

# MATEMATIKA ÉS STATISZTIKA DATA SCIENCE-HEZ



**HIPOTÉZISVIZSGÁLAT**  
**ELMÉLETE + GYAKORLATA**  
**HIPOTÉZIS ALKOTÁS,**  
**STATISZTIKAI TESZTEK ÉS P-ÉRTÉK,**  
**DÖNTÉS**

# HIPOTÉZISVIZSGÁLAT — MIKOR?



**Hipotézisvizsgálatot** akkor végzünk, ha...

- ▶ egy **feltételezést** adatok alapján szeretnénk ellenőrizni/**igazolni**,
- ▶ **össze** akarunk **hasonlítani** két (vagy több) csoportot,
- ▶ meg akarjuk tudni, hogy van-e **valós különbség vagy hatás**, vagy csak véletlen ingadozás





# 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 TESZT VÁLASZTÁSA



## 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 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 ( $n < 50$ )
- ▶ **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

# STATISZTIKAI TESZTEK I.



Statisztikai teszt	Mikor használjuk általában	Feltételek / megjegyzések
<b>Shapiro–Wilk teszt</b>	Kis mintáknál ( $n < 50$ , de akár 2000-ig is működik)	<ul style="list-style-type: none"><li>- Kifejezetten érzékeny a normalitástól való eltérésre</li><li>- Kisebb mintákra ajánlott</li><li>- A leggyakrabban használt normalitás-teszt</li></ul>
<b>Kolmogorov–Smirnov teszt</b>	Nagyobb minták esetén ( $n > 50$ ) vagy ha az eloszlást egy konkrét elméleti eloszláshoz hasonlítjuk	<ul style="list-style-type: none"><li>- Kevésbé érzékeny a kis eltérésekre</li><li>- Előre meg kell adni az elméleti eloszlást (pl. normál)</li><li>- Heterogén minták esetén torzíthat</li></ul>
<b>D'Agostino–Pearson teszt</b>	Közepes/nagy mintáknál, ha a minta nagysága elegendő ( $n > 30$ )	<ul style="list-style-type: none"><li>- A normalitást ferdeség és csúcsosság alapján értékeli</li><li>- Érzékeny a szélsőértékekre</li><li>- Stat. erősebb nagyobb mintánál</li></ul>

# STATISZTIKAI TESZTEK II.



Statisztikai teszt	Mikor használjuk általában?	Adattípus feltétel	Eloszlás feltétel	Egyéb feltételek
<b>Pearson- korreláció (r)</b>	Amikor két folytonos változó közti lineáris kapcsolatot vizsgálunk.	mindkét változó intervallum- vagy arányskála	normális eloszlás	kapcsolat lineáris nincsenek kiugró értékek
<b>Spearman- rangkorreláció (ρ)</b>	Ha a kapcsolat nem feltétlenül lineáris, de monoton (egyik nő, a másik is nő vagy csökken).	legalább egy ordinális (rangsorolható) változó	nem kell normális eloszlás	kapcsolat monoton kiugró értékekre nem érzékeny
<b>Khi-négyzet teszt (<math>\chi^2</math>)</b>	Két kategórikus változó közötti függetlenség vizsgálatára.	mindkét változó nominális	-	megfelelően nagy elemszám (cellagyakoriság > 5) megfigyelések függetlenek
<b>Cramer's V</b>	A Khi-négyzet teszt után, ha a kapcsolat erősségét is mérni akarjuk.	mindkét változó nominális	-	megfelelően nagy elemszám (cellagyakoriság > 5) megfigyelések függetlenek



