Filozofski fakultet u Sarajevu

Statistika u psihologiji

Priručnik za studente

Nermin Đapo i Ratko Đokić

Sarajevo, 2012

Nermin Đapo

Ratko Đokić

STATISTIKA U PSIHOLOGIJI, PRIRUČNIK ZA STUDENTE

Urednik:

Prof. dr. Ivo Komšić

Recenzenti:

Prof. dr. Valentin Bucik

Doc. dr. Dženana Husremović

Izdanje:

Prvo

Izdavač:

Filozofski fakultet u Sarajevu

Sarajevo, 2012

Elektronsko izdanje

CIP - Katalogizacija u publikaciji
Nacionalna i univerzitetska biblioteka
Bosne i Hercegovine, Sarajevo

159.9:519.2(075.8)(076)

ĐAPO, Nermin
Statistika u psihologiji : priručnik za studente [Elektronski izvor] / Nermin Đapo, Ratko Đokić. - Sarajevo : Filozofski fakultet, 2012. - 1 elektronski optički disk (CD-ROM) : tekst, slike ; 12 cm

Nasl. s naslovnog ekrana.

ISBN 978-9958-625-22-0
1. Đokić, Ratko
COBISS.BH-ID 19349510

Nermin Đapo i Ratko Đokić

Statistika u psihologiji

Priručnik za studente

Sarajevo, 2012

Predgovor

Statistika u psihologiji priručnik je namijenjen prvenstveno studentima koji započinju izučavati metodologiju psiholoških istraživanja. Premda se Statistika u nastavnim programima psihologije izučava kao zaseban predmet, statističke metode zapravo su sastavni dio istraživačkog procesa, što se u pisanju ovog priručnika nastojalo posebno naglasiti. Priručnik je pripremljen s osnovnim ciljem da studentima pruži sažeta objašnjenje osnovnih statističkih pojmova, čije poznavanje je preduslov za naprednije nastavne programe iz Statistike, kao i naučno-istraživačku praksu. Pri tome, koristili smo pristup uobičajen u edukaciji iz Statistike na studijama iz društvenih i humanističkih nauka, a kojim se definicije i objašnjenja koncepata daju bez komplikovanih matematičkih izraza i izvoda.

Priručnik se sastoji iz 11 poglavlja koja svojim sadržajem obuhvaćaju desktiptivnu i osnove inferencijalne statistike: Uvod, Grafičko i tabelarno predstavljanje podataka, Mjere centralne tendencije, Mjere varijabiliteta, Osnovni koncepti vjerovatnoće, Normalna raspodjela, Standardna pogreška aritmetičke sredine, Testiranje hipoteza, Testiranje razlika između dvije aritmetičke sredine, Analiza varijance, Korelacija i regresija. Svako poglavlje započinje uvodom u kojem su data osnovna i kratka objašnjenja, koja su kroz primjere u nastavku poglavlja dodatno interpretirana i u kojima je ukazano na njihovu praktičnu primjenu. Svako poglavlje završava zadacima koji služe za vježbanje i ponavljanje gradiva određene teme. Na kraju Priručnika data su rješenja nekih zadataka.

Nadamo se da će ovaj priručnik poslužiti ne samo studentima psihologije već i drugima koji se iz različitih razloga interesuju za Statistiku u psihologiji. Važnim nam se čini naglasiti da ovaj Priručnik ne može zamijeniti udžbenik iz Statistike. Njegovu svrhu treba prepoznati u samom nazivu, dakle, da bude pri ruci studentu koji izučava određene teme iz Statistike.

Na kraju, želimo se zahvaliti svima koji su na različite načine doprinijeli nastanku ovog teksta. U prvom redu, to su generacije studenata Odsjeka za psihologiju Filozofskog fakulteta u Sarajevu kojima je Priručnik i namjenjen. Njih smo najviše osluškivali dok smo radili na ovom tekstu. Statističkim rječnikom rečeno, udio njihove varijance u objašnjenju nastanka ovog Priručnika je najveći. Nadalje, zahvaljujemo se i kolegama sa Odsjeka na podršci da posao započet prije nekoliko godina dovedemo do kraja. Posebnu zahvalnost dugujemo Jadranki Kolenović-Đapo koja je prva pročitala rukopis, dala vrijedne savjete i sugestije i nesebično pomogla u tehničkom dijelu posla važnom za nastanak Priručnika. Zahvaljujemo se Marijani Galić, koja je s znatiželjom i strpljenjem pročitala rukopis te iz pozicije stručnjaka ekonomskih nauka dala korisne sugestije.

Nermin Đapo i Ratko Đokić

Sadržaj

1. Uvod

Pojam „statistika“ ima najmanje dva značenja. Prema jednom značenju, *statistika* označava numeričke vrijednosti kojima se opisuje skup podataka (npr. prosječni školski uspjeh učenika jedne škole, ili raspon rezultata koje ispitanici postižu na testu znanja iz nastavnog predmeta Statistika u psihologiji). Prema drugom značenju, *statistika* je grana matematike i predstavlja skup postupaka koje koristimo za prikupljanje, prezentaciju, analizu i interpretaciju podataka. Na studiju psihologije izučava se **primijenjena statistika**, odnosno metode za deskripciju i analizu podataka izvedene iz osnovnih matematičkih principa. Načini kojim se do njih došlo predmetom su izučavanja **teorijske statistike**.

Postupci koje koristimo u opisu podataka (npr. određivanje broj kategorija, centralne vrijednosti, aritmetičke sredine, itd.) dio su **deskriptivne statistike**, dok je donošenje zaključaka o populaciji na osnovu podataka dobivenih na uzorku dio **inferencijalne statistike**. Deskriptivne statističke postupke koristimo, npr., kada želimo opisati jednu ili više grupa ispitanika, a inferencijanu statistiku kako bi pokazali da li je razlika dobivena na ograničenom skupu podataka vjerovatna i na populacijama.

Statistika je sastavni dio istraživačkog procesa jer provedba istraživanja uključuje prikupljanje, obradu i analizu podataka. Stoga predmet *Statistika u psihologiji* pripada **metodološkoj grupi predmeta**, u koju ubrajamo i Metodologiju psiholoških istraživanja i Psihometriju.

Statistika je važna i u svakodnevnom životu. Poznavanje statističkih pojmova i koncepata, vještine i sposobnosti njihovog korištenja čine **statističku pismenost**. Obzirom da smo svakodnevno izloženi podacima na osnovu kojih donosimo odluke, statistička pismenost pomaže nam da bolje razumijemo svijet u kojem živimo. Ponekad podaci ne odražavaju realnost i mogu nas navesti na donošenje pogrešne odluke. Statistika nam pomaže da između nekoliko odluka izaberemo najbolju.

Varijabla, mjerenje, populacija, uzorak, parametar i statistik dio su osnovnog vokabulara statistike. Stoga ćemo u ovom poglavlju navest osnovne definicije ovih pojmova.

Varijabla (promjenljiva) je svojstvo pojave (osobe, objekta ili događaja) koje se mijenja. Nekim varijablama jednostavno klasificiramo pojave u jednu od grupa ili kategorija (npr. spol – osoba je muškog spola; boja kose – osoba ima kosu plave boje); ove varijable nazivamo **kategorijalnim varijablama**. Drugim varijablama određujemo količinu nečega, i nazivamo ih **kvantitativnim varijablama**. Kvantitativne varijable mogu biti **diskretne** (vrijednosti varijable mogu poprimiti samo neke vrijednosti iz određenog intervala, pri čemu su vrijednosti jasno odijeljene) i **kontinuirane** (teorijski, ove varijable mogu poprimiti bilo koju vrijednost iz određenog intervala; između pojedinih vrijednosti ne postoji skokovit prijelaz kao kod diskretnih varijabli). **Zavisne varijable** se mijenjaju zbog promjene drugih varijabli. One nisu pod kontrolom istraživača. Njihovu promjenu objašnjavamo, opisujemo ili klasificiramo pomoću drugih varijabli. Obično se označavaju slovom Y. **Nezavisne varijable** su pod kontrolom istraživača. Njima se objašanjavaju, opisuju ili klasificiraju promjene zavisne varijable. Obično se označavaju slovom X.

Prema Campbellu[1] *mjerenje* je pridruživanje numeričke vrijednosti pojavi koju posmatramo prema jasno utvrđenim pravilima. Mjerenje u psihologiji je indirektno jer ne postoji psihološki mjerni instrument čija je mjerna jedinica precizno definirana vrijednost mjerenog svojstva. Vrsta **skale mjerenja** određuje izbor numeričkih i grafičkih metoda za deskripciju i analizu podataka U psihologiji koristimo različite skale mjerenja: nominalna (kategorizacija), ordinalna (redoslijed + kategorizacija), intervalna (interval + redoslijed + kategorizacija) i racio (apsolutna nula + interval + redoslijed + kategorizacija).

Potpuni skup podataka (događaja) koji su predmet istraživanja naziva se *populacija*. Deskriptivne vrijednosti dobivene na populaciji nazivamo **parametrima**. Često je nemoguće podatke prikupiti iz populacije. Zapravo, za tako nešto često nema niti potrebe. Dovoljno je na pravilan način podatke prikupiti iz dijela populacije. Dio populacije (podskup populacije) koji koristimo za prikupljanje podataka koji će nam omogućiti zaključivanje o populaciji nazivamo *uzorak*. Deskriptivne vrijednosti dobivene na uzorku nazivamo **statisticima**.

Primjeri 1.1 i 1.2 ilustriraju neke od osnovnih statističkih koncepata koje smo naveli u ovom dijelu.

[1] Campbell, N. (1953). What is science. New York: Dover Publication.

PRIMJER 1.1

Pretpostavimo da je vrijeme izbora i da se ne možete odlučiti između dva predsjednička kandidata. Kriteriji kojeg koristite za izbor je prethodni angažman svakog od kandidata u rješavanju različitih problema zaštite okoline. Kandidat A navodi da je 30 puta glasao za različita zakonska rješenja koja se odnose na zaštitu okoline, dok kandidat B ističe da je glasao u 70% slučajeva za takva zakonska rješenja. Za koga biste glasali?

Za donošenje odluke potrebno je dodatno istraživanje koje uključuje prikupljanje više podataka nego što su ih ponudili kandidati. Kandidat A glasao je za 30 zakonskih rješenja, ali je prešutio da je za 70 glasao protiv. Premda je kandidat B glasao u 70% slučajeva, nije naveo da se u ostalih 30% slučajeva radilo o najvažnijim zakonima.

Na kraju, možemo zaključiti da niti jedan kandidat zapravo nije dobar izbor. Istraživanje i pravilno zaključivanje (drugim riječima statistika) pomoći će nam u donošenju dobrih odluka.

PRIMJER 1.2

Rezultati nekih istraživanja ukazuju da učenici nižeg socio-ekonomskog statusa postižu niži školski uspjeh. Učenici koji žive u ekonomski nepovoljnijim uslovima, imaju manje mogućnosti za kvalitetnije školovanje. Školski psiholog odlučio je istražiti povezanost između određenih socio-ekonomskih karakteristika i školskog uspjeha učenika srednjih škola.

Umjesto na populaciji koju čine svi učenici srednjih škola, istraživanje je provedeno na uzorku. Obzirom da istraživač namjerava zaključivati o populaciji na osnovu podataka prikupljenih na uzorku, nužno je na pravilan način formirati uzorak. U protivnom, mogućnost generalizacije rezultata istraživanja bit će ograničena. Uzorak je mogao biti formiran tako da svaki učenik iz populacije ima podjednake šanse da bude uključen u istraživanje. Međutim, istraživač se odlučio da metodom slučajnog odabira izabere određeni broj škola iz čitave države, i da u tako izabranim školama ispita sve učenike.

Istraživač je, između ostalih podataka, prikupio podatke o stepenu stručne spreme majke, radnom statusu majke i broju članova porodice, te o dobi i spolu učenika. U tabeli 1.2.1. prikazani su podaci za pet učenika. Spol, škola, stepen stručne spreme, radni status i broj članova domaćinstva su diskretne (diskontinuirane) varijable. Vrijednosti varijable „radni status“ jasno su odijeljene u tri kategorije: „zaposlen“, „nezaposlen“ i „penzioner“. Dob i prosječni školski uspjeh su kontinuirane varijable. Dob ispitanika može biti 16,7 (što odgovara uzrastu od 16 godina i 8 mjeseci), ali se može još preciznije izraziti vrijednošću 16,69. Mjerni (kvantitativni) podaci su dob, prosječni školski uspjeh, broj članova domaćinstva, dok su spol, škola, stručna sprema i radni status kategorijalni podaci. Vrijednosti kategorijalnih podataka izražavamo frekvencijama ili procentima za svaku kategoriju (npr. u istraživanju je učestovalo 52% djevojčica i 48% dječaka). Nominalne varijable su škola, spol i radni status. Dob je jedina varijabla mjerena racio skalom mjerenja. Školski uspjeh i broj članova domaćinstva predstavljaju varijable mjerene rang skalom mjerenja.

Tabela 1-1. Podaci za pet učenika

UČENIK	ŠKOLA	SPOL	DOB	PROSJEČNI ŠKOLSKI USPJEH	STEPEN STRUČNE SPREME	RADNI STATUS	BROJ ČLANOVA DOMAĆINSTVA
1	A	M	15,2	2,5	SSS	zaposlen	3
2	C	Ž	14,5	3,7	VSS	nezaposlena	4
50	C	Ž	15,0	4,0	OS	zaposlena	5
275	B	M	16,7	2,0	SSS	penzioner	4
350	D	Ž	17,1	4,7	SSS	nezaposlena	3

Zavisna varijabla u istraživanju je školski uspjeh. Istraživač na osnovu rezultata dobivenih u drugim istraživanjima, teorija i svakodnevnog iskustva smatra da se promjena varijable školski uspjeh može objasniti pomoću niza socio-ekonomskih varijabli, koje, stoga, predstavljaju nezavisne varijable.

ZADACI

1. U tabeli ispod prikazan je dio podataka dobivenih u hipotetičkom istraživanju. Koje varijable su kategorijalne, a koje kvantitativne? Koje su vrijednosti svake od varijabli?

PACIJENT	SPOL	DOB	DIJAGNOZA	VRSTA TERAPIJE	TRAJANJE TERAPIJE (U MJESECIMA)
AB	M	Srednja dob	Anksiozni poremećaj	Kbt	10
CR	Ž	Starija dob	Fobija	Kbt	4
NT	Ž	Starija dob	Anksiozni poremećaj	Geštalt	3
SQ	M	Mlađa dob	Depresivnost	Geštalt	9
TW	Ž	Starija dob	Depresivnost	Taa	15

2. U medicinskim istraživanjima koriste se različite varijable. Koje od dolje navedenih su kategorijalne, a koje kvantitativne?
   a. Spol (ženski, muški).
   b. Dob (godine i mjeseci).
   c. Pušenje (da ili ne).
   d. Sistolni krvni pritisak (mm živinog stuba).
   e. Puls (broj otkucaja u minuti).
   f. Količina šećera u krvi (mmol/l).
   g. Dijagnoza.
   h. Tjelesna temperatura (°C).
   i. Dužina terapije (u danima).
   j. Vrsta terapije.

