Statisztika

Előadások letölthetők a
http://www.cs.elte.hu/~arato/InfBC_stat*.pdf
címről

Lehetséges hibák

- Elsőfajú hiba: H_0 igaz, de elutasítjuk
- Másodfajú hiba: H_0 hamis, de elfogadjuk

		Aktuális helyzet	
		A nullhipotézis igaz	A nullhipotézis hamis
Döntés:	Elfogadjuk a nullhipotézist	Helyes döntés	Másodfajú hiba
	Elutasítjuk a nullhipotézist	Elsőfajú hiba	Helyes döntés

Alapfogalmak

- Emlékeztető: $\mathcal{X}$ mintatér: a minta lehetséges értékeinek halmaza.
- $\mathcal{X} = \mathbf{X}_e \cup \mathbf{X}_k$
- $\mathbf{X}_k$: azon lehetséges értékek halmaza, amelyek megfigyelése esetén elutasítjuk a nullhipotézist.
- Gyakran statisztika segítségével határozzuk meg:

$$T(\mathbf{x}) = \begin{cases} 1 & , \ \mathbf{x} \in \mathbf{X}_k \\ 0 & , \ \mathbf{x} \notin \mathbf{X}_k \end{cases}$$

Véletlenített próba

- Eddig adott megfigyelés esetén egyértelmű volt a döntésünk:

$$T(\mathbf{x}) = \begin{cases} 1 & , \ \mathbf{x} \in \mathbf{X}_k \\ 0 & , \ \mathbf{x} \notin \mathbf{X}_k \end{cases}$$

Véletlenített próba esetén sorsolhatunk is:

$$\Psi(\mathbf{x}) = \begin{cases} 1 & , \ \text{ha } T(\mathbf{x}) > c \\ \gamma & , \ \text{ha } T(\mathbf{x}) = c \\ 0 & , \ \text{ha } T(\mathbf{x}) < c \end{cases}$$

Elsőfajú hiba valószínűsége véletlenített próba esetén

$\vartheta \in \Theta_0$-ra az elsőfajú hiba valószínűsége:

$$P_\vartheta\left(T(\boldsymbol{\xi}) > c\right) + \gamma P_\vartheta\left(T(\boldsymbol{\xi}) = c\right) = E_\vartheta\left(\psi(\boldsymbol{\xi})\right)$$

α a próba terjedelme, ha minden $\vartheta \in \Theta_0$-ra

$$E_\vartheta\left(\psi(\boldsymbol{\xi})\right) \le \alpha$$

α a próba szignifikanciaszintje

(másképp: a próba pontos terjedelme),

$$\sup_{\vartheta \in \Theta_0} E_\vartheta\left(\psi(\boldsymbol{\xi})\right) = \alpha$$

Legerősebb próba egyszerű hipotézis esetében

Egyszerű H_0 és H_1 : $|\boldsymbol{\Theta}_0| = |\boldsymbol{\Theta}_1| = 1$.

ψ a legerősebb α-terjedelmű próba, ha:

$P_{\vartheta_0}\left(T(\boldsymbol{\xi}) > c\right) + \gamma P_{\vartheta_0}\left(T(\boldsymbol{\xi}) = c\right) = E_{\vartheta_0}\left(\psi(\boldsymbol{\xi})\right) \leq \alpha,$

továbbá minden más α-terjedelmű ψ' próbára, annak

másodfajú hibavalószínűsége nagyobb:

$E_{\vartheta_1}\left(1 - \psi(\boldsymbol{\xi})\right) \leq E_{\vartheta_1}\left(1 - \psi'(\boldsymbol{\xi})\right).$

A legerősebb próba

- A legegyszerűbb eset: H_0 és H_1 is egyszerű (egyelemű). A valószínűséghányados (vh.) próba:

$$T(\mathbf{x}) = \begin{cases} 1 & \dfrac{L_1(\mathbf{x})}{L_0(\mathbf{x})} > c \\ \gamma & \dfrac{L_1(\mathbf{x})}{L_0(\mathbf{x})} = c \\ 0 & \dfrac{L_1(\mathbf{x})}{L_0(\mathbf{x})} < c \end{cases}$$

- Állítás (Neyman-Pearson lemma): a vh. próba legerősebb a saját terjedelmével. Minden $0<\alpha<1$-hez létezik ilyen terjedelmű vh. próba. Minden legerősebb próba ilyen alakú.

