Zašto pet odsto?

Damjan Krstajić

(objavljeno 10. novembra 2018. u Politikinom Kulturnom dodatku)

Na čajanci u okolini Kembridža pre manje od sto godina, jedna dama je izjavila da ona razlikuje da li se u šolji engleskog čaja prvo sipalo mleko pa onda čaj, ili obrnuto. Kad se pomešaju mleko i čaj rastvor bi trebao da bude isti i naravno bilo je onih koji joj nisu poverovali. Organizovan je eksperiment da se proveri. Ova čajanka je ostala upamćena, jer je mladi Ronald Fišer bio tamo prisutan i koristio je kasnije ovaj primer u svojoj knjizi. Kako najbolje dizajnirati test da se proveri da li dama nagađa ili stvarno zna? Ako nagađa, može slučajno da pogodi.

Ako generalizujemo ovaj problem, možemo reći da imamo rezultate nekog opita i hipotezu i zanima nas da li dobijeni nalazi podržavaju tu hipotezu. U literaturi se verovatnoća da se takvi nalazi dobiju pod uslovom da je hipoteza tačna naziva p-vrednost. Fišer je u svojim delima spominjao je da ako je p-vrednost manja od 5% ili 1% da su to onda po njemu male verovatnoće, ali da on preferira 5% kao granicu. Koliko sam upoznat nigde eksplicitno nije rekao kako naučnik treba da koristi p-vrednost, već je upotreba zavisila od slučaja do slučaja. Po njemu, na osnovu p-vrednosti možemo da odbacimo neku hipotezu, ali nema automatskog prihvatanja nekih drugih hipoteza.

Jerži Nejman (Jerzy Neyman) i Igon Pirson (Egon Pearson) su zauzeli drugačiji pristup i kreirali teoriju za testiranje hipoteza gde imamo dve hipoteze: A i ne-A. U literaturi one su poznate pod imenima nulta i alternativna hipoteza. Cilj je da se na osnovu rezultata opita prihvati jedna od te dve. To znači da postoji i moguća greška da kad prihvatimo jednu od njih da je druga u stvari tačna. Lepota njihove teorije je što ako unapred odredimo verovatnoće obe greške, koje ne moraju biti iste, može se izračunati granična vrednost neke statistike za rezultate opita, pa u zavisnosti od toga da li su rezultati ispod ili iznad nje, zavisi koju ćemo hipotezu prihvatiti. U njihovom pristupu nema p-vrednosti, već samo verovatnoće grešaka za obe hipoteze koje smo odredili pre početka opita.

Za mene je Nejman-Pirson teorema mesto gde teorijska statistika dodiruje umetnost. U situacijama kad na osnovu neke statistike treba da se odlučimo izmedju dva izbora ona je moćan alat. Na primer, kad treba da testiramo da li je novi lek bolji od starog, mi pre početka trajala moramo da odredimo neko pravilo i granicu na osnovu kojih ćemo kasnije bazirati buduću odluku. Pravilna primena Nejman-Pirson teoreme znači da ako prihvatimo hipotezu da je nov lek bolji od starog da smo onda svesni da postoji verovatnoća da smo pogrešili i da pratimo i proveravamo tu odluku.

Fišer je žestoko kritikovao ovaj pristup i ponavljao da ne možemo na osnovu dobijenih nalaza da odbacimo jednu hipotezu i automatski prihvatimo suprotnu. Nejman i Pirson su se argumentovano branili. Na žalost, to je rezultiralo da je danas u opštoj upotrebi mešavina njihovih pristupa sa kojom, siguran sam, niko od njih se ne bi složio. Pojednostavljeno, izračuna se p-vrednost sa hipotezom A i ako je ona manja od 5%, onda se prihvata hipoteza ne-A. Osim što nije ono što su oni zastupali, gde je ovde problem?

Zašto 5%? A ne na primer 3% ili 6%? To što je Fišeru pre manje od sto godina verovatnoća da se nešto desi bila značajno mala ako je ona manja od 1/20=0.05 to ne znači da je to u kamenu zapisano! Takođe, sistem je rigidan, jer ako dobijete p-vrednost 5.01%, onda se prihvata hipoteza A, a ako je p-vrednost 4.99%, onda se prihvata hipoteza ne-A. U praksi ova razlika u p-vrednostima može da bude razlika u tome da li će se nečiji naučni rad objaviti ili ne.

Dodatno komplikuje i to što naučnici izgleda ne razumeju dobro p-vrednost. Stivn Gudman (Steven Goodman) je pre deset godina objavio rad u kojem je naveo 12 prisutnih zabluda u vezi p-vrednosti, a najčešća je ona gde se misli da je p-vrednost verovatnoća hipoteze A. Da ponovimo, nakon završenog opita i dobijenih rezultata, p-vrednost je verovatnoća da se dobiju ti rezultati ako je hipoteza A tačna, a to nije isto što i verovatnoća hipoteze A!

