

Pozitív korreláció (10. előadás)

Tekintsük a függetlenségvizsgálatot abban az esetben, ha mindkét szempont szerint két osztály van.

H_0 : a két szempont között **nincs pozitív korreláció**

H_1 : a két szempont között **pozitív korreláció** van, azaz $\mathbb{P}(A_1 \cap B_1) > \mathbb{P}(A_1)\mathbb{P}(B_1)$.

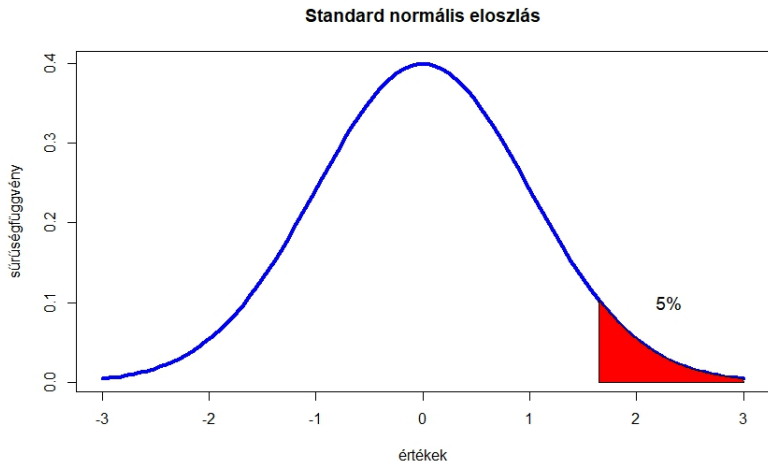
A próbastatisztika (H_0 mellett standard normális eloszlású):

$$z = \sqrt{n} \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1\cdot} \cdot N_{2\cdot} \cdot N_{\cdot 1} \cdot N_{\cdot 2}}}$$

Ha $z > \Phi^{-1}(1 - \alpha)$, akkor elutasítjuk H_0 -t, szignifikáns pozitív korreláció van; különben elfogadjuk H_0 -t, nincs szignifikáns pozitív korreláció.

A p -érték: $1 - \Phi(z)$, ahol $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$.

Az egyoldali z-próba kritikus értéke



Ha $z > \Phi^{-1}(1 - \alpha)$, akkor elutasítjuk H_0 -t, szignifikáns pozitív korreláció van; különben elfogadjuk H_0 -t, nincs szignifikáns pozitív korreláció.

Pozitív korreláció: példa

Vérnyomás-szűrővizsgálatnál a 40 évesnél idősebbek közül 24-nek magas, 62-nek megfelelő volt a vérnyomása, a 40 évesnél nem idősebbek közül 12-nek volt magas, 88-nak megfelelő. Állíthatjuk-e $\alpha = 0,05$ szignifikanciaszint mellett, hogy a 40 évesnél idősebbek között gyakoribb a magas vérnyomás?

Pozitív korreláció: példa

Vérnyomás-szűrővizsgálatnál a 40 évesnél idősebbek közül 24-nek magas, 62-nek megfelelő volt a vérnyomása, a 40 évesnél nem idősebbek közül 12-nek volt magas, 88-nak megfelelő. Állíthatjuk-e $\alpha = 0,05$ szignifikanciaszint mellett, hogy a 40 évesnél idősebbek között gyakoribb a magas vérnyomás?

A_1 : 40 évesnél nagyobb életkor; A_2 : legfeljebb 40 éves életkor.

B_1 : magas vérnyomás; B_2 : megfelelő vérnyomás.

H_0 : nincs pozitív korreláció;

H_1 : pozitív korreláció van.

