

MATEMATIKA ÉS STATISZTIKA DATA SCIENCE-HEZ



**STATISZTIKAI TESZTEK
EREDMÉNYEINEK
ÉRTELMEZÉSE**

HIPOTÉZISVIZSGÁLAT LÉPÉSEI



- 1) Megfelelő **statisztikai teszt** kiválasztása
- 2) **Nullhipotézis** és **alternatív hipotézis** megfogalmazása
- 3) **Teszt statisztika** / **együttható** és **p-érték** kiszámítása
- 4) **Döntés** — Elutasítjuk / Nem utasítjuk el a nullhipotézist

STATISZTIKAI TESZTEK



MI AZ ALAP SZITUÁCIÓ?

1 MINTÁT VIZSGÁLOK

normál eloszlású adatok esetén

- ▶ **T-TEST**, ha kicsi a minta ($n \leq 30$)
- ▶ **Z-TEST**, ha nagy a minta ($n > 30$)

nem normál eloszlású adatok esetén

- ▶ **WILCOXON SIGNED-RANK TEST**

2 FÜGGETLEN MINTÁT VIZSGÁLOK

normál eloszlású adatok esetén

- ▶ **INDEPENDENT T-TEST**
(Student's or Welch's t-test)

nem normál eloszlású adatok esetén

- ▶ **MANN-WHITNEY U-TEST**

3 VAGY TÖBB MINTÁT VIZSGÁLOK

normál eloszlású adatok esetén

- ▶ **ONE-WAY ANOVA TEST**
(post hoc **Tukey-test**)

nem normál eloszlású adatok esetén

- ▶ **KRUSKAL-WALLIS TEST**
(post hoc **Dunn-test**)

NORMALITÁS VIZSGÁLAT

- ▶ **Shapiro-Wilk test** — kis minták normalitásának ellenőrzésére ($n < 500$)
- ▶ **Kolmogorov-Smirnov test** — nagyobb mintáknál, eloszlás összevetésére
- ▶ **D'Agostino-Pearson test** — ferdeség és csúcsosság összevetésére

KORRELÁCIÓ VIZSGÁLAT

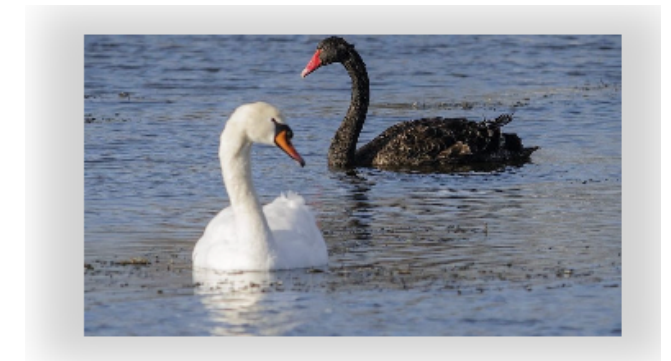
- ▶ **Pearson coeff** — folytonos, normál eloszlású változók
- ▶ **Spearman coeff** — legalább egy ordinális változó és nem kell normál eloszlás
- ▶ **Khi-square test vagy Cramer's V** — kategórikus változók

NORMALITÁS — SHAPIRO-WILK TESZT



H₀ = az adatok normális eloszlásúak (μ és σ ismeretlen)

H₁ = az adatok nem normális eloszlásúak



```
from scipy import stats

# Shapiro-Wilk (for small to medium samples)
W, p = stats.shapiro(df['variable'])
print('Shapiro-Wilk:', W, p)
```

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

ahol:

- $x_{(i)}$ = a minta rendezett értékei (növekvő sorrendben)
- \bar{x} = a minta átlaga
- a_i = súlyok, amelyeket a normál eloszlás elvárt kvantiliseiből származtatunk

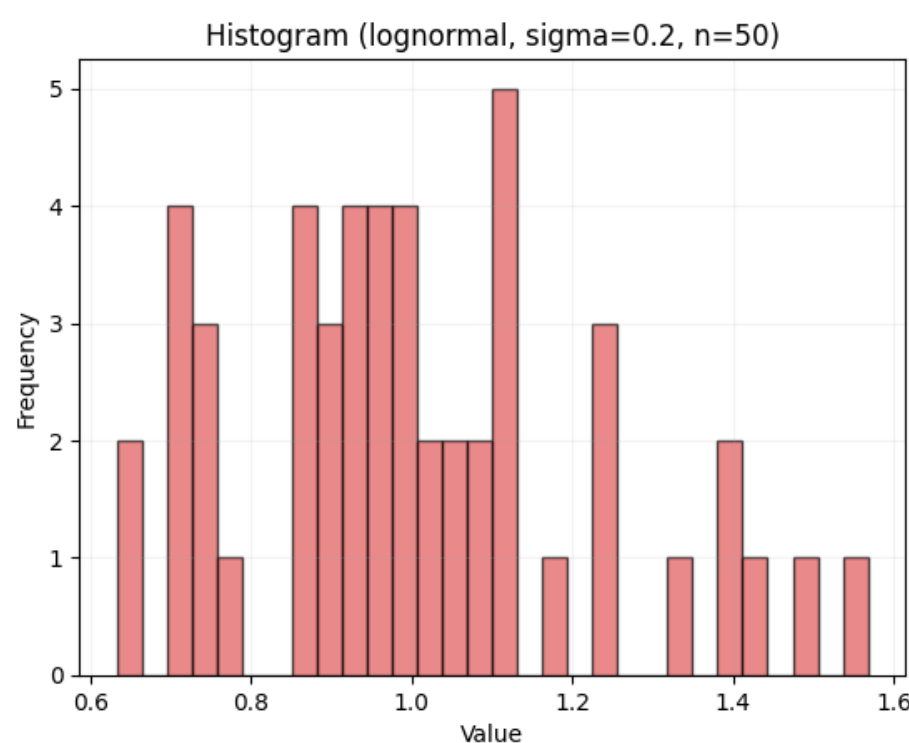
- ▶ $0 < \mathbf{W} \leq 1$ — minél közelebb van 1-hez, annál inkább normális az eloszlás
- ▶ $\mathbf{p} < 0.05 \Rightarrow$ elutasítjuk a **H₀**-t \Rightarrow az adatok nem normális eloszlásúak
- ▶ $0.05 \leq \mathbf{p} \Rightarrow$ most nem utasítjuk el a **H₀**-t

Ez nem bizonyítja a normalitást — csak azt jelzi, hogy nincs elég bizonyíték a normalitás elutasítására (különösen kis mintáknál).

NORMALITÁS — SHAPIRO-WILK TESZT



Miért soroljuk a sample-size biased tesztek közé?



