Populacija i reprezentativni uzorak

Damjan Krstajić

Poslednjih dana smo bombardovani raznim statistika u vezi korona virusa. Čujemo kako je u jednoj zemlji ovoliko obolelo, od kojih onoliko preminulo (sa izračunatim procentom smrtnosti), a opet u drugoj državi drugačije cifre za iste statistike. Spominju se i procenjene brojke inficiranih. Ne vidim ništa sporno ni sumnjivo u datim statistikama, ali primećujem kako ljudi prebrzo donose zaključke ne razumevajući kompleksnost problema.

Da bih lakše objasnio problematiku, uzeću jedan naizgled lakši primer iz druge oblasti. Neka imamo države A, B i C i cilj je da proverimo čiji stanovnici su u proseku najviši. Izvešćemo sledeći eksperiment. Postavićemo posmatrače u najprometnija šetališta u svakom od tri glavna grada i u jednom danu izmeriti visinu svih prolaznika. Na kraju dana sabraćemo sve izračunate visine, podeliti sa brojem prolaznika i tako dobiti prosečnu visinu.

Ispostavilo se da su stanovnici države C u proseku za 5 cm viši od ostalih. Nema sumnje da je statistika tačna, ali da li na osnovu nje smemo da zaključimo da su stanovnici države C u proseku viši od A i B?

Detaljnom proverom podataka ispostavilo se da se u glavnom gradu C održavalo državno prvenstvo u košarci. Košarkaši su iskoristili slobodan dan, prošetali se svojim glavnim gradom i njihove visine su izmerene. Takođe, u glavnom gradu države C tog dana većinom su se šetali muškarci, dok su žene ostale kod kuće. Da li sad možemo da tvrdimo da su stanovnici države C u proseku viši od A i B?

Da bismo tačno znali čiji stanovnici su u proseku najviši trebali bi da izmerimo svakog stanovnika u sve tri države, a to je nemoguće. Kako onda rešiti ovaj naizgled jednostavan problem? Potrebno je da se iz svake države uzme *reprezentativan uzorak* i na osnovu njega izvede zaključak. Šta je to reprezentativan uzorak?

Stručno govoreći imamo *populaciju* čije nas karakteristike interesuju, a to su u našem slučaju svi stanovnici jedne države. Iz populacije biramo podskup koji nazivamo *uzorak*, a on je *reprezentativan* ako tačno odražava karakteristike populacije.

U našem slučaju reprezentativan uzorak bi morao recimo da ima isti procenat žena i muškaraca kao u populaciji. Slično bi trebalo da važi i za procenat mladi i starih, pa da bude podjednak procenat iz svih krajeva zemlje itd. Ukratko, veliki je posao i komplikacija naći reprezentativan uzorak.

Međutim, ima slučajeva kad mi i ne možemo da znamo karakteristike populacije, pa samim tim nemamo ni mogućnost da odaberemo reprezentativan uzorak.

Neka populacija bude skup ljudi obolelih od korona virusa. Mi ne možemo da znamo ko je sve bio oboleo, jer ima onih koji su posle par dana kod kuće, ozdravili i nastavili sa svojim životom ne znajući da su je preležali. Stoga ne možemo ni da znamo da li je uzorak obolelih koji imamo reprezentativan. Samim tim, kao što je pokazano sa merenjem visina, samo na osnovu grubih statistika nije preporučljivo da se izvode zaključci.

Sada u priču ulazi epidemiologija. Malo je poznato da je to nauka čija se metodologija toliko razvila u zadnjih dvadesetak godina da svetski epidemiolozi danas, neformalno govoreći, drže časove statistike nama statističarima kako da pratimo i obrađujemo opservacione studije. Evo par primera.

Džud Perl (Judea Pearl) je razvio teoriju uzročnosti i ukazao na mane statistike, ali Sander Grinland (Sander Greenland) je samo jedan od plejade istaknutih epidemiologa koji su razvijali praktična rešenja za razumevanje uzročnosti u opservacionim studijama. Takođe, u medicinskim istraživanjima imamo konstantan problem nemogućnosti organizovanja randomiziranih trajala (metod za odabir boljeg tretmana), ali su epidemiolozi Džordž Smit (George Davey Smith) i Šah Ebrahim (Shah Ebrahim) ti koji su razvili Mendelevu randomizaciju kao supstitut i podigli epidemiološka istraživanja na viši metodološki nivo.

Kao što se ne bih usudio da išta zaključim u vezi prosečnih visina u državama A, B i C na osnovu dobijenih rezultata eksperimenta, tako se i sada ne mešam u posao epidemiologa. Pošto sam statističar po struci, svestan sam da ne mogu ništa da zaključim na osnovu statistika koje neminovno saznajem o korona virusu preko medija. Stoga, ne iznosim svoje „razumevanje“ situacije drugima, već pratim šta epidemiolozi savetuju i verujem im.

Reference koje podržavaju činjenice spomenute u članku

1. Reprezentativni uzorak

https://www.statisticssolutions.com/what-is-a-representative-sample/

2. Populacija

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

3. Uzorak

https://en.wikipedia.org/wiki/Sample_(statistics)

4. Džud Perl (Judea Pearl)

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

5. Sander Grinland (Sander Greenland)

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

6. Džordž Smit (George Davey Smith) i Šah Ebrahim (Shah Ebrahim)

http://www.bris.ac.uk/social-community-medicine/people/george-davey-smith/index.html

https://www.lshtm.ac.uk/aboutus/people/ebrahim.shah

https://academic.oup.com/ije/article/32/1/1/642797