

“LAŽ je samostalnik, ki ga stopnjujemo:
otroška laž, laž, izjave politikov, statistika!”

0. UVOD

“Statistična obravnava vas dandanes spremlja vse življenje in še potem, začne se pa že pred rojstvom.”

Opredelitev: *statistika je veda, ki proučuje zakonitosti v množičnih pojavih.*

Pri tem *proučevanje* obsega vse aktivnosti, ki so s tem lahko povezane, od načrtovanja raziskav, opredelitve vsebine in ciljev pa ustrezne populacije, oblikovanja vzorca, izbire metod in njihovega razvoja, vse do prikazov in analize dobljenih rezultatov in njihove pravilne interpretacije.

Izvori statistike:

- a) *zgodovinsko*: potrebe, pogojene z organiziranjem družbe; dokazila že iz starih civilizacij.
- b) *formalno*: 19. stoletje, ko se ustalila uporaba izraza.
- c) *metodično*: matematična statistika in druge, predvsem naravoslovne znanosti.

V sodobnem času je generator silovitega razmaha uporabe statističnih metod tudi informacijska doba sama, s sočasnimi pojavom osebnih računalnikov in zmogljive programske opreme za delo z velikimi množicami podatkov, ne glede na področje udejstvovanja. Mnoge znanosti uporabljajo statistične metode kot osnovno orodje pri raziskavah.

Področja statistike:

- a) *družbena* statistika, ki je namenjena spremljanju za družbo pomembnih pojavov;
- b) *uporabna* statistika, kjer gre predvsem za uporabo statističnih metod; običajno prilagojena področju uporabe, npr. *medicinska st.*, *biostatistika*, *zavarovalniška st.*, itd.
- c) *matematična (teoretična)* statistika, ki razvija in proučuje metode, utemeljuje njihove prednosti in omejitve.

Našteta področja se prepletajo vse od ravni posameznika do ravni organizacij. V družbi in po organizacijah obstajajo posebne službe in uradi za statistične obdelave podatkov:

- SURS: Statistični urad republike Slovenije, s sedežem v Ljubljani;
- Eurostat: urad EC (Komisije Evropske skupnosti) za statistično spremljanje dogajanj v EU, sedež v Luxemburgu;
- statistične službe na SPIZ, DURS, ... in v javni upravi;
- statistične dejavnosti v okviru zavodov (šolstvo, zdravstvo, ...);
- statistično spremljanje proizvodnje, nabave in prodaje, servisiranja, po gospodarskih organizacijah;
- agencije in druge organizacije, ki statistično spremljajo javnost in medije, ...

Predavanja o statistiki obstajajo v Sloveniji na skoraj vseh študijskih smereh, na Univerzi v Ljubljani obstajata magistrski in doktorski študij statistike, slednji je namenjen predvsem raziskovalnim usmeritvam po različnih strokah.

Vrste statistike:

- a) *opisna* statistika: zbrane podatke ureja, razvršča in daje osnovne povzetke zbranih informacij s številiškimi in grafičnimi prikazi.
- b) *analitična* statistika: na osnovi zbranih podatkov ocenjuje parametre porazdelitev in opravlja preverjanje veljavnosti statističnih domnev.
- c) *statistično modeliranje*: na osnovi zbranih podatkov gradi modele, ki pojasnjujejo zveze med pojavi, njihov časovni potek ipd.

V delu statistikov se tovrstne dejavnosti prepletajo, uporabi ene vrste sledi druga itd.

Odnos do sorodnih ved

Zmogljivi računalniki omogočajo lahkotno delo z množicami podatkov. Tu se s statističnimi metodami srečujejo računalništvo, informatika pa tudi vse druge vede. Razvijajo se novi pristopi, metode in programska oprema, kot vedno pa vznikajo tudi nova vprašanja, problemi, naloge.

Predmet statistike:

osnovna množica je *populacija* (v dobesednem ali prenešenem pomenu, npr. populacija labodov, ki prezimujejo v Sloveniji, populacija Slovencev z visokim krvnim pritiskom, vse igre z izbrano napravo v igralnici, vse transakcije z izbrano delnico, nakupi v trgovski verigi, ...), njeni elementi so statistične *enote*. Le-te imajo več lastnosti, ki lahko zavzamejo različne vrednosti, so torej spremenljive in njihove "vrednosti" so nekako porazdeljene po populaciji. Te lastnosti imenujemo statistične *spremenljivke* (statistični *znaki*).

Glede na *zalogo vrednosti* (naravo) so statistične spremenljivke

- a) **opisne** (atributivne, kategorične), npr. spol, narodnost, blagovna znamka, barva,...
- b) **številске** (numerične), npr. dopolnjena leta starosti, število ljudi v dvorani, temperatura, trajanje smuka,...

