

Nemparaméteres próbák (10. előadás)

Illeszkedésvizsgálat: a minta egy adott, folytonos eloszlásból származik-e?

Homogenitásvizsgálat: két minta ugyanabból az eloszlásból származik-e?

Egy lehetőség: **diszkrétizáljuk** a megfigyeléseket, vagyis közel azonos hosszúságú intervallumokba osztjuk be őket, és az így kapott diszkrét eloszlásra χ^2 -próbát végzünk. Ha szükséges, a paramétereket maximumlikelihood-módszerrel becsüljük.

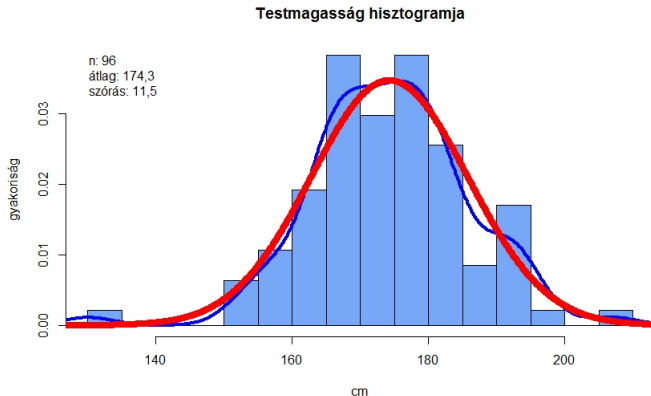
Tapasztalati eloszlásfüggvények távolságát használó próbák:

- Kolmogorov–Szmirnov-próba
- Anderson–Darling-próba (az eltéréseket másképp súlyozzuk)
- Cramér–von Mises próba (az eltéréseket másképp súlyozzuk)

Speciálisan annak ellenőrzésére, hogy egy eloszlás **normális eloszlású**-e:

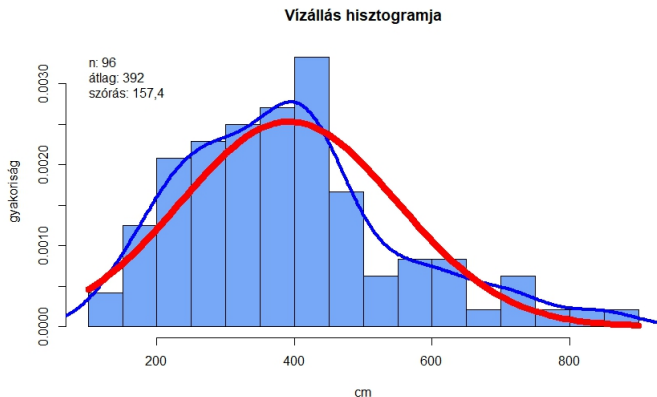
- Lilliefors-próba (a Kolmogorov–Szmirnov-próbán alapul)
- Shapiro–Wilk-próba (a rendezett minta várható értékét és kovarianciamátrixát használja)

Testmagasság és normális eloszlás



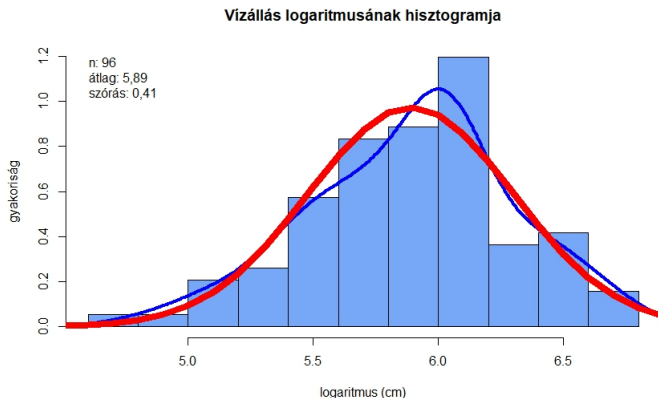
A testmagasság histogramja $n = 96$ elemű mintából, a sűrűségfüggvény becslése Gauss-magfüggvénnyel, és az $\bar{X} = 174,3$ várható értékű és $s_n^* = 11,5$ szórású normális eloszlás sűrűségfüggvénye.

