Uvod u statistiku

Uvod. Priroda statistike. Prikupljanje podataka. Populacija i uzorci. Kratka historija statistike.

Statistika je umjetnost učenja iz podataka. Ona se bavi proučavanjem načina prikupljanja podataka, njihovih pratećih opisa i njihovom analizom, koja često vodi do izvođenja nekog zaključka.

Dio statistike koja se bavi opisom i sumiranjem podataka se naziva deskriptivna statistika.

Dio statistike koja je skoncentrisana na izvođenju zaključaka iz dobijenih podataka se naziva inferencijalna statistika.

Ukupna familija svih elemenata za čije informacije smo zainteresovani se naziva populacija. Podgrupu populacije koju demo detaljno proučavati se naziva uzorak

Za uzorak od k članova populacije kažemo da je slučajan uzorak, a ako su ti članovi izabrani na takav način da su sve vjerovatnoće izbora svih k članova jednake, tada uzorak nazivamo jednostavan slučajan uzorak.

(#) Da li je bolje upisati dijete u osnovnu školu mlađe ili starije dobi? Ovo je sigurno pitanje koje zanima mnoge roditelje kao i ljude koji su odgovorni za postavljanje društvenih pravila. Obrazložiti kako odgovoriti na ovo pitanje.

R.j.

- Posmatrati svoje iskustvo?
- Pričati sa prijateljima o njihovim iskustvima?
  - ni jedno ni drugo nije objektivno
  - treba nam puno veća grupa

- Testirati znanje djece na kraju prvog razreda?
  - starije dijete je najvjerovatnije "pokupilo" malo više znanja kod kuće od mlađeg djeteta u periodu kad nije išlo u školu i zbog toga bi trebalo pokazati bolje rezultate na testu

- Analizirati ukupan broj godina provedenih u školi?
  - mnogi autori tvrde da je ukupan broj godina provedenih u školi do njenog završetka daleko bolja mjera za odgovor na pitanje kada dijete treba upisati u školu.

- Analizirati njihov opšti uspjeh tokom školovanja?

U svim slučajevima moramo prikupiti odgovarajuće infomacije, ili podatke, i onda se ti podaci moraju opisati i analizirati.

(#) Kada je najbolje dijete upisati u školu - kada je mlađe ili kada je starije? Kako odgovoriti na ovo pitanje? Data je tabela

Tabela 1 - Ukupan broj godina provedenih u školi u odnosu na starost upisa

Godina	Mlađa polovina djece		Starija polovina djece	
	Prosječan broj godina u trenutku upisa u školu	Prosječan broj godina provedenih do završetka škole	Prosječan broj godina u trenutku upisa u školu	Prosječan broj godina provedenih do završetka
1956	6,38	13,84	6,62	13,67
1957	6,34	13,80	6,59	13,86
1958	6,31	13,78	6,56	13,79
1959	6,29	13,77	6,54	13,78
1960	6,24	13,68	6,53	13,68
1961	6,18	13,63	6,45	13,65
1962	6,08	13,49	6,37	13,53

(a) U kojoj godini se desila najveća razlika prosječnog broja godina provedenih u školi do njenog završetka između mlađih i starijih prvačića?

(b) Da li je broj godina provedenih u školi do završetka u prosjeku veći kod mlađe starosne grupe ili starije?

Rj.

1956: $|13,84 - 13,67| = 0,17$
1957: $|13,80 - 13,86| = 0,06$
1958: $|13,78 - 13,79| = 0,01$
1959: $|13,77 - 13,78| = 0,01$
1960: $|13,68 - 13,68| = 0$
1961: $|13,63 - 13,65| = 0,02$
1962: $|13,49 - 13,53| = 0,04$

(a) Najveća razlika se desila 1956.

(b) Starija starosna grupa je sljedeće godine provela više u školi u prosjeku od mlađe 1957, 1958, 1959, 1961, 1962

Postoji veći broj godina (u kojoj prosječan broj godina provedenih do završetka škole) kod starije starosne grupe.

(#) Ako želimo naučiti nešto o nečemu, prvo moramo sakupiti odgovarajuće podatke. Navesti nekoliko primjera problema za čije rješenje nam treba sakupljanje odgovarajućih podataka.

Rj.

1. Trenutno stanje ekonomije
2. Procent javnog mijenja o podršku određenog prijedloga
3. Prosječna potrošnja po kilometru novog automobila
4. Efikasnost novog lijeka
5. Korisnost novog načina učenja čitanja kod djece u osnovnoj školi

(#) Sljedeći podaci prikazuju procente odraslih osoba iz BiH, podjeljeni po edukacijskim nivoima, koji su konzumirali cigarete u godinama od 1999 do 2002.