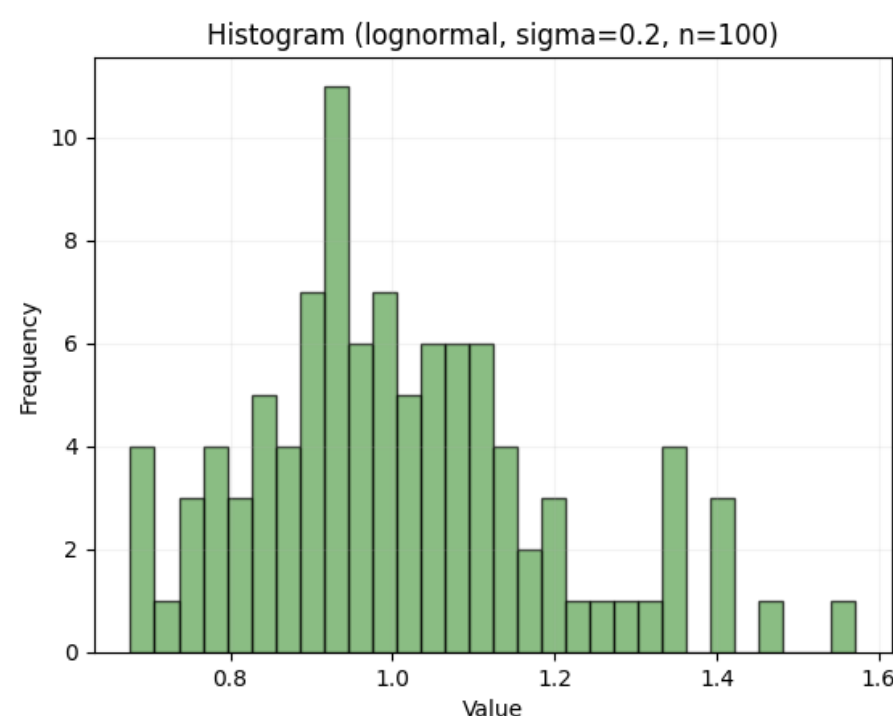
n = 50

$W = 0.96$ és $p = 0.11$



Nem utasítjuk el H_0 -t

még nincs elég ereje a tesztnek
kimutatni a normalitástól való
enyhe eltérést



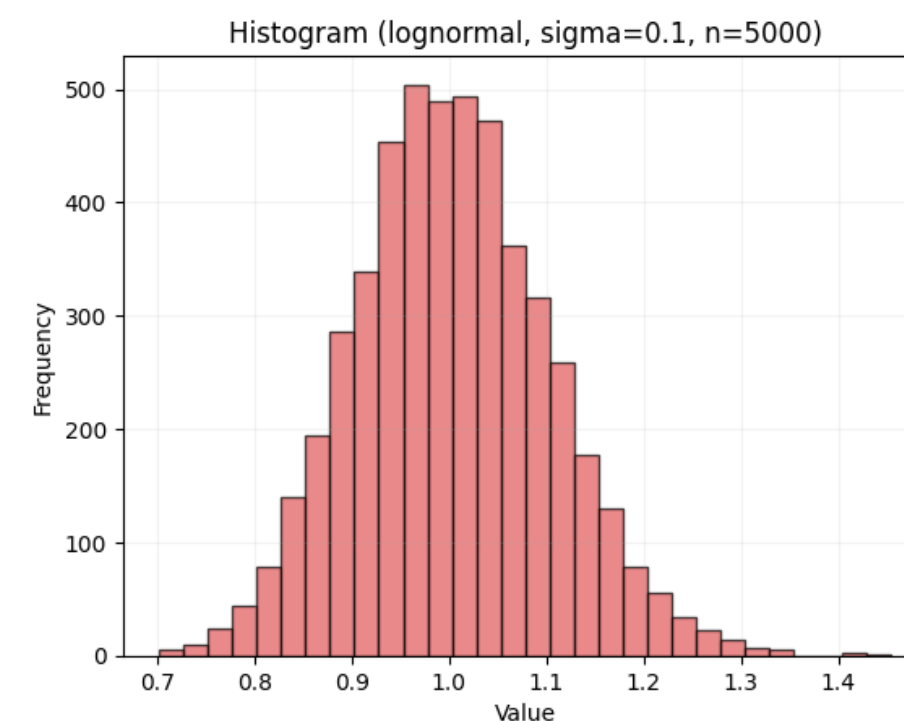
n = 100

$W = 0.97$ és $p = 0.011$



Elutasítjuk H_0 -t

szignifikáns bizonyítékunk van rá,
hogy nem normális eloszlásúak
az adatok



n = 5000

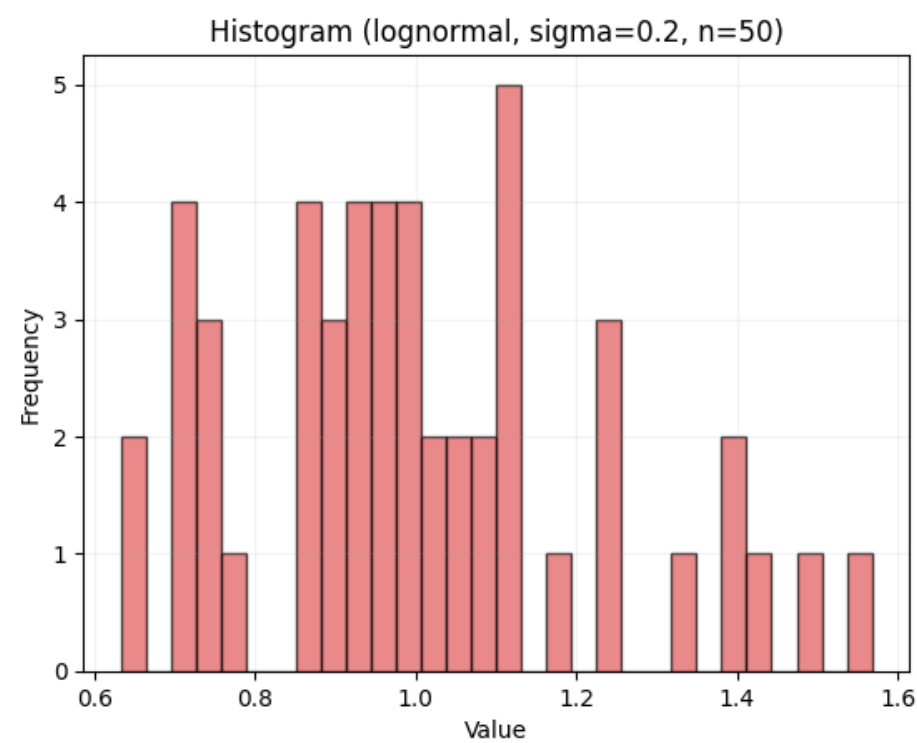
$W \approx 1$ és $p = 6.817e-10$



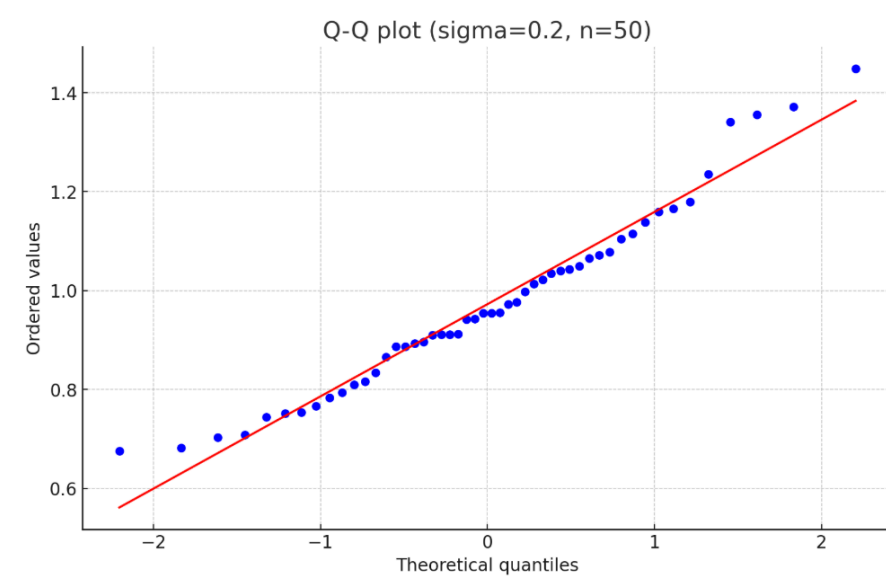
Elutasítjuk H_0 -t

pedig szinte normál eloszlást látunk
a hisztogramon — a teszt már a
nagyon kicsi eltérést is bejelzi