A Duna vízállása



A Duna havi legnagyobb vízállásának hisztogramja (2002–2009, $n = 96$, forrás: Országos Vízelző Szolgálat), a becsült sűrűségfüggvény, és az $\bar{X} = 392$ várható értékű és $s_n^* = 157,4$ szórású normális eloszlás sűrűségfüggvénye – **itt a függetlenség nem teljesen érvényes**

A Duna vízállása



A Duna havi legnagyobb vízállásának **logaritmusának** hisztogramja (2002–2009, $n = 96$, forrás: Országos Vízeljáró Szolgálat), a becsült sűrűségfüggvény, és az $\bar{X} = 392$ várható értékű és $s_n^* = 157,4$ szórású normális eloszlás sűrűségfüggvénye

Tapasztalati eloszlásfüggvény

Az X valószínűségi változó eloszlásfüggvénye az $F : \mathbb{R} \rightarrow [0, 1]$ függvény, melyre

$$F(t) = \mathbb{P}(X \leq t)$$

minden $t \in \mathbb{R}$ -re.

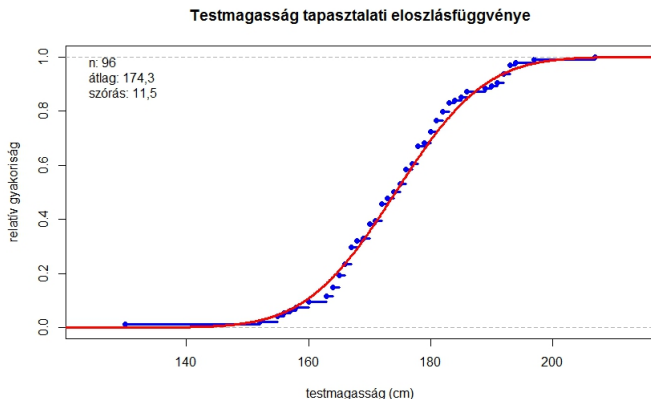
Definíció (Tapasztalati eloszlásfüggvény)

Az X_1, X_2, \dots, X_n minta tapasztalati eloszlásfüggvénye az $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$ függvény, melyre

$$\hat{F}_n(t) = \frac{t\text{-nél nem nagyobb mintaelemek száma}}{n}.$$

(empirical cumulative distribution function)

Tapasztalati eloszlásfüggvény



A testmagasság tapasztalati eloszlásfüggvénye $n = 96$ elemű mintából, és az $\bar{X} = 174,3$ várható értékű és $s_n^* = 11,5$ szórású normális eloszlás eloszlásfüggvénye.

Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat

H_0 : a minta valódi eloszlásfüggvénye F (ami folytonos)

H_1 : a minta valódi eloszlásfüggvénye F -től különböző

Próbastatisztika:

$$D_n = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)|,$$

ahol F_n a minta tapasztalati eloszlásfüggvénye.

Ha $D_n > D_{\text{krit}}$ (vagy $p < \alpha$), akkor elutasítjuk H_0 -t, a minta eloszlásfüggvénye szignifikánsan eltér D -től (itt D_{krit} a megfelelő Kolmogorov–Szmirnov-próba kritikus értéke).

Ha $D_n < D_{\text{krit}}$, (vagy $p > \alpha$) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés F -től.

A normalitás tesztelése: Lilliefors-próba

H_0 : a testmagasság normális eloszlású $m = 174,3$ várható értékkel és $\sigma = 11,5$ szórással

H_1 : a testmagasság eloszlása ettől különböző

Szignifikanciaszint: $\alpha = 0,05$.

A normalitás tesztelése: Lilliefors-próba

H_0 : a testmagasság normális eloszlású $m = 174,3$ várható értékkel és $\sigma = 11,5$ szórással

H_1 : a testmagasság eloszlása ettől különböző

Szignifikanciaszint: $\alpha = 0,05$.

Próbastatisztika (ugyanaz, mint a Kolmogorov–Szmirnov-próbánál):

$$D = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| = 0,068$$

A kritikus érték: 0,09

A p -érték: 0,367

A normalitás tesztelése: Lilliefors-próba

H_0 : a testmagasság normális eloszlású $m = 174,3$ várható értékkel és $\sigma = 11,5$ szórással

H_1 : a testmagasság eloszlása ettől különböző

Szignifikanciaszint: $\alpha = 0,05$.

