

# Korrelációs együtthatók

## Ábrák az R-ben

# Változók típusai

**Független változó:** A minta jellemzésére szolgáló mérőszám, gyakran a minta összeállításának szempontja is.

Független változók: kategóriális változók (nem, anyanyelv), ordinális változók (kor: fiatal, közép, éltés), metrikus változók (évek száma, kockán látható pöttyök).

**Függő változó:** A kísérlet hipotéziseinek tesztelésére szolgáló mérőszám.

Függő változók: számszerű változó, statisztikai teszt alapja, ordinális (pl. természetességi ítéletek) vagy metrikus (pl. távolságok, hőmérsékletek, gyakoriságok).

# Változók összefüggései

Ha két legalább ordinális függő változónk van, vagy egy legalább ordinális független és egy legalább ordinális függő változónk, a kettő összefüggései kifejezhetők korrelációs együtthatókkal.

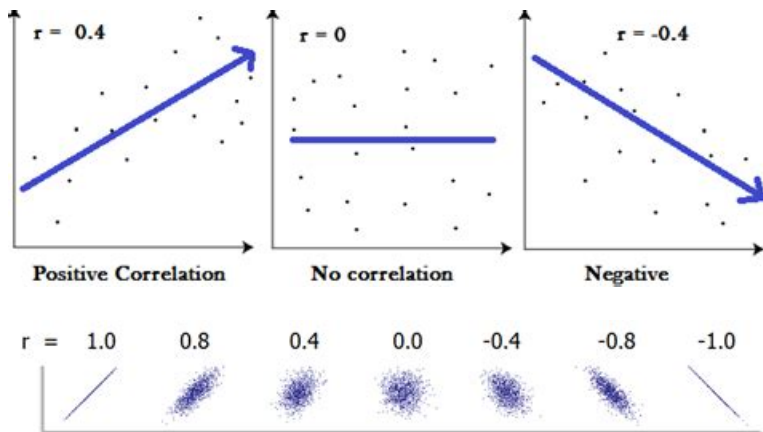
Lehetséges összefüggések  $a$  független és  $b$  függő változók között:

- ▶ Ahogy  $a$  növekszik, úgy nő  $b$  is.
- ▶ Ahogy  $a$  növekszik, úgy csökken  $b$ .
- ▶ A két változó között nem látszik összefüggés.

Erős korreláció lehetséges értelmezései:

1.  $b$  függő változó  $a$  független változó következménye.  
Klasszikus példa: iskolapadban eltöltött évek száma és éves jövedelem.
2. A két változó egy harmadik, figyelmen kívül hagyott vagy még ismeretlen változó függvénye. Híres példa: gólyák száma és születések száma (mindkettő csökkenő tendenciát mutat).

# Korreláció erőssége és iránya



# Korrelációs együttható értelmezése

Tartomány:  $-1$  és  $+1$  között.

- ▶ Ha a két változó szorosan összefügg, a korrelációs együttható  $+1$ -hez közelít.
- ▶ Ha a két változó között szoros fordított összefüggés áll fent, a korrelációs együttható  $-1$ -hez közelít.
- ▶ Ha a két változó között nincs összefüggés, az együttható értéke  $0$  körül mozog.

Együtthatók: Kendall-féle  $\tau$  vagy Spearman-féle  $\rho$  ordinális adatokra, Pearson-féle  $r$  metrikus adatokra.

# Korrelációs együtthatók: Kendall-féle $\tau$ (tau)

Ordinális adatok mérőszáma,  $-1$  és  $+1$  közötti tartományra esik.

- ▶ Előny: kis elemszám esetén megbízhatóbb, mint  $\rho$ .
- ▶ Hátrány: négyzete nem fogható fel determinációs együtthatóként (lásd  $\rho$  és  $r$ ).

Eljárás: elemek sorrendjének „jósága”, osztva a lehetséges párok számával.

**Proverzió (P):**  $y$  vektor elemeinek száma, amelyek a várt sorrendbe illeszkednek.

**Inverzió (I):** az  $a$  halmazhoz rendelt  $b$  értékeknél a várt sorrendtől való eltérés.

Négy településen élő gólyák és születések számának összefüggését vizsgáljuk egy adott évben.

Gólyák száma sorba rendezve Ács, Bő, Cece és Decs településen:  
 $a = [1, 5, 6, 10]$

Születések száma Ács, Bő, Cece és Decs településen:  
 $b = [21, 13, 36, 44]$

Az  $a$  halmazba tartozó gólyaszámok növekvő sorrendbe vannak állítva. Azt vizsgáljuk, hogy a hozzájuk tartozó születésszámok hasonlóan növekvő sorrendet mutatnak-e.

