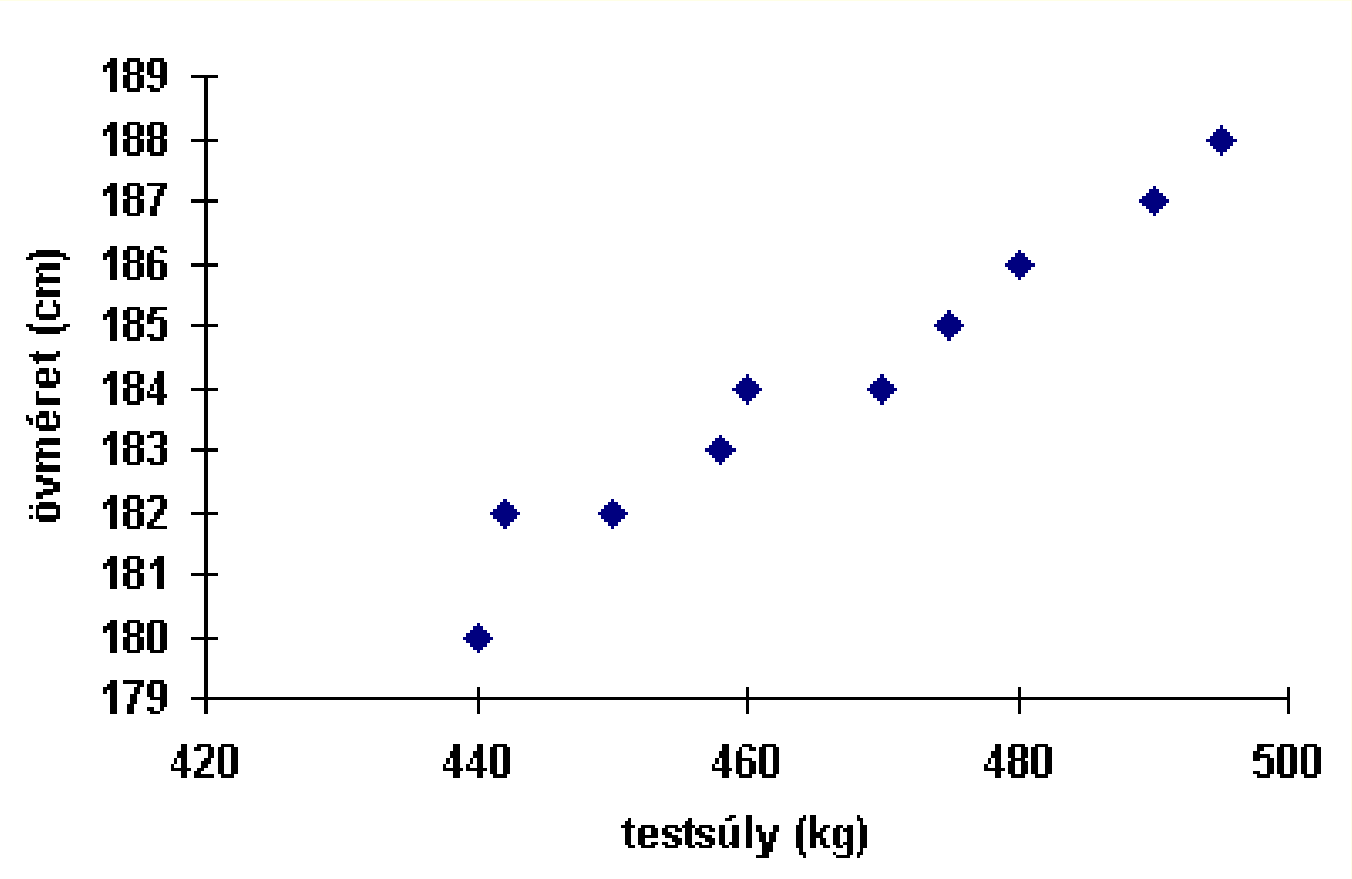
Az ε pedig a véletlentől függő hibatag. Ez okozza a különbözőségeket egy adott X értékhez tartozó Y értékek között.

X a **magyarázó**, vagy **független** változó, míg Y a **magyarázott**, vagy **függő** változó.

A két változó közti kapcsolat **pozitív**, ha növekvő X értékekhez egyre nagyobb Y értékek tartoznak (azaz $\beta > 0$).

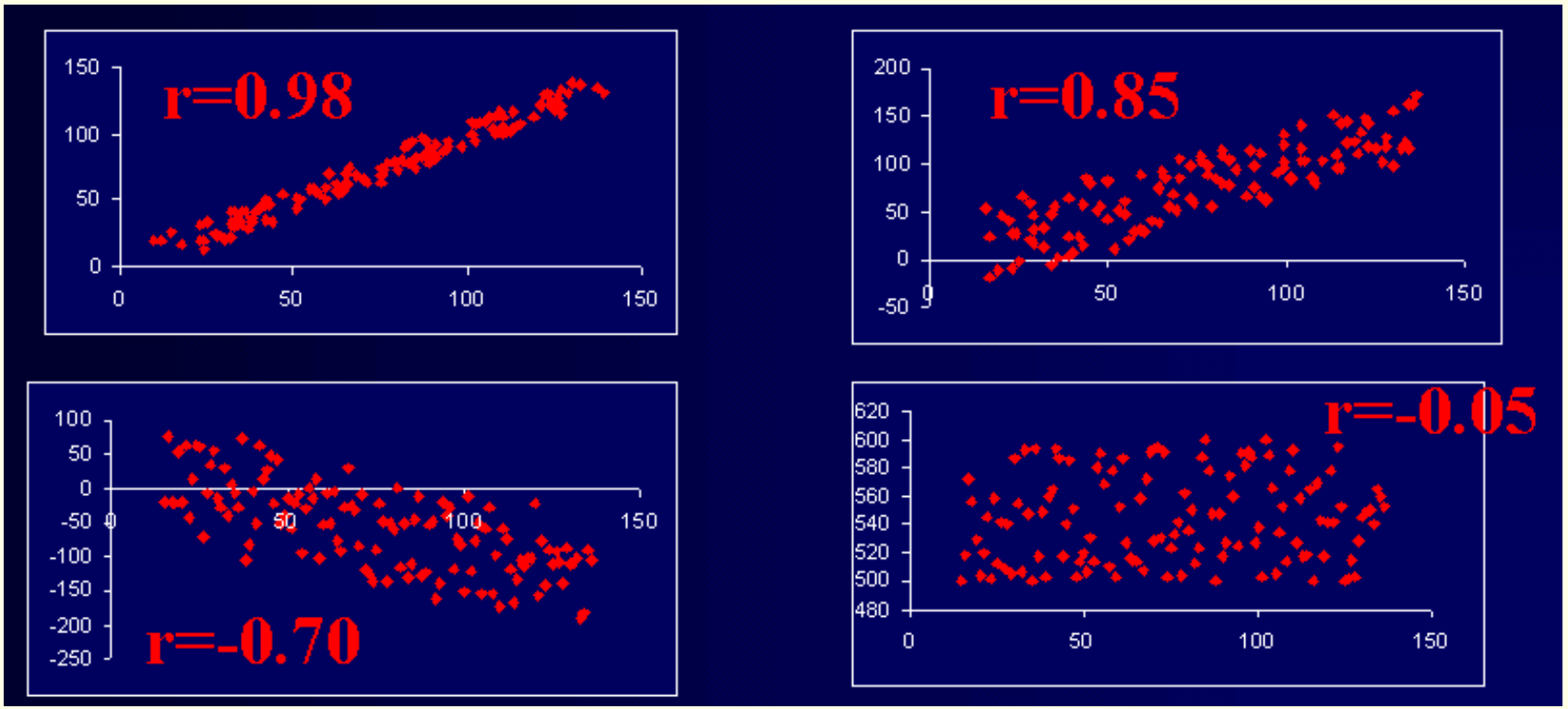
Negatív kapcsolat esetén pedig növekvő X értékekhez egyre kisebb Y értékek tartoznak (vagyis $\beta < 0$).