Tabela 2 - Konzumiranje cigareta u BiH (% od svih odraslih osoba)

	1999	2000	2001	2002
Ukupno	25,8	24,9	24,9	26,0
Spol				
Muškarac	28,3	26,9	27,1	28,7
Žena	23,4	23,1	23,0	23,4
Edukacija				
Nezavršena srednja škola	38,9	32,4	33,8	35,2
Srednja škola	36,4	31,1	32,1	32,3
Viša srednja	32,5	27,7	26,7	29,0
Fakultet	18,2	13,9	13,8	14,5

(a) Za koju grupu postoji lagano opadanje?
(b) Možete li reći da li postoji cjelokupan trend?

Rj.

(a) Ako posmatramo godine od 1999 do 2002 ni za jednu grupu ne možemo reći da postoji lagano opadanje. Najbliže tom cilju bi bile žene.

(b) Možemo reći da postoji cjelokupan trend. U svim slučajevima pušača je 2002 bilo mnogo više nego 2001, a u većini slučajeva se vidi opadanje u periodu od 1999 do 2001.

(#) Statistika se često koristi za dizajniranje odgovarajućeg eksperimenta za prikupljanje podataka. Obrazložiti ovo na primjeru - Na koji način testirati efikasnost novog lijeka o smanjivanju holesterola?

Rj.

- Regrutovati volontere za testiranje.
- Periodično im davati lijek i mjeriti holesterol
- Da li davati lijek svim volonterima?
  - placebo efekat
  - vrijeme neobično toplo (ili hladno) što kao rezultat može imati da pacijenti provode više ili manje vremena napolju
  ...
- Podijeliti volontere u dvije grupe?
  - jednoj grupi davati lijek, drugoj placebo
  - medicinsko osoblje koje mjeri holesterol ne bi smjeli znati podjelu grupa
- Podijeliti volontere na slučajan način ili planirano?

Grupa koja neprima nikakav tretman se naziva kontrolna grupa.

(#) Bacanjem novčića 10 puta, glava je se pojavila 7 puta. Bacanjem nekog drugog novčića 50 puta, glava se pojavila 47 puta. Šta možemo reći o ova dva novčića i kako napraviti vezu između ovog eksperimenta i statističkih podataka.

R.j.

Da bi izveli logičan zaključak iz podataka, veoma je bitno da napravimo neke pretpostavke o šansi (ili vjerovatnoći) njihovog pojavljivanja. Ukupan zbir svih mogućih ovakvih pretpostavki u statistici je poznato pod imenom <u>modeli vjerovatnoće</u> za podatke.

Prvi novčić je najvjerovatnije običan novčić, kod kog a se nekom šansom od 10 bacanja glava pojavila 7 puta.

Drugi novčić nije običan novčić.

(#) U statistici, zanimaju nas određene informacije o familiji elemenata, koju nazivamo populacija. Populacija je obično prevише velika da bi se mogao ispitati svaki njegov član. Pretpostavimo da nas zanima prosječan broj godina ljudi u gradu u kojem živimo. Obrazložiti da li taj odgovor možemo dobiti anketirajući prvih 100 ljudi koji uniđu u gradsku biblioteku?

Rj.

Podgrupu populacije koju u detalje proučavamo nazivamo <u>uzorak</u>. Da bi uzorak davao odgovarajuće informacije o ukupnoj populaciji, mora na neki način reprezentovati tu populaciju.

Anketirajući prvih 100 ljudi koji uniđu u gradsku biblioteku ne možemo dobiti odgovor na postavljeno pitanje — sigurno možemo tvrditi da izabrani uzorak u ovom slučaju ne predstavlja čitavu populaciju (da li su svi ljudi učlanjeni u biblioteku? ko posjećuje biblioteku? studenti i penzioneri, učenici srednjih škola)

Reprezentacija u ovom smislu predstavlja uzorak koji je izabran na takav način da su svi dijelovi populacije na jednak način uključeni u uzorak.

(#) Jedna srednja škola broji 300 učenika u prvom razredu, 500 u drugom razredu i po 600 učenika u trećem i četvrtom razredu. Pretpostavimo da želimo ispitati mišljenje učenika da li odlazak u vojsku mora biti obaveza ili ne, i želimo napraviti uzorak od 100 učenika. Obrazložiti kakav bi izbor za uzorak bio idealan?

R.j.

1 razred = 300

2 razred = 500

3 razred = 600

4 razred = 600 Ukupno učenika: 2000.