Próbastatisztika (ugyanaz, mint a Kolmogorov–Szmirnov-próbánál):

$$D = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| = 0,068$$

A kritikus érték: 0,09

A p -érték: 0,367

Mivel $0,068 = D < D_{\text{krit}} = 0,09$, illetve $p = 0,367 > 0,05 = \alpha$, a szignifikanciaszintet $\alpha = 0,05$ -nek választva **elfogadható**, hogy a testmagasság normális eloszlású a megadott paraméterekkel, nincs szignifikáns eltérés.

Normális eloszlásra vonatkozó próbák: példa

A táblázat a p -értékeket mutatja az egyes minták és próbák esetén.

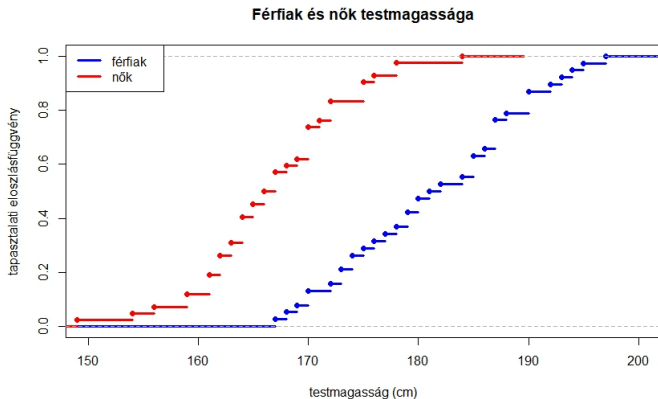
Szignifikanciaszint: $\alpha = 0,05$

minta	Lilliefors (Kolmogorov–Szmirnov)	Shapiro–Wilk
testmagasság	0,367	0,066
max. vízállás	0,014	0,002
max. vízállás logaritmusa	0,22	0,629

Tehát a testmagasság és Duna havi legnagyobb vízállásának logaritmusáról elfogadható, hogy normális eloszlású, a havi legnagyobb vízállás eloszlása viszont szignifikánsan eltér a normális eloszlástól.

A Duna havi legnagyobb vízállásáról azt mondhatjuk, hogy a logaritmusa normális eloszlású, vagyis **lognormális eloszlású**.

Kolmogorov–Szmirnov-próba: homogenitásvizsgálat



A férfiak ($n = 38$ megfigyelés) és nők ($m = 42$ megfigyelés) testmagasságának tapasztalati eloszlásfüggvénye

Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

H_0 : az X_1, \dots, X_n és Y_1, \dots, Y_m minták ugyanabból az eloszlásból származnak

H_1 : a minták különböző eloszlásból származnak.

Próbastatisztika:

$$D_{m,n} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|,$$

ahol \hat{F}_n az X , a \hat{G}_m pedig az Y minta tapasztalati eloszlásfüggvénye.

Ha $D_{m,n} > D_{\text{krit}}$ (vagy $p < \alpha$), akkor elutasítjuk H_0 -t, a minták eloszlása szignifikánsan különböző (itt D_{krit} a megfelelő Kolmogorov–Szmirnov-próba kritikus értéke).

Ha $D < D_{\text{krit}}$, (vagy $p > \alpha$) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés a minták eloszlása között.

A kritikus értékek az alábbi összefüggés alapján határozhatók meg:

$$\lim_{m,n \rightarrow \infty} \mathbb{P} \left(\sqrt{\frac{mn}{m+n}} D_{m,n} \right) = \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 y^2}.$$

Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

H_0 : az X_1, \dots, X_n és Y_1, \dots, Y_m minták ugyanabból az eloszlásból származnak

H_1 : a minták különböző eloszlásból származnak.

Próbastatisztika:

$$D_{m,n} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|,$$

ahol \hat{F}_n az X , a \hat{G}_m pedig az Y minta tapasztalati eloszlásfüggvénye. A nullhipotézist elutasítjuk, ha D nagyobb a kritikus értéknél.

```
> ks.test(ferfi, no, alternative="two.sided")
```

Two-sample Kolmogorov-Smirnov test

data: ferfi and no

D = 0.6754, **p-value = 2.486e-08**

alternative hypothesis: two-sided

A férfiak ($n = 38$ megfigyelés) és a nők ($m = 42$ megfigyelés) testmagasságának eloszlása szignifikánsan különböző.