Megnézzük, hogy az első, 21-es értéket követő értékek magasabbak-e (proverzio, P) vagy alacsonyabbak (inverzio, I).

1. érték:  $P_{b1} = 2, I_{b1} = 1$

2. érték:  $P_{b2} = 2$

3. érték:  $P_{b3} = 1$

Lehetséges páronkénti összehasonlítások száma:  $\frac{n(n-1)}{2} = 6$

A Kendall-féle  $\tau$  kiszámítása:

$$\tau = \frac{P-I}{\frac{n(n-1)}{2}} = \frac{5-1}{6} = 0,6666666667$$



# Spearman-féle *Rho*

Ha az egyik vagy mindkét változó ordinális: Spearman-féle  $\rho$ .

Eljárás: mindkét változó értékeit sorrendbe állítjuk, eltérések különbségét ( $d$ ) négyzetre emeljük és összegezzük, majd az alábbi képlet szerint számítjuk ki:

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

## Példa

IQ és a tévé nézéssel töltött órák száma.

| IQ  | óra | x sorszáma | y sorszáma | d  | d <sup>2</sup> |
|-----|-----|------------|------------|----|----------------|
| 86  | 1   | 1          | 1          | 0  | 0              |
| 97  | 20  | 2          | 6          | -4 | 16             |
| 99  | 28  | 3          | 8          | -5 | 25             |
| 100 | 27  | 4          | 7          | -3 | 9              |
| 101 | 50  | 5          | 10         | -5 | 25             |
| 103 | 29  | 6          | 9          | -3 | 9              |
| 106 | 7   | 7          | 3          | 4  | 16             |
| 110 | 17  | 8          | 5          | 3  | 9              |
| 112 | 6   | 9          | 2          | 7  | 49             |
| 113 | 12  | 10         | 4          | 6  | 36             |

$$\rho = 1 - \frac{6 \cdot 194}{10 \cdot (100 - 1)} = 1 - \frac{1164}{990} = -0,175757$$

# Kovariancia

Ha két mérőszám  $(x, y)$  függ egymástól, akkor ha  $x$  eltér  $x$  átlagától, akkor  $y$  is el fog térni  $y$  átlagától pozitív vagy negatív irányba.

Eljárás:

1. Kiszámítjuk minden egyes pont átlagtól való eltérését.
2. Összegezzük az eltéréseket.
3. Ha a teljes populáció rendelkezésünkre áll, az összeget elosztjuk az összes elem számával,  $n$ -nel.
4. Ha populáció helyett mintával dolgozunk,  $n-1$ -gyel osztunk, mert a minta kovarianciája csak közelítő értéket ad.

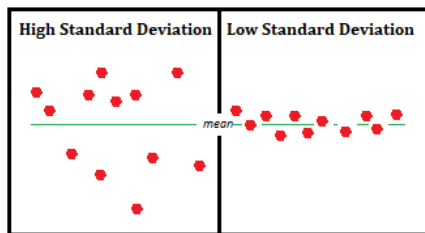
$$\sigma_{x,y} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}) * (y_k - \bar{y})$$

A kovariancia alsó értéke 0, a felső pedig végtelen.

# Szórás (standard deviation, SD)

A korreláció erősségénél figyelembe kell vennünk a két mérőszám szóródását: viszonylag szorosan csoportosulnak az átlag körül, vagy nagy terjedelemben?

Azaz: maga az átlag mennyire jellemzi jól az adatok összességét?  
Kis szórás esetén jó becslést ad, nagy szórás esetén nem.



Ábra forrása: <https://homework.study.com>

# Szórás (standard deviation)

A metrikus skála szóródásának mérőszáma a **szórás**. Alapja: az egyes pontok átlagtól való eltérése, ezek négyzetre emelése és összeadása. Azért emelünk négyzetre, mert az átlagtól való eltérés negatív vagy pozitív iránya nem érdekes, csak a mértéke. Az összegzés után az összegekből gyököt vonunk.

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

A szórás jelentőségéről a normális eloszlás kapcsán részletesen lesz szó.

# Kovariancia standardizálása

A korreláció jellemzésére szükség van egy mérőszámra, ami független a mintamérettől és a mintaátlagtól.