NORMALITÁS — SHAPIRO-WILK TESZT

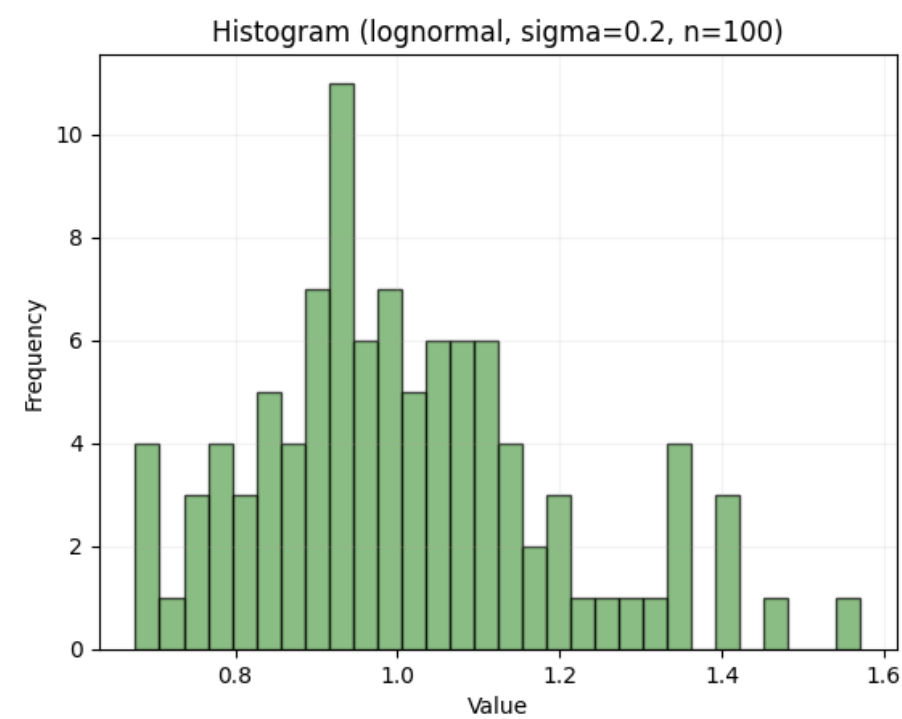


$W = 0.96$ és $p = 0.11$

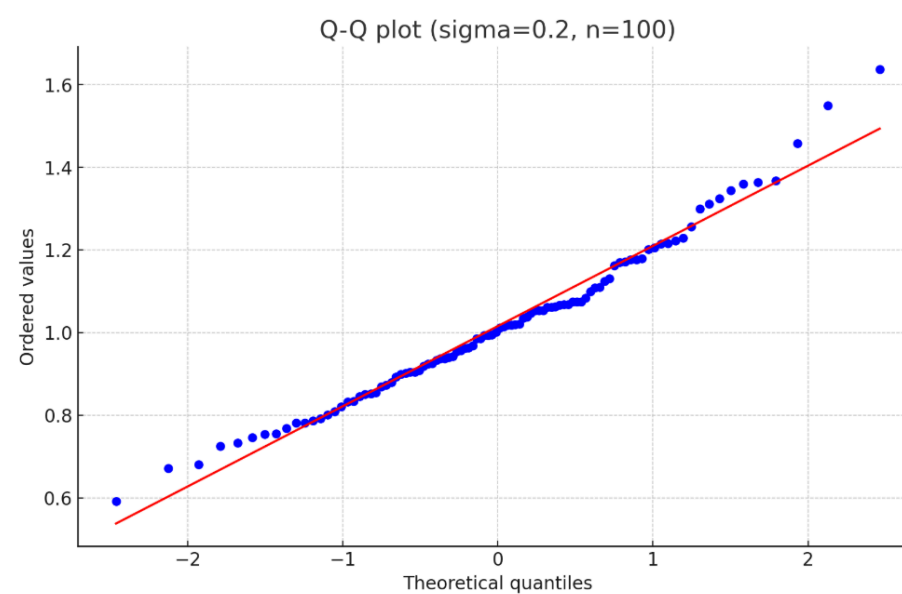


skewness = 0.571

kurtosis = -0.021

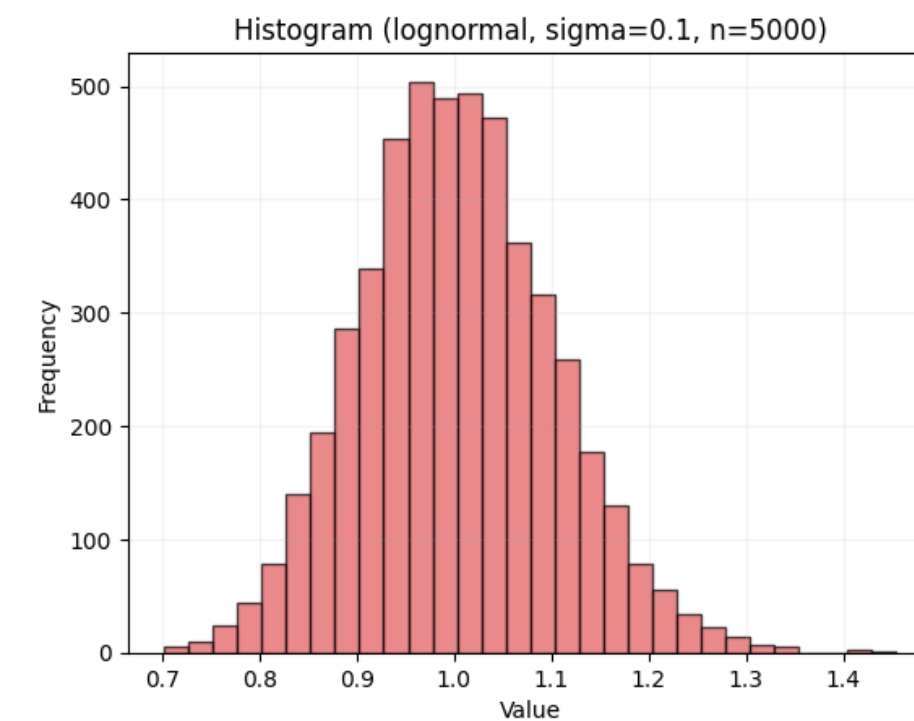


$W = 0.97$ és $p = 0.011$

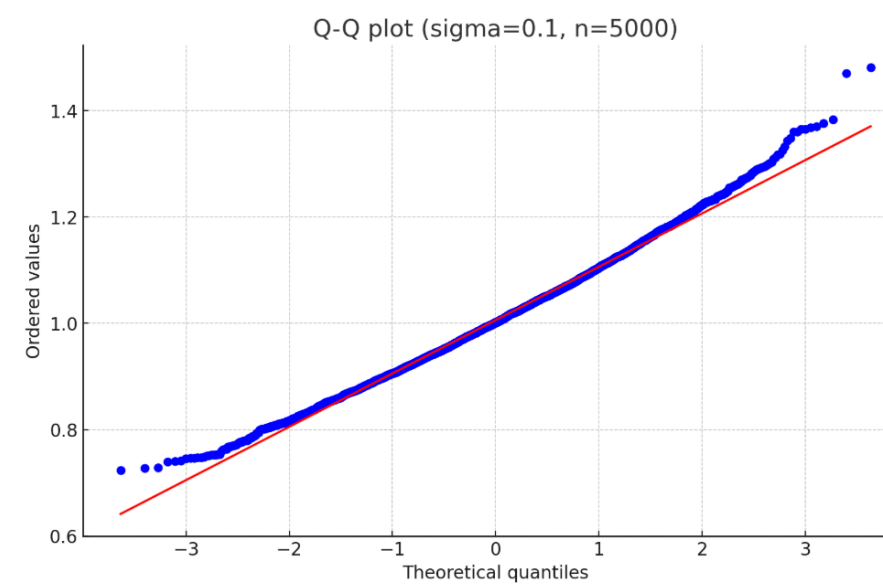


skewness = 0.577

kurtosis = 0.502



$W \approx 1$ és $p = 6.817e-10$



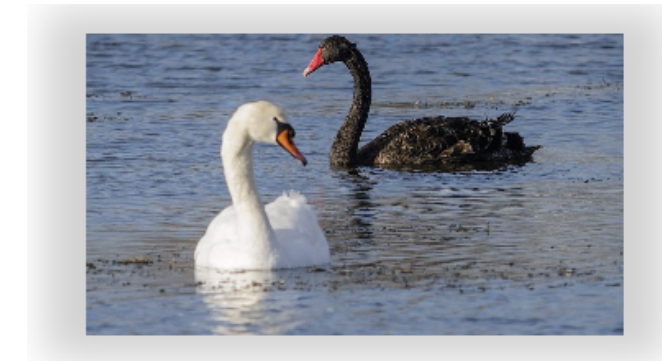
skewness = 0.300

kurtosis = 0.248

NORMALITÁS — KOLMOGOROV-SMIRNOV TESZT

H_0 = az adatok egy megadott eloszlásból származnak
(pl. normális eloszlás **ismert** μ és σ paraméterekkel)

H_1 = az adatok nem a megadott eloszlásból származnak



```
from scipy import stats
import numpy as np

data = df["variable"]

mu0 = 0      # elméleti átlag (pl. 0)
sigma0 = 1   # elméleti szórás (pl. 1)

D, p = stats.kstest(data, 'norm', args=(mu0, sigma0))
print("K-S:", D, p)
```