Homogenitásvizsgálat

A próbastatisztika:

$$\chi^2 = \sum_{k=1}^r \frac{\left(\frac{N_k}{n} - \frac{M_k}{m}\right)^2}{\frac{N_k}{n} + \frac{M_k}{m}} \cdot n \cdot m.$$

A szabadsági fok: $f = r - 1$.

c_{krit} : az f szabadsági fokú χ^2 -próba kritikus értéke α terjedelem mellett.

- $\chi^2 < c_{\text{krit}}$ (azaz $p \geq \alpha$): elfogadjuk H_0 -t, nem találtunk szignifikáns eltérést az eloszlások között.
- $\chi^2 > c_{\text{krit}}$ (azaz a $p < \alpha$): elutasítjuk H_0 -t, az eloszlások szignifikánsan eltérnek.

Homogenitásvizsgálat: példa

Két városban felmérték a háztartások létszámát, az elsőben $n = 249$, a másodikban $m = 301$ elemű mintát vizsgálva. A szignifikanciaszintet $\alpha = 0,05$ -nek választva állíthatjuk-e, hogy a két városban szignifikánsan eltérő a háztartások létszámának eloszlása?

létszám	1	2	3	4	> 5
első város	37	86	54	49	23
második város	45	94	67	56	39
első város, arány	0,15	0,35	0,22	0,2	0,09
második város, arány	0,18	0,38	0,27	0,22	0,16

Homogenitásvizsgálat: példa

Két városban felmérték a háztartások létszámát, az elsőben $n = 249$, a másodikban $m = 301$ elemű mintát vizsgálva. A szignifikanciaszintet $\alpha = 0,05$ -nek választva állíthatjuk-e, hogy a két városban szignifikánsan eltérő a háztartások létszámának eloszlása?

létszám	1	2	3	4	> 4
első város	37	86	54	49	23
második város	45	94	67	56	39

Minden osztályba esik legalább 4 megfigyelés.

$$\chi^2 = \sum_{k=1}^r \frac{\left(\frac{N_k}{n} - \frac{M_k}{m}\right)^2}{\frac{N_k + M_k}{n \cdot m}} \cdot n \cdot m = \left(\frac{(37/249 - 45/301)^2}{37 + 45} + \frac{(86/249 - 94/301)^2}{86 + 94} + \dots + \frac{(23/249 - 39/301)^2}{23 + 39} \right) \cdot 249 \cdot 301 = 2,23.$$

Homogenitásvizsgálat: példa

Két városban felmérték a háztartások létszámát. A szignifikanciaszintet $\alpha = 0,05$ -nek választva állíthatjuk-e, hogy a két városban szignifikánsan eltérő a háztartások létszámának eloszlása?

létszám	1	2	3	4	> 4
első város	37	86	54	49	23
második város	45	94	67	56	39

Az osztályok száma $r = 5$.

$$\chi^2 = 2,23; \quad f = r - 1 = 4; \quad \alpha = 0,05 \quad c_{\text{krit}} = 9,49$$

$\chi^2 = 2,23 < c_{\text{krit}} = 9,49$, elfogadjuk a nullhipotézist, a kétféle homok szemcseméretének eloszlása **nem tér el szignifikánsan**. A p -érték: $p = 0,31 > 0,05$.

Kvantilisek

Az X valószínűségi változó z -kvantilise a legkisebb olyan q szám, melyre teljesül, hogy $\mathbb{P}(X \leq q) \geq z$.

