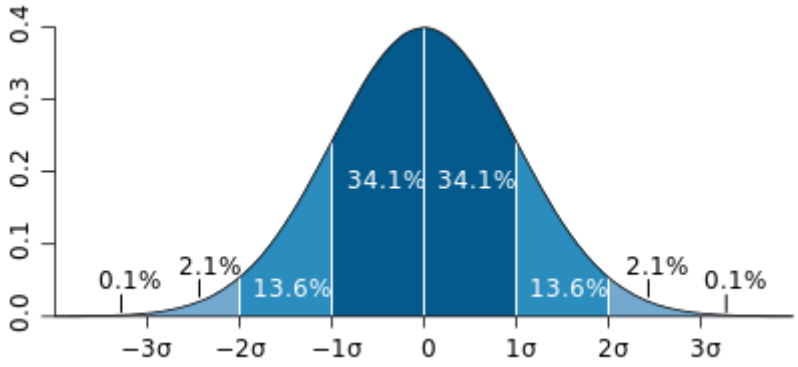
STATISTIKA

Napomena: Ovaj dokument je prilog prezentaciji ''statistika'' gdje su već dati neki osnovni pojmovi, ovdje ih razrađujemo i nadopunjujemo, zato prije čitanja ovog dokumenta preporučujem da se pogleda spomenuta prezentacija.

U kontekstu logike, statistika nam je potrebna da 'sredimo' rezultate dobivene pomoćnim metodama indukcije (napose brojanjem i mjerenjem), a dio o vjerojatnostima je relevantan zbog ispravnog tumačenja dobivenih rezultata. Kad imamo jako puno rezultata brojčano izraženih (neovisno o izvoru), trebamo ih biti u stanju nekako skraćeno prezentirati, a to, u elementarnom obliku, radimo preko *mjera centralne vrijednosti* i *mjera raspršenja.* Prve nam govore kako su rezultati grupirani oko centra, a drugi distribuciju odstupanja od centra (najčešće - prosječno odstupanje od prosjeka).

I.

Svi smo se negdje susreli s poznatom Gaussovom krivuljom (koja predstavlja grafički prikaz normalne distribucije rezultata):



Slika preuzeta s: https://bs.wikipedia.org/wiki/Standardna_devijacija

Što ovo zapravo znači? Znači da rezultati imaju tendenciju ravnomjerno se raspoređivati oko centra (u ovom slučaju aritmetičke sredine –x), a standardna devijacija nam pokazuje prosječno odstupanje od prosjeka, odnosno vjerojatnost da će neki, nasumično odabrani broj u nizu, biti u određenom rasponu (kako vidimo, 68% rezultata je unutar prve standardne devijacije, 95% unutar dvije, a 99% u 3, to je važno kod procjene statističke greške u različitim anketama i sl. ...)

Mnogi rezultati istraživanja dobiveni pomoćnim metodama indukcije imaju oblik ove zvonolike krivulje ili podsjećaju na istu (recimo distribucija IQ-a u populaciji, visina, težina i sl.).

Konkretno, otprilike 50% rezultata je iznad prosjeka, 50% ispod, ako se radi o pravilnoj distribuciji rezultata.

Prvo i osnovno pitanje je, kako ćete vi iz nekog niza, pogotovo ako predstavlja samo slučajni uzorak, 'znati' da populacija tendira zvonolikoj krivulji?
Već smo kod pomoćnih metoda indukcije i metoda znanosti vidjeli što znači reprezentativnost uzorka i zašto treba voditi računa o metodologiji izbora uzorka.
Da ne kompliciramo, pretpostavimo da je neki izabrani uzorak zaista reprezentativan za populaciju.
Na početku, rezultati istraživanja koje ste dobili (vlastitim istraživanjem ili iz sekundarnih izvora) su samo hrpa nesređenih brojki. Iz tog niza brojki ovi podaci nam najčešće mogu pomoći u tumačenju tih rezultata:

A) MJERE CENTRANE VRIJEDNOSTI

1. Mode (Mod- M) - broj koji se najčešće pojavljuje u nekom nizu, recimo u nizu 1,1,2,3,5 mod (M) je 1, u nizu 1,2,3,4,5 moda nema, a u nizu 1,1, 3, 2,2, 5 imamo 2 moda (1 i 2).

