8. KORELACIJA NE ZNAČI KAUZALNOST

Malo statistike koju nikada niste učili

Ponovo ćemo posegnuti za rečnikom:

Statistika je oblast matematike koja se bavi sakupljanjem, analizom interpretacijom i prezentacijom mase numeričkih podataka[1].

Ovde ćemo se dotaći tek nekoliko najosnovnijih statističkih veličina, i to samo onoliko koliko je nepohodno za razumevanje teze iz naslova poglavlja[2], a koju početnik na polju nauke ni jednog jedinog trenutka ne sme da smetne sa uma.

Ponekad u toku predavanja na osnovnim studijama upitam studente: „Znate li šta je aritmetička srednja vrednost?“ Gledaju me začuđeno koliko i ja njih. *Tabula rasa*[3]. „A znate li makar šta je prosek?“ To znaju, uglavnom iz svojih školskih svedočanstava. Sad smo već blizu:

Aritmetička srednja vrednost skupa opservacija/merenja je njihova suma podeljena brojem opservacija:

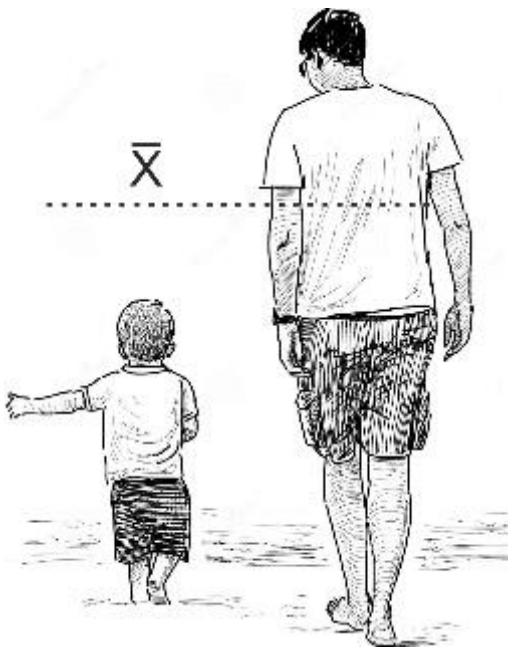
$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

[1] *www.merriam-webster.com*

[2] Za sve ostalo što vam bude neophodno prilikom izrade master rada prepušteni ste samostalnom učenju. Nije nepremostivo ☺

[3] Tabula rasa (lat.) - neispisana tabla.

Na prvi pogled radi se o veoma korisnoj a jednostavno odredivoj statističkoj veličini – iz n merenja vrednosti x prostim aritmetičkim operacijama računamo prosečnu vrednost. Da li je rezultat odista *koristan*?



Hoćemo, na primer, da izračunamo *aritmetičku srednju vrednost* visine čoveka, na osnovu visine jednog trogodišnjeg deteta, 90 cm, i njegovog oca, 177 cm.

$$\bar{X} = \frac{90 + 177}{2} = 133.5$$

Dobijeni rezultat je potpuno besmislen, samim tim i neupotrebljiv. Jasno je i zašto – uzorak, ili *skup opservacija* na kome smo radili je premali, ima samo dva člana.

Proširimo skup opservacija na 1000, recimo. Ali 500 (malih) ljudi izmerimo u obdaništu, a ostalih 500 na ulici, nasumično. I opet dobijamo besmislen rezultat, neka bude 140 cm. Dakle, i dalje nešto nije u redu sa uzorkom, a ne samo njegova veličina. Šta? Pa svaki od izmerenih, malih i velikih, ljudi svojim visinom previše odstupa od dobijene prosečne visine.

Standardna devijacija[4] je statistička mera *odstupanja* elemenata uzorka od njegove aritmetičke srednje vrednosti. Tačnije, koliko *u proseku* elementi uzorka odstupaju od aritmetičke sredine uzorka:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Ukratko, što je standardna devijacija veća, dobijena srednja vrednost nam je neupotrebljivija. Sve ovo lako je uočljivo na posmatranom primeru, uostalom, niko iole razuman neće na gornji način birati uzorak. Šta ukoliko imamo posla sa nedovoljno poznatim prirodnim pojavama, ili ukoliko su nam merenja ograničena što fizičkim što materijalnim resursima? Pa standarna devijacija je prvi znak da li smo na dobrom putu.