$$D = \sup_x |F_n(x) - F_0(x)|$$

ahol:

- $F_n(x)$ = az empirikus (megfigyelt) eloszlásfüggvény
- $F_0(x)$ = a referencia (pl. normális) eloszlásfüggvény
- \sup_x = a legnagyobb abszolút eltérés az egész tartományon

- ▶ $D \in [0, 1]$ — minél kisebb, annál jobb az illeszkedés
- ▶ $p < 0.05 \Rightarrow$ elutasítjuk a H_0 -t \Rightarrow az adatok nem illeszkednek a megadott eloszláshoz
- ▶ $0.05 \leq p \Rightarrow$ most nem utasítjuk el a H_0 -t

Ez nem bizonyítja az illeszkedést — csak azt jelzi, hogy nincs elég bizonyíték az eltérésre.

NORMALITÁS — K-S TESZT LILLIEFORS VARIÁNS

Hogy segít nekünk a Lilliefors variáns?

H₀ = az adatok normális eloszlásból származnak
(μ és σ ismeretlen, az **adatokból becsüljük őket**)

H₁ = az adatok nem normális eloszlásból származnak

```
from statsmodels.stats.diagnostic import lilliefors

D, p = lilliefors(data, dist='norm')
print("Lilliefors:", D, p)
```

