Poglavlje 4

Statistika

Korištenje je riječi **statistika** u svakodnevnom životu najčešće povezano s brojčanim vrijednostima kojima pokušavamo opisati bitne karakteristike nekog skupa podataka.

Statistika, kao znanstvena disciplina, bavi se razvojem metoda prikupljanja, opisivanja i analiziranja podataka te primjenom tih metoda u procesu donošenja zaključaka na temelju prikupljenih podataka.

Statističko istraživanje fokusirano je na skup **objekata**, tj. **jedinki** (ljudi, životinja, biljaka, stvari, država, gradova, poduzeća itd.) i skup odabranih veličina koje se na njima promatraju. Veličine koje se na jedinkama promatraju zovemo **varijablama**. Sve jedinke koje se žele obuhvatiti istraživanjem, tj. o kojima se želi zaključivati, čine **populaciju**. Podaci se prikupljaju da bi se zaključivalo o varijablama na populaciji.

Prije provođenja istraživanja populacija mora biti precizno opisana. Na primjer, ako je varijabla od interesa broj riječi koji je osoba u stanju pročitati u minuti, onda valja precizirati čine li populaciju djeca predškolskog uzrasta u Hrvatskoj ili djeca prvog razreda osnovne škole u Hrvatskoj ili djeca prvog razreda neke specifične škole u Hrvatskoj (npr. O.Š. Tin Ujević u Osijeku) ili djeca prvog razreda osnovne škole u Hrvatskoj koja idu po posebnom programu itd. U svakom slučaju, populacija mora biti precizirana.

Iz definirane populacije uzima se u statističko istraživanje **uzorak**. Podaci moraju biti prikupljeni na uzorku koji je reprezentativan za opisanu populaciju. Reprezentativan uzorak mora odražavati populaciju tj. u njemu trebaju biti zastupljene

sve tipične karakteristike populacije. Najčešći je način odabira jedinki iz populacije u reprezentativan uzorak tzv. **slučajni uzorak**, tj. takav izbor u kojemu svaka jedinka ima jednaku šansu biti izabrana u uzorak.

Izbor jedinki u reprezentativan uzorak iz populacije može se postići pomoću generatora slučajnih brojeva. Da bi ilustrirali tu proceduru, označimo s M broj jedinki u populaciji i numerirajmo jedinke brojevima od 1 do M. Neka je n broj jedinki koje želimo uzeti u uzorak. Potrebno je generirati n međusobno različitih realizacija diskretne uniformne slučajne varijable na skupu $\{1, \ldots, M\}$ i uzeti u uzorak jedinke koje za oznaku imaju dobivene prirodne brojeve. Treba naglasiti da tu proceduru nije uvijek jednostavno provesti u primjeni.

Zadatak 4.1. U studentskoj referadi pribavite popis studenata koji slušaju UVIS ove godine. Numerirajte studente i provedite proceduru izbora 40 studenata u slučajni uzorak. Od svakog studenta odabranog u uzorak prikupite podatke o spolu, broju do sada položenih ispita, broju do sada ostvarenih ECTS bodova, prosječnoj ocjeni svih do sada položenih ispita, ocjeni iz predmeta Diferencijalni račun i Integralni račun. Formirajte bazu podataka u nekom programskom paketu.

U okviru statističke teorije valja se pozabaviti sljedećim metodama:

- metodama prikupljanja podataka,
- metodama opisivanja skupa podataka (deskriptivna statistika),
- metodama statističkog zaključivanja.

Metode prikupljanja podataka ovdje nećemo posebno proučavati. Ipak, valja napomenuti da se podaci mogu prikupiti npr. na osnovu javnih izvora (knjige, časopisi, novine, web), dizajniranog eksperimenta, anketa, promatranjem itd.

4.1 Deskriptivna statistika

U statističkim istraživanjima razlikujemo nekoliko osnovnih tipova varijabli koje se međusobno razlikuju po svojstvima vrijednosti koje mogu poprimiti.

Kvalitativne varijable

Karakteristika je kvalitativnih varijabli da njihove vrijednosti nisu, po svojim svojstvima korištenim u istraživanju, realni brojevi. Tipičan je primjer takve varijable spol osobe. Vrijednosti kvalitativne varijable uobičajeno nazivamo kategorijama. Kategorije kvalitativnih varijabli mogu biti definirane u skladu s potrebama statističkog istraživanja.

Primjer 4.1. *Sljedeće su varijable kvalitativnog tipa:*

- *radna mjesta u školi (spremačica, domar, tajnik, nastavnik, pedagog, ravnatelj),*
- *boja očiju (plava, smeđa, zelena),*
- *krvne grupe (A, B, AB, 0),*
- *spol (m ili ž).*

Numeričke varijable

Numeričke varijable prirodno primaju vrijednosti iz skupa **realnih brojeva**. Tipični su primjeri numeričkih varijabli masa i visina osobe. Međutim, treba naglasiti da se i kategorije kvalitativnih varijabli mogu izražavati brojevima što ih ne čini numeričkim varijablama. Primjerice, spol osobe jedna je kvalitativna varijabla. Kategoriju "ženski spol" možemo označiti s "1", a kategoriju "muški spol" s "2", što može biti korisno prilikom unošenja podataka u bazu. Time smo kategorijama kvalitativne varijable pridružili numeričke vrijednosti, ali samu varijablu nismo učinili numeričkom po njezinim svojstvima.

Primjer 4.2. *Sljedeće su varijable numeričkog tipa:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine,*
- *broj bodova na državnoj maturi iz matematike,*
- *broj ulovljenih komaraca u klopku,*
- *temperatura mora,*
- *koncentracija soli u morskoj vodi.*

Među numeričkim varijablama razlikujemo diskretne i kontinuirane varijable.

Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti.

Primjer 4.3. *Sljedeće su numeričke varijable diskretne:*

- *broj bodova na državnoj maturi iz matematike,*
- *broj ulovljenih komaraca u klopku,*
- *broj dana u godini s temperaturom zraka većom od 35^oC.*

Skup je mogućih vrijednosti kontinuiranih numeričkih varijabli cijeli skup realnih brojeva ili neki interval.

Primjer 4.4. *Sljedeće su numeričke varijable kontinuirane:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine,*
- *temperatura mora,*
- *vodostaj neke rijeke.*

Primjer 4.5. (djelatnici.xls)
Baza podataka djelatnici.xls *sadrži podatke o uzorcima djelatnika dviju konkurentskih tvornica - tvornice A i tvornice B. U tablici s imenom "tvornica A" zabilježene su vrijednosti sljedećih varijabli za djelatnike* **tvornice** *A:*

spol - *kvalitativna varijabla koja sadrži informaciju o spolu (*M - *muški spol,* Z - *ženski spol),*

odjel - *kvalitativna varijabla sadrži naziv odjela u kojemu je djelatnik zaposlen (*TR - *transport,* P- *pakiranje,* IS - *isporuka),*

obrazovanje - *kvalitativna varijabla koja sadrži stručnu spremu djelatnika (*SSS - *srednja stručna sprema,* VŠSS - *viša stručna sprema,* VSS - *visoka stručna sprema),*

dob - *kontinuirana numerička varijabla koja sadrži starost djelatnika u godinama,*

visina - *kontinuirana numerička varijabla koja sadrži visinu djelatnika u centimetrima,*

rukovodstvo - *diskretna numerička varijabla koja sadrži broj godina rada koje je djelatnik proveo na nekoj od rukovodećih pozicija u toj tvornici,*

placa_prije - *kontinuirana numerička varijabla koja sadrži iznos godišnje plaće djelatnika prije reorganizacije poslovnog sustava,*

placa_poslije - *kontinuirana numerička varijabla koja sadrži iznos godišnje plaće djelatnika nakon reorganizacije poslovnog sustava.*

U tablici s imenom "tvornica B", u varijabli placa_konkurencija*, zabilježeni su iznosi godišnje plaće za svakog djelatnika iz uzorka iz* **tvornice** *B.*

U svrhu prikaza podataka i nekih statističkih analiza, vrijednosti se numeričke varijable također mogu svrstati u **kategorije**. Za razliku od kategorija kvalitativnih varijabli, među kategorijama se numeričke varijable uvijek može prepoznati prirodan poredak.

Primjer 4.6. (auto-centar.xls)
Svrha je ovog primjera prikazati mogućnost kategorizacije numeričke varijable. Taj se postupak najčešće provodi stvaranjem nove varijable čije su vrijednosti kategorije kojih je (znatno) manje nego svih mogućih vrijednosti odgovarajuće numeričke varijable. Baza podataka auto-centar.xls *sastoji se od sljedećih varijabli:*

automobili - *diskretna numerička varijabla koja sadrži podatke o broju prodanih automobila u jednom danu za sto promatranih dana. Budući da broj prodanih automobila u jednom danu može biti vrlo mali (npr. samo nekoliko osobnih automobila), ali i vrlo velik (npr. narudžbe automobila za vozni park nekog poduzeća), zaključujemo da diskretna numerička varijabla* automobili *može poprimiti velik broj različitih vrijednosti iz skupa prirodnih brojeva. Zato je u nekim situacijama korisno kategorizirati vrijednosti te varijable prema točno određenom kriteriju. Na primjer, kategorizacija broja prodanih automobila u jednome danu može se napraviti kao što je prikazano varijablom* kategorija.

kategorija - *kvalitativna varijabla koja podatke iz varijable* automobili *svrstava u pet kategorija prema kriteriju prikazanom u tablici 4.1.*

broj prodanih automobila	kategorija
0 - 9	E
10 i 11	D
12 i 13	C
14 i 15	B
16 i više	A

Tablica 4.1: Primjer kategorizacije numeričke varijable automobili

Ordinalne varijable

Karakteristika je ordinalnih varijabli da su one, po svom karakteru, kvalitativne, ali među kategorijama se može uspostaviti prirodan poredak. Tipičan su primjeri takvih varijabli stručna sprema osobe i ocjena u školi.

Primjer 4.7. (matematika.sta)
Baza podataka matematika.sta *sadrži podatke prikupljene anketiranjem studenata nakon održanih predavanja, vježbi, kolokvija te usmenog ispita iz jednog matematičkog kolegija. Prikupljeni podaci organizirani su na sljedeći način:*

prosjek - *varijabla koja sadrži podatke o prosječnoj ocjeni studiranja za* 49 *anketiranih studenata,*

polozeno - *varijabla koja studente svrstava u dvije kategorije s obzirom na to jesu li položili ispit iz promatranog kolegija prema kriteriju prikazanom u tablici 4.2.*

položen/nepoložen ispit	kategorija
položen ispit	1
nepoložen ispit	0

Tablica 4.2: Kategorizacija studenata prema položenosti ispita.

predavanja, vjezbe - *dvije varijable koje prisutnost studenata na predavanjima/vježbama (p/v) svrstavaju u tri kategorije na način prikazan u tablici 4.3.*

prisutnost studenta na p/v	kategorija
student s p/v nije nikada izostao	1
student je s p/v izostao samo jednom	2
student je s p/v izostao barem dva puta	3

Tablica 4.3: Kategorizacija studenata prema broju izostanaka s predavanja/vježbi.

tezina kolegija, materijali - *dvije varijable koje sadrže subjektivne ocjene (u standardnoj skali od 1 do 5) studenata o težini kolegija i dostatnosti dostupnih materijala za pripremanje ispita iz promatranog kolegija.*

Uočimo da se varijabla prosjek *može promatrati kao neprekidna numerička varijabla, varijabla* polozeno *jest kvalitativna, dok se varijable* predavanja, vjezbe, tezina kolegija i materijali *mogu svrstati u ordinalne varijable.*

4.1.1 Metode opisivanja kvalitativnih varijabli

Vrijednosti kvalitativne varijable jesu kategorije. Mjere kojima opisujemo zastupljenost jedne kategorije u uzorku jesu **frekvencija** kategorije i **relativna frekvencija** kategorije.

Frekvencija kategorije broj je izmjerenih vrijednosti varijable koje pripadaju danoj kategoriji.

Relativna frekvencija kategorije broj je izmjerenih vrijednosti varijable koje pripadaju danoj kategoriji podijeljen ukupnim brojem izmjerenih vrijednosti za ispitivanu varijablu.

Pretpostavimo da varijabla može primiti vrijednosti k različitih kategorija, a da se u podacima nalazi n izmjerenih vrijednosti za tu varijablu. Frekvenciju i-te kategorije označit ćemo f_i, a relatvnu frekvenciju dobijemo kao

$$\frac{f_i}{n}.$$

Relativna frekvencija kategorije mjera je zastupljenosti koja daje informaciju o udjelu kategorije u uzorku poznate veličine i često se izražava kao postotak. Frekvencije i relativne frekvencije pojedinih kategorija prikazujemo tablično i grafički.

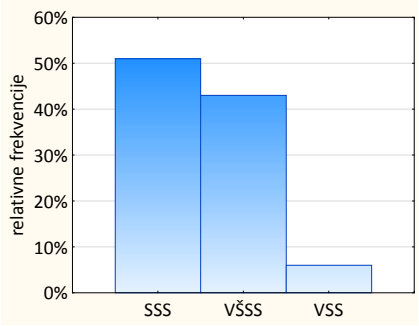
Frekvencije i relativne frekvencije kategorija kvalitativnih varijabli grafički prikazujemo pomoću **stupčastog dijagrama frekvencija** i **stupčastog dijagrama relativnih frekvencija**. U istu svrhu može se koristiti i **kružni dijagram** frekvencija i relativnih frekvencija . Popularan je naziv za isti grafički prikaz "pita".

Primjer 4.8. (djelatnici.xls)
Baza podataka djelatnici.xls *opisana je u primjeru 4.5. Promatajmo kvalitativnu varijablu* obrazovanje *čije su vrijednosti svrstane u tri kategorije:* SSS *- srednja stručna sprema,* VŠSS *- viša stručna sprema,* VSS *- visoka stručna sprema. Zastupljenost tih kategorija u promatranom uzorku od* 100 *djelatnika opisana je tablicom frekvencija i relativnih frekvencija 4.4.*

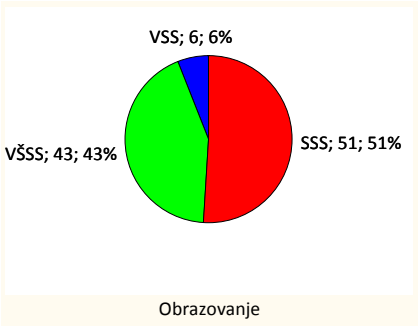
kategorija	frekvencija	relativna frekvencija
SSS	51	$51/100 = 0.51$
VŠSS	43	$43/100 = 0.43$
VSS	6	$6/100 = 0.06$

Tablica 4.4: Tablica frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje.

Grafički prikazi relativnih frekvencija dani su u obliku stupčastog dijagrama na slici 4.1, a prikazi frekvencija i relativnih frekvencija u obliku kružnog dijagrama na slici 4.2.



Slika 4.1: Stupčasti dijagram relativnih frekvencija svih kategorija varijable obrazovanje.



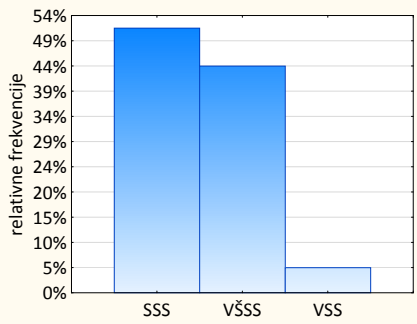
Slika 4.2: Kružni dijagram frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje.

Primjer 4.9. (djelatnici.xls)
Često se u praksi pokazuje korisnim poznavanje zastupljenosti kategorija jedne varijable za svaku od kategorija neke druge kvalitativne varijable proučavane na istom uzorku. U ovom ćemo primjeru tablično i grafički prikazati frekvencije i relativne frekvencije svih kategorija varijable obrazovanje *posebno za ispitanike ženskog spola, a posebno za ispitanike muškog spola iz promatranog uzorka djelatnika.*

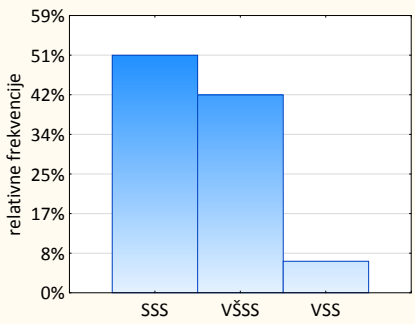
Tablice tako kategoriziranih frekvencija i relativnih frekvencija varijable obrazovanje *prikazane su u tablici 4.5. Stupčasti dijagrami relativnih frekvencija svih kategorija varijable* obrazovanje *za kategorije* Z *i* M *varijable* spol *prikazani su na slici 4.3, a kružni dijagrami frekvencija i relativnih frekvencija na slici 4.4.*

kategorija	spol=Z		spol=M	
	frekvencija	relativna frekvencija	frekvencija	relativna frekvencija
SSS	21	$21/41 = 0.5122$	30	$30/59 = 0.5085$
VŠSS	18	$18/41 = 0.4390$	25	$25/59 = 0.4237$
VSS	2	$2/41 = 0.0488$	4	$4/59 = 0.0678$

Tablica 4.5: Tablica frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.

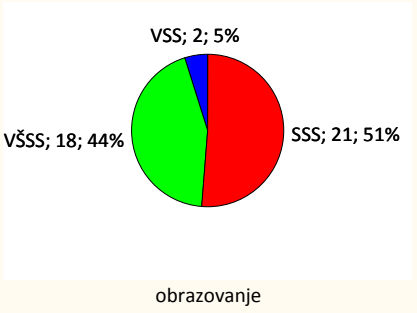


(a) spol=Z

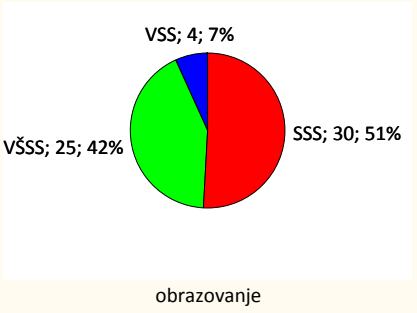


(b) spol=M

Slika 4.3: Stupčasti dijagrami relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.



(a) spol=Z



(b) spol=M

Slika 4.4: Kružni dijagram frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.

4.1.2 Metode opisivanja numeričkih varijabli

Numeričke varijable po svojoj prirodi mogu biti diskretne i neprekidne. U oba slučaja, a posebno kod neprekidnih varijabli, može se dogoditi da u prikupljenim podacima postoji mnogo međusobno različitih vrijednosti. U takvim slučajevima tablični i grafički prikazi uvedeni za kvalitativne varijable mogu biti nedovoljno informativni.

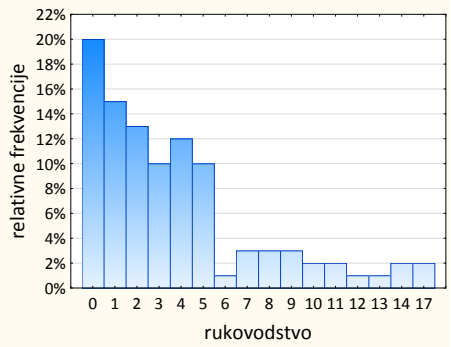
Ako su numeričke varijable diskretne s malo mogućih vrijednosti, za opis podataka možemo koristiti iste metode kao pri opisivanju kvalitativnih podataka, tj. frekvencije i relativne frekvencije, te ih grafički prikazivati stupčastim dijagramima i kružnim dijagramima.

Primjer 4.10. (djelatnici.xls)
Broj zabilježenih vrijednosti diskretnih numeričkih varijabli znatno utječe na preglednost tabličnih i grafičkih prikaza frekvencija i relativnih frekvencija. Analizirajmo diskretnu numeričku varijablu rukovodstvo *baze podataka* djelatnici.xls *iz primjera 4.5 koja prima vrijednosti iz skupa* $\{0, 1, \ldots, N\}$*, gdje je* N *najveći mogući broj godina radnog staža koje djelatnik može provesti na rukovodećoj poziciji. Frekvencije i relativne frekvencije zabilježenih vrijednosti varijable* rukovodstvo *za djelatnike iz promatranog uzorka dane su u tablici 4.6.*

rukovodstvo	frek.	rel. frek.	rukovodstvo	frek.	rel. frek.
0	20	0.20	8	3	0.03
1	15	0.15	9	3	0.03
2	13	0.13	10	2	0.02
3	10	0.10	11	2	0.02
4	12	0.12	12	1	0.01
5	10	0.10	13	1	0.01
6	1	0.01	14	2	0.02
7	3	0.03	17	2	0.02

Tablica 4.6: Tablica frekvencija i relativnih frekvencija svih zabilježenih vrijednosti varijable rukovodstvo.

Vidimo da se varijabla rukovodstvo *na promatranom uzorku djelanika realizirala sa* 16 *različitih vrijednosti čije su relativne frekvencije grafički prikazane stupčastim dijagramom 4.5.*



Slika 4.5: Stupčasti dijagram relativnih frekvencija svih zabilježenih vrijednosti varijable rukovodstvo.

Ako numerička varijabla prima mnogo međusobno različitih vrijednosti, za prikazivanje skupa izmjerenih vrijednosti obično nam neće puno pomoći frekvencije te stupčasti ili kružni dijagrami napravljeni na osnovu svake pojedine izmjerene vrijednosti. Takvi se slučajevi često javljaju ako podaci dolaze iz kontinuiranih numeričkih varijabli. Problem ćemo ilustrirati sljedećim primjerom.

Primjer 4.11. (djelatnici.xls)
Promotrimo varijablu placa_prije *iz baze podataka* djelatnici.xls *koja sadrži godišnje plaće prije reorganizacije poslovanja poduzeća za uzorak od* 100 *djelatnika promatranog poduzeća. Ovdje možemo pretpostaviti da se radi o kontinuiranoj numeričkoj varijabli koja može primiti vrijednosti iz intervala* $\langle 0, x]$. *Kao gornju granicu intervala, tj.* x, *možemo uzeti realan broja za koji pretpostavljamo da u danim uvjetima predstavlja najveću moguću godišnju plaću u tom poduzeću.*

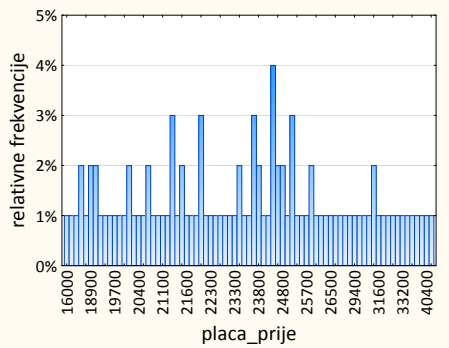
Dio tablice frekvencija i relativnih frekvencija na temelju svih realiziranih vrijednosti prikazan je tablicom 4.7. Od 100 *realizacija zabilježeno je čak* 77 *različitih vrijednosti, a pojedine se vrijednosti pojavljuju s izrazito malim frekvencijama:* 1, 2, 3 *ili* 4. *Iz takvih je tablica teško očitavati frekvencije koje nas zapravo zanimaju (npr. frekvencija ispitanika s godišnjom plaćom manjom od* 20000*). Zbog toga se u tablice uobičajeno dodaje i stupac kumulativnih frekvencija i kumulativnih relativnih frekvencija, kako je prikazano tablicom 4.7.*

Stupčasti dijagram (slika 4.6) ili kružni dijagram frekvencija (relativnih frekvencija) na kojemu su prikazane sve realizirane vrijednosti s odgovarajućom frekvencijom (relativnom frekvencijom), u takvim slučajevima nije osobito informativan.

