

A statisztika alapelvei, hipotézisek

Félév beosztása

- ▶ Hipotézisek, skálatípusok.
- ▶ Eloszlások, szórás.
- ▶ Korrelációs számítás.
- ▶ Normális eloszlás, standard normális eloszlás.
- ▶ Valószínűség, mintavétel, konfidenciatartomány, 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 a mady@nytud.hu emailcímre. Fontos: legfeljebb két házi feladat hiányozhat félév végéig!
- ▶ Zárthelyi dolgozat a félév közepén és a teljes anyagból a félév végén.

Irodalom

Órák anyaga:

phon.nytud.hu/mady/courses/statistics/2025

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 többet beszél, mint Feri bácsi.
- ▶ Az angolok többet olvasnak, mint a franciák.
- ▶ Az éves átlaghőmérséklet ma magasabb, 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 felsorolása, az emberek undora a patkányoktól és a pókoktól, a nők és háziasszonyok szociális helyzetének jellemzése.

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 viszont 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 színesebb 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 száraz, nedves, hideg vagy meleg volt az idő.
- ▶ Irtási akciók gyakorisága.
- ▶ Az egyes irtószerek használatának mértéke korábbi évekkel összehasonlítva.

Független változó: a kísérletező által kontrollált vagy manipulált változók, pl. éghajlat (nedves, meleg), irtások száma, irtószer mennyisége. Egyéb elnevezések: faktor, tényező, magyarázó változó.

Függő változó: a független, kontrollált változóktól függ. Ezek a kísérlet során gyűjtött kvantitatív adatok (pl. szúnyogok száma a tesztelt területen).

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 nevűek 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 sejtethő, 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 éberen 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, azaz a Szerecsen és a Hagyományos kávé után egyaránt kb. 2 órán át maradunk éberek. Ha a statisztikai modell azt mutatja, hogy nagyon kicsi a valószínűsége, hogy a két kávé hatása azonos lenne a fogyasztóra, 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 őszt.

- ▶ Kiktől gyűjtsünk adatokat?
- ▶ Hogyan érhetjük el, hogy kiegyensúlyozott legyen a mintavétel?

Válaszokat most a saját ismeretségi körből gyűjtünk. Próbáljunk eltérő korosztályú és nemű embereket megkérdezni.

Az adatgyűjtés módja: fejenként öt ismerőst megkérünk, hogy 1 és 20 pont között adja meg, mennyire szereti a tavaszt. Öt másik ismerőst az őszelel kérdezőnk. A válaszadóval kapcsolatos adatokat és a pontszámot írásban rögzítjük.

Házi feladat: legalább tíz válasz gyűjtése.

Táblázat készítése Excelben vagy LibreOffice Spreadsheet formátumban.

Oszlopnevek:

1. KERD: a kérdező azonosítója (pl. monogram), mind a tíz megkérdezettnél azonos.
2. NEM: a beszélgetőpartner neve (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.

statisztika elnevezésű mappa létrehozása a saját gépen.

Excel fájl mentése csv-fájlként. Lépések: Mentés másként <

Fájlnev: <valaszadovezetekneve>.csv

Fájl típusa: CSV (pontosvesszővel tagolt) (*.csv)

Mentés a statisztika nevű mappába. Előguró figyelmeztetésre a válasz: IGEN.

Fontos: az oszlopok neve és sorrendje pontosan ez legyen. Az egyes megkérdezetteknel megadott betűk a mintát követve, nagybetűvel szerepeljenek. Ékezet és szóköz ne forduljon elő se a fájlnevben, se a táblázatban.

Beküldés: mady@nytud.hu, 2025. február 17., hétfő estig.

Jövő héttől az R statisztikai programnyelvet fogjuk használni, ezért az órákra mindenki hozzon laptopot.

Letöltés és telepítés: <https://cran.r-project.org>

Bármelyik tükröt (mirror) lehet választani a bal oldalon felül látható CRAN menüpontból. A magyar tükör közvetlen linkje: <https://cran.dnss.org>

A szoftver Windowsra, Macre és Linuxra elérhető. A megfelelő operációs rendszerhez tartozó telepítő fájlt kell letölteni.

Az R-4.4.2-win.exe fájl futtatásakor elég az alapértelmezett beállításokat elfogadni. Megnyitáskor egy fehér terminálablak látszik, ebben fogunk dolgozni jövő héttől.