

# A statisztika alapelvei, hipotézisek

## Objektumok az R-ben

# Félév beosztása

- ▶ Hipotézisek, skálatípusok.
- ▶ Eloszlások, szórás.
- ▶ Korrelációs számítás.
- ▶ Normális eloszlás, standard normális eloszlás.
- ▶ Valószínűség, mintavétel, konfidencia-tartomány, szignifikancia.
- ▶ Nem parametrikus próbák: khi-négyzet, Wilcoxon, Mann-Whitney, Kruskal-Wallis.
- ▶ Lineáris regresszió, variancia és átlag összehasonlítása: F-próba, t-próba.

## Követelmény:

- ▶ házi feladatok, következő hét hétfő estig kell küldeni emailben (legfeljebb kettő hiányozhat félév végéig),
- ▶ zárthelyi dolgozat a félév közepén és a teljes anyagból a félév végén.

# Irodalom

Órák anyaga:

[phon.nytud.hu/mady/courses/statistics/2024](http://phon.nytud.hu/mady/courses/statistics/2024)

Általános, bevezető részekhez:

Reiczigel, J., Harnos, A. & Solymosi, N. (2010): Biostatisztika nem statisztikusoknak. Nagykovácsi: Pars.

Nyelvészeti kérdések, feladatok:

Baayen, R. H. (2008): Analyzing linguistic data: a practical introduction to statistics using R. Cambridge: University Press.

Az R működésének jobb megismeréséhez:

Peter Dalgaard (2008): Introductory statistics with R. New York: Springer.

További modellek, főleg a pszichológiában

Field, Andy, Miles, Jeremy, & Field, Zoë (2012): *Discovering statistics using R*. London: SAGE.

# Statisztikailag tesztelhető állítások

- ▶ Ha négy hétig fogyasztó koktélokat eszünk, soványabbak leszünk.
- ▶ A kétnyelvű hatévesek kognitív teljesítménye jobb, mint az egynyelvűeké.
- ▶ Juli néni egy nap alatt több időt tölt beszéléssel, mint Feri bácsi.
- ▶ Az angolok többet olvasnak, mint a franciák.
- ▶ Az éghajlat ma melegebb, mint 100 évvel ezelőtt.

# Statisztikailag nem tesztelhető állítások

- ▶ A keserűcsokoládé finomabb, mint a tejsoki.
- ▶ A patkány a legrondább állat, a pók szorosan követi.
- ▶ Az emberek régen sokkal szebben beszéltek, mint ma.
- ▶ Az Ödön nevű emberek ellenszenvesek.

Miért?

Ezek az állítások szubjektívek, nem tesztelhetőek számszerű adatokkal, azaz nem mérhetőek.

# Kvantitatív és kvalitatív jellemzés

**Kvalitatív adatok:** megfigyelések részletes leírása, pl. a csokoládétípusok közötti különbségek leírása, az emberek undora a patkányoktól és a pókoktól, a nők és háziasszonyok szociális helyzetének leírása.

**Kvantitatív adatok:** megszámlálható vagy mérhető egységek.

A kvantitatív adatgyűjtést gyakran megelőzi a kvalitatív adatgyűjtés, ami alapján kiválaszthatóak a releváns, azaz tesztelendő változók.

A kvalitatív adatok gyakran kvantifikálhatóak a kérdés átfogalmazásával. Például: a keserűcsokoládé ízének megítélése egy 5-ös skálán. A múltra és személyekre vonatkozó kijelentések nehezen kvantifikálhatóak, sőt, objektívvá sem mindig tehetőek.

# Kezdeti megfigyelések

Egy kísérlet gyakran egy sejtésen vagy megfigyelésen alapul.

- ▶ Idén sokkal több a szúnyog a Balatonnál, mint az előző években.
- ▶ A nők több verbális és nonverbális visszajelzést adnak egy társalgásban, mint a férfiak.
- ▶ Az emberek mostanában világosabb ruhákat hordanak, mint régebben.

# Az elmélet felállítása

A tudományos kísérletekben a meglévő eredmények alapján feltételezhetőek bizonyos magyarázatok, amiket tesztelhetünk. Vegyük a szúnyogokat.

Potenciális magyarázatok:

- ▶ A szúnyogok szaporodási időszakában ideálisak voltak az időjárási viszonyok, ezért sok egyed életben maradt.
- ▶ Kevesebb volt a szervezett irtás.
- ▶ Immunissá váltak egy gyakran használt irtószerre.



# Kísérleti dizájn

Az elméletek tesztelésére összehasonlítható adatokat kell gyűjtenünk:

- ▶ Szúnyogok száma egy adott területen, ahol a szaporodási időszakban napos, nedves, hideg vagy meleg volt az idő.
- ▶ Irtási akciók gyakorisága.
- ▶ Az egyes irtószerek használatának mértéke korábbi évekkel összehasonlítva.