Próbák a normális eloszlás várható értékére: t próba.

- H_0: $m=m_0$, H_1 $m \neq m_0$. Ha nem ismert a szórás (t-próba):

$$t = \sqrt{n}\,\frac{\overline{X} - m_0}{\hat{\sigma}}$$

- ahol

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 /(n-1)}$$

- Kritikus tartomány: $|t| > t_{1-\alpha/2,\, n-1}$. ($H_0$ esetén a próbastatisztika n-1 szabadságfokú, t-eloszlású.)
- Ha egyoldali az ellenhipotézis, akkor a kritikus tartomány $t > t_{1-\alpha,\, n-1}$ ($m > m_0$), illetve $t < -t_{1-\alpha,\, n-1}$ alakú ($m < m_0$). Ezek is legerősebb próbák!

Megjegyzések

- A kétoldali esetre kapott próba nem a legerősebb (ilyenkor nincs is ilyen).
- Ha a minta elemszáma nagy, a t-próba helyett az u-próba is használható (ekkor még a normális eloszlásúságra sincs szükség a centrális határeloszlás tétel miatt).

Kétoldali próbák és konfidencia intervallumok

- A normális eloszlásnál a várható értékre vonatkozó α terjedelmű próbánál láttuk, hogy a H_0: $m=m_0$ hipotézist a H_1: $m \neq m_0$ hipotézissel szemben pontosan akkor fogadjuk el, ha m_0 benne van az $1-\alpha$ megbízhatóságú konfidencia intervallumban.

Kétmintás eset: párosított megfigyelések

- Példa: Van-e különbség Budapest és Cegléd napi átlaghőmérséklete között? H_0: $m_1=m_2$ a nullhipotézis.
- Ha ugyanazon napokról van megfigyelésünk mindkét helyen: nem függetlenek a minták. Ekkor a párok tagjai közötti különbséget vizsgálva, az előző egymintás esetre vezethető vissza a feladat. H_0^*: $m=0$, H_1^*:$m\neq0$ az új hipotézisek.

Kétmintás eset: független minták

Ha ismert a szórás: ($\underline{X}$ n elemű, σ_1 szórású, $\underline{Y}$ m elemű, σ_2 szórású), alkalmazható a kétmintás u-próba

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}}$$

Kritikus tartomány: mint az egymintás esetben

Ha ismeretlenek, de azonosak a szórások:

$$t_{n+m-2} = \sqrt{\frac{nm(n+m-2)}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2}}$$

A szórás vizsgálata kétmintás esetben: F-próba

- H_0: $\sigma_1=\sigma_2$
- Két független, *n*, illetve *m* elemű normális eloszlású minta alapján a próbastatisztika:

  (a korrigált tapasztalati szórásnégyzetek hányadosa)

  $$F=\max\left(\frac{s_1^2}{s_2^2},\frac{s_2^2}{s_1^2}\right)$$

- Kritikus érték: az *n*-1,*m*-1 szabadságfokú F eloszlás 1-α/2 kvantilise (*n* a számlálóbeli, *m* pedig a nevezőbeli minta elemszáma).

Kétmintás t-próba ismét

- Alkalmazható, ha az F-próba elfogadja a szórások azonosságát.
- Ha nem, akkor Welch-próba:

$$t' = \frac{\overline{X} - \overline{Y}}{\sqrt{s_1^{\,2} / n_1 + s_2^{\,2} / n_2}}$$

- H_0 esetén közelítőleg t eloszlású f szabadságfokkal, ahol

$$\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1} \qquad c = \frac{s_1^{\,2} / n}{s_1^{\,2} / n + s_2^{\,2} / m}$$

χ-négyzet próba

- H_0 hipotézis: az A_1, A_2 , ..., A_r teljes eseményrendszerre teljesül $P(A_1)=p_1$, $P(A_2)=p_2$, ..., $P(A_r)=p_r$
- A tesztstatisztika: $$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$

  ami aszimptotikusan r-1 szabadságfokú χ-négyzet eloszlású, ha igaz a nullhipotézis.
- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az r-1 szabadságfokú χ-négyzet eloszlás 1- α kvantilise, elutasítjuk a nullhipotézist.

χ-négyzet próba (folytatás)

- Miért is ez a határeloszlás?