Készítsünk **szórásdiagramot** a fenti táblázat alapján:



Jól látható az összefüggés pozitív jellege.

A következő ábrákon példát láthatunk pozitív, negatív összefüggésre, illetve olyan esetre, amikor nincs összefüggés a két változó között ($\beta = 0$).



2. A hibatag

Az ε véletlen hibatag származhat a mérés pontatlanságából, elhanyagolt hatásokból, valamint a mintavételi hibából. A következőket tételezzük fel ε-ról:

1. ε várható értéke nulla. (Ezért Y várható értéke $\beta X + \alpha$.)
2. ε szórása minden megfigyelt értékre ugyanaz.
3. ε értékei függetlenek az X változótól és egymástól (azaz, minden X érték esetén a fellépő ε értékek függetlenek egymástól). Ezért Y ak-

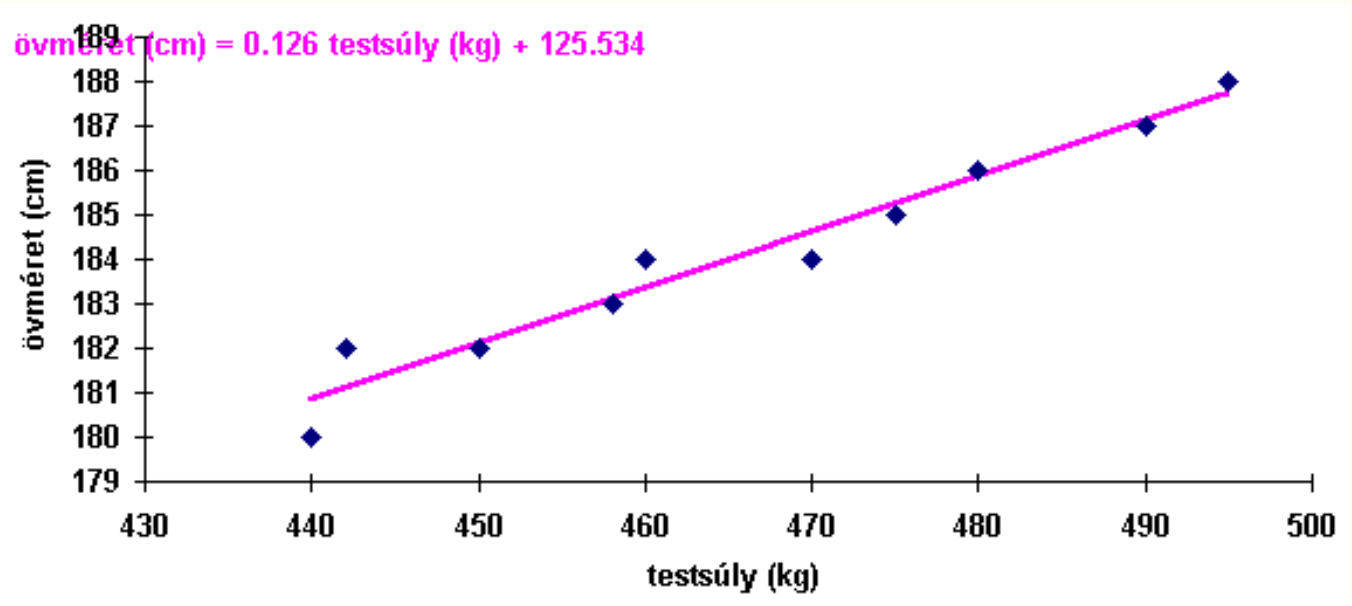
tuális értékét csak a hozzá tartozó X érték befolyásolja, a többi nem.

4. ε normális eloszlású minden X esetén (így adott X esetén Y normális eloszlású).

3. Az α és β paraméterek becslése

Ezeket az X és Y mért értékeiből kell becsülnünk, a **legkisebb négyzetek módszerével**.

A mért értékpárokat reprezentáló (x_i, y_i) pontok és az ezekre legjobban illeszkedő egyenes:



Tehát y_i a mért pontok második koordinátái, Y_i pedig az egyenes x_i-hez tartozó pontjának második koordinátája. A legkisebb négyzetek módszere azt

az egyenest határozza meg, amelyre a

$$\sum_{i=1}^{n}(Y_i - y_i)^2$$

kifejezés a legkisebb. Ez az összeg a becslés hibája.

Jelöljük az α és β becsült értékeit a-val (ez a tengelymetszet) és b-vel (ez a regressziós együttható). Ezek értékei a minta alapján:

$$\boxed{b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},}$$

$$\boxed{a = \bar{y} - b\bar{x},}$$

ahol n a megfigyelések száma.

A példánkban: $b = 0,125$ és $a = 125.5$.

Nyilvánvaló, hogy a és b értéke függ azon pontoktól, amelyekre illesztjük az egyenest. Tehát a és b is

valószínűségi változó.

