

Skriveni uzroci

Damjan Krstajić

(objavljeno 31. marta 2018. u Politikinom Kulturnom dodatku)

Statistika je naučna oblast matematike i ne dozvoljava logičke protivrečnosti. Međutim, to ne znači da paradoksa nema u statistici, ili ga nije bilo ... Evo primera. Na slici 1 su prikazani rezultati jedne studije.

SVI PACIJENTI	OPORAVLJENO	NIJE OPORAVLJENO	PROCENAT OPORAVLJENIH
NOVA TERAPIJA	20	20	50% (20/40)
STARA TERAPIJA	16	24	40% (16/40)

Slika 1

Zanima nas da li je nova terapija bolja od stare i upoređujemo procenat oporavljenih pacijenata. U praksi slične studije su sa većim brojevima, ali radi jednostavnosti mi ćemo upotrebiti male brojeve. Imamo ukupno 80 pacijenata od kojih je pola primilo jednu, a pola drugu terapiju. Očigledno je da nova terapija daje bolje rezultate, jer sa njom dobijamo 50% oporavljenih, dok sa starom 40%. Zapisali smo i pol svakog pacijenta i na slici 2 su prikazani rezultati oporavka za muškarce i žene zasebno.

MUŠKARCI	OPORAVLJENO	NIJE OPORAVLJENO	PROCENAT OPORAVLJENIH
NOVA TERAPIJA	18	12	60% (18/30)
STARA TERAPIJA	7	3	70% (7/10)

ŽENE	OPORAVLJENO	NIJE OPORAVLJENO	PROCENAT OPORAVLJENIH
NOVA TERAPIJA	2	8	20% (2/10)
STARA TERAPIJA	9	21	30% (9/30)

Slika 2

Sada vidimo problem. Sumirani rezultati (slika 1) nam govore da nova terapija daje 10% više oporavljenih pacijenata, dok ako gledamo svaki pol pojedinačno (slika 2) dobijamo obrnuto.

Ovaj paradoks je dobio naziv po Edvardu Simpsonu (Edward H. Simpson) koji je 1951. godine objavio rad o ovome. Trideset godina kasnije Denis Lindlij (Dennis V. Lindley) i Melvin Novik (Melvin R. Novick) su detaljno analizirali ovu problematiku sa zaključkom da rešenje zavisi od konteksta. Po njima, u nekim slučajevima treba prihvatiti šta važi za sve (slika 1), a u drugim za podgrupe (slika 2). Za ovaj naš primer, koji je identičan njihovom u radu, oni tvrde da ne može da se prihvati da je nova terapija bolja. Prvo, primećuju da oboleli muškarci, bez obzira koju terapiju primali, imaju bolji oporavak od obolelih žena (60% i 70% nasuprot 20% i 30%). To znači da je moguće da pol utiče na oporavak od bolesti. Drugo, iako imamo podjednak broj žena i muškaraca u uzorku, više muškaraca je primilo novu terapiju od žena (30 nasuprot 10). Možemo da zamislamo situaciju gde lekari radije daju novu nepoznatu terapiju onima koji imaju veće šanse da se oporave. Stoga je moguće da je pol uticao i na to ko prima terapiju. Lindlij i Novik smatraju da zbog svega toga pol treba da se izvuče kao zaseban faktor u analizi i da se posmatra slika 2.

Da li bismo isto odlučili, ako bismo na slici 2, umesto po polu, imali pacijente podeljene u zavisnosti od toga na koju nogu ustaju ujutru? Tada bismo imali da nova terapija daje poboljšanje za sve pacijente, ali ne za one koji ustaju na desnu nogu, niti za one koji ustaju na levu. U tom slučaju bismo prihvatili sumirane rezultate (slika 1), jer nema smisla razlikovati pacijente po tome na koju nogu ujutru ustaju. Ovo je jednostavan primer, jer nema uzročnosti, ali to ne znači da ako postoje uzročno-posledične veze da je uvek tačan odgovor po podgrupama.

Džud Perl (Judea Pearl) je relativno skoro u svojoj knjizi o kauzalitetu dokazao da sa razumevanjem uzročnosti Simpsonov paradoks ima objašnjenje i jedinstveno rešenje. Naveo je sva moguća uzročno-posledična scenarija u kojima može da se pojavi ovaj paradoks i u kojim slučajevima je tačno da se prihvati sumirani rezultat (slika 1), a u kojim po podgrupama (slika 2). Njegov dokaz uopšte nije jednostavan, ali ono što je bitno je da njegovo rešenje ne postoji u okviru statistike, već samo uzimajući u obzir uzročnost. Perl je 2011. godine, za doprinos razvoju uzročnosti, dobio Turingovu nagradu, jednu od najprestižnijih priznanja u računarstvu.

Simpsonov paradoks može da se izbegne sa dobrim dizajnom eksperimenta. Ako postoje indicije da pol utiče na terapiju, onda je trebalo osmisliti studiju tako da pola muškaraca prima novu, a druga polovina staru terapiju, i da isto važi i za žene. U tom slučaju ne bi bilo moguće da dobijemo Simpsonov paradoks ako gledamo sumirane rezultate i po polovima. Ključno pitanje pri svakom dizajnu eksperimenta je da li smo sve moguće uzroke uzeli u obzir. Ako nismo, onda je moguće da postoji Simpsonov paradoks u našim podacima i da grešimo.

Lično sam bio upoznat sa ovim paradoksom pre nego što ga je Perl razjasnio. Tada sam prihvatio da nema rešenja za njega i osećao nelagodnost u vezi saznanja da samo na osnovu slike 1, mi bismo bez kolebanja doneli pogrešnu odluku da je nova terapija bolja. Ništa prijatnija mi nisu ni sledeća pitanja na koja ni danas nemam odgovor: Kako u budućnosti mogu da znam, samo gledajući sliku 1, da ne postoji neki skriveni uzrok sličan slici 2? Koliko puta sam do sada doneo pogrešnu odluku samo na osnovu slike 1?

Simpsonov paradoks je primer gde identični podaci mogu da daju suprotne zaključke u zavisnosti od razumevanja uzročnosti i gde ne postoji rešenje u okviru statistike. Kao statističar po struci, voleo bih da je drugačije i zahvalan sam Perlu što je dao rešenje. To i dalje ne menja činjenicu da nepoznati skriveni uzroci vise kao Damaklov mač iznad statistike.

Reference koje podržavaju činjenice spomenute u članku

1. Simpsonov paradoks

https://en.wikipedia.org/wiki/Simpson%27s_paradox

2. Rad od Edvarda Simpsona

<http://www.jstor.org/stable/2984065>

3. Rad od Lindlija i Novika

<http://www.jstor.org/stable/2240868>

4. Džud Perl

<http://bayes.cs.ucla.edu/home.htm>

5. Knjiga o kauzalitetu

<http://bayes.cs.ucla.edu/BOOK-2K/>

6. Džud Perlov rad o Simpsonovom paradoksu objavljen posle knjige

<https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2014.876829?journalCode=utas20>

7. Džud Perl dobitnik Turingove nagrade

http://amturing.acm.org/award_winners/pearl_2658896.cfm