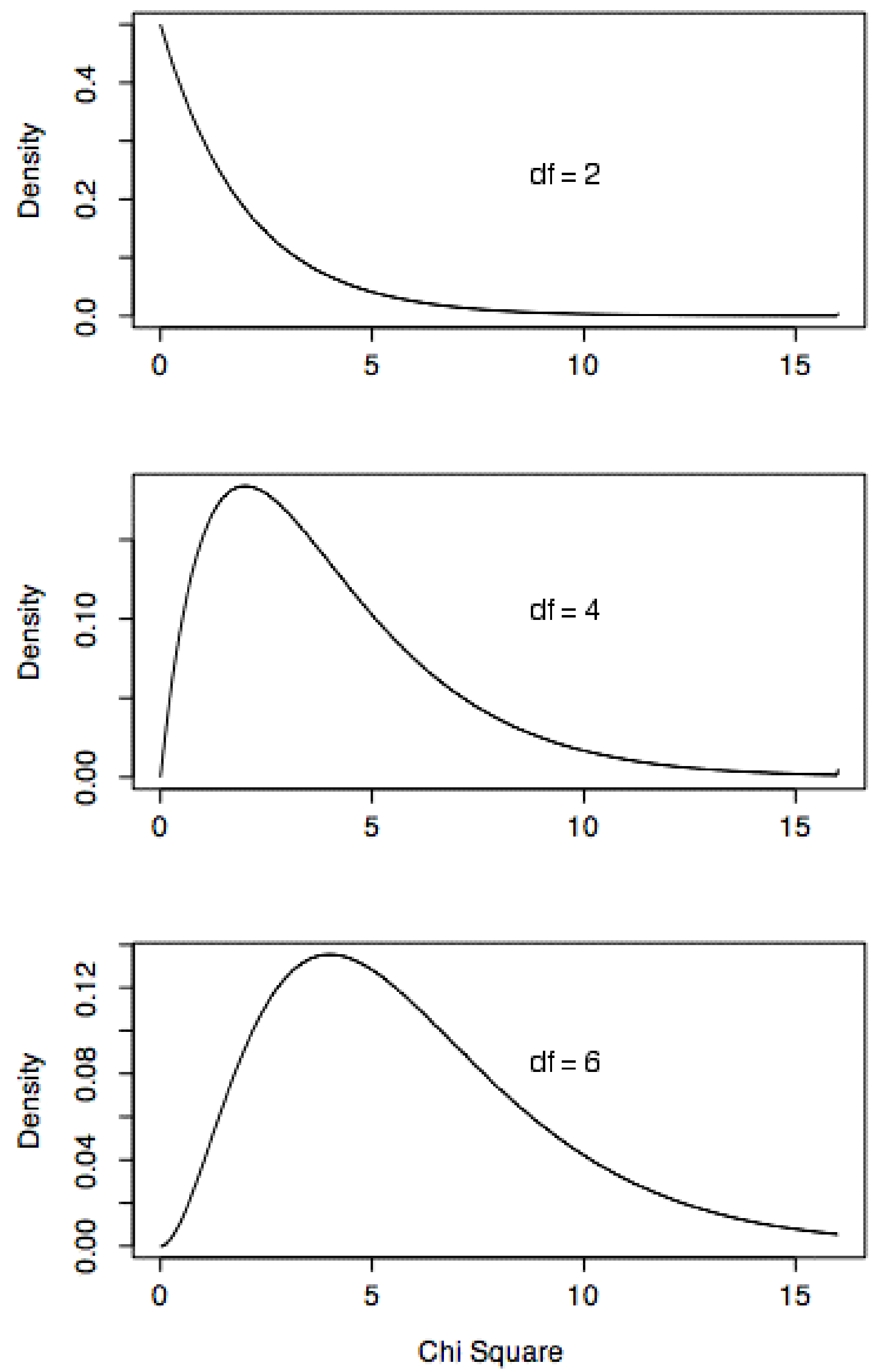
$r=2,\ H_0: P(A)=p,\ \nu: A$ gyakorisága n kísérletből

$$\chi^2=\frac{(\nu-np)^2}{np}+\frac{((n-\nu)-n(1-p))^2}{n(1-p)}=\frac{(\nu-np)^2}{np}+\frac{(\nu-np)^2}{n(1-p)}=\frac{(\nu-np)^2}{np(1-p)}$$

