Statistika i uzročnost

Damjan Krstajić

Šta bih ja, kao statističar sa preko dve decenije iskustva, voleo da se zna u vezi statistike? Simpsonov paradoks. Definitivno. Evo primera. Na slikama 1 i 2 su prikazani rezultati jedne iste studije. Zanima nas da li je nova terapija bolja od stare i upoređujemo procenat oporavljenih pacijenata.

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

Slika 1

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)

S**lika 2**

Proverite. Sabiranjem podataka iz tabela za muškarce i žene na slici 2 dobija se tabela na slici 1. Dakle, u pitanju su isti podaci koji su prikazani na obe slike, a opet su kontradiktorni u vezi zaključka da li je nova terapija bolja.

Zbunjeni ste? Koje je rešenje? Denis Lindlij (Dennis V. Lindley) i Melvin Novik (Melvin R. Novick) su 1981. godine objavili rad u kojem su detaljno analizirali ovu problematiku sa zaključkom da rešenje zavisi od konteksta. Prema njihovom mišljenju, u ovom konkretnom slučaju treba prihvatiti rezultate po podgrupama (slika 2). Lepo, a šta ako smo samo znali sumirane rezultate (slika 1), kako smo mogli da znamo da grešimo? Nismo.

Mišljenja sam da je poznavanje Simpsonovog paradoksa krucijalno za razumevanje kako se uz pomoć podataka, dakle samo sa statistikom, ne može zaključiti uzročnost. To je lako reći, ali koju odluku u praksi da donesemo kad se sledeći put suočimo samo sa slikom 1?

Samo na osnovu informacija iz slike 1, jedino što možemo da kažemo je da je nova terapija bolja. Međutim, nije isto ako zaključimo da „*podaci govore da je nova terapija bolja*“ i kad rezimiramo „*pretpostavljajući da u našim podacima nema skrivenih faktora koji bi ukazali na suprotno, trenutno prihvatamo hipotezu da je nova terapija bolja“.*

Kad su u pitanju manje važne odluke, prihvatićemo šta se može zaključiti iz slike 1 znajući da možda grešimo i nadajući se da nemamo peh da baš u našim podacima postoji Simpsonov paradoks. Međutim, kad su u pitanju životi ljudi (medicina) ili velika suma novca i moć (političke i marketing kampanje u SAD), ne treba da nas začudi kako se dosta ulaže da se proveri postojanje *„skrivenih faktora koji bi ukazali na suprotno“,* pre nego što se donese odluka o terapiji ili kampanji.

Što se samog Simpsonovog paradoksa tiče, Džud Perl (Judea Pearl) je u svojoj knjizi *Causality* dokazao da paradoks ima objašnjenje i jedinstveno rešenje sa razumevanjem uzročnosti. Zainteresovani mogu da pročitaju detalje i u njegovom javno dostupnom komentaru objavljenom 2014. godine u časopisu *The American Statistician*. Važno je naglasiti da on nije našao rešenje u okviru statistike, već uvođenjem novih termina i novog jezika za uzročnost.

Perl je dobrim delom svoje karijere radio na razvoju Bajesovih mreža i propagirao je verovatnosni pristup u veštačkoj inteligenciji, o čemu je napisao i knjigu. Međutim, u jednom trenutku je shvatio da samo sa verovatnoćom i statistikom ne može da se razume uzročnost i počeo je da

uvodi novitete koji su delom bili u suprotnosti sa onim što je do tada objavljivao.

Po meni, njegov glavni doprinos je nova matematičku operacija *uradi()* (eng. d*o()*). Njena tačna definicija nije jednostavna i zainteresovani mogu da je nađu u njegovoj knjizi *Causality,* kao i u nekoliko njegovih javno dostupnih radova. Na pitanje zašto je ona neophodna, umesto odgovora, postavio bih kontra-pitanje. Koliko ste uzročno-posledičnih odnosa rešili u vašem svakodnevnom životu tako što ste samo gledali? Pažljiv posmatrač i poznavalac može dosta njih da pretpostavi, ali da bi ih dokazao mora da *uradi()* nešto. Primera radi, ako ne radi sijalica u kolima, mi možemo da pretpostavimo da je ona izgorela ili da sa dovodnom žicom nešto nije u redu ili nešto treće, ali pukim posmatranjem ne možemo proveriti uzrok i rešiti problem.

Dok je Perl razvijao teorijske osnove uzročnosti, za čiji doprinos je 2011. godine dobio Tjuringovu nagradu, jednu od najprestižnijih priznanja u računarstvu, druga grupa naučnika je isto radila na uzročnosti, ali pre svega na tome kako da se u praksi izbegnu *„skriveni faktori koji bi ukazali na suprotno“*. Njihovi predlozi promena nisu bili toliko revolucionarni, kao Perlovi, ali su zato bili lakši za primenu i stoga popularniji. Rubinov uzročni model, nazvan po američkom statističaru Donaldu Rubinu, samo je jedan primer.

Uzročnost je mlada naučna oblast, a za njen razvoj u praksi su zaslužne potrebe moderne medicine da se pokaže uspešnost raznih terapija. Koliko je meni poznato, uzročnost se danas ozbiljno primenjuje samo u medicini (klinički trajali i epidemiologija) i u političkim naukama, gde se na kampanje metodološki gleda kao na terapije nad ljudima.

Tačno je da ima naučnika i kod nas i u svetu koji svesno zloupotrebljavaju statistiku, ali moram da priznam da nije lako ni iskrenim ljubiteljima nauke. Većina današnjih praktičnih studija o uzročnosti u nauci su dosta skupa. Šta da se radi onda? Ne znam. Gledam prijatelja naučnika, nefrologa u Zagrebu, kako posvećeno sa kolegama iz raznih krajeva sveta radi na pronalaženju uzroka u svojoj oblasti i stičem utisak da je saradnja sa drugim iskrenim ljubiteljima nauke deo rešenja.

Zašto bih voleo da su svi upoznati sa Simpsonovim paradoksom? Prvo, upozorava nas da samo na osnovu statistike možemo pogrešiti pri zaključivanju o uzrocima. Drugo, ukazuje na to da sve aktuelnija priča o nenadgledanoj veštačkoj inteligenciji, koja iz puke mase podataka nešto

pouzdano zaključuje, može biti opasna. Treće, možda i neke zainteresuje uzročnost.

Reference koje podržavaju činjenice spomenute u članku

1. Simpsonov paradoks

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

2. Rad od Lindlija i Novika

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

3. Džud Perl

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

4. Knjiga *Causality*

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

5. Džud Perlov komentar o Simpsonovom paradoksu objavljen u *The American Statistican*

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

6. Neki od javno dostupnih radova o *uradi()* operatoru

https://arxiv.org/abs/1210.4852

https://www.degruyter.com/view/journals/jci/7/1/article-20192002.xml

https://www.jstor.org/stable/43288500?seq=1

7. Džud Perl dobitnik Tjuringove nagrade

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

8. Donald Rubin

https://statistics.fas.harvard.edu/people/donald-b-rubin

9. Rubinov uzročni model

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