Centralno u statistici

Damjan Krstajić

(objavljeno 27. februara 2021. u Politikinom Kulturnom dodatku)

Probajte prilikom upoznavanja sa nepoznatom osobom, da na pitanje čime se bavite, odgovorite da ste statističar. Kakva je reakcija? Oduševljenje? Potpitanja? Probajte takođe da saznate od bivših i sadašnjih studenata koji predmet im je bio najdosadniji na fakultetu. U mom nereprezentativnom uzorku ispitanika, na vrhu liste stoji moja naučna oblast, statistika.

Pretpostavljam da svaka nauka ima svoje čari, a uzrok tome što neke u našim očima nemaju, jeste što su nam pogrešno predavane. Pod pogrešno, ne mislim netačno, već pogrešno u smislu razumevanja. Zainteresovao sam se za statistiku tek nakon završetka studiranja, radeći u privredi i postavljajući pitanja zašto se nešto primenjuje. Upoznavao sam se sa zanimljivim naučnim radovima (Bajesova statistika, uzročnost) koja dovode u pitanje ono što sam učio na fakultetu. Paradoksalno, tek kad sam uvideo mane statistike, počeo sam da spoznajem i njenu jedinstvenu vrednost, a tu glavno mesto zauzima *centralna granična teorema*.

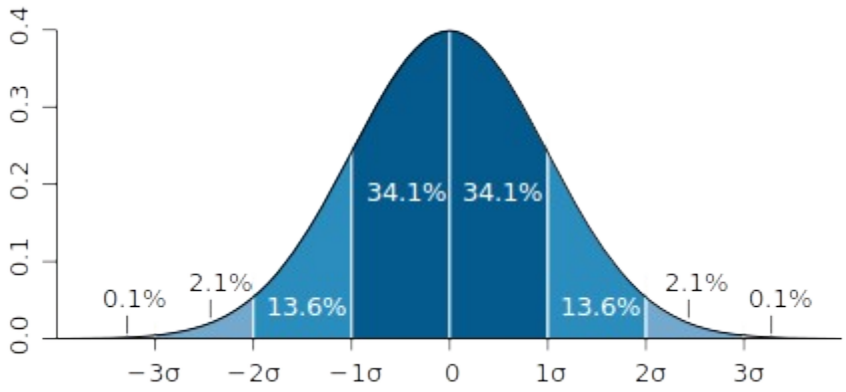
Neka posmatramo visinu muškaraca u jednom društvu. Visina varira i izdvojićemo dve mere za njen opis: aritmetička sredina i standardna devijacija. Aritmetička sredina (μ) nekog skupa je kad saberemo sve vrednosti tog skupa i podelimo sa brojem članova, a standardna devijacija (σ) je mera koliko vrednosti u tom skupu odstupaju od aritmetičke sredine (formula ovde nije bitna). Neka bude da u našem društvu visina muškaraca ima μ=180cm, σ=10cm.

Normalna ili Gausova raspodela je u srži centralne granične teoreme. Ona je definisana nimalo jednostavnom funkcijom f(x) koja zavisi od (μ, σ).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

.

Priznajem, pre ćemo uplašiti nekog sa jednačinom normalne raspodele nego zaseniti lepotom, ali njena vrednost se kasnije vidi. Na slici 1 je prikazana kriva funkcije f(x). Cela površina ispod krive je 100%, a verovatnoća da naša pojava (sa normalnom raspodelom) bude u nekom intervalu je procenat te površine.



Neka bude da visina muškaraca ima normalnu raspodelu sa μ=180cm, σ=10cm. Tada je verovatnoća da slučajan muškarac ima visinu u intervalu (170cm, 180cm)= (μ-σ,μ) tačno 34.1%. Slično, verovatnoća da muškarac ima visinu u intervalu (160cm, 200cm)=(μ-2σ,μ+2σ) je 95.4% (13.6%+34.1% +34.1%+13.6%). Uz pomoć slike 1, kolika je verovatnoća da muškarac ima visinu u intervalu (200cm, 210cm)=(μ+2σ,μ+3σ)?