$N_{11} = 24$; $N_{12} = 62$; $N_{21} = 12$; $N_{22} = 88$; $n = 186$.

$$z = \sqrt{n} \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}} = \sqrt{186} \cdot \frac{24 \cdot 88 - 62 \cdot 12}{\sqrt{86 \cdot 100 \cdot 36 \cdot 150}} = 2,74.$$

Pozitív korreláció: példa

Vérnyomás-szűrővizsgálatnál a 40 évesnél idősebbek közül 24-nek magas, 62-nek megfelelő volt a vérnyomása, a 40 évesnél nem idősebbek közül 12-nek volt magas, 88-nak megfelelő. Állíthatjuk-e $\alpha = 0,05$ szignifikanciaszint mellett, hogy a 40 évesnél idősebbek között gyakoribb a magas vérnyomás?

A_1 : 40 évesnél nagyobb életkor; A_2 : legfeljebb 40 éves életkor.

B_1 : magas vérnyomás; B_2 : megfelelő vérnyomás.

H_0 : nincs pozitív korreláció;

H_1 : pozitív korreláció van.

$N_{11} = 24$; $N_{12} = 62$; $N_{21} = 12$; $N_{22} = 88$; $n = 186$.

$$z = \sqrt{n} \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}} = \sqrt{186} \cdot \frac{24 \cdot 88 - 62 \cdot 12}{\sqrt{86 \cdot 100 \cdot 36 \cdot 150}} = 2,74.$$

Mivel $2,74 > \Phi^{-1}(0,95) = 1,645$, így elutasítjuk a nullhipotézist. A nagyobb életkor és a magas vérnyomás között **szignifikáns pozitív** korreláció van. A p -érték: $1 - \Phi(2,74) = 0,003 < 0,05$.

Pozitív korreláció

A függetlenség vagy a pozitív korreláció vizsgálatánál a következőket érdemes figyelembe venni.

- minden osztályba essen legalább 6 megfigyelés
- a pozitív korreláció **nem jelent ok-okozati összefüggést**
- ha sok mennyiséget vizsgálunk, előre kell eldönteni (az adatok ismerete nélkül), hogy hol keressük a pozitív összefüggést: öt mennyiség között 10 pár van, így jó eséllyel lesz olyan pár, ahol tévesen szignifikáns összefüggést vagy pozitív korrelációt találhatunk ($\alpha = 0,05$ szignifikanciaszintet választva)

χ^2 -próba: homogenitásvizsgálat

Legyenek X, Y valószínűségi változók, A_1, \dots, A_r teljes eseményrendszer.

H_0 : $\mathbb{P}(X \in A_k) = \mathbb{P}(Y \in A_k)$ minden $k = 1, 2, \dots, r$ -re.

H_1 : van legalább egy k , melyre $\mathbb{P}(X \in A_k) \neq \mathbb{P}(Y \in A_k)$.

$X_1, \dots, X_n, Y_1, \dots, Y_m$ független minta, melyre $X_i \sim X, Y_i \sim Y$.

N_k az A_k gyakorisága az \underline{X} mintában;

M_k az A_k gyakorisága az \underline{Y} mintában.

Ha $N_k \geq 4$ vagy $M_k \geq 4$ nem teljesül, osztályokat vonunk össze.

A próbastatisztika:

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

Homogenitásvizsgálat

A próbastatisztika:

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

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 két város háztartásainak méretének eloszlása **nem tér el szignifikánsan**. A p -érték: $p = 0,31 > 0,05$.

Nem-paraméteres próbák

Ha egy ismeretlen mennyiségnek nem csak a várható értékét vagy szórását vizsgáljuk, az alábbi kérdések is fontosak:

- 1 **Illeszkedésvizsgálat:** a minta egy adott, folytonos eloszlásból származik-e? Például, igaz-e, hogy egy véletlenszerűen választott ember havi jövedelme a minimálbérrel osztva egyes típusú Pareto-eloszlású $\alpha = 2,5$ paraméterrel?
- 2 **Normalitás tesztelése:** igaz-e, hogy egy minta normális eloszlásból származik? 100 ember testmagasságát megmérve mikor mondhatjuk, hogy elfogadható ez a feltételezés, és mikor állíthatjuk, hogy a testmagasság eloszlása szignifikánsan eltér a normális eloszlástól?
- 3 **Homogenitásvizsgálat:** két minta ugyanabból az eloszlásból származik-e? Például: megkérdezzük két város 100 – 100 véletlenszerűen választott lakóját a jövedelméről. Állíthatjuk-e az adatok alapján, hogy a két városban a jövedelmek eloszlása szignifikánsan eltérő? A két eloszlás akkor egyezik meg, ha minden t -re igaz, hogy a legfeljebb t jövedelműek aránya megegyezik a két esetben.