Ako imamo podatke $(x_1, \ldots, x_n)$, *onda je* **kumulativna frekvencija** *podatka* x_i, $i = 1, \ldots, n$, *broj svih podataka iz* $(x_1, \ldots, x_n)$ *koji su manji ili jednaki* x_i. *Analogno se definira i kumulativna relativna frekvencija podatka.*

iznos plaće	frekvencija	kumulativna frek.	relativna frekvencija	kumulativna rel. frek.
16000	1	1	0.01	0.01
⋮	⋮	⋮	⋮	⋮
19800	1	15	0.01	0.15
20000	1	16	0.01	0.16
20200	2	18	0.02	0.18
⋮	⋮	⋮	⋮	⋮
42400	1	100	0.01	1

Tablica 4.7: Tablica frekvencija i relativnih frekvencija svih sto zabilježenih vrijednosti varijable placa_prije.



Slika 4.6: Stupčasti dijagram relativnih frekvencija svih realiziranih vrijednosti varijable placa_prije.

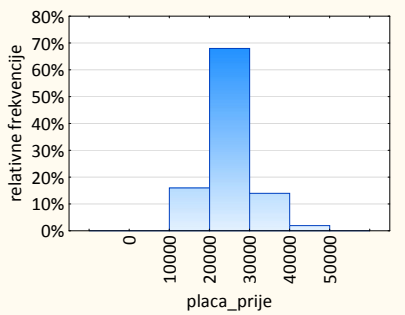
U svrhu dobivanja preglednih i korisnih stupčastih dijagrama za podatke iz kontinuiranih numeričkih varijabli, izmjerene je vrijednosti potrebno kategorizirati. To znači da velik skup podataka podijelimo u nekoliko disjunktnih intervala po kriteriju za koji smatramo da će nam dati željene rezultate. Stupčasti dijagram tada smještamo u koordinatni sustav tako da prikazujemo stupiće nad tim intervalima s površinom koja odgovara relativnoj frekvenciji podataka sadržanih u odgovarajućem intervalu. Duljina intervala informacija je koja je također prikazana stupčastim dijagramom jer odgovara širini stupića. Takav stupčasti dijagram zovemo **histogram**.

Primjer 4.12. (djelatnici.xls)
Promotrimo ponovno varijablu placa_prije *iz baze podataka* djelatnici.xls. *Razvrstajmo vrijednosti u disjunktne intervale duljine* 10000 *počevši od nule. Tako dobiveni tablični prikaz frekvencija i relativnih frekvencija dan je tablicom 4.8, a pripadni histogram slikom 4.7. Takakv histogram*

jasno ilustrira činjenicu da najviše djelatnika u uzorku ima godišnju plaću od 20000 *do* 30000 *novčanih jedinica, dok je plaća iz intervala* 40000 *do* 50000 *rijetkost. Intervale za kategorizaciju u takvim i sličnim slučajevima obično radimo tako da bi zadovoljili potrebe za prezentiranjem informacija koje želimo istaknuti.*

iznos plaće	frekvencija	relativna frekvencija
$[0, 10000\rangle$	0	0
$[10000, 20000\rangle$	15	0.15
$[20000, 30000\rangle$	69	0.69
$[30000, 40000\rangle$	14	0.14
$[40000, 50000\rangle$	2	0.02

Tablica 4.8: Tablica frekvencija i relativnih frekvencija kategoriziranih izmjerenih vrijednosti varijable placa_prije.



Slika 4.7: Histogram relativnih frekvencija kategoriziranih izmjerenih vrijednosti varijable placa_prije.

Za numeričke varijable možemo definirati numeričke karakteristike koje imaju logičnu interpretaciju i mogu se iskoristiti s ciljem prikazivanja skupa podataka. Ovdje ćemo definirati neke od najčešće korištenih numeričkih karakteristika skupa podataka.

Aritmetička sredina podataka

Aritmetička sredina niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom

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

Aritmetička sredina numerička je karakteristika koja pripada mjerama centralne tendencije, tj. mjeri "srednju vrijednost" podataka.

Primjer 4.13. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8.$$

S obzirom da ih ima ukupno 9*, aritmetička sredina (eng. mean) tog skupa izmjerenih vrijednosti jest*

$$\frac{1.2 + 2.1 + 3.2 + 4.3 + 5.4 + 6.5 + 7.6 + 8.7 + 9.8}{9} \approx 5.42.$$

Medijan podataka

Da bismo razumjeli i odredili medijan potrebno je prvo poredati izmjerene vrijednosti $x_1, x_2, \ldots, x_n$ varijable X po veličini (u rastućem poretku, tj. od manjeg prema većem). Medijan je također jedna mjera centralne tendencije numeričkih podataka, a ima značenje izmjerene vrijednosti koja se nalazi na sredini niza podataka kada je on uređen po veličini, tj. barem pola podataka manje je ili jednako medijanu, a istovremeno je barem pola podataka veće ili jednako medijanu. Način njegovog izračuna ovisi o tome imamo li **neparan** ili **paran** broj izmjerenih vrijednosti varijable. Ukoliko imamo **neparan broj** izmjerenih vrijednosti, onda postoji vrijednost koja je na srednjoj poziciji u uređenom skupu, pa nju definiramo kao medijan.

Primjer 4.14. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3.$$

Prvo te vrijednosti poredamo po veličini:

$$1, 1, 2, 2, 2, \mathbf{2}, 3, 5, 5, 6, 7.$$

S obzirom da ih ima ukupno 11*, medijan je vrijednost koja je na šestoj poziciji u tako dobivenom nizu, tj. broj* 2*.*

Ukoliko imamo **paran broj** izmjerenih vrijednosti varijable, onda ne postoji podatak koji je na srednjoj poziciji jer srednju poziciju "zauzimaju" dva podatka. Medijan se tada definira kao polovište tih dvaju podataka (tj. aritmetička sredina tih dvaju podataka).

Primjer 4.15. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Prvo te vrijednosti poredamo po veličini:

$$1, 1, 2, 2, 2, \mathbf{2}, \mathbf{3}, 3, 5, 5, 6, 7.$$

S obzirom da ima 12 *podataka, "sredinu" čine šesti i sedmi podatak, tj. vrijednosti* 2 *i* 3*. Medijan je tog skupa podataka sredina tih dvaju brojeva, tj. medijan je* $(2 + 3)/2 = 2.5$*.*

Postotna vrijednost podataka, donji i gornji kvartil

Medijan odgovara pedeset postotnoj vrijednosti s obzirom da je barem 50% podataka manje ili jednako medijanu i barem 50% podataka veće ili jednako od medijana. Postotna vrijednost za neki izabrani broj $p \in \langle 0, 100\rangle$, označimo je x'_p, definira se poštujući zahtjev da je barem $p\%$ izmjerenih vrijednosti manje ili jednako x'_p, dok je barem $(100-p)\%$ vrijednosti veće ili jednako x'_p. Dvadeset pet postotna vrijednost zove se donji kvartil, a sedamdeset pet postotna vrijednost zove se gornji kvartil. Analogno, kao i kod računanja medijana, ako se na traženoj poziciji za računanje postotne vrijednosti nalaze dva podatka u uređenom skupu izmjerenih vrijednosti, postotnu vrijednost određujemo kao njihovu sredinu. Donji i gornji kvartil mjere su koje pripadaju grupi mjera raspršenosti podataka.

Primjer 4.16. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 6, 1, 3, 7, 3, 3, 3, 3.$$

Prvo te vrijednosti poredamo po veličini:

$$1, 1, 2, 3, 3, 3, 3, 3, 5, 6, 6, 7.$$

Želimo li odrediti donji kvartil, potrebno je prvo odrediti četvrtinu podataka (25%). S obzirom da imamo 12 *podataka, četvrtinu (25%) čine tri podatka. Treći je podatak u gornjem skupu broj* 2*, a četvrti* 3*. Donji kvartil jest* 2.5*. Deveti je broj u gornjem skupu podataka broj* 5*, a deseti* 6*, stoga je gornji kvartil* 5.5*.*

Najmanja i najveća vrijednost, raspon podataka

Raspon podataka mjera je koja pokazuje koliko su numerički podaci raspršeni, tj. to je jedna od mjera raspršenosti podataka. Definiran je kao razlika najveće i najmanje vrijednosti u skupu mjerenih vrijednosti varijable (tj. razlika maksimalne i minimalne izmjerene vrijednosti varijable). Ako su $x_1, x_2, \ldots, x_n$ izmjerene vrijednosti varijable, označimo najmanju od njih (minimum) $x_{\min}$, a najveću (maksimum) $x_{\max}$.

Primjer 4.17. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da je vrijednost 1 *najmanja izmjerena vrijednost, a* 7 *najveća. Prema tome, raspon ovog skupa izmjerenih vrijednosti je* $7 - 1 = 6$*.*

U mnogim je primjerima zanimljivo promatrati **maksimalno odstupanje izmjerenih vrijednosti varijable od "prosjeka", tj. aritmetičke sredine** izmjerenih vrijednosti. Ta je numerička karakteristika definirana kao veći od brojeva $(\overline{x}_n - x_{\min})$ i $(x_{\max} - \overline{x}_n)$, tj. broj

$$\max\{(\overline{x}_n - x_{\min}), (x_{\max} - \overline{x}_n)\}.$$

Primjer 4.18. *Neka su* $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ *izmjerene vrijednosti neke varijable. Tada je*

$$x_{\min} = 1, \quad x_{\max} = 7, \quad \overline{x}_n = \frac{1+2+5+6+5+1+2+7+2+2+3+3}{12} = 3.25.$$

Maksimalno odstupanje izmjerenih vijednosti te varijable od njihovog prosjeka jest

$$\max\{3.25 - 1, 7 - 3.25\} = \max\{2.25, 3.75\} = 3.75.$$

Varijanca i standardna devijacija podataka

Varijanca i standardna devijacija također pripadaju grupi mjera raspršenosti podataka. One karakteriziraju raspršenost podataka oko aritmetičke sredine. Varijanca niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable definirana je izrazom:

$$\overline{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2,$$

a standardna devijacija jest kvadratni korijen varijance, tj.

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

Primjer 4.19. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8.$$

Iz primjera 4.13 znamo da je aritmetička sredina tog skupa podataka približno jednaka 5.42. *Varijanca tog skupa podataka jest*

$$\overline{s}_n^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 \approx 7.87,$$

a standardna devijacija

$$\overline{s}_n = \sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2} \approx 2.81.$$

Mod podataka

Mod je vrijednost iz niza izmjerenih vrijednosti varijable kojoj pripada najveća frekvencija, tj. izmjerena je najviše puta. Mod ne mora biti jedinstven.

Primjer 4.20. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da je vrijednost 2 *izmjerena najviše puta (četiri puta), pa je* 2 *mod tog skupa podataka.*

Primjer 4.21. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 3, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da su najviše puta izmjerene dvije vrijednosi - 2 *i* 3 *izmjerene su točno četiri puta. Dakle, mod tog skupa podataka nije jedinstven.*

Korištenjem numeričkih karakteristika numeričkih varijabli skup mjerenih vrijednosti može se prikazati grafički pomoću **kutijastog dijagrama** (eng. *box plot*, *boxplot* ili *box-and-whisker plot*).

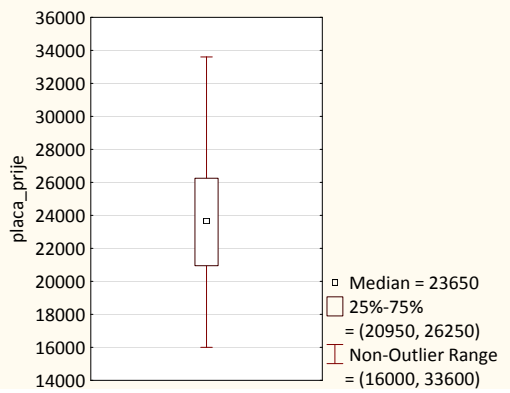
Kutijastim dijagramom prikazujemo odnos pet numeričkih karakteristika skupa izmjerenih vrijednosti: minimalnu vrijednost, donji kvartil, medijan, gornji kvartil i maksimalnu vrijednost.

Primjer 4.22. (djelatnici.xls)
Numeričke karakteristike skupa izmjerenih vrijednosti varijable placa_prije *iz baze podataka* djelatnici.xls *prikazane su u tablici 4.9.*

veličina uzorka	aritmetička sredina	mod	frek. moda	st. dev.	varijanca
100	24522	24600	4	5105.801	26069208.1
minimum	**donji kvartil**	**medijan**	**gornji kvartil**	**maksimum**	**raspon**
16000	20950	23650	26250	42400	26400

Tablica 4.9: Deskriptivna statistika varijable placa_prije.

Odnos minimuma, donjeg kvartila, medijana, gornjeg kvartila i maksimuma izmjerenih vrijednosti varijable placa_prije *prikazan je kutijastim dijagramom 4.8.*



Slika 4.8: Kutijasti dijagram na bazi medijana za izmjerene vrijednosti varijable placa_prije.

Iz tablice 4.9 i kutijastog dijagrama 4.8 možemo izvesti sljedeće i slične zaključke:

- *najniža godišnja plaća u uzorku iznosi* 16000*, a najviša* 42400*,*
- *bar* 25% *ispitanika iz uzorka ima plaću manju ili jednaku* 20950*,*
- *bar* 25% *ispitanika iz uzorka ima plaću veću ili jednaku* 26250*,*
- *bar* 50% *ispitanika iz uzorka ima plaću manju ili jednaku medijanu, tj.* 23650*,*
- *bar* 50% *ispitanika iz uzorka ima plaću veću ili jednaku* 23650*.*

Na kutijastom dijagramu mogu se označiti i takozvane **stršeće vrijednosti** ako postoje, a radi se o podacima koji su po svojoj vrijednosti značajno veći ili manji u odnosu na druge izmjerene vrijednosti promatrane varijable. Pojavljivanje je stršećih vrijednosti najčešće vezano uz jedan od sljedećih razloga:

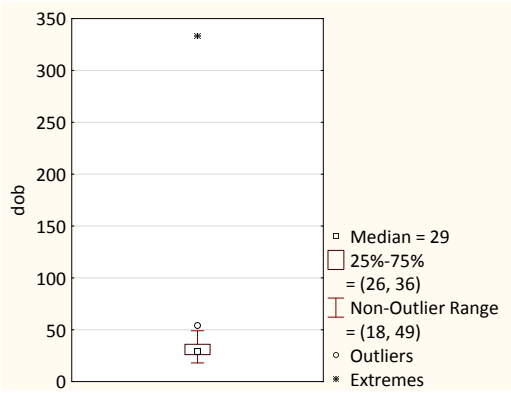
- podatak je ili netočno izmjeren ili krivo unesen u bazu podataka,
- podatak dolazi iz druge populacije (ne iz populacije koju promatramo u kontekstu problema kojega proučavamo) - npr. ako u varijablu čije su izmjerene vrijednosti godišnje plaće 1000 poreznih obveznika u Hrvatskoj upišemo godišnju plaću Microsoftovog managera iz SAD-a, taj će podatak biti stršeća vrijednost,
- podatak je točno izmjeren i unesen u bazu, ali predstavlja rijetku pojavu u populaciji - npr. ako se u varijabli čije su izmjerene vrijednosti koncentracije glukoze u krvi za 1000 osoba nađe točno izmjerena vrijednost 46.7, taj ćemo podatak smatrati stršećom vrijednošću jer se radi o vrlo visokoj koncentraciji glukoze koja se rijetko pojavljuje.

Primjer 4.23. (djelatnici.xls)
Varijabla dob *iz baze podataka* djelatnici.sta *za svakog ispitanika iz uzorka djelatnika promatranog poduzeća sadrži informaciju o dobi u godinama. Iz deskriptivne statistike te varijable (tablica 4.10) vidimo da je dob od* 333 *godine stršeći podatak i zaključujemo da se radi o podatku koji je pogrešno upisan u bazu podataka.*

veličina uzorka	aritmetička sredina	mod	frek. moda	st. dev.	varijanca
100	33.83	28	12	31.05	964.28
minimum	**donji kvartil**	**medijan**	**gornji kvartil**	**maksimum**	**raspon**
18	26	29	36	333	315

Tablica 4.10: Deskriptivna statistika varijable dob.

Osim iz tablice 4.10, stršeću vrijednost među izmjerenim vrijednostima varijable dob *mogli smo detektirati i pomoću kutijastog dijagrama na bazi medijana.*



Slika 4.9: Kutijasti dijagram na bazi medijana s prikazom stršećih vrijednosti za izmjerene vrijednosti varijable dob.

Kao što vidimo iz kutijastog dijagrama 4.9, i dob od 54 *godine prepoznata je kao stršeći podatak. Budući da je sasvim razumljivo da promatrano poduzeće može imati djelatnika starog* 54 *godine, taj podatak smatramo točno izmjerenim i točno upisanim u bazu podataka, no radi se o dobi koja se rijetko pojavljuje u populaciji djelatnika tog poduzeća. Tako iz kategorizirane tablice frekvencija i relativnih frekvencija (tablica 4.11) varijable* dob *vidimo da je takvih djelatnika samo* 0.03%.

dob	frekvencija	relativna frekvencija
$[18, 28\rangle$	37	0.37
$[28, 38\rangle$	44	0.44
$[38, 48\rangle$	15	0.15
$[48, 58\rangle$	3	0.03
$[58, 333]$	1	0.01

Tablica 4.11: Kategorizirana tablica frekvencija i relativnih frekvencijna varijable dob.

4.1.3 Metode opisivanja ordinalnih varijabli

Ordinalne se varijable najčešće zadaju tako da mogu primiti samo nekoliko međusobno različitih vrijednosti i one su po svojoj prirodi kvalitativne, zbog čega su metode opisivanja kvalitativnih varijabli primjenjive u potpunosti i na ordinalne varijable. S obzirom da se vrijednosti ordinalinih varijabli izražavaju brojčano, često se u primjeni može vidjeti da se za ordinalne varijable izražavaju, komentiraju i koriste numeričke karakteristike podataka. Ponekad to ima smisla, ali moramo upozoriti da treba dobro razmisliti u svakom pojedinom slučaju je li i kako je moguće iskoristiti

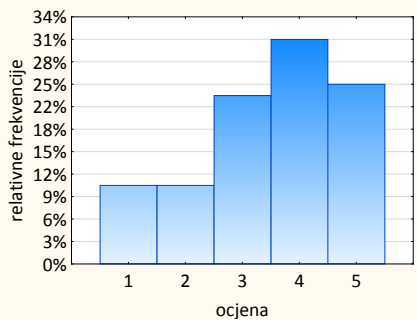
informaciju koju daje odabrana numerička karakteristika.

Primjer 4.24. *U hrvatskom su obrazovnom sustavu ocjene su od* 1 *do* 5. *Prilikom ocjenjivanja znanja na pisanim testovima nerijetko se ocjena* 1 *postiže za manje od* 40% *mogućih bodova, dok su ostale ocjene raspoređene na intervalu od* 40% *do* 100% *tako da svakoj ocjeni odgovara bodovni interval iste duljine. Pretpostavimo li da je prosjek svih postignutih ocjena 1.5, možemo li taj broj tumačiti u odnosu na znanje mjereno brojem bodova koje se ocjenjuje? Ipak, odredimo li da je medijan 1.5, možemo zaključiti da je barem* 50% *učenika postiglo manje od* 40% *bodova na testu.*

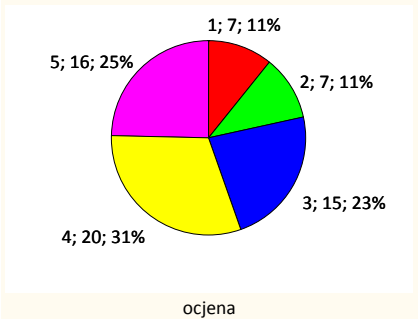
Primjer 4.25. (ocjena.xls)
Varijabla ocjena *baze podataka* ocjena.xls *ordinalna je varijabla koja sadrži ocjene na usmenom ispitu iz jednog kolegija za uzorak od* 65 *studenata iste generacije. Dakle, varijabla* ocjena *prima vrijednosti iz skupa* $1, 2, 3, 4, 5$. *Frekvencije i relativne frekvencije zabilježenih vrijednosti varijable* ocjena *dane su u tablici 4.12. Relativne frekvencije grafički su prikazane stupčastim dijagramom, a frekvencije i relativne frekvencije kružnim dijagramom na slici 4.10.*

ocjena	frekvencija	relativna frekvencija
1	7	$7/65 \approx 0.11$
2	7	$7/65 \approx 0.11$
3	15	$3/13 \approx 0.23$
4	20	$4/13 \approx 0.31$
5	16	$16/65 \approx 0.25$

Tablica 4.12: Tablica frekvencija i relativnih frekvencija svih zabilježenih vrijednosti varijable ocjena.



(a) stupčasti dijagram



(b) kružni dijagram

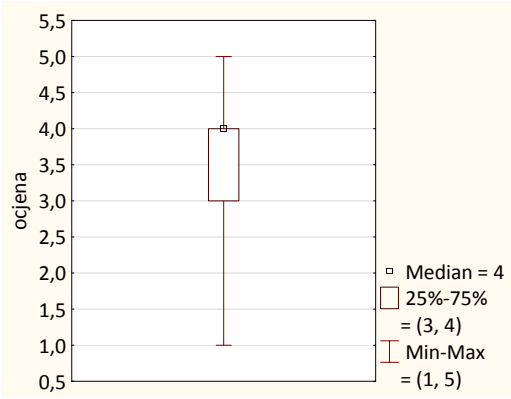
Slika 4.10: Grafički prikazi frekvencija i relativnih frekvencija svih zabilježenih vrijednosti varijable ocjena.