A b eloszlása normális, várható értéke β, szórása pedig

$$\frac{\sigma}{s_x \cdot \sqrt{n}},$$

ahol

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

σ pedig az ε hiba elméleti szórása.

Az a eloszlása normális, várható értéke α, szórása pedig

$$\frac{\sqrt{\sigma \overline{x^2}}}{s_x \cdot \sqrt{n}},$$

ahol még

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

Megfigyelések:

- Ha ε szórása (σ) kicsi, akkor b szórása kicsi.
- Minél több adatunk van, b szórása annál kisebb (sok adat esetén a pozitív és negatív hibák kiegyenlítik egymást).
- b szórása annál kisebb, minél nagyobb X szórása.

3.1. ε szórásának becslése

Legyen $e_i := y_i - (\alpha + \beta x_i)$. Ennek alapján az ε szórásának becslése:

$$\boxed{s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}}$$

Itt $n-2$-vel osztunk, mert két becsült paraméter szerepel a képletben.

Példánkban $s_\varepsilon = 0.558$.

Mivel ismerjük ε szórásának becslését, meg tudjuk becsülni a és b szórását is:

$$\boxed{s_b = \frac{s_\varepsilon}{\sqrt{n} \cdot s_x}}, \qquad \boxed{s_a = \frac{s_\varepsilon \cdot \overline{x^2}}{\sqrt{n} \cdot s_x}}$$

Példánkban $s_a = 4.52$ és $s_b = 0.0097$.

4. A paraméterek konfidencia-intervallumai

Az a és b paraméterekre vonatkozó 95%-os konfidenciaintervallumok:

$$\boxed{[b - 2s_b, b + 2s_b]}, \qquad \boxed{[a - 2s_a, a + 2s_a]}$$

Magához az illesztett egyeneshez is megadható 95%-os konfidenciaintervallum (konfidencia sáv).

5. A korrelációs és determinációs együttható

Kérdés: az előbbi módon adatainkra illesztett egyenes mennyire jól írja le a kapcsolatot? Mennyire szorosan illeszkednek a pontok az egyeneshez?

Milyen erős lineáris kapcsolat van a két változónk között?

A kapcsolat erősségét a **korrelációs együttható** méri. Ezt $R(X, Y)$ jelöli. Tulajdonságai:

- $-1 \leq R(X, Y) \leq 1$
- $R(X, Y)$ előjele egyenlő β előjelével.
- Ha a megfigyelések pontosan illeszkednek az egyeneshez, akkor
  - pozitív irányú kapcsolat esetén $R(X, Y) = 1$,
  - negatív irányú kapcsolat esetén $R(X, Y) = -1$.

- Ha $R(X, Y) = 0$, akkor a két változó korrelálatlan.
- $R(X, Y)$ lehet 0 akkor is, ha létezik kapcsolat X és Y között, de az nem lineáris.

A korrelációs együttható becslése a mintából:

Megfigyeléseink: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) (n darab). Ekkor

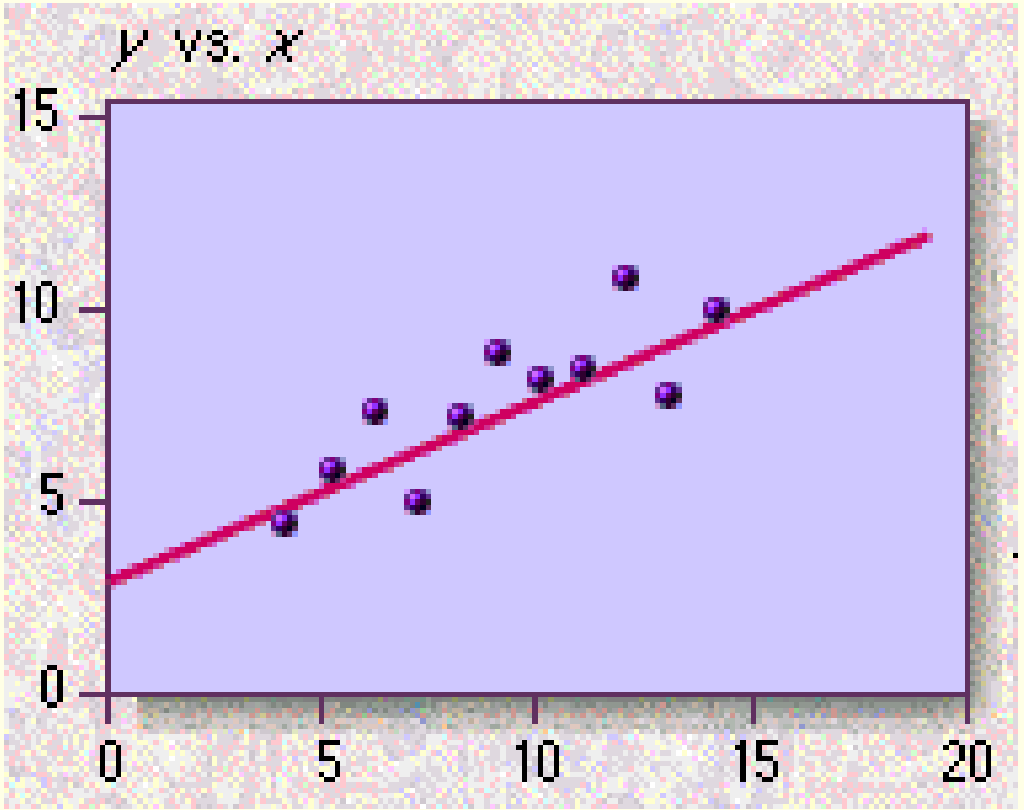
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y},$$

ahol $\bar{x}$, $\bar{y}$ a változók mintaátlaga, s_x és s_y a becsült szórások. Belátható, hogy

