

Nemparaméteres próbák: illeszkedésvizsgálat és homogenitásvizsgálat

1. 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 az 1. ábrán látható, egy $n = 23$ elemű mintáé pedig a 3. 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)

1.1. Tapasztalati eloszlásfüggvény

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

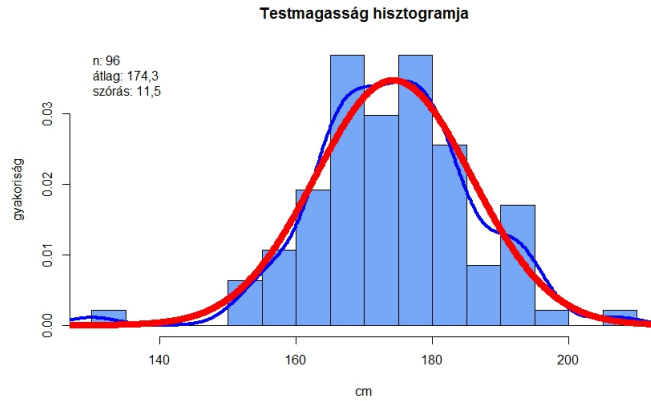
$$F(t) = \mathbb{P}(X \leq t)$$

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

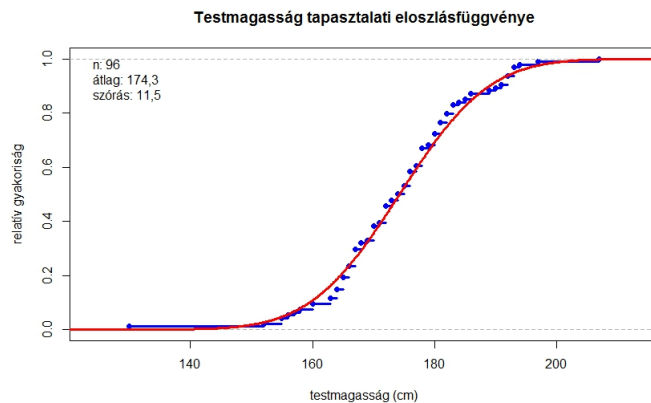
1.1. Definíció (Tapasztalati eloszlásfüggvény (empirical cumulative distribution function)).

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

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



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



2. 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.

1.2. 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

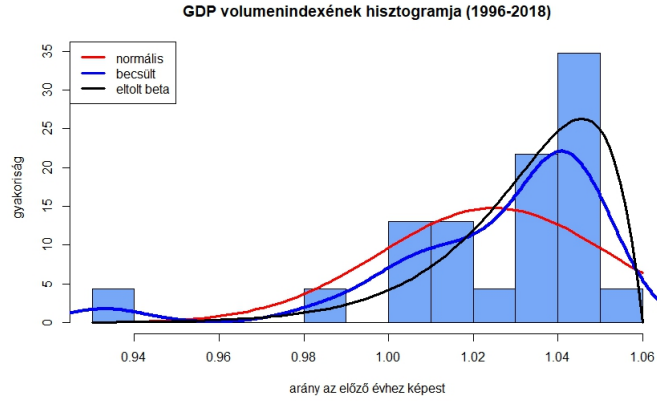
H_1 : a minta valódi eloszlásfüggvénye G -től különböző

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.

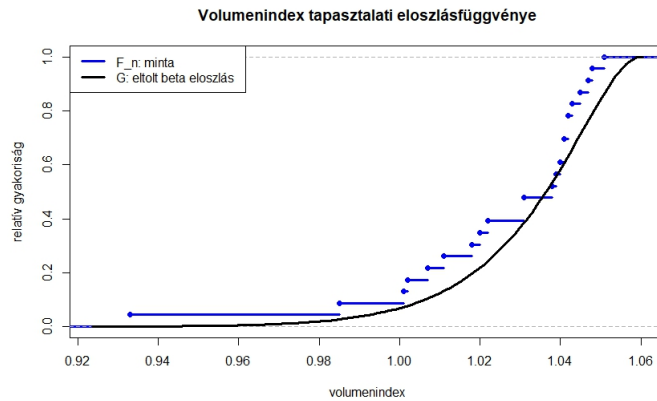
Próbastatisztika, ami a tapasztalati eloszlásfüggvény és G távolságát méri, úgy, hogy a legnagyobb különbséget veszi, abszolút értékben:

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

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



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



4. ábra. 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)

(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 $D_n > D_{krit}$ (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 $D_n < D_{krit}$, (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):

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

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 3. ábrán látható ennek a sűrűségfüggvénynek és a hisztogramnak az összehasonlítása, a 4. ábrán pedig a tapasztalati eloszlásfüggvénynek és az eltoló 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.