Numeričke karakteristike podataka varijable ocjena *prikazane su u tablici 4.13.*

veličina uzorka	aritmetička sredina	mod	frek. moda	st. dev.	varijanca
65	3.48	4	20	1.63	1.28
minimum	**donji kvartil**	**medijan**	**gornji kvartil**	**maksimum**	**raspon**
1	3	4	4	5	4

Tablica 4.13: Deskriptivna statistika varijable ocjena.

Odnosi minimuma, donjeg kvartila, medijana, gornjeg kvartila i maksimuma izmjerenih vrijednosti varijable ocjena *prikazani su kutijastim dijagramom 4.11.*



Slika 4.11: Kutijasti dijagram na bazi medijana za izmjerene vrijednosti varijable ocjena.

Aritmetičku sredinu, varijancu i standardnu devijaciju ne bismo komentirali jer ne znamo koliko se stvarno znanje krije iza svake ocjene, ali na temelju ostalih numeričkih karakteristika možemo izvesti sljedeće i slične zaključke:

- *na tom uzorku postignuta je i najniža i najviša moguća ocjena*
- *bar* 25% *studenata iz uzorka ocijenjeno je ocjenom* 3 *ili manjom od* 3, *dok je bar* 75% *studenata ocijenjeno ocjenom* 3 *ili većom.*
- *bar* 50% *studenata iz uzorka ocijenjeno je ocjenom* 4 *ili* 5.

4.2 Statistički model

Pretpostavka je statističkog zaključivanja da se donose zaključci o varijablama od kojih baram jedna ima slučajan karakter, a na temelju prikupljenih ralizacija za te varijable. Zbog toga je slučajna varijabla (slučajni vektor) matematički objekt koji je temelj svakoga statističkog zaključivanja.

Da bi primjenjivali metode statističkog zaključivanja, prva je pretpostavka postavljanje statističkog modela koji koristimo za modeliranje prikupljenih podataka. Zadati statistički model znači opisati poznate karakteristike slučajnog vektora za kojega smatramo da naši podaci čine jednu realizaciju. S obzirom da je slučajni vektor određen svojom funkcijom distribucije, statističkim je modelom opisano ono što je unaprijed poznato o funkciji distribucije slučajnog vektora kojim modeliramo podatke, odnosno **statistički model jest familija funkcija distribucije koja se uzima u obzir za zaključivanje u danom problemu**.

Najjednostavniji su modeli koji se koriste u statistici razni modeli **jednostavnoga slučajnog uzorka**.

Definicija 4.1. *Statistički model zovemo model* **jednostavnoga slučajnog uzorka** *iz funkcije distribucije F ako za slučajni vektor $(X_1, \ldots, X_n)$, čiju realizaciju čine podaci $(x_1, \ldots, x_n)$, vrijedi:*

- *slučajne varijable $X_1, \ldots, X_n$ nezavisne su,*
- *sve slučajne varijable $X_1, \ldots, X_n$ imaju istu funkciju distribucije F.*

Kod takvih modela promatranu veličinu smatramo slučajnom varijablom s funkcijom distribucije F, a vrijednosti varijable izmjerene na jedinkama iz uzorka nezavisne su realizacije te slučajne varijable.

Zbog jednostavnosti izražavanja u nastavku teksta ćemo koristiti termin **slučajni uzorak** za slučajni vektor $(X_1, \ldots, X_n)$ statističkog modela, a termin **uzorak** za njegovu realizaciju $(x_1, \ldots, x_n)$, tj. podatke.

U nastavku ćemo opisati nekoliko karakterističnih statističkih modela koji se vrlo često koriste u praksi, kao i primjere za koje su ti modeli prikladni.

4.2.1 Problem proporcije

Da bismo pojasnili probleme koji se mogu razmatrati u okviru naziva "problem proporcije" i definirali statistički model koji koristimo, poslužit ćemo se primjerom.

Primjer 4.26. *Imamo proizvodnu traku koja proizvodi proizvod A, ali s određenim (nepoznatim) postotkom škarta. Nepoznati postotak škarta označimo s θ.*

Skupo je kontrolirati svaki proizvod koji je proizveden na toj traci, pa za zaključivanje o postotku škarta uzimamo slučajni uzorak od n proizvoda.

Uzimanje jednog proizvoda u uzorak rezultira slučajnom varijablom

$$X = \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}, \ \theta \in [0,1],$$

koja daje jedinicu ako je proizvod škart, a u suprotnom nulu. Ako pod proporcijom smatramo omjer broja neispravnih proizvoda proizvedenih na toj traci i ukupnog broja proizvoda proizvedenih na toj traci u danom periodu, onda vjerojatnost $P\{X = 1\} = \theta$ odgovara proporciji škarta u populaciji.

S obzirom da se na traci stalno proizvode novi proizvodi, ako ne dolazi do bitnih promjena u procesu proizvodnje tijekom vremena, možemo smatrati da svako uzimanje proizvoda u uzorak predstavlja realizaciju isto distribuirane slučajne varijable kao X te da su te slučajne varijable međusobno nezavisne.

Podaci dobiveni u opisanom postupku čine uređenu n-torku nula i jedinica koja je realizacija slučajnog vektora $(X_1, \ldots, X_n)$ n. j. d. slučajnih varijabli distribuiranih kao X (vidi Bernoullijevu shemu).

Skup svih mogućih ishoda slučajnog vektora opisanog u prethodnom primjeru jest

$$\mathcal{R}(X_1, \ldots, X_n) = \{(x_1, \ldots, x_n) : x_i \in \{0, 1\}\}.$$

Distribucija je zadana s

$$\begin{aligned} p(x_1, \ldots, x_n; \theta) &= P\{X_1 = x_1, \ldots, X_n = x_n\} \\ &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}, \ \ (x_1, \ldots, x_n) \in \mathcal{R}(X_1, \ldots, X_n). \end{aligned}$$

Tako nastao slučajni vektor zovemo **jednostavni slučajni uzorak iz Bernoullijeve distribucije** s parametrom $\theta \in [0, 1]$.

Ako s F_θ označimo funkciju distribucije tog slučajnog uzorka za dani θ, onda statistički model u okviru kojega se bavimo zaključivanjem o proporciji čini familija funkcija distribucije:

$$\mathcal{P} = \{F_\theta : \theta \in [0, 1]\}.$$

Zovemo ga **statistički model jednostavnoga slučajnog uzorka iz Bernoullijeve distribucije**.

Uočimo da je taj model $\mathcal{P}$ indeksiran parametrom θ, pa takav model zovemo **parametarski**.

Model slučajnog uzorka iz Bernoullijeve distribucije možemo koristiti u mnogo drugih praktičnih primjera.

Primjer 4.27. *U primjeru 4.26 uzorak od n proizvoda konstruiran je uzimanjem proizvoda s proizvodne trake. To možemo interpretirati kao odabir konačnog uzorka iz beskonačne populacije. U slučaju odabira konačnog uzorka iz konačne populacije, npr. uzorka od n vijaka iste vrste iz skladišta koje sadrži N takvih vijaka, moramo razlikovati dva slučaja:*

(1) *izvlačenje jednog po jednog vijka iz konačnog skupa vijaka, pri čemu se izvučeni element vraća u skup,*

(2) *izvlačenje jednog po jednog vijka iz konačnog skupa vijaka, pri čemu se izvučeni element ne vraća u skup.*

Slučaj (1).
Označimo s v broj neispravnih vijaka u tom skladištu. To znači da je proporcija neispravnih vijaka $\theta = v/N$. Označimo s X slučajnu varijablu koja se realizira nulom u slučaju da pri odabiru jednog vijka iz skladišta izvučemo ispravan, a jedinicom u slučaju da izvučemo neispravan vijak. Dakle, vidimo da je $P\{X = 1\} = v/N$, a $P\{X = 0\} = 1 - (v/N)$, tj. distribucija slučajne varijable X dana je tablicom

$$X = \begin{pmatrix} 0 & 1 \\ 1-(v/N) & v/N \end{pmatrix}.$$

Ako iz tog skladišta na slučajan način odaberemo n vijaka na način opisan u (1) i za svaki izvučeni vijak zabilježimo je li ispravan (0) ili neispravan (1), dobijemo jednu uređenu n-torku nula i jedinica koju shvaćamo kao jednu realizaciju slučajnog vektora $(X_1, \ldots, X_n)$ sa slikom $\mathcal{R}(X) = \{(x_1, \ldots, x_n) : x_i \in \{0, 1\}\}$. Zbog načina provođenja izvlačenja opisanog pod (1), komponente su slučajnog vektora $(X_1, \ldots, X_n)$ međusobno nezavisne i sve su jednako distribuirane kao slučajna varijabla X. Distribucija tog slučajnog vektora zadana je izrazom

$$p(x_1, \ldots, x_n; v) = \prod_{i=1}^{n} P\{X_i = x_i\} = \prod_{i=1}^{n} \left(\frac{v}{N}\right)^{x_i} \left(1 - \frac{v}{N}\right)^{n-x_i}. \tag{4.1}$$

Prema tome, $(X_1, \ldots, X_n)$ jest jednostavni slučajni uzorak iz Bernoullijeve distribucije s parametrom $\theta = v/N$, a familija pripadnih funkcija distribucije $\{F_\theta : \theta \in [0, 1]\}$ određenih distribucijom 4.1 pripadni je parametarski statistički model. Iz tog primjera vidimo da se parametarski statistički modeli za jednostavni slučajni uzorak iz populacije koja nije konačna i za jednostavni slučajni uzorak iz konačne populacije konstruiran na način opisan pod (1) podudaraju.

Slučaj (2).
Opišimo sada konstrukciju slučajnog uzorka iz konačne populacije na način opisan pod (2), tj. bez vraćanja u uzorak prethodno izvučenih elemenata. Slučajne varijable $X_1, \ldots, X_n$, kojima modeliramo ishode izvlačenja vijaka pri čemu nas zanima je li izvučeni vijak dobar ili lošs, nisu nezavisne. Slučajna varijabla X_1 jednako je distribuirana kao slučajna varijabla X opisana u prethodnom dijelu primjera, no slučajne varijable $X_2, \ldots X_n$ nemaju takvu distribuciju. Naime,

$$X_2|_{X_1=a_1} = \begin{pmatrix} 0 & 1 \\ 1 - \frac{v-a_1}{N-1} & \frac{v-a_1}{N-1} \end{pmatrix}, \quad a_1 \in \{0, 1\},$$

$$X_3|_{X_1=a_1, X_2=a_2} = \begin{pmatrix} 0 & 1 \\ 1 - \frac{v-a_1-a_2}{N-2} & \frac{v-a_1-a_2}{N-2} \end{pmatrix}, \quad a_1, a_2 \in \{0, 1\},$$

odnosno općenito za $i \in \{3, \ldots, N\}$

$$X_i|_{X_1=a_1,\ldots,X_{i-1}=a_{i-1}} = \begin{pmatrix} 0 & 1 \\ 1 - \frac{v-\sum_{j=1}^{i-1} a_j}{N-(i-1)} & \frac{v-\sum_{j=1}^{i-1} a_j}{N-(i-1)} \end{pmatrix}, \quad a_1, \ldots, a_{i-1} \in \{0, 1\}.$$

Distribucija slučajnog vektora $(X_1, \ldots, X_n)$ *dana je na sljedeći način:*

$$p(x_1, \ldots, x_n; v) = P\{X_1 = x_1, \ldots X_n = x_n\} =$$
$$= P\{X_1 = x_1, \ldots X_{n-1} = x_{n-1} | X_n = x_n\} P\{X_n = x_n\} = \ldots =$$
$$= P\{X_1 = x_1\} P\{X_2 = x_2 | X_1 = x_1\} \prod_{i=3}^{n} P\{X_i = x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\}. \tag{4.2}$$

Očito distribucija slučajne varijable X_1 *i uvjetne distribucije slučajnih varijabli* $X_2, \ldots, X_n$ *ovise o nepoznatoj proporciji neispravnih vijaka u velikom skladištu, tj. o parametru* $\theta = v/N \in [0, 1]$. *Prema tome, familija pripadnih funkcija distribucije* $\{F_\theta : \theta \in [0, 1]\}$ *određenih distribucijom 4.2 pripadni je parametarski statistički model.*
Uočimo da distribuciju 4.2 možemo aproksimirati distribucijom 4.1 ukoliko je

$$\frac{v - \sum_{j=1}^{i-1} a_j}{N - (i-1)} \approx \frac{v}{N}.$$

Riječima rečeno, za veliki N *i mali* v, *parametarski statistički model temeljen na odabiru* n *elemenata iz populacije veličine* N *bez vraćanja prethodno izvučenih elemenata u tu populaciju možemo aproksimirati parametarskim statističkim modelom temeljenim na odabiru* n *elemenata iz iste populacije s vraćanjem prethodno izvučenih elemenata u tu populaciju.*

4.2.2 Problem očekivanja i varijance normalne distribucije

Za mnogo kvantitativnih slučajnih pojava može se pretpostaviti da imaju normalnu distribuciju. Činjenica je to koju se može potkrijepiti centralnim graničnim teoremima. Naime, u centralnom graničnom teoremu (vidi poglavlje 3.5.2) navedeno je da se suma n. j. d. slučajnih varijabli koje imaju konačnu varijancu asimptotski ponaša po zakonu normalne distribucije.

Te jake pretpostavke o nezavisnosti i jednakoj distribuiranosti slučajnih varijabli koje se sumiraju mogu se u velikoj mjeri relaksirati, a da zaključak o asimptotski normalnom ponašanju sume ostane istinit (vidi npr. [26]).

Kao posljedicu takvih rezultata imamo činjenicu da vrlo često pojave, čiji je slučajan karakter rezultat sume puno slučajnih utjecaja koje ne možemo posebno analizirati, imaju distribuciju toliko blisku normalnoj distribuciji da u statističkoj analizi ne možemo primijetiti razlike između stvarne distribucije i aproksimativne normalne distribucije.

Za takve pojave provodimo statističko zaključivanje kao da su normalno distribuirane.

Primjer 4.28. *Beskonačna traka pakira šećer u pakovanja na kojima je deklarirana neto masa od* 1 *kg. Međutim, dovoljno preciznom vagom moguće je utvrditi da neto masa šećera pokazuje odstupanja od* 1 *kg. Po karakteru slučajne greške koja nastaje pri pakiranju šećera na tekućoj*

vrpci, može se smatrati da masa šećera u pakiranjima delariranim kao 1 *kg zapravo ima normalnu distribuciju.*
Uzimanje jednog pakiranja i precizno vaganje rezultira podatkom koji je realizacija normalne slučajne varijable s očekivanjem μ *i varijancom* σ^2*. Uzimamo li s trake* n *pakiranja i važemo im sadržaj, imamo podatke* $(x_1, x_2, \ldots, x_n)$ *koji su jedna realizacija n. j. d. slučajnog vektora* $(X_1, X_2, \ldots, X_n)$*.*

Slučajni vektor $(X_1, X_2, \ldots, X_n)$ zovemo **jednostavni slučajni uzorak iz normalne distribucije s parametrima** μ **i** σ^2, ako su $X_1, X_2, \ldots, X_n$ nezavisne i jednako distribuirane slučajne varijable s distribucijom koja je $\mathcal{N}(\mu, \sigma^2)$.

Ako je $F_{(\mu,\sigma^2)}$ funkcija distribucije normalne slučajne varijable za dani (μ, σ^2), onda je funkcija distribucije jednostavnoga slučajnog uzorka iz te distribucije (zbog nezavisnosti):

$$\mathbf{F}_{(\mu,\sigma^2)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{(\mu,\sigma^2)}(x_i).$$

Dakle, statistički model u okviru kojega se bavimo zaključivanjem o očekivanju i varijanci normalne distribucije jest familija funkcija distribucija

$$\mathcal{P} = \{\mathbf{F}_{(\mu,\sigma^2)} : (\mu, \sigma^2) \in \mathbb{R} \times \langle 0, \infty \rangle\}, \tag{4.3}$$

a zovemo ga **statistički model jednostavnoga slučajnog uzorka iz normalne distribucije**.

Uočimo da je taj model $\mathcal{P}$ također indeksiran parametrom, ovoga puta dvodimenzionalnim (μ, σ^2), pa je i takav model parametarski.

Model slučajnog uzorka iz normalne distribucije možemo koristiti u mnogo praktičnih primjera.

Primjer 4.29. *Prosječnu godišnju količinu kiše koja padne na određenom području ima smisla promatrati kao slučajnu veličinu, tj. modelirati slučajnom varijablom. Naime, neka su* X_i*,* $i \in \{1, \ldots, 365\}$*, slučajne varijable kojima modeliramo količinu kiše koja na promatranom području padne tijekom* i*-tog dana u neprijestupnoj godini. Tada prosječnu količinu kiše koja na tom području padne tijekom jedne neprijestupne godine modeliramo slučajnom varijablom*

$$Y = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad n = 365,$$

za koju zbog centralnog graničnog teorema možemo smatrati da ima normalnu distribuciju s parametrima μ *i* σ^2*. Dakle, ako raspolažemo podacima o prosječnoj godišnjoj količini kiše na nekom području za* 100 *neprijestupnih godina, zapravo raspolažemo jednom realizacijom slučajnog vektora* $(Y_1, \ldots, Y_{100})$*. Ako su istinite pretpostavke o nezavisnosti i jednakoj distribuiranosti slučajnih varijabli* $Y_1, \ldots, Y_{100}$*, ima smisla u zaključivanju o prosječnoj godišnjoj količini kiše na tom području koristiti statistički model jednostavnoga slučajnog uzorka iz normalne distribucije s očekivanjem* μ *i varijancom* σ^2*.*

4.2.3 Jednostavna linearna regresija

Da bismo lakše objasnili statistički model, poslužit ćemo se jednim primjerom.

Primjer 4.30. *U jednom istraživanju želi se odrediti ovisi li prosječna dnevna potrošnja goriva zaposlene osobe u Hrvatskoj, koja je vlasnik automobila i ima vozačku dozvolu, o udaljenosti mjesta stanovanja osobe od radnog mjesta. Pod pojmom se "prosječna dnevna potrošnja goriva" ovdje podrazumijeva ukupna potrošnja goriva u godini dana podijeljena s* 365.

Udaljenost mjesta stanovanja od radnog mjesta za svaku pojedinu osobu može se smatrati točno određenom, tj. determinističkom veličinom. Međutim, prosječnu dnevnu potrošnju goriva svakako treba tretirati kao slučajnu veličinu. S obzirom na to da se radi o prosječnoj potrošnji, zahvaljujući centralnom graničnom teoremu, za potrebe statističkog zaključivanja moći ćemo se osloniti na činjenicu da je ta veličina normalno distribuirana.

Statistički model za zaključivanje o vezi između determinističke veličine x (udaljenost mjesta stanovanja od radnog mjesta) i slučajne veličine Y_x (prosječna dnevna potrošnja goriva), izgradit ćemo u ovom primjeru postavljanjem linearne veze između tih veličina:

$$Y_x = ax + b + \varepsilon_x, \quad a, b \in \mathbb{R},$$

gdje ε_x predstavlja slučajnu varijablu greške modela za dani x.

U navedenom primjeru, kao i u mnogim drugim primjerima sličnog tipa, prikupljanjem podataka u uzorak imamo n uređenih parova brojeva

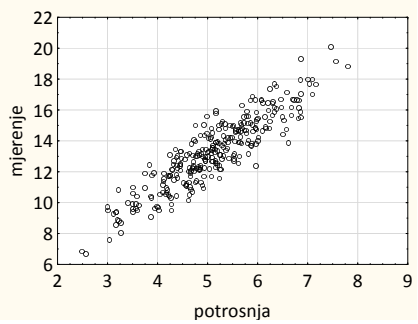
$$(x_1, y_1), \ldots, (x_n, y_n)$$

koji predstavljaju realizacije uređenih parova koje ćemo označiti

$$(x_1, Y_1), \ldots, (x_n, Y_n),$$

a kod kojih je druga komponenta slučajna varijabla. Uočimo da slučajne varijable iz tih parova, tj. $Y_1, \ldots, Y_n$ ne moraju biti jednako distribuirane. Naprotiv, pretpostavka je da njihova distribucija ovisi o iznosu prve koordinate iz para.

Primjer 4.31. (automobili.xls)
Varijabla potrosnja *baze podataka* automobili.sta *sadrži podatke o potrošnji goriva novog modela automobila pri brzini od* 110 *km/h za* 300 *nezavisnih mjerenja, dok varijabla* mjerenje *sadrži vrijednosti nekog parametra izmjerenog na tehničkom pregledu tog automobila nakon takve vožnje, a za kojeg se pretpostavlja da bi kod tehnički ispravnog automobila trebao biti linearno povezan s prosječnom potrošnjom automobila pri velikim brzinama. U kontekstu jednostavne linearne regresije izmjerene vrijednosti varijable* potrosnja *označavamo s $x_1, \ldots x_{300}$, a izmjerene vrijednosti varijable* mjerenje *označavamo s $(y_1, \ldots, y_{300})$ i smatramo ih realizacijama nezavisnih slučajnih varijabli $Y_1, \ldots, Y_{300}$ kojima modeliramo rezultate provedenih mjerenja. Parove (x_1, y_1),..., (x_{300}, y_{300}) izmjerenih vrijednosti varijabli* potrosnja *i* mjerenje *grafički prikazujemo dijagramom raspršenosti (eng. scatterplot) 4.12.*



Slika 4.12: Dijagram raspršenosti za izmjerene vrijednosti varijabli potrosnja i mjerenje.

U postupku ćemo modeliranja pretpostaviti sljedeće:

- $EY_i = ax_i + b$ za svaki $i = 1, \ldots, n$, tj. ako je $\varepsilon_i = Y_i - ax_i - b$, onda je $E\varepsilon_i = 0$.

- $Y_1, \ldots, Y_n$ jesu međusobno nezavisne i imaju istu varijancu, označimo je σ^2. Ako je ε vektor $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^\tau$, onda je matrica kovarijanci tog vektora oblika $\sigma^2 I$ (I je oznaka za jediničnu matricu n-tog reda).

- $Y_1, \ldots, Y_n$ su normalno distribuirane slučajne varijable. Dakle, $Y_i \sim \mathcal{N}(ax_i + b, \sigma^2)$.