2. Aritmetička sredina (Ẋ) – Broj koji dobijete tako što zbrojite vrijednost članova nekog niza i podijelite s brojem članova, npr. u nizu 1,2,3, 4, 5 zbroj vrijednosti je 15, broj članova je 5, dakle, aritmetička sredina je 3…

3. Medijan (Md) – Broj koji neki niz, poredan po veličini, dijeli točno na pola, npr. 1,4, 2, 3, 1, 5. Md= 2,5 (1,1,2,3,4,5, 2+3/2= 2,5). Kad imate neparan broj brojeva, lako je odrediti medijan, nakon što ste niz poredali po veličini. Ako je paran broj članova u nizu, uzmete 2 'središnja' broja, u našem slučaju to su 2 i 3, njihov prosjek je vaš medijan…

Što ovi brojevi znače? - Ako postoje značajne varijacije između moda, medijana i aritmetičke sredine, već sama ta činjenica vam govori da je lako moguće da distribucija rezultata nije pravilna, sigurni ste da su rezultati oko centra grupirani nepravilno.

Npr. kod prosječne težine ili IQ-a, opaziti ćete da između moda, medijana i aritmetičke sredine nema velikih razlika, iz toga se često događaju pogreške da ljudi distribuciju plaća često zamišljaju kao zvonoliku krivulju. Npr. znaju da je prosječna plaća u RH oko 5 500 kn i pretpostavljaju da to znači da otprilike 50% ljudi prima veću plaću od toga, 50% manju. Kad znate da je mod oko 3 600 kn, a medijan 4100, već naslućujete istinu (da distribucija plaća ne odgovara Gaussovoj krivulji i da je gore prisutno mišljenje - zabluda). Isto vrijedi za distribuciju plaća u BiH, kao i u većini zemalja. Zato je važno poznavati mjere centralne vrijednosti.

Vrlo često iz mjera centralne vrijednosti ne možemo saznati sve relevantne informacije o nekom nizu, napose kad uspoređujemo neke nizove različitih brojčanih vrijednosti, zato se računaju i mjere raspršenja.

B) MJERE RASPRŠENJA

Neovisno o tome je li distribucija pravilna ili ne, mjere raspršenja nam daju dodatne informacije o nizu (ako je pravilan, hoće li zvonolika krivulja biti 'spljoštena' ili ne, ako je nepravilan, kakve su i kolike te nepravilnosti. Oni koji su od vas malo pozornije pratili statistike zaraze korona virusom, često su mogli čuti izraz 'izravnati krivulju' – konkretno i s mjerama i bez njih će broj zaraženih bez cjepiva biti isti, ali vrijeme zaraze će se produžiti, time i krivulja 'spljoštiti' pa zdravstveni sustav neće biti pretrpan, to je barem bila ideja, a kad pročitate naredne retke, iz podataka o broju umrlih i zaraženih lako ćete zaključiti koliko uspješno je realizirana)…

1. Raspon varijacije (V) – Broj koji dobijete tako što oduzmete broj najmanje vrijednosti u nekom nizu od broja najveće vrijednosti. U nizu 1,1, 2,4, 5, 3 V= 4 (5-1).

2. Standardna devijacija (S) – Izražava prosječno odstupanje rezultata od prosjeka.
U nizu 1, 2, 3, 4, 5 S je = (2+1 +0 +1 +2 = 6/5= 1.2). (Aritmetička sredina u danom nizu je 3, prvi broj odstupa od nje za 2, drugi broj za 1, treći broj ne odstupa, četvrti broj odstupa za 1, peti broj odstupa za 2. Obzirom da računamo prosječno odstupanje od prosjeka, u ovom slučaju aritmetičke sredine, zbrojimo ta odstupanja i podijelimo s brojem članova, i tako dobijemo gore navedeni rezultat).

Zašto je ovo bitno? Uzmimo za primjer da su gore navedeni brojevi ocjene Vašeg školarca. Da ima 5 trojki, S = 0, odnosno nema odstupanja od prosjeka, a aritmetička sredina je ista (3). Vama S ustvari govori ima li varijacija u nizu i kolike su one u *prosjeku.* To nam kaže i koliko je neki rezultat udaljen od centra u terminima standardnih devijacija (što vidimo na priloženoj Gaussovoj krivulji).

3. Koeficijent disperzije (K) – Predstavlja omjer standardne devijacije i aritmetičke sredine. (S/X) Ovo je iznimno važno pri uspoređivanju različitih nizova, S i X i usporedbi varijacija u tom nizu. U našim primjerima, S i X u apsolutnom iznosu nikad ne prelaze 5 jer smo uzeli relativno male nizove.
No što ako želimo usporediti varijabilnost ocjena Vašeg školarca (1,2,3,4,5) s varijabilnošću posjećenosti utakmica Hajduka u sezoni 2017/2018? (Najčešće uspoređujemo neke nizove koji su nam relevantni za usporedbu, ovakve nizove najčešće nemamo razlog uspoređivati, no navedeni su kao ilustracija, da možete usporediti bilo što).
Recimo da je prosječna posjećenost (X) Hajdukovih utakmica 30000 navijača, S je 1200. Znači li to da je varijabilnost u posjećenosti 1000x veća, nego u navedenom nizu ocjena?

