OSNOVE STATISTIKE

Definicija i podjele

Statistika je znanost o prikupljanju, organiziranju, interpretiranju i prezentiranju podataka i dijeli se na **teorijsku** i **primijenjenu** statistiku.

Teorijska statistika bavi se razvijanjem statističkih metoda, dok se **primijenjena statistika** bavi primjenjivanjem statističkih metoda koje je razvila teorijska statistika na konkretnim problemima.

Statistiku možemo grubo podijeliti i na:

- **Deskriptivnu statistiku**: koja **opisuje uzorak** uređivanjem, grupiranjem, tabeliranjem i grafičkim prikazivanjem njegovih vrijednosti (sredine, medijana, maksimalne, minimalne vrijednosti,…).
- **Inferencijalnu statistiku**: koja **na temelju uzorka** (dijela informacija, podskupa) donosi vjerojatnosne sudove o **cjelini** (populaciji, osnovnom skupu) koja je nama nepoznata. **Parametar** je mjera koja opisuje cijelu populaciju, koji je uglavnom nepoznat, budući da rijetko možemo promatrati cijelu populaciju. Parametre često označavamo grčkim slovima. **Statistika** je mjera izračunata na temelju uzorka kao procjena populacijskog parametra. Statistika uzorka se često označava s latinskim slovima.

Statističke zamke

1. **Donošenje zaključaka o velikoj populaciji na temelju malog uzorka** –rečenica tipa “Moj je ujak pušio cigarete cijeli život i doživio je 90 godina, stoga pušenje nije štetno” nije statistički značajno. Ako 10 dana pratimo vezu između naše glavobolje i primijetimo da kada pada kiša boli nas glava a kada je sunčano ne boli, također ništa statistički ne dokazuje jer jednostavno imamo premali uzorak. Srećom statističari su razvili jasna pravila o veličini uzorka na temelju koje možemo statistički značajno zaključivati
2. **Donošenje odluka na temelju uzorka koji nije slučajan** – rečenica “Rock zvijezde umiru mlade; vidi Buddy Hollya, Jimi Hendrixa, Janis Joplin, Jima Morrisona, Kurt Cobaina“ nije slučajan uzorak jer smo u uzorak uključili samo one rock zvijezde koje i jesu umrle mlade. Mora se paziti i kada generaliziramo na temelju povijesnih studija npr. ljudi koji su preživjeli srčani udar ako nemamo podatke o onima koji nisu imali srčani udar.
3. **Pridavanje važnosti rijetkim događajima iz velikog uzorka** – Malo vjerojatni događaji mogu se desiti ako se uzme dovoljno veliki uzorak tako npr. rečenica “Moj je šef dobio na lutriji, njegov je sustav nepogrešiv” upada u statističku

zamku, jer mnogo ljudi igra lutriju, netko mora i dobiti, što ništa ne govori o njegovom sustavu igranja, već je proizvod vjerojatnosti.

4. **Korištenje loših istraživačkih metoda –** loša istraživačka metoda može donijeti do krivih rezultata tako npr. ako profesor u razredu pita studente "Tko zna kako se rješava problem optimizacije pomoću Lagrangeovog multiplikatora?" dobit će mali potvrdni odaziv jer se studenti često ne vole eksponirati pred kolegama kako ne bi ispalo da se prave „važni". U tom bi slučaju bilo bolje dobiti odgovore kroz neku drugu metodu ispitivanja, npr. kroz anonimni upitnik.
5. **Pretpostavljanje uzročno/posljedične veze na temelju opažanja** – asocijacija između dvaju pojava ne dokazuje ujedno i uzročnost; tako npr. izjava "Morski psi najčešće napadaju između 12 i 14 sati" može biti kriva ako se uzme u obzir da se ljudi najčešće i kupaju u tom vremenskom razdoblju, stoga ne čudi da je u tom razdoblju bilo i najviše napada morskih pasa. Da bi se došlo do zaključka da morski psi češće napadaju od 12-14 sati u odnosu na ostale dijelove dana trebalo bi provesti strogo kontrolirani eksperiment.
6. **Generaliziranje na pojedinca temeljeno na opažanjima o grupi** – rečenica "Muškarci su viši od žena" vrijedi samo u statističkom smislu, u prosjeku. Postoje mnoge žene koje su više od većine muškaraca. Umjesto da se koncentriramo na statističku generalizaciju (srednja visina) u tom bi slučaju bilo bolje koncentrirati se na koliko ima preklapanja između grupa.
7. **Pridavanje praktične važnosti svakom statističkom značajnom istraživanju** – statistička značajnost ne mora odražavati i praktičnu važnost. Npr. istraživanje provedeno od Austrijske vojske na 500 000 regruta otkrilo je da regruti rođeni u proljeće viši su u prosjeku za 0.6 cm u odnosu na one rođene u jesen. Hoće li taj podatak promijeniti planove roditelja kako bi imali 0.6 cm višeg sina? Ili, otkrilo se da novi model tvrdog diska može bez greške raditi 10% dulje od starog modela koji je bez greške mogao kontinuirano raditi 160 godina. Iako statistički značajno, ima li to praktičnu vrijednost u donošenju odluke o kupnji tvrdog diska?