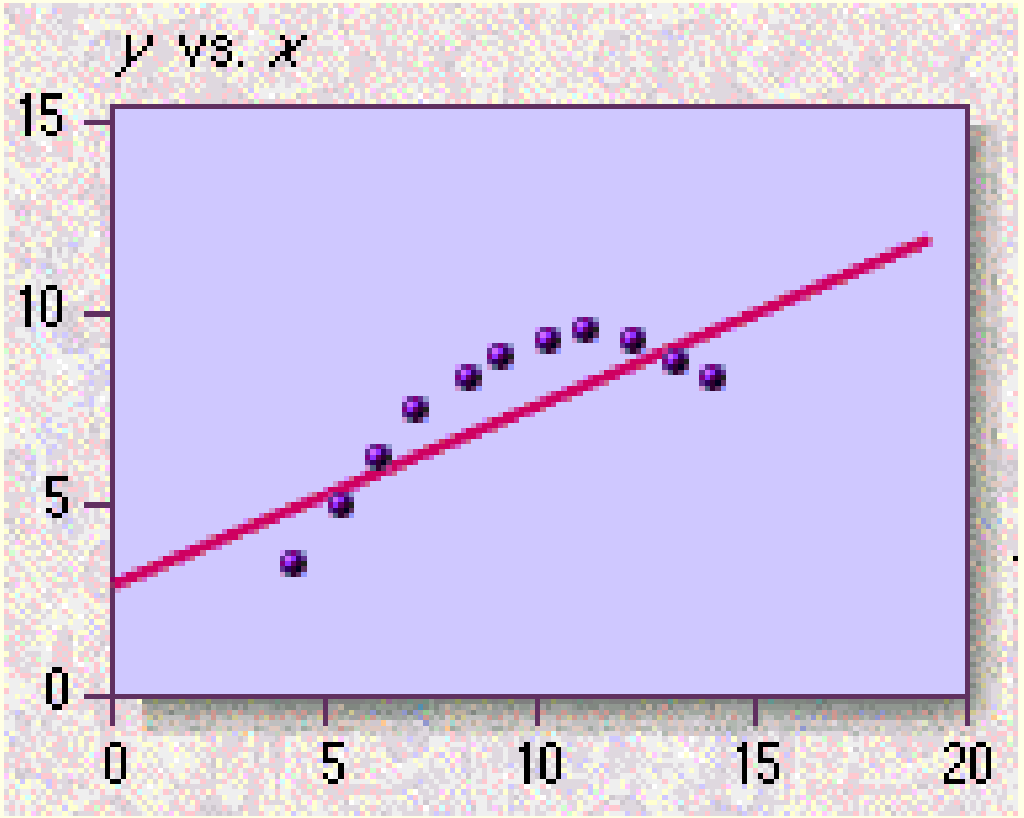
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

Francis Anscombe (Yale University): négy olyan adathalmaz, amelyek főbb jellemzői (korreláció, a, b) ugyanazok. Azonban a két változó közti kapcsolat teljesen más jellegű.

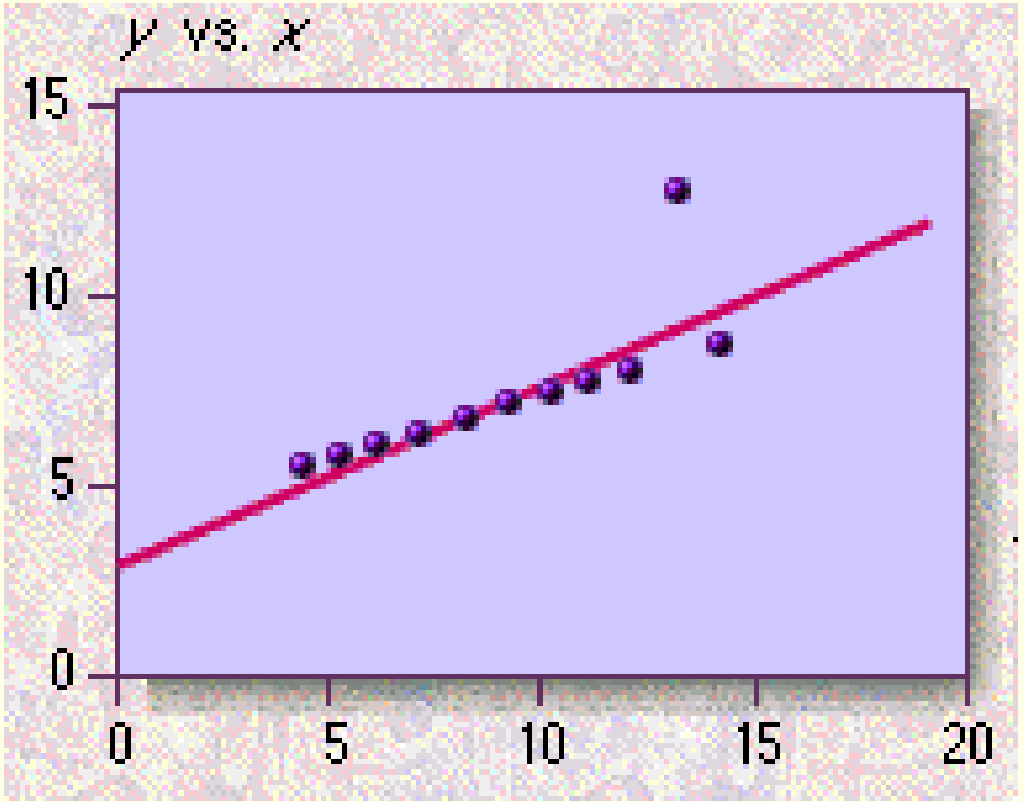
1. Jó illeszkedés:



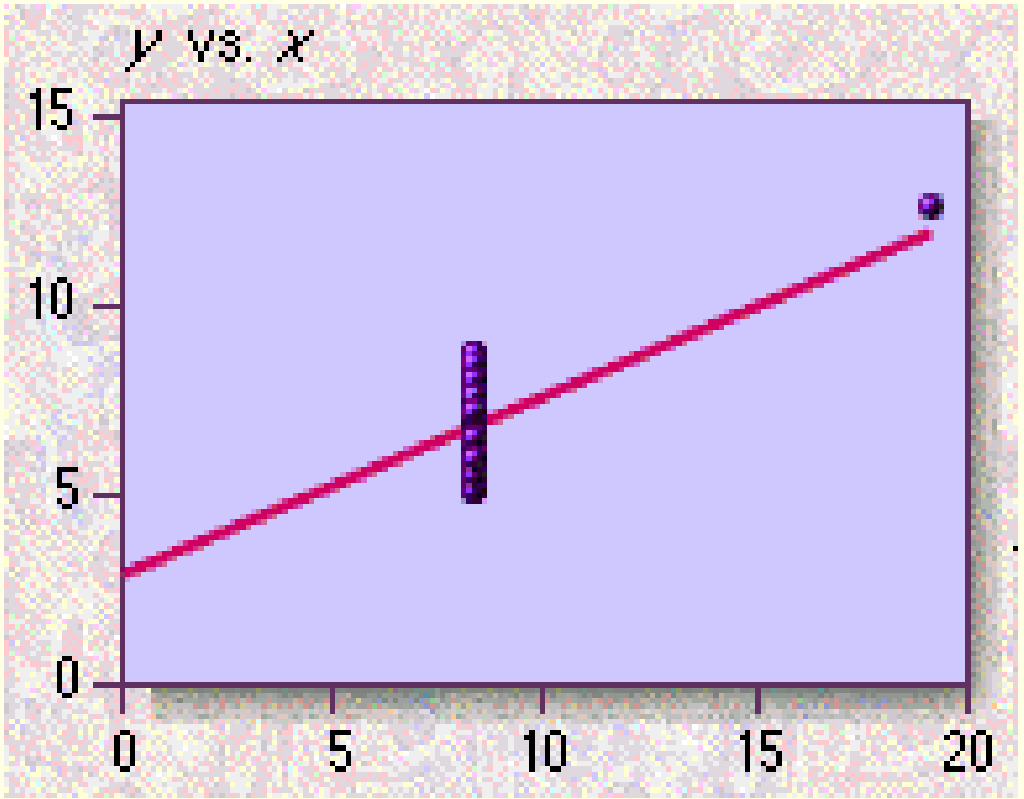
2. Görbe:



3. Kiütő érték:



4. Izolált pont:



Hipotézisvizsgálat a korrelációs együtthatóra:

$$H_0: \qquad R(X,Y) = 0,$$
$$H_1: \qquad R(X,Y) \neq 0.$$

A vizsgálat menete hasonló az eddigiekéhez.

Például: egy 10 elemű mintában két dolgot mértünk. Legyen $r_{xy} = 0.35$. A hipotézisvizsgálat során annak valószínűségét számítjuk ki, hogy ekkora értéket kapjunk, ha a két változó korrelálatlan. Ha ez a valószínűség kicsi (pl. $<$ 95%), akkor elutasítjuk a

H_0-t; ellenkező esetben elfogadjuk.

Nézzük meg, hogy pontosan mit számolunk ki. Itt is eltérés négyzetösszegeket vizsgálunk, mint egy általános varianciaanalízis esetén. Az Y változó teljes eltérés négyzetösszege:

$$SSQ_t = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Ennek szabadsági foka $n - 1$, hiszen $\bar{y}$ becsült paraméter.

A reziduumok négyzetösszege:

$$SSQ_r = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.$$

Ennek a szabadsági foka $n - 2$, hiszen a is és b is becsült paraméter.

Belátható, hogy minden esetben

$$SSQ_t \leq SSQ_r.$$

Az $SSQ_r - SSQ_t$ különbséget regressziós négyzetösszegnek, vagy magyarázott négyzetösszegnek, vagy a modellhez tartozó négyzetösszegnek is szokták ne-

vezni.

Két mennyiséget szokás figyelni:

1. $$\boxed{F = \frac{SSQ_r - SSQ_t}{\frac{SSQ_r}{n-2}}}$$

Ez nagy, ha X és Y között szoros a kapcsolat; egyébként kicsi.

2. $$\boxed{R^2 = \frac{SSQ_r - SSQ_t}{SSQ_r}}$$

Ez 0 és 1 közti értékeket vehet fel, és F-hez hasonlóan viselkedik.

Ha X-ről és Y-ról tudjuk, hogy normális eloszlásúak, akkor bebizonyítható, hogy ha semmi kapcsolat nincs közöttük, akkor a fenti F-statisztika F-eloszlású $(1, n-2)$ szabadsági fokokkal, így hipotézisvizsgálatot végezhetünk az alábbi hipotézisekre:

H_0 : X és Y független
(a regresszió a "semmit" magyarázza),

H_1 : X és Y között van kapcsolat.

Az R^2-ről pedig belátható, hogy a korrelációs együttható négyzete. Ezt szokták **determinációs**

együtthatónak is nevezni, és azt fejezi ki, hogy az X változó milyen mértékben határozza meg az Y változó értékét.

Példánk esetén: $F = 162.22$ és $R^2 = 0.9546$.

6. Nemparaméteres próbák

A hipotézisvizsgálatok során eddig mindig egy populáció valamely paraméteréről (átlag, szórás) kellett döntenünk.

Most magáról a populáció eloszlásáról döntünk. Az ilyen típusú – paramétertől független – vizsgálatokat

nem-paraméteres próbáknak nevezzük.

6.1. Illeszkedésvizsgálat

Nagyon gyakran felmerül az a kérdés, hogy hogyan illeszkedik az a változó, amit vizsgálunk, a modellünkhöz. Ez az illeszkedésvizsgálat.

Példa. Azt az elképzelésünket (modellünket) szeretnénk tesztelni, hogy egy dobókocka szabályos (azaz minden szám egyforma (1/6) valószínűséggel fordulhat elő).

Formálisan felírva:

$$H_0: \quad \text{a kocka szabályos,}$$
$$H_1: \quad \text{a kocka nem szabályos.}$$

Tegyük fel, hogy 60-szor dobtunk a kockával, és a következő eredményeket kaptuk:

3	4	5	4	3	1	6	6	3	4
6	4	4	5	3	2	1	1	2	3
6	3	4	5	1	3	2	2	5	1
5	4	3	6	3	3	4	3	4	4
2	1	3	3	4	3	4	1	5	5
1	5	4	4	5	3	4	4	3	2

Mivel feltevésünk az, hogy a kocka szabályos, az egyes értékek dobásának várt gyakorisága 10-10. Ahhoz, hogy a tényleges adatainkat össze tudjuk hasonlítani a várt értékekkel, össze kell számolnunk, hogy a valóságban melyik érték hányszor fordult elő

(megfigyelt gyakoriság).

Érték	Megfigyelt gyakoriság	Várt gyakoriság
1	8	10
2	6	10
3	16	10
4	17	10
5	9	10
6	4	10

Már a táblázatból látszik, hogy túl sok a 3-as és a 4-es, és túl kevés a 6-os.

Nyilvánvaló, hogy ha a változónk csupán 2-3 értéket

vehet fel, akkor ha abból 1 kiugróan nagy gyakoriséggal fordul elő, akkor az nagyon gyanús.

Ha azonban elég sok lehetséges értékünk van, akkor ennek már nincs olyan nagy jelentősége.