Statistički model za zaključivanje o parametrima a, b i σ^2 definirat ćemo korištenjem familije dopuštenih funkcija distribucije za slučajni vektor $(Y_1, \ldots, Y_n)$. Njegova funkcija distribucije određena je funkcijom gustoće

$$f(y_1, \ldots, y_n; a, b, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \prod_{i=1}^{n} e^{-\frac{(y_i - ax_i - b)^2}{2\sigma^2}} =$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ax_i - b)^2}$$

u kojoj su tri nepoznata parametra a, b i σ^2, dok su $x_1, \ldots, x_n$ izmjerene vrijednosti prve komponente uređenih parova podataka.

Takav statistički model zovemo **klasičan jednostavan linearan regresijski model**.

4.3 Procjena parametra

U parametarskim statističkim modelima osnovu statističkog zaključivanja čini zaključivanje o parametrima. Jedan je od osnovnih problema njihova procjena na temelju realizacije uzorka iz danog statističkog modela. Za procjene parametra koristimo procjenitelje.

Definicija 4.2. *Neka je $\mathcal{P} = \{F_\theta : \theta \in \mathbf{\Theta} \subseteq \mathbb{R}^k\}$ parametarski statistički model, θ k-dimenzionalan parametar, a $\mathbf{\Theta} \subseteq \mathbb{R}^k$ prostor parametra, tj. skup svih dozvoljenih vrijednosti nepoznatog parametra θ. Neka je $(X_1, \ldots, X_n)$ slučajni vektor s distribucijom iz $\mathcal{P}$ i $t : \mathbb{R}^n \to \mathbf{\Theta}$. Slučajni vektor $\mathbf{T} = t(X_1, \ldots, X_n)$ jest* **procjenitelj** *za θ.*[1]

Primjer 4.32. *U modelu je jednostavnoga slučajnog uzorka iz Bernoullijeve distribucije s parametrom $p \in [0, 1]$ relativna frekvencija jedinice jedan procjenitelj za p.*

Primjer 4.33. *U modelu jednostavnoga slučajnog uzorka iz $\mathcal{N}(\mu, \sigma^2)$ označimo*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

$$\overline{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Tada je $(\overline{X}_n, \overline{S}_n^2)$ jedan procjenitelj za (μ, σ^2), a $(\overline{X}_n, S_n^2)$ drugi procjenitelj za taj parametar.

Problem je kako izabrati funkciju t procjenitelja u pojedinom statističkom modelu. Takvim pitanjima bavi se matematička statistika. Ovdje ćemo samo navesti nekoliko svojstava koja su vrlo poželjna da ih procjenitelj parametra ima i intuitivno ilustrirati zašto su ta svojstva korisna.

Definicija 4.3. *Neka je $\mathcal{P} = \{F_\theta : \theta \in \mathbf{\Theta} \subseteq \mathbb{R}^k\}$ parametarski statistički model, θ k-dimenzionalan parametar. Procjenitelj T* **nepristran** *je ako za svaki $\theta \in \mathbf{\Theta}$ vrijedi*

$$E_\theta T = E_\theta\, t(X_1, \ldots, X_n) = \theta.$$

Ukoliko je nepoznati parametar jednodimenzionalan, a predloženi procjenitelj nepristran, poželjno je da takav procjenitelj ima što manju varijancu. Naime, varijanca je jedna mjera raspršenosti oko očekivanja pa je za očekivati da će se procjena s većom vjerojatnošću realizirati u malim okolinama oko stvarne vrijednosti parametra

[1] Uočite da distribucija procjenitelja ovisi o vrijednosti nepoznatog parametra θ. Da bi tu činjenicu naglasili, vjerojatnost skupa $\{\mathbf{T} \in A\}$ označavamo s $P_\theta\{\mathbf{T} \in A\}$, a očekivanje od $\mathbf{T}$ s $E_\theta\mathbf{T}$.

(tj. očekivanja procjenitelja) ako je varijanca manja. Zbog toga, uspoređujući dva nepristrana procjenitelja, kažemo da je **efikasniji** onaj koji ima manju varijancu. Osim toga, poželjno je da varijanca nepristranog procjenitelja konvergira u nulu s porastom veličine uzorka pa za jednodimenzionalan parametar definiramo svojstvo konzistentnosti sljedećom definicijom.

Definicija 4.4. *Neka je $\mathcal{P} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ parametarski statistički model, a θ jednodimenzionalan parametar. Procjenitelj T* **konzistentan** *je ako za pripadni niz procjenitelja $T_n = t(X_1, \ldots, X_n)$, $n \in \mathbb{N}$, za svaki $\varepsilon > 0$ i $\theta \in \Theta$ vrijedi*

$$\lim_{n\to\infty} P_\theta\{|T_n - \theta| \geq \varepsilon\} = 0.$$

Primjer 4.34. *Neko poduzeće sve svoje finalne proizvode raspoređuje po kvaliteti u tri skupine: I, II i III kategoriju. Iskustveno je poznato da su vjerojatnosti da finalni proizvod bude I kategorije i da finalni proizvod bude II kategorije jednake. Ovu vjerojatnost označimo s θ. Dakle, slučajna varijabla koja opisuje kategoriju kvalitete proizvoda može se zadati na sljedeći način:*

$$X = \begin{pmatrix} 1 & 2 & 3 \\ \theta & \theta & 1-2\theta \end{pmatrix}.$$

Parametar θ može primiti bilo koju vrijednost iz intervala $[0, 1/2]$ da bi ta slučajna varijabla bila dobro definirana. Problem koji se postavlja pred statističara procjena je parametra θ na temelju n-dimenzionalnog slučajnog uzorka.

U skladu s prethodnim primjerom za očekivati je da se relativna frekvencija neke kategorije može iskoristiti u svrhu procjene s obzirom da θ ima značenje vjerojatnosti da se realizira proizvod prve kategorije, ali također i vjerojatnost da se realizira proizvod druge kategorije. Tako se može razmišljati o procjeniteljima:

$$T_1 = \frac{f_1}{n}, \quad T_2 = \frac{f_2}{n},$$

gdje je f_1 frekvencija jedinice, a f_2 frekvencija dvojke u uzorku.

Uočimo da su oba procjenitelja nepristrana. Naime, $f_1 \sim \mathcal{B}(n, \theta)$ i $f_2 \sim \mathcal{B}(n, \theta)$ pa je

$$E_\theta\, T_1 = E_\theta\, T_2 = \theta.$$

Možemo također izračunati i varijance:

$$\mathrm{Var}_\theta\, T_1 = \mathrm{Var}_\theta\, T_2 = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}.$$

Korištenjem se slabog zakona velikih brojeva lako vidi da su oba procjenitelja konzistentna. Na temelju tih analiza ne možemo preferirati niti jednog od navedenih dvaju procjenitelja. Međutim, razmotrimo kao varijantu i procjenitelja koji je vezan uz relativnu frekvenciju treće kategorije proizvoda (f_3):

$$T_3 = \frac{1}{2}\left(1 - \frac{f_3}{n}\right).$$

S obzirom da je $f_3 \sim \mathcal{B}(n, 1-2\theta)$, $E_\theta\, T_3 = \theta$, to je i nepristran procjenitelj za θ. Varijanca je

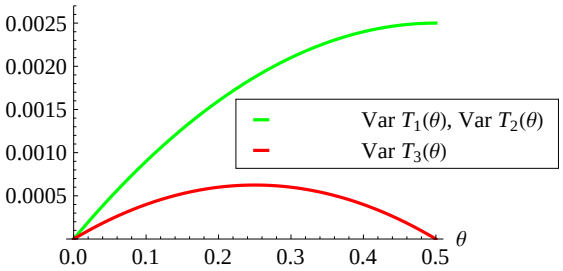
$$\mathrm{Var}_\theta\, T_3 = E_\theta(T_3 - \theta)^2 = \frac{1}{4n^2}\mathrm{Var}_\theta\, f_3 = \frac{1}{4n^2} n(1-2\theta)2\theta = \frac{\theta}{2n}(1-2\theta).$$

Dakle, to je također konzistenatan procjenitelj. Usporedimo li varijance od T_1 i T_2 s varijancom od T_3 za svaki $\theta \in \Theta$ i $n \in \mathbb{N}$, vidimo da je

$$\operatorname{Var}_\theta T_3(\theta, n) \leq \operatorname{Var}_\theta T_1(\theta, n) = \operatorname{Var}_\theta T_2(\theta, n),$$

pa je T_3 efikasniji procjenitelj za θ (slika 4.13). Napomenimo da taj rezultat nije neočekivan s obzirom da procjenitelj T_3 u sebi sadrži informacije koje nose oba prethodna procjenitelja jer vrijedi:

$$T_3 = \frac{1}{2}(T_1 + T_2).$$



Slika 4.13: Varijance procjenitelja T_1, T_2 i T_3 za $n = 100$.

4.3.1 Procjena proporcije

Neka je $(X_1, \ldots, X_n)$ jednostavni slučajni uzorak iz Bernoullijeve distribucije

$$X = \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}, \quad \theta \in [0, 1].$$

Tada je relativna frekvencija jedinice, tj.

$$T = \frac{f_1}{n}$$

jedan nepristran i konzistentan procjenitelj za parametar θ koji ima značenje vjerojatnosti realizacije jedinice.

Zaista, s obzirom da slučajna varijabla koja daje broj uspjeha u n nezavisnih Bernoullijevih pokusa ima binomnu distribuciju s parametrima n i θ, to je $f_1 \sim \mathcal{B}(n, \theta)$. Vrijedi:

$$E_\theta \frac{f_1}{n} = \frac{n\theta}{n} = \theta.$$

Konzistentnost je posljedica slabog zakona velikih brojeva.

Primjer 4.35. *Promotrimo igraću kockicu sa svojstvom da se pri jednom bacanju broj* $i \in \{1,\dots,6\}$ *okrene s vjerojatnošću* p_i*, pri čemu je*

$$p_i > 0, \quad \forall i \in \{1,\dots,6\}, \qquad \text{i} \qquad \sum_{i=1}^{6} p_i = 1.$$

Želimo procijeniti vjerojatnost da se pri bacanju te kockice okrene šestica, stoga označimo s 1 *događaj "okrenula se šestica", a s* 0 *događaj "nije se okrenula šestica". U tom kontekstu rezultat bacanja kockice modeliramo Bernoullijevom slučajnom varijablom*

$$X = \begin{pmatrix} 0 & 1 \\ 1-p_6 & p_6 \end{pmatrix},$$

gdje je p_6 *nepoznati parametar kojega ćemo procijeniti.*
Pretpostavimo da smo kockicu bacili nezavisno sto puta zaredom. Realizaciju tih 100 *bacanja modeliramo slučajnim vektorom* $(X_1,\dots X_{100})$ *čije su komponente nezavisne i jednako distribuirane kao slučajna varijabla* X*, tj.* $(X_1,\dots X_{100})$ *je jednostavni slučajni uzorak iz Bernoullijeve distribucije s parametrom* p_6*. Nadalje, pretpostavimo da se u tih* 100 *bacanja šestica realizirala* 17 *puta. To znači da je realizacija slučajnog uzorka* $(X_1,\dots X_{100})$ *uređena* 100*-torka koja se sastoji od* 17 *jedinica i* 83 *nule. Slijedi da je relativna rekvencija pojavljivanja šestice, broj* 17/100*, procjena vjerojatnosti* p_6*. Analogno bismo postupali pri procjeni bilo koje od preostalih vjerojatnosti* p_i*,* $i \in \{1,\dots,5\}$*.*

Primjer 4.36. (djelatnici.xls)
Varijabla dob *baze podataka* djelatnici.xls *sadrži informacije o godinama starosti za svakoga od* 100 *djelatnika iz reprezentativnog uzorka zaposlenika tvornice A. Označimo s* θ *vjerojatnost da je slučajno odabrani djelatnik te tvornice stariji od* 30 *godina te procijenimo* θ*. Znamo da svaki djelatnik te tvornice ima ili najviše* 30 *godina (oznaka* 0*) ili je stariji od* 30 *godina (oznaka* 1*) pa rezultat ispitivanja starosti u tom kontekstu možemo modelirati Bernoullijevom slučajnom varijablom*

$$X = \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}, \quad \theta \in [0,1].$$

Prema tome, podatke iz varijable dob *shvaćamo kao jednu realizaciju jednostavnog slučajnog uzorka* $(X_1,\dots,X_{100})$ *iz te Bernoullijeve distribucije i na temelju te realizacije parametar* θ *procjenjujemo relativnom frekvencijom jedinica, tj. relativnom frekvencijom djelatnika iz uzorka koji su stariji od* 30 *godina. Kako takvih djelatnika u uzorku ima* 46*, parametar* θ *procjenjujemo brojem* $46/100 = 0.46$*.*

4.3.2 Procjena očekivanja

Neka je $(X_1, X_2, \dots, X_n)$ jednostavni slučajni uzorak iz distribucije F_μ, gdje je $\mu \in \mathbb{R}$ nepoznato očekivanje. Dakle, $(X_1, X_2, \dots, X_n)$ je vektor n. j. d. slučajnih varijabli iz distribucije F_μ. Pretpostavimo da varijanca distribucije F_μ također postoji i označimo je sa σ^2.

Ako razmatramo problem procjene očekivanja, možemo iskoristiti **aritmetičku sredinu uzorka**

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

koja je nepristran i konzistentan procijenitelj za očekivanje. Naime,

$$E_\mu \overline{X}_n = \mu, \quad \mathrm{Var}_\mu\, \overline{X}_n = \frac{\sigma^2}{n},$$

a konzistentnost slijedi iz slabog zakona velikih brojeva.

Primjer 4.37. *Ocjenu znanja studenata nekog fakulteta na usmenom ispitu iz jednog kolegija modeliramo diskretnom slučajnom varijablom X sa slikom $\mathcal{R}(X) = \{1, 2, 3, 4, 5\}$ i funkcijom distribucije F_μ, gdje je μ nepoznato očekivanje koje želimo procijeniti. Pretpostavimo da od svih studenata fakulteta koji su ocijenjeni na usmenom ispitu tog kolegija odabiremo reprezentativan uzorak od 65 studenata. Frekvencije ocjena na tom uzorku dane su tablicom frekvencija 4.14.*

ocjena	frekvencija
1	7
2	7
3	15
4	20
5	16

Tablica 4.14: Tablica frekvencija ocjena studenata na usmenom ispitu iz jednog kolegija.

Ocjene prikupljene na uzorku, predstavljene kao uređena 65-torka, čine jednu realizaciju jednostavnog slučajnog uzorka $(X_1, \ldots, X_{65})$ iz distribucije F_μ. Procjenitelj za očekivanje od X jest slučajna varijabla $\overline{X}_{65}$. Njezinu realizaciju na uzorku $(x_1, \ldots, x_{65})$ označimo s $\overline{x}_{65}$. Prema tome, procjena očekivanja slučajne varijable X na temelju tablice frekvencija 4.14 jest

$$\overline{x}_{65} = \frac{1 \cdot 7 + 2 \cdot 7 + 3 \cdot 15 + 4 \cdot 20 + 5 \cdot 16}{65} \approx 3.48.$$

Primjer 4.38. (kolokvij.xls)
U bazi podataka kolokvij.sta *nalaze se rezultati dvaju kolokvija iz nekog kolegija (varijable* kolokvij_1 *i* kolokvij_2*) te ukupan broj bodova ostvaren na kolokvijima (varijabla* bodovi_ukupno*) za reprezentativan uzorak od 70 studenata koji su pristupili tim kolokvijima. Na svakom od kolokvija bilo je moguće ostvariti najviše 100 bodova. Dakle, na obama kolokvijima bilo je moguće ostvariti najviše 200 bodova.*
Usmjerimo se na ukupan broj bodova ostvaren na obama kolokvijima. Njega modeliramo kontinuiranom slučajnom varijablom X s funkcijom distribucije F_μ, gdje je μ nepoznato očekivanje koje želimo procijeniti.
Postignuti ukupni bodovi za promatrani uzorak od 70 studenata (podaci iz varijable bodovi_ukupno*) jedna su realizacija jednostavnog slučajnog uzorka $(X_1, \ldots, X_{70})$ iz distribucije F_μ. Prema tome, procjenitelj $\overline{X}_{70}$ za očekivanje od X realizira se aritmetičkom sredinom podataka iz varijable* bodovi_ukupno*, tj. procjena je očekivanja od X $\overline{x}_{70} \approx 96.9$.*

Primjer 4.39. (djelatnici.xls)
Varijabla placa_prije *baze podataka* djelatnici.xls *sadrži iznos godišnje plaće za reprezentativan uzorak od 100 djelatnika tvornice A. Godišnju plaću djelatnika te tvornice modeliramo kontinuiranom slučajnom varijablom X s funkcijom distribucije F_μ, gdje je μ nepoznato očekivanje koje želimo procijeniti.*

Iznosi godišnjih plaća za promatrani uzorak od 100 *djelatnika (podaci iz varijable* placa_prije*) jedna su realizacija jednostavnog slučajnog uzorka* $(X_1, \ldots, X_{100})$ *iz distribucije* F_μ*. Prema tome, procjenitelj* $\overline{X}_{100}$ *za očekivanje od* X *realizira se aritmetičkom sredinom podataka iz varijable* placa_prije*, tj. procjena je očekivanja od* X $\overline{x}_{100} \approx 24522$*.*

4.3.3 Procjena varijance

Neka je $(X_1, X_2, \ldots, X_n)$ jednostavni slučajni uzorak iz distribucije F_θ, gdje je $\theta = \sigma^2 > 0$ nepoznata varijanca. Dakle, $(X_1, X_2, \ldots, X_n)$ je vektor n. j. d. slučajnih varijabli iz distribucije F_θ. Označimo s μ očekivanje te distribucije.

Ako razmatramo problem procjene varijance, možemo iskoristiti **varijancu uzorka**

$$\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Međutim, taj procjenitelj nije nepristran. Naime,

$$\begin{aligned}\overline{S}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu) - (\overline{X}_n - \mu)]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(\overline{X}_n - \mu)^2 - \frac{2}{n}(\overline{X}_n - \mu)\sum_{i=1}^{n}(X_i - \mu).\end{aligned}$$

Budući da je

$$\sum_{i=1}^{n}(\overline{X}_n - \mu)^2 = n(\overline{X}_n - \mu)^2, \quad \sum_{i=1}^{n}(X_i - \mu) = n(\overline{X}_n - \mu),$$

slijedi da je

$$\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X}_n - \mu)^2.$$

Dakle,

$$E_\theta\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}E_\theta[(X_i - \mu)^2] - E_\theta[(\overline{X}_n - \mu)^2] = \frac{1}{n}n\sigma^2 - \frac{\sigma^2}{n},$$

tj.

$$E_\theta\overline{S}_n^2 = \frac{n-1}{n}\theta.$$

Ako napravimo malu korekciju varijance uzorka i definiramo novog procjenitelja, **korigiranu varijanca uzorka**

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

njegovo očekivanje bit će upravo jednako varijanci obilježja koje proučavamo, pa ćemo tako dobiti zadovoljeno svojstvo nepristranosti.
Zaista,

$$E_\theta S_n^2 = E_\theta \left(\frac{n}{n-1} S_n^2 \right) = \theta. \tag{4.4}$$

Može se pokazati (vidi npr. [21]) da je S_n^2 ujedno i konzistentan procjenitelj.

Primjer 4.40. (kolokvij.xls)
Neka je X diskretna slučajna vrijabla iz primjera 4.38 kojom smo modelirali ukupan broj bodova studenta postignutih na dvama kolokvijima iz nekog kolegija. Znamo da je procjena očekivanja od X dana s $\overline{x}_{70} \approx 96.9$. Za procjenu varijance koristimo standardne procjenitelje: varijancu uzorka $\overline{S}_{70}^2$ i korigiranu varijancu uzorka S_{70}^2. Za realizaciju uzorka iz funkcije distribucije od X, danu u varijabli bodovi_ukupno*, $\overline{S}_{70}^2$ realizira se procjenom*

$$\overline{s}_{70}^2 = 43.56,$$

a S_{70}^2 procjenom

$$s_{70}^2 = 44.19.$$

Primjer 4.41. (djelatnici.xls)
Standardni tablični kalkulatori i programski paketi za statističku analizu imaju ugrađenu funkciju za procjenu varijance definiranu na temelju korigirane varijance uzorka, tj. na temelju procjenitelja S_n^2. Procijenimo na taj način varijancu neprekidne slučajne varijable X kojom modeliramo godišnju plaću zaposlenika u jednoj tvornici (primjer 4.39). Za zabilježene vrijednosti slučajne varijable X na uzorku od 100 zaposlenika te tvornice (varijabla placa_prije*) korigirana varijanca uzorka S_{100}^2 realizira se procjenom $s_{100}^2 \approx 26069208.1 = 5105.8^2$.*

4.3.4 Procjena funkcije distribucije

Neka je $(X_1, X_2, \ldots, X_n)$ jednostavmo slučajni uzorak iz nepoznate funkcije distribucije F slučajne varijable X. Statistički model takvog slučajnog uzorka ne možemo smatrati parametarskim s obzirom da familija dozvoljenih funkcija distribucije nije indeksirana nepoznatim parametrom, nego je potpuno neodređena. Takav statistički model zvat ćemo **neparametarski**.

U takvom modelu opisat ćemo jedan način procjene funkcije distribucije F. Preciznije, za svaki pojedini $x \in \mathbb{R}$ procjenu vrijednosti funkcije distribucije, $F(x)$.

S obzirom da je $F(x)$ po definiciji $P\{X_i \leq x\}$, za svaki $i \in \{1, \ldots, n\}$ i za dani x zapravo imamo problem procjene vjerojatnosti uspjeha (uspjeh znači da se realizirao događaj $\{X_i \leq x\}$) pa se možemo poslužiti rezultatima iz procjene proporcije.

Dakle, neka je x fiksan realan broj, a F funkcija distribucije slučajne varijable X. Definirajmo slučajnu varijablu

$$I_{\{X\leq x\}} = \begin{cases} 1 \ , \ X \leq x \\ 0 \ , \ X > x. \end{cases}$$

Njezina je distribucija:

$$I_{\{X\leq x\}} = \begin{pmatrix} 0 & 1 \\ 1-F(x) & F(x) \end{pmatrix}, \quad F(x) \in [0,1].$$

Uočimo da ovdje $F(x)$ igra ulogu parametra distribucije slučajne varijable $I_{\{X\leq x\}}$.