3. U primjerima datim ispod: a) identificirajte varijable i b) za svaku varijablu odredite njen tip (kategorijalna ili kvantitativna).
   a. Za svako od 150 novorođenčadi utvrđeni su: spol, porođajna težina, datum i vrijeme rođenja.
   b. Stomatolog je za potrebe istraživanja izmjerio dužine (u mm) zadnjeg donjeg molara kod 10 pacijenata.
   c. Tokom kontrole 250 vozila policija je utvrdila da u 50 slučajeva vozač nije vezan sigurnosnim pojasom, u 10 slučajeva vozilo je bilo u tehnički neispravnom stanju, dok u 30 vozila nije nađena prva pomoć. Za sva vozila zabilježene su registarske tablice, brizna kojom se vozilo kretao prema kontrolnoj tački i spol vozača.
   d. Na sistematskom pregledu djeci je izmjerena visina i težina, urađen pregled pluća, abdomena, ekstremiteta, oftalmološki pregled, te laboratorijski nalazi (sedimentacija, broj eritrocita i broj leukocita). Djeca su na pregled došla u pratnji jednog ili oba roditelja.
   e. Prije nego što je započeo sa rezanjem, Edvin je pet puta izmjerio dimenzije obrisa kocke ucrtanih na kartonu.
   f. Biolog je prebrojao listove na svakoj od 20 biljaka.
   g. Emir je deset puta u toku dana mjerio krvni pritisak. Svaki put dobio je drugačije vrijednosti.

h. U Osnovnoj školi „Wilhelm Wundt” utvrđen je broj razreda, odjeljenja i učenika.
i. Psiholog je posmatrao desetero djece i mjerio vrijeme koje svako od njih provede igrajući se sa plišanim medvjedićem.
j. Deset atletičara trčalo je na 100 metara. Prvi na cilj došao je A.N, drugi N.Z, dok je zadnji na cilj došao W.P.

4. U dolje navedenim primjerima odredite nivo (skalu) mjerenja.
   a. Broj pogrešaka koje pacov načini prilikom prolaska kroz labirint.
   b. Spol.
   c. Dob.
   d. Udaljenost između Sarajeva i Konjica (u km).
   e. Vrijeme koje je potrebno motoru da postigne brzinu od 100 km/h.
   f. Stepen zadovoljstva životom koji ispitanici procjenjuju na skali od 7 jedinica (1 – uopšte nisam zadovoljan/a; 7 – ekstremno sam zadovoljan/a).
   g. Da li osoba koristi ili ne koristi zamjenice prvog lica (mene, meni, nas, nama) tokom 10-minutne konverzacije?
   h. Koliko puta u toku 30-minutne konverzacije osoba upotrijebi prvo lice jednine.
   i. Nacionalna pripadnost.
   j. Rezultati na testu depresije (maksimalni rezultat je 100).
   k. Rangiranje 5 tema u skladu sa stepenom u kojem je osoba upoznata sa temama (teme se rangiraju od “najbolje upoznat” do “najslabije upoznat”).
   l. Djetetov izbor “ružne igračke” ili “lijepe igračke”.
   m. Broj tačno riješenih zadataka na testu znanja iz fizike.
   n. Vrste psiholoških fobija.
   o. Tjelesna temperatura (izražena u °C).
   p. Samopoštovanje, izraženo na skali sa 5 stupnjeva.
   q. Godišnji prihod u KM.
   r. Teoretska psihoterapijska orijentacija.
   s. Osvojeno mjesto na takmičenju u latinoameričkom plesu.
   t. Srčani ritam izražen kao broj otkucaja u minuti.
   u. Broj eritrocita u krvi.
   v. Kućni brojevi u ulici Wilhelma Wundta (npr. ulica Wilhelma Wundta br. 25).
   w. Vrijeme reakcije (u sekundama).
   x. Prosječna školska ocjena iz fizike na kraju školske godine.
   y. Broj riječi zapamćenih na testu pamćenja.
   z. Temperatura tijela izražena u Kelvinima.
   aa. Vrsta kazne u kriminološkoj praksi (npr. uvjetna, mjera zatvora).
   bb. Vrijeme koje dijete provede u igranju sa svojim vršnjakom (u minutama).
   cc. Broj godina školovanja.
   dd. Bračni status ispitanika.

5. Odredite skale mjerenja u zadacima 1, 2 i 3.

6. Provjerite tačnost sljedećih tvrdnji:

a. Nominalna skala služi za klasificiranje pojava.	DA	NE
b. Intervalne skale imaju jednake intervale između mjernih jedinica (tačaka na skali).	DA	NE
c. Na rang skali vrijednosti su poredane od najmanje do najveće.	DA	NE
d. Na omjernim skalama, veći broj uvijek znači "više" pojave koja se mjeri.	DA	NE
e. Intervalne skale ne podrazumijevaju svojstvo "redoslijeda".	DA	NE
f. Na intervalnim skalama, rezultat nula znači odsustvo fenomena koji se mjeri.	DA	NE
g. Kada se koristi nominalna mjerna skala, najviše što se može reći je da je jedan rezultat veći od drugog.	DA	NE
h. Rang skalom ne možemo klasificirati pojave.	DA	NE
i. Na intervalnim skalama rezultat nula nije moguć.	DA	NE

7. Za sljedeće rezultate mjerenja identificirajte originalno korištenu skalu, a potom transformirajte rezultate sa skala viših nivoa mjerenja na rezultate skala nižih nivoa mjerenja:
   a. Ben je na ispitu iz Statistike dobio ocjenu 9,1; Den 7,5; Ken 9,9; Jen 8,7; Ren 5 (nije uspio položiti ispit); Wen 10.
      a.1. Transformirajte ove rezultate na nominalnu skalu.
      a.2. Transformirajte ove rezultate na rang skalu.
   b. Carl Lewis istrčao je dionicu od 100 m za 9,86 s; Tyson Gay za 9,69; Donovan Bailey za 9,84 s; Usain Bolt za 9,58; Leroy Burrell za 9,85 s; Maurice Green za 9,79 s; dok je Asaffa Powell odustao prije dolaska na cilj.
      b.1. Transformirajte ove rezultate na nominalnu skalu.
      b.2. Transformirajte ove rezultate na rang skalu.
   c. Zavod za socijalnu zaštitu utvrdio je sljedeće ekonomske razrede za rangiranje stanovništva: mjesečni prihodi od 0 do 370 KM spadaju u razred ispodprosječnih novčanih primanja. Primanja od 371 do 670 KM spadaju u prosječna primanja, dok se mjesečna primanja viša od 670 KM registruju kao nadprosječna. Domaćinstvo Jung ima mjesečna primanja od 230 KM, Wundt 567 KM, Skinner 1.045 KM; Adler 746 KM, Erickson 984 KM, Terman 350 KM i Lindsay 650 KM.
      c.1. Transformirajte ove rezultate na nominalnu skalu.
      c.2. Transformirajte ove rezultate na rang skalu.

2. Grafičko i tabelarno predstavljanje podataka

Zamislite da ispred sebe imate podatke o školskom uspjehu i pokazateljima socio-ekonomskog statusa za više od 2000 učenika. Šta bi mogli zaključiti na osnovu podataka? Ko postiže bolji školski uspjeh: dječaci ili djevojčice? U kakvom je odnosu školski uspjeh i stepen obrazovanja roditelja? Od samih podataka prikupljenih tokom istraživanja zapravo nemamo mnogo koristi. Podaci koji nisu sistematizirani i uređeni nazivaju se «sirovi podaci».

Podatke je potrebno organizirati i prikazati tako da ih možemo opisati, analizirati, interpretirati. U tu svrhu koristimo numeričke i grafičke postupke pomoću kojih organiziramo i reprezentiramo podatake na jasan, ekonomičan i razumljiv način. **Numeričkim postupcima** izračunavamo određene vrijednosti kojima opisujemo uzorak; izračunate vrijednosti nazivamo statisticima. Numerički postupci pružaju precizne i objektivne informacije o podacima. **Grafičkim postupcima** vizuelno predstavljamo podatke. Za razliku od numeričkih postupaka, oni uključuju detaljnije informacije o nekim karakteristikama podataka, npr. obliku distribucije.

Izbor načina prikazivanja podataka zavisi od korištene skale mjerenja. Kategorijalne podatke grafički predstavljamo u stupčastim i torta dijagramima. Numeričke podatke možemo predstaviti kroz tabelarni prikaz distribucije frekvencija, „stablo i listovi" (engl. *steam and leaf*) i box-plot prikaz. Numerički podaci grafički se prikazuju pomoću histograma i procentualne kumulativne krive.

U primjerima koji slijede prikazani su najčešće korišteni grafički i tabelarni postupci organiziranja i prikazivanja podataka.

PRIMJER 2.1

Na grupi od 100 ispitanika primjenjen je Test opće informiranosti. Podaci (broj tačnih odgovora) prikazani su ispod.

24	27	32	20	20	15	20	20	19	22
18	22	27	28	15	20	14	24	24	19
24	9	30	25	19	21	13	20	22	20
19	20	25	24	26	21	19	16	22	22
14	12	25	25	12	17	19	21	18	14
15	17	18	18	17	15	20	19	22	16
19	12	24	17	10	21	21	14	19	22
16	8	21	20	22	15	21	22	16	18
22	30	27	21	17	25	19	20	19	19
15	27	20	24	21	16	18	16	19	16

Podatake ćemo poredati po veličini, a zatim za svaki odrediti koliko se puta pojavljuje. Uvrštavanjem ovih vrijednosti u tabelu, podatke ćemo urediti u vidu tabele **distribucije frekvencija negrupiranih podataka** (tabela 2.1.1).

Tabela 2.1.1: Distribucije frekvencija negrupiranih podataka

rezultat	f
8	1
9	1
10	1
12	3
13	1
14	4
15	6
16	7
17	5
18	6
19	13
20	12
21	9
22	10
24	7
25	5
26	1
27	4
28	1
30	2
32	1

Distribucije frekvencija negrupiranih podataka praktičan je način organiziranja i prikazivanja manjeg skupa podataka. Za veći skup podataka, koristimo tabelarno prikazivanje pomoću **distribucije frekvencija grupiranih podataka**. Podatke grupiramo na sljedeći način:

- Odredimo totalni raspon rezultat.
- Raspon rezultata podijelimo na određeni broj razreda tako da razredi imaju istu veličinu. Broj razreda je obično između 5 i 20.
- Odredimo veličinu intervala razreda prema izrazu: totalni raspon/ broj razreda. Sredina razreda treba da koincidira sa stvarnim rezultatima. Granice razreda ne bi trebale koincidirati sa stvarnim rezultatima.
- Odredimo broj podataka za svaki razred.
- Razrede i frekvencije uvrstitimo u tabelu.

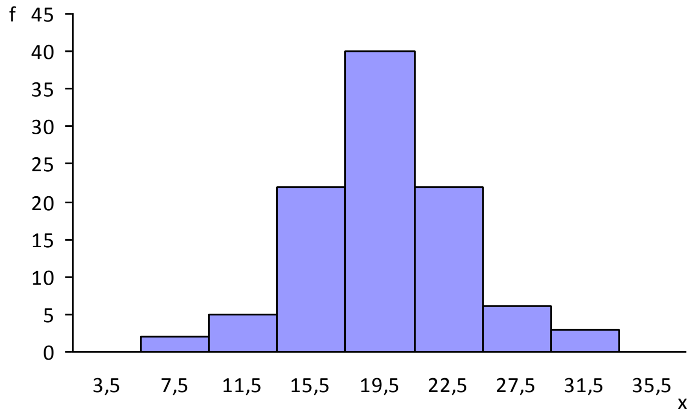
U tabeli 2.1.2 prikazana je distribucije frekvencija grupiranih podataka.

Table 2.1.2: Distribucije frekvencija grupiranih podataka

razred	f
6-9	2
10-13	5
14-17	22
18-21	40
22-25	22
26-29	6
30-33	3
total	100

Grafički prikaz ovako organiziranih podataka naziva se **histogram**. Histogram distribucije frekvencija grupiranih podataka prikazan je na slici 2.1.1.

Slika 2.1.1. Distribucija frekvencija grupiranih podataka



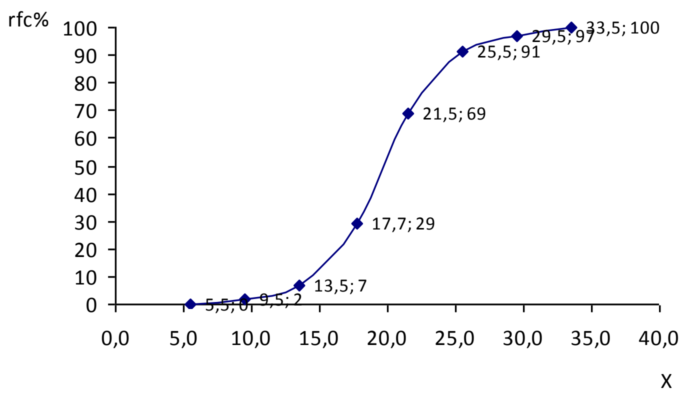
Ponekad se koristi tabelarni/grafički prikaz distribucije **relativnih i kumulativnih frekvencija**. Relativne frekvencije određuju se djeljenjem frekvencije datog razreda sa ukupnim brojem podataka. Izražava se kao proporcija ili postotak. Kumulativne frekvencije određujemo sukcesivnim dodavanjem frekvencija razreda ispod datog razreda. Izražava se kao frekvencija ili postotak. U tabeli 2.1.3. prikazane su distribucije relativnih i kumulativnih frekvencija. Grafički prikaz distribucije relativnih kumulativnih frekvencija (u procentima) dat je na slici 2.1.2.

Tabela 2.1.3: Distribucije relativnih i kumulativnih frekvencija

razred	f	rf	rf(%)	cf	rcf	rcf(%)
6-9	2	0,02	2,0%	2	0,02	2,0%
10-13	5	0,05	5,0%	7	0,07	7,0%
14-17	22	0,22	22,0%	29	0,29	29,0%
18-21	40	0,40	40,0%	69	0,69	69,0%
22-25	22	0,22	22,0%	91	0,91	91,0%
26-29	6	0,06	6,0%	97	0,97	97,0%
30-33	3	0,03	3,0%	100	1	100,0%
total	100	1	100%			

Distribuciju kumulativnih frekvencija koristimo kada trebamo odrediti položaj podatka u distribuciji svih podataka, odnosno položaj ispitanika kojem podatak pripada u grupi svih ispitanika. Za određivanje položaja nekog podatka koristimo centile i decile. **Centili** su vrijednosti koje skup rezultata dijele na 100 jednakih dijelova (frakcije rezultata od 1%). Prvi centil obuhvaća 1% najnižih rezultata, drugi centil 1% sljedećih... dvadeseti centil obuhvaća onih 1% koji su na dvadesetom mjestu od najnižeg rezultata. Ako neki rezultat pada u 60. centil, to znači da je 60% rezultata jednake ili niže vrijednost, a 40% više vrijednosti. **Decili** su vrijednosti koje skup rezultata dijele na 10 jednakih dijelova (frakcije rezultata od 10%). Prvi decil obuhvaća 10% najnižih (ili najviših) rezultata, drugi decil sljedećih 10%, itd. Prvi centil je početna vrijednost za prvi decil, 10. centil je početna vrijednost za drugi decil, dok je 90. centil početna vrijednost za deseti decil.

Slika 2.1.2. Distribucija procentualnih kumulativnih frekvencija



Tabelarnim i grafičkim prikazom kumulativnih frekvencija možemo odrediti koliko se podataka nalazi u određenom intervalu (npr. u intervalu od 5,5 do 21,5 nalazi se 69% podataka, tj. 69 podataka). Nadalje, možemo odrediti granice intervala u kojem se nalazi određeni procenat (ili broj) podataka (npr. 9% najviših vrijednosti nalazi se u intervalu od 21,5 do kraja distribucije, tj. do 33,5).

