A legfontosabb félreértések a p-érték kapcsán

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 12.

A legfontosabb félreértés a p-érték kapcsán

- Az első és legfontosabb félreértés: a p-érték *csak és kizárólag* a mintavételi hibát jellemzi, a világon semmit, de semmit nem mond a nem-mintavételi hibákról
- …miközben lehet, hogy egy kutatásban a nem-mintavételi hibák dominálnak!
- (Ebben az esetben a p-értékek közlése akár kifejezetten félrevezető is lehet!)
- Közel nem igaz tehát, hogy a p-érték (és társai) valamiféle univerzális metrikái lennének egy kutatás potenciális hibázásának

Második legfontosabb félreértés: egy kis motiváció

- Terroristák a városban…
- Mellesleg mitől medikális ez a példa?
- …ritka betegségek szűrése!
- Na és mi köze a p-értékekhez?
- Íme…

A második legfontosabb félreértés: p-érték = hibavalószínűsége

- Az egyszerűség kedvéért a kísérletünknek csak két kimenete lehet: a gyógyszer hatásos és hatástalan
- Kétféle hiba van, legyen a hatástalan hatásosnak minősítése 5, a hatásos hatástalannak minősítése 20% valószínűségű
- Az utóbbi azt jelenti, hogy 80% az erő…

- …az előbbi épp a választott szignifikanciaszint
- A hatástalanság a priori valószínűsége 90%

Mi az a prior valószínűség?

- „Előzetes” valószínűség: *még mielőtt* egyáltalán megkezdenénk a vizsgálatot, mennyi a valószínűsége annak, hogy hatástalan a gyógyszer
- Hogy micsoda? Hát ennek mi értelme? Pont azért csináljuk a kutatást, mert nem tudjuk, hogy hat-e…!
- Korábbi vizsgálatok, kutatáson kívül egyéb okok, például más területek eredményei (biológiai megfontolások), stb.
- Ha valaki szeretné valami konkréthoz kötni, akkor gondoljon arra, hogy egy számos már sikeres gyógyszerrel rendelkező gyógyszercsalád egy minimálisan módosított új tagjánál ez a valószínűség nagy, annak viszont, hogy a hiperpulzatív mágneses térrel kvantumtranszformált rezgőkristály hat, ez a prior valószínűség csekély

Egy kis fejszámítás

- Hogy matematikai részletek nélkül megindokoljam ezt, végezzünk egy gondolatkísérletet
- 1000-szer szimuláljuk ezt a világot (avagy 1000 párhuzamos univerzumot tekintünk)
- $1000 \cdot 0{,}1 = 100$-szor hatásos lesz a gyógyszer…
  - … ebből $100 \cdot 0{,}2 = 20$-szor hatástalannak minősítjük, a maradék 80 esetben hatásosnak
- $1000 \cdot 0{,}9 = 900$-szor hatástalan lesz a gyógyszer…
  - … ebből $900 \cdot 0{,}05 = 45$-szor hatásosnak minősítjük, a maradék 855 esetben hatástalannak
- Összességében $80 + 45 = 125$ esetben lesz hatásosnak minősítve a gyógyszerünk
- A hibarány, tehát, hogy a hatásosnak minősítések közül mekkora arányban hatástalan valójában a gyógyszer $45/125 = 36\%$ – *marhára nem* 5%!

Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014 Nov 19;1(3):140216.

A hatás

Power of study (proportion (%) of time we reject null hypothesis if it is false)	Percentage of "significant" results that are false positives		
	P=0.05	P=0.01	P=0.001
80% of ideas correct (null hypothesis false)			
20	5.9	1.2	0.10
50	2.4	0.5	0.05
80	1.5	0.3	0.03
50% of ideas correct (null hypothesis false)			
20	20.0	4.8	0.50
50	9.1	2.0	0.20
80	5.9	1.2	0.10
10% of ideas correct (null hypothesis false)			
20	69.2	31.0	4.30
50	47.4*	15.3	1.80
80	36.0	10.1	1.10
1% of ideas correct (null hypothesis false)			
20	96.1	83.2	33.10
50	90.8	66.4	16.50
80	86.1	55.3	11.00

*Corresponds to assumptions in table 2.

Azaz: elborult ötleteknél még egy kimondottan alacsony p-érték is jelentheti azt, hogy valószínűleg *nem* vethető el a nullhipotézis (az ilyeneknél nagyon erős bizonyíték kell)!

Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ. 2001 Jan 27;322(7280):226-31.