Iz problema procjene proporcije znamo da je dobar procjenitelj za parametar Bernoullijeve distribucije relativna frekvencija jedinice, a u slučaju varijable $I_{\{X\leq x\}}$ to je relativna frekvencija skupa $\{X \leq x\}$. Označimo

$$\hat{F}(x) = \frac{f_{\{X\leq x\}}}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\leq x\}}.$$

Tada je

$$E(\hat{F}(x)) = F(x)$$

za svaki pojedini $x \in \mathbb{R}$ i

$$Var\,(\hat{F}(x)) = \frac{F(x)(1-F(x))}{n},$$

pa možemo govoriti o nepristranosti i konzistentnosti procjenitelja $\hat{F}(x)$ za svaki pojedini $x \in \mathbb{R}$. Tako definiran procjenitelj funkcije distribucije na temelju realizacije jednostavnog slučajnog uzorka zove se **empirijska funkcija distribucije**.

Primjer 4.42. djelatnici.xls

Diskretna numerička varijabla rukovodstvo *baze podataka* djelatnici.xls *iz primjera 4.5 prima vrijednosti iz skupa* $\{0, 1, \ldots, N\}$*, gdje je* N *najveći mogući broj godina radnog staža koje djelatnik može provesti na rukovodećoj poziciji. Frekvencije i relativne frekvencije zabilježenih vrijednosti varijable* rukovodstvo *za djelatnike iz promatranog uzorka dane su u tablici 4.6. Procijenimo funkciju distribucije diskretne slučajne varijable* X *kojom modeliramo varijablu* rukovodstvo*. U tu svrhu promotrimo kumulativne frekvencije zabilježenih vrijednosti te varijable (tablica 4.15).*

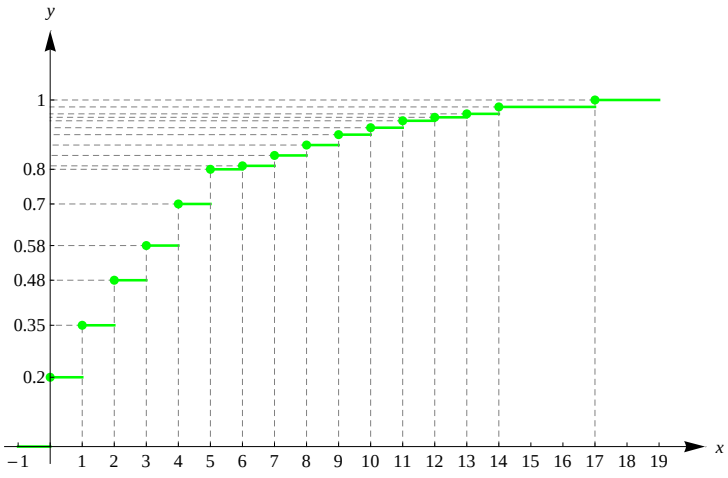
rukovodstvo	frek.	kumulativna frek.	rukovodstvo	frek.	kumulativna frek.
0	20	0.20	8	3	0.87
1	15	0.35	9	3	0.9
2	13	0.48	10	2	0.92
3	10	0.58	11	2	0.94
4	12	0.7	12	1	0.95
5	10	0.8	13	1	0.96
6	1	0.81	14	2	0.98
7	3	0.84	17	2	1

Tablica 4.15: Tablica frekvencija i kumulativnih frekvencija svih zabilježenih vrijednosti varijable rukovodstvo.

Iz tablice 4.15, preciznije, iz stupca s kumulativnim frekvencijama, slijedi da je procjena funkcije distribucije slučajne varijable X, tj. njezina empirijska funkcija distribucije, definirana izrazom

$$\hat{F}(x) = \frac{f_{\{X \le x\}}}{100} = \begin{cases} 0 & , \quad x < 0 \\ 0.2 & , \ 0 \le x < 1 \\ 0.35 & , \ 1 \le x < 2 \\ 0.48 & , \ 2 \le x < 3 \\ 0.58 & , \ 3 \le x < 4 \\ 0.7 & , \ 4 \le x < 5 \\ \vdots & \vdots \quad \vdots \\ 1 & , \quad x \ge 17. \end{cases}$$

Graf funkcije $\hat{F}(x)$ prikazan je na slici 4.14.

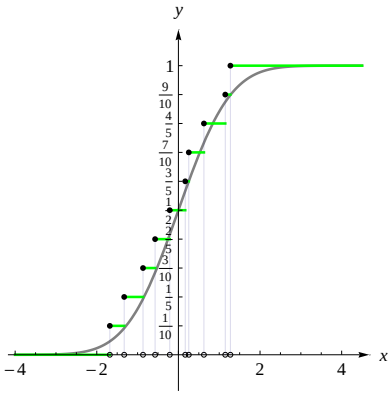


Slika 4.14: Empirijska funkcija distribucije slučajne varijable X iz primjera 4.44.

Primjer 4.43. *Funkcija distribucije standardne normalne slučajne varijable definirana je pravilom*

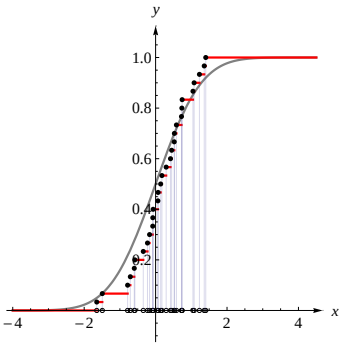
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \quad x \in \mathbb{R}.$$

Graf te funkcije prikazan je na slici 2.20 u drugom poglavlju. U ovom ćemo primjeru na temelju simuliranih vrijednosti iz standardne normalne distribucije (tj. simulirane realizacije n-dimenzionalnog jednostavnog slučajnog uzorka iz standardne normalne distribucije) ilustrirati ponašanje procjene $\hat{F}$ za F u ovisnosti o veličini uzorka.
Na temelju jedne simulacije 10 *vrijednosti iz te distribucije dobili smo graf funkcije $\hat{F}(x)$ prikazan na slici 4.15 (stepenasti graf).*

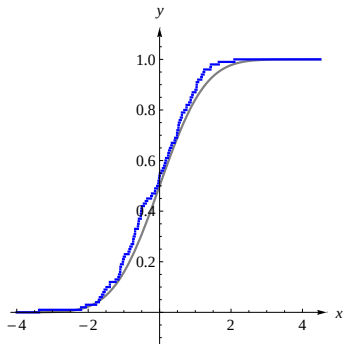


Slika 4.15: Grafovi funkcije distribucije i empirijske funkcije distribucije ($n = 10$) standardne normalne slučajne varijable.

Grafovi empirijskih funkcija distribucije standardne normalne slučajne varijable za 30 *i* 100 *simuliranih vrijednosti prikazani su na slikama 4.16 i 4.17.*



Slika 4.16: Grafovi funkcije distribucije i empirijske funkcije distribucije ($n = 30$) standardne normalne slučajne varijable.



Slika 4.17: Grafovi funkcije distribucije i empirijske funkcije distribucije ($n = 100$) standardne normalne slučajne varijable.

Korištenjem nekog programskog paketa simulirajte 1000 *podataka iz standardne normalne distribucije i skicirajte empirijsku funkciju distribucije. Usporedite ju sa stvarnom funkcijom distribucije.*

Primjer 4.44. (djelatnici.xls)
Procijenimo funkciju distribucije neprekidne slučajne varijable X iz primjera 4.39 kojom modeliramo godišnju plaću djelatnika jedne tvornice. Varijabla placa_prije *sadrži mnogo različitih realizacija slučajne varijable X. Uočimo da su frekvencije svih realizacija male. Empirijska funkcija distribucije temeljena na svih* 77 *različitih realizacija slučajne varijable X može se lako očitati korištenjem relativnih kumulativnih frekvencija. Tako, npr. $\hat{F}(18400) = 0.05$ (tablica 4.16).*

plaća	frekvencija	kumulativna frekvencija
16000	1	1
18100	1	2
18300	1	3
18400	2	5
⋮	⋮	⋮
24500	1	57
24600	4	61
⋮	⋮	⋮
42400	1	100

Tablica 4.16: Tablica kumulativnih frekvencija plaća djelatnika iz primjera 4.39.

4.3.5 Procjena parametara u jednostavnoj linearnoj regresiji

Ideja za procjenu parametara u linernom regresijskom modelu (poglavlje 4.2.3) temelji se na jednostavnom zahtjevu da treba minimizirati sumu kvadrata odstupanja

eksperimentalnih od teorijskih vrijednosti.

Neka je $(x_1, y_1), \ldots, (x_n, y_n)$ dana realizacija slučajnog uzorka iz regresijskog problema. Onda su $y_1, \ldots, y_n$ tzv. eksperimentalne vrijednosti, tj. realizacije slučajnog vektora $(Y_1, \ldots, Y_n)$ kojega želimo opisati modelom

$$Y_i = ax_i + b + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.5}$$

kao što je navedeno u poglavlju 4.2.3. S obzirom da ε_i predstavljaju slučajne varijable koje opisuju grešku modela, teorijske su vrijednosti $y(x_i) = ax_i + b$. Jedan procjenitelj za nepoznati vektorski parametar (a, b) može se dobiti minimizacijom izraza

$$S(a, b) = \sum_{i=1}^{n} (y(x_i) - y_i)^2$$

po $(a, b) \in \mathbb{R} \times \mathbb{R}$.

Uvedimo oznake:

$$\mathbf{x} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad \beta = [a, b]^\tau, \quad \mathbf{y} = [y_1, \ldots, y_n]^\tau.$$

Rješenje je toga minimizacijskog problema

$$\hat{\beta} = (\mathbf{x}^\tau \mathbf{x})^{-1} \mathbf{x}^\tau \mathbf{y}. \tag{4.6}$$

Uvrstimo li slučajni vektor $\mathbf{Y} = [Y_1, \ldots Y_n]^\tau$ u izraz za $\hat{\beta}$ umjesto $\mathbf{y}$, imamo procjenitelja za nepoznati parametar β.

Može se pokazati da je tako definiran procjenitelj za β nepristran. Za dokaze i ostala svojstva procjenitelja vidi npr. [21] ili [29].

Primjer 4.45. (automobili.xls)
Baza podataka automobili.xls *opisana je u primjeru 4.12. U istom primjeru opisano je i značenje izmjerenih vrijednosti varijabli* potrosnja *i* mjerenje *u jednostavnom linearnom regresijskom modelu:* $x_1, \ldots x_{300}$ *prosječne su potrošnje promatranog tipa automobila pri vožnji po autocesti brzinom* 110 *km/h (varijabla* potrosnja*), a* $y_1, \ldots, y_{300}$ *realizacije su slučajnog vektora* $(Y_1, \ldots, Y_{300})$ *kojim modeliramo rezultate mjerenja nekog parametra na tehničkom pregledu tog automobila nakon te vožnje, a za kojega se pretpostavlja da bi kod tehnički ispravnog automobila trebao biti*

linearno povezan s prosječnom potrošnjom automobila pri velikim brzinama. Uvrštavanjem podataka iz baze automobili.xls *u izraz 4.6 i primjenom nekog statističkog softvera možemo dobiti procjene parametara a i b regresijskog modela 4.5 u ovom primjeru:*

$$\widehat{a} = 2.138, \quad \widehat{b} = 2.349.$$

Slijedi da modelom određen rezultat mjerenja promatranog parametra na tehničkom pregledu tog automobila nakon vožnje u opisanim uvjetima računamo pomoću pravila

$$y(x) = \widehat{a}x + \widehat{b} = 2.138x + 2.349.$$

4.4 Procjena parametra pouzdanim intervalom

U ovom poglavlju pretpostavit ćemo da je statistički model za podatke parametarski s **jednodimenzionalnim** parametrom θ kojega želimo procijeniti. Parametar procjenjujemo korištenjem procjenitelja T. S obzirom da je parametar jednodimenzionalan, procjenitelj je slučajna varijabla. Procjena je na temelju podataka samo jedna realizacija procjenitelja. Koristeći distribuciju procjenitelja u danom statističkom modelu moguće je dobiti i druge informacije o stvarnoj vrijednosti procijenjenog parametra, ne samo jedan broj koji predstavlja procjenu.

U ovom poglavlju opisat ćemo koncept procjene parametra **pouzdanim intervalom** i interpretaciju takve procjene te ilustrirati procjenu pouzdanim intervalom na nekoliko primjera.

Pouzdani interval za procjenu parametra po svojoj definiciji nije interval s realnim brojevima kao granicama nego interval kojemu su granice slučajne varijable pa ga možemo zvati **slučajni interval**.

Definicija 4.5. *Neka je $\gamma \in \langle 0, 1\rangle$, a slučajni vektor $(X_1, \ldots, X_n)$ s distribucijom iz parametarske familije $\mathcal{P} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ neka određuje statistički model. Ako postoje dva procjenitelja $D = d(X_1, \ldots, X_n)$ i $G = g(X_1, \ldots, X_n)$ za paramear θ sa svojstvima:*

- $d(x_1, \ldots, x_n) \leq g(x_1, \ldots, x_n)$, *za sve* $(x_1, \ldots, x_n) \in \mathcal{R}(X_1, \ldots, X_n)$,
- $P_\theta\{D \leq \theta \leq G\} \geq \gamma$, *za sve* $\theta \in \Theta$,

onda kažemo da je slučajni interval $[D, G]$ pouzdani interval za θ pouzdanosti γ.

Kao što se može prepoznati iz definicije, pouzdani interval određuje se na temelju zahtjeva da se stvarna vrijednost parametra nalazi u slučajnom intervalu s vjerojatnošću barem γ.

Ako na temelju podataka izračunamo realizacije procjenitelja D i G koji su granice slučajnog intervala, dobit ćemo običan interval s realnim brojevima kao granicama. Dakle, realizacija slučajnog intervala običan je interval realnih brojeva. Iz definicije pouzdanog intervala pouzdanosti γ jasno je da će, pod pretpostavkom adekvatnosti statističkog modela, u približno $100\gamma\,\%$ slučajeva izračunati interval realnih brojeva sadržavati stvarnu vrijednost parametra.

Dakle, interval pouzdanosti γ jest slučajni interval, tj. granice su mu slučajne varijable. Jedna realizacija intervala pouzdanosti γ, određena na osnovu podataka (realizacije slučajnog vektora statističkog modela), običan je interval realnih brojeva. Uobičajneno je u praksi i tu realizaciju pouzdanog intervala također zvati pouzdani interval. Međutim, važno je znati razliku između pouzdanog intervala kao slučajnog intervala i njegove realizacije - običnog intervala realnih brojeva.

U nastavku ilustriramo postupak određivanja pouzdanih intervala na nekoliko primjera. Zainteresirani čitatelj više o tom problemu može pronaći u literaturi koja opširnije obrađuje teme iz matematičke statistike kao npr. [2], [9], [16], [21], [22], [24].

4.4.1 Procjena očekivanja pouzdanim intervalom za velike uzorke

Neka je $(X_1, \ldots, X_n)$ jednostavni slučajni uzorak iz zadane distibucije, označimo je F_μ, za koju je poznata varijanca i iznosi σ^2, a očekivanje je μ i želimo ga procijeniti iz jedne realizacije uzorka.

Aritmetička sredina uzorka $\overline{X}_n$ jest jedan nepristran i konzistentan procjenitelj za očekivanje. Osim toga, na temelju centralnog graničnog teorema poznato je da standardizirana aritmetička sredina, tj.

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma},$$

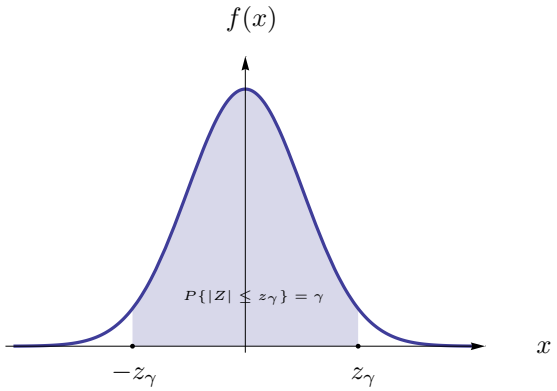
ima asimptotski standardnu normalnu distribuciju. Označimo sa $Z \sim \mathcal{N}(0,1)$. Korištenjem tih rezultata konstruirat ćemo pouzdani interval za očekivanje na temelju jednostavnoga slučajnog uzorka iz distribucije F_μ.

Neka je $\gamma \in \langle 0, 1\rangle$ izabrana pouzdanost i z_γ broj za koji vrijedi

$$P\{|Z| \leq z_\gamma\} = \gamma.$$

Uočimo da vrijednost γ predstavlja površinu ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $[-z_\gamma, z_\gamma]$ (slika 4.18), tj.

$$P\{|Z| \leq z_\gamma\} = \frac{1}{\sqrt{2\pi}} \int_{-z_\gamma}^{z_\gamma} e^{-x^2/2}\, dx = \gamma.$$



Slika 4.18: Vjerojatnost $P\{|Z| \leq z_\gamma\}$.

S obzirom da distribuciju od $\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$ možemo dobro aproksimirati standardnom normalnom distribucijom za velike n, vrijedi:

$$P_\mu\left\{-z_\gamma \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq z_\gamma\right\} = P_\mu\left\{\overline{X}_n - z_\gamma \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right\}.$$

Dakle, vrijedi:

$$P_\mu\left\{\mu \in \left[\overline{X}_n - z_\gamma \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right]\right\} \approx \gamma,$$

pa slučajmi interval $\left[\overline{X}_n - z_\gamma \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right]$ možemo smatrati pouzdanim intervalom za μ pouzdanosti γ ukoliko imamo veliku veličinu uzorka n.

Postupak procjene očekivanja pouzdanim intervalom na temelju realizacije $(x_1, \ldots, x_n)$ jednostavnog slučajnog uzorka iz distribucije s varijancom σ^2, tj. standardnom devijacijom σ, možemo opisati sljedećim koracima:

- odrediti broj z_γ za koji vrijedi da je $P\{|Z| \leq z_\gamma\} = \gamma$, gdje je $Z \sim \mathcal{N}(0,1)$,
- izračunati interval po formuli

$$\left[\overline{x}_n - z_\gamma \frac{\sigma}{\sqrt{n}},\ \overline{x}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right]. \tag{4.7}$$

Interval će izračunat tim postupkom u približno $100\gamma\%$ slučajeva sadržati stvarnu vrijednost očekivanja μ.

U praksi najčešće ne znamo stvarnu vrijednost standardne devijacije σ. S obzirom da je taj rezultat izveden za velike uzorke, može se također dokazati da će korištenje procjenitelja za standardnu devijaciju (korijena korigirane varijance uzorka) umjesto standardne devijacije u izvodu formule za pouzdani interval, dati analognu asimptotsku distribuciju. Iz tog se razloga u praksi koristi procjena standardne devijacije umjesto stvarne vrijednosti standardne devijacije ako je ona nepoznata.

Primjer 4.46. (kolokvij.xls)
Baza podataka kolokvij.sta *ukratko je opisana u primjeru 4.38. Procijenimo intervalom pouzdanosti* $\gamma = 0.95$ *očekivanje diskretne slučajne varijable* X *kojom modeliramo ukupan broj bodova ostvaren na tim dvama kolokvijima. Da bismo izračunali realizaciju traženog intervala pouzdanosti, trebaju nam sljedeće vrijednosti*

$$n = 70, \quad \overline{x}_{70} \approx 96.9, \quad s_{70} \approx 44.19, \quad z_{0.95} \approx 1.96,$$

gdje su $\overline{x}_{70}$ *i* s_{70} *redom procjene očekivanja i varijance slučajne varijable* X*. Korištenjem izraza 4.7 slijedi da je traženi interval*

$$[86.36, 107.44].$$

Procjena istog očekivanja intervalom pouzdanosti $\gamma = 0.97$ *zahtijeva poznavanje vrijednosti* $z_{0.97} \approx 2.17$*. Tražena je realizacija intervala pouzdanosti* $\gamma = 0.97$ *interval*

$$[85.19, 108.61].$$

Uočimo da se za veću vrijednost γ *interval pouzdanosti realizira širim intervalom realnih brojeva nego za manju vrijednost* γ*.*

Primjer 4.47. (djelatnici.xls)
Procijenimo intervalom pouzdanosti $\gamma = 0.97$ *očekivanje neprekidne slučajne varijable iz primjera 4.39 kojom modeliramo godišnju plaću djelatnika jedne tvornice (varijabla* placa_prije*). Da bismo izračunali realizaciju traženog intervala pouzdanosti, trebaju nam sljedeće vrijednosti*

$$n = 100, \quad \overline{x}_{100} \approx 24522, \quad s_{100} \approx 5105.8, \quad z_{0.97} \approx 2.17,$$

gdje su $\overline{x}_{100}$ *i* s_{100} *redom procjene očekivanja i varijance slučajne varijable* X *određene u primjerima 4.39 i 4.41. Korištenjem izraza 4.7 slijedi da je traženi interval*

$$[23414, 25630].$$

4.4.2 Procjena proporcije pouzdanim intervalom za velike uzorke

U statističkom modelu koji smo nazvali "problem proporcije" zaključujemo o vjerojatnosti da se realizira broj 1 na temelju n-dimenzionalnoga jednostavnog slučajnog uzorka iz Bernoullijeve distribucije.

Nepristran i konzistentan procjenitelj koji u tu svrhu koristimo jest relativna frekvencija jedinice, tj.

$$\frac{f_1}{n}.$$

Slično prethodnom poglavlju, centralni granični teorem ovdje omogućuje određivanje asimptotske distribucije standardizirane relativne frekvencije jedinice, pa ćemo na temelju tog rezultata konstruirati pouzdani interval za proporciju.