NE. Pravi razmjer varijabilnosti nam otkriva K, u prvom slučaju: 1,2/3 (S/X) K=0.4, u drugom slučaju: 1200/30000 = 0.04, odnosno varijabilnost u drugom nizu je 10X manja nego u prvom!

Ove statističke mjere nam puno kažu o nekom nizu, čak i kad ne znamo sve rezultate u tom nizu, to je osobito bitno kad npr. imamo niz od milijun brojki, pamtiti i prezentirati SVE rezultate bi bilo vrlo nepraktično, uz pomoć ovih veličina već imamo hrpu relevantnih informacija o samom nizu, a treba nam samo 6 brojeva (ako nema preklapanja)…

II.
OSNOVE VJEROVATNOSTI

A) Jednostavne vjerojatnosti

Većina nas je u stanju izračunati jednostavne vjerojatnosti nekih događaja, npr. u igrama na sreću – Šansa da će doći 'glava' pri bacanju novčića je 50% (dvije su varijante, pismo ili glava i svaka ima jednaku vjerojatnost). Šansa da dobijemo 5 na standardnoj kocki je 1/6 ili otprilike 16,6 %. No i ove vjerojatnosti često nas znaju dovesti u zabludu, npr. kad želimo znati kolika je šansa da 3 puta zaredom dobijemo 5 na standardnoj kocki ili da 3x zaredom dođe 'glava'. Točno je da u svakom pokušaju imamo jednake izglede (1/6, odnosno ½), no za izračun ovakvih vjerojatnosti se mora uzeti u obzir nešto što se u statistici zove ***zakon multiplikacije***.
Šansa prvog događaja je 1/6x1/6x1/6 = 1/216 odnosno samo 0,46%!
U drugom slučaju su nam malo bolji izgledi: 1/2x1/2x1/2 = 1/8 = 12,5 %!
Šansa da ćemo iz 'standardnog' špila od 52 karte izvući As karo je 1/52… Šansa da ćete izvući A, K, Q ili J karo (bilo koju od tih karata) jeste 4/52 odnosno 1/13, ovo se naziva ***zakon adicije*** (1/52 + 1/52 +1/52+1/52).
Kad analizirate nepovezane događaje i kalkulirate šanse da vam se 'posloži', upotrebljavate zakon multiplikacije, na temelju takvih izračuna se često prezentiraju šanse da je život nastao slučajno i sl.

B) Složene vjerojatnosti

Kako izračunati vjerojatnost pojavljivanja određenih brojeva ili kombinacija u nekom nizu brojeva? Za razbuđivanje, možete učenicima ponuditi ovaj zadatak:
Procijenite koji niz u bacanju novčića, kad isti bacamo 6 puta, je najvjerojatniji (P pismo, G glava):

A) P P P G G G
B) P G P G G P
C) P P P P P P
D) G G G G G G

Odgovor: Svaki niz je jednako vjerojatan. Ovdje se isprepliću onaj subjektivni osjećaj vjerojatnosti i objektivna statistika – kako je moguće da je niz D jednako vjerojatan kao i niz B, kad smo već naučili zakon multiplikacije i znamo da je šansa za niz D 1/64?! (Sjetite se pravilnih modusa silogizma i računanja mogućih kombinacija tamo)!

Stvar je u tome da svaka, točno zadana kombinacija, ovdje ima jednaku vjerojatnost pojavljivanja (1/64) jer i u nizu B imamo točno zadan redoslijed, ali intuitivno (subjektivno) pretpostavljamo da će u nizu biti izmiješane glave i pisma i to nepravilnim redom. Ovo je nešto složenija statistika jer treba uočiti da se ne pita kolika je vjerojatnost pojavljivanja barem jednog pisma (ili glave) u 6 pokušaja.

Kad bi se to pitalo, račun bi išao ovako: $1 – (1/2)^6$ i dobiti ćemo preko 98%! Zato ljudi često C i D niz odmah eliminiraju kad odgovaraju na pitanje iz uvoda.

Dakle, ako ste krivo odgovorili na uvodno pitanje, nije tragedija, to se jako često događa (i često je pokazatelj određenog iskustva s praktičnim određivanjem vjerojatnosti).