Valahogyan az összes eltérést figyelembe kell vennünk ahhoz, hogy biztosan megállapíthassuk a vizsgált változóról, hogy az megfelel-e az elméletileg várt eloszlásnak. Ehhez tekintsük a következő statisztikát:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i},$$

ahol

- k: az osztályok száma (ennyi különböző értéke lehet a változónknak),
- f_i: az i-edik érték megfigyelt gyakorisága,
- e_i: az i-edik érték (igaz nullhipotézis esetén) várt gyakorisága.

A képletből kitűnik, hogy valójában az elméleti értéktől való eltéréseket összegezzük, méghozzá a négyzetre emelésekkel egyformán figyelembe véve a pozitív és negatív eltéréseket.

Behelyettesítve a képletbe:

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(6-10)^2}{10} + \ldots + \frac{(4-10)^2}{10} = 14.2.$$

Ha a megfigyelt gyakoriságok messze vannak a várttól, akkor ez az összeg nagy lesz, ha azonban közel vannak, akkor kicsi. Így χ^2 megad egy mértéket a megfigyelt és a várt gyakoriságok távolságának mérésére.

Természetesen a dobások véletlenszerűsége miatt még szabályos kocka esetén sem fogunk pontosan $\chi^2 = 0$-t kapni. A χ^2 értékre meg kell engednünk egy bizonyos intervallumot, amelybe ha beleesik, akkor

még elég nagy a valószínűsége annak, hogy a kocka szabályos. Úgy is lehet fogalmazni: mekkora a valószínűsége annak, hogy szabályos kocka esetén ilyen eredményt kapjunk.

A hipotézisvizsgálatoknál megismert módszerrel ki kell számítanunk, hogy mekkora a valószínűsége annak, hogy szabályos kocka esetén $\chi^2 \geq 14.2$.

Ez a valószínűség 1.4%.

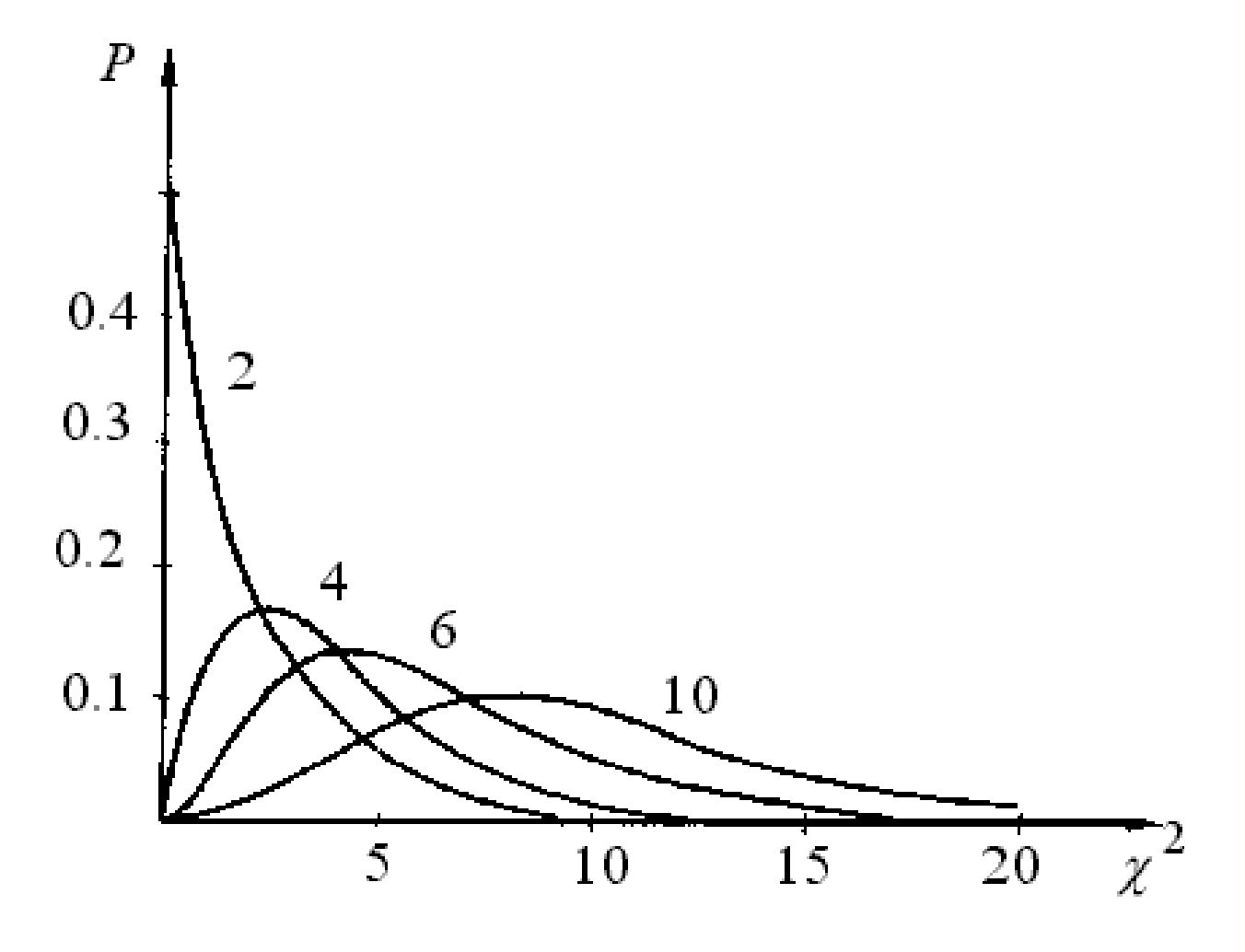
Ezt hívjuk **megfigyelt szignifikanciaszintnek**.

Honnan találjuk meg egyes χ^2 értékekhez tartozó valószínűségeket?

Ehhez be kell vezetni a χ^2-eloszlásfüggvényt, melyet a következő ábrán mutatunk be.

A t-eloszláshoz hasonlóan itt is egy görbeseregről van szó. Azt, hogy a görbeseregből melyiket kell kiválasztanunk, a szabadsági fok mondja meg. A szabadsági fok illeszkedésvizsgálat esetén egyenlő az osztályok száma minusz 1-gyel, esetünkben $k - 1 = 6 - 1 = 5$.

Ezek után elő kell venni a χ^2-eloszlás táblázatát, és meg kell benne nézni a számolt χ^2 értékhez tartozó valószínűséget. (A számítógépes programcsomagok közlik a megfelelő valószínűséget.)