Neka je $(X_1, \ldots, X_n)$ jednostavni slučajni uzorak iz Bernoullijeve distribucije s parametrom $\theta \in \langle 0, 1\rangle$. Tada je

$$\frac{f_1}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Osim toga, $f_1 \sim \mathcal{B}(n, \theta)$, pa je

$$E_\theta \frac{f_1}{n} = \theta, \quad \mathrm{Var}_\theta \frac{f_1}{n} = \frac{\theta(1-\theta)}{n}.$$

Prema centralnom graničnom teoremu sada slijedi da slučajna varijabla

$$\sqrt{n}\frac{\frac{f_1}{n} - \theta}{\sqrt{\theta(1-\theta)}}$$

ima asimptotski standardnu normalnu distribuciju. Označimo sa $Z \sim \mathcal{N}(0,1)$ i z_γ odaberimo kao u prethodnom poglavlju, tj. $P\{|Z| \leq z_\gamma\} = \gamma$. Sada vrijedi:

$$\gamma \approx P_\theta \left\{ \left| \sqrt{n}\frac{\frac{f_1}{n} - \theta}{\sqrt{\theta(1-\theta)}} \right| \leq z_\gamma \right\}.$$

Nejednadžba

$$\left| \sqrt{n}\frac{\frac{f_1}{n} - \theta}{\sqrt{\theta(1-\theta)}} \right| \leq z_\gamma$$

bit će zadovoljena ako je $\theta \in [\theta_1, \theta_2]$, gdje su θ_1 i θ_2 rješenja jednadžbe

$$\theta^2(n + z_\gamma^2) - \theta(2f_1 + z_\gamma^2) + \frac{f_1^2}{n} = 0.$$

Pod pretpostavkom da je n velik, u rješenjima je te kvadratne jednadžbe n mnogo veći od z_γ pa se zanemarivanjem članova koji sadrže z_γ^2 dobije:

$$\theta_1 \approx \frac{f_1}{n} - z_\gamma\sqrt{\frac{1}{n}\frac{f_1}{n}(1 - \frac{f_1}{n})}, \;\; \theta_2 \approx \frac{f_1}{n} + z_\gamma\sqrt{\frac{1}{n}\frac{f_1}{n}(1 - \frac{f_1}{n})}.$$

Za danu je realizaciju toga jednostavnog slučajnog uzorka uobičajeno relativnu frekvenciju jedinice označiti s $\hat{p}$, a relativnu frekvenciju nule s $\hat{q}$. Korištenjem takvih oznaka i dobivenih rezultata navodimo postupak za računanje pouzdanog intervala proporcije na temelju realizacije jednostavnoga slučajnog uzorka iz Bernoullijeve distribucije:

- odrediti broj z_γ za koji vrijedi da je $P\{|Z| \leq z_\gamma\} = \gamma$, gdje je $Z \sim \mathcal{N}(0,1)$,
- izračunati interval po formuli:

$$\left[\hat{p} - z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}}\right]. \tag{4.8}$$

Primjer 4.48. *Neka je X diskretna slučajna varijabla kojom modeliramo ocjene znanja studenata jednog fakulteta na usmenom ispitu iz nekog kolegija (primjer 4.37). Procijenimo intervalom pouzdanosti $\gamma = 0.95$ proporciju studenata koji usmeni ispit polože ocjenom izvrstan. Ovdje se realizacija ocjene izvrstan smatra "uspjehom", a realizacija bilo koje druge ocjene "neuspjehom", pa iz tablice 4.14 slijedi da je*

$$\hat{p} = \frac{16}{65}, \quad \hat{q} = 1 - \frac{16}{65} = \frac{49}{65}.$$

Kako je $n = 65$, a $z_{0.95} \approx 1.96$, primjenom izraza 4.8 slijedi da je realizacija traženog intervala pouzdanosti interval

$$[0.14, 0.35].$$

Primjer 4.49. (kolokvij.xls)
Neka je X diskretna slučajna varijabla kojom modeliramo ukupan broj bodova studenta ostvaren na dvama kolokvijima iz nekog kolegija (primjer 4.38). Procijenimo intervalom pouzdanosti $\gamma = 0.95$ proporciju studenata koji su na kolokvijima ukupno ostvarili barem 100 od mogućih 200 bodova. Ovdje se ukupan broj bodova veći ili jednak 100 smatra "uspjehom", a ukupan broj bodova manji od 100 "neuspjehom", pa iz podataka dostupnih u varijabli bodovi_ukupno *slijedi da je*

$$\hat{p} = \frac{39}{70} \approx 0.56, \quad \hat{q} = 1 - \frac{39}{70} = \frac{31}{70} \approx 0.44.$$

Kako je $n = 70$, a $z_{0.95} \approx 1.96$, primjenom izraza 4.8 slijedi da je realizacija traženog intervala pouzdanosti interval

$$[0.44, 0.68].$$

Primjer 4.50. (djelatnici.xls)
Neka je X neprekidna slučajna varijabla iz primjera 4.39 kojom modeliramo godišnju plaću djelatnika jedne tvornice. Procijenimo intervalom pouzdanosti $\gamma = 0.98$ proporciju djelatnika koji godišnje zarađuju više od 30000. Ovdje se realizacija godišnje plaće veće od 30000 smatra "uspjehom", a realizacija godišnje plaće manje ili jednake 30000 "neuspjehom". Iz podataka za varijablu placa_prije *slijedi da je*

$$\hat{p} = 0.16, \quad \hat{q} = 1 - 0.16 = 0.84.$$

Budući da je $n = 100$, a $z_{0.98} \approx 2.33$, primjenom izraza 4.8 slijedi da je traženi interval

$$[0.075, 0.245].$$

4.5 Testiranje statističkih hipoteza

Općenito, cilj je testiranja hipoteze odlučiti o istinitosti ili neistinitosti slutnje koju zovemo hipoteza. **Statistička hipoteza** slutnja je koja se odnosi na statistički model.

Primjer 4.51. *O proporciji pušača izabranog lokaliteta može se zaključivati na temelju reprezentativnog uzorka koji predstavlja jednu realizaciju jednostavnog slučajnog uzorka iz Bernoullijeve distribucije s parametrom $\theta \in \langle 0, 1\rangle$. Nepoznati je parametar θ upravo proporcija pušača populacije. Jedna je slutnja/hipoteza koju možemo testirati npr. da je proporcija pušača u toj populaciji manja od 20%.*

Postupak testiranja hipoteza uvijek počinje prevođenjem slutnje koja nas zanima u statističku hipotezu. To znači formiranje statističkog modela u okviru kojega ćemo zaključivati i izricanje hipoteze u terminima koji se odnose na odabrani statistički model. S obzirom da je statistički model $\mathcal{P}$ familija funkcija distribucije koje su dozvoljene prilikom zaključivanja o zadanom problemu, očigledno je da postavljena hipoteza zapravo određuje jedan podskup od $\mathcal{P}$. U prethodnom je primjeru statistički model odabran, a hipotezu možemo iskazati kao podskup distribucija modela za koji vrijedi nejednakost $\theta \leq 0.2$.

Statističku hipotezu standardno označavamo s $\mathcal{H}$. Testirati hipotezu znači donijeti odluku o tome hoćemo li $\mathcal{H}$ odbaciti ili ne.

Odluku u postupku testiranja hipoteza donosimo na temelju odabranog kriterija i realizacije uzorka. Odabrati kriterij zapravo znači definirati pravilo za odbacivanje hipoteze. S obzirom da se odluka donosi na temelju realizacije slučajnog vektora $(X_1, \ldots, X_n)$ statističkog modela, pravilo mora biti iskazano u terminima tog slučajnog vektora . U tu svrhu koristimo pojam "statistika".

Definicija 4.6. *Neka je $(X_1, \ldots, X_n)$ slučajni vektor statističkog modela $\mathcal{P}$ na temelju kojega donosimo zaključke i neka je $t : \mathcal{R}(X_1, \ldots, X_n) \to \mathbb{R}^k$ zadana funkcija. Slučajni vektor/varijablu $T = t(X_1, \ldots, X_n)$ zovemo* **statistika** *za taj statistički model.*

Uočimo da je i procjenitelj jedna statistika, ali s dodatnim zahtjevom na skup vrijednosti funkcije t.

Slično kao kod definiranja pouzdanih intervala za procjenu nepoznatog parametra, kod odabira pravila za testiranje hipoteza koristimo svojstva statistike kao slučajnog vektora/varijable, a prilikom donošenja odluke koristimo odabranu statistiku i pravilo te realizaciju statistike na podacima.

Pravilo po kojemu odbacujemo postavljenu hipotezu podijelit će skup svih mogućih realizacija slučajnog vektora statističkog modela na dva disjunktna dijela. Uobčajeno ih označavamo s C_r i C_r^c, pri čemu je C_r dio koji odgovara odbacivanju postavljene hipoteze. C_r zovemo **kritično područje**.
S obzirom da postavljena hipoteza $\mathcal{H}$ izdvaja podskup distribucija iz statističkog modela $\mathcal{P}$, uobičajeno je uvesti i drugu hipotezu u postupku testiranja. Ta hipoteza obuhvatit će sve one distribucije koje nisu sadržane u $\mathcal{H}$. Zbog toga često govorimo o testiranju dviju hipoteza u statističkom testu. Jednu od njih zovemo **nul-hipoteza** i označavamo s $\mathcal{H}_0$, a drugu **alternativna hipoteza** i označavamo s $\mathcal{H}_1$. Alternativna hipoteza jest ona koju prihvaćamo u slučaju odbacivanja nul-hipoteze. Za statistički model $\mathcal{P}$ vrijedi:

$$\mathcal{P} = \mathcal{H}_0 \cup \mathcal{H}_1.$$

4.5.1 Pogreške statističkog testa

Odluka koja je donesena statističkim testom može biti pogrešna ili ispravna. Pritom se mogu dogoditi dva tipa pogrešne odluke:

pogreška I. tipa: **odbaciti $\mathcal{H}_0$ ako je ona istinita,**
pogreška II. tipa: **ne odbaciti $\mathcal{H}_0$ ako je $\mathcal{H}_1$ istinita.**

Vjerojatnosti pogreške prvog tipa i pogreške drugog tipa ovise o stvarnoj distribuciji slučajne varijable o kojoj testiramo hipotezu, tj. te su vjerojatnosti zapravo funkcije koje distribuciji iz $\mathcal{P}$ pridružuju broj, ali one nisu definirane na istoj domeni. Vjerojatnost pogreške prvog tipa može se računati za svaku distribuciju iz $\mathcal{H}_0$, dok se vjerojatnost pogreške drugog tipa računa za svaku distribuciju iz $\mathcal{H}_1$. Ilustrirajmo taj postupak na primjeru.

Primjer 4.52. *Pretpostavimo da želimo zaključivati o proporciji pušača izabranog lokaliteta na temelju jednostavnoga slučajnog uzorka iz Bernoullijeve distribucije s parametrom $\theta \in \langle 0,1\rangle$ (primjer 4.51). Neka je $\mathcal{P} = \{F_\theta : F_\theta$ Bernoullijeva, $\theta \in \langle 0,1\rangle\}$ pripadni statistički model. Taj model parametarski je, pa i hipoteze možemo izražavati u terminima parametra. Imamo slutnju da je proporcija pušača manja od 20%. Statistička hipoteza koja opisuje ovu slutnju jest podskup od $\mathcal{P}$ za koji vrijedi $\theta \leq 0.2$. Dakle,*

$$\mathcal{H}_0 : \ \theta \leq 0.2,$$
$$\mathcal{H}_1 : \ \theta > 0.2.$$

Za taj parametarski model podjela $\mathcal{P}$ na $\mathcal{H}_0$ i $\mathcal{H}_1$ ekvivalentna je podjeli skupa dozvoljenih vrijednosti parametra $\langle 0,1\rangle$ na dva dijela: $\langle 0, 0.2]$ i $\langle 0.2, 1\rangle$. Vjerojatnost pogreške prvog tipa i vjerojatnost pogreške drugog tipa sada možemo promatrati kao funkcije od nepoznatog parametra. Ako je odabran postupak za odbacivanje nul-hipoteze, znači da je definirano kritično područje C_r, pa se te funkcije mogu definirati na sljedeći način:

- *vjerojatnost pogreške prvog tipa α:*

$$\alpha : \langle 0, 0.2] \to [0,1], \;\; \alpha(\theta) = P_\theta\{(X_1, \dots X_n) \in C_r\},$$

- *vjerojatnost pogreške drugog tipa β:*

$$\beta : \langle 0.2, 1\rangle \to [0,1], \;\; \beta(\theta) = P_\theta\{(X_1, \dots X_n) \in C_r^c\}.$$

U prehodnom primjeru uočimo da se funkcija β može prikazati kao

$$\beta(\theta) = 1 - P_\theta\{(X_1, \dots X_n) \in C_r\}.$$

Objedinjeni prikaz obiju funkcija vjerojatnosti pogreške postiže se definiranjem funkcije jakosti testa.

Funkcija jakosti testa, u oznaci π, definirana je za svaku distribuciju iz $\mathcal{P}$ (odnosno za svaku dozvoljenu vrijednost parametra ako je statistički model parametarski) kao

$$\pi(F) = P_F\{(X_1, \dots, X_n) \in C_r\},\ F \in \mathcal{P},$$

odnosno za parametarski statistički model

$$\pi(\theta) = P_\theta\{(X_1, \dots, X_n) \in C_r\},\ \theta \in \Theta.$$

Uočimo da je

$$\alpha(F) = \pi(F), \quad \forall F \in \mathcal{H}_0,$$
$$\beta(F) = 1 - \pi(F), \quad \forall F \in \mathcal{H}_1.$$

Ilustrirajmo primjerom funkciju jakosti testa i njezinu vezu s vjerojatnostima pogreške.

Primjer 4.53. *U problemu zaključivanja o proporciji pušača na nekom lokalitetu (primjeri 4.51 i 4.52) želimo testirati hipotezu da je proporcija pušača manja od 20%. Statistički model i hipoteze opisani su u primjeru 4.52.*

$$\mathcal{H}_0 : \ \theta \leq 0.2$$
$$\mathcal{H}_1 : \ \theta > 0.2$$

Odaberimo sljedeće pravilo odbacivanja nul-hipoteze:

C_r: odbacit ćemo nul-hipotezu ako je relativna frekvencija pušača u uzorku veća od 22%.

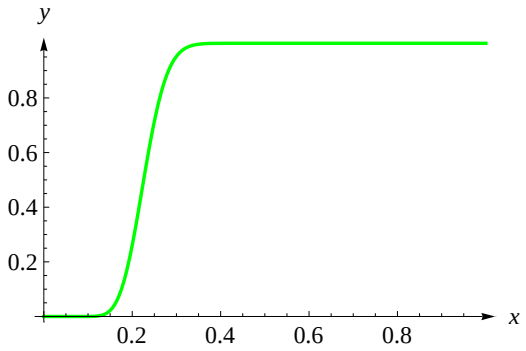
Uz standardnu oznaku f_1 za frekvenciju pušača, funkcija je jakosti za takav test dana izrazom

$$\pi(\theta) = P_\theta\{(X_1, \dots, X_n) \in C_r\} = P_\theta\left\{\frac{f_1}{n} > 0.22\right\}.$$

Pod pretpostavkam statističkog modela znamo da je $f_1 \sim \mathcal{B}(n, \theta)$. Dakle, vrijedi:

$$P_\theta\left\{\frac{f_1}{n} > 0.22\right\} = P_\theta\{f_1 > 0.22n\} = \sum_{k > 0.22n} \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Graf te funkcije za n=100 prikazan je slikom 4.19.



Slika 4.19: Funkcija jakosti testa za kritično područje $\frac{f_1}{n} > 0.22$.

S obzirom da je $\pi(0.2) = 0.261067$, *vidimo da je maksimalna vjerojatnost pogreške prvog reda*

$$\max_{\theta \in \mathcal{H}_0} \alpha(\theta) = 0.261067,$$

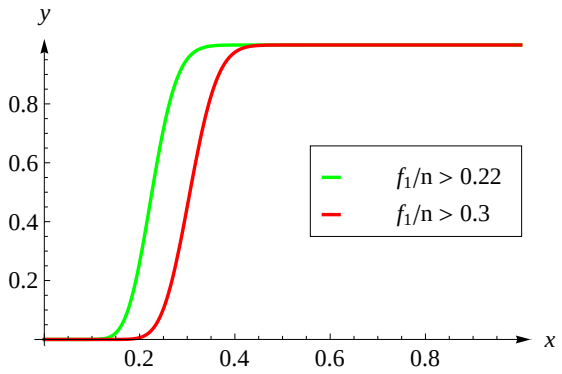
dok je maksimalna vjerojatnost pogreške drugog reda za tako definirano kritično područje

$$\max_{\theta \in \mathcal{H}_1} \beta(\theta) = 1 - 0.261067 = 0.738933$$

Odabirom drugog kritičnog područja tako da granicu za odbacivanje nul-hipoteze pomaknemo u veće vrijednosti, npr.

C_r*: odbacit ćemo nul-hipotezu ako je relativna frekvencija pušača u uzorku veća od 30%,*

možemo smanjiti maksimalnu vjerojatnost pogreške prvog tipa, ali će se povećati maksimalna vjerojatnost pogreške drugog tipa. Grafovi funkcija jakosti testa za oba navedena kritična područja i $n = 100$ *prikazani su slikom 4.20.*



Slika 4.20: Funkcije jakosti testa za kritična područja $\frac{f_1}{n} > 0.22$ i $\frac{f_1}{n} > 0.3$.

Za novo je kritično područje

$$\max_{\theta\in\mathcal{H}_0}\alpha(\theta)=0.00605934$$
$$\max_{\theta\in\mathcal{H}_1}\beta(\theta)=1-0.00605934=0.993941.$$

Povećanje maksimalne vjerojatnosti jedne pogreške ako želimo smanjiti maksimalnu vjerojatnost druge pogreške, ilustrirano prethodnim primjerom, nije neuobičajno. Dapače, može se pokazati da je nemoguće odabrati kritično područje koje bi minimiziralo maksimalnu vjerojatnost obiju pogrešaka po svim mogućim kritičnim područjima. Zbog toga se odabir kritičnog područja vrši tako da se dopušta istraživaču izbor maksimalne vjerojatnosti pogreške prvog tipa koju želi prihvatiti. Te se vrijednosti se uglavnom biraju između brojeva 0.01, 0.05 ili 0.1. Odabrana maksimalna vjerojatnost pogreške prvog tipa zove se **razina značajnosti testa** ili **nivo signifikantnosti testa** i standardno označava s α. Uz izabranu razinu značajnosti testa, test se dizajnira uz nastojanje da se maksimalna vjerojatnost pogreške drugog tipa učini što manjom. Dakle, maksimalna je vjerojatnost pogreške drugog tipa temelj za teorijsku analizu testa i odabir test-statistike i najčešće se ne iskazuje u primjeni testa. Primjena se oslanja na činjenicu da je odabrani test kreiran kao optimalan za dani skup pretpostavki.

Uzimajući u obzir da ćemo biti u mogućnosti birati maksimalnu vjerojatnost pogreške prilikom odbacivanja nul-hipoteze, to je informacija koju u primjeni testa referiramo. Npr. reći ćemo da **odbacujemo nul-hipotezu na nivou značajnosti** α, tj. da odbacujemo nul-hipotezu uz vjerojatnost najviše α da smo pri tome pogriješili. Ako pravilo testa primijenjeno na podatke sugerira da ne odbacimo nul-hipotezu, prilikom primjene testa obično nemamo dostupnu informaciju koliko iznosi maksimalna vjerojatnost da smo pogriješili. Zato ćemo tada reći kako podaci ne podupiru tvrdnju da $\mathcal{H}_0$ treba odbaciti.

Takav neravnopravan odnos između nulte i alternativne hipoteze prilikom kreiranja statističkog testa upućuje na činjenicu da nije svejedno kako smo izbrali nultu i alternativnu hipotezu i pripadni test. Ukoliko je moguće, uputno je u primjeni birati statistički test kojemu alternativna hipoteza odgovara tvrdnji koju želimo dokazati.

Iako je teorija odabira optimalne test-statistike vrlo razvijena (vidi npr. [18], [2], [21]) ovdje se nećemo baviti metodama za odabir optimalnih statistika za testiranje hipoteza odabranog modela. U nastavku samo ilustriramo intuitivnu metodu odabira postupka za testiranje hipoteza na nekoliko primjera bez analize maksimalne vjerojatnosti pogreške drugog tipa. Zainteresirani čitatelj više o tom problemu može

pronaći u literaturi koja opširnije obrađuje teme iz matematičke statistike kao npr. [2], [9], [16], [21], [22], [24].

4.5.2 Testiranje hipoteze o očekivanju za velike uzorke

U ovom poglavlju ilustrirat ćemo jedan od statističkih testova koji možemo koristiti prilikom testiranja hipoteza o iznosu očekivanja distribucije slučajne varijable na temelju koje je formiran statistički model jednostavnoga slučajnog uzorka. Problem ilustriramo sljedećim primjerom.

Primjer 4.54. *Pretpostavimo da želimo provjeriti je li očekivana vrijednost vremena čekanja u redu studentske menze u vrijeme ručka veća od pet minuta. U tu svrhu, od sto slučajno izabranih studenata koji odlaze na ručak u studentsku menzu prikupljamo podatke o vremenu čekanja. Tako dolazimo do podataka* $(x_1, \ldots, x_{100})$*. Na temelju tih podataka aritmetičkom sredinom procijenili smo očekivanje slučajne varijable iz koje potječu podaci - procjena je iznosila* 6.5 *minuta. Znajući iz prethodnih proučavanja te slučajne varijable da je njezina varijanca* 25*, želimo testirati slutnju da je očekivano vrijeme čekanja veće od 5 minuta.*

Test kojim možemo testirati takvu slutnju izvest ćemo na temelju statističkog modela jednostavnog slučajnog uzorka $(X_1, \ldots, X_n)$ iz distribucije F kojoj je poznata varijanca i iznosi σ^2, a očekivanje μ jest nepoznato.

Postavimo nultu i alternativnu hipotezu na sljedeći način:

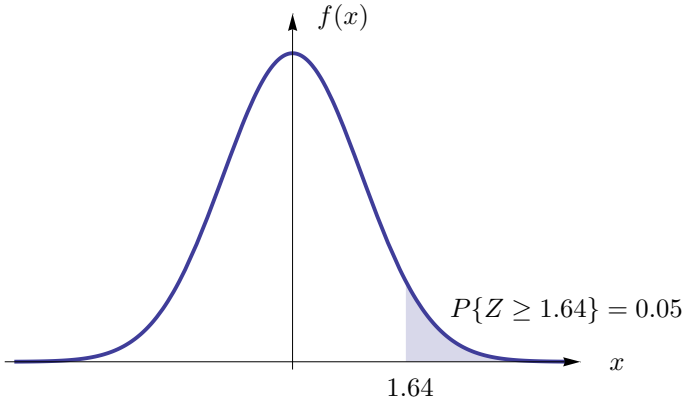
$$\mathcal{H}_0\colon\ \mu = \mu_0$$
$$\mathcal{H}_1\colon\ \mu > \mu_0.$$

Ako je $\mathcal{H}_0$ istinita hipoteza, a uzorak velik, onda je distribucija artmetičke sredine uzorka približno normalna s očekivanjem μ_0 i varijancom σ^2/n (centralni granični teorem!). Dakle, pod pretpostavkom je istinitosti nul-hipoteze distribucija od

$$Z' = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n}$$

približno standardna normalna i očekuje se realizacija od Z' blizu ili manje od nule (slika 4.21). Uočimo primjerice da se realizacije veće ili jednake 1.64 pojavljuju s vjerojatnošću približno 0.05, tj. da je

$$P\{Z' \geq 1.64\} \approx 0.05.$$

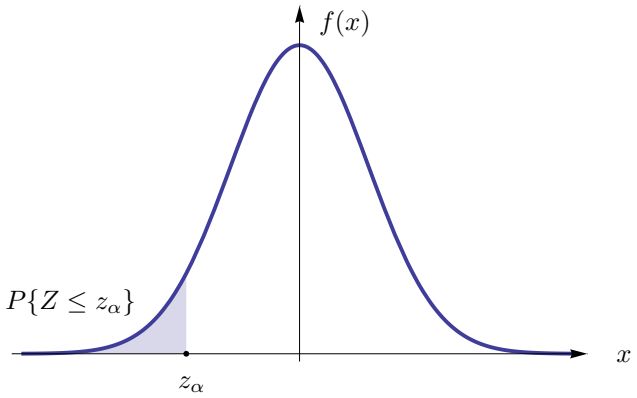


Slika 4.21: Površina ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $[1.64, \infty)$.