U statistici, stručno govoreći imamo *populaciju* čija karakteristika nas interesuje, recimo visina ljudi. Kako ne možemo sve ljude da izmerimo, iz populacije biramo podskup koji nazivamo *uzorak*, i na osnovu njega pokušavamo da razumemo populaciju.

Visina čoveka je retka pojava u prirodi za koju se pokazalo da ima normalnu raspodelu. Za većinu slučajnih pojava u prirodi, mi ne možemo da znamo po kojoj zakonitosti se pojavljuju. Prinos hibrida pšenice po hektaru, broj ljudi koji dnevno posećuje ekspozituru banke, broj dana dok se ne pokvari neki uređaj – sve su to pojave sa nama nepoznatim zakonitostima. Bolje razumevanje tih pojava može da nam pomogne u odlučivanju. Koliko šaltera da bude u ekspozituri? Koji garantni rok da bude za uređaj? I slično.

Tu sad u pomoć dolazi centralna granična teorema. Ako posmatramo događaje koji su nezavisni (visina jednog čoveka ne zavisi od visine drugog) i imaju istu raspodelu (zakonitost po kojom se pojavljuju), što veći uzorak iz populacije da uzmemo, to će njegova aritmetička sredina imati raspodelu

bliže normalnoj raspodeli. Drugim rečima, mi i ne moramo da znamo zakonitost po kojoj se pojavljuje slučajna pojava X. Dovoljno je da imamo veliki uzorak, da izračunamo srednju vrednost i standardnu devijaciju iz uzorka i onda aritmetičku sredinu X možemo da aproksimiramo normalnom raspodelom.

Zamislite prvu polovinu 20. veka, bez računara i digitrona. Kako na osnovu uzorka da izračunate verovatnoću neke slučajne pojave? Teško da ćemo moći da izračunamo pojedinačne verovatnoće događaja neke pojave, jer ne znamo njenu zakonitost. Međutim, zahvaljujući centralnoj graničnoj teoremi, moći ćemo da izračunamo verovatnoću aritmetičke sredine uzorka te slučajne pojave. Trik je da ne baratamo sa pojedinačnim vrednostima neke pojave, već sa njenom aritmetičkom sredinom. Dovoljno je bilo da imamo papir, olovku i tabelu normalne raspodele. Usuđujem se reći da je ovo doprinelo revoluciji u nauci 20. veka i maltene neizbežnoj primeni statistike u njoj.

U udžbenicima statistike (ovo će biti poznato onima koji su imali zadovoljstvo da je polažu kao ispit) dosta se spominje „pretpostavimo da je osnovni skup normalno raspodeljen“. Međutim, nisam nigde video objašnjenje, a priznajem nisam ni pitao dok sam studirao, zašto se to toliko pretpostavlja. Moje objašnjenje, koje naravno ne mora biti tačno, jeste da su autori prvih udžbenika i priručnika statistike tokom prve polovine 20. veka koristili taj termin, jer su pretpostavili da će istraživač naravno koristiti aritmetičku sredinu neke pojave, koja je normalno raspodeljena.

Od polovine 20. veka pa do sada, pojava računara i veća količina podataka je promenila u velikoj meri statistiku, ali nekako udžbenici statistike kao da su ostali isti, a objašnjenja maltene izgubljena.

Postoji jasan i nedvosmislen dokaz centralne granične teoreme u kojoj se aritmetičke sredine nezavisnih slučajnih pojava sa istom raspodelom mogu aproksimirati normalnom raspodelom. Međutim, šta je to toliko magično u definiciji njene funkcije?

Kao da je prilikom stvaranja ovog sveta dozvoljeno raznim slučajnim promenljivima da imaju svoje zakonitosti, ali samo pod uslovom da se njihove aritmetičke sredine pridržavaju normalne raspodele.

Reference koje podržavaju činjenice spomenute u članku

1. Normalna ili Gausova raspodela

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

2. Centralna granična teorema

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