$$D = \sup_x |F_n(x) - F_0(x)|$$

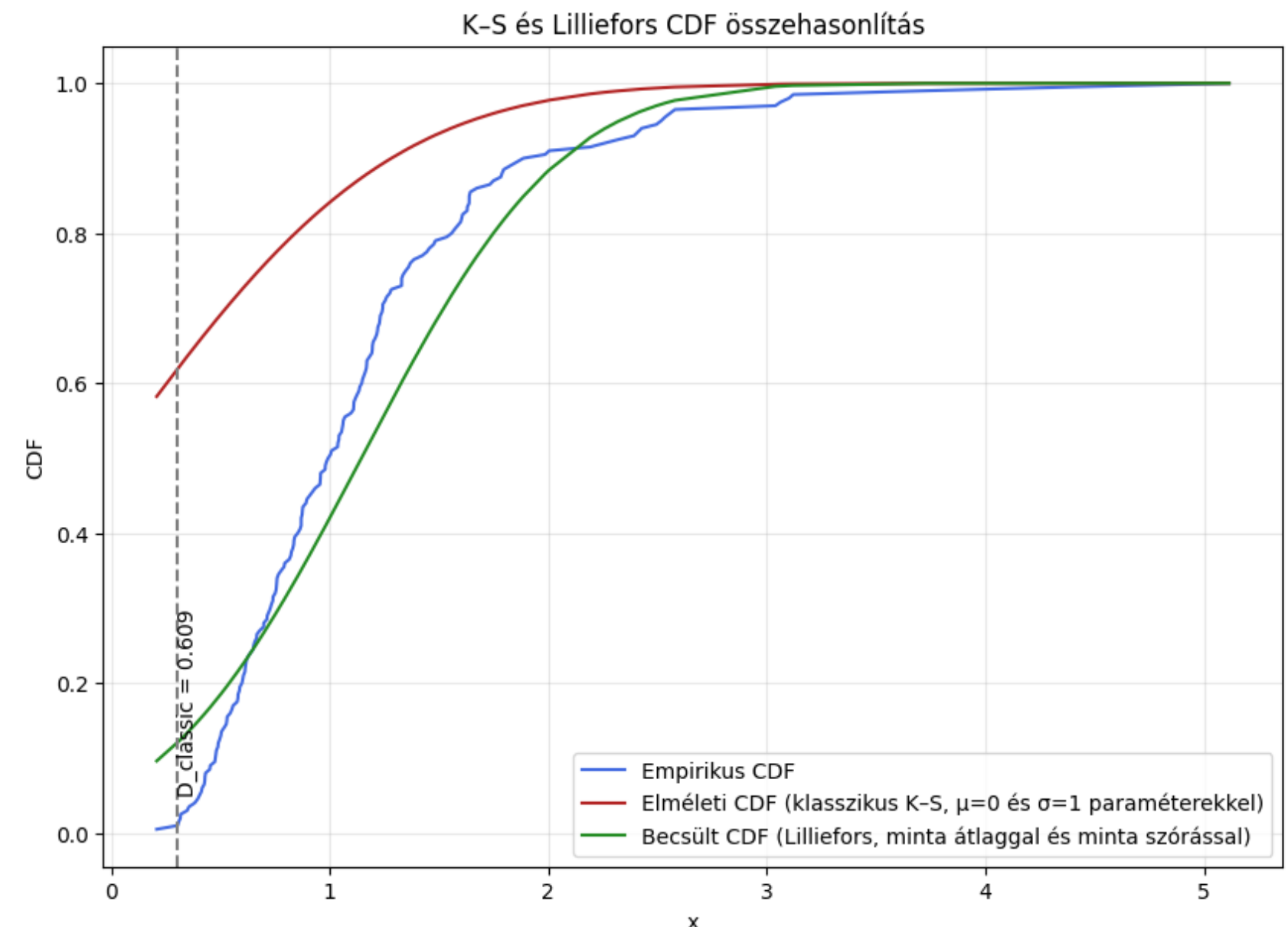
ahol:

- $F_n(x)$ = az empirikus (megfigyelt) eloszlásfüggvény
- $F_0(x)$ = a referencia (pl. normális) eloszlásfüggvény
- \sup_x = a legnagyobb abszolút eltérés az egész tartományon

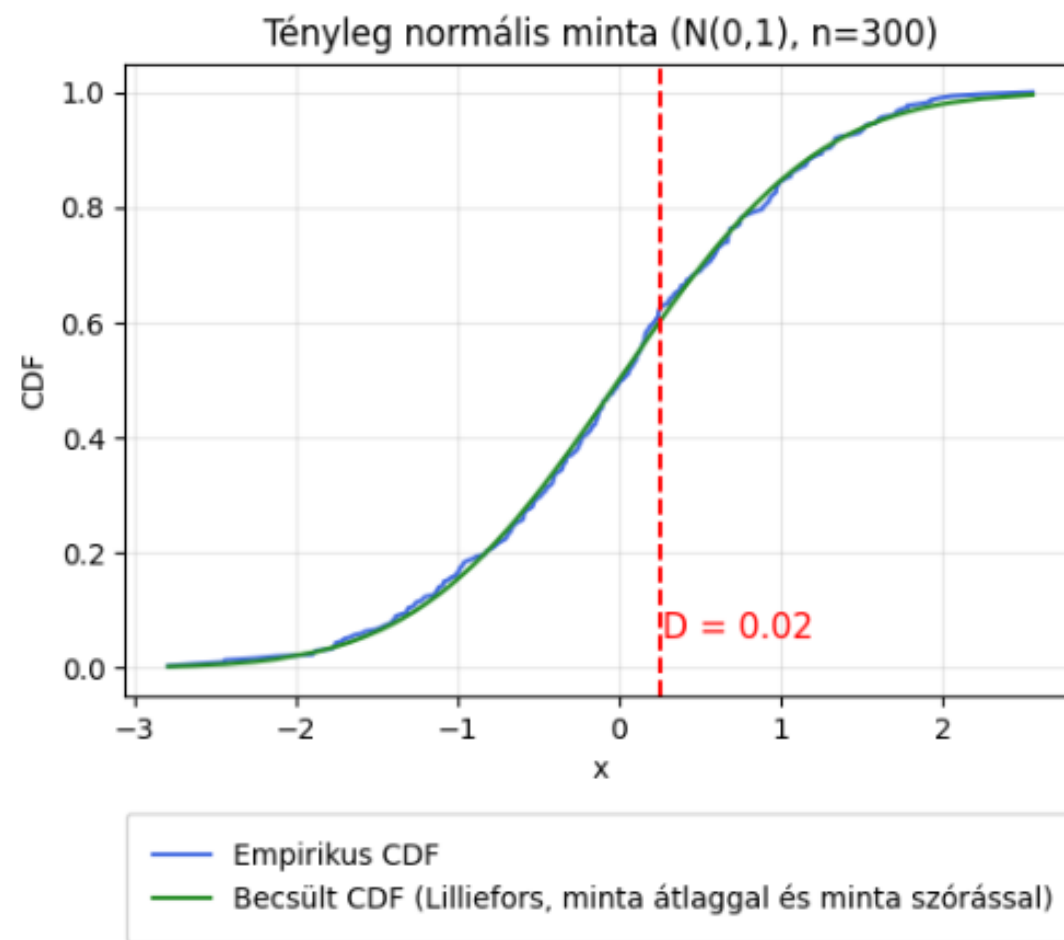
$$F_0(x) = \Phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)$$