$\xi_i=1$, ha az i. kísérletnél A bekövetkezik, 0 különben

$$\nu=\sum_{i=1}^{n}\xi_i,\ E\xi_i=p,\ D^2\xi_i=p(1-p),$$

$$\chi^2=\left(\frac{\sum_{i=1}^{n}\xi_i-nE\xi_1}{\sqrt{n}D\xi_1}\right)^2\xrightarrow[n\to\infty,\,\text{eloszlásban}]{}\chi_1^2$$


Density
0.0
0.2
0.4
df = 2
0
5
10
15
Density
0.00
0.10
df = 4
0
5
10
15
Density
0.00
0.04
0.08
0.12
df = 6
0
5
10
15
Chi Square

Példa (kockadobás)

- 36 kockadobás eredménye

Szám	Megfigyelt	np_i	$\frac{(v_i - np_i)^2}{np_i}$
1	8	6	0.667
2	5	6	0.167
3	9	6	1.500
4	2	6	2.667
5	7	6	0.167
6	5	6	0.167

$n = 36,\ r = 6$

$$\sum_{i=1}^{6} \frac{\left(\nu_i - np_i\right)^2}{np_i} \sim \chi_5^2$$

$$\sum_{i=1}^{6} \frac{\left(\nu_i - np_i\right)^2}{np_i} = 5.333$$

$$P\left(\chi_5^2 > 5.333\right) = 0.377 \Rightarrow$$

Nem tudjuk a szabályosság hipotézisét elutasítani!

Példa (számítógépek népszerűsége)

- 100 amerikai diák

Számí-tógép	Megfigyelt	np_i	$\frac{(\nu_i - np_i)^2}{np_i}$
IBM	47	33.333	5.604
Macintosh	36	33.333	0.213
Egyéb	17	33.333	8.003

$$n = 100,\ r = 3$$

$$\sum_{i=1}^{3} \frac{\left(\nu_i - np_i\right)^2}{np_i} \sim \chi_2^2$$

$$\sum_{i=1}^{3} \frac{\left(\nu_i - np_i\right)^2}{np_i} = 13.820$$

$$P\left(\chi_2^2 > 5.99\right) = 0.05 \Rightarrow$$

Elutasítjuk az egyforma kedveltség hipotézisét!

χ-négyzet próba illeszkedésvizsgálatra

- Illeszkedésvizsgálat:

$H_0 : \xi_1, \ldots, \xi_n$ F eloszlásfüggvényűek

- Visszavezetjük az előző esetre

$$A_i = \{\xi \in C_i\}, i = 1, 2, \ldots, r, \bigcup_i C_i = \mathbf{R}$$

Diszkrét esetben gyakran: $A_i = \{\xi = x_i\}, i = 1, 2, \ldots, r$

Példa

- Mi lehet egy vezető által okozott károk számának eloszlása?
- Poisson eloszlású-e?

Kár-szám	0	1	2	3	4	5	6	7	>7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006

Becsléses χ-négyzet próba

- H_0 hipotézis: az A_1, A_2 , …, A_r teljes eseményrendszerre teljesül:

$$P(A_i) = p_i(\vartheta_1, \ldots, \vartheta_s), i = 1, 2, \ldots, r$$

$\vartheta_1, \ldots, \vartheta_s$ ismeretlen paraméterek.

A tesztstatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} \xrightarrow[n\to\infty]{} \chi^2_{r-s-1},$$

ahol

$$\hat{p}_i = p_i(\hat{\vartheta}_1, \ldots, \hat{\vartheta}_s).$$

Példa (folyt.)

Kár-szám	0	1	2	3	4	5	6	7	>7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006
np_i *Poisson*	128 433	18 218	1 292	61	2,2	0,06	0,001	3E-05	5E-07	
Np_i *Neg. bin.*	129 541	16 237	1 962	234	28	3,3	0,39	0,05	0,006	

$$n = 148006,\ r = 5$$

$$A_i = \{\xi = i\}, i = 0, 1, 2, 3$$

$$A_4 = \{\xi \geq 4\}$$

Poisson eset:

$$\hat{\lambda} = 0.709$$

$$\sum_{i=0}^{4} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} \sim \chi^2_{5-1-1}$$

$$\sum_{i=0}^{4} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} > 200$$

$$P\left(\chi^2_3 > 17.7\right) = 0.05\% \Rightarrow$$

Elutasítjuk Poisson eloszlás hipotézisét!