Nem-paraméteres próbák

Egy lehetőség: **diszkrétizáljuk** a megfigyeléseket, vagyis közel azonos hosszúságú intervallumokba osztjuk be őket (például jövedelmi kategóriákba), és ezután χ^2 -próbát végzünk. Ha szükséges, a paramétereket maximumlikelihood-módszerrel becsüljük. Ebben az esetben viszont a végeredmény akár függhet is az intervallumok (kategóriák) kialakításától.

Nem-paraméteres próbák

Egy lehetőség: **diszkrétizáljuk** a megfigyeléseket, vagyis közel azonos hosszúságú intervallumokba osztjuk be őket (például jövedelmi kategóriákba), és ezután χ^2 -próbát végzünk. Ha szükséges, a paramétereket maximumlikelihood-módszerrel becsüljük. Ebben az esetben viszont a végeredmény akár függhet is az intervallumok (kategóriák) kialakításától.

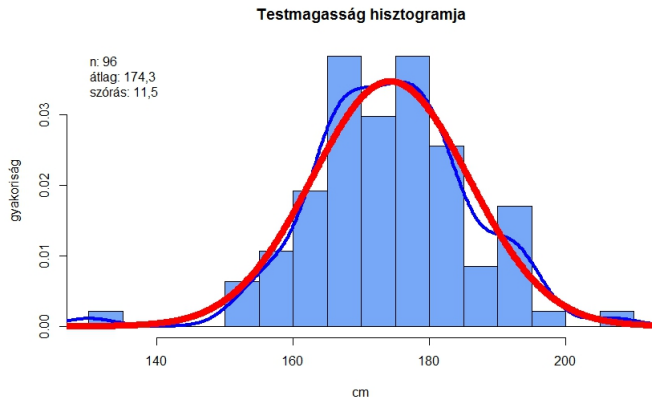
Tapasztalati eloszlásfüggvények távolságát használó próbák:

- 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)
- leíró statisztikai eszközökkel: ferdeségi, csúcossági együtthatók kiszámítása (skewness, kurtosis)

Testmagasság és normális eloszlás



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

Tapasztalati eloszlásfüggvény

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

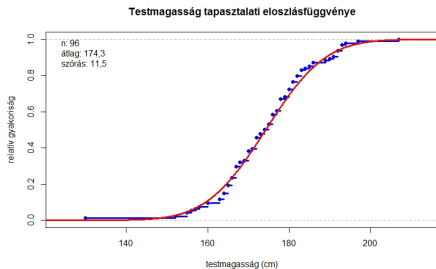
$$F(t) = \mathbb{P}(X \leq t)$$

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

Definíció

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

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



Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat

Legyen G **egy rögzített, folytonos eloszlásfüggvény**, vagyis $G : \mathbb{R} \rightarrow [0, 1]$ monoton növekvő, folytonos, $-\infty$ -beli limesze 0, ∞ -beli limesze 1.

H_0 : a minta valódi eloszlásfüggvénye G , azaz $\mathbb{P}(X_1 \leq t) = G(t)$ minden t -re

H_1 : a minta valódi eloszlásfüggvénye G -től különböző

Emlékeztetőül: a Glivenko–Cantelli-tétel, vagyis a statisztika alaptétele szerint $\hat{F}_n(t)$, azaz a mintában a t -nél nem nagyobb mintaelemek aránya $n \rightarrow \infty$ esetén X_1 eloszlásfüggvényéhez konvergál – ezért $\hat{F}_n(t)$ -t hasonlítjuk össze G -vel.

Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat

Próbastatisztika, ami a tapasztalati eloszlásfüggvény és G távolságát méri, úgy, hogy a legnagyobb különbséget veszi, abszolút értékben:

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

ahol F_n a minta tapasztalati eloszlásfüggvénye. H_0 teljesülése esetén D_n eloszlása (megfelelő normálás után) Kolmogorov–Szmirnov-eloszlású.

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, ez táblázatból kiolvasható.

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

Ha $n \geq 35$, akkor a kritikus értékre az alábbi közelítés adható (α szignifikanciaszint mellett):

$$D_{\text{krit}} \approx \frac{\sqrt{\log(4/\alpha)}}{\sqrt{n}}.$$

Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat

Kolmogorov–Szmirnov-próba, példa. Tekintsük a GDP volumenindexének (az előző évi érték osztva az aktuális értékkel) adatait 1993–2018 között (évenként van egy megfigyelésünk). Elfogadható-e, hogy az eloszlás egy $a = 70, b = 2$ paraméterű Beta-eloszlás 0,06-tal eltolva? Ez azt jelentené, hogy a sűrűségfüggvény egy megfelelő polinom.

A próbát elvégezve:

```
ks.test(gdp-0.06, "pbeta", 70, 2)
```

```
One-sample Kolmogorov-Smirnov test
```

```
data:  gdp - 0.06
```

```
D = 0.1666, p-value = 0.5456
```

```
alternative hypothesis:  two-sided
```

Az eloszlásfüggvények közötti legnagyobb különbség tehát 0,167 (talán $t = 1,022$ vagy 1,045 körül).

Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat

Kolmogorov–Szmirnov-próba, példa. Tekintsük a GDP volumenindexének (az előző évi érték osztva az aktuális értékkel) adatait 1993–2018 között (évenként van egy megfigyelésünk). Elfogadható-e, hogy az eloszlás egy $a = 70, b = 2$ paraméterű Beta-eloszlás $0,06$ -tal eltolva? Ez azt jelentené, hogy a sűrűségfüggvény egy megfelelő polinom.

A próbát elvégezve:

```
ks.test(gdp-0.06, "pbeta", 70, 2)
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: gdp - 0.06
```

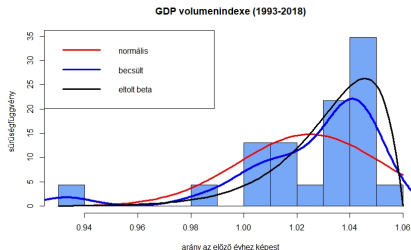
```
D = 0.1666, p-value = 0.5456
```

```
alternative hypothesis: two-sided
```

Az eloszlásfüggvények közötti legnagyobb különbség tehát $0,167$ (talán $t = 1,022$ vagy $1,045$ körül).

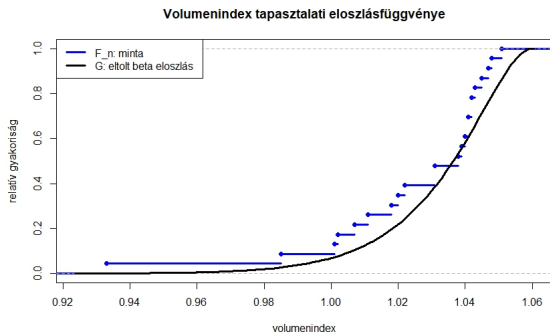
A p -érték több $0,05$ -nél, így a hipotézis elfogadható.

Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat



A GDP volumenindexének (az érték osztva az előző évi értékkel) 1993-2018 közötti értékeinek hisztogramja, a becsült normális eloszlás és a becsült sűrűségfüggvény illetve az eltoló Beta-eloszlás sűrűségfüggvénye (az adatok forrása: KSH)

Kolmogorov–Szmirnov-próba: illeszkedésvizsgálat



A GDP volumenindexének (az érték osztva az előző évi értékkel) 1993-2018 közötti értékeinek tapasztalati eloszlásfüggvénye és a megadott G eloszlásfüggvény (az adatok forrása: KSH)

A normalitás tesztelése: Lilliefors-próba

A normális eloszlás paramétereit először meg kell becsülni az adatok alapján.

