# 1
# A statisztika alapelvei, hipotézisek
# Objektumok az R-ben

# Félév beosztása

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

Ha marad rá idő, kedv, kitartás:

- Lineáris regresszió.
- Varianciaanalízis (ANOVA).
- Kevert modellek.
- Gépi tanulás: clusterek, döntési fák.

# Irodalom

Órák anyaga:
`clara.nytud.hu/~mady/courses/statistics/ba`
Bevezetés az R-be, FAQ, teljes kézikönyv (*help* teljes anyaga pdf-ben).


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

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

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

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

# Statisztikailag tesztelhető állítások

# Statisztikailag tesztelhető állítások

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

# Statisztikailag nem tesztelhető állítások

# Statisztikailag nem tesztelhető állítások

- A keserűcsokoládé finomabb, mint a tejcsoki.
- A patkány a legrondább állat, a pók szorosan követi.
- A nők arra vannak teremtve, hogy ellássák a háztartást és a gyerekeket.

Miért?

# Statisztikailag nem tesztelhető állítások

- A keserűcsokoládé finomabb, mint a tejcsoki.
- A patkány a legrondább állat, a pók szorosan követi.
- A nők arra vannak teremtve, hogy ellássák a háztartást és a gyerekeket.

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

# Kvantitatív és kvalitatív leírás

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

**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.

A kvalitatív adatok gyakran kvantifikálhatóak a kérdés átfogalmazásával. Például: a keserűcsokoládé ízének megítélése egy 5-ös skálán, háziasszonyok és női vezetők elégedettségi mutatói.

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

# Kezdeti megfigyelések

Egy kísérlet gyakran egy sejtésen vagy megfigyelésen alapul.

- Idén több darázs van, mint a korábbi években.
- A férfiak általában inkább felhívnak valakit, a nők SMS-t küldenek.
- A nők több verbális és nonverbális visszajelzést adnak egy társalgásban, mint a férfiak.

# Az elmélet felállítása

A tudományos kísérletekben a meglévő eredmények alapján feltételezhetőek bizonyos magyarázatok, amiket tesztelhetünk. A darazsak esetében:

Potenciális magyarázatok:

- A darazsak szaporodási időszakában ideálisak voltak az időjárási viszonyok, ezért sokan életben maradtak.
- Immunissá váltak egy gyakran használt irtószerre.

# Kísérleti dizájn

Az elméletek tesztelésére összehasonlítható adatokat kell gyűjtenünk:

- Darazsak száma egy adott területen, ahol a szaporodási időszakban napos, nedves, hideg vagy meleg volt az idő.
- Az aktuális irtószerek bevetésének kezdeti időszakával való összehasonlítás.

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

A kísérlet során gyűjtött adatok neve (pl. darazsak száma) **függő változó**, mert függ az adatgyűjtés körülményeitől.

# Háttér

Fallibilizmus (Popper): az igazolásból nem következik az igazság.

Alaptézisek:

- Ismételt megfigyelésből nem lehet levezetni, hogy valami törvényszerű. Ha csak fehér hattyút látok, abból nem következik, hogy minden hattyú fehér.
- Falszifikáció: keress fekete hattyút! Amíg az eredmény negatív (nem láttunk fekete hattyút), addig nem dőlt meg az 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 $\rightarrow$ 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.
Tudok egy tavat, ahonnan kitelepítették a fekete hattyúkat. Oda megyek kísérletezni ☺

# Hipotézisek

**Hipotézis:** feltételezés. Itt: előzetes válasz tudományos kérdésfeltevésekre.

**Kísérleti hipotézis:** a változók viszonyára vonatkozó állítás.

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

# Hipotézisállítás

Szeretnénk bizonyítani, hogy az általunk forgalomba hozott Szerecsen kávé hosszabb időn keresztül frissen tartja a fogyasztóit, mint a Hagyományos kávé(k).

**Kiindulási vagy nullhipotézis ($H_0$):** A Szerecsen és a Hagyományos kávé fogyasztása után a tesztalanyok azonos ideig maradnak frissek.
**Ellenhipotézis ($H_1$):** A Szerecsen kávét fogyasztók átlagosan fél órával később érezték magukat újra fáradtnak.

Miért?

# Hipotézisállítás

Szeretnénk bizonyítani, hogy az általunk forgalomba hozott Szerecsen kávé hosszabb időn keresztül frissen tartja a fogyasztóit, mint a Hagyományos kávé(k).

**Kiindulási vagy nullhipotézis ($H_0$):** A Szerecsen és a Hagyományos kávé fogyasztása után a tesztalanyok azonos ideig maradnak frissek.
**Ellenhipotézis ($H_1$):** A Szerecsen kávét fogyasztók átlagosan fél órával később érezték magukat újra fáradtnak.

Miért? „Keress fekete hattyút!" Amíg nem találsz, addig a korábbi hipotézis (minden hattyú fehér) marad életben.

Statisztikai eljárás: nullhipotézis tesztelése. Ha $H_0$ valószínűsége a megadott küszöbnél nagyobb $\rightarrow$ megtartjuk $H_0$-t. Ha $H_0$ valószínűsége csekély $\rightarrow$ elutasítjuk, és feltételezzük, hogy $H_1$ igaz.

# Alapfogalmak

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

**Reprezentantív mintavétel:** fontos tulajdonságainak arányában megfelel a populáció megfelelő tulajdonságainak. Véletlenszerű: minden kiválasztott elem egyforma valószínűséggel kerülhet bele a mintavételbe.

**Irányítottan reprezentantív mintavétel:** populáció eloszlásának leképezése, csoportokon belül véletlenszerű kiválasztás. Pl. egyetemisták nem, szakirány, évfolyam szerint súlyozva.

# Csoportos kísérlet

Egy külföldi kolléga szerint a magyarban van egy olyan szabály, hogy a *Viszontlátásra* köszönésre *Viszlát* a válasz. Ha viszont valaki a *Viszontlátásra* kifejezést használja, a beszélgetőpartner *Viszlát*-tal fog elköszönni.

Hogyan lehet ezt tesztelni?

- Kikből áll a populáció?
- Hogyan gyűjtsünk adatokat?
- Hogyan érhetjük el, hogy kiegyensúlyozott legyen a mintavétel?
- Mire van szükség a reprezentatív mintavételhez?
- Mik az adatgyűjtés buktatói?

R

Eredeti programnyelv S, ebből licensz alapú S+, ennek felel meg állítólag az S-, azaz R. Fejlesztők: **R**oss Ihaka és **R**obert Gentleman.

Letöltés: www.r-project.org, onnan elérhető tükrök.

Windows GUI: 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,...)`

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

Adatok beadása:
`a = c(1,2,3,4)`
`c()`: concatenate = kösd össze

() 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]`.

() 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].

() 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].

String változók idézőjelek között:
`d = c("n","n","f","f")`

() 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].

String változók 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

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.

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)`.

# 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)
```

# 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)
```

$\Rightarrow$ hibajelzés.

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

# Feladatok

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()`.

# Feladatok

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