PRIMJER 2.2

U tabeli ispod prikazani su podaci prikupljeni od 29 pacijenata koji se liječe na Pedijatrijskoj klinici. Prikupljeni su podaci za dob pacijenta (izražena u godinama), spol (1-dječaci; 2-djevojčice), tjelesnu težinu TTEZ (u kg), puls i sistolni krvni pritisak SKP. U ovom primjeru, podatke ćemo organizirati i predstaviti na različite načine zavisno od tipa varijabli.

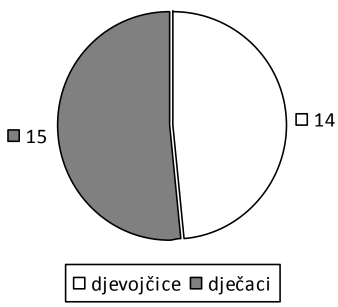
Tabela 2.2.4: Podaci za dob, spol, tjelesnu težinu TTEZ (u kg), puls i sistolni krvni pritisak SKP

r/br.	DOB	SPOL	TTEZ	PULS	SKP
1	0,3	1	4	90	65
2	0,2	1	3	178	65
3	0,3	2	4	140	87
4	0,3	1	5	120	60
5	0,3	1	5	155	65
6	0,3	1	5	150	70
7	0,3	2	4	156	65
8	0,6	1	8	163	88
9	0,9	1	8	128	105
10	1,1	2	7	110	95
11	5,1	1	28	160	99
12	1,5	1	10	110	86
13	2,2	2	12	110	95
14	1,5	1	10	110	86
15	1,5	2	12	110	86

r/br	DOB	SPOL	TTEZ	PULS	SKP
16	5,4	2	22	100	100
17	6,5	2	23	96	99
18	12,7	1	51	79	109
19	10,1	1	24	93	104
20	16,2	2	52	74	101
21	15,6	1	35	100	128
22	9,4	2	28	88	86
23	15,3	2	45	90	100
24	9,1	2	45	110	112
25	9,7	1	30	90	110
26	9,4	2	20	160	99
27	12,2	2	25	100	112
28	6,0	2	20	117	115
29	16,0	2	47	81	128

Varijabla SPOL je kategorijalna, i stoga su podaci prikazani u obliku **torta dijagrama** (slika 2.2.3). Kategorije varijable spol (djevojčice i dječaci) označene su na različit način. Pored svake kategorije naveden je broj podataka.

Slika 2.2.3. Spolna zastupljenost ispitanika



Varijabla DOB je kontinuirana, a vrijednosti su izražene kao decimalni brojevi. Za predstavljanje podataka koristit ćemo „stablo i listovi" prikaz (tabela 2.2.5). „Stablo" sadržava brojeve sa lijeve strane decimalnog zareza, a „listovi" brojeve sa desne strane decimalnog zareza.

Tabela 2.2.5. "Stablo i listovi" prikaz dobi ispitanika

0	2	3	3	3	3	3	3	6	9	(9)
1	1	5	5	5						(4)
2	2									(1)
5	1	4								(2)
6	0	5								(2)
9	1	4	4	7						(4)
10	1									(1)
12	2	7								(2)
15	3	6								(2)
16	0	2								(2)

1 | 1 znači 1,1 godina

U zagradama su prikazane frekvencije „listova", tj. broj podataka za svaku vrijednost „stabla". „Stablo i listovi" prikaz omogućava vizuelnu impresiju o distribuciji podataka. Ako zamislimo da prikaz rotiramo za 90° u smjeru obrnutom od smjera kazaljke na satu, dobit ćemo prikaz koji je veoma sličan šipkastom dijagramu. Iz prikaza možemo vidjeti da je najveći broj pacijenata mlađih od godinu dana (ukupno devet), zatim onih sa devet godina (4 pacijenta) itd.

Varijabla „TJELESNA TEŽINA" je kontinuirana. Stoga distribuciju podataka ove varijable možemo predstaviti pomoću „stablo i listovi" prikaza. Vrijednosti varijable izražene su kao cijeli brojevi sa jednom ili dvije cifre. Stoga u „stablo i listovi" prikazu 2|4 znači 24 kg, dok 0|6 znači 6 kg. U tabeli 2.2.6 dat je „stablo i listovi" prikaz vrijednosti tjelesne težine.

Tabela 2.2.6. „Stablo i listovi" prikaz tjelesne težine

0	3	4	4	4	5	5	5	7	8	8	(10)
1	0	0	2	2							(4)
2	0	0	2	3	4	5	8	8			(8)
3	0	5									(2)
4	5	5	7								(3)
5	1	2									(2)

0 | 3 znači 3 kg

2 | 0 znači 20 kg

Podaci dobiveni mjerenjem pulsa ispitanika najprije su grupirani u razrede. Formirano je šest razreda (prvi od 61 do 80, zadnji od 161 do 180), nakon čega su određeni podaci koji pripadaju svakom razredu. Tabelarni prikaz distribucije frekvencija rezultata grupiranih u razrede prikazan je u tabeli 2.2.7). Na sličan način formirana je distribucija frekvencija rezultata grupiranih u razrede varijable sistolni krvni pritisak (tabela 2.2.8).

Tabela 2.2.7: Distribucija frekvencija rezultata varijable PULS

Razred	f
61-80	2
81-100	10
101-120	8
121-140	2
141-160	5
161-180	2

Tabela 2.2.8: Distribucije frekvencija rezultata SKP

Razred	f
60-69	5
70-79	1
80-89	6
90-99	5
100-109	6
110-119	4
120-129	2

PRIMJER 2.3

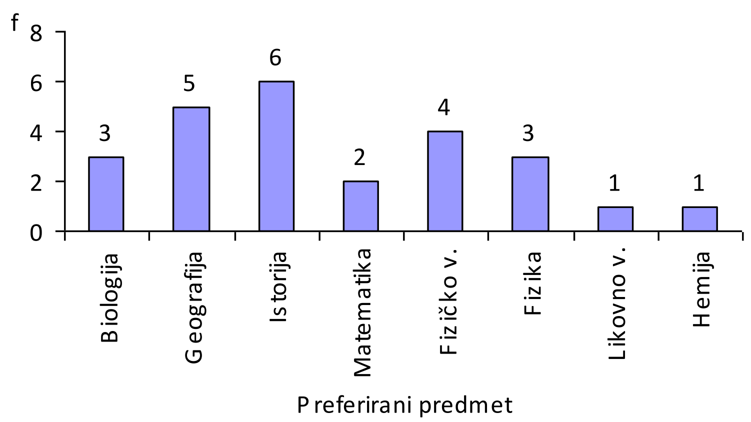
Treći razred „1. Osnovne škole“ broji ukupno 25 učenika. Na pitanje „Koji vam je omiljeni školski predmet?“, troje učenika navelo je biologiju, petoro geografiju, šestoro historiju, dvoje matematiku, četvoro tjelesni odgoj, troje fiziku, jedno likovnu kulturu i jedno hemiju. Dobivene podatke naprije ćemo tabelarno prikazati (tabela 2.3.9). U prvi stupac tabele unijeti su predmeti, a u drugi broj učenika koji su navedene predmete naveli kao omiljene.

Tabela 2.3.9: Podjela učenika IIIc odjeljenja prema preferiranom školskom predmetu

Preferirani školski predmet	f
Biologija	3
Geografija	5
Historija	6
Matematika	2
Tjelesni odgoj	4
Fizika	3
Likovna kultura	1
Hemija	1
total	25

S obzirom da je varijabla „PREFERIRANI PREDMET“ kategorijalna, za grafički prikaz koristit ćemo stupčasti dijagram (slika 2.3.4).

Slika 2.3.4. Podjela učenika IIIc odjeljena prema preferiranom školskom predmetu



Na školskom sistematskom pregledu mjerena je visina svakog od ovih 25 učenika. Utvrđene su sljedeće visine učenika, zaokružene na cijeli centimetar:

131	134	136	135	139
135	137	133	138	136
136	134	135	136	134
137	135	132	135	134
135	137	135	132	136

Podaci se mogu predstaviti tabelarno u obliku negrupirane distribucije frekvencija (tabela 2.3.10) ili pomoću „stablo i listovi“ prikaza (slika 2.3.11).

Tabela 2.3.10: Visina učenika IIIc odjeljenja

Visina učenika (cm)	f
131	1
132	2
133	1
134	4
135	7
136	5
137	3
138	1
139	1
total	25

Tabela 2.3.11. “Stablo i listovi” prikaz visina učenika IIIc odjeljenja

13	1							(1)
13	2	2						(2)
13	3							(1)
13	4	4	4	4				(4)
13	5	5	5	5	5	5	5	(7)
13	6	6	6	6	6			(5)
13	7	7	7					(3)
13	8							(1)
13	9							(1)

13	1	znači 131 cm
13	8	znači 138 cm

Iz Tabele 2.3.10 možemo vidjeti da je najniži učenik u razredu visok 131 cm, a najviši 139 cm. Dakle utvrđeni raspon rezultata je 8 (139 – 131). Obzirom da je utvrđeni raspon rezultata mali, te da imamo relativno mali broj ispitanika (N=25), tj. da već na temelju tabele 2.3.10 imamo dobar pregled distribucije, rezultate nije potrebno grupirati u razrede. Tabeli 2.3.10 dodat ćemo vrijednosti relativnih i kumulativnih frekvencija [rf, rf (%), cf, rcf i rcf (%)] koje će nam omogućiti dodatne informacija o distribuciji visina učenika (tabela 2.3.12).

Iz tabele 2.3.12 mogli bismo izvući zaključak da je 20% učenika visoko 136 cm – procenat rezultata u tom razredu [rf(%)] je upravo 20. Ipak, preciznije bi bilo reći da je 20% učenika visoko između 135,5 i 136,5 cm. Naime, zbog nepreciznosti našeg mjerenja (tj. zaokruživanja rezultata na cijeli cm), visina učenika koji su visoki između 135,5 i 136,5 cm bit će registrirana kao visina od 136 cm. Dakle, iako je prikazani rezultat 136 cm, stvarni rezultat kreće se u intervalu omeđenom stvarnom donjom granicom 135,5 cm i stvarnom gornjom granicom 136,5 cm.

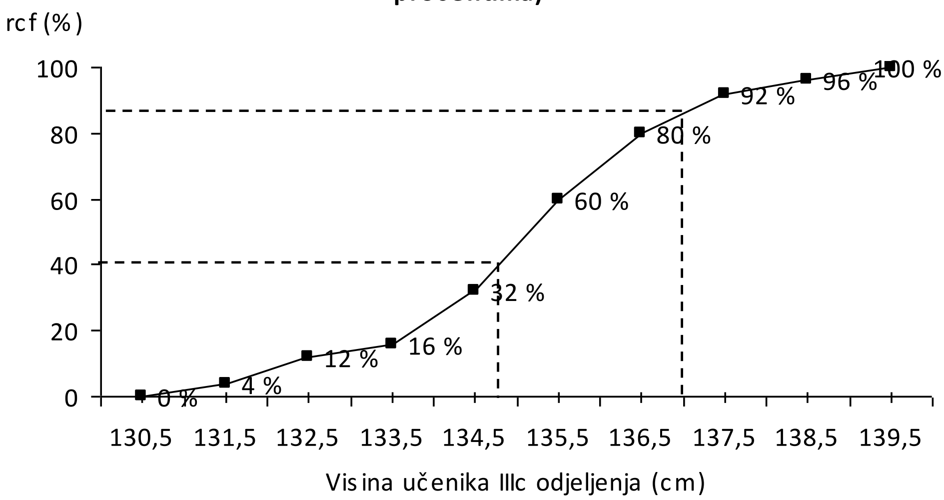
Tabela 2.3.12. Distribucija visina učenika IIIc odjeljenja

Visina učenika (cm)	f	rf	rf(%)	cf	rcf	rcf(%)
131	1	0,04	4	1	0,04	4
132	2	0,08	8	3	0,12	12
133	1	0,04	4	4	0,16	16
134	4	0,16	16	8	0,32	32
135	7	0,28	28	15	0,6	60
136	5	0,2	20	20	0,8	80
137	3	0,12	12	23	0,92	92
138	1	0,04	4	24	0,96	96
139	1	0,04	4	25	1	100
total	25	1	100			

Nadalje, iz tabele 2.3.12 možemo zaključiti da je 40% učenika visoko najmanje 136 cm (preciznije 135,5 cm) – ukoliko saberemo procenat učenika sa visinom jednakom ili većom od 136 cm dobit ćemo upravo 40 procenata distribucije.

Podatke možemo predstaviti i pomoću ogive (slika 2.3.5)

Slika 2.3.5. Distribucije relativnih kumulativnih frekvencija (u procentima)



Koristeći se ogivom, tj. spajajući preko ogive rezultat sa odgovarajućom kumulativnom procentualnom vrijednošću i obratno, možemo doći do zaključaka kao što su: 40% učenika je niže od oko 134,75 cm; nešto manje od 20% učenika je visoko najmanje 137 cm; nešto više od 40% učenika svojom visinom spada u interval između dva spomenuta rezultata (134,75 i 137 cm).

PRIMJER 2.4

Određivanje centila i decila ilustrirat ćemo na primjeru 2.1. Odredit ćemo: a) u koji centil pada rezultat 22, b) rezultat koji pada na 25 centil, c) centil rezultata 19, d) rezultat koji odgovara 60 centilu i e) granice 7. decila. Za određivanje zadatih vrijednosti koristit ćemo tabelu 2.1.3 i grafički prikaz distribucije relativnih kumulativnih frekvencija (slika 2.1.2).

a) U koji centil pada rezultat 22

Centil možemo odrediti pomoću formule:

Centil rezultata = (Rang rezultata / N) x 100.

Npr. ako je među 90 rezultata neki rezultat 40 po redu, onda se taj rezultat nalazi u: (40/90) x 100 = 44. centilu.

Kod rezultata grupiranih u razrede centil se određuje prema sljedećoj formuli:

$$\text{centil} = RKF_D + (X - D)\frac{RF_R}{i}$$

gdje je:

X – rezultat za koji tražimo centil,

RKF_D – relativna kumulativna frekvencija (%) rezultata ispod razreda u kojem je rezultat X,

D – prava donja granica razreda u kojem se nalazi rezultat X,

i – interval,

RF_R – relativna frekvencija (%) rezultata u razredu u kojem se nalazi rezultat X.

Nakon što odredimo poznate vrijednosti: X=22; RKF_D=69,0; D=21,5; i=4; RF_R=22,0, i uvrstimo ih u izraz, izračunat ćemo centil rezultata X=22:

$$\text{centil} = RKF_D + (X - D)\frac{RF_R}{i} = 69{,}0 + (22 - 21{,}5)\frac{22{,}0}{4} = 71{,}75$$

Možemo zaključiti da se ispod rezultata X=22 (uključujući i rezultat 22) nalazi 71,75% rezultata (tj. ispitanika), dok se iznad ovog rezultata nalazi 28,25% rezultata (tj. ispitanika).

b) Koji rezultat pada na 25 centil

Za određivanje rezultata koji pada na 25 centil primijenit ćemo sljedeći izraz:

$$X = D + \left(\frac{\text{centil} \times N}{100} - f_D \right) \times \frac{i}{f_R}$$

gdje je:

Centil – zadati centil

D – prava donja granica razreda u kojem je rezultat X,

f_D – ukupan broj rezultata ispod razreda u kojem je rezultat X,

f_R – broj rezultata u razredu u kojem je rezultat X,

i – interval.

Nakon što odredimo poznate vrijednosti: Centil=25; D=13,5; f_D=7; f_R=22; i=4, i uvrstimo ih u izraz, izračunat ćemo rezultat koji pada na 25 centil:

$$X = D + \left(\frac{\text{centil} \times N}{100} - f_D \right) \times \frac{i}{f_R} = 13{,}5 + \left(\frac{25 \times 100}{100} - 7 \right) \times \frac{4}{22} = 16{,}77$$

Rezultat 16,77 dijeli distribuciju na dva dijela, tako da je ispod ovog rezultata 25% rezultata (tj. ispitanika), a iznad 75%.