1.3. 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 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):

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

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 2. 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 3. ábrához tartozó, volumenindexre vonatkozó példában:

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

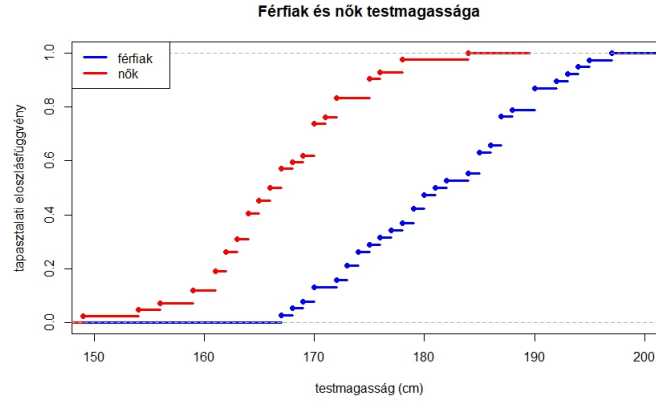
Lilliefors (Kolmogorov-Smirnov) normality test

```
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.



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

2. 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 hasonlób” (a nagy számok törvénye alapján), így a tapasztalati eloszlásfüggvények legnagyobb eltérésén alapul az eljárás.

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:

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

2.1. 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 próbatasztika 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.

2.2. 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 próbatasztika 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

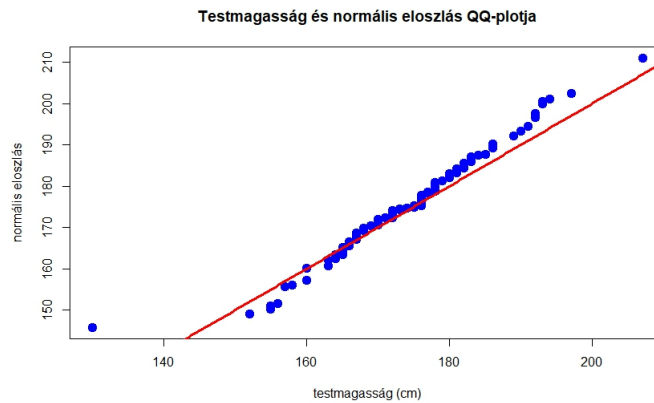
A próbatasztika értéke ugyanaz, mint a kétoldali esetben volt, hiszen itt pozitív volt a különbség, az abszolút értéke saját maga. A p -érték kisebb 0,05-nél, a nullhipotézist elutasítjuk. A férfiak ($n = 38$ megfigyelés) testmagassága szignifikánsan nagyobb nőkéénél ($m = 42$ megfigyelés) a sztochasztikus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

3. QQ-plot

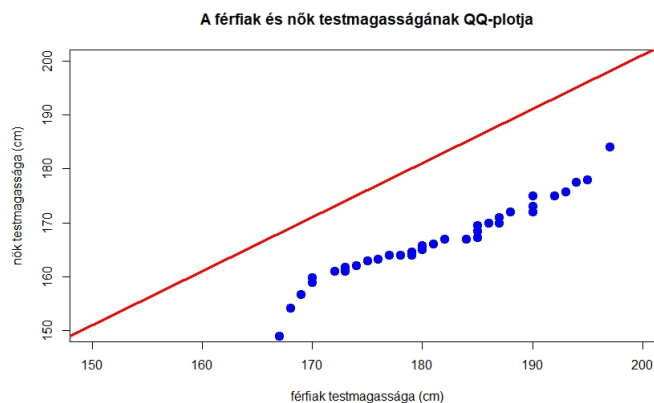
Annak vizsgálatára, hogy két minta ugyanabból az eloszlásból származik-e (homogenitásvizsgálat) a leíró statisztikában a QQ-plot is gyakran használt ábrázolási mód.

- 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 eloszlásfüggvények, 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

4. Előjelpróba

Azt, hogy két mennyiség közül tipikusan melyik a nagyobb (erről volt szó a 2.2. szakaszban az egyoldali Kolmogorov–Szmirnov-próbánál), úgy is vizsgálhatjuk, hogy nem azt kérdezzük, hogy egy adott t értéknél milyen valószínűséggel kisebbek, hanem hogy egymáshoz hasonlítjuk őket, és azt kérdezzük, hogy annak valószínűsége, hogy az első nagyobb lesz, mint a második, ugyanaz-e, mint annak valószínűsége, hogy a második nagyobb, mint az első.

Ha például kétszer megismételjük ugyanazt a kísérletet, függetlenül, ugyanolyan körülmények között, akkor ez a két valószínűség a szimmetria miatt ugyanaz, a két kísérlet sorrendje valójában felcserélhető. Ha viszont például X egy véletlenszerűen választott fővárosi munkavállaló jövedelme, míg Y egy véletlenszerűen választott vidéki munkavállaló jövedelme, akkor $\mathbb{P}(X > Y) > \mathbb{P}(X < Y)$, hiszen bármelyik eset előfordulhat, de (feltételezve, hogy a fővárosban tipikusan nagyobbak a jövedelmek), az első eset valószínűbb. Vagy például egy tóparti szálloda forgalma legyen májusban X , júniusban Y . Ekkor $\mathbb{P}(X > Y) < \mathbb{P}(X < Y)$, hiszen feltehetően a júniusi forgalom nagyobb. Ráadásul ez a kérdés akkor is értelmes, ha X és Y nem függetlenek egymástól.

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 nullhipotézis:

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

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

Legyen W az olyan párok száma, amikre $Y_i > X_i$. A H_0 nullhipotézis teljesülése esetén ez binomiális eloszlású, n renddel és $p = 0,5$ paraméterrel, hiszen minden pár esetében egymástól függetlenül $0,5$ valószínűséggel teljesül az egyenlőtlenség. (Mivel az eloszlások folytonosak, annak valószínűsége, hogy $X = Y$, nulla lesz, ezzel nem kell számolnunk.) Mivel a binomiális eloszlás független azonos eloszlású indikátorok összege, érvényes rá a centrális határeloszlástétel, így a nullhipotézis esetén W -ből a várható értékét ($n/2$ -t) levonva, majd a szórásával ($\sqrt{np(1-p)} = \sqrt{n/4}$) osztva a kapott valószínűségi változó a standard normális eloszláshoz tart. Ezt közelítésként használva feltételezzük, hogy az alábbi z mennyiség eloszlása közelítőleg standard normális eloszlás, ha n elég nagy, és a standard normális eloszlás kvantiliseit, vagyis a z -próba kritikus értékeit használjuk a próbához.

A próbastatisztika legyen tehát

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

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 $1 - \Phi(z)$.

