

Függetlenségvizsgálat, homogenitásvizsgálat, nemparaméteres próbák

1. Függetlenségvizsgálat

Ez az eljárás annak eldöntésére szolgál, hogy két szempont szerinti osztályba sorolás független-e egymástól. Például: egy véletlenszerűen választott embert iskolai végzettség és jövedelmi kategória szerint osztályokba sorolva független-e a két szempont.

Két szempont szerint soroljuk osztályokba a megfigyeléseket.

Első szempont: A_1, \dots, A_r (teljes eseményrendszer, pontosan az egyik következik be, például: iskolai végzettség szerinti kategóriák).

Második szempont: B_1, \dots, B_s (ez egy másik teljes eseményrendszer, például: jövedelmi kategóriák).

H_0 : a két szempont **független** egymástól, azaz $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(B_j)$ minden i, j -re.

H_1 : a nullhipotézis nem igaz, a két szempont között **összefüggés** van.

N_{ij} : hány olyan megfigyelés van, melyre A_i és B_j teljesül.

$N_{i\cdot} = \sum_{j=1}^s N_{ij}$ (azaz az A_i gyakorisága); $N_{\cdot j} = \sum_{i=1}^r N_{ij}$ (azaz B_j gyakorisága); n pedig az összes megfigyelés száma. Ekkor a próbastatisztika, mely H_0 mellett $f = (r-1)(s-1)$ szabadsági fokú χ^2 -eloszláshoz tart:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(N_{ij} - \frac{N_{i\cdot} \cdot N_{\cdot j}}{n})^2}{\frac{N_{i\cdot} \cdot N_{\cdot j}}{n}}.$$

Ha a függetlenség teljesül, akkor a $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(B_j)$ egyenletben a valószínűségeket a relatív gyakoriságokkal helyettesítve:

$$\frac{N_{ij}}{n} \approx \frac{N_{i\cdot}}{n} \cdot \frac{N_{\cdot j}}{n} \quad \Leftrightarrow \quad N_{ij} \approx \frac{N_{i\cdot} \cdot N_{\cdot j}}{n}.$$

Ebből adódik, hogy a χ^2 számlálója a nullhipotézistől való eltérést méri.

A szabadsági fok $f = (r-1)(s-1)$.

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

- $\chi^2 < c_{\text{krit}}$ (azaz a $p \geq \alpha$): elfogadjuk H_0 -t, **nem találtunk szignifikáns összefüggést** a szempontok között.
- $\chi^2 > c_{\text{krit}}$ (azaz a $p < \alpha$): elutasítjuk H_0 -t, az adatok **szignifikáns összefüggést** mutatnak.

Ha $r = s = 2$, a próbastatisztika az alábbi egyszerűbb alakra hozható:

$$\chi^2 = \frac{n(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\cdot} \cdot N_{\cdot 1} \cdot N_{\cdot 2} \cdot N_{2\cdot}}.$$

1.1. Függetlenségvizsgálat: példa

H_0 : a hőmérséklet és a csapadékmennyiség **független**; H_1 : a hőmérséklet és a csapadékmennyiség között **összefüggés** van.

	meleg	átlagos	hideg
esős	15	10	5
átlagos	10	10	20
száraz	5	20	5

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(N_{ij} - \frac{N_{i.} \cdot N_{.j}}{n})^2}{\frac{N_{i.} \cdot N_{.j}}{n}} = \frac{(15 - \frac{30 \cdot 30}{100})^2}{\frac{30 \cdot 30}{100}} + \frac{(10 - \frac{30 \cdot 40}{100})^2}{\frac{30 \cdot 40}{100}} + \dots + \frac{(5 - \frac{30 \cdot 30}{100})^2}{\frac{30 \cdot 30}{100}} = 22,92$$

$n = 100$, $f = (r - 1) \cdot (s - 1) = 2 \cdot 2 = 4$, $\alpha = 0,05$, $c_{\text{krit}} = 9,49$