Pretpostavimo da se u našem slučaju Z' realizirala brojem $\hat{z}$. U uvjetima istinitosti hipoteze $\mathcal{H}_0$ očekujemo realizacije od Z' bliske 0 ili manje od 0. Veći iznosi realizacija od Z' idu u prilog alternativnoj hipotezi, ali to ne znači da se ne mogu dogoditi i ako je H_0 istinita hipoteza. Naime, ako je H_0 istinita hipoteza, vjerojatnost da se slučajna varijabla Z' realizira brojem većim ili jednakim $\hat{z}$ iznosi približno $P\{Z \geq \hat{z}\}$.

Sada zaključujemo na sljedeći način. Ako je $\mathcal{H}_0$ istinita hipoteza, realizacije veće ili jednake $\hat{z}$ mogu se pojaviti, a vjerojatnost je za to približno $P\{Z \geq \hat{z}\}$. Dakle, ako odbacimo nul-hipotezu, vjerojatnost je da ćemo time pogriješiti najviše $P\{Z \geq \hat{z}\}$. Ukoliko je taj broj manji od standardno prihvaćenih vrijednosti za maksimalnu vjerojatnost pogreške prvog tipa (tj. nivoa značajnosti testa), hipotezu $\mathcal{H}_0$ možemo odbaciti. U suprotnom kažemo da nemamo dovoljno argumenata za odbacivanje hipoteze $\mathcal{H}_0$.

Na sličan bismo način bismo proveli postupak testiranja u slučaju da je alternativna hipoteza oblika $\mathcal{H}_1 : \mu < \mu_0$. Tada vjerojatnost $p = P_{\mu_0}\{Z' \leq \hat{z}\} \approx P\{Z \leq \hat{z}\}$, $Z \sim \mathcal{N}(0, 1)$, uspoređujemo s razinom značajnosti α, koja je u tom slučaju površina ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $(-\infty, z_\alpha]$ (slika 4.22).



Slika 4.22: Površina ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $(-\infty, z_\alpha]$.

Objašnjeni postupak općenito zapisujemo na sljedeći način:

- Nul-hipoteza:
$$\mathcal{H}_0 : \ \mu = \mu_0,$$
- Test-statistika:
$$Z' = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$
Ovdje je n veličina uzorka, $\overline{X}_n$ aritmetička sredina uzorka, a σ poznata standardna devijacija.
- U uvjetima istinitosti nul-hipoteze očekujemo da je realizacija od Z' (označit ćemo je $\hat{z}$) blizu ili veća od 0 jer varijabla Z' ima približno standardnu normalnu distribuciju. Ako označimo $Z \sim \mathcal{N}(0,1)$, na osnovu realizacije $\hat{z}$ statistike Z' na podacima možemo odrediti takozvanu p-vrijednost kao:
$$p = P\{Z \geq \hat{z}\}, \text{ ako je alternativna hipoteza oblika } \mathcal{H}_1 : \mu > \mu_0,$$
$$p = P\{Z \leq \hat{z}\}, \text{ ako je alternativna hipoteza oblika } \mathcal{H}_1 : \mu < \mu_0.$$
- Tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α. U slučaju da je $p < \alpha$, odbacujemo nul-hipotezu na razini značajnosti α. Ako je $p > \alpha$, zaključujemo da nemamo dovoljno informacija koje bi poduprle odluku o odbacivanju nul-hipoteze.

Treba napomenuti da je pretpostavka o poznatoj varijanci iz te procedure nerealna za primjene. S obzirom da je taj rezultat izveden za velike uzorke, može se također

dokazati da će korištenje korigirane varijance uzorka umjesto varijance u izvodu formule za test-statistiku dati analognu asimptotsku distribuciju. Iz tog se razloga u praksi koristi procjena za varijancu umjesto stvarne vrijednosti varijance ako je ona nepoznata.

Primjer 4.55. (kolokvij.xls)
Promotrimo ponovno diskretnu slučajnu varijablu X kojom modeliramo ukupan broj bodova koje je student ostvario na dvama kolokvijima iz nekog kolegija. Iz primjera 4.38 znamo da je procjena njezinog očekivanja realan broj 96.9. *Korištenjem istih podataka testirajmo hipotezu da je očekivani ukupni broj bodova ostvaren na tim dvama kolokvijima manji od* 100. *U tu svrhu postavimo statističke hipoteze na sljedeći način:*

$$\mathcal{H}_0: \ \mu = 100,$$
$$\mathcal{H}_1: \ \mu < 100.$$

Da bismo donijeli odluku na nivou značajnosti $\alpha = 0.05$, računamo vrijednost $\hat{z}$ test-statistike Z' temeljenu na podacima iz varijable bodovi_ukupno*:*

$$\hat{z} = \frac{\overline{x}_{70} - \mu_0}{s_{70}/\sqrt{70}} = \frac{96.9 - 100}{44.19/\sqrt{70}} = -0.59.$$

Odavde upotrebom kalkulatora vjerojatnosti slijedi da je pripadna p-vrijednost

$$p = P\{Z < \hat{z}\} = 0.28,$$

što je veće od zadane razine značajnosti $\alpha = 0.05$. Zaključujemo da, na razini značajnosti $\alpha =$ 0.05, ne možemo tvrditi da je očekivani ukupni broj bodova ostvaren na tim dvama kolokvijima manji od 100.

Primjer 4.56. (djelatnici.xls)
Neka je X neprekidna slučajna varijabla iz primjera 4.39 kojom modeliramo godišnju plaću djelatnika jedne tvornice. U primjeru 4.39 pokazali smo da je $\overline{x}_{100} = 24522$ procjena njezinog očekivanja, a u primjeru 4.41 da je $s^2_{100} = 26069208.1$ procjena varijance na temelju odabranog uzorka.
Želimo provjeriti je li očekivanje od X veé od 23000 *na razini značajnosti $\alpha = 0.05$. U tu svrhu postavljamo statističke hipoteze:*

$$\mathcal{H}_0: \ \mu = 23000,$$
$$\mathcal{H}_1: \ \mu > 23000.$$

Da bismo donijeli odluku, računamo vrijednost $\hat{z}$ test-statistike Z' na temelju podataka za varijablu placa_prije.

$$\hat{z} = \frac{\overline{x}_{100} - \mu_0}{s_{100}/10} = \frac{24522 - 23000}{\sqrt{26069208.1}/10} \approx 2.98.$$

Odavde upotrebom kalkulatora vjerojatnosti slijedi da je pripadna p-vrijednost

$$p = P\{Z > \hat{z}\} = 0.0014,$$

što je manje od zadane razine značajnosti $\alpha = 0.05$. Stoga zaključujemo da, na nivou značajnosti $\alpha = 0.05$, prihvaćamo alternativnu hipotezu, tj. tvrdimo da je očekivanje godišnje plaće djelatnika u promatranoj tvornici veće od 23000.

4.5.3 Testiranje hipoteze o proporciji za velike uzorke

O proporciji ćemo, u ovom primjeru statističkog testa, zaključivati na temelju n-dimenzionalnoga jednostavnog slučajnog uzorka iz Bernoullijeve distribucije. Neka je slučajni pokus modeliran Bernoullijevom slučajnom varijablom s tablicom distribucije

$$X = \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, \quad p \in (0,1), \quad q = 1 - p.$$

Testirat ćemo hipotezu o vrijednosti parametra p koji ima značenje vjerojatnosti realizacije "uspjeha" (tj. jedinice) u jednom izvođenju danog pokusa. U tom postupku koristimo relativnu frekvenciju realiziranih "uspjeha" (jedinica) kao procjenitelja za vjerojatnost p i označavamo ju s $\widehat{p}$.

Postavimo nultu hipotezu na sljedeći način:

$$\mathcal{H}_0 : \ p = p_0.$$

Iskoristimo test-statistiku

$$Z' = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

U uvjetima istinitosti nul-hipoteze centralni granični terem garantira da slučajna varijabla Z' ima približno standardnu normalnu distribuciju za velike n. Ako je $Z \sim \mathcal{N}(0,1)$, na osnovu realizacije $\hat{z}$ na našem uzorku možemo odrediti p-vrijednost kao:

- $p = P\{Z \geq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : p > p_0$,
- $p = P\{Z \leq \hat{z}\}$ ako je alternativna hipoteza oblika $\mathcal{H}_1 : p < p_0$.

Tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α. U slučaju da je $p < \alpha$, na razini značajnosti α odbacujemo nul-hipotezu $\mathcal{H}_0$. Ako je $p > \alpha$, nemamo dovoljno informacija koje bi poduprle odluku o odbacivanju nul-hipoteze.

Veličina uzorka za provođenje tog testa treba biti najmanje takva da interval

$$\left[p_0 - 3\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + 3\sqrt{\frac{p_0(1-p_0)}{n}} \right]$$

ne sadrži ni 0 ni 1.

Primjer 4.57. *Iskoristimo podatke o ocjeni studenata na usmenom dijelu ispita izabranog kolegija iz primjera 4.37. Zanima nas je li proporcija studenata koji taj usmeni ispit polažu ocjenom izvrstan manja od* 0.2. *Iz tablice 4.14 znamo da je relativna frekvencija studenata koji su ispit položili izvrsnom ocjenom* $\hat{p} = 0.16$. *Postavljamo statističke hipoteze:*

$$\mathcal{H}_0 : p = 0.2,$$
$$\mathcal{H}_1 : p < 0.2.$$

Da bismo donijeli odluku, računamo vrijednost $\hat{z}$ *test-statistike* Z'*:*

$$\hat{z} = \frac{0.16 - 0.2}{\sqrt{\frac{0.16(1-0.16)}{65}}} = -0.88.$$

Odavde slijedi da je pripadna p-vrijednost

$$p = P\{Z < \hat{z}\} = 0.19,$$

što je veće od zadane razine značajnosti $\alpha = 0.05$. *Zaključujemo da na razini značajnosti* $\alpha = 0.05$ *ne možemo poduprijeti hipotezu da je proporcija studenata koji taj usmeni ispit polažu ocjenom izvrstan manja od 20%.*

Primjer 4.58. (kolokvij.xls)
Vratimo se bazi podataka kolokvij.xls *i testirajmo je li proporcija studenata koji su na kolokvijima ostvarili ukupno barem* 100 *bodova veća od* 50% *na razini značajnosti* $\alpha = 0.05$. *Iz primjera 4.49 znamo da je relativna frekvencija takvih studenata* $\hat{p} = 0.56$, *pa su statističke hipoteze oblika*

$$\mathcal{H}_0 : p = 0.5,$$
$$\mathcal{H}_1 : p > 0.5.$$

Računamo vrijednost $\hat{z}$ *test-statistike* Z'*:*

$$\hat{z} = \frac{0.56 - 0.5}{\sqrt{\frac{0.25}{70}}} = 1.004.$$

Odavde slijedi da je pripadna p-vrijednost

$$p = P\{Z > \hat{z}\} = 0.16$$

što je veće od zadane razine značajnosti $\alpha = 0.05$. *Stoga zaključujemo da ne odbacujemo nul-hipotezu na razini značajnosti* $\alpha = 0.05$, *tj. na nivou značajnosti* $\alpha = 0.05$ *ne možemo tvrditi da je više od* 50% *studenata ostvarilo barem* 100 *bodova na kolokvijima.*

Primjer 4.59. (djelatnici.xls)
Vratimo se primjeru 4.39 i testirajmo je li proporcija zaposlenika koji imaju godišnju plaću manju od 30000 *veća od 75% na razini značajnosti* $\alpha = 0.05$. *Iz tablice frekvencija za varijablu* placa_prije *vidimo da je relativna frekvencija takvih djelatnika* $\hat{p} = 0.86$, *pa su statističke hipoteze oblika*

$$\mathcal{H}_0 : p = 0.75,$$
$$\mathcal{H}_1 : p > 0.75.$$

Računamo vrijednost $\hat{z}$ *test-statistike* Z'*:*

$$\hat{z} = \frac{0.86 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{100}}} = 2.54.$$

Odavde slijedi da je pripadna p-vrijednost

$$p = P\{Z > \hat{z}\} = 0.0055,$$

što je manje od zadane razine značajnosti $\alpha = 0.05$. *Stoga zaključujemo da odbacujemo nul-hipotezu na razini značajnosti* $\alpha = 0.05$, *tj. na razini značajnosti* $\alpha = 0.05$ *možemo tvrditi da više od 75% djelatnika te tvornice ostvaruje godišnju plaću manju od* 30000.

4.5.4 Testiranje hipoteze o jednakosti očekivanja

Do sada prezentirani testovi kreirani su na temelju najjednostavnijega statističkog modela, tj. modela jednostavnog slučajnog uzorka iz zadane jednodimenzionalne distribucije. U ovom poglavlju navodimo dva testa koja se na prvi pogled ne uklapaju u takav model, ali se, prikladnim pristupom problemu, mogu svesti na njega.

Zavisni uzorci

Za ilustraciju problema navodimo prvo primjer.

Primjer 4.60. *Farmaceutska tvrtka želi testirati učinkovitost novog lijeka na smanjenje razine triglicerida u krvi. U tu svrhu odabire osobe u reprezentativnu skupinu za testiranje i izvodi mjerenje razine triglicerida prije uzimaja lijeka. Iste osobe koriste lijek u zadano vrijeme. Nakon toga se ponovo mjeri razina triglicerida. Dobiveni podaci prikazani su sljedećom tablicom:*

pacijent	prije	poslije
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

Istraživači bi htjeli dobiti odgovor na pitanje dolazi li do promjene razine triglicerida nakon uzimanja lijeka.

Analogni problemi pojavljuju se često u primjenama. Kažemo da u takvim slučajevima analiziramo istu karakteristiku (varijablu) prije i nakon tretmana.

Statistički model za zaključivanje možemo postaviti korištenjem slučajnog vektora tako da parove (x_i, y_i), $i = 1, \ldots, n$, smatramo međusobno nezavisnim realizacijama slučajnog vektora (X, Y), gdje komponenta X opisuje varijablu na populaciji prije tretmana, a Y varijablu na populaciji nakon tretmana. Tako nastao slučajni uzorak $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ niz je nezavisnih jednako distribuiranih slučajnih vekotra čije komponente X_i i Y_i, $i \in \{1, \ldots, n\}$, ne moraju biti nezavisne, što komplicira statistički model. Međutim, problem analize razlika u distribuciji varijable prije i poslije tretmana može se postaviti i tako da formiramo slučajnu varijablu razlike:

$$D = X - Y.$$

Ako je očekivanje slučajne varijable D različito od nule, to je svakako dokaz da su razlike u distribuciji između X i Y bitne. U tu svrhu proširimo tablicu podataka novom varijablom "razlike" i promatramo statistički model jednostavnoga slučajnog uzorka iz distribucije slučajne varijable D.

br	tretman 1	tretman 2	razlike
1	x_1	y_1	$d_1 = x_1 - y_1$
2	x_2	y_2	$d_2 = x_1 - y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$d_n = x_1 - y_n$

Postavimo hipotezu:

$$\mathcal{H}_0 : \mu_D = 0.$$

Ukoliko je uzorak dovoljno velik, u poglavlju 4.5.2 opisali smo proceduru za testiranje te hipoteze u kombinaciji s alternativnom hipotezom $\mu_D > 0$ ili $\mu_D < 0$. U tom slučaju za varijancu od D možemo koristiti korigiranu varijancu uzorka iz distribucije od D, kao što je navedeno u poglavlju 4.5.2.

Primjer 4.61. (djelatnici.xls)
U bazi podataka djelatnici.xls *raspolažemo podacima o godišnjim plaćama za uzorak od* 100 *djelatnika jedne tvornice (tvornice A). Varijabla* placa_prije *sadrži iznose godišnjih plaća za djelatnike iz uzorka prije reorganizacije poslovnog sustava, a varijabla* placa_poslije *iznose godišnjih plaća za isti uzorak djelatnika nakon reorganizacije. Budući da se ovdje radi o istom uzorku djelatnika čije godišnje plaće pratimo prije i poslije reorganizacije posla, kažemo da se radi o zavisnim uzorcima. Označimo s $X^{(1)}$ slučajnu varijablu kojom modeliramo godišnju plaću djelatnika tvornice A prije reorganizacije, a s $X^{(2)}$ slučajnu varijablu kojom modeliramo godišnju plaću djelatnika te tvornice poslije reorganizacije. Zanima nas razlikuju li se na nivou značajnosti $\alpha = 0.05$ očekivanja plaća djelatnika te tvornice prije i poslije reorganizacije posla. Prije nego postavimo potrebne statističke hipoteze, procijenimo, na temelju realizacija jednostavnih slučajnih uzoraka iz distribucija od $X^{(1)}$ i $X^{(2)}$, očekivanja i varijance od $X^{(1)}$, $X^{(2)}$ i $D = X^{(1)} - X^{(2)}$. Procjene su dane u tablici 4.17.*

slučajna varijabla	procjena očekivanja	procjena varijance
$X^{(1)}$	24522	26069208.1
$X^{(2)}$	24986.85	26252789.5
D	-464.85	116302.63

Tablica 4.17: Procjene očekivanja i varijance slučajnih varijabli $X^{(1)}$ i $X^{(2)}$.

Postavimo statističke hipoteze:

$$\mathcal{H}_0\text{: } \mu_D = 0,$$
$$\mathcal{H}_1\text{: } \mu_D < 0.$$

Da bismo donijeli odluku o tome odbacujemo li na nivou značajnosti $\alpha = 0.05$ nul-hipotezu ili ne, računamo vrijednost $\hat{z}$ test-statistike Z' temeljem procjena očekivanja i varijance slučajne varijable D i vrijednosti $\mu_0 = 0$:

$$\hat{z} = \frac{-464.85}{\sqrt{116302.63}/10} = -13.63.$$

Odavde, upotrebom kalkulatora vjerojatnosti, slijedi da je pripadna p-vrijednost

$$p = P\{Z < \hat{z}\} < 10^{-6},$$

što je manje od zadanog nivoa značajnosti $\alpha = 0.05$, stoga odbacujemo nul-hipotezu na nivou značajnosti $\alpha = 0.05$ i prihvaćamo alternativnu hipotezu, tj. tvrdimo da je očekivanje godišnje plaće djelatnika tvornice A prije reorganizacije bilo manje od očekivanja godišnje plaće djelatnika iste tvornce nakon reorganizacije.

Nezavisni uzorci

Za ilustraciju problema prvo navedimo jedan primjer.

Primjer 4.62. *Ured za kvalitetu na nekom fakultetu želi provjeriti je li došlo do bitne promjene u distribuciji ocjene iz kolegija Matematika 1 koju su ostvarili studenti generacije 2009./2010. u odnosu na generaciju 2008./2009. U tu svrhu prikuplja podatke o ocjenama na uzorcima studenata i modelira ocjenu iz tog kolegija kao slučajnu varijablu i to X za generaciju 2008./2009., a Y za 2009./2010. Također, pretpostavlja da su X i Y nezavisne slučajne varijable. Prikupljeni podaci o ocjenama $(x_1, \ldots, x_n)$ i $(y_1, \ldots, y_m)$ realizacije su jednostavnih slučajnih uzoraka iz distrbucije od X, odnosno iz distribucije od Y.*

Statistički model opisan u tom primjeru sastoji se od distribucija slučajnog vektora $(X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m)$ nezavisnih, ali ne nužno jednako distribuiranih slučajnih varijabli. Međutim, prvi dio vektora, tj. $(X_1, X_2, \ldots, X_n)$ jest jednostavni slučajni uzorak iz distribucije slučajne varijable X, dok je drugi dio, $(Y_1, Y_2, \ldots, Y_m)$, jednostavni slučajni uzorak iz distribucije slučajne varijable Y.

Jednostavan način potvrde postojanja razlike u distribucijama potvrda je postojanja razlike u očekivanjima tih distribucija. Ovdje ćemo navesti jedan test kojim možemo testirati upravo takve hipoteze, tj. hipoteze o razlici u očekivanjima.

U tu svrhu pretpostavimo da su veličine uzoraka (n i m) velike te uočimo da su $\overline{X}_n$ i $\overline{Y}_m$ dvije slučajne varijable s asimptotski normalnim distribucijama. Osim toga, $E(\overline{X}_n) = \mu_X$, dok je $E(\overline{Y}_m) = \mu_Y$, $\operatorname{Var}\overline{X}_n = \frac{\sigma_X^2}{n}$, $\operatorname{Var}\overline{Y}_m = \frac{\sigma_Y^2}{m}$, a $\overline{X}_n$ i $\overline{Y}_m$ nezavisne su slučajne varijable.

Neka je $D = \overline{X}_n - \overline{Y}_m$. Uočimo da je

$$E\,D = \mu_X - \mu_Y, \quad \operatorname{Var} D = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

Sada možemo testirati slutnju o postojanju razlika u očekivanjima korištenjem distribucije slučajne varijable D. Postavimo statističku hipotezu:

$$\mathcal{H}_0 : \mu_D = 0.$$

Statistička teorija dokazuje da, u uvjetima istinitosti $\mathcal{H}_0$, standardizirani oblik slučajne varijable D, tj. slučajna varijabla

$$Z' = \frac{D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

ima asimptotski standardnu normalnu distribuciju. Ukoliko varijance nisu unaprijed poznate (što je u praksi slučaj), moramo modificirati Z' korištenjem varijance uzorka kao procjenitelja za varijancu, pa koristimo statistiku

$$Z'' = \frac{D}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}},$$

koja također ima asimptotski standarnu normalnu distribuciju ako je $\mathcal{H}_0$ istinita hipoteza.

