

1

Hipotézisek, skálatípusok Objektumok az R-ben

Félév beosztása

1. Hipotézisek, skálatípusok.
2. Eloszlások, szórás.
3. Korrelációs számítás.
4. Normális eloszlás, standard normális eloszlás.
5. Valószínűség, mintavétel, konfidencia-tartomány, szignifikancia.
6. Variancia és átlag összehasonlítása: F-próba, t-próba.
7. Nem parametrikus tesztek: khi-négyzet, Wilcoxon, Mann-Whitney, Kruskal-Wallis.
8. Lineáris regresszió.
9. Varianciaanalízis (ANOVA).
10. Kevert modellek.
11. Gépi tanulás: clusterek, döntési fák.

órák anyaga itt:

`clara.nytud.hu/~mady/courses/statistics/intro`

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, Harnos & Solymosi (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 tejsoki.
- ▶ A patkány a legrondább állat, a pók szorosan követi.
- ▶ A nők arra vannak teremtvé, 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 teremtvé, 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: megszámlálható 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 visszacsatolást 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 igaz a hipotézis, hogy minden hattyú fehér.
- ▶ Alapfeltétel: megdönthetőség. A kísérleti módszert úgy kell megválasztani, hogy a hipotézis 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.

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

Feltételezzük, hogy a felnőtt férfiak magasabbak, mint a felnőtt nők.

Munkahipotézis (H_1): A felnőtt férfiak átlagos magassága bizonyos valószínűséggel nagyobb, mint a nők átlagos magassága.

Kiindulási vagy nullhipotézis (H_0): A felnőtt nők és felnőtt férfiak függőleges testmérete **egyforma**. Miért?

Hipotézisállítás

Feltételezzük, hogy a felnőtt férfiak magasabbak, mint a felnőtt nők.

Munkahipotézis (H_1): A felnőtt férfiak átlagos magassága bizonyos valószínűséggel nagyobb, mint a nők átlagos magassága.

Kiindulási vagy nullhipotézis (H_0): A felnőtt nők és felnőtt férfiak függőleges testmérete **egyforma**. Miért? „Keress fekete hattyút!”

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.

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

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

Változók

- ▶ **Kvalitatív:** valamilyen tulajdonság (februárban született, nő, romungró, ige stb.).
- ▶ **Diszkrét:** megszámlálható, véges (hibák száma egy tesztben, életkor években megadva).
- ▶ **Folytonos:** adott intervallumban akármilyen valós szám.
- ▶ **Kategóriák vagy csoportok:** változók összefoglalása (pl. 25 és 34 év között). Egyszerűbb kezelés, de információvesztés.

Skálatípusok

Nominális skála: változó értékei megkülönböztethetők, de semmilyen viszonyban nem állnak egymással. (Nem, vallás, hajszín, szófaj.)

Ordinális skála: értékek rangsorolhatóak, de az egyes elemek távolsága nem egyenlő vagy nem értelmezhető. (Végzettség, osztályzat.)

Metrikus skálák: egy adott mértékegység többszöröse. A mértékegység részei és többszöröse is értelmezhetőek, tehát a távolság értelmezhető és összehasonlítható.

Intervallumskála: nullpontja önkényes (pl. Celsius fok), mérőszámok különbsége igen, de aránya nem értelmezhető. **Húsztíz fok nem kétszer olyan meleg, mint tíztíz fok.**

Arányskála: nulla pont fizikailag definiált, arányok is értelmezhetőek (távolság, tömeg, energia, Kelvin fok).

