Zašto većina eksperimentalnih istraživanja u "soft" znanostima nije istinita

Zvonimir Šikić[1]

Većina eksperimentalnih istraživanja u „soft" znanostima svoju potvrdu vidi u *p*-vrijednosti manjoj od 5 %. To je vjerojatnost dobivenog ishoda eksperimenta pod uvjetom da je hipoteza istraživanja netočna (vjerojatnost tzv. lažno pozitivnog rezultata). Zvuči razumno. Naime, ako je hipoteza netočna, dogodilo se nešto što je jako malo vjerojatno, a to znači da je hipoteza vrlo vjerojatno točna (mnogi čak misle da je hipoteza tada 95 % točna). Uostalom, tu je metodu predložio R. Fisher [3], otac moderne statistike.

Ipak, pokušaji repliciranja eksperimentalnih rezultata s *p*-vrijednostima manjim od 5 %, koji su objavljeni u vodećim psihološkim časopisima, utvrdili su da ih je uspješno replicirano manje od polovice [2]. Sve je rađeno vrlo pažljivo i uz pomoć izvornih autora. Slični neuspjesi repliciranja utvrđeni su i u drugim područjima. Pokušaji repliciranja „krucijalnih" rezultata u istraživanju raka pokazali su da ih je uspješno replicirano 11 % [1]. O tome je još 2005. pisao J. Ioannidis u radu pod naslovom „Why Most Published Research Findings are False" [6]:

Istraživanja nisu najprikladnije predstavljena i sažeta p-vrijednostima, ali, nažalost, raširenо je mišljenje da bi medicinski istraživački članci trebali biti interpretirani samo na temelju p-vrijednosti.

O čemu je tu riječ? Prije svega uočite da 5 % mogućnosti za lažni rezultat znači da će se on u prosjeku pojaviti jednom u 20 eksperimenata. Istraživači mogu eksperiment ponavljati dok ne dođu do željenog rezultata (u prosjeku će trebati 20 pokušaja, ali katkada će ih biti dovoljno svega par). Ako i isključimo svjesnu prijevaru, istraživači neuspjele pokušaje mogu odbaciti nesvjesno, kao neuspjele zbog raznih tehničkih razloga. Osim toga, ako ne dođu do *p*-potvrde svoje hipoteze, istraživači rezultat neće objaviti, ali će prosječno svaki 20. slučajno doći do *p*-potvrde i svoj će rezultat objaviti. Na kraju vidimo samo ono što je objavljeno. Moguće je da istraživač nakon eksperimenta svoju hipotezu zamijeni novom hipotezom za koju je $p < 5$ %. To je matematički nekorektno, iako mnogi istraživači to ne znaju – katkada je neznanje

[1]Zvonimir Šikić, Sveučilište u Zagrebu

put do uspjeha, što objašnjava čestu pojavu da ljudi odbijaju da im se nešto objasni. Aplikaciju za razne vrste ovakvih *p*-hackinga možete naći na blogu N. Silvera [8].

Na sreću, danas postoji jak pokret protiv slijepe upotrebe *p*-vrijednosti. Sve više istraživača razumije da ne smiju zanemarivati neuspjele eksperimente, da je važno prijaviti i negativne rezultate, da ne smiju usklađivati svoje hipoteze s ishodima eksperimenta itd. Još je važnije da to razumije sve veći broj urednika znanstvenih časopisa i da u skladu s tim odlučuju što će objaviti. Mnogi časopisi više ne prihvaćaju malu *p*-vrijednost kao argument za valjanost istraživanja, npr. *Basic and Applied Social Psychology* [10], [4]. Respektabilna znanstvena udruženja, poput American Statistical Association, daju službene izjave o nekorektnom korištenju *p*-vrijednosti [11]. Medicinski časopisi objavljuju članke poput „A Dirty Dozen: Twelve *P*-Value Misconceptions" [5]. I tako dalje.