A probléma oka

- A kutya tehát ott van elásva, hogy mi a prior valószínűsége
- A p-érték „előnye", hogy ez nem kell hozzá
- De emiatt nem is azt méri, amit sokan gondolnának!
- Azt is lehet, de ahhoz kell a prior valószínűség: Bayes-tétel

$$\mathbb{P}\left(H_0|\text{Minta}\right) = \frac{\mathbb{P}\left(\text{Minta}|H_0\right) \cdot \mathbb{P}\left(H_0\right)}{\mathbb{P}\left(\text{Minta}\right)}$$

- Út a Bayes-faktorokhoz

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999 Jun 15;130(12):995-1004.

Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999 Jun 15;130(12):1005-13.

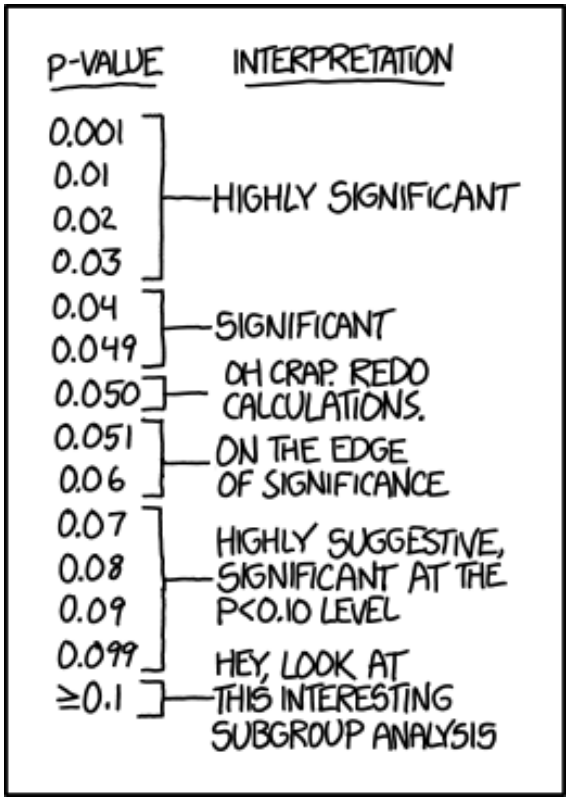
Cromwell-elv

- De: még a legelborultabb ötletekhez se rendeljünk 0 prior valószínűséget, rendeljünk nagyon picit, de ne 0-t
- „Ahhoz se 0 prior valószínűséget rendeljünk, hogy sajtból van a Hold, különben egy hadseregnyi, sajttal visszatérő űrhajós sem fog meggyőzni minket" (Dennis Lindley)
- Tehát ahhoz, hogy hat a hiperpulzatív mágneses térrel kvantumtranszformált rezgőkristály se 0 prior valószínűséget rendeljünk! Rendeljünk 10^{-10}-t, 10^{-100}-t, 10^{-1000}-t, de *ne* nullát; ez a Cromwell-elv
- Ez tökéletesen racionális: nagyon erős bizonyíték kell, hogy elhiggyük, hogy hat (összhangban azzal, hogy az alapkutatási eredmények ennek ordítóan ellentmondanak), de közben nem is lehetetlen, tehát ha sorban meggyógyít mindenkit, akkor be fogjuk látni, hogy tényleg hat

Frekventista és bayes-i statisztika

- Az előzőekből látható, hogy miért gondolják a legtöbben, hogy jobb lenne bayes-i keretet használni az orvosi vizsgálatok kiértékelésére
- Mégis, szinte kizárólagos a frekvencionista elv használata:
  - A bayes-i statisztika számításigényes (ma már nem lényeges ellenérv, van számítási kapacitás könnyen elérhetően)
  - Apróbb filozófiai és statisztikai ellenérvek (pl. a valószínűség bayes-i interpretációja vagy, hogy minden hipotézist azonosítani kell)
  - De a legeslegfontosabb: az orvosok gyomra nem veszi be a prior valószínűség használatát (mondván, hogy az „szubjektív" – noha a frekvencionista keretnek nem kevesebb feltevése van, csak azok nem ilyen látványosak)
  - És végül a tehetetlenség: a frekvencionista iskola rendkívül jól kidolgozott elméletileg és gyakorlatilag is, kitűnő számítógépes támogatással rendelkezik, meg egyáltalán, 100 éven keresztül megszámlálhatatlan sok vizsgálatban ezt alkalmazták, ezt tanulta szinte mindenki

A p-értékek minősítése, a nagyságának a jelentősége



Aki azt gondolná, hogy ez egy vicc: `https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/`