A táblázatokban általában csak bizonyos valószínű-

ségekhez tartozó χ^2 értékeket adnak meg. Ha 5%-os szignifikanciaszinttel dolgozunk (azaz esetünkben azt mondjuk, hogy akkor utasítjuk el a kocka szabályosságára vonatkozó hipotézisünket, ha szabályos kockát feltételezve kisebb, mint 5% a valószínűsége annak, hogy olyan eredményt kapjunk, amilyet kaptunk), akkor ki kell keresnünk a táblázatból az 5%-hoz és a megfelelő szabadsági fokhoz tartozó értéket, és össze kell hasonlítani a számolttal. Ha a **számolt nagyobb a táblázatbeli értéknél**, akkor **elutasítjuk** hipotézisünket.

6.2. Normalitásvizsgálat

Ezt a tesztet használhatjuk a normalitás ellenőrzésére is, azaz arra, hogy a vizsgált eloszlás normális-e. Ilyen esetekben meg kell mondanunk, hogy melyik normális görbéhez illesztjük az adatainkat. Tudjuk, hogy a normális eloszlást két paraméter határozza meg, a várható érték és a szórás. Ezeket a paramétereket a mintából becsülhetjük.

Ha becsüljük az eloszlás paramétereit, akkor megváltozik a megfelelő χ^2 -eloszlás szabadsági foka, méghozzá úgy, hogy annyit le kell vonnunk még az

osztályszám -1-ből, ahány paramétert becsültünk.

Példa.

Egy nagy áruházban felmérést készítettek arról, hogy a vásárlók mekkora összeget költöttek. 500-as mintát vizsgáltak. A megfigyelések átlaga 10, szórása 2 volt. Az adatokat a következő táblázat tartalmazza. Ezek alapján levonhatjuk-e azt a következtetést, hogy a költött összeg normális eloszlású? (Legyen $\alpha = 5\%$.)

AMOUNT SPENT	f_0	AREA	f_e	$(f_0 - f_e)^2/f_e$
Less than $6	20	0.02	11.40	6.49
$6 - $8	60	0.14	67.95	0.93
$8 - $10	140	0.34	170.65	5.50
$10 - $12	120	0.34	170.65	15.03
$12 - $14	90	0.14	67.95	7.16
$14 or more	70	0.02	11.40	301.22
Total	500		500	336.33

Most 6 osztályunk van. Az első osztály várt gyakoriságát számítjuk ki most (a normalitás esetén):

$$P(X < 6) = P\left(\frac{X-10}{2} < \frac{6-10}{2}\right) = 0.0228,$$

tehát

$$e_1 = 0.0228 \cdot 500 = 11.4.$$

A többi várt gyakoriság is hasonlóan számítható ki.

A szabadsági fok most 5, és a kritikus érték 11.07.

A χ^2 értéke pedig 363.33. Mivel ez jóval nagyobb a kritikus értéknél, ezért a normalitást el kell vetnünk.

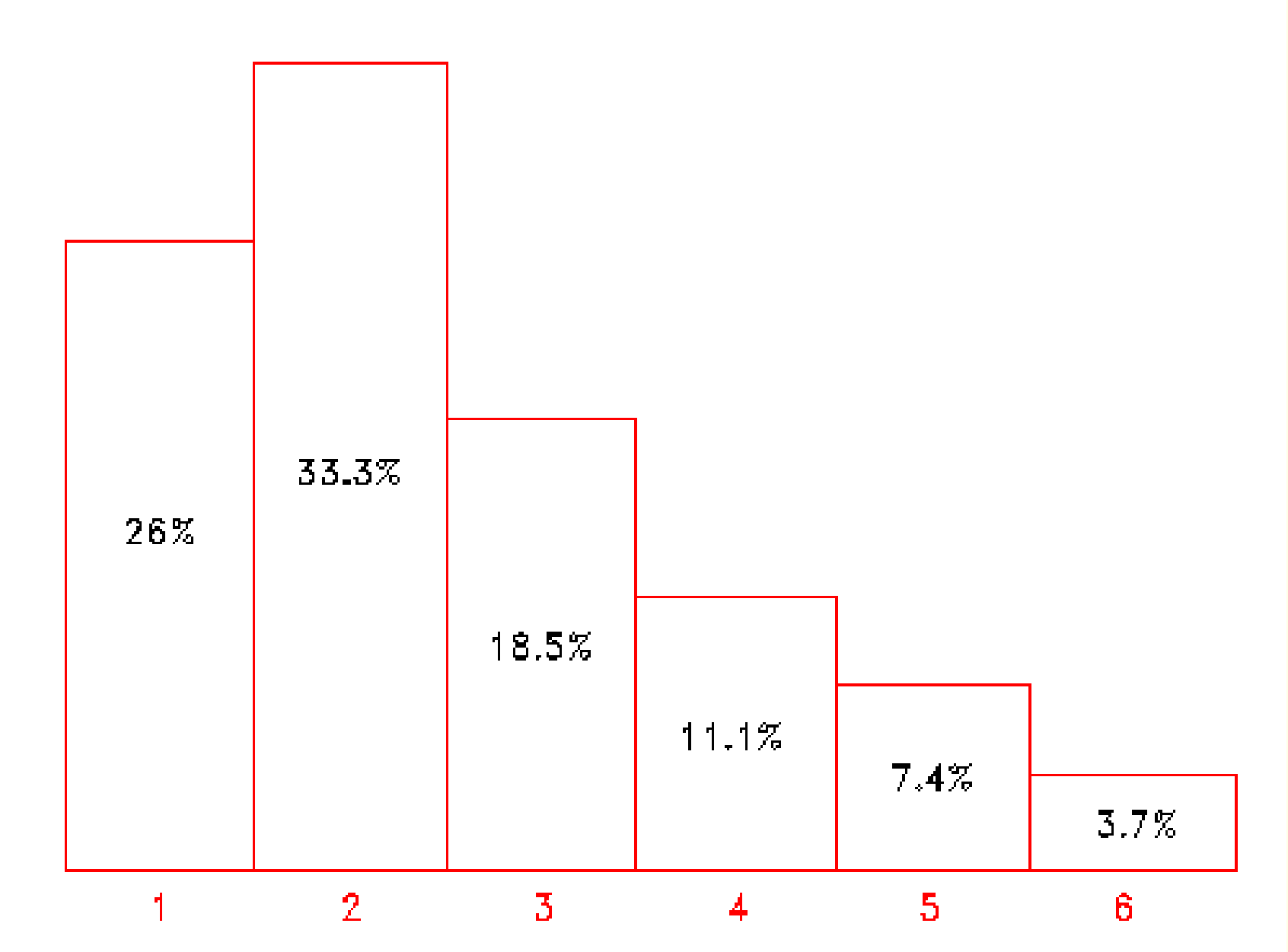
6.3. Homogenitásvizsgálat

Most a mintánkat nem egy elméleti eloszláshoz akarjuk hasonlítani, hanem egy másik mintához.