5. Wilcoxon-próba

Az előző hipotézisvizsgálati feladatban egy másik eljárást is gyakran használnak.

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), folytonos eloszlásúak.

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

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

- 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:

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

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

- Ha a mintaelemszám elég nagy, a

$$z = \frac{W}{\sqrt{\frac{k(k+1)(2k+1)}{6}}}$$

mennyiségre kétoldali z -próbát alkalmazhatunk, a kritikus érték ebben az esetben $1 - \Phi^{-1}(1 - \alpha/2)$, ahol α a szignifikanciaszint.

- Itt is lehet egyoldali ellenhipotézist is vizsgálni, akkor az egyoldali Wilcoxon-próba kritikus értékére van szükség, illetve a közelítő esetben az egyoldali z -próbának megfelelően járhatunk el.

Wilcoxon-próba, példa. Hat tóparti szálloda májusi és júniusi bevételét mutatja az alábbi táblázat (millió forintban). Az egyes szállodák bevételét egymástól függetlennek tekintjük, de a májusi és a júniusi érték egy szálloda esetében összefügghet. Nincsenek egyenlő értékek a párokon belül, így nem kell elhagyni mintaelemeket.

Kétoldali ellenhipotézist vizsgálunk. Legyen X a májusi, Y a júniusi bevétel:

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

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

szálloda	A	B	C	D	E	F
májusi bevétel (X_j)	20,3	19,3	16,5	22,4	23,8	18,5
júniusi bevétel (Y_j)	25,2	22,9	14,3	26,3	21,7	22,1
a különbség abszolút értéke ($ X_j - Y_j $)	4,9	3,6	2,2	3,9	2,1	3,6
rang (R_j)	6	3,5	2	5	1	3,5
a különbség előjele ($\operatorname{sgn}(Y_j - X_j)$)	+1	+1	-1	+1	-1	+1

Ezután:

$$W = \sum_{j=1}^k \operatorname{sgn}(Y_j - X_j) \cdot R_j = 6 + 3,5 - 2 + 5 - 1 + 3,5 = 15.$$

Bár most a mintaelemszám nem elég nagy, a példa kedvéért a közelítő összeget használva $k = 6$ -tal (hiszen hat pár van):

$$z = \frac{W}{\sqrt{\frac{k(k+1)(2k+1)}{6}}} = \frac{15}{\sqrt{6 \cdot 7 \cdot 136}} = 1,57.$$

A kétoldali z -próba esetén az $\alpha = 0,05$ -höz tartozó kritikus érték: $1 - \Phi^{-1}(0,025) = 1,96$. Mivel tehát $|z|$ kisebb a kritikus értéknél, a nullhipotézist elfogadjuk, annak valószínűsége, hogy a májusi bevétel nagyobb a júniusinál, nem tér el szignifikánsan annak valószínűségétől, hogy a júniusi szignifikánsan nagyobb a májusinál.

Az előjelpróbával az (1) egyenlet alapján ($n = 6$ a párok száma):

$$z = \frac{W - n/2}{\sqrt{n/4}} = \frac{4 - 3}{\sqrt{6/4}} = 0,817$$

adódik, hiszen ott W az olyan párok száma, ahol $Y_j > X_j$. Erre is z -próbát végezhetünk, a kritikus érték most is $1,96$, ezzel az eljárással is elfogadjuk a nullhipotézist.

Házi feladat április 22., 8:15-ig. Sorsoljunk két független, 100 elemű standard normális eloszlású mintát, legyen X_1, X_2, \dots, X_{100} az egyik minta, Z_1, \dots, Z_{100} a másik minta. Legyen továbbá $Y_j = X_j + a|Z_j|$, ahol $a \geq 0$ később megválasztandó valós szám.

Keressünk olyan a számot, amire elfogadható az $(X_1, Y_1), (X_2, Y_2), \dots, (X_{100}, Y_{100})$ minta alapján, hogy $\mathbb{P}(X_j > Y_j) = \mathbb{P}(X_j < Y_j)$, és keressünk olyan a számot is $\alpha = 0,05$ mellett, amire nem fogadható el ez a nullhipotézis.

Milyen $a \geq 0$ számokra teljesül az $\mathbb{P}(X_j > Y_j) = \mathbb{P}(X_j < Y_j)$ egyenlőség?

Házi feladat április 15., 8:15-ig, megoldás. A házi feladathoz begyűjtött adatsor alapján végezzük el az alábbi hipotézisvizsgálati feladatokat.

- Elfogadható-e $\alpha = 0,05$ szignifikanciaszint mellett, hogy az egy ember által az elmúlt egy hónapban nézett sorozatok száma Poisson-eloszlású?
- Számítsuk ki a nézett sorozatok számának mediánját a teljes adatsorból, legyen ez m . Állíthatjuk-e $\alpha = 0,05$ valószínűséggel, hogy aközött, hogy egy véletlenszerűen kiválasztott ember nő, illetve hogy legalább m sorozatot nézett, szignifikáns összefüggés van?

Az adatsor 41 nő és 32 férfi válaszait tartalmazta, így $n = 73$. Az adatsor átlaga lesz a λ paraméter maximumlikelihood-becslése: $\hat{\lambda} = 2,23$. A legalább 4-es értékeket összevonva $r = 5$ osztály alakult ki. A \hat{p}_k kiszámítása: `dpois(k, lambda = 2.23)`, illetve az utolsót úgy választjuk, hogy 1 legyen az összeg.

sorozatok száma	0	1	2	3	≥ 4
emberek száma	19	16	15	7	16
\hat{p}_k	0,11	0,24	0,27	0,2	0,18
$n\hat{p}_k$ (Poisson($\hat{\lambda}$))	8,03	17,52	19,71	14,6	13,14