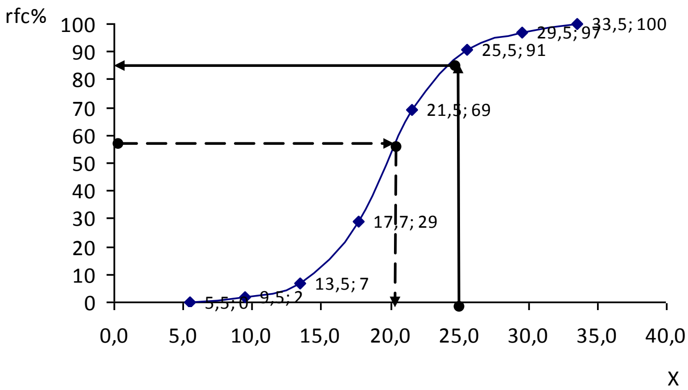
c) Odrediti centil rezultata 25

Za određivanje centila rezultata X=25 koristi ćemo grafički prikaz distribucije relativnih kumulativnih frekvencija (slika 2.4.7) Sa apscise ćemo povući okomitu liniju od tačke koja pada na X=25, do mjesta presjeka za procentualnom ogivom. Zatim ćemo od ove tačke povući liniju paralelnu sa apscisom, do ordinate. Tačka u kojoj ova linija sječe ordinatu odgovara traženom centilu. Tako je Centil ≈ 90%

d) Odrediti rezultat koji odgovara 60 centilu

Za određivanje rezultata koji odgovara 60. centilu koristi ćemo procentualnu ogivu (slika 2.4.7). Sa ordinate ćemo povući okomitu liniju od tačke koja pada na rcf%=60, do mjesta presjeka sa procentualnom ogivom. Zatim ćemo od ove tačke povući liniju paralelnu sa ordinatom, do apscise. Tačka u kojoj ova linija sječe apscisu odgovara traženom rezultatu. Tako je X ≈ 20.

Slika 2.4.7. Distribucija procentualnih kumulativnih frekvencija



e) Koje su granice 7. decila?

Sedmi decil počinje sa 60-im a završava sa 69,99-im centilom. Stoga je potrebno odrediti rezultate koji padaju u 60 i 69,99 centil.

$$X_{60centil} = D + \left(\frac{centil \times N}{100} - f_D \right) \times \frac{i}{f_R} = 17,5 + \left(\frac{60 \times 100}{100} - 29 \right) \times \frac{4}{40} = 20,6$$

$$X_{69,99centil} = D + \left(\frac{centil \times N}{100} - f_D \right) \times \frac{i}{f_R} = 17,5 + \left(\frac{69,99 \times 100}{100} - 29 \right) \times \frac{4}{40} = 21,6$$

Granice 7. decila su: 20,6 – 21,6.

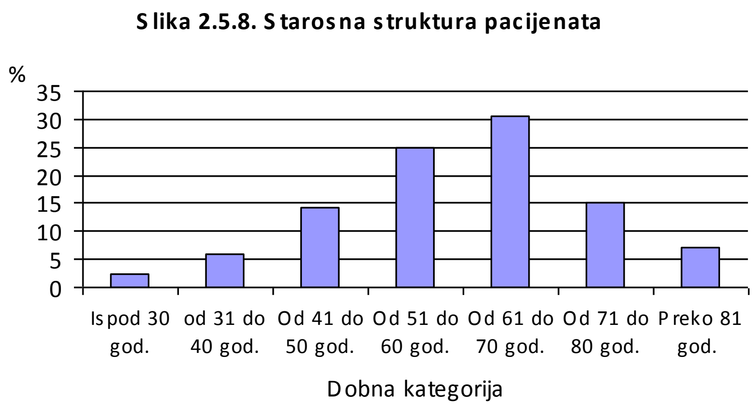
PRIMJER 2.5

Kantonalni Zdravstveni zavod Sarajevo proveo je istraživanje sa ciljem utvrđivanja karakteristika populacije pacijenata sa dijagnosticiranom hipertenzijom te utvrđivanja načina na koji liječnici opće prakse tretiraju ove pacijente. Anketiranje je provedeno na reprezentativnom uzorku od 300 pacijenata iz Kantona.

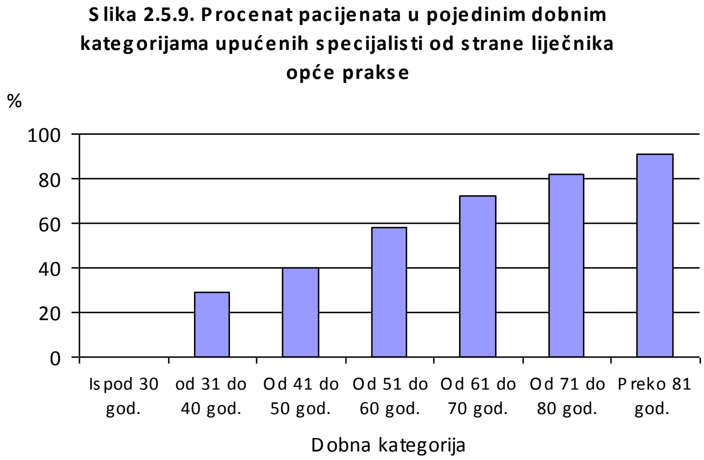
Na priloženim graficima prezentirani su dobiveni rezultati istraživanja prema pojedinim varijablama.

Prva varijabla je spol – kategorijalna varijabla utvrđena na nominalnoj skali. Obzirom na prirodu mjerenja u ovoj varijabli, prikladan način grafičkog prikazivanja podataka je torta dijagram:

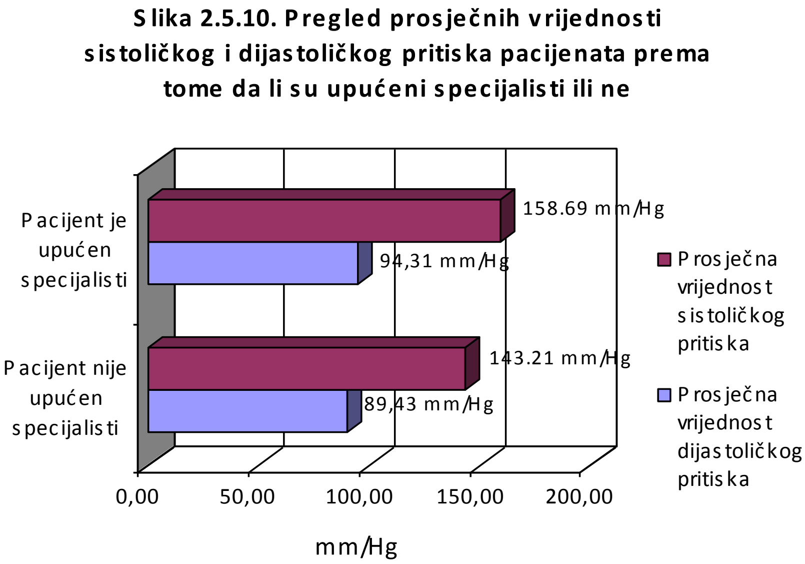
Naredni skup podataka odnosi se na starosnu strukturu pacijenata. Starosna strukura ispitanika je kvantitativna varijabla – iako ispitanike razvrstavamo u jednu od kategorija prema njihovoj starosnoj dobi, na temelju pripadnosti određenoj kategoriji, ispitanici mogu dobiti i odgovarajući rang. Tako, npr., svi ispitanici mlađi od 30 godina će dobiti rang „1“, ispitanici starosti od 31 do 40 godina rang „2“, ispitanici u narednoj starosnoj skupini rang „3“...



Prilikom utvrđivanja procenata pacijenata pojedinih starosnih uzrasta koji su upućeni liječniku specijalisti (slika 2.5.9) ponovo smo se koristili mjerenjem na nominalnoj skali – ispitanici su svrstani u jednu od dvije kategorije: „Upućen specijalisti/ Nije upućen specijalisti“, a potom je u svakoj od kategorija utvrđen broj ispitanika.



Slika 2.5.10 predstavlja kombinaciju kategorijalne i kontinuirane varijable. Varijabla „Upućen specijalisti" je, kako smo već naveli, kategorijalna, dok je varijabla „Krvni (sistolički/ dijastolički) pritisak" kontinuirana.



ZADACI

1. Na grupi od 30 učenika primjenjen je test inteligencije. Dobiveni su sljedeći rezultati:

70	85	100	114	93	89	126	76	142	104
68	73	109	82	97	79	116	121	84	100
93	74	95	113	96	83	70	110	69	91

   a. Prikažite tabelarno distribuciju frekvencija rezultata grupiranih u 8 razreda. Neka donja granica prvog razreda bude 67.
   b. Koje su prave granice zadnjeg razreda?
   c. Konstruirajte „stablo i listovi“dijagram.
   d. Nacrtajte histogram.
   e. Grafički prikažite distribuciju relativnih kumulativnih frekvencija.

2. U tabeli ispod prikazana je dobna distribucija 230 članova planinarskog društva.

Dob	<15	16-19	20-29	30-39	40-49	50-59	> 60
Broj članova (f)	16	25	40	47	50	45	

   a. Koliko je članova starijih od 60 godina? Izračunatu vrijednost unesite u tabelu.
   b. Podatke prikažite u tabeli sa pravim donjim i gornjim granicama razreda.
   c. Nacrtajte histogram podataka.

3. Nivo hemoglobina izmjeren je na grupi od 40 pacijenata. Dobivene su sljedeće vrijednosti mjerenja, zaokružene na prvu decimalu:

9,2	9,7	9,7	10,3	10,9	10,9	11,4	11,4	11,4	11,7
11,9	12,1	12,4	12,6	12,8	13	13,1	13,2	13,7	13,8
14	14,1	14,5	14,6	14,6	14,8	14,9	15,1	15,1	15,3
15,6	15,9	16	16,4	16,5	16,6	16,7	16,8	16,8	16,9

   a. Formirajte tabelu rezultata grupiranih u razrede.
   b. Nacrtajte histogram rezultata.

4. U tabeli ispod prikazana je distribucija školskih ocjena 520 učenika petih razreda.

Školski uspjeh	1	2	3	4	5
Broj učenika (f)	2	35	240	150	93

Formirajte nove distribucije (u odnosu na distribuciju prikazanu u tabeli) tako da:

a. se u svakom razredu nalazi podjednak broj učenika,
b. je više učenika koji su postigli niže ocjene,
c. je više učenika koji su postigli više ocjene.

5. U dva kruga grupnog nadmetanja Lige šampiona postignuta su 72 gola. U tabeli ispod prikazan je broj pogodaka tokom različitih perioda utakmice. Grafički prikažite podatke iz tabele.

minuta	broj golova
1-15	8
16-30	9
31-45	5
Dodatno vrijeme (1)	3
46-60	9
61-75	15
76-90	20
Dodatno vrijeme (2)	3

6. Na osnovu stem-and-leaf prikaza odredite sirove podatke.

žene	stem	muškarci
3 3 2 1 1 1	6	4 8
6 4 3 3 2 2 1 1 1 1 1	7	2 5 5 9
9 3 3 1	8	1 1 1 1 3 4 5 5 6 6 6 8 9 9
5 5 3 2	9	2 2 3 6 8

7. VIa razred Osnovne škole „Sigmund Freud“ broji ukupno 25 učenika. Na pitanje „Koji vam je omiljeni školski predmet?“, učenici su dali sljedeće odgovore: njih troje navelo je maternji jezik, četvoro hemiju, petoro tjelesni odgoj, dvoje matematiku, šestoro historiju, dvoje fiziku, jedno tehnički odgoj i dvoje geografiju.

a. Predstavite ove podatke tabelarno. Navedite procenat slučajeva za svaku kategoriju.
b. Podatke predstavite grafički pomoću stupčastog dijagrama.

8. Na školskom sistematskom pregledu izmjerena je visina 25 učenika iz prethodnog zadatka. Utvrđene su sljedeće visine učenika, zaokružene na cijeli centimetar:

148	158	158	153	141
134	152	148	134	145
158	142	137	152	148
134	140	153	155	146
152	138	137	141	152

   a. Ove rezultate predstavite u obliku jednostavne distribucije frekvencija.
   b. Rezultate predstavite pomoću „stablo i listovi“ prikaza. Opišite distribuciju.
   c. Rezultate prezentirajte u vidu grupirane distribucije frekvencija.
   d. Rezultate predstavite grafički, pomoću histograma.
   e. U tabeli sa grupiranom distribucijom frekvencija dodajte kolone sa relativnim frekvencijama [rf], relativnim frekvencijama u procentima [rf(%)], kumulativnim frekvencijama [cf], relativnim kumulativnim frekvencijama [rcf] i relativnim kumulativnim frekvencijama u postocima [rcf(%)].
   f. Relativne kumulativne frekvencije izražene u procentima [rcf(%)] prikažite grafički, pomoću OGIVE.
   g. Nastavnik tjelesnog odgoja želi za svoj školski košarkaški tim testirati 20% najviših učenika iz razreda. Koliko će učenika pozvati na testiranje? Koja će biti visina najnižeg učenika kojeg će pozvati na testiranje?
   h. Školski ljekar sumnja da kod djece koja su u VI razredu niža od 137 cm može postojati određena hormonalna neravnoteža u organizmu koja usporava rast. Kako se ovaj poremećaj tretira vitaminskom terapijom, ljekar želi znati za koji broj učenika iz razreda treba naručiti vitamine?

9. Kompanija želi zaposliti 10 novih radnika na poziciji istraživača tržišta te je u novinama objavila oglas na koji se javilo ukupno 46 kanditata. Svi kandidati su pozvani na test namjenjen utvrđivanju nivoa njihovog znanja iz statistike (maksimalni rezultat na testu iznosio je 40).

   Dobiveni su sljedeći rezultati:

19	25	26	27	28	29	30	31	32	39
20	25	26	27	28	29	30	31	32	
23	25	26	27	28	29	30	32	35	
24	25	27	27	29	29	30	32	35	
25	26	27	27	29	29	31	32	35	

   Napravite tabelarni prikaz distribucije frekvencija rezultata grupiranih u razrede. Utvrdite relativne frekvencije [rf], relativne frekvencije u procentima [rf(%)], kumulativne frekvencije [cf], relativne kumulativne frekvencije [rcf] i relativne kumulativne frekvencije u postocima [rcf(%)].

a. Grupirane frekvencije predstavite grafički, pomoću histograma.
b. Relativne kumulativne frekvencije izražene u procentima [rcf(%)] prikažite grafički, pomoću ogive.

Koristeći tabelu grupiranih frekvencija ili ogivu, odgovorite na naredna pitanja:

a. Koji rezultat (otprilike) dijeli distribuciju na 50% gornjih i donjih rezultata (50% slabijih i 50% boljih kandidata).
b. Kompanija će zaposliti one kandidate koji su na testu iz statistike skupili najmanje 34 boda. Koliko kandidata će odmah dobiti ponudu za posao? Koliko je to procenata ukupno prijavljenih kandidata?
c. Kako kompaniji treba ukupno 10 novih uposlenika, odlučeno je se na dodatni trening iz statistike pozove duplo više kandidata nego što ima preostalih upražnjenih radnih mjesta. Kandidat koji će biti pozvani na trening su oni koji imaju najbolje rezultate na testu kada se isključe oni kandidati koji su već dobili ponudu za posao. Koliko kandidata će biti pozvano na dodatni trening? Koji su granični rezultati (donji i gornji) koji određuju ovu skupinu? Koji je to procenat ukupnog broja prijavljenih kandidata?