# STATISZTIKAI TESZTEK III.



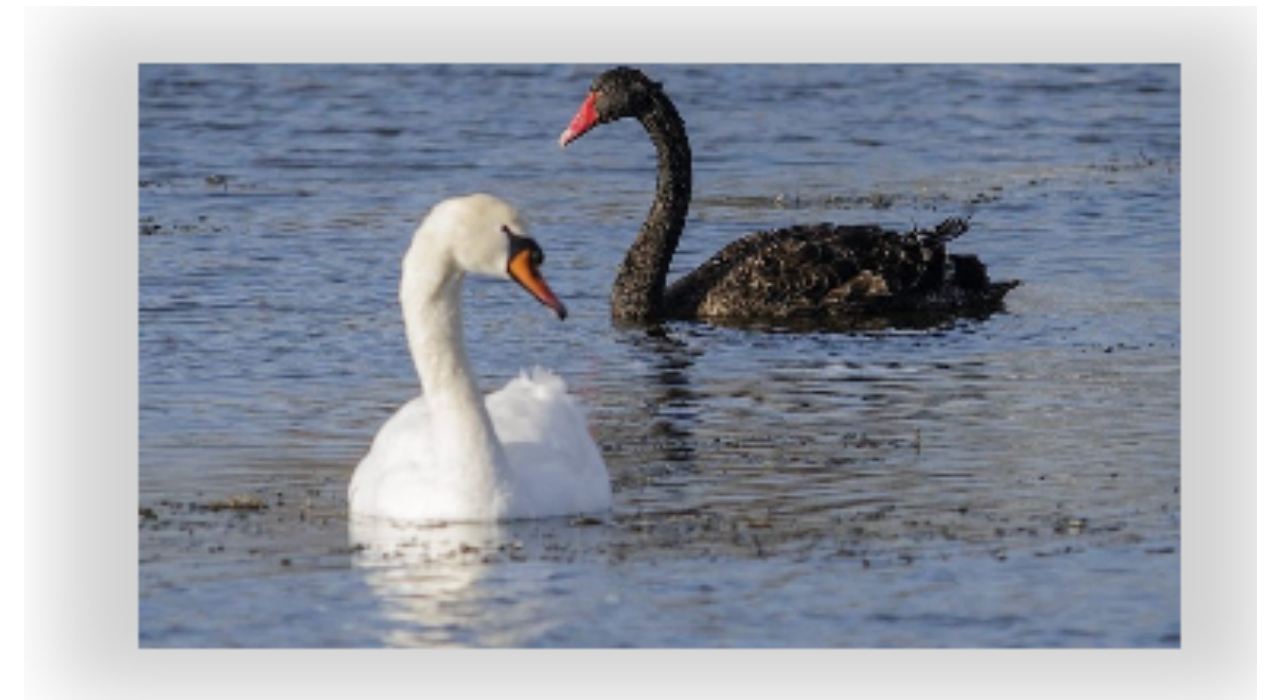
Statisztikai teszt	Mikor használjuk általában?	Adattípus feltétel	Eloszlás feltétel	Milyen statisztikai jellemzővel dolgozik?
<b>T-teszt (egy minta)</b>	Egy csoport átlagát hasonlítjuk össze egy ismert értékkel — kis minta ( $n < 30$ ).	Folytonos	Normális eloszlás	Átlag
<b>Z-teszt</b>	Egy csoport átlagát hasonlítjuk össze egy ismert értékkel — nagy minta ( $n > 30$ ), ismert a szórás.	Folytonos	Normális eloszlás	Átlag
<b>Wilcoxon signed- rank teszt</b>	Egy minta párosított értékeinek mediánját hasonlítjuk össze, ha nem normális az eloszlás.	Folytonos Ordinális	Nem kell normális eloszlás	Medián / rang
<b>két független mintás T-teszt</b>	Két független csoport átlagát hasonlítjuk össze.	Folytonos	Normális eloszlás	Átlag
<b>Mann–Whitney U-teszt</b>	Két független csoportot hasonlítunk össze, ha az adatok nem normál eloszlásúak.	Folytonos Ordinális	Nem kell normális eloszlás	Medián / rang
<b>One-way ANOVA teszt</b>	Három vagy több csoport átlagát hasonlítjuk össze.	Folytonos	Normális eloszlás	Átlag (varianciaelemzés alapján)
<b>Kruskal–Wallis teszt</b>	Három vagy több csoport mediánját hasonlítjuk össze, ha az adatok nem normális eloszlásúak.	Folytonos Ordinális	Nem kell normális eloszlás	Medián / rang