Podaci

Statistički podaci prikazuju se u obliku varijabli. **Varijabla** je svojstvo sudionika ili situacije koja poprima **različite** vrijednosti. Npr. spol je varijabla jer ima dvije vrijednosti (modaliteta) M/Ž, starost je varijabla ako imamo sudionike različitih starosti, itd. Etnička grupa nije varijabla ako u istraživanju sudjeluju samo Europljani, ili spol nije varijabla ako u istraživanju sudjeluju samo žene; tada su to **konstante.**

Prema **ulozi** koju poprimaju **u modelu** varijable se dijele na:

- **Nezavisne varijable** koje u ovisnosti o vrsti analize zovemo i faktorima, kovarijatima, tretmanskim varijablama, egzogenim varijablama
- **Zavisne varijable** koje zovemo i endogenim varijablama mjere utjecaj nezavisnih varijabli

Deskriptivna statistika

Mjere centralne tendencije

- **Aritmetička sredina** – najčešća je mjera centralne tendencije. Ona je zbroj vrijednosti koje poprima varijabla podijeljeno s brojem opažanja:

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

Aritmetička sredina je točka ravnoteže distribucije jer ima svojstvo da udaljenosti između sredine i vrijednosti koje poprima varijabla je uvijek nula, tj.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

za populaciju kao i za uzorak neovisno o izgledu distribucije (asimetrija, zaobljenost,…). Prikladna je za normalno distribuirane numeričke varijable.

- **Medijan –** središnja je vrijednost nakon što smo ih poredali po veličini. Posebno je koristan kada imamo ekstremne, netipične vrijednosti i često se koristi kao mjera centralne tendencije ordinalnih varijabli**.**
- **Mod** – je vrijednost koja se najčešće ponavlja kod određene varijable. Jedina je mjera centralne tendencije koja ima smisla za kategorijalne (kvalitativne) varijable. Ponekad je mod koristan i za opisivanje diskretne numeričke varijable s malim brojem vrijednosti, kao npr. trostupanjske Likertove ljestvice.

Mjere disperzije

- **Raspon varijacije (Range)** – je razlika između najveće i najmanje vrijednosti koje varijabla poprima. To je najgrublja mjera disperzije. Osjetljiva je na ekstremne vrijednosti, zato je prikladnija za ordinalne varijable nego li za ostale numeričke varijable.
- **Standardna devijacija**: zovemo je i korijen drugog momenta oko sredine jer:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

Standardna je devijacija nenegativna jer su devijacije oko sredine kvadrirane. Ako su sve vrijednosti jednake sredini tada je standardna devijacija jednaka nuli. Standardna se devijacija može uspoređivati samo između varijabli koje su izražene u istim jedinicama mjere i ako sredine između varijabli nisu jako različite.