(Ilyen eset áll elő akkor is, ha nem tudjuk a két minta átlagát összehasonlítani t-próbával, mivel nem teljesül a normalitás feltétele.)

Tipikus példa a tehenek esetén a megtermékenyüléshez szükséges inszeminálások számának eloszlása. Ez tipikusan nem normális, hanem ferde eloszlású.

Osztály		Gyakoriság	
(inszeminálások száma)	egyed-szám	relatív	%
1	7	0.26	26
2	9	0.333	33.3
3	5	0.185	18.5
4	3	0.111	11.1
5	2	0.074	7.4
6	1	0.037	3.7


26%
33.3%
18.5%
11.1%
7.4%
3.7%
1
2
3
4
5
6

Az eloszlás nem normális volta miatt csak azt tudjuk megvizsgálni, hogy a két minta eloszlása azonosnak tekinthető-e. Formálisan:

H_0 : a két minta azonos eloszlású,

H_1 : a két minta nem azonos eloszlású.

A vizsgálat technikája hasonló, mint az illeszkedés-vizsgálatnál. Itt is az egyes osztályok gyakoriságát hasonlítjuk össze (a két mintából készített hisztogramot). Azt nézzük meg, hogy az eltérésük nagyobb-

e annál, ami a mintavételezési hiba miatt azonos eloszlásoknál is elfogadható.

Az eltérést a

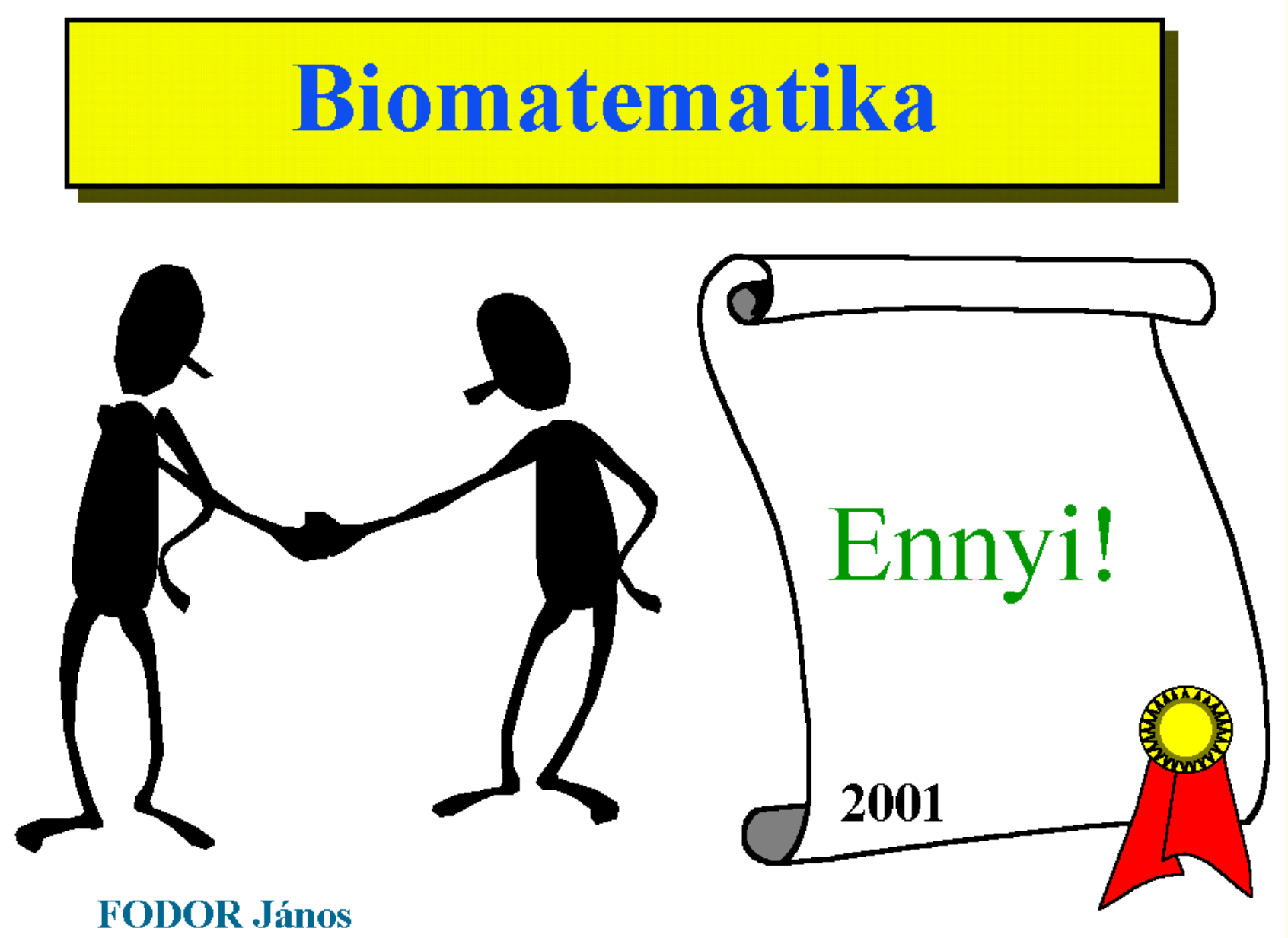
$$\chi^2 = n_1 \cdot n_2 \cdot \sum_{i=1}^{k} \frac{1}{f_{1i} + f_{2i}} \left(\frac{f_{1i}}{n_1} - \frac{f_{2i}}{n_2} \right)^2$$

összeggel mérjük, ahol

$$n_1 = \sum_{i=1}^{k} f_{1i}, \quad n_2 = \sum_{i=1}^{k} f_{2i}.$$

Az f_{1i}, f_{2i} az első ill. második minta i-edik osztályának gyakorisága, k az osztályok száma.

Most is ki kell keresnünk a χ^2-eloszlás táblázatából a megfelelő valószínűséghez (általában 5%) és szabadsági fokhoz tartozó értéket (esetünkben a szabadsági fok szintén $k-1$).


Biomatematika
Ennyi!
2001
FODOR János