A tapasztalati z -kvantilis a legkisebb olyan q szám, melyre $\hat{F}_n(q) \geq z$ teljesül, vagy egy másik lehetőség (ezen kívül is több definíciót szoktak használni):

Definíció (Tapasztalati kvantilis)

Legyen $X_1^* \leq X_2^* \leq \dots \leq X_n^*$ rendezett minta, és $z \in [0, 1]$ adott szám. Ekkor a minta tapasztalati z -kvantilise:

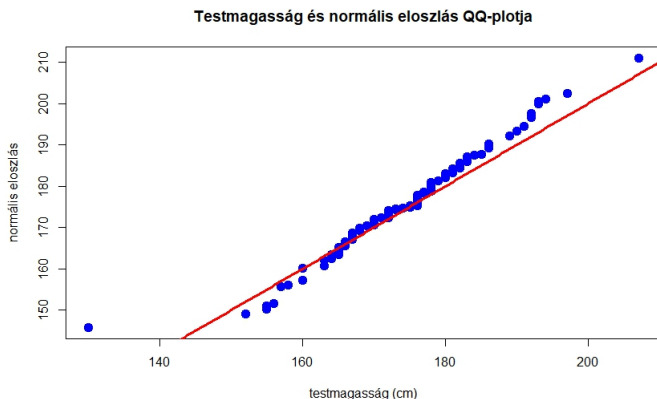
$$\hat{q}_z = X_{\lfloor z(n+1) \rfloor}^* + (z(n+1) - \lfloor z(n+1) \rfloor) \cdot (X_{\lfloor z(n+1) \rfloor + 1}^* - X_{\lfloor z(n+1) \rfloor}^*).$$

Első kvartilis: $z = 1/4$ -kvantilis, harmadik kvartilis: $z = 3/4$ -kvantilis, a medián pedig a $z = 1/2$ -hez tartozó tapasztalati kvantilis.

QQ-plot

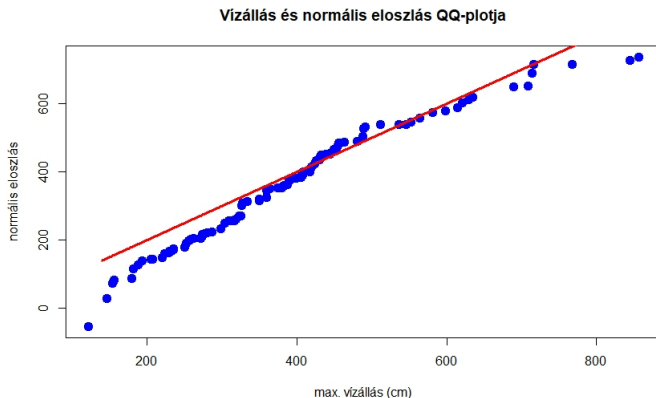
- a QQ-plot két minta eloszlásának az összehasonlítására szolgál, a kvantilisok összehasonlításával
- ha a tapasztalati z-kvantilis az első mintában q_1 , a másodikban q_2 , akkor a (q_1, q_2) pontba kerül egy pont
- minél inkább egyezik a két minta eloszlása, annál közelebb lesz a QQ-plot az $y = x$ egyeneshez

A testmagasság és a becsült normális eloszlás QQ-plotja



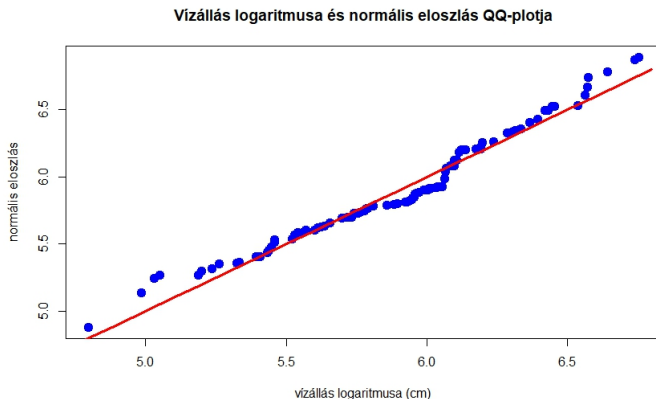
A testmagasság adatok és egy szintén 96 elemű, $\bar{X} = 174,3$ várható értékű és $s_n^* = 11,5$ szórású normális eloszlású minta QQ-plotja