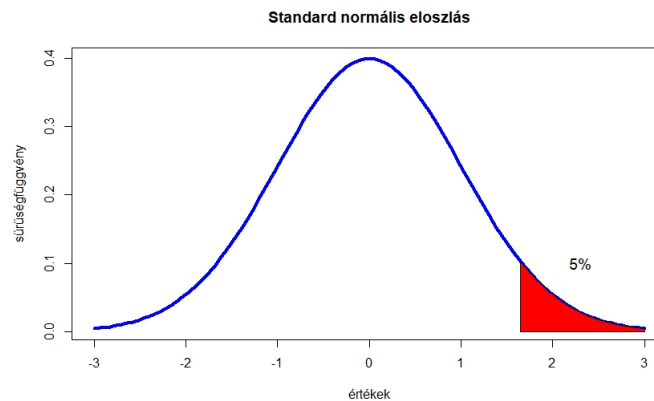
$22,917 > c_{\text{krit}} = 9,49$, illetve $p = 0,00013 < \alpha = 0,05 \Rightarrow$ elutasítjuk a nullhipotézist, szignifikáns összefüggés van a két szempont között.

1.2. Pozitív korreláció

Tekintsük a függetlenségvizsgálatot abban az esetben, ha mindkét szempont szerint két osztály van. Ekkor az is értelmes kérdés, hogy „milyen irányú” az összefüggés, például igaz-e, hogy az A_1 esemény a B_1 -gyel pozitívan korrelál, azaz egyszerre nagyobb valószínűséggel következnek be, mint amit függetlenség esetén várnánk (ez utóbbi a két valószínűség szorzata lenne). Például, egy embert véletlenszerűen kiválasztva, van-e pozitív korreláció az alábbi két esemény között: van egyetemi végzettsége, a bruttó havi jövedelme legalább 400000 forint. Ez szorosan kapcsolódik a függetlenségvizsgálathoz, de nem χ^2 -próbát, hanem z -próbát tudunk alkalmazni. Itt is minden osztályban kell, hogy legyen legalább 5 megfigyelés.

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

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



1. ábra. A standard normális eloszlás és a z -próba kritikus értéke $\alpha = 0,05$ esetén

A próbastatisztika (H_0 mellett standard normális eloszlású):

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

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 $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$. Vagyis a p -érték $\int_z^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx$. Ez annak valószínűsége, hogy a standard normális eloszlás z -nél nagyobb, vagyis az 1. ábrán a z -től jobbra lévő terület.

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.

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

Minden osztályba esik legalább 5 megfigyelés ($N_{ij} \geq 5$ minden i, j -re), alkalmazható a függetlenségvizsgálat.

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

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: $1 - \Phi(2,74) = 0,003 < 0,05$.

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

- minden osztályba essen legalább 5 megfigyelés
- a pozitív korreláció **nem jelent ok-okozati összefüggést**
- ha sok mennyiséget vizsgálunk, vagy előre el kell dönteni, 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); vagy alaposabban meg kell vizsgálni a kapott pozitív összefüggéseket (hiszen lehetnek tévesek); vagy a tévedés kockázatával számolva lehet használni a kapott összefüggéseket.

2. 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 $N_k \geq 5$ vagy $M_k \geq 5$ nem teljesül, osztályokat vonunk össze.

A próbatasztika:

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

Ha H_0 igaz, és $n \rightarrow \infty$, akkor χ^2 eloszlása az $f = r - 1$ szabadsági fokú χ^2 -eloszláshoz konvergál eloszlásban.

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.

2.1. 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

Minden osztályba esik legalább 5 megfigyelés.

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

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 városban az egy háztartásban élők számának eloszlása **nem tér el szignifikánsan**. A p -érték: $p = 0,31 > 0,05$.

