Mnoštvo podataka u nauci je ozbiljan problem

Damjan Krstajić

(objavljeno 10. septembra 2022. u Danasu (dodatak Nedelja)

Živimo u dobu u kojem imamo pristup podacima kao nikad do sada. Njihov obim je toliki da postoji izraz u engleskom jeziku (Big Data) sa kojim se želi naglasiti da je u pitanju veoma velika količina podataka.

Neko bi mogao pomisliti da je ovo novo doba idealno za statističare, ali bi se prevario. Temelji današnje naučne metodologije u statistici su postavljeni prvom polovinom 20. veka, kada nije bilo ni računara, ni puno podataka na raspolaganju. Statistički metodi koji su nam danas dostupni nisu koncipirani da se nose sa velikom količinom podataka. To predstavlja ozbiljan problem za današnju nauku, a malo ljudi je toga svesno.

Naviknuti smo da preko medija čujemo razne rezultate statističkih analiza koji su pre svega deskriptivne prirode (proseci, prikazi trendova, itd.). Međutim, statistika, kao naučna oblast matematike, primenjuje se u eksperimentalnim naukama kad je potrebno da se prihvati ili odbije neka hipoteza. Recimo, u medicini se uz pomoć statistike odlučuje da li da se primeni ova ili ona terapija.

Kad imamo neku hipotezu, na primer, da je nova terapija bolja od stare, nije dovoljno da podaci govore tome u prilog. Potrebno je takođe izračunati neku meru naše moguće greške pri prihvatanju hipoteze da je nova terapija bolja. Statistika nam pomaže da to procenimo i ako je sve zadovoljavajuće, onda kažemo da postoji statistički značajna razlika između nove i stare terapije. Međutim, postoji detalj u procesu odlučivanja koji je bitan da bismo razumeli zašto sa velikom količinom podataka sve ovo počinje da gubi smisao.

Istaknuti svetski statističari prve polovine 20. veka (Fišer, Nojman, Pirson) podrazumevali su da će naučnici prvo dizajnirati eksperiment pre nego što bilo šta urade. Pojednostavljeno, naučnik bi prvo trebalo da definiše kolika je razlika po njemu značajna, onda bi se na osnovu te očekivane razlike i procenjene varijabilnosti, uz pomoć matematičke formule, odredila potrebna veličina uzorka za taj eksperiment.

U našem slučaju to bi bio broj pacijenata u studiji neophodnih da procenimo da li je nova terapija bolja od stare. Kad se dostigne taj broj pacijenata uključenih u studiju, zaustavlja se sa primanjem novih pacijenata i izračunavaju se statistike da se utvrdi da li je razlika između nove i stare terapije statistički značajna.

E sad dolazimo do trik dela. Da li u studiji smemo da imamo više pacijenata od broja definisanog u njenom dizajnu? U principu ne.

Statistički značajna razlika nije neka fiksna velika razlika, već veličina koja je određena dizajnom eksperimenta. Ona zavisi od veličine uzorka i varijabilnosti u podacima. Što je veći uzorak, razlika može da bude manja, da bi ona bila statistički značajna. Drugim rečima, bilo koja razlika, ma koliko mala bila, sa dovoljno velikim uzorkom postaće statistički značajna.

Na prvi pogled ovo je suprotno intuiciji. Zar sa više podataka mi nismo sigurniji u ono što treba da procenimo? Jeste. Sa što više podataka mi ćemo biti sigurniji da razlika između dva skupa postoji, ma koliko ona mala bila.

Ovo otvara mogućnost ozbiljnim manipulacijama statistikom u nauci.

Očigledna mana ovog pristupa sa dizajnom eksperimenta jeste da početno definisanje razlike, ono što naučnik podrazumeva da je značajno u ovom slučaju, jeste arbitrarno. Važno je napomenuti da u kliničkim ispitivanjima nezavisna ustanova to proverava i ona odobrava dizajn pre početka eksperimenta. Međutim, u drugim oblastima, ima naučnika koji prvo odrade eksperiment bez dizajna i onda traže statističku značajnost u svojim podacima, jer je to često uslov da se njihov naučni rad objavi.

U doba kad nije bilo puno podataka, a ovo važi i za neke naučne oblasti i danas, problem je bio kako generisati traženu veličinu uzorka, jer to zahteva dosta vremena i novca. Međutim, šta da radimo sa naučnim oblastima u kojima baratamo sa velikom količinom podataka? Tamo će skoro svaka razlika biti statistički značajna, a samim tim, tamo se gubi smisao testiranja hipoteza.

Pojednostavljeno, statistička metodologija 20. veka nije spremna za velike uzorke. Od onog što se uči u današnjim udžbenicima statistike, osim deskriptivnih statistika (aritmetička sredina, varijansa, itd.), malo šta drugo je primenjivo na veliku količinu podataka.

Danas su svetski statističari svesni ove problematike i ozbiljno se radi na pronalaženju rešenja, ali su ona trenutno, koliko uspevam da vidim, sva u

vidu nekih korekcija koja važe od slučaja do slučaja. Važno je naglasiti da sve te korekcije nisu rešenje problema, već više kao neka brana da se ne može svašta objavljivati.

Danas u svetu imamo ozbiljna ulaganja u naučna istraživanja koja se oslanjaju na veliku količinu podataka. U pitanju su veoma moćni računari i skupi instrumenti. Velike su očekivanja od Big Data u nauci. Nisam siguran da se razume da bez mogućnosti testiranja hipoteza, ostaje nam samo deskriptivna statistika. Mišljenja sam da razvoj u nauci nauštrb naučne metodologije nije pravi razvoj. Ko nam garantuje da sa Big Data u nauci nećemo dobiti Big Lie (veliku laž) u nauci?

Stoga ne treba da nas začudi da uporedo sa zapostavljanjem naučne metodologije, zadnjih godina dejta sajentist (Data Scientist) postaje sve više traženo zanimanje, a postoji i nešto što se zove dejta sajens (Data Science), koja je bliska mašinskom učenju i obradi velike količine podataka. Zašto je nastala potreba za posebnom (takozvanom) naukom dejta sajens? Zar već ne postoji prava nauka o podacima koja se zove statistika?

U današnjoj nauci postoji ozbiljan problem zloupotrebe statistike, a ako se još uzme u obzir da nam je sa velikim uzorcima otežana pravilna upotreba statistike, sumnjam da ćemo sa velikom količinom podataka biti bliži naučnoj istini. Osim, ako se u statistici ne pojave neki novi geniji kao što su bili Fišer, Nojman i Pirson, i dođe do promene paradigme u njoj.