Az illeszkedésvizsgálat alkalmazása folytonos eloszlásokra

- A teljes eseményrendszer a számegyenes felosztása révén jön létre.
- Ügyeljünk arra, hogy minden intervallum közel azonos valószínűségű legyen.
- Ha paraméterbecslés szükséges, ML módszer alkalmazható.

χ-négyzet próba homogenitásvizsgálatra

- Homogenitásvizsgálat:

$H_0 : \xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_m$ ugyanolyan eloszlásúak

- Hasonlóan járunk el, mint korábban

$$\bigcup_{i=1}^{r} C_i = \mathbf{R}$$

$$\nu_i = \left|\left\{ j : \xi_j \in C_i \right\}\right|, \mu_i = \left|\left\{ j : \eta_j \in C_i \right\}\right|, \ i = 1, 2, \ldots, r,$$

A tesztstatisztika:

$$\chi^2 = nm \sum_{i=1}^{r} \frac{\left(\nu_i / n - \mu_i / m \right)^2}{\nu_i + \mu_i} \xrightarrow[n,m\to\infty]{} \chi^2_{r-1}$$

Ki tanul jobban?

2009. január 5-ei vizsga

Jegy	Férfi	Nő	Összesen
1	47	4	51
2	11	1	12
3	11	2	13
4	9	2	11
5	8	2	10
Összesen	86	11	97
Átlag	2,1	2,7	2,1

$C_1 = \{1;2\}, C_1 = \{3;4;5\}$

$\nu_i = \left|\{j : \xi_j \in C_i\}\right|, \mu_i = \left|\{j : \eta_j \in C_i\}\right|, i = 1,2,$

$\nu_1 = 58, \nu_2 = 28, \mu_1 = 5, \mu_2 = 6, n = 86, m = 11$

A tesztstatisztika:

$$\chi^2 = 86 \cdot 11 \left(\frac{\left(\frac{58}{86} - \frac{5}{11}\right)^2}{58+5} + \frac{\left(\frac{28}{86} - \frac{6}{11}\right)^2}{28+6} \right) = 2.071$$

$P(\chi_1^2 > 2.71) = 10\% \Rightarrow$

Nem tudjuk elutasítani az egyforma képesség hipotézisét!

χ-négyzet próba függetlenségvizsgálatra

- H_0 hipotézis: az $A_1, A_2, \ldots, A_r$ és $B_1, B_2, \ldots, B_s$ teljes eseményrendszerekre teljesül a függetlenség.

$$\sum_{i,j} \frac{\left(\nu_{ij} - np_i q_j\right)^2}{np_i q_j}$$

- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az rs-1 szabadságfokú χ-négyzet eloszlás 1- α kvantilise, elutasítjuk a nullhipotézist.

Becsléses eset

- Általában, ha az illesztendő eloszlást nem ismerjük – csak a családját - becsüljük a paramétereit. Ekkor a próbastatisztika szabadságfoka annyival csökken, ahány paramétert becsültünk.
- Függetlenségvizsgálatnál általában nem ismerjük a teljes eseményrendszer tagjainak valószínűségét, így $r-1+s-1$ valószínűséget kell becsülnünk. A szabadságfok ekkor tehát

  $rs-1-r-s+2=(r-1)(s-1)$.

$\nu_{ij} : A_i B_j$ gyakorisága

$\nu_{i\bullet} : A_i$ gyakorisága

$\nu_{\bullet j} : B_j$ gyakorisága

A tesztstatisztika

$$n \sum_{i,j} \frac{\left(\nu_{ij} - \frac{\nu_{i\bullet} \nu_{\bullet j}}{n} \right)^2}{\nu_{i\bullet} \nu_{\bullet j}} \xrightarrow[n \to \infty]{} \chi^2_{(r-1)(s-1)}$$

$r = s = 1$ esetben

$$n \frac{\left(\nu_{11} \nu_{22} - \nu_{12} \nu_{21} \right)^2}{\nu_{1\bullet} \nu_{2\bullet} \nu_{\bullet 1} \nu_{\bullet 2}} \xrightarrow[n \to \infty]{} \chi^2_1$$

Példa

- http://onlinestatbook.com/case_studies/diet.html
- http://onlinestatbook.com/chapter14/contingency.html