# NULLHIPOTÉZIS & ALTERNATÍV HIPOTÉZIS



**H<sub>0</sub>** általában azt állítja, hogy **nincs** hatás vagy különbség.

- ▶ fontos, hogy az állítás cáfolható legyen:  
"Minden hattyú fehér."
- ▶ a cél nem az, hogy bizonyítsuk, hanem hogy megpróbáljuk akár csak egyetlen egy ellenpéldával cáfolni == megtalálni a fekete hattyút



**H<sub>1</sub>** általában azt állítja, hogy **van** hatás vagy különbség.



# NULLHIPOTÉZIS & ALTERNATÍV HIPOTÉZIS



Statisztikai teszt	Nullhipotézis	Alternatív hipotézis
Shapiro–Wilk teszt	Az adatok normális eloszlásúak.	Az adatok nem normális eloszlásúak.
Kolmogorov–Smirnov teszt	A minta eloszlása nem különbözik a normál eloszlástól (vagy megadott eloszlástól).	A minta eloszlása szignifikánsan eltér a normál (vagy megadott) eloszlástól.
D’Agostino–Pearson teszt	Az adatok normális eloszlásúak (ferdeség és csúcsosság alapján).	Az adatok nem normális eloszlásúak.
Pearson korrelációs együttható	Nincs lineáris kapcsolat a két változó között ( $r = 0$ ).	Van lineáris kapcsolat a két változó között ( $r \neq 0$ ).
Spearman korrelációs együttható	Nincs monoton kapcsolat a két változó között ( $\rho = 0$ ).	Van monoton kapcsolat a két változó között ( $\rho \neq 0$ ).
Khi-négyzet teszt ( $\chi^2$ )	A kategóriák között nincs kapcsolat (a változók függetlenek).	A kategóriák között van kapcsolat (a változók nem függetlenek).

# NULLHIPOTÉZIS & ALTERNATÍV HIPOTÉZIS



Statisztikai teszt	Nullhipotézis	Alternatív hipotézis
T-teszt (egy minta)	A minta átlaga nem tér el a populációs átlagtól.	A minta átlaga eltér a populációs átlagtól.
Z-teszt	A minta átlaga nem tér el a populációs átlagtól.	A minta átlaga eltér a populációs átlagtól.
Wilcoxon signed-rank teszt	A páros megfigyelések között nincs különbség a mediánban.	A páros megfigyelések között van különbség a mediánban.
két független mintás T-teszt	A két független csoport átlaga megegyezik.	A két független csoport átlaga eltér.
Mann–Whitney U-teszt	A két független minta eloszlása megegyezik.	A két független minta eloszlása eltér.
One-way ANOVA teszt	A csoportok átlagai megegyeznek.	Legalább egy csoport átlaga eltér a többtől.
Kruskal–Wallis teszt	A csoportok eloszlása (mediánja) megegyezik.	Legalább egy csoport eloszlása (mediánja) eltér a többtől.

# TESZT STATISZTIKA ÉS P-ÉRTÉK



**Teszt statisztika** egy adott statisztikai értéket határoz meg, mely megadja az összefüggést a vizsgált változók / csoportok között.

**p-érték** annak a valószínűsége, hogy a kapott összefüggés csupán a véletlen műve (és  $H_0$  mégis igaz).



# TESZT STATISZTIKA ÉS P-ÉRTÉK



## Normalitás tesztek

```
from scipy import stats

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

# Kolmogorov-Smirnov (compare to normal distribution)
z = (df['variable'] - df['variable'].mean()) / df['variable'].std(ddof=1)
D, p = stats.kstest(z, 'norm')
print('Kolmogorov-Smirnov:', D, p)

# D'Agostino-Pearson (based on skewness and kurtosis)
K2, p = stats.normaltest(df['variable'])
print('D'Agostino-Pearson:', K2, 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 elvart kvantiliseiből származtatunk

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

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

$$K^2 = Z_1^2 + Z_2^2$$

ahol:

- $Z_1$  = a ferdeséghez (skewness) tartozó z-score
- $Z_2$  = a csúcsossághoz (kurtosis) tartozó z-score

# TESZT STATISZTIKA ÉS P-ÉRTÉK



## Korreláció vizsgálatnál