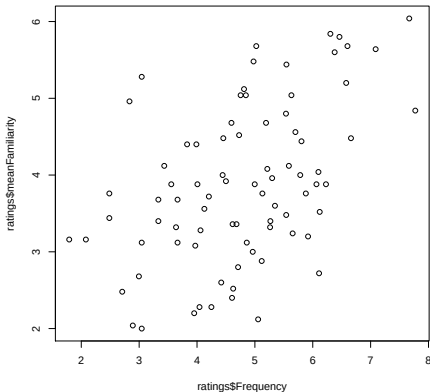
A kovariancia standardizálása, azaz az elemszámtól való függetlenítése a szórással való osztással történik. Az így kapott korrelációs együttható a Pearson-féle  $r$ .

$$r = \frac{\sum_{k=1}^n (x_k - \bar{x}) * (y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2} * \sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}}.$$

A Pearson-féle  $r$  a normális eloszlású metrikus adatok korrelációs együtthatója. Minimum értéke  $-1$ , maximuma  $+1$ , hasonlóan a Kendall-féle  $\tau$  és Spearman-féle  $\rho$  együtthatókhoz.

## Példa

Angol beszélőket megkérdeztek, hogy mennyire jól ismernek bizonyos növényeket és állatokat. Az ismertségi fokot 7-es skálán adták meg. Függ-e az ismertségi fok a szavak (növény- vagy állatnév) gyakoriságától?



## Példa

A szógyakoriság metrikus arányskála, hiszen van értelmezhető nullpontja (soha nem fordul elő), és az arányok is valósak (tíz előfordulás kétszer annyi, mint öt). Az ismertséget mérő 7-es skálát szintén metrikusnak lehet tekinteni. Ezért az összefüggést a Pearson-féle  $r$ -rel vizsgáljuk.

Az ábrán látszik, hogy minél nagyobb  $x$ , annál nagyobb  $y \rightarrow$  pozitív korreláció.

$$r = 0.48$$

$\rightarrow$  erős pozitív korreláció.



# Korreláció erőssége

$r$  értelmezése

- +0,70 és 1 között: nagyon erős pozitív összefüggés
- +0,40 és 0,69 között: erős pozitív összefüggés
- +0,30 és 0,39 között: közepesen erős pozitív összefüggés
- +0,20 és 0,29 között: gyenge pozitív összefüggés
- +0,19 és  $-0,19$  között: nincs vagy elhanyagolható összefüggés
- $-0,20$  és  $-0,29$  között: gyenge negatív összefüggés
- $-0,30$  és  $-0,39$  között: közepesen erős negatív összefüggés
- $-0,40$  és  $-0,69$  között: erős negatív összefüggés
- $-0,70$  és  $-1$  között: nagyon erős negatív összefüggés

Itt lehet gyakorolni pontdiagramok alapján az együttható megtippelését:

<https://correlation.streamlit.app/>

# Determinációs együttható

$d$ : a mért  $y$  értékek varianciájából mekkora részt (hány százalékot) magyaráz meg a regressziós becslések varianciája, vagyis a regressziós becslés milyen mértékben mutatja meg a  $b$  változó viselkedését.

Ha az összefüggés lineáris, az együttható:  $r^2$ .

Előbbi példa:

$$d = r^2 = 0,48^2 = 0,2304$$

Azaz: az összefüggés 23%-ban magyarázza meg  $y$  változó viselkedését  $x$  függvényében.

R

# Házi feladat

Kvartilisek listázása: `quantile(evszakok$kor)`

Megtekintés dobozdiagrammal nemenként:

`boxplot(kor~nem, evszakok)`

Az interkvartilis félterjedelem a doboz alsó és felső széle közötti különbség **fele**. A férfiaknál a medián (vízszintes fekete vonal) ehhez képest valamivel alacsonyabb, a nőknél sokkal alacsonyabb. Vagyis: mindkét nemben többségben van a fiatalabb korosztály az idősebbekhez képest.

Kumulatív gyakoriságok

`barplot(cumsum(table(evszakok$kor)))`

Az értékek nem egyenletesen nőnek: a legfeljebb 19 évesek száma jóval nagyobb a legfeljebb 18 éveseknél, 40 év fölött pedig egyre lassabban nőnek a kumulatív gyakoriságok.

# Házi feladat

Dobozdiagramok nemenként elválasztva a gyakoriság és ismertség adataira:

```
boxplot(Frequency~Class, ratings)
```

```
boxplot(meanFamiliarity~Class, ratings)
```

Szóhosszok százalékos arányban:

```
barplot(table(ratings$Length)/81*100)
```

# Programcsomagok telepítése az R-ben

Mivel az R nyílt forráskódú szoftver, bárki fejleszthet hozzá csomagokat. Elérhető csomagok listája az R mirror oldalakon, Packages menüpont alatt.

Telepítés interneten keresztül:

```
install.packages("languageR")
```

mirror kiválasztása (minél közelebb, annál kevesebb adatforgalmat generálunk).

Linux: csomagokat érdemes superuser-ként telepíteni, akkor a root könyvtárból bárhonnán elérhetők lesznek. Vagyis az R megnyitása:  
`sudo R`.