A vízállás és a becült normális eloszlás QQ-plotja



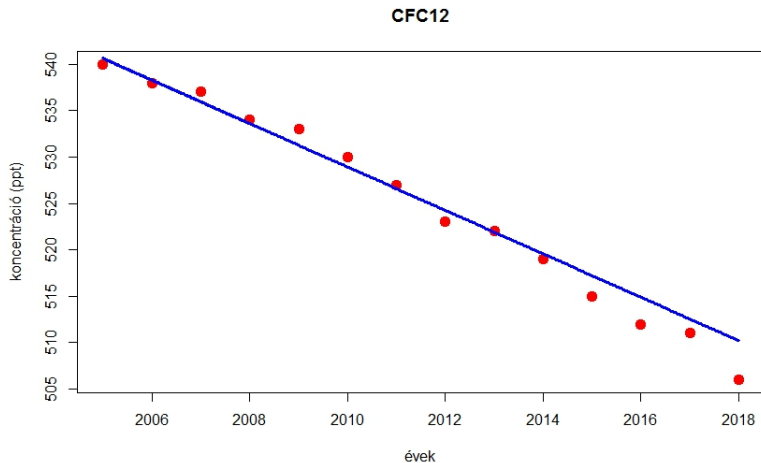
A havi legnagyobb vízállás adatok és egy szintén 96 elemű, $\bar{X} = 352$ várható értékű és $s_n^* = 157,4$ szórású normális eloszlású minta QQ-plotja (ez szignifikánsan eltért a normális eloszlástól)

A vízállás logaritmusának és a becsült normális eloszlás QQ-plotja



A havi legnagyobb vízállás adatok és egy szintén 96 elemű, $\bar{X} = 5,89$ várható értékű és $s_n^* = 0,41$ szórású normális eloszlású minta QQ-plotja (ez nem tért el szignifikánsan a normális eloszlástól)

Lineáris regresszió



A CFC-12 (freon) gáz koncentrációja az Antarktison és az adatokra illesztett egyenes (forrás: ESRL, USA)

Lineáris regresszió

Egyenes illesztése a **legkisebb négyzetek módszerével**:

Állítás (Lineáris regresszió)

Legyenek $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ adott számpárok. Azokat az a és b együtthatókat keressük, melyre a

$$h^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

mennyiség minimális. Ennek megoldása:

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2}; \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

A példában: $\hat{a} = -2,63$; $\hat{b} = 5807,7$ (a b együttható neve: intercept)

Lineáris modell: példa R-ben

```
> cfc12<-c(540, 538, 537, 534, 533, 530, 527, 523, 522, 519, 515, 511, 506)
```

```
> ev<-c(seq(from=2005, to=2018, by=1))
```

```
> summary(lm(cfc12 ~ ev))
```

```
Call:  lm(formula = cfc12 ~ ev)
```

Residuals:	Min	1Q	Median	3Q	Max
	-1.8571	-0.8736	0.2088	0.8709	1.6483

Lineáris modell: példa R-ben (folytatás)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5807.73626	159.19290	36.48	1.15e-13 ***
ev	-2.62637	0.07914	-33.19	3.55e-13 ***

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.194 on 12 degrees of freedom

Multiple R-squared: 0.9892, Adjusted R-squared: 0.9883

F-statistic: 1101 on 1 and 12 DF, p-value: 3.554e-13

Lineáris modell

Definíció (Lineáris modell)

Legyenek $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ valószínűségi változók, és tegyük fel, hogy valamely a, b valós számokra

$$Y_i = aX_i + b + \varepsilon_i,$$

ahol $\varepsilon_1, \dots, \varepsilon_n$ független $N(0, \sigma^2)$ normális eloszlású valószínűségi változók. Az így kapott (X_i, Y_i) párok együttes eloszlását lineáris modellnek nevezzük.

Az X_i valószínűségi változókat magyarázó változóknak, az ε_i valószínűségi változókat hibának szokták nevezni.

Becslések a lineáris modellben

Állítás

A lineáris modellben az a, b együtthatók maximumlikelihood-becslése a következőképpen írható:

$$\hat{a} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{k=1}^n (X_k - \bar{X})^2}; \quad \hat{b} = \bar{Y} - \hat{a}\bar{X}.$$

Továbbá, ezek a becslések torzítatlan becslései az a és b paramétereknek. A hiba szórásának becslése:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2.$$

A becslések szórása:

$$D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}; \quad D(\hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

Előrejelzés a lineáris modellben

Állítás

Legyen x^* adott szám. A lineáris modellből kapott előrejelzés az Y véletlen folyamat x^* pontban felvett értékére:

$$\hat{a}x^* + \hat{b}.$$

Az előrejelzés szórása:

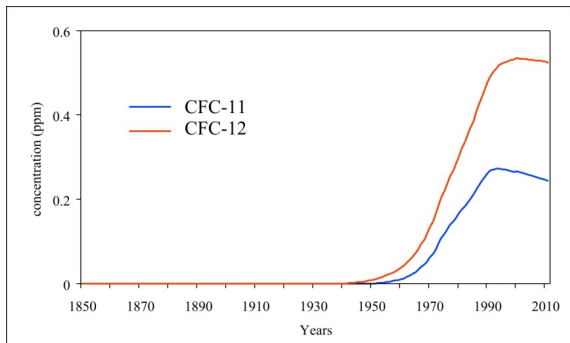
$$D(\hat{a}x^* + \hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

Az előrejelzés szórásának becslésekor a σ értéket gyakran $\hat{\sigma}$ -val helyettesítik.

A példában: előrejelzés $x^* = 2019$ -re:

$$\hat{a} \cdot x^* + \hat{b} = -2,63 \cdot 2019 + 5807,7 = 497,7.$$

Előrejelzés a lineáris modellben



A CFC-11 és CFC-12 (freon) gáz koncentrációja (forrás: elte.promt.hu)

Reziduálisok

Reziduálisok: $Y_i - \hat{a}X_i - \hat{b}$ (ezeknek a négyzetösszege minimális)

A teljes ingadozás (total sum of squares): $\sum_{j=1}^n (Y_j - \bar{Y})^2$.

Definíció

A megmagyarázott ingadozás részaránya (coefficient of determination):

$$R^2 = \frac{[\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})]^2}{[\sum_{k=1}^n (X_k - \bar{X})^2][\sum_{k=1}^n (Y_k - \bar{Y})^2]}.$$

Az R^2 értéke 0 és 1 közé esik.

Értelmezés: minél közelebb van 1-hez, annál inkább jó közelítést ad a lineáris modell. Ugyanakkor R érzékeny a kiugró értékekre, néhány kiugró esetén R^2 lecsökken.

A példában: $R^2 = 0,98$, vagyis jól illeszkedik a lineáris modell.

Konfidenciaintervallumok

$1 - \alpha$ megbízhatósági szintű konfidenciaintervallum a -ra:

$$\left(\hat{a} - t_{n-2, \alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}, \hat{a} + t_{n-2, \alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right),$$

ahol $t_{n-2, \alpha}$ az $f = n - 2$ szabadsági fokú α terjedelmű kétoldali t -próba kritikus értéke.

Konfidenciaintervallumok

$1 - \alpha$ megbízhatósági szintű konfidenciaintervallum a -ra:

$$\left(\hat{a} - t_{n-2, \alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}, \hat{a} + t_{n-2, \alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right),$$

ahol $t_{n-2, \alpha}$ az $f = n - 2$ szabadsági fokú α terjedelmű kétoldali t -próba kritikus értéke.

Az x^* pontban az előrejelzett érték becslése $\hat{a} \cdot x^* + \hat{b}$.

$1 - \alpha$ megbízhatósági szintű konfidenciaintervallum $ax^* + b$ -re, azaz az x^* -ban felvett érték várható értékére:

$$\left(\hat{a}x^* + \hat{b} \pm t_{n-2, \alpha} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \right).$$

Az $a = 0$ hipotézis ellenőrzése

Lineáris modell: $Y_i = aX_i + b + \varepsilon_i$. Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan eltér 0-tól?

$$H_0: a = 0 \quad H_1: a \neq 0$$

Az $a = 0$ hipotézis ellenőrzése

Lineáris modell: $Y_i = aX_i + b + \varepsilon_i$. Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan eltér 0-tól?

$$H_0: a = 0 \quad H_1: a \neq 0$$

Kétoldali t -próbát végezhetünk az alábbi próbastatisztikával és $f = n - 2$ szabadsági fokkal:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

Ha $|t| > t_{n-2, \alpha}$, azaz $p < \alpha$, akkor elutasítjuk H_0 -t, az egyenes meredeksége szignifikánsan eltér 0-tól (itt $t_{n-2, \alpha}$ az α szignifikanciaszintű $f = n - 2$ szabadsági fokú kétoldali t -próba kritikus értéke).