Upošteva je *način merjenja* je pomembna razvrstitev spremenljivk v:

- a) **imenske** (nominalne): možne vrednosti so le imena (npr. narodnost, spol);
- b) **urejenostne** (ordinalne): glede na vrednosti je možno primerjanje in razvrščanje (npr. mnenjske sodbe, razvrstitev po dogovorjenem kriteriju, ...);
- c) **intervalske**: vrednosti so številске, a razmerja med vrednostmi nimajo vsebine, npr. temperatura, ocene pri pouku,...
- d) **razmernostne**: razmerja vrednosti so pomembna (npr. dolžina, teža, število ozdravljenih pacientov,...), običajno so vrednosti nenegativne.

Opisne spremenljivke so imenske ali urejenostne, številске pa intervalske, razmernostne ali urejenostne.

Glede na *moč* (velikost) *zaloge vrednosti* deli matematična statistika (po analogiji s teorijo verjetnosti) številске spremenljivke na *diskretne* (končno ali števno neskončno različnih vrednosti) in *zvezne* (zaloga vrednosti vsebuje z različnima vrednostima tudi vse vmesne vrednosti). Možne so tudi kombinacije obojih.

Pri razvrstitvi glede na način merjenja *količina informacije*, ki jo vsebujejo podatki, pada od razmernostnih preko intervalskih in urejenostnih do imenskih spremenljivk. Trajanje nekega dogajanja (čas) lahko obravnavamo kot zvezno razmernostno spremenljivko, a če zaokrožamo vrednosti, je le diskretna in glede na naš občutek morda celo le urejenostna (kratko, povprečno, dolgo trajanje).

Posledica tega je, da so metode za delo z nominalnimi spremenljivkami najbolj grobe in tudi najširše uporabljive, vendar nam npr. v primeru številskih spremenljivk ne dajejo tako dobrih rezultatov kot metode, ki upoštevajo številsko naravo le-teh. Po drugi strani imamo za delo s številskimi spremenljivkami na voljo mnogo računskih in analitičnih metod, ki v primeru imenskih spremenljivk nimajo pomena.

Za statistično obravnavo določenih pojavov je potrebno včasih obravnavati vso populacijo, pogosto pa to ni potrebno ali pa celo ni mogoče. V takšnih primerih opazujemo vrednosti le na delu populacije, ki ga imenujemo **vzorec**. Pogosto je za smiselnost uporabe statističnih metod potrebno, da je vzorec **naključen**, kar pomeni, da je iz populacije dobljen “na slepo”. Torej naj bi izbira potekala tako, da ima vsaka statistična enota enako možnost biti izbrana v vzorec. Tudi temu pogoju pogosto ni mogoče zadostiti (medicina, statistična kontrola kvalitete, ...).

Pri vzorcih pogosto pričakujemo, da bodo verno odražali porazdelitev vrednosti opazovane spremenljivke (na celotni populaciji), govorimo o njihovi (želeni) reprezentativnosti. Pri opazovanju celotne populacije je to samoumevno, pri vzorcih pa je to le cilj. Posebna skrb je potrebna, da naključni vzorec upošteva tudi morebitne “plasti” populacije, torej morebitne druge lastnosti statističnih enot, ki bi lahko vplivale na vrednost opazovane spremenljivke.

Obstaja posebna veda, **načrtovanje poskusov** (angl. *experimental design*, tudi design of experiments, DoE), ki se ukvarja z aktivnostmi pred samim začetkom statistične analize in obsega:

- a) načrtovanje (statističnih) eksperimentov glede na zastavljene cilje,
- b) proučuje (predvsem učinkovite) metode vzorčenja.

Vse večja informatiziranost je privedla tudi do pojava nove vede, *podatkovnega rudarjenja* (data mining), ki pa se, kot ime namiguje, posveča pridobivanju zelenih, koristnih, pomembnih, ... informacij iz velikih zbirk podatkov. Pri tem delu se seveda ta veda prepleta s statistično analizo.

Ob koncu statistične raziskave so izsledki pogosto zanimivi za javnost, ki ni vedno podkovana v razumevanju statističnih izrazov in metod. Rezultate je zato potrebno primerno predstaviti, posebna skrb naj bi veljala izogibanju možnosti različnih interpretacij. Prav na tem mestu obstaja možnost nehotenih spodrseljajev, pa tudi namernega zavajanja. Razširjenost programskih paketov za statistične raziskave, pa tudi enostavnejših orodij za delo s pregledicami, kot je npr. Excel, omogoča izvajanje mnogih statističnih metod tudi osebam, ki nimajo primerne statistične izobrazbe, kar lahko včasih pripelje do vsebinskih napak.

Na spletu je mogoče najti obilo informacij o statistiki, o programski opremi za raziskovalno, predvsem pa za uporabno ukvarjanje s statističnimi metodami, o statističnih metodah, vključite se lahko v tečaje za samostojno ali vodeno izobraževanje, ... Tudi dela za tiste, ki statistiko obvladajo, ne manjka.