Középértékek

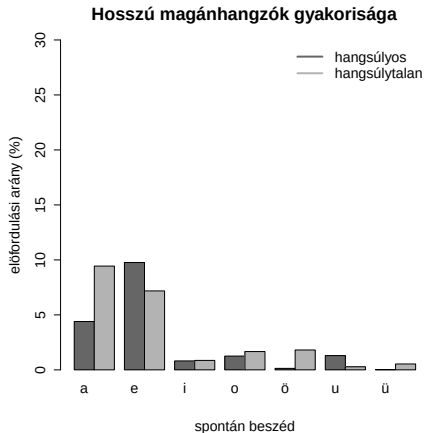
- ▶ **Módusz:** a mintában a legnagyobb gyakorisággal előforduló adatérték.
- ▶ **Medián:** a növekvő sorba rendezett adatok közül a középső. Ha az n mintaelemszám páros, a két középső érték átlaga.
- ▶ **Átlag:** mintabeli adatok számtani közepe.

Nominális skála: módusz, ordinális skála: medián, metrikus skála: átlag.

Alacsonyabb skálára érvényes statisztikai módszerek mindig alkalmazhatóak a magasabbakra, de információvesztéssel jár(hat)nak.

Középértékek: modusz

A mintában előforduló leggyakoribb kategória. Minden skálatípusra alkalmazható.



Középértékek: medián

Egy sorozat középső eleme. Ha az n elemből álló sorozat elemszáma páros, akkor a medián a két középső elem átlaga. Nominális adatokra NEM számolható medián.

Hány ismerőse van a Facebook-os ismerőseimnek?

11 véletlenszerűen kiválasztott ismerős ismerőseinek száma:

546 388 724 269 113 467 682 178 149 382 196

Sorba rendezett értékek:

113 149 178 196 269 382 388 467 546 682 724

Középső érték: 6. elem = 382.

12 ismerős esetén a 6. és 7. elem átlaga a medián.

Középértékek: átlag (számtani közép)

Az összes érték átlaga, azaz az értékek összege osztva az elemszámmal.

A Facebook-os ismerőseim ismerőseinek száma:

$$\text{átlag} = (546+388+724+269+113+467+682+178+149+382+196)/11 = 382,1818$$

Az átlag egy statisztikai modell, nem feltétlenül tükröz reális adatokat. Senkinek nincs 0,1818-ad ismerőse.

Fontos: átlagot csak parametrikus adatokra lehet számolni, amelyek ekvidisztánsak, azaz egyenlő távolságra vannak egymástól.

Iskolai osztályzatok?

Középértékek: átlag (számtani közép)

Az összes érték átlaga, azaz az értékek összege osztva az elemszámmal.

A Facebook-os ismerőseim ismerőseinek száma:

$$\text{átlag} = (546+388+724+269+113+467+682+178+149+382+196)/11 = 382,1818$$

Az átlag egy statisztikai modell, nem feltétlenül tükröz reális adatokat. Senkinek nincs 0,1818-ad ismerőse.

Fontos: átlagot csak parametrikus adatokra lehet számolni, amelyek ekvidisztánsak, azaz egyenlő távolságra vannak egymástól.

Iskolai osztályzatok? Az 1-es és 2-es különbsége nagyobb, mint a 4-esé és 5-ösé, ezért nem ekvidisztáns skála.

Medián vagy átlag?

Képzeljük el, hogy a 11 ismerősünk közül valaki csak tegnap iratkozott fel a Facebook-ra, ezért még csak 11 barátja van. Egy másik ismerős híres színésznő, és 5439 ismerőse van.

átlag =

$$(11+149+178+196+269+382+388+467+546+682+5429)/11 = 791,5455$$

Ha 11 helyett 111 ismerős ismerőseit vizsgáljuk, kiderül, hogy kevés embernek van 11 vagy 5429 ismerőse - ezek szélső vagy extrém értékek (ld. később).

Az adatok eloszlását érdemes ábrázolni, illetve az átlag mellett a mediánt is kiszámolni, ami robusztusabb, mert kevésbé érzékeny a szélső értékekre.

A fenti adatok mediánja szintén 382, ami realisabban tükrözi az adatok közéértékét.

R

Eredeti programnyelv S, ebből licenz 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)
```

⇒ 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 a móduszt, mediánt, átlagot, és a dobások összegét.

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

Feladatok

Dobjunk 100-at, majd 1000-et. Hasonlítsuk össze a középértékeket.