Ha $|t| \leq t_{n-2, \alpha}$, azaz $p \geq \alpha$, akkor elfogadjuk H_0 -t, az egyenes meredeksége nem tér el szignifikánsan 0-tól.

Az $a = 0$ hipotézis ellenőrzése

Lineáris modell: $Y_i = aX_i + b + \varepsilon_i$

$H_0: a = 0$ $H_1: a \neq 0$

Az $a = 0$ hipotézis ellenőrzése

Lineáris modell: $Y_i = aX_i + b + \varepsilon_i$

$H_0: a = 0$ $H_1: a \neq 0$

Kétoldali t -próbát végezhetünk az alábbi próbastatisztikával:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

A példában

$$t = -33,19; \quad \alpha = 0,05; \quad n = 14; \quad f = n - 2 = 12; \quad c_{\text{krit}} = 2,19.$$

Az $a = 0$ hipotézis ellenőrzése

Lineáris modell: $Y_i = aX_i + b + \varepsilon_i$

$H_0: a = 0$ $H_1: a \neq 0$

Kétoldali t -próbát végezhetünk az alábbi próbastatisztikával:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

A példában

$$t = -33,19; \quad \alpha = 0,05; \quad n = 14; \quad f = n - 2 = 12; \quad c_{\text{krit}} = 2,19.$$

Mivel $|t| = 33,19 > c_{\text{krit}} = 2,19$, elutasítjuk a nullhipotézist, az egyenes meredeksége szignifikánsan eltér 0-tól. A p -érték: $p = 3,6 \cdot 10^{-13} < 0,05 = \alpha$.

Az $a \leq 0$ hipotézis ellenőrzése

Lineáris modell: $Y_i = aX_i + b + \varepsilon_i$. Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan nagyobb 0-nál?

$$H_0: a \leq 0 \quad H_1: a > 0$$

Az $a \leq 0$ hipotézis ellenőrzése

Lineáris modell: $Y_i = aX_i + b + \varepsilon_i$. Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan nagyobb 0-nál?

$$H_0: a \leq 0 \quad H_1: a > 0$$

Egyoldali t -próbát végezhetünk az alábbi próbastatisztikával és $f = n - 2$ szabadsági fokkal:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

Ha $t > \bar{t}_{n-2, \alpha}$, azaz $p < \alpha$, akkor elutasítjuk H_0 -t, az egyenes meredeksége szignifikánsan több 0-nál (itt $\bar{t}_{n-2, \alpha}$ az α terjedelmű $f = n - 2$ szabadsági fokú egyoldali t -próba kritikus értéke α szignifikanciaszint mellett).

Ha $t \leq \bar{t}_{n-2, \alpha}$, azaz $p \geq \alpha$, akkor elfogadjuk H_0 -t, az egyenes meredeksége nem szignifikánsan pozitív.

Többváltozós lineáris regresszió (multiple linear regression)

Az Y változót fejezzük ki az X_1, \dots, X_p valószínűségi változók lineáris függvényeként, de az együtthatókat ismeretlenek tekintjük:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i,$$

ahol ε_i független $N(0, \sigma^2)$ normális eloszlású valószínűségi változók. Vektoros formában: $\underline{Y} = \underline{X}\underline{a} + \underline{\varepsilon}$, ahol \underline{X} az $X_{i,j}$ megfigyelésekből készített mátrix.

Többváltozós lineáris regresszió (multiple linear regression)

Az Y változót fejezzük ki az X_1, \dots, X_p valószínűségi változók lineáris függvényeként, de az együtthatókat ismeretlennek tekintjük:

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i,$$

ahol ε_i független $N(0, \sigma^2)$ normális eloszlású valószínűségi változók. Vektoros formában: $\underline{Y} = X\underline{a} + \underline{\varepsilon}$, ahol X az $X_{i,j}$ megfigyelésekből készített mátrix.

Ezután az a együtthatók becslése: $\hat{a} = (X^T X)^{-1} X^T \underline{Y}$.

Ekkor is megfelelő próbastatisztikával t -próbával tesztelhetők az $a_i = 0$ hipotézisek, vagyis ellenőrizhető, hogy az Y mely mennyiségektől függ szignifikánsan.