10. Klinički centar registovao je učestalost prijema pacijenata koji su se žalili na probleme sa respiratornim sistemom tokom godine dana. Tokom te godine zabilježeno je 1500 prijema. Ukoliko se želi saznati u kojem godišnjem dobu je veći broj prijema, na koji način će se podaci prikazati? Koji će se grafički prikaz koristiti?

11. U prodavnici tehničkih uređaja kupci su u knjigu utisaka mogli upisivati primjedbe koje imaju na pružene usluge. Primjedbe su razvrstane u pet grupa. Tokom godinu dana registrovano je 500 žalbi. Na koji način se grafički mogu prikazati registrirane primjedbe? Ukoliko nas zanima u kojem mjesecu je zabilježen najveći broj primjedbi (bez obzira na vrstu primjedbe), koji ćemo grafički prikaz izabrati?

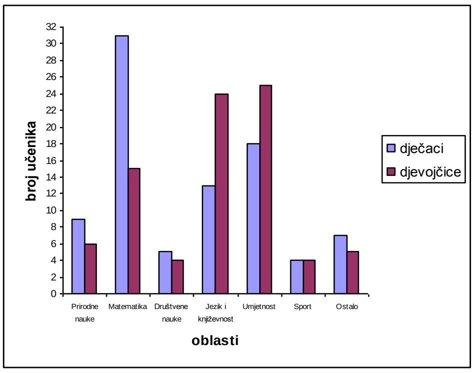
12. Telefonska kompanija prikupila je podatke o dobi i spolu svojih korisnika. Najmlađi ima 14 godina, a najstariji 75. Na koji način se prikupljeni podaci mogu prikazati?

13. U istraživanju su prikupljeni podaci o spolu, tjelesnoj težini i dnevnoj količini unesenih kalorija. Na koji način se podaci mogu prikazati?

14. Ukoliko želimo uporediti procenat podataka koji pripadaju jednoj kategoriji, u odnosu na procenat podataka koji pripadaju drugoj kategoriji, koji prikaz ćemo koristiti?

15. Psiholog je prikupio podatke o verbalnim sposobnostima muškaraca i žena. Koji načini prikazivanja podataka će koristiti ukoliko želi uporediti distribucije rezultata ove dvije grupe?

16. Nastavnici osnovnih škola predlagali su učenike za program za poticanje nadarenosti. Za svakog učenika predložili su oblast za koju smatraju da je učenik nadaren. Na slici ispod prikazani su dobiveni rezultati.



Koliko je dječaka, a koliko djevojčica predloženo za pojedine oblasti? Za koju oblast je bilo najviše prijedloga, a za koju najmanje?

17. Studenti su tokom mjesec dana vodili dnevnik spavanja. Između ostalih vrijednosti, u dnevnik su unosili podatke o vremenu kada idu na spavanje, vremenu kada se bude, dužini spavanja, broju snova i vrsti snova (radi jednostavnosti, pretpostavimo da su trebali odrediti da li je san bio prijatan ili neprijatan). Na koji način se prikupljeni podaci mogu prikazati?

3. Mjere centralne tendencije

Utvrđivanje mjera centalne tendencije numerički je postupak deskripcije podataka. Mjere centralne tendencije predstavljaju vrijednosti koje odražavaju centralno mjesto distribucije podataka; to je vrijednost koja je tipična, tj. reprezentira skup podataka. U svakodnevnom jeziku za mjere centralne tendencije koristimo izraz „prosjek", ali je ovaj termin neprecizan jer uključuje više mjera. Tri najvažnije mjere centralne tendencije su: mod, medijana i aritmetička sredina.

Mod (dominantna vrijednost) je najučestalija vrijednost u distribuciji. Ako je u distribuciji podataka jedna dominantna vrijednost, takvu distibuciju nazivamo unimodalnom. Ako su dvije dominantne vrijednosti, distribucija je bimodalna i tako redom.

U skupu podataka: x=(2,4,5,4,6,3), vrijednost koja se najčešće pojavljuje je 4; stoga je: Mod=4.

Medijana (centralna vrijednost) je vrijednost koja distribuciju dijeli na dva jednaka dijela, tj. vrijednost koja se u nizu podataka poredanih po veličini nalazi tačno u sredini. Položaj medijane određujemo preko izraza:

Položaj medijane = (N+1)/2

U skupu podataka: x=(2,7,4,5,6,7,2), položaj medijane je (7 + 1)/2 = 4, što znači da se centralna vrijednost nalazi na četvrtom mjestu niza rezultata poredanih po veličini:

x=(2,2,4,5,6,7,7)

Centralna vrijednost iznosi: C = 5

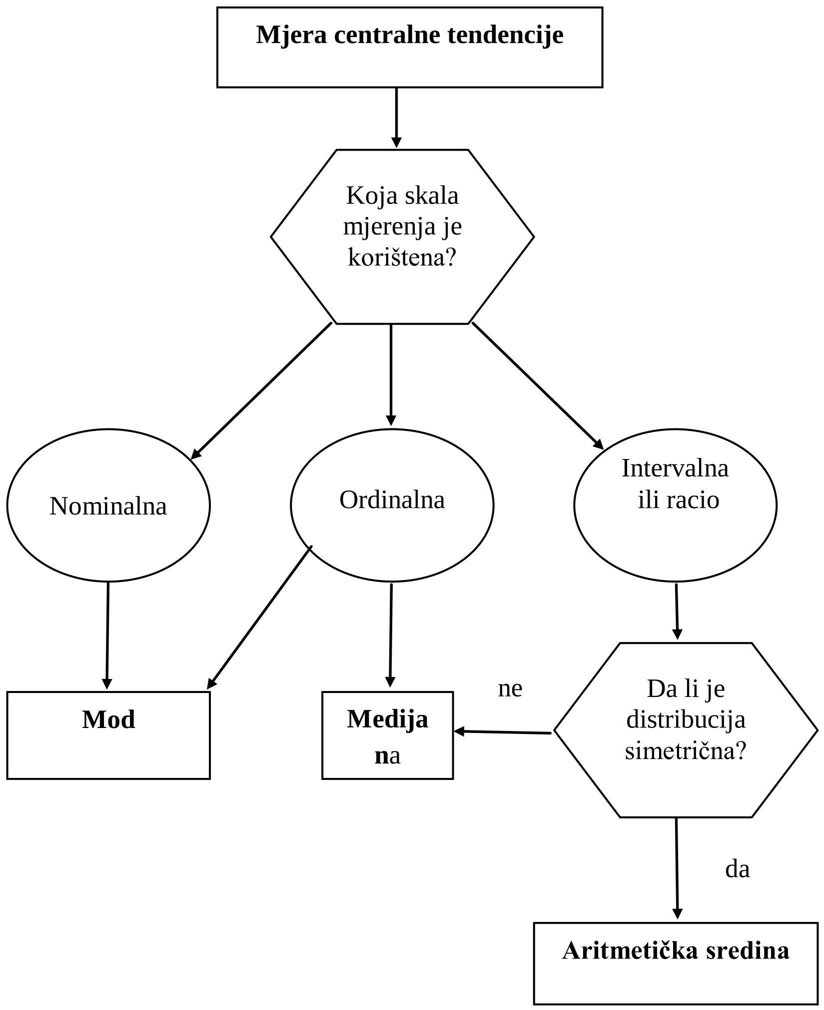
Aritmetička sredina je suma svih rezultata podijeljena brojem rezultata. Aritmetička sredina predstavlja **težište rezultata** i stoga je osjetljiva na ekstremne vrijednosti rezultata. Aritmetička sredina određuje se prema izrazu:

$$M= \Sigma X/N$$

Za skup podataka: x=(2,7,4,5,6,7,2), aritmetička sredina iznosi: M=(2+7+4+5+6+7+2)/7=4,71

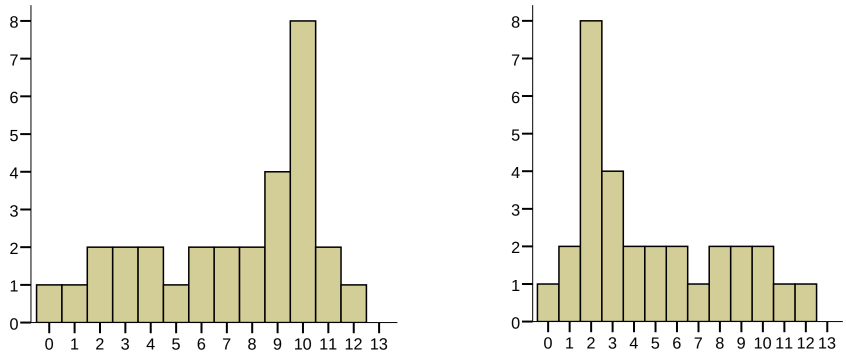
Kod simetrične, unimodalne distribucije mod, medijana i aritmetička sredina bit će približno jednake vrijednosti.

Izbor mjera centralne tendencije zavisi od skale mjerenja i oblika distribucije. Ako je korištena nominalna skala mjerenja kao mjeru centralne tendencije koristimo mod, dok kod ordinalnih skala mjerenja koristimo medijanu. Kada su podaci dobiveni na intervalnoj ili racio skali mjerenja, kao mjeru centralne tendencije koristimo aritmetičku sredinu, ali samo ukoliko je distribucija simetrična; u protivnom koristimo medijanu. Ispod je dat shematski prikaz izbor mjera centralne tendencije.



Distribucija je asimetrična ako je veći broj podataka koncentriran na jednoj strani skale, a manji broj na drugoj strani. Kod **pozitivno asimetrične** distribucije relativno je veći broj nižih vrijednosti, a kod **negativno asimetrične** distribucije relativno je veći broj viših vrijednosti.

Slika 3.3.1: Primjeri negativno i pozitivno asimetričnih distribucija



Skjunis (eng. *skewness*) koristimo kao mjeru (a)simetričnosti distribucije. Izračunava se prema izrazu:

$$\text{skjunis} = \frac{N}{N-2} \times \frac{\sum (X_i - M)^3}{(N-1) \times s^3}$$

Ako je vrijednost skjunisa pozitivna, distribucija je pozitivno asimetrična. Ako je vrijednost skjunisa negativna, distribucija je negativno asimetrična. Ako je vrijednost skjunisa 0, distribucija je simetrična. Uz vrijednost skjunisa potrebno je odrediti i standardnu pogrešku skjunisa!

Kod asimetričnih distribucija aritmetička sredina je pomjerena od medijane u smjeru dužeg kraka distribucije.

PRIMJER 3.1

Na grupi od 20 ispitanika primjenjen je Test znanja iz statistike. Rezultati (broj tačnih odgovora) prikazani su ispod.

24	27	32	20	20	15	20	20	19	22
18	22	27	28	15	20	14	24	24	19

Koliko iznosi mod, medijana i aritmetička sredina?

Podatke treba najprije urediti. Koristit ćemo distribuciju negrupiranih rezultata, prikazanu u tabeli 3.1.1.

Tabela 3.1.1. Distribucija negrupiranih rezultata

rezultat	f
14	1
15	2
18	1
19	2
20	5
22	2
24	3
27	2
28	1
32	1
total	20

1. Mod

Rezultat koji se najučestalije pojavljuje je 20 (pet puta). Prema tome:

$$\text{Mod} = 20$$

2. Medijana

U tabeli distribucije negrupiranih rezultata, rezultati su već poredani po veličini. Poziciju medijane odredit ćemo preko izraza: (N+1)/2.

Položaj medijane = (20+1)/2 = 10,5

Medijana se nalazi na 10,5 mjestu, tj. na polovini rastojanja između 10 i 11 mjesta. Rezultat 19 nalazi se na šestom mjestu, a rezultat 20 od sedmog do jedanaestog mjesta (ukupno je pet rezultata 20). Na dvanaestom i trinaestom mjestu je rezultat 22, itd. Budući da 10. i 11. mjesto zauzima jedan te isti rezultat, X=20, medijana iznosi upravo toliko, tj.

C = 20

3. Aritmetička sredina

Aritmetičku sredinu odredit ćemo preko izraza:

$$M = \frac{\sum X}{N}$$

S obzirom da se pojedini rezultati pojavljuju više puta, gornjem izrazu dodat ćemo vrijednosti frekvencija i dobiti sljedeću formulu:

$$M = \frac{\sum f_i X_i}{N}$$

$$M = \frac{14\times1+15\times2+18\times1+19\times2+20\times5+22\times2+24\times3+27\times2+28\times1+32\times1}{20}$$

Aritmetička sredina iznosi:

M = 21,5

Poređenjem izračunatih mjera centralne tendencije uočavamo da je: M>C=D, što ukazuje da je distribucija rezultata asimetrična.

PRIMJER 3.2

Za podatke iz primjera 2.1 odredite aritmetičku sredinu i medijanu. Tabeli distribucije grupiranih rezultata dodat ćemo vrijednosti sredine razreda (X'), a zatim pomnožiti svaku sredinu razreda sa frekvencijom pripadajućeg razreda (tabela 3.2.1).

Tabela 3.2.1. Distribucije frekvencija grupiranih podataka

razred	f	X	f X
6-9	2	7,5	15
10-13	5	11,5	57,5
14-17	22	15,5	341
18-21	40	19,5	780
22-25	22	23,5	517
26-29	6	27,5	165
30-33	3	31,5	94,5
	Σf_i= 100		$\Sigma f_i X_i$=1970

Vrijednosti $\Sigma f_i X_i$' i Σf_i uvrstit ćemo u formulu i izračunati aritmetičku sredinu.

$$M = \frac{\sum f_i X_i}{N} = \frac{1970}{100} = 19{,}7$$

(napomena: $\Sigma f_i = N$)

Medijanu možemo odrediti na dva načina: računskim postupkom ili očitavanjem iz grafičkog prikaza.

Za izračunavanje medijane potrebno je formirati distribucije kumulativne i procentualne relativne kumulativne frekvencije (tabela 3.2.2).

Tabela 3.2.2. Distribucije frekvencija grupiranih podataka

razred	f	cf	rcf(%)
6-9	2	2	2%
10-13	5	7	7%
14-17	22	29	29%
18-21	40	69	69%
22-25	22	91	91%
26-29	6	97	97%
30-33	3	100	100%
total	100		

Primijenit ćemo sljedeću formulu:

$$X = D + \left(\frac{centil \times N}{100} - f_D \right) \times \frac{i}{f_R}$$

Medijana se nalazi u razredu 18-21 jer se do prave gornje granice ovog razreda nalazi 69% rezultata. Prava donja granica ovog razreda je D=17,5; u ovom razredu nalazi se f_R=40 rezultata; do razreda 18-21 ukupno je f_D=29 rezultata; interval razreda iznosi i=4. Centil centralne vrijednosti je 50-i.

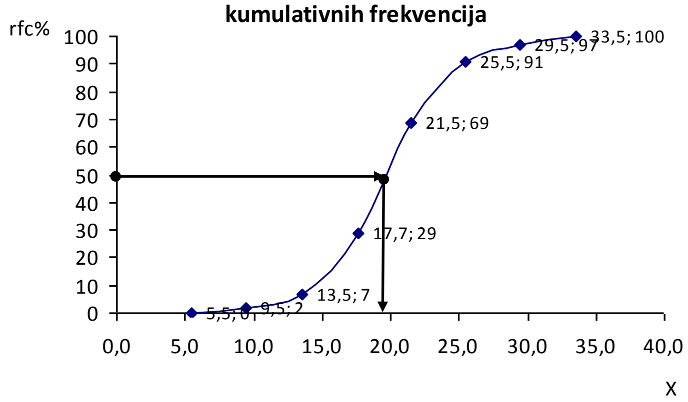
Nakon što uvrstimo vrijednosti u formulu:

$$C = 17,5 + \left(\frac{50 \times 100}{100} - 29 \right) \times \frac{4}{40}$$

izračunat ćemo vrijednost medijane: C = 19,6.