3. A tapasztalati eloszlásfüggvényen alapuló próbák

Az alábbi kérdések gyakran felvetődnek, ha egy ismeretlen mennyiségnek nem csak a várható értékét és szórását szeretnénk vizsgálni (mint például a t -próba esetében), hanem az eloszlását pontosabban is:

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? 100 ember jövedelmét felmérve mikor fogadható el ez a feltételezés, és mikor állíthatjuk, hogy a jövedelem szignifikánsan eltér ettől az eloszlástól?
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? Egy $n = 96$ elemű minta hisztogramja a 2. ábrán látható, egy $n = 23$ elemű mintáé pedig a 4. ábrán.
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.

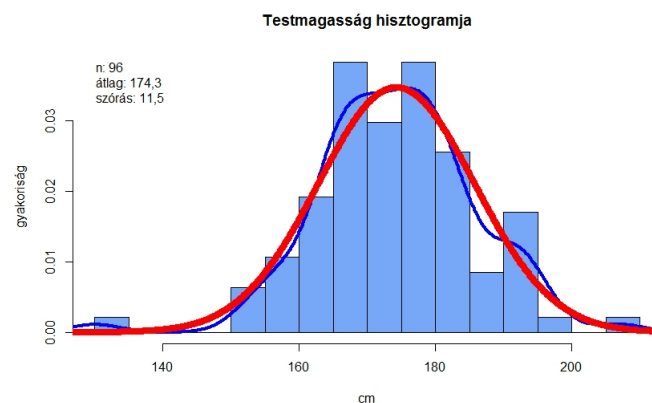
Az első és a harmadik kérdés megválaszolására egy lehetőség: **diszkretizáljuk** a megfigyeléseket, vagyis közel azonos hosszúságú intervallumokba osztjuk be őket (például jövedelmi kategóriákba), és az így kapott diszkrét eloszlásra χ^2 -próbát végzünk. Ha szükséges, a paramétereket maximumlikelihood-módszerrel becsüljük. Ebben az esetben viszont a végeredmény akár függhet is az intervallumok (kategóriák) kialakításától. Az alábbi módszerek ebből a szempontból „stabilabbak”, nincs bennük ilyen jellegű tetszőleges választás.

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)



2. ábra. 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.

3.1. Tapasztalati eloszlásfüggvény

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

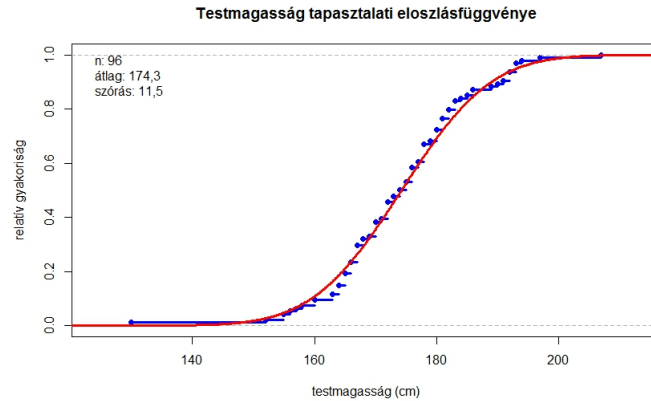
$$F(t) = \mathbb{P}(X \leq t)$$

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

3.1. Definíció (Tapasztalati eloszlásfüggvény (empirical cumulative distribution function)).

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

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



3. ábra. 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.

3.2. 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. Azt szeretnénk eldönteni az X_1, X_2, \dots, X_n minta alapján, hogy a vizsgált mennyiség (ezt jelöljük X -szel) valódi eloszlását a G függvény írja-e le, azaz igaz-e, hogy

$$F_X(t) = G(t) \quad \Leftrightarrow \quad \mathbb{P}(X \leq t) = G(t) \quad \text{minden } t\text{-re.}$$

H_0 : a minta valódi eloszlásfüggvénye G

H_1 : a minta valódi eloszlásfüggvénye G -től különböző

Az $\hat{F}_n(t)$ tapasztalati eloszlásfüggvény tehát a t -nél nem nagyobb mintaelemek aránya a megfigyelt mintában, míg $G(t)$ a nullhipotézis mellett annak valószínűsége, hogy egy mintaelem legfeljebb t . Ha tehát a nullhipotézis teljesül, akkor $\hat{F}_n(t)$ és $G(t)$ nagy valószínűséggel közel lesznek egymáshoz minden t -re, ahogy az a statisztika alaptételéből (Glivenko–Cantelli-tétel) is következik.