- $\hat{\mu}$: mintaátlag
- $\hat{\sigma}$: minta szórása

► **D** \in [0, 1] — minél kisebb, annál jobb az illeszkedés



NORMALITÁS – K-S TESZT LILLIEFORS VARIÁNS

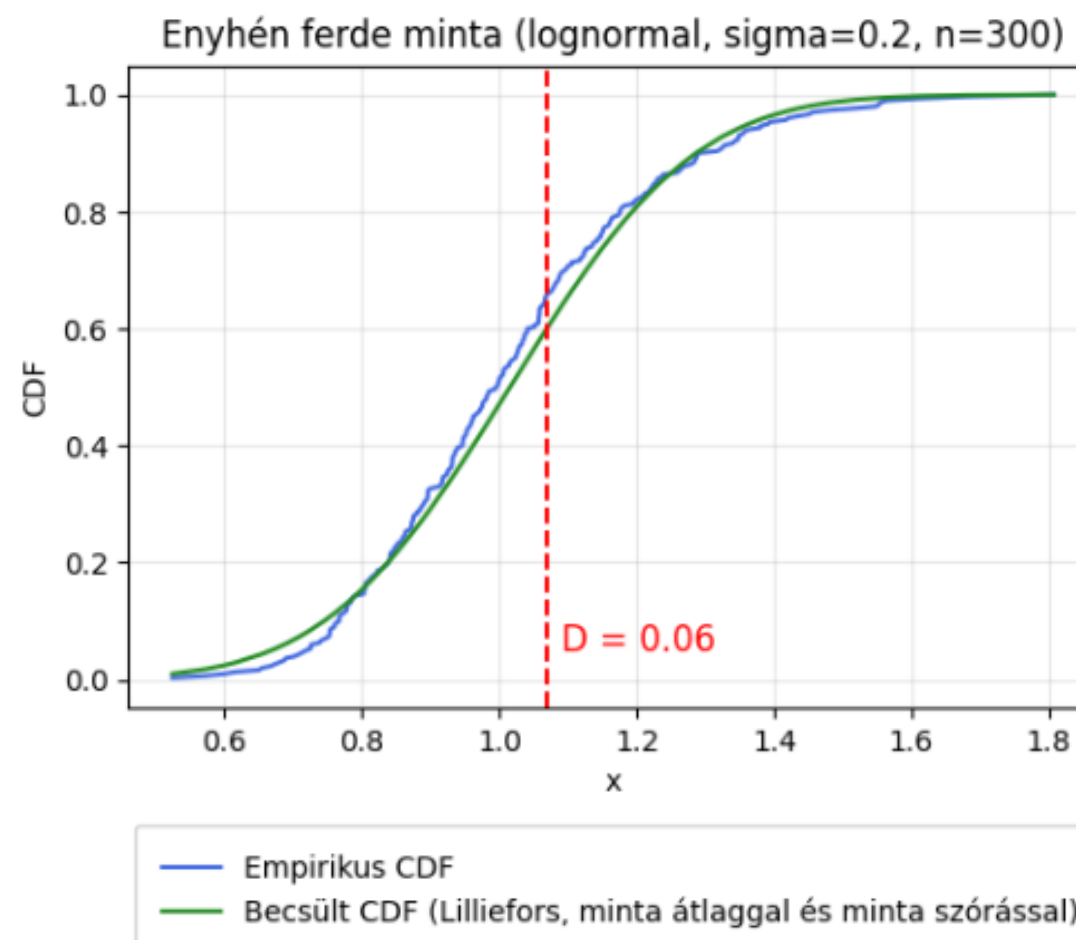


$D = 0.02$ és $p = 0.96$



Nem utasítjuk el H_0 -t

Normálisnak tűnő mintával
van dolgunk

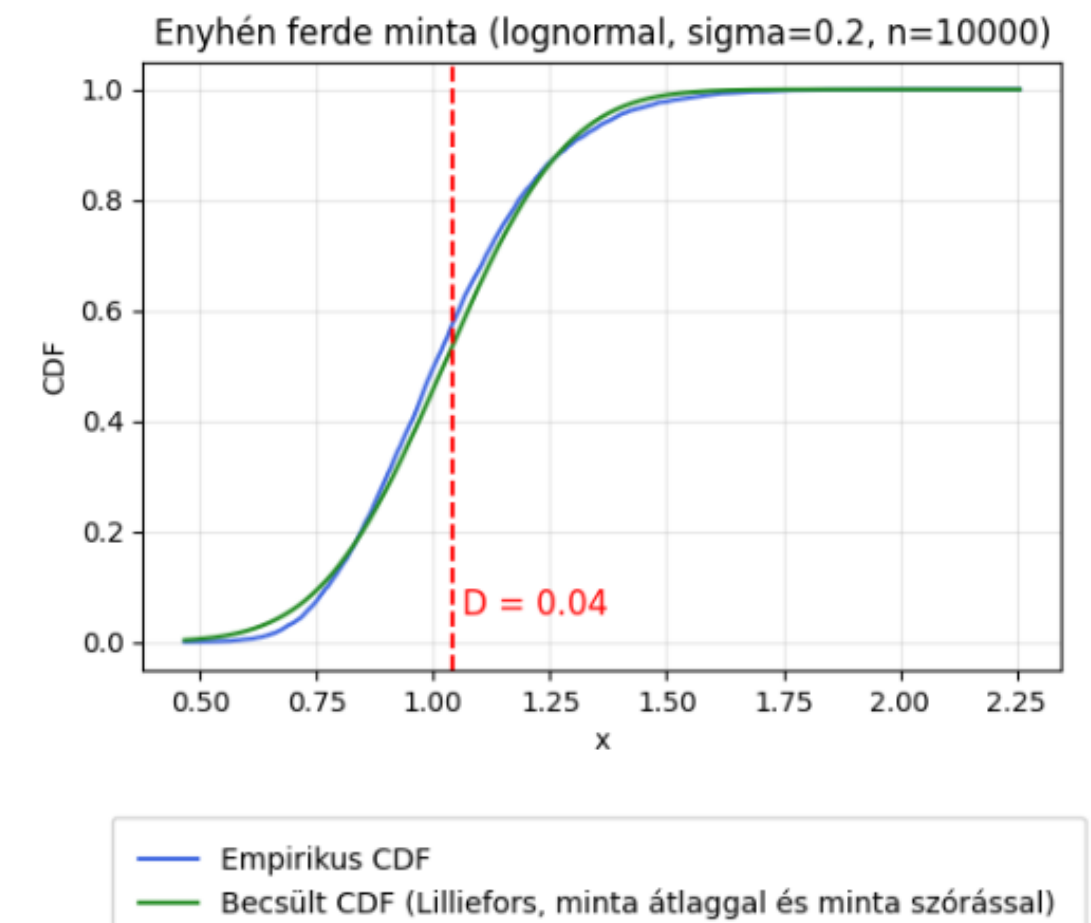


$D = 0.06$ és $p = 0.03$



Elutasítjuk H_0 -t

szignifikáns bizonyítékunk van rá,
hogy nem normális eloszlású a minta



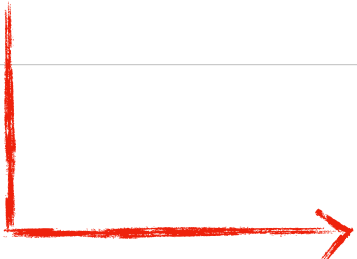
$D = 0.04$ és $p = 0.001$

nagy elemszámnál ($n=10\ 000$)
sem zuhan be olyan hirtelen
a p érték, mint a S-W tesztnél