Opća formula kod ovakvih problema: od 1 oduzeti vjerojatnost da se vaš događaj neće dogoditi na potenciju broja pokušaja

Npr.: šansa da u 4 bacanja jednom dođe 3?
$1- (5/6)^4$ = cca 52% ** (5/6 je vjerojatnost da neće doći 3 u bacanju, 4 bacanja su u pitanju zato na četvrtu…) Na sličan način možete izračunati bilo koje šanse određene brojem pokušaja i šansom primarnog događaja.

Šansa da u 5 bacanja s dvije kocke dođe zbroj 7? $1- (30/36)^5$ = približno 0, 60 (60%)… (Ovdje budite pažljivi, druge kombinacije mogu imati drugačije šanse!)

Šansa da u 3 pokušaja izvučete bilo kakvog kralja iz standardnog špila karata od 52 karte (gdje prilikom neuspješnog pokušaja vraćate kartu špil pa ne vrijedi zakon adicije): $1-(12/13)^3$, kad ne bi vraćali kartu u špil, račun bi izgledao ovako: 1/13 + 4/51 + 2/25 Dobrim matematičarima koji poznaju ovo nije preteško računati vjerojatnosti pojavljivanja mnogih kombinacija karata, zato su među rijetkim nepoželjnim gostima kockarnica!

Kako procijeniti vjerojatnost dobitka na lotu, ako igrate 7/39? Ako nije zadan redoslijed brojeva, već samo trebate pogoditi brojeve, onda je vaša vjerojatnost:
1 : 39!/ 7! (32!)* (otprilike 1 naprema 15 milijuna) (generalna formula za ovakve slučajeve: n!/ r!(n-r)! , ako je zadan redoslijed, onda bi to išlo: 1/39 x 1/38 x 1/37… (što su znatno manje šanse). Objašnjenje principa ovog računa bi uzelo previše vremena i previše je komplicirano za razinu statistike potrebne prosječnom čovjeku, pa je navedeno samo kao ilustracija činjenice da se preko statistike svašta da izračunati, iako nam se na prvi pogled ne čini tako.

- *! Označava faktorijelu, recimo 4! = 4x3x2x1
- ** Učenici često pitaju zašto u primjeru s kockom ne vrijedi zakon adicije, ako mi je šansa u prvom pokušaju 1/6, zar u 2 pokušaja nije 1/6 + 1/6?

Ne! Jer bi po takvoj statistici iz 6 pokušaja morali dobiti 6, vjerojatnost bi bila 1. (odnosno 100%), zato se ovakve vjerojatnosti računaju tako da od 1 (sigurnosti), oduzmete vjerojatnost pojavljivanja negativnog ishoda na potenciju broja pokušaja, tako da teoretski u 444 pokušaja, nijednom ne morate dobiti 6, iako su šanse za to 1 – $(5/6)^{444}$, odnosno vrlo male, ali ne i nepostojeće!

*** Neke učenike posebno zanima kako se određuju koeficijenti na kladionici, ti koeficijenti su rezultat objektivnih statistika i subjektivne procjene jer vjerojatnost ishoda 1 ili X ili 2 nije ovdje 1/3, kao što bi bilo da se radi o potpuno slučajnoj vjerojatnosti - Poker, kladionice i slične preokupacije su često tako popularne upravo zato što su kombinacija igračevog znanja i čiste sreće, u omjeru koji je ponekad teško statistički izraziti…

Na državnoj maturi znaju se pojaviti zadaci iz biologije koji uključuju primjenu logike i statistike, primjerice, kolika je šansa da će od troje djece bar jedno biti žensko?
Odgovor: 7/8 (87,5%).
(Kako smo dobili gornji broj? Šansa da budu sva muška djeca su 1/2x1/2x1/2, odnosno 12,5% pa je lako iz toga zaključiti rezultat, šansa bilo kojeg zadanog redoslijeda djece, tipa Ž,M,Ž bi bila 12,5% . Čak i ako vam ne ide matematika, možete ispisati svih 8 mogućih kombinacija kad imate troje djece i lako uvidjeti zašto je ovo ovako).
Šansa da u 3 nepovezane obitelji koje imaju po troje djece, bude bar po jedno žensko? 7/8x7/8/x7/8 (primjena zakona multiplikacije). Kombinacijama adicije i multiplikacije možete i računati šanse da bude po x djece sa zadanom bojom očiju (ako znate boju očiju roditelja)…

Završna napomena: Nije mi cilj stvoriti novu, statistički osviještenu, generaciju kockara i kladioničara, ali primjeri s kartama ili kockom su ljudima najbliži jer teško da netko nije igrao čovječe ne ljuti se ili neku kartašku igru…

Mile Logara