Y közelítése X függvényével

- Gyakori eset, hogy nem ismerjük a számunkra érdekes mennyiség (Y) pontos értékét (pl. holnapi részvényárfolyam, vízállás, időjárás). Van viszont információnk hozzá kapcsolódó mennyiségről (X, mai értékek).
- Feladat: olyan f_0 megtalálása, amelyre $f_0(X)$ a lehető legjobb közelítése Y-nak.
- Matematikailag: f_0 a megoldása a szélsőérték-problémának (legkisebb négyzetes becslés). $\min_f E\left(Y - f(X)\right)^2$

Valószínűségszámításból tanultak

$E\left(Y-a\right)^2$ minimumhelye: EY

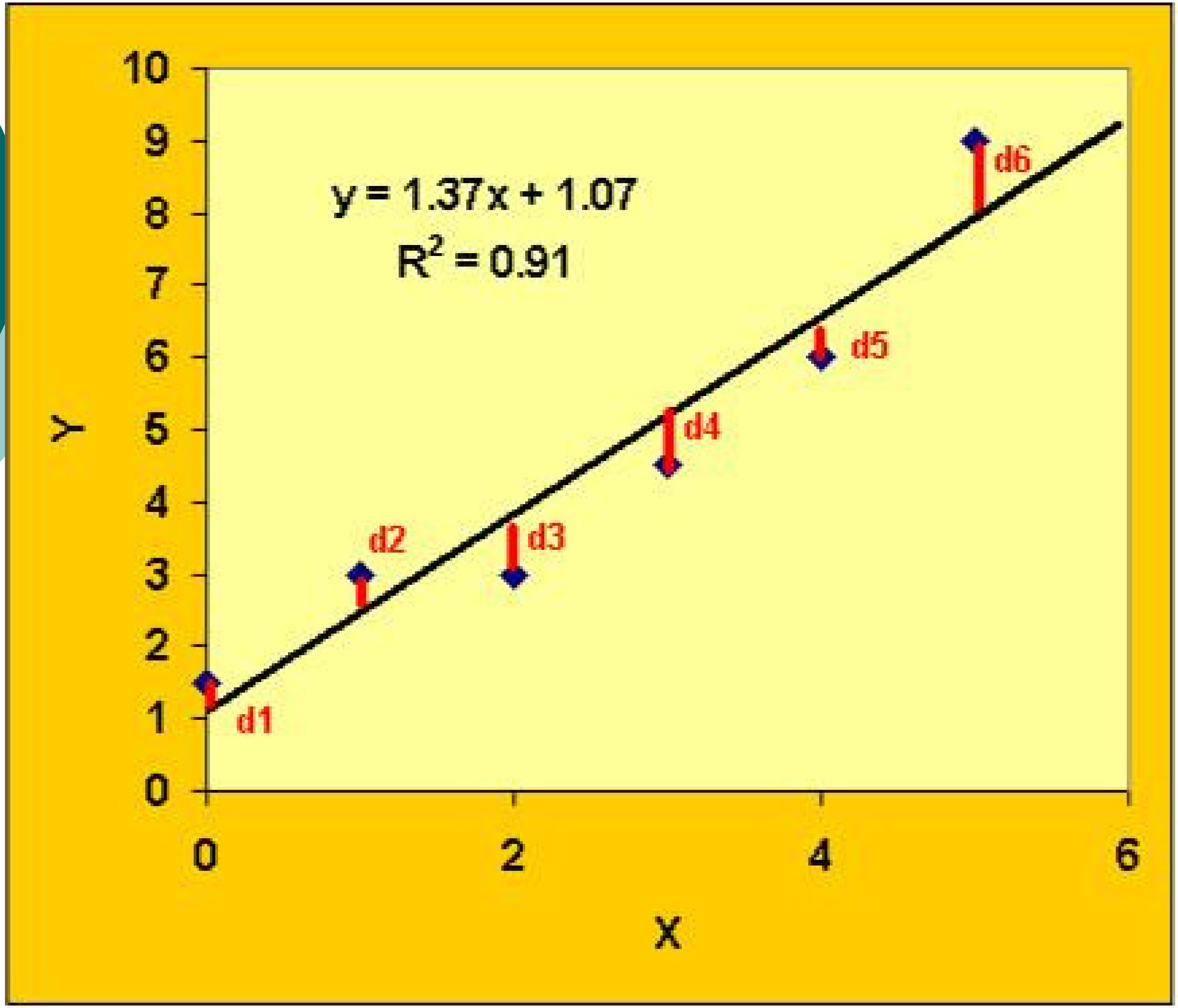
$E\left(Y-f(X)\right)^2$ minimumhelye: $f_0(x)=E\left(Y \middle| X=x\right)$