Korištenjem statistike Z'' testiramo hipotezu $\mathcal{H}_0$ u kombinaciji s alternativnom hipotezom $\mathcal{H}_1$:$\mu_D > 0$ ili $\mathcal{H}_1$:$\mu_D < 0$ analognim postupkom kao u poglavlju 4.5.2.

Primjer 4.63. (djelatnici.xls)
U bazi podataka djelatnici.xls *raspolažemo podacima o godišnjim plaćama za uzorke od po* 100 *djelatnika iz dviju konkurentskih tvornica - tvornice A (varijable* placa_prije *i* placa_poslije*, pogledati primjer 4.61) i tvornice B (varijabla* placa_konkurencija*). Budući da se radi o uzorcima djelatnika iz dviju različitih tvornica, zaključujemo da se radi o nezavisnim uzorcima.*
Označimo ponovno s $X^{(1)}$ *i* $X^{(2)}$ *slučajne varijable kojima modeliramo godišnje plaće djelatnika tvornice A prije i poslije reorganizacije posla, redom, te s* Y *slučajnu varijablu kojom modeliramo godišnju plaću djelatnika tvornice B. U varijablama* placa_prije, placa_poslije *i* placa_konkurencija *sadržane su realizacije jednostavnih slučajnih uzoraka* $(X_1^{(1)}, \ldots, X_{100}^{(1)})$, $(X_1^{(2)}, \ldots, X_{100}^{(2)})$ *i* $(Y_1, \ldots Y_{100})$ *iz distribucija slučajnih varijabli* $X^{(1)}$, $X^{(2)}$ *i* Y, *redom. Zanima nas razlikuju li se, na nivou značajnosti* $\alpha = 0.01$, *očekivanje plaće djelatnika tvornice A prije reorganizacije posla od očekivanja plaće djelatnika tvornice B. Potražimo također i odgovor na analogno pitanje za očekivanje plaće djelatnika tvornice A nakon reorganizacije.*
Procjene očekivanja i varijance slučajnih varijabli $X^{(1)}$ *i* $X^{(2)}$ *dane su u tablici 4.17. Procjene očekivanja i varijanci slučajnih varijabli* Y, $D_1 = \overline{X}_{100}^{(1)} - \overline{Y}_{100}$ *i* $D_2 = \overline{X}_{100}^{(2)} - \overline{Y}_{100}$ *dane su u tablici 4.18.*

slučajna varijabla	procjena očekivanja	procjena varijance
Y	25432.24	316895.23
D_1	-910.24	26386103.33
D_2	-445.39	26629684.73

Tablica 4.18: Procjene očekivanja i varijanci slučajnih varijabli Y, D_1 i D_2.

Postavimo statističke hipoteze:

$$\mathcal{H}_0\colon \mu_{D_i} = 0,$$
$$\mathcal{H}_1\colon \mu_{D_i} < 0,$$

gdje je $i \in \{1, 2\}$. *Da bismo donijeli odluku o tome odbacujemo li na nivou značajnosti* $\alpha = 0.01$ *nul-hipoteze ili ne, računamo vrijednosti* $\hat{z}_i$ *test-statistike* Z'' *na temelju procjena očekivanja i varijanci slučajnih varijabli* D_1 *i* D_2*:*

$$\hat{z}_1 = \frac{-910.24}{\sqrt{26386103.33}} = -1.77, \qquad \hat{z}_2 = \frac{-445.39}{\sqrt{26629684.73}} = -0.86.$$

Odavde, upotrebom kalkulatora vjerojatnosti, slijedi da su pripadne p-vrijednosti redom

$$p_1 = P\{Z < \hat{z}_1\} = 0.078, \qquad p_2 = P\{Z < \hat{z}_2\} = 0.39.$$

Budući da je p_2 *veći od* 0.01 *na nivou značajnosti* $\alpha = 0.01$*, ne možemo tvrditi da je očekivanje godišnje plaće djelatnika tvornice A nakon reorganizacije manje od očekivanja godišnje plaće djelatnika tvornice B. Osim toga,* p_1 *je također veći od 0.01, pa takvu tvrdnju ne možemo poduprijeti niti ako uspoređujemo godišnje plaće djelatnika tvornice A prije reorganizacije s godišnjim plaćama djelatnika tvornice B. Uočimo da* p_2 *ima veću "težinu" od* p_1 *u neodbacivanju hipoteze* $\mathcal{H}_0$.

4.6 Zadaci

Zadatak 4.2. (tlak.xls)
Baza podataka tlak.xls sadrži podatke o krvnom tlaku dobivene anketiranjem reprezentativnog uzorka pacijenata jedne liječničke ordinacije:

- varijable spol i dob sadrže informacije o spolu i broju godina za svakog ispitanika,
- varijable sistolicki-tlak i dijastolicki-tlak sadrže vrijednosti sistoličkog i dijastoličkog tlaka za svakog ispitanika,
- varijabla tlak klasificira vrijednosti sistoličkog i dijastoličkog tlaka u tri kategorije: N - nizak tlak, O - normalan tlak, P - povišen tlak,
- varijabla puls sadrži broj otkucaja srca u minuti (puls) za svakog ispitanika,
- varijabla opce-stanje sadrži subjektivnu ocjenu (u standardnoj skali od 1 do 5) vlastitog zdravstvenog stanja svakog ispitanika.

Na temelju podataka sadržanih u toj bazi riješite sljedeće zadatke.

a) Odredite tablice frekvencija i relativnih frekvencija, nacrtajte i proanalizirajte histograme frekvencija i relativnih frekvencija te kružni dijagram s prikazom relativnih frekvencija za podatke sadržane u varijabli opce-stanje. Kolike su frekvencija i relativna frekvencija ispitanika koji su svoje opće zdravstveno stanje ocijenili barem ocjenom 4?
Rješenje: 39 ispitanika, tj. njih je 78% svoje opće zdravstveno stanje ocijenilo barem ocjenom 4.

b) Odredite tablice frekvencija i relativnih frekvencija za podatke sadržane u varijabli opce-stanje posebno za kategoriju ispitanika ženskog spola i kategoriju ispitanika muškog spola te nacrtajte pripadne histograme frekvencija i relativnih frekvencija. Također, nacrtajte histograme frekvencija i relativnih frekvencija za podatke sadržane u varijabli opce-stanje kategorizirane po vrijednostima varijable tlak (N, O, P). Proanalizirajte dobivene histograme.

c) Odredite i ukratko protumačite sljedeće numeričke karakteristike podataka sadržanih u varijabli dob: aritmetičku sredinu, medijan, donji i gornji kvartil, mod, raspon i standardnu devijaciju. Je li mod jedinstven? Koliko iznosi maksimalno odstupanje podataka sadržanih u varijabli dob od njihove aritmetičke sredine? Nacrtajte i detaljno proanalizirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli dob. Obrazložite svoj odgovor.
Rješenje: $\overline{x}_{50} = 40.22$, $x_{min} = 11$, $x'_{25} = 23$, $x'_{50} = 43$, $x'_{75} = 51$, $x_{max} = 83$, $s_{50} = 17.69$.

d) Nacrtajte i detaljno proanalizirajte kutijasti dijagram na bazi medijana za podatke sadržane u varijabli dob. Obrazložite svoj odgovor.

e) Crtanjem i analizom kutijastog dijagrama na bazi medijana neosjetljivog na stršeće vrijednosti i kutijastog dijagrama na bazi medijana osjetljivog na stršeće vrijednosti donesite zaključak o tome pojavljuju li se među podacima sadržanima u varijabli puls stršeće vrijednosti ili ne. Ako ste se uvjerili u njihovo postojanje, korištenjem kategoriziranih tablica frekvencija odredite sve prisutne stršeće vrijednosti među podacima u varijabli puls. Kako biste neutralizirali njihov utjecaj na numeričke karakteristike podataka?
Rješenje: stršeće su vrijednosti varijable puls 800 i 1006.

Zadatak 4.3. (glukoza.xls)
Baza podataka (glukoza.xls) za reprezentativan uzorak pacijenata jedne internističke klinike sadrži informacije o dobi (varijabla dob), koncentraciji glukoze u krvi (varijabla koncentracija) i tome je li izmjerena koncentracija glukoze normalna ili povišena (varijabla kategorija: N - normalna koncentracije, P - povišena koncentracija). Riješite sljedeće zadatke i sva rješenja interpretirajte u kontekstu promatranog problema.

a) Procijenite proporciju pacijenta promatrane klinike koji su stariji od 30 godina te proporciju pacijenata starih barem 50 ali manje od 60 godina.
Rješenje: proporcija je pacijenta starijih od 30 godina $90/102 = 0.88$, a proporcija pacijenata starih barem 50, ali manje od 60 godina jest $21/102 = 0.21$.

b) Intervalom pozdanosti 95% procijenite proporciju pacijenata koji imaju povišenu koncentraciju glukoze u krvi.
Rješenje: $[0.61, 0.79]$.

c) Procijenite očekivanje, varijancu i standardnu devijaciju slučajne varijable kojom modeliramo koncentraciju glukoze u krvi pacijenata.
Rješenje: $\overline{x}_{102} = 7.69$, $s^2_{102} = 0.75$, $s_{102} = 2.78$.

d) Intervalom pozdanosti 95% procijenite očekivanu koncentraciju glukoze u krvi pacijenata.
Rješenje: $[7.15, 8.24]$.

e) Možete li na nivou značajnosti $\alpha = 0.01$ tvrditi da je proporcija pacijenata koji imaju povišenu koncentraciju glukoze manja od 0.8? Koji ste test koristili i zašto?
Rješenje: $p = 0.0044$.

f) Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da je očekivana koncentracija glukoze u krvi veća od normalne, tj. 6 mMol/L? Koji ste test koristili i zašto?
Rješenje: $p = 0$.

g) Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da je očekivana dob pacijenata koji imaju povišenu koncentraciju glukoze u krvi veća od očekivane dobi onih koji imaju normalnu koncentraciju glukoze? Koji ste test koristili i zašto?
Rješenje: $p = 0.59$.

Zadatak 4.4. (zdravlje.xls)
Baza podataka zdravlje.xls sadrži neke podatke o zdravstvenom stanju reprezentativnog uzorka stanovnika jednog mjesta:

varijable godine i spol sadrže podatke o starosti u godinama i spolu ispitanika;

varijabla zdravlje sadrži subjektivne ocjene vlastitog zdravstvenog stanja ispitanika;

varijabla broj-pregleda sadrži informacije o ukupnom broju zdravstvenih pregleda svakog ispitanika u tekućoj kalendarskoj godini;

varijabla dodatno-zdravstveno sadrži podatke o dodatnom zdravstvenom osiguranju svakog ispitanika (1 - ispitanik je dodatno osiguran; 0 - ispitanik nije dodatno osiguran);

varijabla cijena sadrži cijenu u kunama najskupljeg zdravstvenog pregleda svakog ispitanika (u tekućoj kalendarskoj godini).

Riješite sljedeće zadatke i sva rješenje interpretirajte u kontekstu promatranog problema.

a) Procijenite proporciju stanovnika promatranog mjesta koji su svoje zdravstveno stanje ocijenili ocjenom većom od dva, ali manjom of pet.
Rješenje: proporcija stanovnika koji su svoje zdravstveno stanje ocijenili ocjenom većom od dva, ali manjom od pet $30/51 = 0.59$.

b) Intervalom pozdanosti 97% procijenite proporciju stanovnika koji nemaju dodatno zdravstveno osiguranje.
Rješenje: $[0.59, 0.86]$.

c) Procijenite očekivanje, varijancu i standardnu devijaciju slučajne varijable kojom modeliramo subjektivnu ocjenu zdravstvenog stanja stanovnika.
Rješenje: $\overline{x}_{51} = 3.27$, $s^2_{51} = 1.36$, $s_{51} = 1.17$.

d) Intervalom pozdanosti 95% procijenite očekivanu ocjenu zdravstvenog stanja stanovnika.
Rješenje: $[2.95, 3.6]$.

e) Možete li na razini značajnosti $\alpha = 0.01$ tvrditi da je proporcija stanovnika koji imaju dopunsko zdravstveno osiguranje manja od 0.15? Koji ste test koristili i zašto?
Rješenje: $p = 0.421$.

g) Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da je očekivana dob stanovnika koji nemaju dopunsko zdravstveno osiguranje manja od očekivane dobi onih koji imaju dopunsko zdravstveno osiguranje? Koji ste test koristili i zašto? Što zaključujete ako za razinu značajnosti testa uzmete $\alpha = 0.01$ i zašto?
Rješenje: $p = 0.03$.

Zadatak 4.5. (matematika.xls)
Baza podataka (matematika.xls) sadrži podatke prikupljene anketiranjem reprezentativnog uzorka studenata jedne generacije nakon održanih predavanja, vježbi, kolokvija te usmenog ispita iz jednog matematičkog kolegija. Prikupljeni podaci organizirani su na sljedeći način:

varijabla prosjek sadrži podatke o prosječnoj ocjeni na studiju za svakog od anketiranih studenata,

varijabla polozeno za svakog anketiranog studenta sadrži informaciju o tome je li položio usmeni ispit iz promatranog kolegija (oznaka 1) ili nije (oznaka 0),

varijable predavanja i vjezbe sadrže informaciju o redovitosti pohađanja nastave iz promatranog kolegija (oznaka 1 - student s p/v nije nikada izostao, oznaka 2 - student je s p/v izostao samo jednom, oznaka 3 - student je s p/v izostao barem dva puta),

varijable tezina_kolegija i materijali sadrže subjektivne ocjene (u standardnoj skali od 1 do 5) promatranih studenata za težinu kolegija i dostatnost dostupnih materijala za pripremanje ispita iz promatranog kolegija.

Riješite sljedeće zadatke i sva rješenja interpretirajte u kontekstu promatranog problema.

a) Procijenite proporciju studenata čija je prosječna ocjena na studiju veća od tri.
Rješenje: proporcija studenata čija je prosječna ocjena studiranja veća od tri jest $42/49 = 0.86$.

b) Intervalom pozdanosti 95% procijenite proporciju studenata čija je prosječna ocjena na studiju najviše tri.
Rješenje: $[0.045, 0.24]$.

c) Procijenite očekivanje, varijancu i standardnu devijaciju slučajne varijable kojom modeliramo prosječnu ocjenu na studiju kojeg pohađaju ispitanici.
Rješenje: $\overline{x}_{49} = 3.98$, $s^2_{49} = 0.565$, $s_{49} = 0.752$.

d) Intervalom pozdanosti 95% procijenite očekivanu prosječnu ocjenu na studiju kojeg pohađaju ispitanici.
Rješenje: $[3.76, 4.19]$.

e) Možete li na razini značajnosti $\alpha = 0.01$ tvrditi da je proporcija studenata koji su redovito pohađali predavanja veća od 0.7? Koji ste test koristili i zašto?
Rješenje: $p = 0.413$.

Zadatak 4.6. (komarci.xls)
Baza podataka komarci.xls sadrži dio rezultata proučavanja komaraca u jednom močvarnom području (dostupni su podaci za 210 mjerenja na istoj lokaciji):

varijable brojM i brojZ redom sadrže broj muških i ženskih jedinki komaraca uhvaćenih u klopku;

varijabla mjesec sadrži mjesečevu mijenu (M - mlađak, U - uštap) za svako mjerenje;

varijabla doba-dana sadrži doba dana u kojem je mjerenje obavljeno (P - predvečerje, N - noć, S - svitanje);

varijabla svjetlost sadrži tip osvjetljenja pri mjerenju;

varijabla temperatura sadrži temperaturu zraka pri kojoj je mjerenje izvršeno;

varijabla rel-vlaznost sadrži relativnu vlažnost zraka za vrijeme mjerenja.

Riješite sljedeće zadatke i sva rješenja interpretirajte u kontekstu promatranog problema.

a) Procijenite proporciju mjerenja u kojima je izbrojeno manje od 50 muških jedinki komaraca te proporciju mjerenja u kojima je izbrojeno više od 50 ženskih jedinki komaraca.
Rješenje: proporcija mjerenja u kojima je izbrojeno manje od 50 muških jedinki komaraca jest $202/210 = 0.962$, a proporcija je mjerenja u kojima je izbrojeno više od 50 ženskih jedinki komaraca je $39/210 = 0.186$.

b) Intervalom pozdanosti 95% procijenite proporciju mjerenja u kojima je izbrojeno barem 50, ali manje od 100 ženskih jedinki komaraca.
Rješenje: $[0.022, 0.083]$.

c) Procijenite očekivanje, varijancu i standardnu devijaciju slučajne varijable kojom modeliramo temperaturu zraka tijekom jednog mjerenja.
Rješenje: $\overline{x}_{210} = 21.89$, $s^2_{210} = 6.76$, $s_{210} = 2.6$.

d) Intervalom pozdanosti 97% procijenite očekivanu temperaturu tijekom jednog mjerenja.
Rješenje: $[21.51, 22.29]$.

e) Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da je proporcija mjerenja u kojima je izbrojeno manje od 50 muških jedinki komaraca veća od 0.95? Koji ste test koristili i zašto?
Rješenje: $p = 0.2$.

f) Možete li na nivou značajnosti $\alpha = 0.01$ tvrditi da je očekivani broj muških jedinki komaraca izbrojenih u mjerenju manji od 20? Koji ste test koristili i zašto?
Rješenje: $p = 0.99$.

g) Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da se očekivani broj muških jedinki komaraca izbrojenih u jednom mjerenju razlikuje od očekivanog broja ženskih jedinki izbrojenih u jednom mjerenju? Koji ste test koristili i zašto?
Rješenje: $p = 0.009$.

Zadatak 4.7. (gradjevina.xls)
Baza podataka gradjevina.xls sadrži neke informacije o reprezentativnom uzorku koji se sastoji od 100 srednje velikih građevinskih poduzeća u jednoj tranzicijskoj zemlji:

varijabla godina_osnivanja za svako od 100 poduzeća iz uzorka sadrži godinu kada je poduzeće osnovano,

varijable zaposleni2007, zaposleni2008 i zaposleni2009 sadrže podatke o broju zaposlenika u tih 100 poduzeća u 2007., 2008. i 2009. godini,

varijabla prosjecna_starost sadrži prosjecnu dob zaposlenika u 2009. godini,

varijable motivacija_placa i napredovanje redom sadrže subjektivne ocjene kadrovske službe poduzeća o tome u kolikoj je mjeri visina plaće motivacijski faktor za uspješno obavljanje posla te u kolikoj mjeri uspješno obavljanje poslovnih zadataka utječe na mogućnost napredovanja na bolje radno mjesto unutar poduzeća,

varijable placa2007, placa2008 i placa2009 sadrže iznose prosječnih plaća zaposlenika u 2007., 2008. i 2009. godini,

varijable troskovi2007, troskovi2008 i troskovi2009 sadrže iznose troškova poduzeća u 2007., 2008. i 2009. godini,

varijable prihodi2007, prihodi2008 i prihodi2009 sadrže iznose ostvarenih prihoda u 2007., 2008. i 2009. godini,

Riješite sljedeće zadatke i sva rješenja interpretirajte u kontekstu promatranog problema.

a) Odredite tipove i pripadne skupove vrijednosti slučajnih varijabli kojima modeliramo broj zaposlenika, njihovu prosječnu dob, prosječnu godišnju plaću zaposlenika u poduzeću, godišnje troškove poduzeća te godišnje prihode poduzeća.

b) Procijenite vjerojatnost da slučajno odabrano srednje veliko građevinsko poduzeće u promatranoj zemlji ima više od 50 zaposlenika u 2007., 2008. te 2009. godini?
Rješenje: proporcije su 0.83 za 2007., 0.93 za 2008. te 0.95 za 2009. godinu.

c) Procijenite očekivanje i varijancu slučajne varijable kojom se modelira prosječna plaća zaposlenika u srednje velikom građevinskom poduzeću u toj zemlji u 2009. godini.
Rješenje: $\overline{x}_{100} = 600.13$, $s_{100} = 194.63$.

d) Kategorizirajte podatke kojima raspolažete te odlučite ima li smisla modelirati prosječnu godišnju plaću u 2009. godini kao normalnu slučajnu varijablu. Ako smatrate da ima, korištenjem normalne distribucije s procijenjenim vrijednostima očekivanja i varijance odredite vjerojatnost da je u 2009. godini u slučajno odabranom poduzeću srednje veličine u toj zemlji prosječna plaća bila viša od 500 eura. Istu vjerojatnost izračunajte i korištenjem procijenjene (empirijske) distribucije te slučajne varijable te usporedite rezultate.
Rješenje: Iz histograma relativnih frekvencija vidimo da normalna distribucija nije prikladna za modeliranje tih podataka, a to sugeriraju i izračunate tražene vjerojatnosti: iz empirijske distribucije te slučajne varijable, označimo ju s X, slijedi da je $P(X > 500) = 0.66$, a ako X modeliramo kao $\mathcal{N}(600.13, 194.63^2)$, slijedi da je $P\{X > 500\} = 0.3$.)

e) Intervalom pouzdanosti 95 % procijenite proporciju srednje velikih građevinskih poduzeća u toj zemlji u kojima je prosječna mjesečna plaća veća od aritmetičke sredine plaća zabilježenih u varijabli placa2009.
Rješenje: $[0.343, 0.537]$.

f) Procijenite očekivanje, varijancu i standardnu devijaciju slučajne varijable kojom modeliramo prosječnu plaću u 2009. godini.
Rješenje: $\overline{x}_{100} = 600.13$, $s^2_{100} = 37879.1$, $s_{100} = 194.63$.

g) Intervalom pouzdanosti 95 % procijenite očekivanje slučajne varijable kojom se modelira prosječna mjesečna plaća zaposlenika u 2009. godini.
Rješenje: $[561.51, 638.75]$.

h) Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da je proporcija zaposlenika koji u 2009. godini imaju plaću višu od očekivane manja od 0.5? Koji ste test koristili i zašto?
Rješenje: $p = 0.12$.

i) Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da je očekivana plaća u 2009. godini veća od 650 eura? Koji ste test koristili i zašto?
Rješenje: $p = 0.012$.

j) Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da postoji razlika u očekivanoj prosječnoj plaći u građevinskim poduzećima srednje veličine u toj zemlji u 2008. i 2009. godini pod pretpostavkom da razlike prosječnih plaća u 2008. i 2009. godini možemo modelirati normalnom slučajnom varijablom? Koji ste test koristili i zašto?
Rješenje: $p = 0.164$.

k) Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da je proporcija srednje velikih građevinskih poduzeća u toj zemlji, koja imaju više od 150 zaposlenih, veća za 2009. nego za 2008. godinu? Koji ste test koristili i zašto?
Rješenje: $p = 0.4245$.