U ovom slučaju naivan izbor za uzorak bi bio da nasumice izaberemo 100 osoba i sprovedemo anketu među njima.

Puno bolji izbor za uzorak bi bio da prvo izračunamo koliko osoba da izaberemo iz svakog razreda.

Kako su proporcije učenika

1 god. $\frac{300}{2000} = 0,15$ 2 god $\frac{500}{2000} = 0,25$

3 god $\frac{600}{2000} = 0,30$ 4 god $\frac{600}{2000} = 0,30$

ovo nam nameće da iz prvog razreda uzmemo 15, iz drugog 25, iz trećeg 30 i iz četvrtog 30 osoba.

Tek onda da izaberemo na slučajan način učenike.

(#) Medicinski istraživač, pokušavajući da ustanovi efikasnost novog lijeka, je počeo testirati lijek zajedno sa placebom. (u ovom smislu riječ placebo predstavlja lijek koji nema nikakva svojstva). Da bi bio siguran da su dvije grupe volontera pacijenata, oni koji će dobiti lijek i oni koji će dobiti placebo, što je moguće više slične, istraživač je odlučio da se ne pouzda u slučajnost nego da pažljivo pregleda volontere i onda sam formira grupe. Da li je ovaj pristup preporučljiv? Zašto ili zašto nije?

Lj.

Ovaj pristup nije preporučljiv. Istraživač koji pokušava da nauči o efikasnosti novog lijeka nebi trebao znati koji pacijenti primaju novi lijek a koji primaju placebo - u suprotnom, istraživač koji ima takvo znanje će tim znanjem biti "opterećen" i tokom istraživanja kao i u analizi dobijenih podataka će nesvjesno praviti dodatna zapažanja o efikasnosti novog lijeka.

(#) Sljedeće sedmice će se održati izbori, i uzimajući uzorak iz glasačke populacije mi želimo predvidjeti koja će stranka pobijediti. Koja od sljedećih metoda za izbor će proizvesti reprezentativan uzorak?

(a) Anketiranje svih ljudi glasačke dobi koji posjećuju ženske košarkaške utakmice.

(b) Anketiranje svih ljudi glasačke dobi koji napuštaju luksuzni gradski restoran.

(c) Dobiti kopiju registracijske liste glasača, na slučajan način izabrati 100 imena, i sprovesti anketu.

(d) Iskoristiti rezultat televizijskog telefon programa, u kojem voditelj traži od gledalaca da ga nazovu i saopšte mu svoj izbor.

(e) Iskoristiti imena iz telefonskog imenika i anketirati te ljude.

R.

(a) uzorak nije reprezentativan (ZAŠTO?)

(b) Uzorak nije reprezentativan (ZAŠTO?)

(c) Uzorak je reprezentativan i može se čak poboljšati (KAKO?)

(d) Uzorak nije reprezentativan.

(e) Uzorak je reprezentativan u zavisnosti koji telefonski imenik posmatramo – (a) imenik fiksnih telefona
(b) imenik mobilnih telefona

(#) Univerzitet planira da štampa izvještaj o svojim diplomiranim studentima u kome će sadržati informacije o njihovim godišnjim platama. Na slučajan način su izabrali 200 nedavnih diplomanata i poslali im upitnike koji su prilagođeni njihovim trenutnim poslovima. Međutim, od ujih 200, samo je 86 upitnika vraćeno. Pretpostavimo da je izvještaj o prosječnim platama 75000 KM.

(a) Hoće li Univerzitet pogriješiti u mišljenju ako zaključi da je 75000 KM prosječna plata diplomiranih studenata? Obrazložiti razlog koji se nalazi iza vašeg odgovora.

(b) Ako je vaš odgovor pod (a) pozitivan, možete li smanjiti neki skup uslova koji su povezani sa grupom koja je vratila popunjene upitnike iz kojih bi se složili da je 75000 KM dobra aproksimacija?

Rj:

(a) Da, Univerzitet će pogriješiti u mišljenju.
Diplomanti koji su vratili upitnike ne moraju reprezentovati ukupnu populaciju diplomiranih studenata.

(b) Ako je broj upitnika koji su vraćeni približno jednako 200 (200 upitnika je poslano) tada bi aproksimacija bila puno bolja.