Grafičko određivanje medijane radimo pomoću procentualne ogive (slika 3.2.1). Obzirom da medijana dijeli distibuciju na dva jednaka dijela, potrebno je na apscisi (x) očitati rezultat koji odgovara 50-om procentu na ordinati (rcf %).

Slika 3.2.1. Grafičko određivanje centralne vrijednosti iz distribucija procentualnih kumulativnih frekvencija



Rezultat koji odgovara 50-om centilu iznosi nešto malo više od 19,5. Čitav postupak radi se na milimetarskom papiru na kojem možemo očitati tačnu vrijednost. Kada bi tako uradili, uvjerili bismo se da se radi o rezultatu C=19,6, kojeg smo dobili i računskim putem.

Obzirom da su vrijednosti aritmetičke sredine i medijane podjednake (M≈C), možemo zaključiti da je distribucija rezultata simetrična. Zaista, uvidom u oblik histograma primjećujemo da je oblik distribucije simetričan (slika 2.1.1).

PRIMJER 3.3

Za varijablu DOB iz primjera 2.2 odredit ćemo mod, medijanu i aritmetičku sredinu. Za određivanje mjera centralne tendencije koristit ćemo „stablo i listovi" prikaz (tabela 3.3.1).

Tabela 3.3.*1 Stem and leaf* prikaz dobi pacijenata

0	2	3	3	3	3	3	3	6	9	(9)
1	1	5	5	5						(4)
2	2									(1)
5	1	4								(2)
6	0	5								(2)
9	1	4	4	7						(4)
10	1									(1)
12	2	7								(2)
15	3	6								(2)
16	0	2								(2)
									Σ=	29

1. Mod

Iz „stablo i listovi“ prikaza odredit ćemo rezultat koji se javlja najveći broj puta. Za vrijednost stabla „0“, frekvencija je 9. Vrijednost „lista“ 3 javlja se šest puta. Možemo zaključiti da je:

Mod = 0,3

2. Medijana

U prikazu rezultati su poredani po veličini (od 0,2 do 16,2). Položaj medijane određujemo preko izraza (N+1)/2. Medijana se nalazi na 15 mjestu. Petnaesto mjesto odredit ćemo sabiranjem broja listova za pojedine vrijednosti stabla, sve dok zbir ne bude 15 ili dok ne pređemo vrijednost 15. Dakle, 9 (za 0) + 4 (za 1) + 1 (za 2) + 2 (za 5), što iznosi 16. Na 16. mjestu nalazi se vrijednost 5,4. Jedno mjesto ispod, tj. na 15. mjestu, nalazi se vrijednost 5,1. Dakle:

C = 5,1

3. Aritmetička sredina

Aritmetičku sredinu izračunat ćemo koristeći formulu: $M=\Sigma X/N$. Uvrštavanjem vrijednosti u formulu, dobit ćemo da aritmetička sredina iznosi:

M = 5,9

Izračunate vrijednosti mjera centralne tendencije se razlikuju. Najveća razlika je između moda i ostale dvije mjere. Koju ćemo mjeru centralne tendencije izabrati ako skup podataka želimo numerički opisati? Iz *stem and leaf* prikaza možemo vidjeti da je oblik distribucije asimetričan. Dakle, aritmetička sredina nije najbolji izbor. Mod također ne bi adekvatno reprezentirao skup podataka jer se vrijednost 0,3 javlja samo šest puta, što je malo u odnosu na ukupan broj podataka. Najbolji reprezentant podataka je medijana.

PRIMJER 3.4

Za visine učenika IIIc odjeljenja, potrebno je utvrditi deskripitive vrijednosti– mod, medijanu i aritmetičku sredinu. U tu svrhu koristit ćemo tabelu 3.4.1.

Tabela 3.4.1: Distribucija visina učenika IIIc odjeljenja

Visina učenika (cm)	f	rf	rf(%)	cf	rcf	rcf(%)
131	1	0,04	4	1	0,04	4
132	2	0,08	8	3	0,12	12
133	1	0,04	4	4	0,16	16
134	4	0,16	16	8	0,32	32
135	7	0,28	28	15	0,6	60
136	5	0,2	20	20	0,8	80
137	3	0,12	12	23	0,92	92
138	1	0,04	4	24	0,96	96
139	1	0,04	4	25	1	100
Σ	25	1	100			

1. *Mod*

Rezultat sa najvećom frekvencijom je X=135. Prema tome:

Mod = 135

2. *Medijana*

Obzirom da su u Tabeli 3.4.1 rezultati poredani prema veličini (od 131 cm do 139 cm), vrijednost medijane možemo utvrditi prema formuli za utvrđivanje njenog položaja u skupu podataka (N+1)/2. Medijana se nalazi na 13. mjestu u nizu podataka poredanih prema veličini. Uvidom u distribuciju kumulativnih frekvencija (cf) vidimo da rezultat 135 zauzima pozicije od 9. do 15. mjesta. Prema tome, rezultat koji se nalazi na 13. mjestu i koji predstavlja medijanu distribucije također iznosi 135 cm.

Do istog zaključka smo mogli doći i uvidom u distribuciju kumulativnih relativnih frekvencija u procentima [crf (%)]. Vidimo da je zaključno sa rezultatom 134 u distribuciji „akumulirano“ 32% rezultata. „Uključenjem“ rezultata 135 „akumulira“ se dodatnih 28%, tj. ukupno 60% rezultata. Dakle, rezultat koji dijeli distribuciju na pola (50% donjih i 50% gornjih rezultata) je upravo rezultat 135.

3. Aritmetička sredina

Aritmetička sredina iznosi:

$$M = \frac{\sum f_i X_i}{N} = \frac{3377}{25} = 135{,}08$$

Vidimo da su tri izračunate deskriptive vrijednosti praktično jednake na temelju čega zaključujemo da je riječ o distribuciji koju možemo opisati kao simetričnu.

PRIMJER 3.5

Nastavnik matematike u školi „Grbavica II“ dao je učenicima trećih razreda test iz matematike, na kojem su učenici ostvarili rezultate (izražene kao broj skupljenih bodova, pri čemu je maksimalana broj bodova na testu iznosio 50) prikazane u tabeli grupiranih rezultata (tabela 3.5.1).

Tabela 3.5.1: Distribucija grupiranih rezultata

rezultat	f	rf	rf (%)	cf	crf	crf (%)
0-4	7	0,07	7,14	7	0,07	7,14
5-9	0	0,00	0,00	7	0,07	7,14
10-14	0	0,00	0,00	7	0,07	7,14
15-19	0	0,00	0,00	7	0,07	7,14
20-24	0	0,00	0,00	7	0,07	7,14
25-29	0	0,00	0,00	7	0,07	7,14
30-34	0	0,00	0,00	7	0,07	7,14
35-39	0	0,00	0,00	7	0,07	7,14
40-44	0	0,00	0,00	7	0,07	7,14
45-49	0	0,00	0,00	7	0,07	7,14
50-54	0	0,00	0,00	7	0,07	7,14
55-59	0	0,00	0,00	7	0,07	7,14
60-64	0	0,00	0,00	7	0,07	7,14
65-69	9	0,09	9,18	16	0,16	16,33
70-74	13	0,13	13,27	29	0,30	29,59
75-79	61	0,62	62,24	90	0,92	91,84
80-84	7	0,07	7,14	97	0,99	98,98
85-89	1	0,01	1,02	98	1,00	100,00
Σ	98	1	100			

Nastavnik je odlučio izračunati mjeru centralne tendencije koja najbolje reprezentira prosječnu vrijednost. Izračunao je medijanu i aritmetičku sredinu.

1. *Medijana*

Iz gornje tabele može se zaključiti da je 50% distribucije „akumulirano“ u razredu sa graničnim vrijednostima 75 i 79 (do donje stvarne granice ovog razreda nalazi se 29,59% rezultata u distribuciji, a do njegove gornje stvarne granice 91,84%; dakle, tačka u kojoj se distribucija dijeli na pola je negdje u intervalu između rezultata 75 i 79). Na taj način dobijamo vrijednosti potrebne za formulu za određivanje centralne vrijednosti (tj. 50. centila).

$$C = 74{,}5 + \left(\frac{50 \times 98}{100} - 29 \right) \times \frac{5}{61} = 76{,}14$$

2. *Aritmetička sredina*

Kako bi izračunao aritmetičku sredinu rezultata, nastavnik je svaki pojedini razred u grupiranoj distribuciji predstavio njegovom srednjom vrijednošću, koju je potom pomnožio sa frekvencijom razreda, konačno sumirajući dobivene vrijednosti.

Tabela 3.5.2. Srednje vrijednosti i frekvencije razreda

razred	X	f	X f
0-4	2	7	14
5-9	7	0	0
10-14	12	0	0
15-19	17	0	0
20-24	22	0	0
25-29	27	0	0
30-34	32	0	0
35-39	37	0	0
40-44	42	0	0
45-49	47	0	0
50-54	52	0	0
55-59	57	0	0
60-64	62	0	0
65-69	67	9	603
70-74	72	13	936
75-79	77	61	4697
80-84	82	7	574
85-89	87	1	87
total		98	6911

Koristeći formulu:

$$M = \frac{\sum f_i X_i}{N}$$

Nastavnik je utvrdio konačnu vrijednost aritmetičke sredine: M = 70,52.

Obzirom da je medijana distribucije veća od aritmetičke sredine (76,14 u odnosu na 70,52 – razlika od skoro 6 bodova), nastavnik je zaključio da je distribucija rezultata negativno asimetrična (rezultati se u većoj mjeri grupiraju na desnoj strani distribucije; ekstremne vrijednosti se javljaju na lijevoj strani distribucije, tj. idu prema negativnom kraju x ose, odakle i odrednica – negativna asimetrija). Iz tabele grupiranih rezultata, jasno je da 93% učenika ostvaruje rezultate između 65 i 89 bodova, dok samo njih 7% ima rezultate niže od 4 boda te su oni i razlog niže aritmetičke sredine (uzrokuju da se ukupni učinak grupe čini slabijim nego što stvarno jeste). Iz tabele je također jasno da je najveći broj učenika (njih čak 62,24%) postigao rezultat između 75 i 79 bodova – što je i interval iz kojeg dolazi medijana. Dakle, zbog asimetričnosti distribucije, bolji pokazatelj postignuća učenika imaćemo ako se oslonimo na medijanu, a ne na aritmetičku sredinu.

ZADACI

1. Ispod su prikazane vrijednosti ekstraverzije za 30 učenika.

20	10	14	20	15	15	15	15	15	23
16	16	17	17	18	21	18	19	19	20
8	20	20	14	21	18	21	23	23	16

Izračunajte mod, medijanu i aritmetičku sredinu.

2. Na istoj grupi učenika utvrđene su i vrijednosti psihoticizma. Podaci su prikazani ispod.

4	0	5	0	4	1	13	1	6	2
2	2	2	2	2	2	3	3	3	3
0	4	1	4	0	5	6	1	6	1

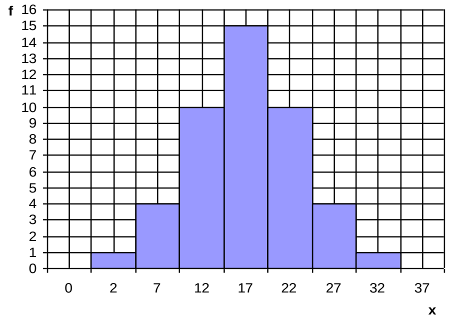
Izračunajte mod, medijanu i aritmetičku sredinu.

3. U tabeli ispod prikazana je distribucija grupiranih rezultata.

Razred	f
0-4	1
5-9	14
10-14	12
15-19	2
20-24	1

Izračunajte medijanu i aritmetičku sredinu.

4. Na slici je prikazan histogram rezultata koje je grupa učenika postigla na testu znanja iz matematike. Na osnovu histograma formirajte distibuciju grupiranih rezultata (veličina intervala i=5), a zatim, iz grupiranih podataka, odredite medijanu i aritmetičku sredinu.



5. Na osnovu distibucije grupiranih rezultata iz zadatka 3.4 grafičkim putem odredite medijanu.

6. U tabeli ispod navedene su bosanskohercegovačke rijeke i dužine njihovih tokova (u km).

Rijeka	Dužina toka (u km)
Sava	331
Una	212
Unac	58
Sana	140,4
Vrbas	192
Vrbanjka	70,5
Ukrina	53,3
Bosna	271
Krivaja	65,5
Spreča	112,3
Usora	77
Tolisa	56,2
Tinja	69
Drina	346
Ćehotina	33
Lim	40
Prača	57
Drinjača	78,5
Janja	53,3
Neretva	218
Trebišnjica	96,5

Koliko iznosi prosječna dužina toka gore navedenih rijeka? Izračunajte aritmetičku sredinu i medijanu. Koju mjeru centralne tendencije biste izabrali za odgovor na pitanje o prosječnoj dužini toka gore navedenih rijeka?

7. U tabeli ispod navedena su površine bosanskohercegovačkih jezera (u km2).

Jezero	Površina (u km^2)
Buško	55,8
Blidnje	3,2
Boračko	0,26
Jablaničko	13,3
Modrac	17,1
Plivsko	1,15
Perućačko	12,4
Ramsko	15,3
Zvorničko	8,1

Koliko iznosi prosječna površina gore navedenih jezera?

8. Izračunajte mod, medijanu i aritmetičku sredinu za podatke iz zadatka 2.1. Šta na osnovu izračunatih vrijednosti mjera centralne tendencije možete zaključiti o distribuciji rezultata? Koja mjera centralne tendencije najbolje reprezentira skup podataka?

9. Izračunajte mjere centralne tendencije za podatke iz zadatka 2.2.

10. Izračunajte medijanu i aritmetičku sredinu za podatke iz zadatka 2.3.

11. Koristeći „stablo i listovi“ prikaz iz zadatka 2.6 odredite mod i medijanu za žene i muškarce.

12. Nastavnik fizike u OŠ „Sigmund Freud“ je učenicima šestih razreda (uključujući i VIa razred) zadao test. Učenici su ostvarili sljedeće rezultate (izražene kao broj skupljenih bodova; maksimalan broj bodova iznosio je 50):

Broj bodova na testu iz fizike	f
0	4
1	2
3	1
25	2
26	1
27	1
28	1
29	4
30	4
33	4
34	5
35	7
36	10
37	19
38	19
39	6
40	4
41	3
45	1
total	98

a. Utvrdite vrijednost moda za gornju distribuciju.
b. Grupirajte distribuciju frekvencija gornjih rezultata te izračunajte vrijednost medijane i aritmetičke sredine.
c. Načinite grafički prikaz distribucije.
d. Opišite distribuciju (posebno obratite pažnju na izbor mjere centralne tendencije koja najpreciznije reprezentira distribuciju).

13. Nastavno vijeće OŠ „Sigmund Freud" je roditeljima učenika šestih razreda poštom poslalo anketni upitnik sa ciljem da prikupi preciznije informacije o socijalno-ekonomskom statusu porodica iz kojih dolaze njihovi učenici. Upitnik je sadržavao i pitanja o mjesečnim primanjima svakog od roditelja učenika (pri tome je roditeljima garantirana stroga diskrecija datih odgovora). Ispod su prikazani prikupljeni odgovori o mjesečnim primanjima majki učenika, zaokruženi na jedinice od 50 KM (N=98).