A kísérletező által kontrollált változók (napos, hideg terület stb.) neve **független változó**. Egyéb elnevezések: faktor, tényező, magyarázó változó.

A kísérlet során gyűjtött kvantitatív adatok neve (pl. szúnyogok száma a tesztelt területen) **függő változó**, mert függ az adatgyűjtés körülményeitől.

# A statisztikai elemzések háttere

Fallibilizmus (Popper): egy állítás igazolásából nem következik az igazság („minden hattyú fehér”).

Alaptézisek:

- ▶ Egyszeri vagy ismételt megfigyelésből nem lehet levezetni, hogy valami törvényszerű. Ha csak fehér hattyút látok, abból még nem következik, hogy minden hattyú fehér.
- ▶ Falszifikáció: keress fekete hattyút! Amíg az eredmény negatív (nem láttunk fekete hattyút), addig nem dőlt meg a hipotézis, hogy minden hattyú fehér.
- ▶ Alapfeltétel: megdönthetőség. A kísérleti módszert úgy kell megválasztani, hogy a hipotézis, amennyiben helytelen, megdönthető legyen.

# Követelmények

**Operacionalizálás:** kérdésfeltevés úgy, hogy empirikusan megfigyelhető adatok alapján megválaszolható legyen → körültekintő kísérlettervezés.

Állítás: a mai emberek igénytelenebbül beszélnek, mint a régiek.  
Mérőszámok? Összehasonlíthatóság?

**Reprodukálhatóság:** a kísérleti dizájn és a felhasznált módszerek alapján az eredményeknek megismételhetőeknek kell lenniük.  
Előfeltétel: módszerek részletes leírása.

**Objektivitás:** függetlenség a kísérletvezetőtől és a kísérlet körülményeitől.

Az Ödönöket reggel 6-ra hívjuk be tesztelni, a más nevéek választhatnak időpontot 😊

# Alapfogalmak

**Populáció vagy sokaság:** a vizsgálandó elemek összessége, véges vagy végtelen. A teljes populáció vizsgálata többnyire lehetetlen.

**Reprezentatív mintavétel:** fontos tulajdonságainak arányában megfelel a populáció megfelelő tulajdonságainak, pl. kor szerinti eloszlás. Véletlenszerű: minden elem egyforma valószínűséggel kerülhet bele a mintavételbe, pl. Magyarország összes országbérlet-tulajdonosa.

**Írányítottan reprezentatív mintavétel:** populáció eloszlásának többszemponútú leképezése, csoportokon belül véletlenszerű kiválasztás. Pl. országbérlet-tulajdonosok nem, kor és a lakóhely mérete szerint súlyozva.

# Hipotézisek

**Hipotézis:** feltételezés. Itt: előzetes megfigyelésen alapuló válasz tudományos kérdésfeltevésekre („úgy sejthető, hogy...”).

**Kísérleti hipotézis:** a változók mennyiségi viszonyára vonatkozó állítás, pl. 2020-ban több volt a szúnyog, mint a megelőző 9 év átlaga.

**Statisztikai vagy sztochasztikus hipotézis:** egy adott esemény bizonyos körülmények között egy bizonyos valószínűséggel bekövetkezik.

A statisztikai tesztelés alapja: a felállított hipotézis **ellentétét** próbáljuk igazolni, a falszifikálhatóság kívánalmának eleget téve („keress fekete hattyút”).

# Hipotézisállítás

Cél: szeretnénk igazolni, hogy az általunk forgalomba hozott Szerecsen kávé hosszabb időn keresztül frissen tartja a fogyasztóit, mint a Hagyományos kávé.

**Nullhipotézis ( $H_0$ ):** A Szerecsen és a Hagyományos kávé fogyasztása után a tesztalanyok **azonos** ideig maradnak frissek.

**Alternatív hipotézis ( $H_1$ ):** a Szerecsen kávé fogyasztók később érzik magukat újra fáradtnak.

Vagyis: azt próbáljuk bizonyítani statisztikai modellekkel, hogy NINCS különbség a két mintára vonatkozó mérések között, vagyis a Szerecsen és a Hagyományos kávé után egyaránt kb. 2 órán át maradunk frissek. Ha a statisztikai modell azt mutatja, hogy nagyon kicsi a valószínűsége, hogy a két kávéhoz tartozó minták egyazon populációhoz tartozzanak, akkor elvetjük a nullhipotézist, és megtartjuk az alternatív hipotézist.

# Közös kísérlet

Gyakran elhangzik, hogy a tavasz a legkellemesebb évszak: egyre hosszabbak a nappalok, zöldül a természet, stb.