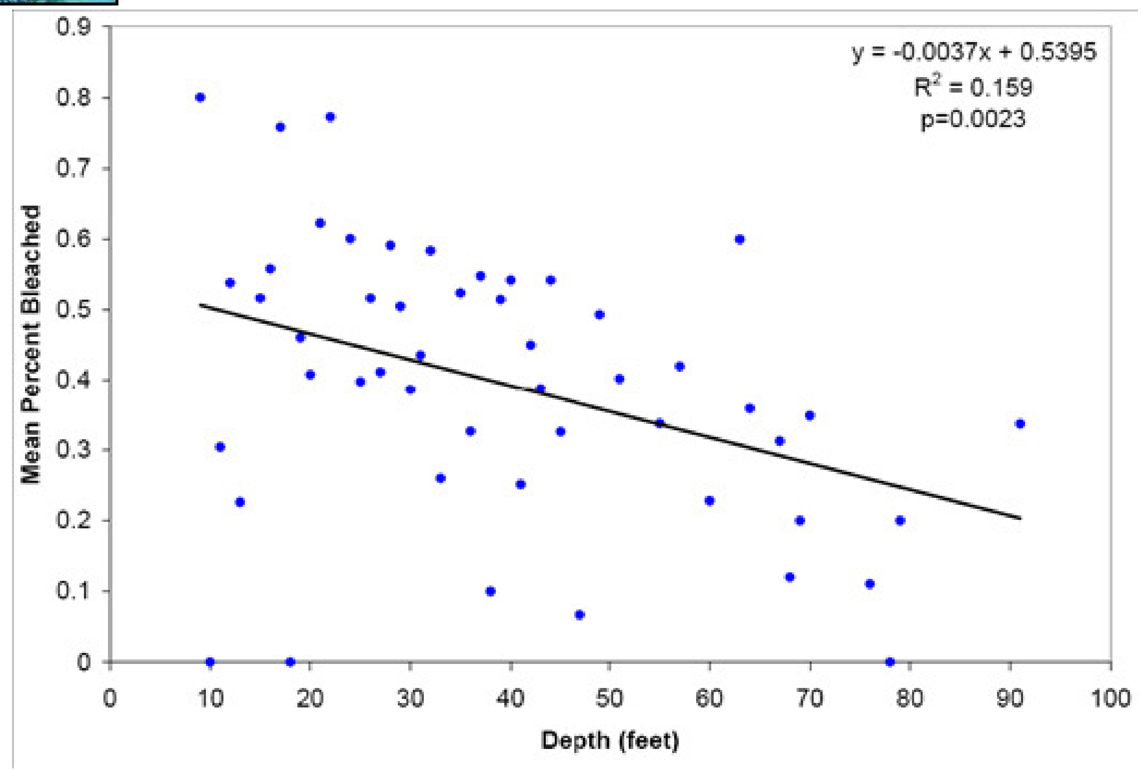
lineáris függvények esetében:

$E\left(Y-aX-b\right)^2$ minimumhelye:

$$a=\frac{\mathrm{cov}(X,Y)}{D^2X}=\frac{\mathrm{corr}(X,Y)DY}{DX}$$

$$b=EY-aEX$$


y = 1.37x + 1.07
R2 = 0.91
d1
d2
d3
d4
d5
d6
Y
X
0
1
2
3
4
5
6
7
8
9
10
0
2
4
6


y = -0.0037x + 0.5395
R^2 = 0.159
p=0.0023
Mean Percent Bleached
Depth (feet)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
10
20
30
40
50
60
70
80
90
100

Lineáris modell

- $Y_i = aX_i + b + \varepsilon_i$ (X_i a magyarázó változó értéke, ε_i független, azonos eloszlású hiba. $E(\varepsilon_i)=0$, általában feltesszük, hogy normális eloszlásúak)
- *a*,*b* a becsülendő együtthatók

Megoldás

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

Szórások

$$D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, D(\hat{b}) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Az x* pontban előrejelzett érték $\hat{a}x^* + \hat{b}$

és ennek szórása

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

A szórásbecslésnél σ helyett annak becsült értékét használjuk:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2}{n-2}$$