No, što je s Fisherovim argumentom iz prvog odlomka? On je jednostavno netočan. Fisher zapravo nije ni tvrdio da je točan, nego ga je samo naveo kao indiciju da je istraživanje obećavajuće. S vremenom se „obećavajuće" pretvorilo u „dokazano". Zašto je do toga došlo, složeno je pitanje povijesti odnosa frekvencijskog i bajesovskog razumijevanja vjerojatnosti, o čemu više uskoro. No, odmah možemo reći što slijedi iz Bayesove formule:

$$\frac{P(H/r)}{P(-H/r)} = \frac{P(H)}{P(-H)} \frac{P(r/H)}{P(r/-H)}.$$

$P(H / r)$ vjerojatnost je da je hipoteza H istinita, pod uvjetom da je eksperiment dao rezultat r (tzv. aposteriorna vjerojatnost od H ili njezin posterior).

$P(-H / r)$ vjerojatnost je da je hipoteza H neistinita, pod uvjetom da je eksperiment dao rezultat r (tzv. aposteriorna vjerojatnost od $-H$ ili njezin posterior).

$P(H)$ vjerojatnost je da je hipoteza H istinita, neovisno od eksperimenta (tzv. apriorna vjerojatnost od H ili njezin prior).

$P(-H)$ vjerojatnost je da je hipoteza H neistinita, neovisno od eksperimenta (tzv. apriorna vjerojatnost od $-H$ ili njezin prior).

$P(r / H)$ vjerojatnost je da eksperiment da rezultat r, pod uvjetom da je hipoteza H istinita.

$P(r / -H)$ vjerojatnost je da eksperiment da rezultat r, pod uvjetom da je hipoteza H neistinita.

Hipoteza H je rezultatom eksperimenta r to bolje potvrđena što je $P(H / r)$ veći u odnosu na $P(-H / r)$. Dakle, izraz s desne strane Bayesove formule treba biti što veći da bi hipoteza bila što vjerojatnija. Kada vjerujemo u pravilo $p < 5$ %, zapravo vjerujemo da je $P(r / -H) < 5$ % dovoljan uvjet da desna strana bude velika (naime, $P(r / -H)$ zapravo je *p*-vrijednost). No, očito je da veličina od $P(H/r)/P(-H/r)$ ne ovisi samo o $P(r / -H)$ pa ta vrijednost sama po sebi ne znači mnogo. Na primjer, ako je $P(-H)$ bitno veća od $P(H)$, onda čak i kada je $P(r / H)$ veća od $P(r / -H)$, činjenica

da je $P(r/–H) < 5$ % ne govori ništa o $P(H / r) / P(–H / r)$. Drugim riječima, ako prije eksperimenta imate valjane razloge da hipotezu H držite malo vjerojatnom (tj. da omjer apriornih vjerojatnosti $P(H) / P(–H)$ držite malim) onda $p < 5$ % nije nikakav argument.

To i nije naročito složen zaključak pa zaista čudi da je itko ikada pomislio da je $p < 5$ % nekakav argument. Problem je u tome što su mnogi (tzv. frekventisti) mislili da vjerojatnosti hipoteza ($P(H / r)$, $P(–H / r)$, $P(H)$ i $P(–H)$) uopće nemaju smisla.

Dakle, problem je u tome što je vjerojatnost nekog događaja i kako je odrediti? Na primjer, kako odrediti vjerojatnost da rezultat bacanja kovanice bude glava (analiza koja slijedi preuzeta je iz [9] pa su tamo i odgovarajuće reference).

Ako o kovanici ne znamo ništa, vjerojatnost glave može biti bilo što između $H = 0$ i $H = 1$ (između ta dva ekstrema leži i slučaj „poštene" kovanice $H = 1/2$). Naivni način određenja te vjerojatnosti jest da je identificiramo s relativnom frekvencijom. Dakle, kovanicu bacimo npr. 12 puta, pa ako je 3 puta pala glava, utvrdimo da je ta vjerojatnost 3/12, tj. $H = 1/4$. To je naivno jer svatko zna da i „poštena" kovanica s $H = 1/2$ može u 12 bacanja dati 3 glave. Dakle, kako na temelju obavljenog eksperimenta, 3 glave u 12 bacanja, možemo procijeniti je li vjerojatnost glave 1/2 (ili 1/4 ili što već želimo procijeniti)?