Szerintünk is a tavasz hat a legjobban az emberek hangulatára, és ezt szeretnénk kísérlettel igazolni.

Hogyan lehet ezt tesztelni? Javasolt módszer: értékelés egy 1 és 20 közötti pontszámmal, hogy a megkérdezett mennyire szereti a tavaszt. Ezt hasonlítsuk össze azzal, hogy mások mennyire szeretik az ősz.

- ▶ Kiktől gyűjtsünk adatokat? Az egyszerűség kedvéért a saját ismeretségi körből.
- ▶ Hogyan érhetjük el, hogy kiegyensúlyozott legyen a mintavétel? Próbáljunk eltérő korosztályú embereket megkérdezni.

Az adatgyűjtés módja: fejenként öt ismerőst megkérünk, hogy 1 és 20 pont között adja meg, mennyire szereti a tavaszt. Öt másik ismerőst pedig az ősről kérdezzük.



Feladat: legalább tíz válasz gyűjtése.

Táblázat készítése Excelben vagy LibreOffice Spreadsheet formátumban. Oszlopnevek:

1. K: a kérdező azonosítója (pl. monogram),
2. NEM: a beszélgetőpartner neme (N, F),
3. KOR: a beszélgetőpartner feltételezett kora (egész szám),
4. EVSZAK: tavaszra vagy ősze vonatkozik a pontszám (T, O),
5. PONT: az elhangzó pontszám lejegyzése.

Mentés: csv-fájlként, UTF8-as kódolással,  
valaszadovezetekneve.csv fájlként.

Leadás: mady@nytud.hu, 2024. február 19., hétfő estig.

R

Eredeti programnyelv S, ebből licenz alapú S+, ennek felel meg az S-, azaz R, mint eggyel korábbi betű az ABC-ben. Fejlesztők: **R**oss Ihaka és **R**obert Gentleman.

Letöltés: [www.r-project.org](http://www.r-project.org), onnan elérhető tükrök.

Windows GUI (graphical user interface): személyre szabott telepítés: eldönthető, hogy terminál és ábrák egy ablakba kerüljenek vagy kettőbe.

Linux: általában alapcsomag része, ha nem, repositoryból letölthető. Nincs GUI, megnyitás terminálablakban R paranccsal.

() függvény jele, így különböztetjük meg a változóktól.  
function(argument1, argument2, ...)

Adatok beadása:

```
a = c(1,2,3,4)
```

c(): concatenate = kösd össze

```
b = c(3,5,2,1)
```

a és b vektor egydimenziós, x-edik eleme a[x], b[x].

```
m = cbind(a,b) cbind = bind as columns
```

Mátrix m: kétdimenziós, első változó a sor, a második az oszlop.

Egész oszlop: m[,1], egész sor: m[2,], egy adott cella: m[1,2].

Karakterváltozók idézőjelek között:

```
d = c("n", "n", "f", "f")
```

Lehet = helyett <- jelet is használni, c <- cbind(a,b), sőt,  
fordítva is: cbind(a,b) -> c. R-specifikus, hardcore  
felhasználók így adják meg leírásokban.

Honnan tudjuk, milyen változó?

```
class()
```

pl. numeric, character, matrix

Ha bizonyos változókkal bizonyos műveleteket végzünk, az R néha átalakítja őket!

```
e = cbind(a,d)
```

Ok: mátrix csak egyféle változótípusból állhat, ezért a számból betű lesz.

```
e = data.frame(a,b,d)
```

Oszlopok: változók, osztályuk lekérdezhető így: `class(e[,3])`.

# Feladat

Dobjunk fel egy dobókockát tizenkétszer, és írjuk le, milyen számokat kaptunk.

## Ha nincs dobókockánk...

Egész számok véletlenszerű generálása:

```
sample()
```

Argumentumok és jelentés: `?sample` vagy `help(sample)`.

Két argumentum megadása kötelező: `x` = miből húzzon, `size` = hányszor.

`x` lehet `c(1,2,3,4,5,6)` vagy `1:6` vagy `seq(1,6)`.

```
kocka = sample(1:6,12)
```

⇒ hibajelzés.

Figyelem! Default argumentum szerint egy számot csak egyszer lehet kihúzni: `replace=FALSE`. Megoldás: `replace=TRUE`.

# 1. feladat

Foglaljuk össze egy mátrixban, hányadikra hányat dobtunk. Így nézzen ki:

1	6
2	6
3	5
4	6
.	.
.	.
.	.

Adjuk meg, hogy hányszor dobtunk valamilyen számot, valamint adjuk meg az átlagukat és az összegüket.

Hasznos parancsok: `cbind()`, `rbind()`, `table()`, `mean()`, `sum()`.



## 2. feladat

Dobjunk 100-at, majd 1000-et. MiHasonlítsuk össze az átlagokat.