H_0 : a minta normális eloszlásból származik (valamilyen m, σ paraméterekkel)

H_1 : a minta eloszlása nem normális eloszlás

Legyen \bar{X} a mintaátlag, s_n^* a korrigált tapasztalati szórás, \hat{G} pedig az m várható értékű és σ szórású normális eloszlás eloszlásfüggvénye: $\hat{G}(t) = \Phi((t - \bar{X})/s_n^*)$. Ekkor a próbastatisztika (ugyanaz, mint a Kolmogorov–Szmirnov-próbánál):

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

Ha $D_n > \bar{D}_{\text{krit}}$ (vagy $p < \alpha$), akkor elutasítjuk H_0 -t, a minta eloszlása szignifikánsan eltér a normális eloszlástól (itt \bar{D}_{krit} a megfelelő Lilliefors-próba kritikus értéke).

Ha $D_n < \bar{D}_{\text{krit}}$, (vagy $p > \alpha$) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés a normális eloszlástól.

A normalitás tesztelése: Lilliefors-próba

A korábbi ábrához tartozó, 96 elemű, testmagasságra vonatkozó példában:

```
require(nortest)
```

```
> lillie.test(testmagassag)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: testmagassag
```

```
D = 0.0609, p-value = 0.5307
```

Mivel $0,068 = D < D_{\text{krit}} = 0,09$, illetve $p = 0,5307 > 0,05 = \alpha$, a szignifikanciaszintet $\alpha = 0,05$ -nek választva **elfogadható**, hogy a testmagasság normális eloszlású valamilyen paraméterekkel, nincs szignifikáns eltérés a normális eloszlástól.

A normalitás tesztelése: Lilliefors-próba

Ugyanakkor GDP volumenindexére vonatkozó példában

```
> lillie.test(gdp)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data:  gdp
```

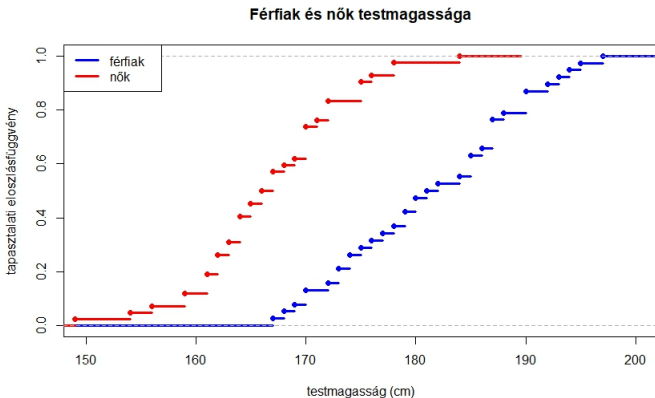
```
D = 0.2055, p-value = 0.01287
```

Itt $p < 0,05$, vagyis a nullhipotézist elutasítjuk, a volumenindex eloszlása szignifikánsan eltér a normális eloszlástól.

Megjegyzés: a rendezett minta kovarianciamátrixát használó Shapiro–Wilk-próbánál a testmagasság esetében $p = 0,36$, míg a gdp volumenindexe esetében $p = 0,0002$. Ilyenkor érdemes lehet részletesebben megnézni, hogy melyik próbánál mit kell feltelezni, milyen az adatsor (vannak-e például kiugró értékek).

Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

Állíthatjuk-e, hogy a férfiak és a nők testmagasságának **eloszlása** szignifikánsan eltérő? Ez a kérdés nem csak a várható értékre és a szórásra vonatkozik, hanem magára az eloszlásra.



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, azaz minden t valós számra teljesül, hogy $\mathbb{P}(X_j \leq t) = \mathbb{P}(Y_j \leq t)$.

H_1 : a minták **különböző eloszlásból** származnak, azaz van olyan t valós szám, amire $\mathbb{P}(X_j \leq t) \neq \mathbb{P}(Y_j \leq t)$.