Mjesečna primanja majki učenika (KM)	f	Mjesečna primanja majki učenika (KM)	f
350	1	950	7
400	1	1.000	4
450	3	1.050	5
500	4	1.100	3
550	1	1.150	5
600	3	1.200	5
650	4	1.250	7
700	6	1.350	3
750	6	1.400	3
800	2	7.550	1
850	9	8.000	1
900	13	10.600	1

a. Utvrdite vrijednost moda, medijane i aritmetičke sredine za gornju distribuciju.
b. Načinite grafički prikaz distribucije.
c. Opišite distribuciju (posebno obratite pažnju na izbor mjere centralne tendencije koja najpreciznije reprezentira distribuciju; obrazložite svoj izbor).

14. Izračunajte mjere centralne tendencije za podatke iz zadatka 2.8.

15. Izračunajte mjere centralne tendencije za podatke iz zadatka 2.9.

16. Izračunajte mjere centralne tendencije za podatke iz zadatka 2.10.

17. Izračunajte mjere centralne tendencije za podatke iz zadatka 2.11.

18. Izračunajte mjere centralne tendencije za podatke iz zadatka 2.12.

19. Ispitivanjem znanja iz Statistike u psihologiji na grupi od 30 studenata dobiveni su sljedeći rezultati:

121	132	119	126	110	122	121	130	113	123
114	123	117	124	120	129	114	123	117	124
119	125	116	124	118	125	120	128	119	127

Izračunajte mjere centralne tendencije.

20. Kako bi se ispitao efekat supstance QWR11 na psihomotorne sposobnosti, ispitanicima su najprije utvrđene psihomotorne sposobnosti, zatim su primili odgovarajuću dozu QWR11, te su ponovo izmjerene psihomotorne sposobnosti. Rezultati dva mjerenja prikazani su ispod.

Ispitanik	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Prije	43	20	17	30	25	19	34	28	23	41	26	40	16	23	34
poslije	45	16	20	33	30	19	33	25	26	40	28	36	15	26	32

Grafički prikažite distribucije rezultata. Izračunajte mjere prosjeka.

4. Mjere varijabiliteta

Pored tendencije grupiranja podataka oko neke srednje vrijednosti, postoji i **tendencija variranja tj. raspršenja** podataka oko srednje vrijednosti. Stoga u numeričkoj deskripciji skupa podataka, pored mjere centralne tendencije koristimo i mjere varijabiliteta. Sljedeći primjer ilustrira nužnost da pored mjere centralne tendencije treba odrediti i mjeru varijabiliteta.

U skupovima podataka A i B Mod = C = M = 70.

A:	*53*	*62*	*70*	*70*	*75*	*78*	*82*
B:	*10*	*20*	*50*	*70*	*70*	*120*	*150*

Numerička deskripcija na temelju mjere centralne tendencije može nas navesti na pogrešan zaključak da se radi o sličnim skupovima. Zapravo, podaci se prilično razlikuju. Raspon rezultata skupa B znatno je veći od raspona rezultata skupa A.

Koristimo različite **mjere varijabiliteta:** broj kategorija, raspon rezultata, inerkvartilni raspon, varijancu i standardnu devijaciju.

Raspon rezultata je razlika između najveće i najmanje vrijednosti u skupu podataka.

$$\text{Raspon} = X_{max} - X_{min}$$

Za skup podataka: x=(53,62,70,70,75,78,82), razlika između najvišeg i najnižeg rezultata je: 82 – 53 = 29. Dakle, raspon rezultata iznosi 29.

Na osnovu raspona rezultata ne možemo saznati ništa o raspodjeli podataka unutar raspona jer se raspon računa samo preko ekstremnih vrijednosti.

U skupovima podataka A i B raspon je jednak i iznosi 70. U skupu B veći broj rezultata grupiran je oko centralne vrijednosti.

A:	*10*	*20*	*50*	*60*	*70*	*80*
B:	*10*	*50*	*50*	*50*	*50*	*80*

Interkvartilni raspon (ili raspršenje) je raspon u kojem se nalazi 50% središnjih rezultata. Za razliku od raspona rezultata interkvartilni raspon nije osjetljiv na ekstremne rezultate. Za određivanje kvartilnog raspršenja, podatke treba poredati po veličini, od najmanjeg do najvećeg. Zatim se odrede rezultati koji distribuciju dijele na četiri jednaka dijela. Prvi takav rezultat naziva se **prvi kvartil (označava se sa Q_1)** i distribuciju dijeli tako da je ispod ¼ (tj. 25%) najnižih rezultata, a iznad ¾ (75%) viših rezultata. Drugi utvrđeni rezultat distribuciju dijeli na dva jednaka dijela (centralna vrijednost). Treći rezultat naziva se **treći kvartil (označava sa Q_3)**, i distibuciju dijeli tako da je ispod ovog rezultata ¾, tj 75% nižih rezultata, a iznad ¼, tj. 25% najviših rezultata.

Interkvartilini raspon izračunava se preko razlike Q_3 i Q_1, tj:

$$IQR = Q_3 - Q_1$$

Poluinterkvartilni raspon izračunava se kada kvartilni raspon podijelimo sa dva**:**

$$SQR = (Q_3 - Q_1)/2,$$

i predstavlja grubu procjenu prosječne udaljenosti između medijane i prvog kvartila i između medijane i trećeg kvartila.

Dati su skupovi podataka A i B.

A: 1 2 5 6 7 8

B: 1 3 4 5 5 7 8

A) Skup ćemo podijeliti na dva jednaka dijela.
Prva polovina: 1, 2, 5; druga polovina: 6, 7, 8.

Prvi kvartil je medijana prve polovine, Q_1=2; treći kvartil je medijana druge polovine, Q_3=7. Interkvartilni raspon jednak je: QR=7-2=5; SQR=2,5

B) Skup ćemo podijeliti na dva dijela (medijana skupa je isključena):
Prva polovina: 1, 3, 4; druga polovina: 5, 7, 8.

Q_1=3; Q_3=7. Interkvartilni raspon jednak je: QR=7-3=4; SQR=2

Standardna devijacija i varijanca su mjere varijabiliteta koje se temelje na udaljenostima svakog rezultata od aritmetičke sredine. **Varijanca** skupa podataka je prosjek kvadriranih odstupanja rezultata od aritmetičke sredine:

$$v = \frac{\sum (X_i - M)^2}{N - 1}$$

Standardna devijacija jednaka je drugom korijenu iz varijance:

$$s = \sqrt{\frac{\sum (X_i - M)^2}{N - 1}}$$

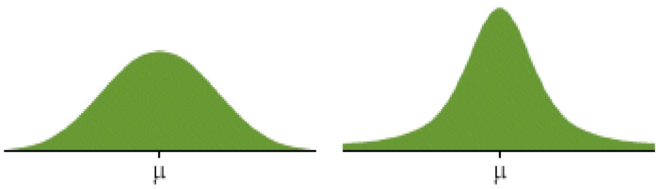
Za skup podataka: x=(1,2,3,5,6,7), varijanca i standardna devijacija iznose:

$$v = \frac{\sum (X_i - M)^2}{N - 1} = \frac{(1-4)^2 + (2-4)^2 + (3-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2}{6-1} = 5{,}6$$

v=5,6; s=2,37

Standardnu devijaciju statistički je opravdano računati kada je: distribucija rezultata normalna ili barem simetrična i kada je korištena intervalna ili racio skala mjerenja.

Varijabilnost podataka utiče na spljoštenost distribucije. Što je raspršenje veće, spljoštenost je veća i obratno - što je raspršenje manje, spljoštenost je manja. S obziron na raspršenost, distribucije mogu biti **platokurtične** (spljoštene) i **leptokurtične** (izdužene).



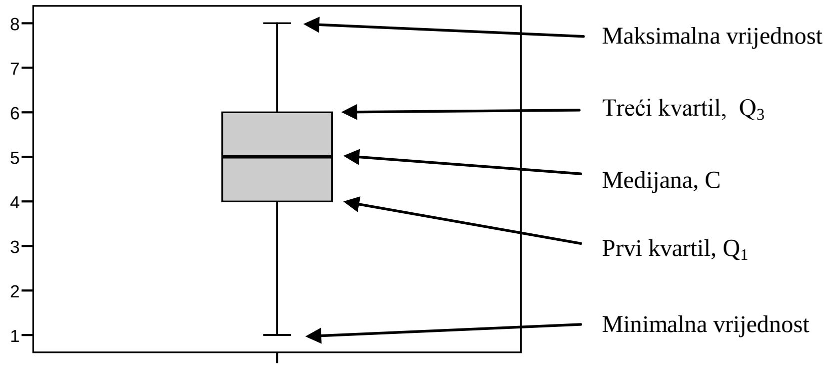
Platokurtična leptokurtična

Kao mjeru spljoštenosti (ili izduženosti) distribucije koristimo **kurtozis** (eng. *kurtosis*):

$$\text{kurtozis} = \frac{N(N-1)}{(N-2)(N-3)} \times \frac{\sum (X_i - M)^4}{(N-1) \times s^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

Ako je vrijednost kurtozisa pozitivna, distribucija je leptokurtična. Ako je vrijednost kurtozisa negativna, distribucija je platokurtična. Ako je vrijednost kurtozisa 0, distribucija je simetrična. Kako je to slučaj i kod skjunisa, sama vrijednost kurtozisa ne ukazuje na odstupanja oblika distribucije. Naime, potrebno je odrediti standardnu pogrešku kurtozisa!

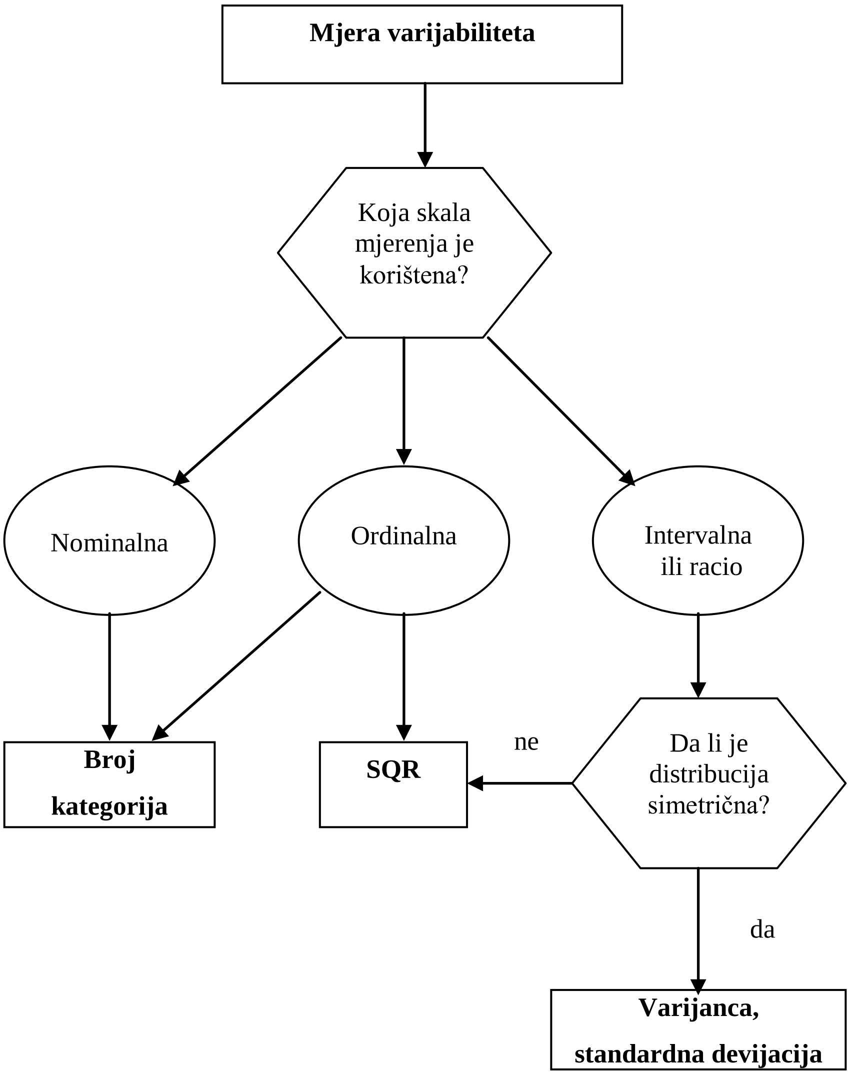
Koristan način numeričkog sumiranja podataka je navođenje **pet vrijednosti** koje pružaju važne informacije o distribuciji. Te vrijednosti su: medijana, prvi kvartil, treći kvartili, minimalna i maksimalna vrijednost. Ovaj način numeričke deskripcije podakata naziva se „**sumiranje sa pet brojeva**“. Sumiranje skupa podataka sa pet brojeva može se transformisati u grafički prikaz kojeg nazivamo „**box-plot“** prikaz.



Krajevi pravougaonika predstavljaju kvartile, a linija koja dijeli pravougaonik medijanu. U slučaju kada je distribucija rezultata simetrična (kao što je ovdje slučaj) linija koja predstavlja medijanu dijeli pravougaonik na dva jednaka dijela. Minimalni i maksimalni rezultat označeni su krajevima linija okomitih na pravougaonik.

S obzirom da box-plot prikaz na jednostavan i ekonomičan način pruža uvid u oblik distribucije (simetričnost i spljoštenost), predstavlja popularan način sumiranja rezultata dobivenih u istraživanju.

Kao i kod mjera centralne tendencije**, izbor mjera varijabiliteta** zavisi od skale mjerenja i oblika distribucije. Ako je korištena nominalna skala mjerenja kao mjeru varijabiliteta koristimo broj kategorija. Kod podataka dobivenih ordinalnom skalom mjerenja koristimo poluinterkvartilni raspon (SQI). Kada su podaci dobiveni na intervalnoj ili racio skali mjerenja, kao mjeru varijabiliteta koristimo varijancu, tj. standardnu devijaciju (ali samo ukoliko je distribucija simetrična); u protivnom koristimo poluinterkvartilni raspon. Ispod je dat shematski prikaz izbora mjera varijabiliteta.


Mjera varijabiliteta
Koja skala mjerenja je korištena?
Nominalna
Ordinalna
Intervalna ili racio
Broj kategorija
SQR
ne
Da li je distribucija simetrična?
da
Varijanca, standardna devijacija

PRIMJER 4.1

Za podatke iz primjera 1.3 izračunat ćemo interkvartilni raspon i standardnu devijaciju, a skup podataka opisat ćemo sa „pet brojeva“. Rezultati (broj tačnih odgovora na Testu znanja iz statistike) prikazani su ispod.