Koeficijent varijacije (CV) – Koeficijent varijacije je relativna mjera disperzije koja može biti izražena kao postotak sredine pomoću:

$$CV = 100 \times \frac{s}{\bar{x}}.$$

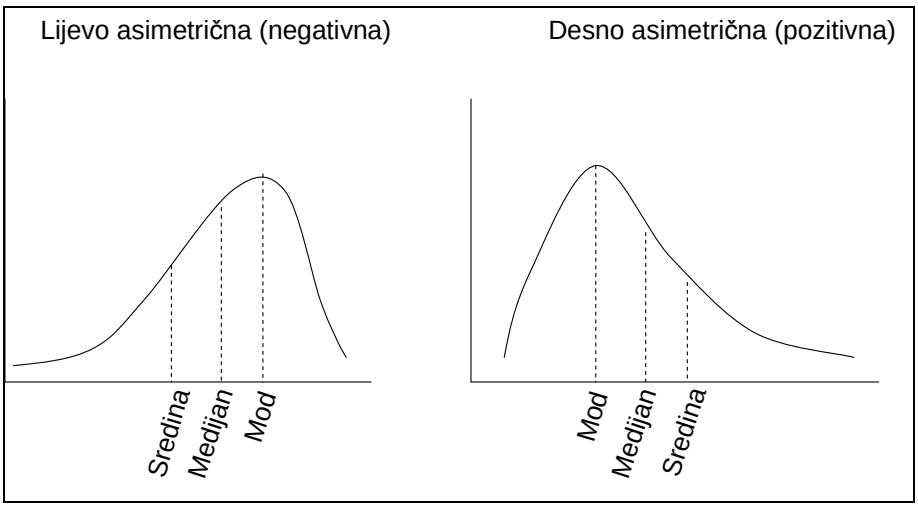
Koristi se za usporedbu disperzije u podacima između varijabli koje imaju različite jedinice mjere ili bitno različite sredine. **Slabosti** su mu da nije definiran ako je sredina jednaka nuli ili ako je negativna. Ako je standardna devijacija varijable veća od njezine sredine tada koeficijent varijacije je veći od 1 ili veći od 100 ako je izražen u postotku.

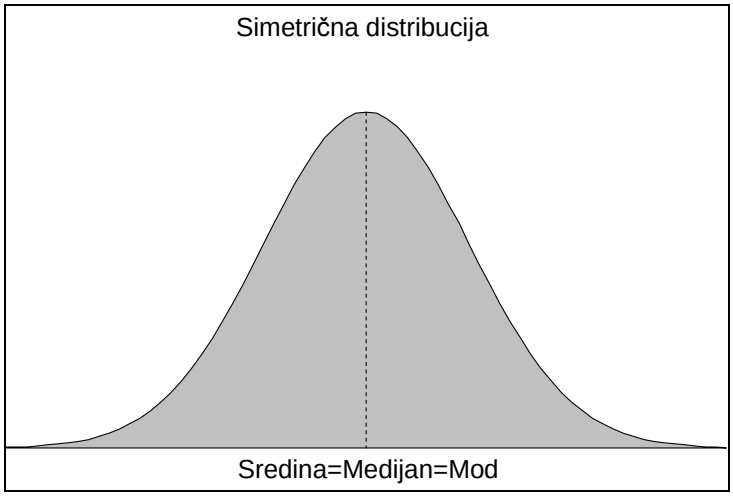
Mjere asimetrije i zaobljenosti

- **Mjera asimetrije (Skewness)** – mjeri asimetriju distribucije pomoću trećeg momenta oko sredine:

$$Skewness = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3 / n}{s^2}.$$

Smatramo da nemamo problema s asimetrijom distribucije ako je vrijednost asimetrije u intervalu između [-1,1]. Ako je manja od -1 tada govorimo o negativnoj asimetriji ili lijevo asimetričnoj distribuciji, ako je veća od 1 tada govorimo o pozitivnoj ili desno asimetričnoj distribuciji. Ako promatramo ponašanje mjere centralne tendencije u ovisnosti o simetriji distribucije imamo sljedeće situacije:





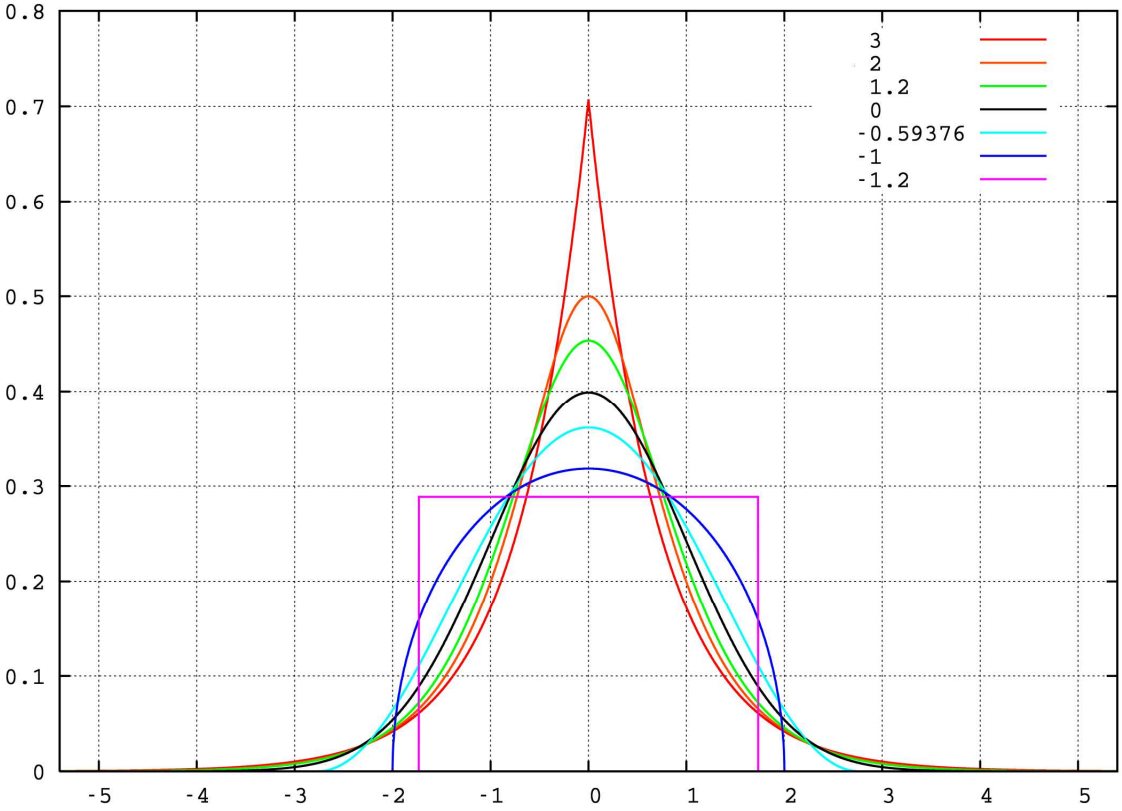
U slučaju potpuno simetrične distribucije aritmetička sredina je jednaka modu i medijanu, dok u slučaju lijevo asimetrične distribucije sredina će biti manja vrijednosti od medijana koji će biti manji od moda, i obrnuto, kod desno asimetrične distribucije sredina će biti veće vrijednost medijana a medijan će biti veći od moda.

- **Mjera zaobljenosti (Kurtosis) –** mjeri zaobljenost distribucije pomoću četvrtog momenta oko sredine:

$$Kurtosis = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 / n}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 / n \right)^2}.$$

Budući da imamo vrijednosti na četvrti eksponent, zbroj je uvijek rastući i Kurtosis je uvijek pozitivan. Često se izračunava tzv. **Excess** Kurtosis koji umanjuje Kurtosis za 3, tj. Excess Kurtosis=Kurtosis-3; na taj se način izjednačava Excess Kurtosis normalno distribuirane varijable s nulom te možemo zaključivati slično kao i kod mjere asimetrije; ako je Excess Kurtosis u intervalu [-1,1] smatramo da varijabla je približno normalno distribuirana, dok ako je izvan tog intervala smatramo da imamo problema sa zaobljenošću distribucije. Ako je Excess Kurtosis veći od jedinice imamo šiljastu distribuciju, a ako je manje od -1 tada imamo tupu distribuciju.

Izgled distribucije za različite vrijednosti Excess Kurtosis-a



Inferencijalna statistika

Inferencijalnom statistikom:

- Procjenjuju se populacijski **parametri** na temelju **statistike** uzorka (sredina, proporcija, varijanca, koeficijenti regresijskog pravca,…).
- Testiraju se **hipoteze**

Procjena populacijskih parametara

- **Procjena točke**; procjena vrijednosti parametra
- **Procjena intervala**; procjenjuje se raspon vrijednosti koji se zove **interval pouzdanosti.**

Istraživačko pitanje nema unaprijed definiranu **tvrdnju**. Istraživačko pitanje je npr. „Postoji li razlika u ocjenama iz ekonometrije između studenata koji su dobili bolje i lošije ocjene iz statistike?“. Postoje:

- **Pitanja razlike**: uspoređujemo razlike u vrijednostima varijable između dvije ili više grupa.
- **Pitanja asocijacije:** analiziramo kreću li se varijable zajedno, postoji li između njih jaka veza, utječu li jedna na drugu, možemo li na temelju kretanja jedne varijable predvidjeti kretanje druge varijable?