A próbastatisztika, ami H_0 esetén Kolmogorov–Szmirnov-eloszlású:

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

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 közelíthetők:

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

Homogenitásvizsgálat: példa

H_0 : az X_1, \dots, X_n és Y_1, \dots, Y_m minták ugyanabból az eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása megegyezik.

H_1 : a minták különböző eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása eltérő.

A próbastatisztika a tapasztalati eloszlásfüggvények legnagyobb eltérése, az ábra alapján $t = 174$ környékén lehet:

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

```
> 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
```

Homogenitásvizsgálat: példa

H_0 : az X_1, \dots, X_n és Y_1, \dots, Y_m minták ugyanabból az eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása megegyezik.

H_1 : a minták különböző eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása eltérő.

A próbastatisztika a tapasztalati eloszlásfüggvények legnagyobb eltérése, az ábra alapján $t = 174$ környékén lehet:

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

```
> 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 p -érték kisebb 0,05-nél, a nullhipotézist elutasítjuk, a férfiak és a nők **testmagasságának eloszlása szignifikánsan különböző.**

Egyoldali eset

H_0 : az X és Y valószínűségi változók ugyanabból az eloszlásból származnak

H_1 : minden t valós számra $F(t) = \mathbb{P}(X \leq t) \leq G(t) = \mathbb{P}(Y \leq t)$, ahol F az X , a G pedig az Y eloszlásfüggvénye. Azaz $X \geq Y$ sztochasztikusan.

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

ahol \hat{F}_n az X , a \hat{G}_m pedig az Y minta tapasztalati eloszlásfüggvénye. A nullhipotézist elutasítjuk, ha D nagyobb az egyoldali Kolmogorov–Szmirnov-próba kritikus értékénél.

Egyoldali eset

H_0 : az X és Y valószínűségi változók ugyanabból az eloszlásból származnak

H_1 : minden t valós számra $F(t) = \mathbb{P}(X \leq t) \leq G(t) = \mathbb{P}(Y \leq t)$, ahol F az X , a G pedig az Y eloszlásfüggvénye. Azaz $X \geq Y$ sztochasztikusan.

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

ahol \hat{F}_n az X , a \hat{G}_m pedig az Y minta tapasztalati eloszlásfüggvénye. A nullhipotézist elutasítjuk, ha D nagyobb az egyoldali Kolmogorov–Szmirnov-próba kritikus értékénél.

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