Ali nedovoljan. Uostalom, sve prethodno odnosi se na merenja jedne promenljive i zaključivanja o njenom ponašanju na osnovu statističke analize, dok i u nauci i u tehnici istražujući najčešće pokušavamo da zaključimo nešto o međusobnim odnosima dve ili više promenljivih veličina. Za početak, da li neka veza uopšte postoji?

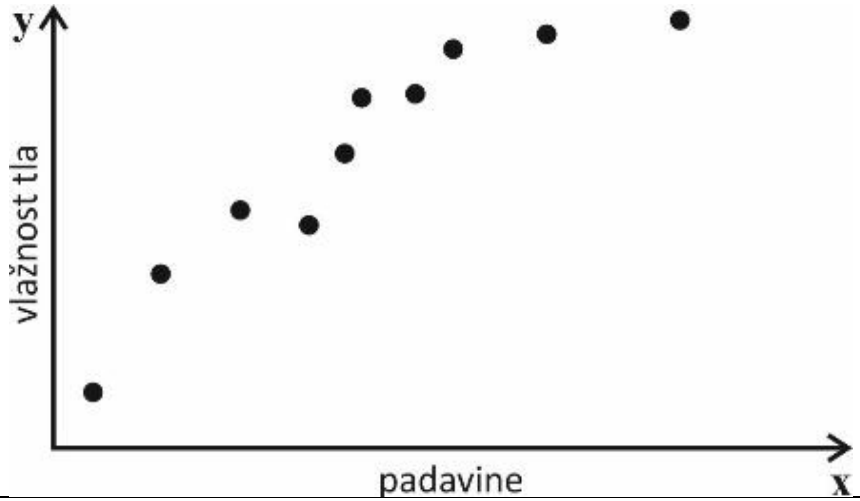
Da li vlažnost tla zavisi od visine padavina, na primer? Zavisi, *jasno je da* zavisi, reći će svako iole *razuman*. Za nauku nedovoljno[5]. Stručnjak će nas odmah podsetiti da to zavisi i od sastava tla, od prethodne zasićenosti tla vodom, od spoljašnje temperature, od evapotranspiracije... trista čuda o kojima niste razmišljali kad ste *zdravorazumski* odgovorili sa da. I stručnjak govori isključivo zdravorazumski, ali su njegovo znanje i iskustvo veći od vašeg. Klimate glavom i gubite interesovanje. S druge strane, naučnik, makar u pokušaju, nabrajajući sve od čega bi vlažnost tla *mogla* da zavisi, shvata da je pred velikim problemom. Previše promenljivih je u igri, od kojih je neke u stanju lako da

[4] Ili srednje kvadratno odstupanje.
[5] Podsetite se prvog poglavlja i varki „zdravog razuma“.

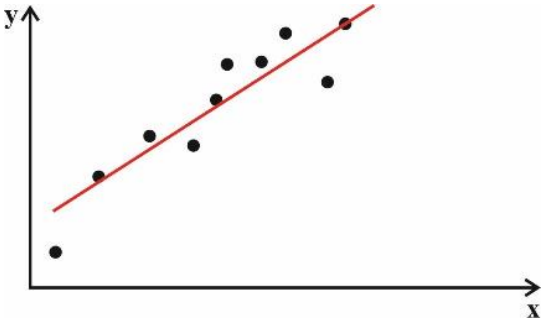
izmeri, neke teže, a neke su mu skoro nedostupne. A njega zanima samo odgovor na jednostavno pitanje – da li vlažnost tla zavisi od visine padavina.

U ovakvom razmišljanju prvo što pada u oči jeste da je vlažnost tla *verovatno* posledica padanja kiše. Stoga će padavine za početak proglasiti nezavisno promenljivom x, a vlažnost tla zavisno promenljivom y. I početi da meri[6], simultano padavine i vlažnost tla. Rezultate njegovog merenja možemo grafički predstaviti ovako:

Rezultat je manje-više očekivan – sa visinom padavina raste i vlažnost tla. Ali kako? Mereni parovi vrednosti ne pokazuju nekakvu preciznu pravilnost, jedino „dokazuju“ onu staru narodnu – kad je kiša, blato je ☺.