To Fisherova metoda testiranja hipoteza (npr. hipoteze da je $H = 1/2$) čini na sljedeći način:

1. Provedemo eksperiment, tj. kovanicu bacimo unaprijed zadani broj puta, na primjer 12 puta.

2. Odredimo prostor mogućih ishoda tog eksperimenta. U našem slučaju to je 2^{12} nizova glava i pisama duljine 12. Rezultat eksperimenta jedan je od tih ishoda, npr. PPGPPPGGPPPP.

3. Rezultat eksperimenta sažmemo u jedan numerički podatak r, u ovom slučaju to je broj glava koje se pojavljuju u ishodu eksperimenta. Taj sažeti podatak o rezultatu eksperimenta zove se statistika testa. U našem primjeru $r = 3$.

4. Izračunamo vjerojatnosti svih mogućih statistika, pod uvjetom da vrijedi naša hipoteza $H = 1/2$ koja se naziva nul-hipotezom. Sve te vjerojatnosti $P(r / H)$, za $r = 0, 1, 2, ..., 12$ i $H = 1/2$ označimo kraće s $P(r)$ i elementarnim računom nađemo

$$P(r) = \binom{12}{r}\left(\frac{1}{2}\right)^{r}\left(\frac{1}{2}\right)^{12-r},$$

što daje sljedeću razdiobu tih vjerojatnosti:

r	$P(r)$	r	$P(r)$	r	$P(r)$	r	$P(r)$
0	0.0002441406	3	0.053710938	6	0.225585938	9	0.053710938
1	0.0029296875	4	0.120849609	7	0.193359375	10	0.0161132813
2	0.0161132813	5	0.193359375	8	0.120849609	11	0.0029296875
						12	0.0002441406

5. Pogledamo rezultate koji su se mogli pojaviti, a koji su uz našu nul-hipotezu ekstremniji od rezultata koji se uistinu pojavio. Preciznije, to su rezultati čija je vjerojatnost manja ili jednaka vjerojatnosti rezultata koji se stvarno pojavio. Izračunamo vjerojatnost p da dođe do takvog ekstremnog rezultata. U našem se primjeru stvarno pojavilo $r = 3$ glave u 12 bacanja, pa su „ekstremni rezultati" $r = 0,1,2,3,9,10,11,12$. To su podvučene vrijednosti u gornjoj razdiobi. Dakle, p je zbroj svih podvučenih vrijednosti, tj. $p = 0.15 = 15\ \%$.

6. Nul-hipoteza se odbacuje ako je $p < 5\ \%$. Dakle, naša nul-hipoteza o „poštenoj" kovanici nije odbačena našim eksperimentom, jer je $15\ \% > 5\ \%$.

Neki statističari preporučuju 1 % ili čak 0.1 % kao kritičnu vrijednost p. Prihvaćena kritična vrijednost zove se razinom signifikantnosti testa, a za nul-hipotezu kaže se da je odbačena na toj razini signifikantnosti ako je vrijednost p manja ili jednaka prihvaćenoj kritičnoj vrijednosti.

Razmislimo što zapravo znači da je „nul-hipoteza odbačena na nekoj razini signifikantnosti". To znači da je rezultat eksperimenta pao u određeno područje koje je proglašeno „područjem odbacivanja". Što to govori o nul-hipotezi? Danas je standardno Neymanovo gledište da odbacivanje ili neodbacivanje nul-hipoteze nije nikakav matematički (čak ni induktivni) zaključak, nego tek „instrukcija o induktivnom ponašanju". Ako se, uz razinu signifikantnosti 1 %, ponašamo prema toj instrukciji, onda ćemo u prosjeku, na duge staze, istinitu hipotezu odbaciti (tj. učinit ćemo grešku I. tipa) ne više od jednom u 100 puta.