```
Two-sample Kolmogorov-Smirnov test
```

```
data: ferfi and no
```

```
 $D^- = 0.6754$ , p-value = 1.243e-08
```

```
alternative hypothesis: the CDF of x lies below that of y
```

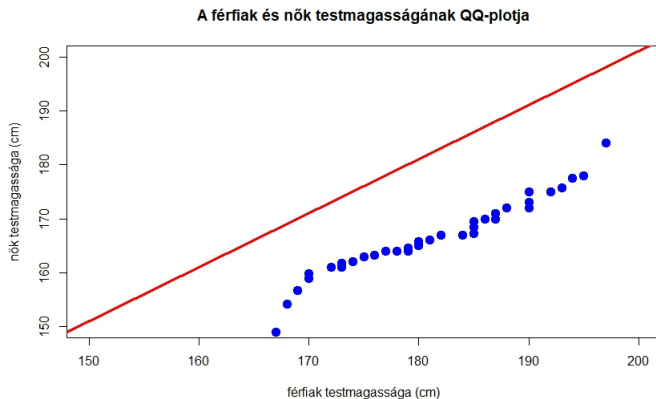
A férfiak testmagasságának eloszlása sztochasztikus értelemben szignifikánsan nagyobb, mint a nőké.

QQ-plot

Annak vizsgálatára, hogy két minta ugyanabból az eloszlásból származik-e (homogenitásvizsgálat) a leíró statisztikában a QQ-plot is gyakran használt ábrázolási mód.

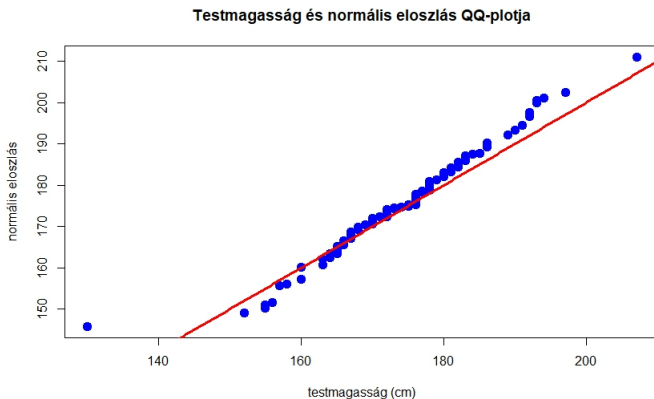
- a QQ-plot két minta eloszlásának az összehasonlítására szolgál, a kvantilisek összehasonlításával
- ha a tapasztalati z -kvantilis az első mintában q_1 , a másodikban q_2 , akkor a (q_1, q_2) pontba kerül egy pont
- minél inkább egyezik a két minta eloszlása, annál közelebb lesznek a tapasztalati eloszlásfüggvényik, ezért annál közelebb lesznek az ugyanahhoz a z -hez tartozó kvantiliseik egymáshoz, vagyis annál közelebb lesz a QQ-plot az $y = x$ egyeneshez.

QQ-plot



QQ-plot a férfiak és nők testmagasságának összehasonlítására, $n = 96$ elemű minta alapján

QQ-plot



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

Előjelpróba

Legyen $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ olyan minta, melyben a felsorolt párok függetlenek egymástól (de a párok két eleme nem feltétlenül független), és folytonos eloszlásúak.

Kétoldali ellenhipotézis:

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

Legyen W az olyan párok száma, amikre $Y_i > X_i$. H_0 esetén ez binomiális eloszlású, n renddel és $p = 0,5$ paraméterrel, hiszen minden pár esetében egymástól függetlenül $0,5$ valószínűséggel teljesül az egyenlőtlenség. A binomiális eloszlást a centrális határeloszlástétel alapján normális eloszlással közelítve jutunk az alábbi eljáráshoz. Legyen

$$z = \frac{W - n/2}{\sqrt{n/4}}. \quad (1)$$

Elutasítjuk a nullhipotézist, ha $|z| > \Phi^{-1}(1 - \alpha/2)$, különben elfogadjuk. A p -érték, ugyanúgy, ahogy a z -próbánál, $2(1 - \Phi(|z|))$ lesz.

Az egyoldali esetben: $H_0 : \mathbb{P}(X > Y) \geq \mathbb{P}(X < Y)$. $H_1 : \mathbb{P}(X > Y) < \mathbb{P}(X < Y)$.

Elutasítjuk a nullhipotézist, ha $z > \Phi^{-1}(1 - \alpha)$, különben elfogadjuk. A p -érték ilyenkor $1 - \Phi(z)$.

Wilcoxon-próba

Az előző hipotézisvizsgálati feladatban egy másik eljárást is gyakran használnak.

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

- Hagyjuk el azokat a párokat, ahol $X_j = Y_j$. Marad k pár.
- A megmaradt k párt állítsuk az $|Y_j - X_j|$ szerint növekvő sorrendbe. Minél nagyobb az eltérés, annál nagyobb súllyal fog számítani.
- Minden párra számítsuk ki, hogy hányadik ebben a sorrendben, legyen ez R_j . Az 1 a legkisebb, k a legnagyobb különbség. Ha egyenlők vannak, mindegyik azonos sorszámot kapjon, a megfelelő sorszámok átlagát.
- Ezt az R_j rangot szorozzuk meg $Y_j - X_j$ előjével, majd ezeket adjuk össze:

$$W = \sum_{j=1}^k \text{sgn}(Y_j - X_j) \cdot R_j.$$

- A W -t a Wilcoxon-próba kritikus értékeihez hasonlíthatjuk.