[6] Čim počne kiša ☺



Ukoliko pretpostavimo da je gornja, ili bilo koja, zavisnost dve promenljive linearna, moguće je povući pravu liniju koja najbolje opisuje tu zavisnost. Šta uopšte znači „najbolje opisuje“? Sigurno nećemo povlačiti linije lenjirom i iz toga izvlačiti naučne zaključke, već ćemo ponovo u pomoć pozvati statistiku, i analitički odrediti jednačinu prave.

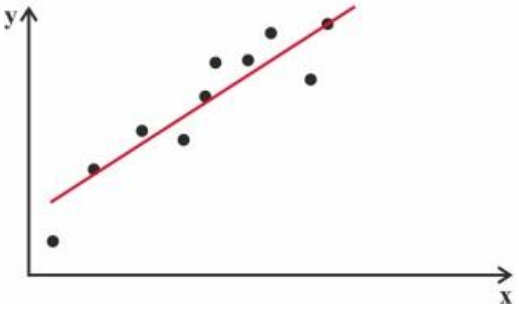
Regresiona prava je linearna zavisnost oblika $Y = aX + b$ kod koje je srednje kvadratno odstupanje, odnosno standardna devijacija najmanja. Parametri regresione prave dobijaju se kao:

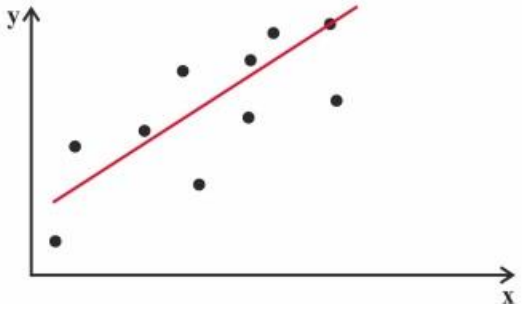
$$a = \bar{y} - b\bar{x}$$

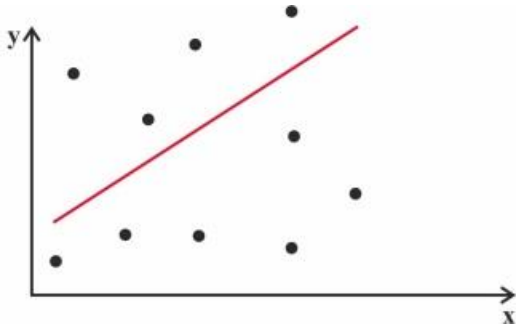
$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Gde su $\bar{x}$ i $\bar{y}$ srednje vrednosti promenljivih a n broj merenja, odnosno veličina uzorka.

Očigledno, regresionu pravu možemo analitički odrediti za bilo kakav uzorak.







Pa čak i ovaj sa treće slike u kojem su tačke razbacane tako da je već na prvi pogled jasno da promenljive nisu ni u kakvoj vezi.

Koeficijent korelacije je parametar koji nam pokazuje stepen zavisnosti između promenljivih, odnosno veličinu disperzije promenljivih oko regresione linije. Računa se kao:

$$K = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Koeficijent korelacije ima vrednost koja se kreće u rasponu od -1 do +1. Ako promenljive nisu povezane, K je jednak nuli. Kada većim vrednostima nezavisno promenljive x, odgovaraju i veće vrednosti zavisno promenljive y, odnosno, opadanjem vrednosti nezavisne x, opadaju i vrednosti zavisne y – onda je to pozitivna korelacija ($K>0$). Obrnuto, kada većim vrednostima nezavisno promenljive x, odgovaraju manje vrednosti zavisno promenljive y , odnosno opadanjem vrednosti nezavisne x rastu vrednosti zavisne y – onda je to negativna korelacija ($K<0$).