(Možemo se, kao J. Neyman i E. Pearson [7], brinuti i o greškama II. tipa, tj. o prihvaćanju neistinite hipoteze. Vjerojatnost greške II. tipa je vjerojatnost odbacivanja istinite alternativne hipoteze H_a prihvaćanjem neistinite nul-hipoteze H_0. Komplement razine signifikantnosti odbacivanja hipoteze H_a zove se snagom testa (dok se, u tom kontekstu, razina signifikantnosti odbacivanja nul-hipoteze H_0 zove veličinom testa). Idealno bismo trebali maksimizirati snagu i minimizirati veličinu testa. No, taj je ideal nekonzistentan. Smanjenje veličine testa sa sobom nosi povećanje njegove snage i obratno.)

Osim proizvoljnosti određenja „područja odbacivanja", nekonzistentnosti redukcije veličine i ekspanzije snage testa, te ispitivanja samo jedne hipoteze, Fisherova metoda ima i drugih problema. Na primjer, različito odabrane statistike testa mogu iz istog eksperimentalnog rezultata izvesti različite zaključke. To je zloglasni problem „izbora statistike".

Tu je i problem „pravila zaustavljanja". Razmislimo opet o kovanici bačenoj 12 puta, s ishodom koji sadrži 3 glave i 9 pisama. Gore opisani test ne možemo ni početi primjenjivati ako ne znamo je li eksperimentator planirao kovanicu baciti točno 12 puta. Što ako ju je planirao bacati do pojave 3 glave ili dok mu ne dosadi bacanje? Tada prostor mogućih ishoda iz 2. koraka izgleda drukčije, a s time i svi preostali koraci,

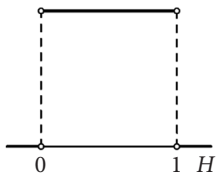
pa i konačni zaključak. Sama činjenica da su se u 12 bacanja pojavile 3 glave nije dovoljna. Moramo znati kako je zamišljen eksperiment koji je doveo do tog rezultata, sam rezultat nije dovoljan.

No, ključni problem Fisherove metode je da evaluira **samo jednu hipotezu**, tzv. nul-hipotezu (u našem primjeru „poštenje" kovanice), uzimajući u obzir **sve rezultate eksperimenta koji su se mogli dogoditi**. A nas zapravo zanima evaluacija **svih mogućih hipoteza** na temelju **jednog jedinog rezultata eksperimenta koji se stvarno dogodio**.

Zar ne postoji metoda koja evaluira ono što nas zanima? Postoji, i čak je povijesno prethodila Fisherovoj. Bitno je informativnija i matematički objašnjiva. Uveo ju je Laplace. Najprije ćemo pokazati kako se njome rješava naš problem testiranja kovanice, a zatim ćemo objasniti zašto ju je zamijenila neinformativna i matematički neobjašnjiva Fisherova metoda.

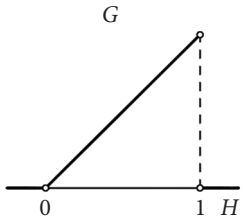
Laplace bi problem kovanice rješavao primjenjujući Bayesov teorem (ovdje nije bitno da razumijete tu metodu, bitno je da vidite što je njezin rezultat). Ako prije eksperimenta nemamo nikakvog razloga da preferiramo bilo koju vrijednost $H \in [0, 1]$ kao vjerojatnost glave, onda je vjerojatnost $P(H)$ (prije eksperimenta) uniformno distribuirana (sve vrijednosti od H jednako su vjerojatne):

$$P(H) = \begin{cases} 1 & 0 \leq H \leq 1 \\ 0 & \text{inače} \end{cases}$$



Bacimo li kovanicu jednom i padne li glava, Laplace bi iz Bayesove formule izračunao da je $P(H / G)$, tj. vjerojatnost hipoteze H (koja tvrdi da vjerojatnost glave iznosi H), nakon što je jednom pala glava, distribuirana na sljedeći način:

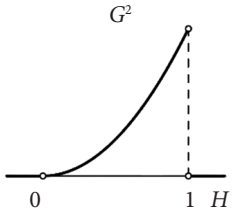
$P(H / G)$



Dakle, vjerojatnost hipoteze H uniformno raste kada H raste od 0 do1 (najmanja je za $H = 0$, a najveća za $H = 1$, ali i druge hipoteze H imaju svoje vjerojatnosti).

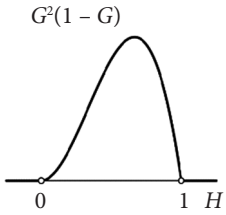
Bacimo li kovanicu još jednom i padne li opet glava, vjerojatnost hipoteze H nakon što je dva puta pala glava, $P(H / GG)$, distribuira se prema Bayesovoj formuli na sljedeći način:

$P(H / GG)$



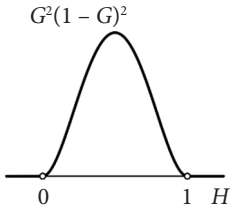
Padne li treći put pismo, vjerojatnost hipoteze H, nakon što je dva puta pala glava i jednom pismo, distribuira se prema Bayesovoj formuli na sljedeći način:

$P(H / GGP)$

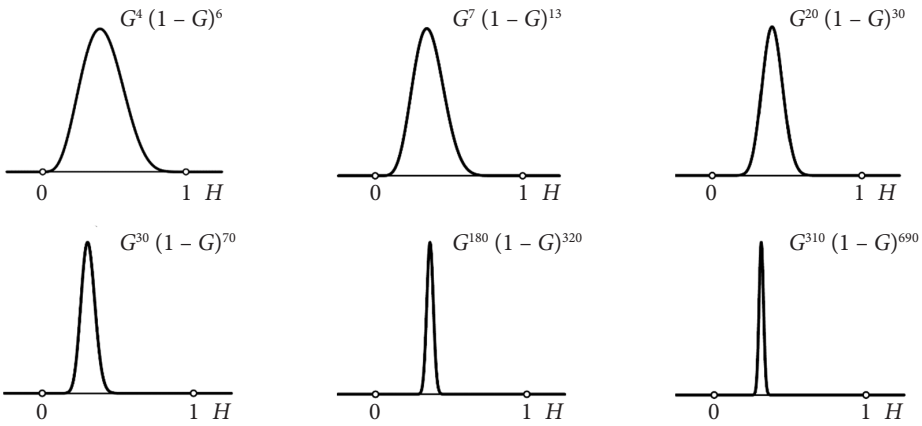


Još jedno pismo i vjerojatnost hipoteze H distribuira se na sljedeći način:

$P(H / GGPP)$



I tako dalje. Sljedeće slike pokazuju kako se distribucija vjerojatnosti hipoteze H mijenja sa sve više i više podataka o rezultatima bacanja kovanice. Položaj maksimalne vjerojatnosti sve se više stabilizira što je broj podataka veći. Osim toga, sa sve većim brojem podataka distribucija postaje sve uža i uža. Ako nakon 1000 bacanja padne 310 glava i 690 pisama, vjerojatnost da je $H = 1/4$ bit će daleko najveća (v. graf). Primijetite da je za svaki $H \in [0, 1]$ izračunata vjerojatnost toga H.



Što bi bilo da smo krenuli od neke druge početne distribucije hipoteza H, a ne od uniformne (npr. da smo prije eksperimenta izučavanjem kovanice zaključili da je $H > 1/2$ vjerojatnije nego $H < 1/2$)? Uz isti niz dobivenih glava i pisama dobili bismo isti rezultat. Naime, naše preduvjerenje o kovanici u početku bi snažno utjecalo na distribuciju od H (ako ste uvjereni da je kovanica „poštena", tj. da je $H = 1/2$, onda vas 3 glave u 12 bacanja neće pokolebati), no kako broj bacanja raste, rezultati bacanja nadjačaju sva preduvjerenja (30 000 glava u 120 000 bacanja pokolebat će vašu vjeru u $H = 1/2$). To je intuitivno jasno, a matematički slijedi iz Laplaceove primjene Bayesovog teorema.