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

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

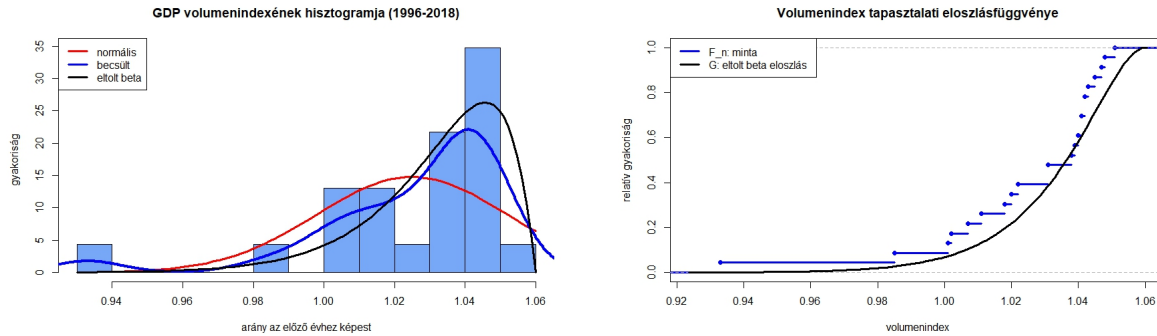
ahol F_n a minta tapasztalati eloszlásfüggvénye. Bár \hat{F}_n és G függvények (valós számokhoz rendelnek 0 és 1 közötti számokat, mint a 4. ábrán), D_n egyetlen szám lesz, hiszen az eltérés szuprémumát vettük (ez sok esetben megegyezik a maximummal, tehát a legnagyobb különbség a két függvény értéke között, abszolút értékben értve). 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):

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



4. ábra. A GDP volumenindexének (az érték osztva az előző évi értékkel) 1993-2018 közötti értékeinek histogramja, a becslt normális eloszlás és a becslt sűrűségfüggvény, illetve az eltolt Beta-eloszlás sűrűségfüggvénye (balra), illetve a tapasztalati eloszlásfüggvény és a megadott G eloszlásfüggvény (az adatok forrása: KSH)

Kolmogorov–Szmirnov-próba, példa. Tekintsük a GDP volumenindexének (az előző évi érték osztva az aktuális értékkel) adatait 1993–2018 között. 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 Beta-eloszlás az értékeit a $[0, 1]$ intervallumon veszi fel, ezért van szükség az eltolásra. A Beta-eloszlás sűrűségfüggvénye a $[0, 1]$ intervallumon belül (ezen kívül 0): $x^{a-1}(1-x)^{b-1}$, most tehát ez az $x^{69}(1-x)$ polinom. A 4. ábrán látható ennek a sűrűségfüggvénynek és a histogramnak az összehasonlítása, a 4. ábrán pedig a tapasztalati eloszlásfüggvénynek és az eltolt Beta-eloszlás G eloszlásfüggvénynek az összehasonlítása.

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.

3.3. A normalitás tesztelése: Lilliefors-próba

Speciális módszerek használhatók annak eldöntésére, hogy egy eloszlás normális eloszlású-e. Ilyenkor azonban nem egy adott eloszlásfüggvényről van szó, amivel az F_n tapasztalati eloszlásfüggvényt össze tudjuk hasonlítani, hiszen a normális eloszlás várható értéke és szórása ebben az esetben ismeretlen paraméter. Először tehát ezeket kell megbecsülni a mintából, és utána tudjuk az összehasonlítást elvégezni.

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 korigált 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):

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

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.