```
from scipy import stats

# Pearson correlation (linear relationship)
r, p = stats.pearsonr(df['x'], df['y'])
print('Pearson:', r, p)

# Spearman rank correlation (monotonic relationship)
rho, p = stats.spearmanr(df['x'], df['y'])
print('Spearman:', rho, p)

# Chi-square test for independence (categorical variables)
from scipy.stats import chi2_contingency
import pandas as pd

tbl = pd.crosstab(df['cat1'], df['cat2'])
chi2, p, dof, expected = chi2_contingency(tbl)
print('Chi-square:', chi2, p)
```

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

ahol:

- $x_i, y_i$ : megfigyelt értékpárok
- $\bar{x}, \bar{y}$ : minták átlaga

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

ahol:

- $d_i = \text{rang}(x_i) - \text{rang}(y_i)$ : az  $i$ -edik megfigyelés rangkülönbsége
- $n$ : a megfigyelések száma

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

ahol:

- $O_{ij}$ : megfigyelt gyakoriság a cellában
- $E_{ij} = \frac{(\text{row}_i \times \text{col}_j)}{N}$ : elvárt gyakoriság függetlenség esetén
- $r, c$ : a táblázat sorainak és oszlopainak száma



# TESZT STATISZTIKA ÉS P-ÉRTÉK



## 1 mintás statisztikai tesztek

```
from scipy import stats

# One-sample t-test (compare sample mean to population mean)
t, p = stats.ttest_1samp(df['variable'], popmean=0)
print('One-sample t-test:', t, p)

# Z-test (known population std or large sample)
from statsmodels.stats.weightstats import ztest
z, p = ztest(df['variable'], value=0)
print('Z-test:', z, p)

# Wilcoxon signed-rank (nonparametric)
W, p = stats.wilcoxon(df['before'], df['after'])
print('Wilcoxon signed-rank:', W, p)
```

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

ahol:

- $\bar{X}$ : mintaátlag
- $\mu_0$ : elméleti (nullhipotézis szerinti) átlag
- $s$ : minta szórása
- $n$ : minta elemszáma

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

ahol:

- $\bar{X}$ : a minta átlaga
- $\mu_0$ : a nullhipotézis szerinti populációs átlag
- $\sigma$ : a populáció szórása (ismert)
- $n$ : a minta elemszáma

1. Számítsd ki a különbségeket:  $d_i = x_i - M_0$
2. Hagyjuk el a nullákat, és rendezzük a különbségek abszolút értékeit.
3. Adjunk rangokat az abszolút értékekhez.
4. Számoljuk a pozitív és negatív rangösszegeket:  $W^+$  és  $W^-$ .
5. A tesztstatisztika:

$$W = \min(W^+, W^-)$$



# TESZT STATISZTIKA ÉS P-ÉRTÉK



## Két mintás statisztikai tesztek

```
from scipy import stats

# Independent t-test (parametric)
t, p = stats.ttest_ind(df['groupA'], df['groupB'], equal_var=False)
print('Independent t-test:', t, p)

# Mann-Whitney U-test (nonparametric)
U, p = stats.mannwhitneyu(df['groupA'], df['groupB'], alternative='two-sided')
print('Mann-Whitney U-test:', U, p)
```

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

ahol:

- $s_1, s_2$ : mintaszórások
- ha a varianciák egyenlők → **Student t-teszt**
- ha nem → **Welch t-teszt** (nem feltételezi azonos varianciát)

**Szabadságfok (Welch-féle közelítés):**

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

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

ahol:

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

# TESZT STATISZTIKA ÉS P-ÉRTÉK



## Több mintás statisztikai teszteknel

