

1 Uvod

1.1 Što je statistika?

- U svakodnevom životu: prikupljanje, registriranje i prikaz podataka o nekom fenomenu: dnevne cijene na zagrebačkim tržnicama, nosivost teretnih brodova naručenih od hrvatskih brodogradilišta, broj postignutih golova u nogometnom kolu,...
- Statistika se bavi **analizom** dobivenih podataka te **metodama izvođenja zaključaka** o promatranom fenomenu na osnovi takve analize.
- Zaključci izvedeni statističkom analizom su nesigurni jer se zasnivaju na **nepotpunim** podacima (npr. procjena broja nezaposlenih u državi dobiva se ispitivanjem uzoraka od nekoliko tisuća ljudi) ili na podacima koji u sebi sadrže **slučajnu komponentu** (npr. ako promatramo rast borova posijanih iz iste grupe sjemenki, tada nakon godine dana ti borovi neće biti jednako visoki iako su rasli pod istim vremenskim uvjetima i na istom tlu).
- Statistika se prvenstveno bavi situacijama u kojima se pojavljivanje nekog događaja ne može predvidjeti sa sigurnošću.

Statistika je skup ideja i metoda koje se upotrebljavaju za prikupljanje i interpretaciju podataka u nekom području istraživanja, te za izvođenje zaključaka u situacijama gdje su prisutne nesigurnosti i varijacije.

1.2 Elementi znanstvenog istraživanja

- (a) **Specifikacija cilja.** Ciljevi istraživanja su razni. Npr. istraživanje koliko studenti prosječno potroše novaca za slobodne aktivnosti tokom tjedna, ili istraživanje kemijskih svojstava krutih otpadaka neke tvornice i njihov utjecaj na okolinu.

- (b) **Sakupljanje informacija (podataka).** Informacije se obično sakupljaju u obliku podataka koji numerički mjere neke karakteristike.
- (c) **Analiza podataka.** Pažljiva analiza podataka kritična je za potvrdu da je dobiveno novo znanje te za vrednovanje zaključaka.

1.3 Uloga statistike u znanstvenom istraživanju

1. **Dizajn eksperimenata** (*experimental design*) je grana statistike koja se bavi planiranjem eksperimenata i sakupljanjem podataka. Npr. u mnogim područjima znanosti eksperimenti su skupi te je unaprijed potrebno pažljivo odrediti tip i količinu potrebnih podataka.
2. **Deskriptivna statistika** (*descriptive statistic*) ili opisna statistika je grana statistike koja se bavi predočavanjem i opisivanjem glavnih karakteristika sakupljenih podataka (tablice, grafikoni, histogrami, srednje vrijednosti, . . .)
3. **Statističko zaključivanje** (*inferntial statistics*) je vrednovanje informacija sadržanih u podacima i ocjena novog znanja dobivenog iz tih podataka (procjena parametara promatrane populacije, testiranje statističkih hipoteza, . . .)

1.4 Primjeri statističkih problema

Primjer 1 *Bacamo novčić 100 puta i bilježimo rezultat. Na kraju imamo 60 pisama i 40 glava. Sumnjamo da je novčić neispravan. Želimo biti 95% sigurni u to. Kako možemo provjeriti hipotezu o neispravnosti novčića?*

Primjer 2 *U nekom gradu u SAD-u, u glasačke liste upisano je 25000 glasača. Ispitivanjem slučajnog uzorka od 1600 osoba želimo procijeniti postotak demokrata. Nakon ispitivanja pokazuje se da u uzorku ima 917 demokrata. Očito je da ćemo postotak demokrata procijeniti sa $\frac{917}{1600} \cdot 100\% \approx 57.3\%$. Da li to znači da u cijelom gradu ima 57.3% demokrata? Više? Manje? Kolika je moguća greška?*

Primjer 3 Kontrolom istovrsnih proizvoda izrađenih na dva različita stroja ustanovljeno je sljedeće:

	stroj A	stroj B
<i>dobri proizvodi</i>	240	380
<i>loši proizvodi</i>	20	24
ukupno	260	404

Postotak loših proizvoda dobivenih na stroju A je $\frac{20}{260} \cdot 100\% \approx 7.7\%$, a na stroju B $\frac{24}{404} \cdot 100\% \approx 5.9\%$. Da li to znači da se strojevi A i B **značajno razlikuju** u proizvodnji loših proizvoda?