(#) 1662 Engleski trgovac John Graunt je izdao knjigu pod naslovom "Prirodna i politička opažanja napravljena na osnovu smrtovnica". Sljedeća tabela prikazuje ukupan broj smrti u Engleskoj i broj smrti koju je izazvala kuga za pet različitih godina kuge, i ta tabela je uzeta iz ove knjige

Godina	Broj pokopa	Broj smrti od kuge
1592	25 886	11 503
1593	17 844	10 662
1603	37 294	30 561
1625	51 758	35 417
1636	23 359	10 400

Graunt je koristio Londonske smrtovnice za procjenu populacije grada. Npr. da bi ustanovio populaciju Londona 1660, Graunt je promatrao domaćinstva u određenim općinama Londona i ustanovio da, u prosjeku, postoje približno 3 smrti za svakih 88 ljudi. Djeljenjem ovog sa 3. dolazimo do zaključka da, u prosjeku, postoji 1 smrt za svaka $\frac{88}{3}$ čovjeka. Kako su Londonske smrtovnice pokazivale 13 200 smrti u Londonu te godine, Graunt je procjenio da je populacija Londona oko $13\,200 \cdot \frac{88}{3} = 387\,200$.

Kritikovati Graunt-ovu metodu za procjenu populacije Londona. Koju implicitnu pretpostavku je on napravio?

Rj. Graunt-ova implicitna pretpostavka je ta da općine koje je on promatrao čine dobru reprezentaciju ukupne populacije Londona.

Ključni pojmovi termini iz ove lekcije:

Statistika - umjetnost učenja iz podataka.

Deskriptivna statistika - dio statistike koja se bavi opisom i sumiranjem podataka

Inferencijalna statistika - dio statistike koja se bavi izvođenjem zaključaka iz dobijenih podataka

Modeli vjerovatnoće - matematičke pretpostavke koje se odnose na mogućnost pojavljivanja različite vrijednosti podataka

Populacija - familija elemenata koji nas zanimaju

Uzorak - podgrupa populacije koja će biti studirana

Slučajan uzorak veličine k - uzorak izabran na takav način da sve podgrupe veličine k imaju istu vjerovatnoću da budu izabrane

Slojevit slučajan uzorak - uzorak dobijen dijeljenjem populacije na različite podpopulacije i onda uzimanje slučajnog uzorka iz svake podpopulacije.

Zadaci za vježbu

1. Neki istraživač pokušava da ustanovi koliki broj godina u prosjeku danas ljudi dožive u BiH. Da bi dobio ovaj podatak, istraživač 30 dana čita čitulje (smrtovnice) iz Dnevnog avaza, i bilježi godinu smrti ljudi iz BiH. Obrazložiti da li će ovakav pristup dovesti do reprezentativnog uzorka?

2. Ako je u zadatku 1 (u prethodnom zadatku) dobijen odgovor da je prosječan broj godina umrle osobe danas u BiH 82,4 godine, kakav zaključak možemo izvesti iz ovoga.

3. Da bi odredili postotak ljudi u vašem gradu koji konzumiraju cigarete, odlučili ste anketirati ljude sa jedne od sljedećih lokacija
(a) bilijar sala;
(b) kuglane;
(c) tržnog centra;
(d) biblioteke.)
Koje od ovih potencijalnih mjesta za anketiranje će najvjerovatnije kao vezultat imati realnu aproksimaciju željenog postotka? Zašto?

4. U novinama je izašao članak iz ministarstva saobraćaja, o sadržaju oblačenja pješaka koji su poginuli u saobraćajnim nesrećama koje su se odvijale u noćnim

sabima, gdje stoji da je 80% žrtvi nosilo crno obojenu odjeću a da je 20% nosilo svijetlo obojenu odjeću. Na kraju članka je izveden zaključak da je noćna sigurnije nositi svijetlo-obojenu odjeću.

(a) Da li je ovaj zaključak opravdan? Objasniti.

(b) Ako je vaš odgovor pod (a) ne, koje dodatne informacije nam trebaju prije izvođenja konačnog zaključka.

5. Londonske smrtnice pokazuju da se desilo 12 246 smrtnih slučajeva 1658. Pretpostavljajući da je istraživanje Londonskih opština pokazalo približno 2 procenta populacije smrti te godine, iskoristiti Graunt-ovu metodu za procjenu londonske populacije 1658. (Grant-ova metoda je opisana u tekstu jednog od ranije urađenih zadataka).

6. Pretpostavimo da ste prodavači osiguranja u 1662 godini, kada je štampana Graunt-ova knjiga. Objasniti na koji način bi mogli iskoristiti njegove podatke o godini smrti ljudi.

7. Objasniti zašto mislite da bi studiranje statistike moglo pomoći u vašoj struci? Na koji način je mislite iskoristiti u svom daljem radu?