Važi opšte pravilo: što je vrednost koeficijenta proste linearne korelacije bliža jedinici, to je međuzavisnost među posmatranim pojavama jača. Koeficijent korelacije nikada nema vrednosti 1 ili -1, jer to bi značilo da između pojava postoji matematička, a ne statistička veza.

Dakle, da bismo mogli da zaključimo da između dve promenljive, bilo da se radi o dve prirodne pojave ili o bilo kakvom paru simultano merenih vrednosti, postoji statistička zavisnost, nije dovoljno da merimo, posmatramo merne tačke u koordinatnom sistemu ili konstruišemo regresionu pravu, neophodno je da postoji i dovoljno veliki

koeficijent korelacije, bilo da je blizak 1 ili -1. Koliko veliki zavisi o kojoj naučnoj disciplini se radi i koliko je složena pojava koju tretiramo, odnosno koliko trenutno „neuhvatljivih“ promenljivih je u igri[7]. Generalno, „jakim korelacijama“ smatraju se sve one kod kojih je $K > 0.8$ odnosno $K < -0.8$.

Šta je kauzalnost

Uzročnost ili *kauzalitet* (lat.*causalitas*) jeste filozofski koncept koji pokušava da objasni odnos uzroka i posledice. U užem smislu, uzročnost je nužna veza uzroka i posledice, koja podrazumeva da kad god je prisutan uzrok, tad je nužno prisutna i posledica (Pavlović, 1997).

Uzročnost, odnosno kauzalnost, je osnova *induktivnog* zaključivanja i pruža osnovu za očekivanje da će se ista veza uzroka i posledice uvek ponoviti pod istim uslovima. Nauka pretpostavlja da uzročno-posledična veza prožima sve pojave u prirodnoj i društvenoj stvarnosti, i da svaka posledica ima svoj uzrok. Zadatak je pojedinih nauka da na pretpostavci uzročnosti istraže i ustanove konkretne uzročno-posledične veze u iskustvenom materijalu[8].

Zapažanje da uzročna povezanost nema u svim oblastima bića isti karakter uslovilo je razlikovanje više tipova kauzaliteta (Pavlović, 1997):

- *Fizički ili mehanički kauzalitet* – kada uzrok izaziva učinak dodirom, sudarom ili nekim drugim dejstvom čiji se intenzitet može izračunavati ili meriti;
- *Organski ili biološki kauzalitet* – objašnjava vezu uzroka i posledice u organskoj prirodi;

[7] Setite se od čega smo pošli - šta je sa sastavom tla, prethodnom zasićenošću tla vodom, spoljašnjom temperaturom, od evapotranspiracijom...?
[8] *kauzalitet*, Filozofijski rečnik, Matica Hrvatska, Zagreb 1984.

- *Teleološki*[9] *kauzalitet* – koji ne podrazumeva da uzrok mora vremenski prethoditi učinku, već da krajnja svrha ima dejstvo na tok i prirodu događanja. Mnogi autori odriču zasnovanost teleološkog kauzaliteta, naročito pristalice mehaničkog kauzaliteta;
- *Psihološki kauzalitet* – koji utvrđuje vezu motiva i delanja u ljudskom ponašanju i omogućava objašnjenje, predviđanje i razumevanje individualnog ili kolektivnog ponašanja.
- *Teološki ili voluntaristički kauzalit*et – koji smatra da su efekti nepredvidljivi ukoliko su uzroci dati u obliku neke slobodne volje. Mnogi naučnici osporavaju ovaj tip kauzaliteta.

Za nas su, u tehničkim naukama, dakako, od najvećeg interesa fizičke odnosno mehaničke kauzalnosti.

