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: klaszterelemzés, döntési fák.

Irodalom

Órák anyaga:
`clara.nytud.hu/~mady/courses/statistics/2021`

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 tejcsoki.
- 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 leírás

Kvantitatív adatok: megszámolható vagy mérhető egységek.

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

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. A múltra és egyes 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 ritkábban tüsszentenek nyilvánosan, mint korábban.

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 sokan életben maradtak.
- 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 időtartama.
- 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 adatok neve (pl. szúnyogok száma a tesztelt területen) **függő változó**, mert függ az adatgyűjtés körülményeitől.

Háttér

Fallibilizmus (Popper): egy állítás igazolásából 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.
Az Ödönöket reggel 6-ra hívjuk be tesztelni, a más nevűek választhatnak időpontot ☺

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, pl. Magyarország összes 1. éves egyetemistája.

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

Hipotézisek

Hipotézis: feltételezés. Itt: előzetes megfigyelésen alapuló válasz tudományos kérdésfeltevésekre.

Kísérleti hipotézis: a változók viszonyára vonatkozó állítás, pl. 2020-ban több volt a szúnyog, mint korábban.

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

Hipotézisállítás

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

Kiindulási vagy nullhipotézis (H_0): A Szerecsen és a Hagyományos kávé fogyasztása után a tesztalanyok azonos ideig maradnak frissek.
Ellenhipotézis vagy alternatív hipotézis (H_1): a Szerecsen kávét fogyasztók később érzik magukat újra fáradtnak.

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

Vagyis: azt próbáljuk bizonyítani statisztikai modellekkel, hogy NINCS különbség két vagy több mintára vonatkozó mérések között. Ha a modell azt mutatja, hogy nagyon valószínűtlen, hogy a minták egyazon populációhoz tartozzanak, akkor elvetjük a nullhipotézist, és fenntartjuk az alternatív hipotézist.

Közös kísérlet

Egy külföldi kolléga megfigyelése szerint a magyarban van egy olyan szabály, hogy a *viszontlátásra* köszönésre *viszlát* a válasz. Ha viszont valaki a *viszlát* kifejezést használja, a beszélgetőpartner válasza *viszontlátásra* lesz.

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?

Feladat február 15., hétfő estig: adatok elküldése emailben.

R

Eredeti programnyelv S, ebből licensz 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, 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,...)`

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].

Karaktervváltozók 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)
```

$\Rightarrow$ 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.