Primijetimo još jednom da, za razliku od Fisherove metode koja odbacuje ili ne odbacuje samo jednu hipotezu (u našem primjeru hipotezu da je $H = 1/2$), Laplaceova metoda daje vjerojatnost za svaki H, dakle za svaku moguću hipotezu. To je točno ono što smo htjeli.

Kako je uopće došlo do toga da Fisherova ograničena, neprecizna i nematematička metoda zamijeni Laplaceovu generalnu, preciznu i matematičku metodu?

Laplaceova metoda temelji se na tome da su vjerojatnosti **racionalne procjene uvjerenja koje se temelje na raspoloživim podatcima**. Nije odmah jasno zašto bi tako shvaćene vjerojatnosti zadovoljavale uobičajene aksiome vjerojatnosti koje su Laplace i njegovi sljedbenici koristili i iz kojih slijedi Bayesov teorem. Ako vjerojatnosti shvatimo kao granične relativne frekvencije, te je aksiome lako dokazati, iako je sam pojam granične relativne frekvencije nekonzistentan. Kako bilo, statističari s kraja 19. stoljeća i s početka 20. stoljeća počeli su inzistirati na tome da se vjerojatnosti moraju shvaćati isključivo kao granične relativne frekvencije u tzv. slučajnim eksperimentima. Nažalost, takvo shvaćanje vjerojatnost hipoteze čini nelegitimnim pojmom. Hipoteza je istinita ili lažna, a ne nešto što može biti manje ili više vjerojatno. Dakle, $H = 1/2$ ili $H \neq 1/2$, i nema smisla govoriti o vjerojatnosti da je $H = 1/2$, a to je temelj Laplaceova pristupa. Rezultat je Fisherova metoda.

Spomenimo na kraju da od 50-tih godina 20. stoljeća **znamo** da vjerojatnost shvaćena kao **racionalna procjena uvjerenja koje se temelji na raspoloživim podacima** zadovoljava standardne aksiome vjerojatnosti. To je dokazano Coxovim teoremom (vidi [9]) i više nema razloga da ograničenu, nepreciznu i nematematičku Fisherovu metodu upotrebljavamo umjesto generalne, precizne i matematičke Laplaceove metode.

Literatura:

1. C. G. Begley & L. M. Ellis, Raise standards for preclinical cancer research, Nature vol. 483, 531–533 (2012.)
2. J. Bohannon, Many psychology papers fail replication test, vol.28, 910-911 (2015.)
3. R. A. Fisher, The Design of Experiments, Oliver & Boyd (1935.)

4. R. D. Fricker, K.Burke, X. Han, W. H. Woodall, Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban, The American Statistician, 73:sup1, 374-384, (2019.)

5. S. Goodman, A dirty dozen: twelve p-value misconceptions, Semin Hematol. vol. 45/3, 135-40 (2008.)

6. J. P. A. Ioannidis, Why most published research findings are false, PLoS Med. vol. 2/8, 0696-0701 (2005.)

7. J. Neyman, E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Phil. Trans. R. Soc. Lond. A. 231, 694–706 (1933.)

8. N. Silver http://fivethirtyeight.com/features/science-isnt-broken/#part2.

9. Z. Šikić, What Is Probability And Why Does It Matter, European Journal of Analytic Philosophy, 10/1, 21-43 (2014.)

10. D. Trafimow, M. Marks, Editorial, Basic and Applied Social Psychology, 37:1, 1-2 (2015.)

11. R. L. Wasserstein & N. A. Lazar, The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133 (2016.)