A harmadik legfontosabb félreértés: a bizonyíték hiánya nem a hiány bizonyítéka

- A kétféle döntés nem szimmetrikus: a hatástalanságot erősen el tudjuk vetni, de olyan nincs, hogy erősen elfogadjuk
- Ha 100 kezeltből 0 hal meg, 100 kontrollból 100, akkor nagyon erősen (megj.: nem biztosan, ugye!) el tudjuk vetni a feltevést, hogy a gyógyszer hatástalan
- De ha 100 kezeltből és 100 kontrollból egyaránt 50 hal meg, még akkor sem mondhatjuk, hogy nagyon erősen elfogadjuk, hogy hatástalan
- Miért? Mert erő is van a világon! Simán lehet hatása, de 2×100 betegből kimutathatatlan (gondoljuk arra, ha csak 2×10 betegen találtuk volna ugyanezt!)
- A „nem tudtuk bizonyítani, hogy van hatás” tehát *nagyon nem ugyanaz*, mint az, hogy „bizonyítottuk, hogy nincs hatás”!

Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995 Aug 19;311(7003):485. Alderson P. Absence of evidence is not evidence of absence. BMJ. 2004 Feb 28;328(7438):476-7.

Egy példa: dexmedetomidine jót tesz lélegeztetett szeptikus betegeknél?

Key Points – Meaning

Treatment with dexmedetomidine in patients with sepsis did not improve either ventilator-free days or 28-day mortality.

Számszerűen

Mortality at 28 days was not significantly different in the dexmedetomidine group vs the control group (19 patients [22.8%] vs 28 patients [30.8%]; hazard ratio, 0.69; 95% CI, 0.38-1.22; P = .20).

Jobbik eset, mert ők legalább tudják

Among patients requiring mechanical ventilation, the use of dexmedetomidine compared with no dexmedetomidine did not result in statistically significant improvement in mortality or ventilator-free days. However, the study may have been underpowered for mortality, and additional research may be needed to evaluate this further.

Kawazoe Y, Miyamoto K, Morimoto T. Effect of Dexmedetomidine on Mortality and Ventilator-Free Days in Patients Requiring Mechanical Ventilation With Sepsis: A Randomized Clinical Trial. JAMA. 2017 Apr 4;317(13):1321-1328.

Az „underpowered" kutatások problémái

- Első ránézésre a szponzor baja, ha a kutatás ereje kicsi (hat a gyógyszer, mégis kidobják az ablakon a pénzt, hogy ezt kimutassák)
- Tehát az orvosi tudásunk csak olyan irányban lehet tőle hibás, hogy valamiről azt hisszük, hogy nem hat, pedig igen
- Ez nem így van:
  - Fontos lenne tudni, hogy egyáltalán volt-e esély kimutatni valamit – ugyanis nem csak hatásosságot vizsgálunk, hanem biztonságosságot is!
  - Még a hatásosságnál sem teljesen így van: a kis erejű vizsgálatok, ha találnak is hatást, az várhatóan nagyobb lesz (az ilyenekben nagyobb ingadozás, a hatásosság kimutatásához nagyobb hatás kell: a fals hatás kimutatásának a valószínűsége így ugyanannyi lesz, csak épp ha megtörténik, akkor a kimutatott hatás nagyobb lesz)
  - Azaz felülbecsüljük a hatást!

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafo MR. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013 May;14(5):365-76.

A p-érték aszimmetriája

- Ha nagyon kicsi a p-érték, az nagyon erősen amellett szól, hogy van hatás; minél kisebb, annál inkább

- De ez fordítva nem igaz: a nagy p-érték (ad abszurdum a $p = 1$) sem mond semmit arról, hogy *mennyire* erősen gondoljuk azt, hogy nincs hatás!
- Gondoljunk bele: 347/5033 (6,89%) vs. 351/5033 (6,98%), p=0,87 és 5/19 (26,32%) vs. 4/14 (28,57%), p=0,89 – ez a kettő biztos ugyanaz?!
- A dolog tehát aszimmetrikus, a hatástalanságban való bizonyosságot nem jellemzi a p-érték
- Bayes-i keretben jól kezelhető
- Illetve megint csak: ha nem is térünk át a bayes-i keretre, legalább preferáljuk a konfidenciaintervallumokat a p-értékekkel szemben

Hoekstra R, Monden R, van Ravenzwaaij D, Wagenmakers EJ. Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. PLoS One. 2018 Apr 25;13(4):e0195474.