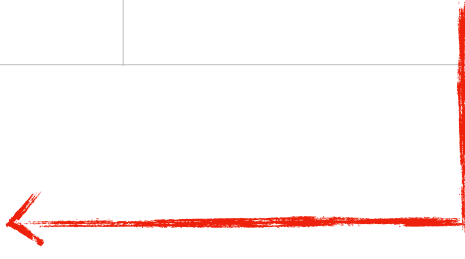
NORMALITÁS TESZTEK — ÖSSZEFOGLALÁS



	Shapiro-Wilk teszt	Kolmogorov-Smirnov teszt	K-S teszt Lilliefors variáns
Mit tesztel?	Az adatok normális eloszlásúak-e ismeretlen μ , σ mellett.	Illeszkedés egy megadott eloszláshoz (előre megadott μ és σ mellett).	Illeszkedés egy megadott eloszláshoz (mintából becsült μ és σ mellett).
Mit mér?	A rendezett mintapontok és az elméleti normál kvantilisek kapcsolatát (kvantilis-korreláció).	A legnagyobb eltérést a minta CDF és az elméleti CDF között.	Ugyanúgy a legnagyobb CDF-eltérést, csak a mintára illesztett normál eloszlás CDF-jétől.
Erősségek	<ul style="list-style-type: none">▸ Kis és közepes mintán legtöbb power▸ Jól észleli a ferdeséget, kurtózist▸ Általános célú normalitás-teszt	<ul style="list-style-type: none">▸ Bármilyen eloszlásra alkalmazható (nem csak normálra!)▸ Egyszerű, gyors, jól értelmezhető	<ul style="list-style-type: none">▸ p értéket reálisabban adja meg▸ Jó kompromisszum S-W vs K-S között▸ Robusztus tail-problémák esetén
Gyengeségek	<ul style="list-style-type: none">▸ Nagy mintán túlérzékeny	<ul style="list-style-type: none">▸ Nagyon ritkán akarunk az elméleti normál eloszláshoz ($\mu=0$ és $\sigma=1$) illeszteni	<ul style="list-style-type: none">▸ Kis mintán gyenge▸ Nagy mintán túlérzékeny



VIZUÁLIS DIAGNOSZTIKA
(hisztogram, QQ-plot)
SKEWNESS, KURTOSIS



STATISZTIKAI TESZTEK



MI AZ ALAP SZITUÁCIÓ?

1 MINTÁT VIZSGÁLOK

normál eloszlású adatok esetén

- ▶ **T-TEST**, ha kicsi a minta ($n \leq 30$)
- ▶ **Z-TEST**, ha nagy a minta ($n > 30$)

nem normál eloszlású adatok esetén

- ▶ **WILCOXON SIGNED-RANK TEST**

2 FÜGGETLEN MINTÁT VIZSGÁLOK

normál eloszlású adatok esetén

- ▶ **INDEPENDENT T-TEST**
(Student's or Welch's t-test)

nem normál eloszlású adatok esetén

- ▶ **MANN-WHITNEY U-TEST**

3 VAGY TÖBB MINTÁT VIZSGÁLOK

normál eloszlású adatok esetén

- ▶ **ONE-WAY ANOVA TEST**
(post hoc **Tukey-test**)

nem normál eloszlású adatok esetén

- ▶ **KRUSKAL-WALLIS TEST**
(post hoc **Dunn-test**)

NORMALITÁS VIZSGÁLAT

- ▶ **Shapiro-Wilk test** — kis minták normalitásának ellenőrzésére ($n < 500$)
- ▶ **Kolmogorov-Smirnov test** — nagyobb mintáknál, eloszlás összevetésére
- ▶ **D'Agostino-Pearson test** — ferdeség és csúcsosság összevetésére

KORRELÁCIÓ VIZSGÁLAT

- ▶ **Pearson coeff** — folytonos, normál eloszlású változók
- ▶ **Spearman coeff** — legalább egy ordinális változó és nem kell normál eloszlás
- ▶ **Khi-square test vagy Cramer's V** — kategórikus változók

GYAKOROLJUNK!



Közösen megoldandó feladatok:

[7_alkalom_csoport_osszehasonlito_tesztek.ipynb](#)

CSOPORT ÖSSZEHAISONLÍTÓ TESZTEK — MANN-WHITNEY U TESZT

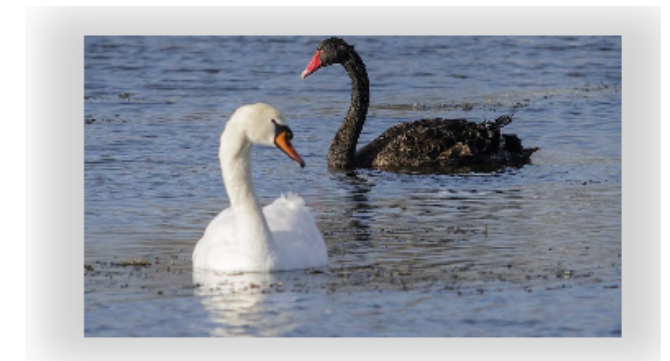


Mikor használjuk?

- ▶ ha két független csoportot akarunk összehasonlítani
- ▶ ha az adatok nem normális eloszlásúak, sok az outlier, megfigyelhető ferdeség

H₀ = a két csoport azonos eloszlásból származik
(nincs különbség a mediánok, illetve a rangeloszlások között)

H₁ = a két csoport különböző eloszlásból származik
(az egyik csoport elemei rendszerint nagyobb értéket vesznek fel)



```
from scipy import stats
```

```
group1 = df["group1_var"]
```

```
group2 = df["group2_var"]
```

```
U, p = stats.mannwhitneyu(group1, group2, alternative='two-sided')
```

```
print("Mann-Whitney U:", U, "p-value:", p)
```

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

ahol:

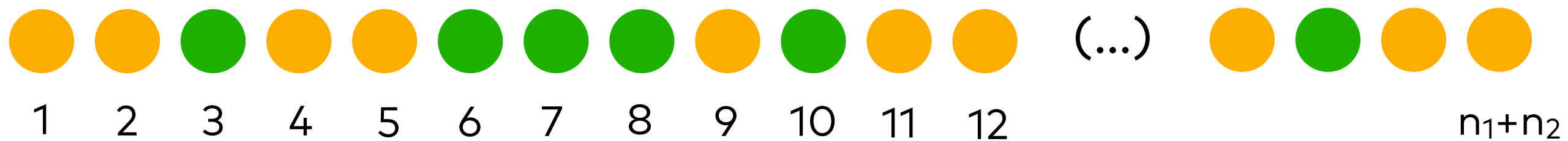
- R_1 : az első minta rangösszege
- n_1, n_2 : mintaelemszámok

CSOPORT ÖSSZEHAISONLÍTÓ TESZTEK — MANN-WHITNEY U TESZT



A teszt működése intuitívan:

- ▶ összefésüli mindkét csoport adatait — sorba rendezi őket így ad nekik rangot,
- ▶ majd azt vizsgálja, hogy az egyik csoport elemei tendenciózusan előrébb vagy hátrébb helyezkednek-e el a rangsorban



- ▶ $U \in [0, n_1 \cdot n_2]$ — azoknak az eseteknek a száma, amikor az elsőként átadott csoport elemei megelőzik a másodikként átadott csoport elemeit.