```
from scipy import stats

# One-way ANOVA (parametric)
F, p = stats.f_oneway(df['group1'], df['group2'], df['group3'])
print('One-way ANOVA:', F, p)

import pandas as pd
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# df has columns: 'value' (numeric), 'group' (category/label)
tukey = pairwise_tukeyhsd(endog=df['value'], groups=df['group'], alpha=0.05)
print(tukey.summary())
```

```
from scipy import stats

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

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
```

$$F = \frac{MS_B}{MS_W} = \frac{SS_B / (k - 1)}{SS_W / (N - k)}$$

ahol:

- $SS_B$  = Between-group Sum of Squares

$$SS_B = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X})^2$$

- $SS_W$  = Within-group Sum of Squares

$$SS_W = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

- $MS_B, MS_W$ : a négyzetösszegek csoportonkénti átlagai
- $k$ : csoportok száma
- $N$ : összes megfigyelés száma

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \bar{R}_j^2 - 3(N+1)$$

ahol:

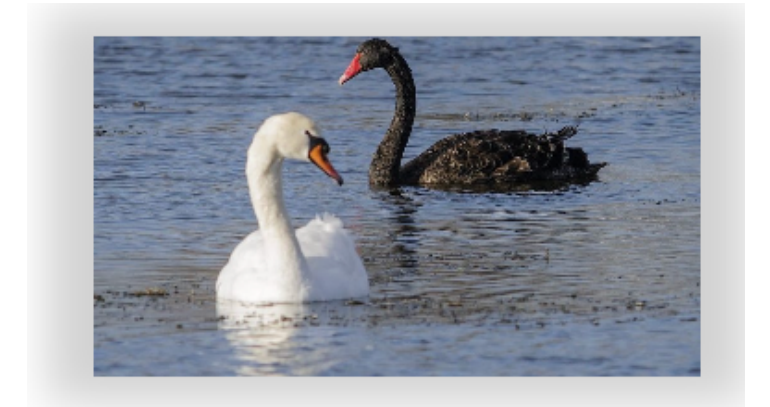
- $N$ : az összes megfigyelés száma
- $k$ : csoportok száma
- $n_j$ : az adott csoport elemszáma
- $\bar{R}_j$ : az adott csoport rangátlagának átlaga

# DÖNTÉS



Nullhipotézis ( $H_0$ ) és alternatív hipotézis állításai:

- ▶  $H_0$  általában azt állítja, hogy **nincs** hatás vagy különbség.