Primjer 4 Kockar je optužen da upotrebljava namještenu kocku. Sačuvani su podaci 60 posljednjih bacanja:

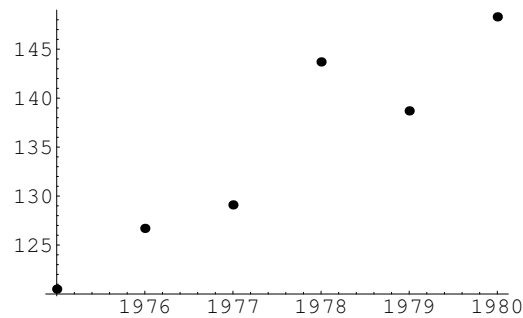
<i>vrijednost</i>	<i>dobivena frekvencija</i>	<i>očekivana frekvencija</i>
1	4	10
2	6	10
3	17	10
4	16	10
5	8	10
6	9	10

Očito je da dobivene frekvencije odstupaju od očekivanih. Da li je to zbog slučajnosti ili zbog namještene kocke? Da bi se kockara optužilo za varanje, potrebna je jaka **evidencija**. U tu svrhu statističar postavlja **hipotezu** da je kocka ispravna i želi izračunati **vjerojatnost odstupanja dobivenih od očekivanih frekvencija**.

Primjer 5 U tablici

<i>godina</i>	1975	1976	1977	1978	1979	1980
<i>broj razvoda (x100)</i>	120.5	126.7	129.1	143.7	138.7	148.3

nalaze se podaci o godišnjem broju razvoda u Engleskoj i Walesu u razdoblju od 1975. do 1980. godine. Podatke prikažemo u Kartezijevom koordinatnom sustavu:

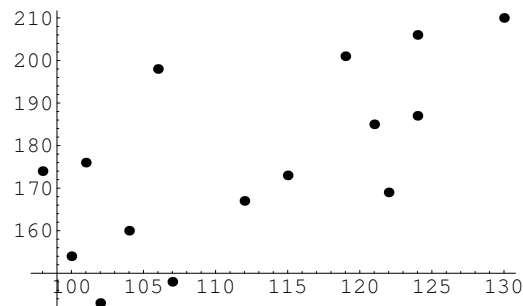
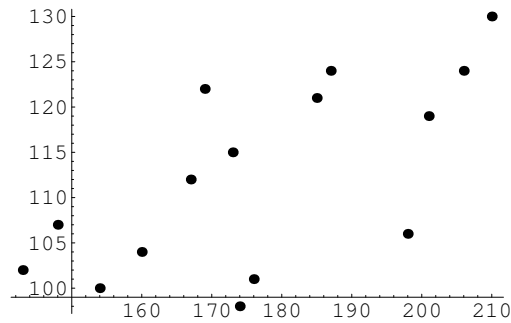


gdje su točke s koordinatama (x_i, y_i) , x_i je godina, a y_i broj razvoda ($i = 1, \dots, 6$). Izgleda da podaci pokazuju neku vrstu **linearnog rastućeg trenda**. To sugerira da se rastući trend može modelirati kao pravac s jednadžbom $y = \alpha + \beta x$, gdje su x godine, a y broj razvoda. Kako možemo procijeniti stopu rasta β ?