0,068

A 3. 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 a megadott paraméterekkel, nincs szignifikáns eltérés a normális eloszlástól.

Ugyanakkor a 4. ábrához tartozó, volumenindexre 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 $p = 0,0002$, vagyis ugyanazokra a következtetésekre juthatunk a két módszerrel.

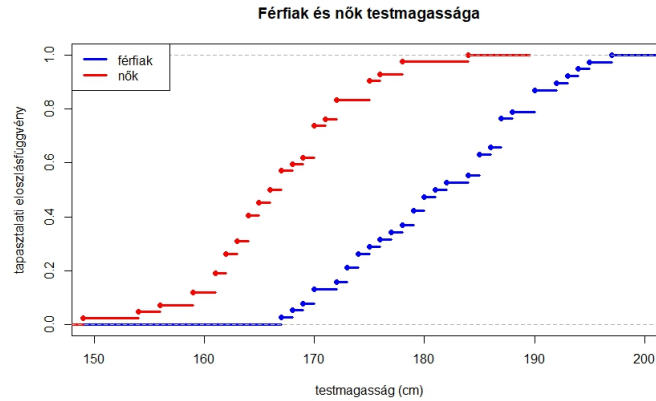
4. Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

Az is gyakran előfordul, hogy nem egy adatsor eloszlását szeretnénk egy adott eloszláshoz hasonlítani, hanem két minta homogenitását teszteljük, vagyis azt szeretnénk eldönteni, hogy a két minta eloszlása megegyező, vagy szignifikánsan eltérő. Például: megmértük néhány férfi és nő testmagasságát. A tapasztalati eloszlásfüggvények az 5. ábrán láthatók. 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.

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 $\mathbb{P}(X_j \leq t) \neq \mathbb{P}(Y_j \leq t)$.

Az egymintás esethez hasonlóan a Kolmogorov–Szmirnov-próba azon alapul, hogy ha a két eloszlás megegyezik, akkor minden t -re a t -nél nem nagyobb mintaelemek aránya hasonló, és a mintaelemek számát növelve „egyre hasonlóbb” (a nagy számok törvénye alapján), így a tapasztalati eloszlásfüggvények legnagyobb eltérésén alapul az eljárás.



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

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:

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

4.1. 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:

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

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 p -érték kisebb 0,05-nél, a nullhipotézist elutasítjuk, férfiak ($n = 38$ megfigyelés) és a nők ($m = 42$ megfigyelés) testmagasságának eloszlása szignifikánsan különböző.

4.2. Kolmogorov–Szmirnov-próba: egyoldali homogenitásvizsgálat

Azt is kérdezhetjük, hogy igaz-e, hogy a két minta közül az egyik „tipikusan” nagyobb értékeket vesz fel, mint a másik. Ezt így fogalmazhatjuk meg az eloszlásfüggvények segítségével: minden t valós számra igaz, hogy

$$\mathbb{P}(X \leq t) \leq \mathbb{P}(Y \leq t).$$

Ez azt jelenti, hogy annak valószínűsége, hogy X egy adott értéknél kisebb, kisebb, mint ugyanez Y esetében, vagyis X inkább nagyobb értékeket vesz fel, mint Y .

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.

Ilyenkor a próbastatisztika szintén a két minta tapasztalati eloszlásfüggvényéből számolható, de most az eltérés előjele is fontos (hiszen arra vagyunk kíváncsiak, hogy $\mathbb{P}(X \leq t) \leq \mathbb{P}(Y \leq t)$ teljesül-e), így abszolút érték nélkül számolunk:

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

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.