```
> pp<-c(0.11, 0.24, 0.27, 0.2, 0.18)
```

```
> adat <-c(19,16,15,7,16)
```

```
> chisq.test(adat, p=pp)
```

Chi-squared test for given probabilities

data: adat

X-squared = 20.8225, df = 4, p-value = 0.0003434

Ebből a χ^2 próbastatisztika értéke megkapható: 20,82. Azonban mivel egy paramétert becsültünk, valójában $f = 5 - 1 - 1$ szabadsági fokú próbát kell végeznünk. Ennél a kritikus érték $\alpha = 0,05$ szignifikanciaszint mellett:

```
> qchisq(0.95, df=3)
```

```
[1] 7.814728
```

Mivel $\chi^2 > c_{\text{krit}}$, elutasítjuk a nullhipotézist, az eloszlás szignifikánsan eltér a Poisson-eloszlástól. A p -értéket is kiszámíthatjuk:

```
> 1-pchisq(20.82, df=3)
```

```
[1] 0.0001147373
```

Ez is azt mutatja, hogy szignifikáns eltérés van a Poisson-eloszlástól.

Az adatsor mediánja: $m = 2$. Az alábbi táblázatot készíthetjük el:

sorozatok száma	≤ 1	≥ 2	összesen
nő	19	22	41
férfi	16	16	32
összesen	35	39	73

```
> A=matrix(c(19, 22, 16, 16), ncol=2)
```

```
> chisq.test(A)
```

Pearson's Chi-squared test with Yates' continuity correction

data: A

X-squared = 0.0055, df = 1, p-value = 0.9407

Mivel a p -érték több 0,05-nél, nem állíthatjuk, szignifikáns összefüggés van a legalább 2 sorozat nézése és aközött, hogy az adott ismerős nő.