Hipoteza istraživanja je **unaprijed** određena **tvrdnja** o vezi između varijabli kao npr.: „Studenti koji su dobili bolju ocjenu iz statistike imaju veću ocjenu iz ekonometrije“, ili „Prosječan IQ muškaraca je određena vrijednost različita od 100“, pa se hipoteza u klasičnoj statistici ili **odbacuje** ili **ne odbacuje**.

Statistička hipoteza: Hipoteza istraživanja se preuređuje u **dvije** statističke hipoteze:

- **Nultu hipotezu, H_0** koja tvrdi da **NEMA efekta ili nema razlike**. Često istraživač očekuje da će je moći odbaciti.
- **Alternativnu hipotezu, H_1** koja označava **prisutnost efekta ili razlike**, npr. ako je H_0: IQ=100 tada H_1: IQ≠100.

Ako nema razlike (efekta) tada može se tvrditi da sredina uzorka prihvatljivo blizu 100. Prisutnost efekta znači da je sredina razumno različita od (ispod ili iznad) 100.

Usmjerenost alternativne hipoteze

Statistička hipoteza H_0:IQ=100 ima alternativnu hipotezu H_1:IQ≠100 koja **nije usmjerena**, ne govori nam o smjeru u kojem ide IQ (veći ili manji od 100). Alternativna hipoteza H_1:IQ≠100 govori samo da populacijska sredina nije jednaka 100.

Alternativne hipoteze H_1: IQ<100; ili H_1:IQ>100 su **usmjerene (ili se ponekad u literaturi zovu jednokračne alternativne hipoteze ili jednosmjerne)**, možemo odbaciti nultu hipotezu samo ako muškarci imaju IQ znatno manji od 100 (prvi slučaj) ili znatno veći od 100 (drugi slučaj). Ako npr. imaju znatno veći od 100 nećemo moći odbaciti usmjerenu alternativnu hipotezu H_1: IQ<100. Oko uporabe usmjerenih i neusmjerenih hipoteza nema konsenzusa među statističarima; jedni kažu da se mora uvijek primjenjivati neusmjerena alternativna hipoteza, neovisno o očekivanim vrijednostima. Drugi misle da ako nemamo a priori očekivanja o vrijednostima koristimo neusmjerenu, dok ako imamo određena očekivanja o vrijednostima tada koristimo usmjerenu. Jedna od prednosti usmjerene alternativne hipoteze je da lakše odbacuje nultu hipotezu (efekt, ili razlika ne moraju biti tako velike kao kod neusmjerene).

Testna statistika: na temelju inferencijalne statistike dobijemo **testnu statistiku,** koju uspoređujemo s kritičnim vrijednostima iz statističkih tablica (teorijska vjerojatnost), kako bismo zaključili o **statističkoj značajnosti.** Alternativni način zaključivanja je pomoću **razine značajnosti testa (*p* vrijednost)** koju moderni statistički softveri automatski prikazuju kraj testne statistike. Razina značajnosti testa pokazuje koja je vjerojatnost da smo učinili grešku ako odbacimo H_0.

Vrste statističkih grešaka

- **Greška prvog tipa:** takvu grešku činimo ako odbacimo hipotezu koja je bila istinita (tj. zaključi se da je alternativna “neistinita” hipoteza istinita). Istraživač sam kontrolira grešku prvog tipa birajući razinu značajnosti na kojoj se testira (α). Drugim riječima **razina značajnosti testa** (*p* vrijednost) pokazuje nam vjerojatnost da smo učinili grešku prvog tipa.
- **Greška drugog tipa**: takvu grešku činimo ako **ne** odbacimo hipotezu koja je neistinita. Vjerojatnost da se učini greška II tipa označavamo s β.