Pre dve godine je Američko statističko društvo izdalo saopštenje o statističkoj značajnosti i p-vrednosti, gde objašnjavaju šta jeste p-vrednost (njihova neformalna definicija je komplikovanija i tačnija od moje u tekstu!) i šta nije. Po njima upali smo u začarani krug. Zašto se u školama podučava granična p-vrednost 5%? Zato što je naučna zajednica koristi i urednici časopisa to traže. Zašto toliko mnogo ljudi i dalje upotrebljava p-vrednost 5%? Zato što su tako učili u školama!

Nedavno su 72 vodeća svetska metodologa (Benjamin *et al.*) objavila rad u Nature Human Behaviour u kojem predlažu da se maksimum za graničnu p-vrednost smanji sa 5% na 0.5%. Njihov zajednički stav je da vodeći uzrok trenutne krize reproducibilnosti u nauci nije na adekvatan način sagledan. Statistički standardi uz pomoć kojih se u mnogim naučnim disciplinama tvrdi da je došlo do otkrića su za njih previše niski. Povezivanje statističke značajnosti sa p-vrednost od 5%, po njima, dovodi do velikog procenta lažno pozitivnih rezultata u radovima.

Iako među ovim naučnicima ima onih koje veoma cenim, usuđujem se reći da mi ovo više izgleda kao glasan vapaj nego kao rešenje. Po meni, dok god naučnici ne budu proveravali jedni drugima rezultate, bilo kakvo smanjenje granične p-vrednosti će samo u početnom periodu smanjiti priliv radova sa lažno pozitivnim rezultatima. Valja imati na umu Gudartov zakon (Charles Goodhart) koji kaže:„*Kad neka mera postane cilj, onda prestaje da bude dobra mera.*“

Mišljenja sam da je Fišer u osnovi bio u pravu. Kao što su neki pronalasci u fizici doprineli stvaranju atomske bombe, a da pronalazačima to nije bio cilj, tako ideja da se uz pomoć statistike može odbaciti jedna hipoteza i automatski prihvatiti suprotna se otela kontroli i u upotrebi je nešto što više liči na birokratiju nego na nauku.

A šta je sa damom sa početka teksta? Fišer nije spominjao da se taj događaj ikad desio, već ga je koristio kao misaoni eksperiment. Međutim, drugi prisutni su potvrdili da je tada organizovan eksperiment sa 8 šolji engleskog čaja, gde je u nekim prvo sipano mleko, a u drugim prvo čaj. Dama je bila u pravu u vezi svake šolje čaja!

Reference koje podržavaju činjenice spomenute u članku

1. Događaj o dami koja ispija čaj je ispričan u knjizi The Lady Tasting Tea

https://www.amazon.com/Lady-Tasting-Tea-Statistics-Revolutionized/dp/0805071342

2. Fišer je obradio slučaj dame koja ispija čaj u knjizi The Design of Experiments

https://en.wikipedia.org/wiki/The_Design_of_Experiments

3. Fišer je 1926. u svom radu “The arrangement of field experiments” već na drugoj stranici diskutovao o granici 1/20=0.05 I kako on preferira 5%.

https://link.springer.com/chapter/10.1007/978-1-4612-4380-9_8

4. Fišerova knjiga Statistical Methods for Research Workers je izdata u 14 izdanja od 1925. do 1962. godine i imala je veliki uticaj na naučnike u 20. veku. U raznim izdanjima spominjao je 5% kao graničnu p-vrednost.

https://en.wikipedia.org/wiki/Statistical_Methods_for_Research_Workers

5. Više o tome odakle 5% može se ovde naći

http://www.jerrydallal.com/lhsp/p05.htm

6. Nejman-Pirson teorema

http://rsta.royalsocietypublishing.org/content/roypta/231/694-706/289.full.pdf

7.Fišerova kritika Nejmana I Pirsona

https://www.phil.vt.edu/dmayo/PhilStatistics/Triad/Fisher%201955.pdf

8. Nejmanov odgovor Fišeru

https://www.phil.vt.edu/dmayo/personal_website/Neyman-1956.pdf

9. Igon Pirsonov odgovor Fišeru

https://www.phil.vt.edu/dmayo/PhilStatistics/Triad/Pearson%201955.pdf

10. Lepo poređenje Fišerovog, Nejman-Pirson i kombinovanog pristupa

https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00223/full

11. Rad od Stivn Gudmana

https://www.sciencedirect.com/science/article/pii/S0037196308000620

12. Obaveštenje Američkog statističkog društva

http://web9.uits.uconn.edu/lundquis/ASA%20statement%20on%20p%20values.pdf

13. Nedavno objavljen rad od 72 svetska metodologa (Benjamin *et al.*)

https://www.nature.com/articles/s41562-017-0189-z