Primjer 6 U tablici

<i>pacijent</i>	1	2	3	4	5	6	7	8
<i>sistolički tlak</i>	210	169	187	160	167	176	185	206
<i>dijastolički tlak</i>	130	122	124	104	112	101	121	124
<i>pacijent</i>	9	10	11	12	13	14	15	
<i>sistolički tlak</i>	173	143	174	201	198	148	154	
<i>dijastolički tlak</i>	115	102	98	119	106	107	100	

prikazani su podaci dobiveni mjerenjem sistoličkog i dijastoličkog krvnog tlaka (u mm Hg) u 15 pacijenata koji imaju esencijalnu hipertenziju. Krvni tlak u tijelu varira ovisno o otkucaju srca. Sistolički tlak je maksimalan tlak do kojeg dolazi kontrakcijom srca, a dijastolički tlak je minimalan tlak. Prikažimo grafički dobivene podatke.



Prvi graf prikazuje vrijednosti dijastoličkog u odnosu na sistolički tlak, a drugi graf vrijednosti sistoličkog u odnosu na dijastolički tlak. Sa slika možemo zaključiti da postoji tendencija da pacijenti koji imaju visok sistolički tlak imaju također i visok dijastolički tlak i obratno. Položaj točaka sugerira rast slijeva na desno. Bilo bi moguće analizirati podatke kao u prethodnom primjeru. Međutim, javlja se problem koju od dviju varijabli uzeti za zavisnu (ordinata), a koju za nezavisnu (apscisa) varijablu. Podaci u tabeli sugeriraju da ljudi s umjerenom esencijalnom hipertenzijom koji imaju visok sistolički tlak imaju tendenciju prema visokom dijastoličkom tlaku i obratno. To intuitivno znači da su te dvije varijable povezane ili **korelirane** (u našem primjeru one su pozitivno korelirane). Postavlja se pitanje kako mjeriti i kako **testirati** koreliranost.

Primjer 7 Istraživana je veza između pojave povišenog krvnog tlaka i pušenja, tako da je ispitano 180 osoba i rezultati su prikazani u tablici:

	<i>nepušač</i>	<i>blagi pušač</i>	<i>teški pušač</i>	Σ
<i>normalni tlak</i>	48	26	19	93
<i>povišeni tlak</i>	21	36	30	87
Σ	69	62	49	180

Da li je pojava povišenog krvnog tlaka **nezavisna** od pušenja? S kolikom sigurnošću to možemo tvrditi ili opovrgnuti?

1.5 Populacija i uzorak

Iako su primjeri iz prethodnog poglavlja izvučeni iz različitih područja, neke zajedničke karakteristike su očite.

1. U svim područjima proučavanja potrebno je **sakupiti podatke** (eksperimentom, popisom, promatranjem).
2. Izvjesna količina **varijabilnosti** neizbježna je u svim primjerima usprkos sličnim uvjetima.
3. Značajna je činjenica u svim primjerima da je, ili fizički nemoguće, ili praktički neizvedivo sakupiti iscrpni i sveobuhvatni skup podataka u navedenim područjima istraživanja. Npr. kolikogod podataka sakupili eksperimentiranjem, u principu je moguće eksperiment i dalje ponavljati i dobivati nove podatke. U socioekonomskim mjerenjima ispituje se tek mali dio populacije.

Fundamentalna ideja koja proizlazi iz 3 pokazuje razliku između skupa podataka koji su zaista dobiveni procesom sakupljanja i ogromne familije potencijalnih podataka koji se mogu zamisliti u danom kontekstu.

Sakupljeni podaci nazivaju se **uzorkom**.

Familija svih podataka (potencijalnih promatranja) naziva se **populacijom**.

Definicija 1 (Statistička) populacija je potpun skup mogućih mjerenja ili podataka o nekom kvalitativnom svojstvu koji odgovaraju cijeloj familiji jedinki o kojoj treba dati zaključak.

Populacija predstavlja cilj istraživanja i svrha procesa sakupljenih podataka je izvođenje zaključaka o populaciji.

Uzorak is statističke populacije je skup mjerenja (podataka) koja su sprovedena u toku istraživanja.

Ciljevi statistike:

- Zaključivanje o populaciji iz podataka u uzorku.
- Ocjena nesigurnosti (neizvjesnosti) koje su uključene (obuhvaćene) tim zaključivanjem.