- ▶  $H_1$  azt állítja, hogy **van** hatás vagy különbség.



$$p < \alpha^*$$

akkor elutasítjuk a  $H_0$ -t

$\Rightarrow$  van hatás / különbség

$$\alpha \leq p$$

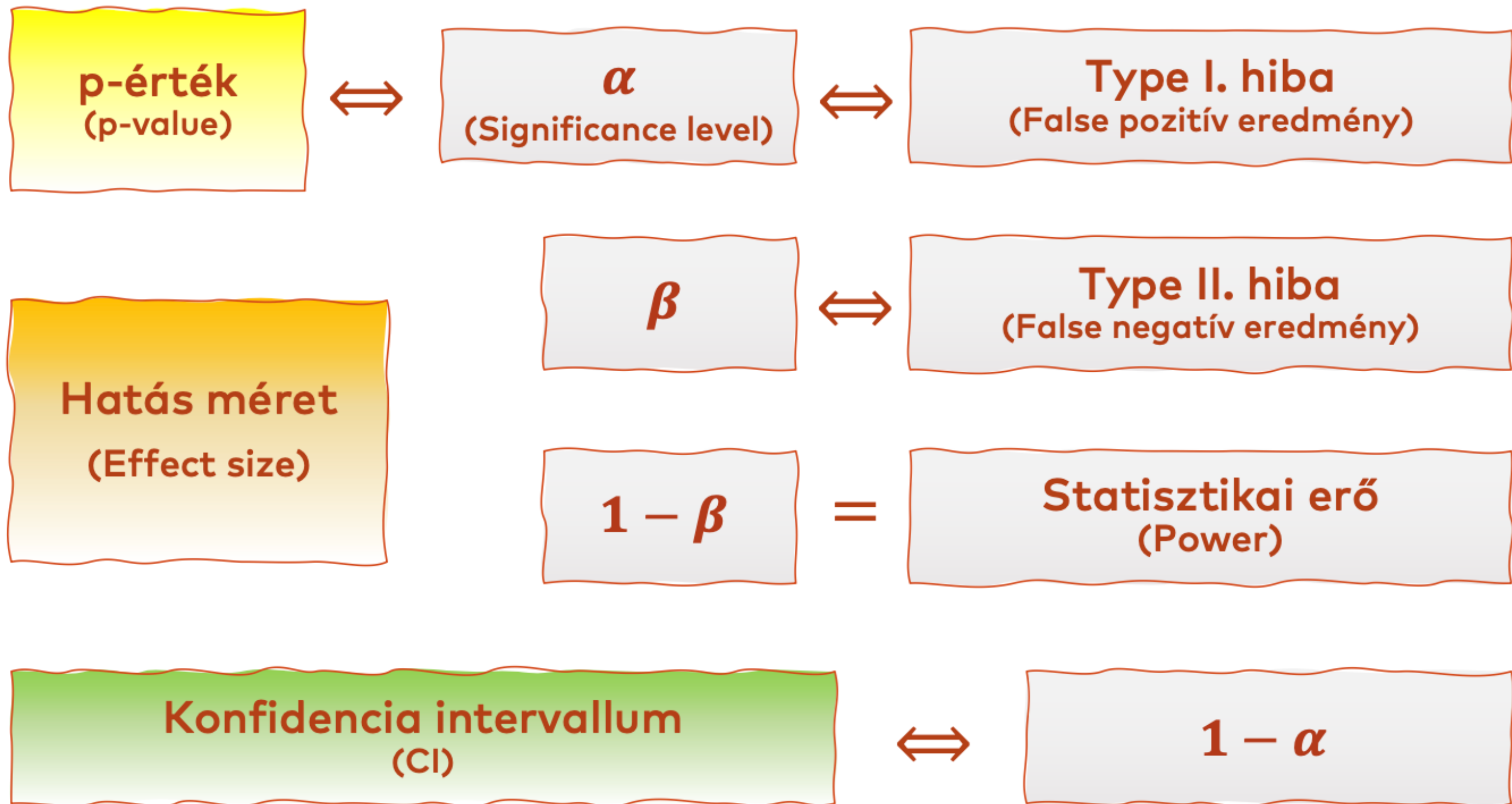
akkor (most) nem elutasítjuk a  $H_0$ -t

$\Rightarrow$  nincs hatás / különbség

\*  $\alpha$  az előre meghatározott szignifikancia szint (általában 0.05), ami megadja, hogy hány % kockázatot vállalunk rá, hogy úgy vetjük el a  $H_0$ -t, hogy az mégis igaz



# TOVÁBBI KAPCSOLÓDÓ FOGALMAK



# KAPCSOLÓDÓ DATAKLUB-OS VIDEÓK



## Statisztika Data Science-hez (4 részes mini sorozat)

<https://dataklub.hu/leckek/szignifikancia/>

- ▶ egy konkrét példa hipotézis vizsgálatra (18:00 — 28:25)
- ▶ az említett kapcsolódó magasabb szintű fogalmak (28:25 - 37:50)
- ▶ gyakorlati példa a magasabb szintű fogalmak megértéséhez (38:30 - 59:22)

<https://dataklub.hu/leckek/korrelacioelemzes/>

- ▶ Pearson & Spearman együtthatók alaposabb vizsgálata, képletek boncolgatása
- ▶ korreláció vizualizációja
- ▶ korreláció és ok-okozati összefüggések viszonya

# GYAKOROLJUNK!



Közösen megoldandó feladat — korreláció vizsgálat:

[5\\_alkalom\\_korrelacio\\_vizsgalat.ipynb](#)



# TOVÁBBI GYAKORLÓ FELADATOK



Önállóan megoldandó feladat az eddigi témakörökből:

`DK_mat_stat_gyakorlo_feladat.ipynb`