Bizarne korelacije

U nastojanju da svojim studentima na primerima pokažu kako postojanje visokog koeficijenta korelacije još uvek ne znači postojanje kauzalnosti između dve pojave, američki profesori statistike uradili su brojna „istraživanja“ na realnim podacima kojima su preko koeficijenta korelacije „dokazali“ sledeće:

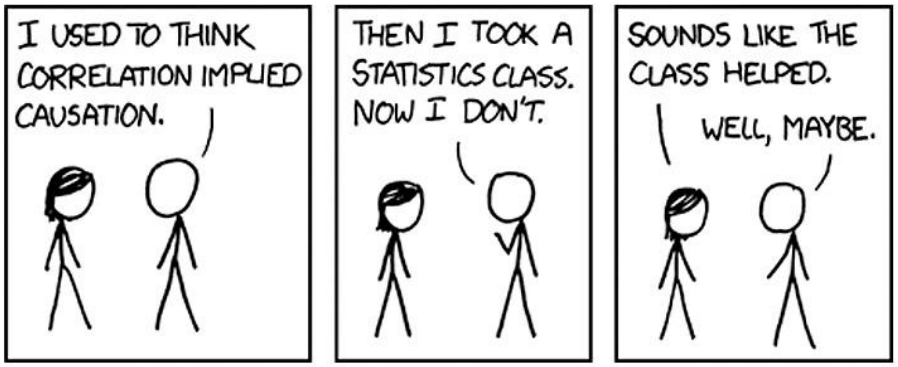
- Povećana potrošnja sladoleda navodi na ubistvo.
- Organska hrana izaziva autizam.
- Uvoz meksičkog limuna je odlična prevencija saobraćajnih nesreća na autoputevima.
- Fejsbuk je izazvao grčku dužničku krizu.
- Broj davljenja u bazenima odlično korelira sa brojem filmova u kojima igra Nikolas Kejdž.
- Povećanje cene čipsa krivo je za povećanu smrtnost ljudi koji ispadnu iz invalidskih kolica.

[9] Teleologija je filozofska doktrina koja smatra da sve što postoji i što se događa ima nekakvu svrhu.

I tako u beskrajno. Možemo i da izvrnemo ove „naučne dokaze“, zamenimo x i y osu i evo novih „otkrića“:

- Ubice posle izvršenog zločina navale na sladoled.
- Osobe sa autizmom obožavaju organsku hranu.
- Saobraćajne nesreće na autoputevima izazvane su nedostatkom meksičkog limuna.
- Grci su se, usled nedostatka novca, masovno okrenuli Fejsbuku.
- Filmovi Nikolasa Kejdža nagone ljude da se bacaju u bazene i dave.
- Povećana smrtnost ispadanjem iz invalidskih kolica izaziva kod ožalošćenih pomamu za čipsom, na šta proizvođači reaguju povećanjem cena.

Jednako je smešno i jednako besmisleno već na prvi pogled. Upravo zato je odlična ilustracija da korelacija nipošto automatski ne znači kauzalnost[10].



[10] I zapitajmo se koliko sličnih, samo naoko manjih besmislica, svakoga dana pročitamo na stranicama štampe.

U životu, naročito u životu mladog naučnika koji planira da otkrije nešto novo, nije sve tako očigledno. Loša korelacija svakako znači da pojave koje razmatra nisu uzročno posledično vezane, odnosno da je na krivom putu. Sa druge strane, dobra korelacija znači da put *može biti* ispravan. Ili, jezikom matematike, dobra korelacija je potreban, ali ne i dovoljan uslov da kauzalitet postoji.

Umesto zaključka

Kada vršimo eksperimente i analiziramo rezultate često brkamo koncepte korelacije i kauzaliteta. Na prvi pogled slično je. I korelacija i kauzalitet zahtevaju po jednu nezavisno i zavisno promenljivu veličinu. U eksperimentu, nezavisno promenljivu predstavlja skup podataka koji mogu biti zadavani i kontrolisani voljom onoga koji ekspeiment vrši. Zavisno promenljivu čini skup podataka koji nastaje pod uticajem spoljnih faktora. Gde spoljni faktor može biti nezavisno promenljiva, ali i ne mora!

Ako pokažemo da postoji zadovoljavajuća korelacija, ponekad možemo pretpostaviti da se zavisno promenljiva menja *isključivo* pod dejstvom nezavisno promenljive. Da li smo u pravu? Tu počinje naučni zaplet.

Kakogod, postoji razlika između *uzroka i posledice* (kauzacija) i *odnosa* dve promenljive (korelacija).

Mislite o tome dok planirate eksperiment, a naročito pred dobijenim statističkim „dokazima“. Stvarni dokaz verovatno će voditi kroz nove eksperimente.