▶ $U = 0$ (...)

▶ $U = n_1 \cdot n_2$ (...)

- ▶ $p < 0.05 \Rightarrow$ elutasítjuk a H_0 -t \Rightarrow van különbség a mediánok (rangeloszlások) között
- ▶ $0.05 \leq p \Rightarrow$ most nem utasítjuk el a H_0 -t

GYAKOROLJUNK!



Közösen megoldandó feladatok:

[7_alkalom_csoport_osszehasonlito_tesztek.ipynb](#)

CSOPORT ÖSSZEHAISONLÍTÓ TESZTEK — KRUSKAL-WALLIS TESZT

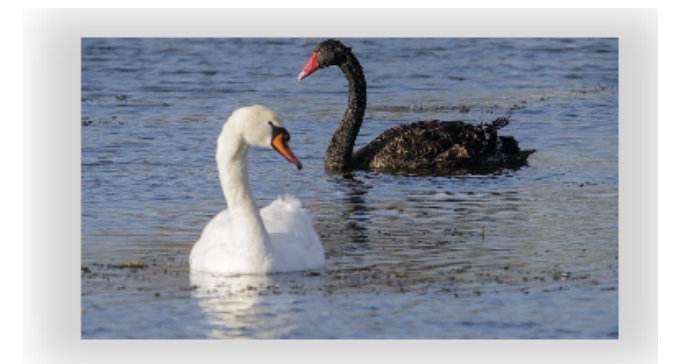


Mikor használjuk?

- ▶ ha kettőnél több független csoportot akarunk összehasonlítani
- ▶ ha az adatok nem normális eloszlásúak, sok az outlier, megfigyelhető ferdeség
- ▶ ha a csoportok szórása nagyon eltér egymástól
- ▶ ha a mintaelemszámok nem egyformák

H₀ = minden csoport azonos eloszlásból származik
(nincs különbség a mediánok, illetve a rangeloszlások között)

H₁ = a csoportok nem azonos eloszlásból származnak
(legalább egy csoport elemei rendszerint kisebb/nagyobb értéket vesznek fel)



```
from scipy import stats

# Kruskal-Wallis (nonparametric)
H, p = stats.kruskal(df['group1'], df['group2'], df['group3'])
print('Kruskal-Wallis:', H, p)
```

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

ahol:

- k = csoportok száma
- n_i = az i -edik csoport elemszáma
- R_i = az i -edik csoport rangösszege
- N = az összes megfigyelés száma

CSOPORT ÖSSZEHAISONLÍTÓ TESZTEK — KRUSKAL-WALLIS TESZT



A teszt működése intuitívan:

- ▶ Összefésüljük az összes adatpontot az összes csoportból
- ▶ Növekvő sorrendbe rendezzük őket → rangokat kapnak
- ▶ Megvizsgáljuk, hogy a rangösszegek mennyire térnek el attól az értéktől, amit akkor kapnának, ha nem lenne valódi különbség köztük

CSOPORT ÖSSZEHAISONLÍTÓ TESZTEK — KRUSKAL-WALLIS TESZT



EMBEREK



A csoport
(gyerekek)

120 cm
125 cm
130 cm



B csoport
(tinédzserek)

140 cm
150 cm
155 cm



C csoport
(felnőttek)

165 cm
170 cm
175 cm

120 cm, 125 cm, 130 cm, 140 cm, 150 cm, 155 cm, 165 cm, 170 cm, 175 cm

1

2

3

4

5

6

7

8

9

$$R_1 = 6$$

$$R_2 = 15$$

$$R_3 = 24$$

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \Rightarrow H = 7.2$$

TINÉDZSEREK



A csoport
(barna hajúak)

148 cm
150 cm
152 cm



B csoport
(szőke hajúak)

147 cm
151 cm
153 cm



C csoport
(vörös hajúak)

149 cm
150 cm
152 cm

147 cm, 148 cm, 149 cm, 150 cm, 150 cm, 151 cm, 152 cm, 152 cm, 153 cm

1

2

3

4.5

4.5

6

7.5

7.5

9

$$R_1 = 2+4.5+7.5 = 14$$

$$R_2 = 1+6+9 = 16$$

$$R_3 = 3+4.5+7.5 = 15$$

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \Rightarrow H \approx 0.09$$

CSOPORT ÖSSZEHAISONLÍTÓ TESZTEK — KRUSKAL-WALLIS TESZT



A teszt működése intuitívan:

- ▶ Összefésüljük az összes adatpontot az összes csoportból
- ▶ Növekvő sorrendbe rendezzük őket → rangokat kapnak
- ▶ Megvizsgáljuk, hogy a rangösszegek mennyire térnek el attól az értéktől, amit akkor kapnának, ha nem lenne valódi különbség köztük

A teszt statisztika értelmezése:

- ▶ $0 \leq H$ — megmutatja, hogy mennyire térnek el egymástól a csoportok rangeloszlásai
 - ▶ nincs felső korlátja, mert minél több a csoport és minél nagyobbak az elemszámok, annál nagyobb eltérések halmozódhatnak fel a rangösszegekben
 - ▶ minél nagyobb a H , annál nagyobb eltérés tapasztalható a rangösszegekben

CSOPORT ÖSSZEHASONLÍTÓ TESZTEK — KRUSKAL-WALLIS TESZT



Döntés a p érték alapján

- ▶ $0.05 \leq p \Rightarrow$ most nem utasítjuk el a H_0 -t
- ▶ $p < 0.05 \Rightarrow$ elutasítjuk a H_0 -t \Rightarrow van különbség a mediánok (rangeloszlások) között
- ▶ legalább egy csoport értékei rendszerint kisebb/nagyobb értéket vesznek fel
- ▶ na! melyik csoporté?



post hoc Dunn teszt

```
import scikit_posthocs as sp

# df has columns 'value' (numeric), 'group' (category/label)
dunn = sp.posthoc_dunn(df, val_col='value', group_col='group', p_adjust='bonferroni')
print(dunn) # pairwise p-values matrix
```

A (gyerekek) és C (felnőttek) csoportok
között van szignifikáns különbség



	A	B	C
A	1.000000	0.540000	0.021900
B	0.540000	1.000000	0.540000
C	0.021900	0.540000	1.000000

GYAKOROLJUNK!



Közösen megoldandó feladatok:

[7_alkalom_csoport_osszehasonlito_tesztek.ipynb](#)