Ennek megvalósítása az előző példa esetében:

H_0 : a férfiak és a nők testmagasságának eloszlása megegyezik

H_1 : ha X egy véletlenszerűen választott férfi testmagassága, Y pedig egy véletlenszerűen választott nőé, akkor 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.

```
> 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 próbastatisztika értéke ugyanaz, mint a kétoldali esetben volt, hiszen itt pozitív volt a különbség, az abszolút értéke saját maga. A p -érték kisebb 0,05-nél, a nullhipotézist elutasítjuk. A férfiak ($n = 38$ megfigyelés) testmagassága szignifikánsan nagyobb nőkénel ($m = 42$ megfigyelés) a sztochasztikus értelemben: $F(t) = \mathbb{P}(X \leq t) \leq G(t) = \mathbb{P}(Y \leq t)$ minden t -re.

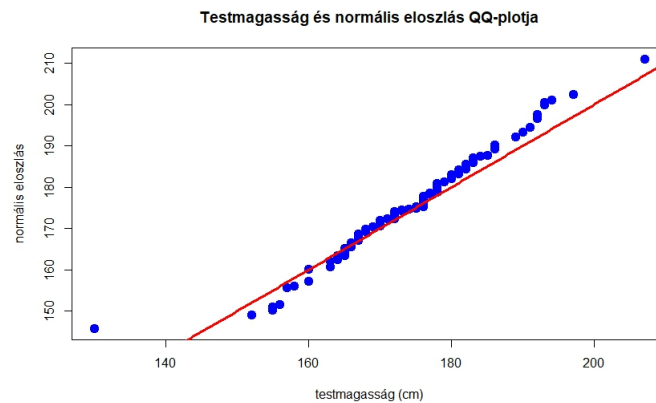
5. QQ-plot

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

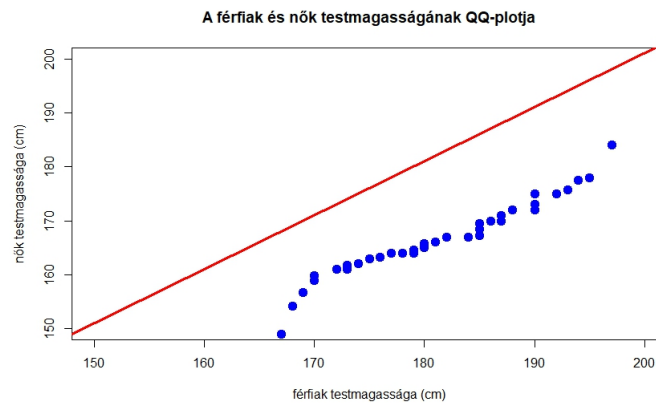
- a QQ -plot két minta eloszlásának az összehasonlítására szolgál, a kvantilisok összehasonlításával
- ha a tapasztalati z -kvantilis az első mintában q_1 , a másodikban q_2 , akkor a (q_1, q_2) pontba kerül egy pont

- minél inkább egyezik a két minta eloszlása, annál közelebb 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.

A 6. ábrán a testmagasság $n = 96$ elemű adatsorának és egy, a becült paraméterekhez tartozó normális eloszlású mintának a QQ-plotja látható. Ez közel van az $y = x$ egyeneshez, annak megfelelően, hogy el tudtuk fogadni azt a hipotézist, hogy a testmagasság normális eloszlású. A 7. ábrán ugyanebből az adatsorból a férfiakhoz, illetve nőkhez tartozó adatsorok QQ-plotja látható. Itt a pontok egyáltalán nem illeszkednek az $y = x$ egyenesre, annak megfelelően, hogy azt láttuk, hogy szignifikánsan eltérők a magasságok a két esetben.



6. ábra. 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



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

Házi feladat április 28., szerda, 9:00-ig Tekintsük az összegyűjtött mintát.

- Állíthatjuk-e $\alpha = 0,05$ szignifikanciaszint mellett, hogy a 25 évesnél fiatalabbak és idősebbek szabadban töltött idejének eloszlásfüggvénye szignifikánsan különböző?
- Az összegyűjtött mintában válasszunk ki két eseményt (például: valaki 25 évesnél fiatalabb, és legalább 3 sorozatot nézett, de lehet más is), és vizsgáljuk meg, hogy van-e köztük szignifikáns pozitív (vagy negatív) korreláció.