Windowson lehet, hogy rendszergazdaként kell megnyitni az R-t a csomag telepítése előtt.

# Csomag betöltése

betöltés (R megnyitása után minden egyes alkalommal):

```
library(languageR)
```

Ha `>` jelet kapunk „válaszként”, akkor a csomag betöltődött az R-be.

Ellenőrzés:

```
search()
```

aktuálisan betöltött csomagok listája.

Elérhető objektumok listája és rövid leírása: `languageR` telepítésének könyvtárában, az `INDEX` fájlban.

## Kétváltozós összefüggések ábrázolása

languageR könyvtár ratings objektuma.

`names(ratings)`: változók (oszlopok) neve.

`head(ratings)`, `tail(ratings)`: első hat és utolsó hat sor (= eset, record).

Hipotézis: A rövidebb állat- és növénynevek gyakoribbak.

Ábrázolás:

```
plot(ratings$Length,ratings$Frequency)
```

első érték: x-tengely, második érték: y-tengely.

Korrelációs együtthatók:

```
cor(ratings$Length,ratings$Frequency)
```

`method = "pearson"` -- default, alternatív:

```
"spearman", "kendall")
```

R érzékeny a kis- és nagybetűkre! Elnevezést lehet rövidíteni, ha különbség egyértelmű, itt elég "p", "s", "k".



# Grafikus paraméterek

Rengeteg paraméteren lehet állítani. Hogyan lehet ezekről tudni?

- ▶ grafikus parancs opcionális argumentumai. Lekérdezés: `?boxplot`, `?plot`, `?barplot` stb.
- ▶ parancsok súgója gyakran utal további hasznos parancsokra, pl. `line()`, `title()`, `abline()` stb.
- ▶ `par()`: rengeteg paraméter, pl. tengelyek feliratozása (felirat mérete, elhelyezése, egységek mérete), tengelyek aránya stb.

Első lépés súgó. Felépítés: (1) kötelező és opcionális argumentumok listája, (2) argumentumok rövid magyarázata, (3) részletek: többnyire innen derül ki a releváns infó, ha még nem ismerjük a parancsot, gyakran hivatkozások is, (4) lásd még - hasznos, esetenként hasznosabb, további parancsok, (5) példák - ezek általában túl bonyolultak, ezért nem túl hasznosak.

## Néhány hasznos paraméter

- ▶ **xlab, ylab:** "x-tengely felirata", "y-tengely felirata".
- ▶ **main:** "Ábra címe".
- ▶ **xlim, ylim:** Ábrázolt értékek tól-ig. Főleg y-tengelynél fontos, ha összehasonlítható ábrákat akarunk. Pl százalékos ábrázolásnál `ylim = c(0,100)`, azaz 0–100%-ig.  
Egyenlőségjel előtti szóköz opcionális.
- ▶ **col:** színek, vagy névvel, vagy számmal. Pl. `col=2` és `col="red"` azonosak.
- ▶ **cex:** `cex.main`, `cex.axis`, `cex.names` stb. Default: `cex=1`, ehhez képest cím, mérőszámok, címkék betűmérete nagyobb (1.3, 1.7) vagy kisebb (0.7).

# Ábra mentése

Mentés pdf-ként:

```
plot(blablabla)
dev.print("cél fájl", device=pdf)
```

Mentés ábraként, pl. png (jpg és tif nem jó, mert képtömörítéssel készül!):

```
png(file = "fajlnev", bg = "white")
plot(blablabla)
dev.off()
```

Alternatíva Windows-os R-ben: menün keresztül.

Ha nem adunk meg elérési útvonalat: mentés aktuális könyvtárba (`getwd()` paranccsal megtudható).

# Boxplotok

Ábrázolás módja: y-tengely: függő változó interkvartilis eloszlása,  
x-tengely: csoportok, esetleg további tagolással.

```
boxplot(függőváltozó~függetlenváltozó)
```

Ha további csoportosítás:

```
boxplot(függőváltozó~csoport*függetlenváltozó)
```

Például:

```
boxplot(ratings$Frequency~ratings$Class*ratings$Complex)
```

Egyszerű és összetett állat- és növénynevek gyakorisága.

Hasznos paraméterek:

`col=c("red","blue")` vagy `col=c(2,4)` – ugyanaz az eredmény.

`names=c("állat","növény")`

# Házi feladatok

1. feladat: ratings fájlban található adatok alapján tetszőleges pontdiagramm készítése és mentése.

2. feladat: ratings fájlban található adatok alapján tetszőleges boxplot készítése.

Cél: cím, angol vagy magyar nyelvű tengelyfelirat, színes ábrák.