24	27	32	20	20	15	20	20	19	22
18	22	27	28	15	20	14	24	24	19

Kao i kod izračunavanja mjera centralne tendencije, podatke ćemo najprije urediti koristeći distribuciju negrupiranih rezultata (prikazanu u tabeli ispod).

rezultat	f
14	1
15	2
18	1
19	2
20	5
22	2
24	3
27	2
28	1
32	1
total	20

a. *Interkvartilni raspon*

Skup podataka sastoji se od 20 rezultata. Ovaj skup podijelit ćemo na dva jednaka dijela:

I dio: 14, 15,15,18,19,19, 20, 20, 20, 20;

II dio: 20, 22, 22, 24, 24, 24, 27, 27, 28, 32

Rezultat koji distribuciju prvog dijela skupa dijeli na dva jednaka dijela iznosi 19 – to je prvi kvartil; rezultat koji drugi dio skupa dijeli na dva jednaka dijela iznosi 24 – to je treći kvartil. Odnosno, Q_1=19; Q_3=24. Interkvartilni raspon iznosi:

$$QR = Q_3 - Q_1 = 24 - 19 = 5$$

$$IQR = 5$$

b. *Standardna devijacija*

Standardnu devijaciju odredit ćemo preko izraza:

$$s = \sqrt{\frac{\sum (X_i - M)^2}{N-1}}$$

S obzirom da se pojedini rezultati pojavljuju više puta, gornjem izrazu dodat ćemo vrijednosti frekvencija i dobiti sljedeću formulu:

$$s = \sqrt{\frac{\sum f_i X^2}{\sum f_i} - M^2}$$

Odredit ćemo vrijednosti koje treba uvrstiti u gornji izraz.

rezultat	f	fX	f X^2
14	1	14	196
15	2	30	450
18	1	18	324
19	2	38	722
20	5	100	2000
22	2	44	968
24	3	72	1728
27	2	54	1458
28	1	28	784
32	1	32	1024
	$\Sigma f_i=20$	$\Sigma f_i X=430$	$\Sigma f_i X^2=9654$

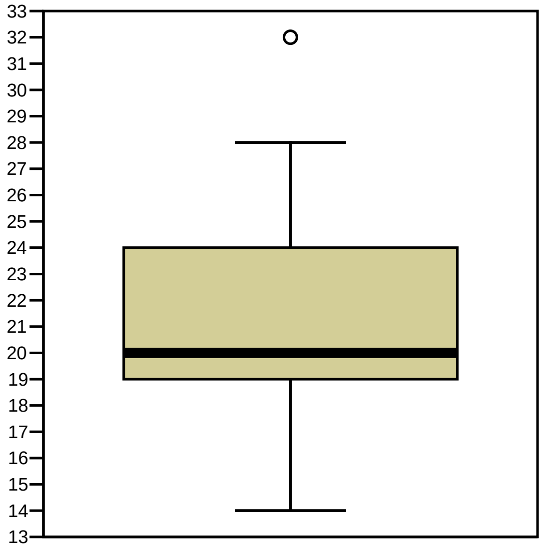
Standardna devijacija iznosi:

$$s = 4{,}43$$

c. *Pet brojeva*

Pet brojeva: X_{min}=14; X_{max}=32; C=20; Q_1=19; Q_3=24

Ispod je dat box-plot prikaz. Iz grafičkog prikaza možemo pročitati gore navedene vrijednosti, ali i neke informacije o distribuciji rezultata. Najprije, uočit ćemo kružić, kojim se označava ekstremna vrijednost. Ekstremne vrijednosti potrebno je posebno tretirati, te je njihova detekcija u deskripciji rezultata izuzetno važna. Linija koja označava medijanu smještena je znatno bliže prvom kvartilu i očigledno je da distribucija nije simetrična. Od prvog kvartila do medijane udaljenost je: C-Q_1=20-19=1. Udaljenost od medijane do trećeg kvartila iznosi: Q_3-C=24-20=4. U rasponu od 19 do 20 nalazi se 25% rezultata (ispod medijane), a isti broj rezultata nalazi se u rasponu od 20 do 24 (iznad medijane). Dakle, raspon od 4 obuhvata jednak broj rezultata kao i raspon od 1.



PRIMJER 4.2

Za podatke iz primjera 3.1 odredite standardnu devijaciju i interkvartilni raspon.

razred	f	X	f X	f X^2
6-9	2	7,5	15	112,5
10-13	5	11,5	57,5	661,3
14-17	22	15,5	341	5286
18-21	40	19,5	780	15210
22-25	22	23,5	517	12150
26-29	6	27,5	165	4538
30-33	3	31,5	94,5	2977
	Σf_i =100			$\Sigma f_i\ X^2$ =40933

Vrijednosti $\Sigma f_i\ X^2$ i Σf_i uvrstit ćemo u formulu i izračunati standardnu devijaciju.

$$s = \sqrt{\frac{\sum f_i X^2}{\sum f_i} - M^2}$$

s= 4,61

Interkvartilni raspon iz grupiranih podataka možemo odrediti na dva načina: računskim postupkom ili očitavanjem iz grafičkog prikaza.

Za izračunavanje prvog i trećeg kvartila potrebno je formirati distribucije kumulativne i procentualne relativne kumulativne frekvencije, što je i urađeno u tabeli ispod.

razred	f	cf	rcf(%)
6-9	2	2	2%
10-13	5	7	7%
14-17	22	29	29%
18-21	40	69	69%
22-25	22	91	91%
26-29	6	97	97%
30-33	3	100	100%
total	100		

Prvi i treći kvartil izračunat ćemo primjenom sljedeće formule:

$$X = D + \left(\frac{\text{centil} \times N}{100} - f_D \right) \times \frac{i}{f_R}$$

odnosno,

$$Q_1 = D + \left(\frac{25 \times N}{100} - f_D \right) \times \frac{i}{f_R} \qquad Q_3 = D + \left(\frac{75 \times N}{100} - f_D \right) \times \frac{i}{f_R}$$

gdje je:

Q_1, Q_3– traženi rezultati (prvi i treći kvartil)

D – prava donja granica razreda u kojem je prvi i treći kvartil,

f_D – ukupan broj rezultata ispod razreda u kojem je prvi i treći kvartil,

f_R – broj rezultata u razredu u kojem je prvi i treći kvartil,

i – interval.

Prvi kvartil nalazi se u razredu 14-17. Prava donja granica ovog razreda je D=13,5; u razredu se nalazi f_R=22 rezultata; do razreda 14-17 ukupno je f_D =7 rezultata; interval razreda iznosi i=4.

$$Q_1 = 13,5 + \left(\frac{25 \times 100}{100} - 7 \right) \times \frac{4}{22} = 16,77$$

Treći kvartil nalazi se u razredu 22-25. Prava donja granica ovog razreda je D=21,5; u razredu se nalazi f_R=22 rezultata; do razreda 14-17 ukupno je f_D =69 rezultata; interval razreda iznosi i=4.

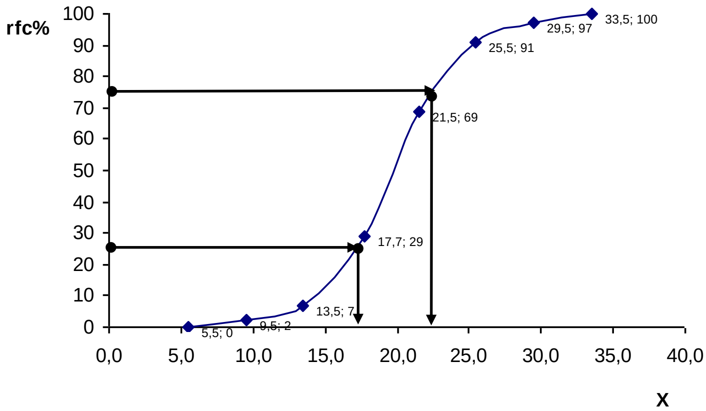
$$Q_3 = 21,5 + \left(\frac{75 \times 100}{100} - 69 \right) \times \frac{4}{22} = 22,23$$

Interkvartilni raspon iznosi:

IQR=22,23-16,77; IQR=5,45

Udaljenost od medijane do prvog kvartila iznosi: 19,6 – 16,77 = 2,83, slično kao i udaljenost od medijane do trećeg kvartila: 22,23 – 19,6 = 2,63. Ovakve vrijednosti mogli smo i očekivati obzirom na simetričnost distribucije. U slučaju idealno simetrične distribucije, ove dvije udaljenosti bile bi identične.

Grafičko određivanje kvartila radimo pomoću procentualne ogive (slika ispod). Obzirom da prvi kvartil dijeli distibuciju na dva dijela, tako da je ispod 25%, a iznad 75% rezultata, potrebno je na apscisi (x) očitati rezultat koji odgovara 25-om procentu na ordinati (rcf %). Slično tome, kako treći kvartil dijeli distibuciju na dva dijela, tako da je ispod 75%, a iznad 25% rezultata, potrebno je na apscisi (x) očitati rezultat koji odgovara 75 procentu na ordinati (rcf %). Na apscisi očitavamo da je Q_1=16,7 i Q_3=22,2. Kao i kod određivanja medijane, čitav postupak radi se na milimetarskom papiru na kojem možemo očitati tačnu vrijednost.



PRIMJER 4.3

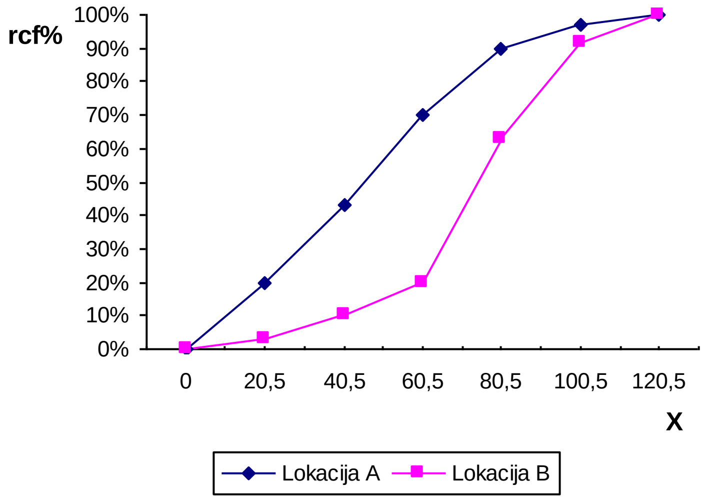
Na dvije lokacije šireg gradskog područja bilježene su brzine 70 automobila (mjerenja su urađena u istom periodu dana). U tabeli ispod prikazani su dobiveni podaci.

Brzina (km/h)	f_A	f_B
0 - 20	14	2
21 - 40	18	5
41 - 60	20	7
61 - 80	13	30
81 - 100	5	20
101 - 120	2	6
totali	70	70

U ovom primjeru želimo numerički opisati podatke. Stoga se nužno nameću dva pitanja, a to su: (1) Koje mjere centralne tendencije i varijabiliteta je potrebno odrediti? i (2) Šta možemo zaključiti na osnovu dobivenih rezultata?

Prvo što uočavamo iz tabelarnog prikaza jeste da su distribucije frekvencija asimetrične, pa je u ovom slučaju opravdano izračunati medijanu, uz koju se kao mjera varijabiliteta izračunava interkvartilni raspon.

Iz grafičkog prikaza procentualnih ogiva odredit ćemo medijanu, prvi i treći kvartil za dvije lokacije.



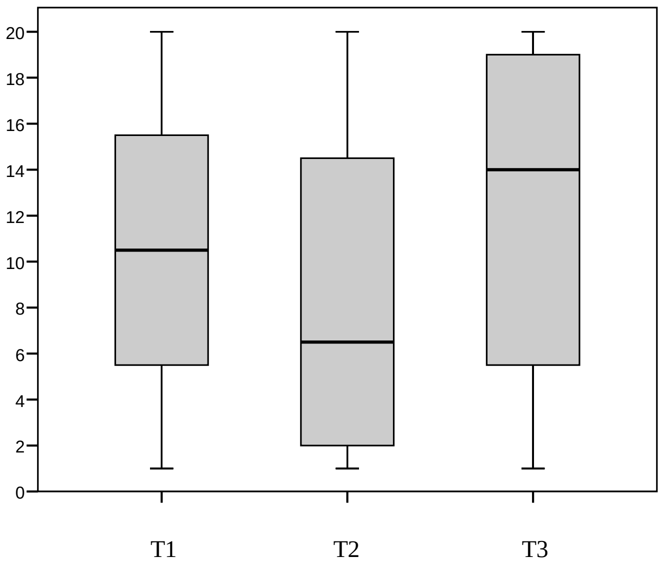
Vrijednosti medijane, prvog i trećeg kvartila koje smo očitali iz grafika navedene su u tabeli ispod.

	A	B
Q1	25,1	62,9
C	46,0	75,2
Q3	65,8	89,4

Prosječna brzina automobila na lokaciji B (C=75,2 km/h) veća je od prosječne brzine na lokaciji A (C=46,0 km/h). Međutim, interkvartilni raspon znatno je veći na lokaciji A (25,1 km/h – 65,8 km/h) u odnosu na lokaciju B (62,9 km/h – 89,4 km/h). Dakle, automobili na lokaciji A se u prosjeku kreću sporije, ali uz veću varijaciju brzina, za razliku od lokacije B, gdje se automobili kreću brže, ali uz manju varijaciju brzina.

PRIMJER 4.4

Na slici ispod dati su box-plot prikazi rezultata koje je jedna grupa ispitanika postigla na tri testa (T1, T2 i T3). Svaki test sadrži 20 zadataka, tako da je ispitanik mogao osvojiti maksimalno 20 bodova. Na osnovu grafičkog prikaza odredit ćemo težine testova.



Težinu testa možemo odrediti prema broju ispitanika koji tačno rješavaju zadatke. Što je veći broj ispitanika koji tačno rješavaju zadatke, to je test lakši, i obratno, što je manji broj ispitanika koji tačno rješavaju zadatke, to je test teži. Kod testa prosječne težine, podjednak je broj ispitanika ispod i iznad prosječne vrijednosti.

Kako bi odredili težine testova, posmatrat ćemo razlike u distribuciji rezultata na tri testa, te utvrditi vrijednosti medijane, prvog i trećeg kvartila. Iz box-plot prikaza očitat ćemo vrijednosti medijane te prvog i trećeg kvartila.

	T1	T2	T3
Q1	5,25	2,0	5,25
C	10,5	6,5	14,0
Q3	15,75	14,75	19,0

Primjećujemo da su udaljenosti između medijane i prvog, odnosno trećeg kvartila različite za T1, T2, T3.

Ukoliko je Q_3–C≈C–$Q_{1,}$ tada je distribucija simetrična (ili gotovo pa simetrična) i box-plot podataka izgleda kao što je prikazano na slici pod T1. Linija koja označava medijanu nalazi se tačno na sredini pravougaonika. Izačunavanjem udaljenosti između medijane i prvog te drugog kvartila uvjerit ćemo se da su jednake (5,25). Ispod i iznad medijane jednak je broj rezultata.

Ukoliko je Q_3–C>C–Q_1, distribucija podataka je pozitivno asimetrična. Udaljenost od medijane do trećeg kvartila (8,25) veća je u odnosu na udaljenost između medijane i prvog kvartila (4,25). Izgleda kao da je gornji dio distribucije (uslovno rečeno, pozitivan kraj) razvučen.

Ukoliko je Q_3–C<C–Q_1, za distribuciju kažemo da je negativno asimetrična. Distribucija rezultata na testu 3 je negativno asimetrična. Udaljenost od medijane do prvog kvartila (8,75) veća je u odnosu na udaljenost između medijane i trećeg kvartila (5). Izgleda kao da je donji dio distribucije (uvjetno rečeno, negativan kraj) razvučen.

I dužine linija koje spajaju pravougaonik sa graničnikom koji predstavlja najviši, odnosno najniži rezultat, ukazuju na simetričnost tj. asimetričnost rezultata. Ako je gornja linija duža od donje, vjerovatno se radi o pozitivno asimetričnoj distribuciji. Ukoliko je donja linija duža od gornje, onda je distribucija rezultata vjerovatno negativno asimetrična. Upravo takve odnose uviđamo i u našem primjeru.

Na osnovu distribucije rezultata možemo zaključiti da je test 2 teži u odnosu na test 3, dok za test 1 možemo tvrditi da je prosječne težine. Na testu 2 postignut je veći broj nižih rezultata, dok je na testu 3 postignut veći broj viših rezultata.

PRIMJER 4.5

Za utvrđene visine učenika našeg IIIc odjeljenja, nakon što su utvrđene mjere centralne tendencije, potrebno je utvrditi i pokazatelje varijabiliteta. Obzirom da je riječ o distribuciji koja se svojim oblikom približava simetričnoj, u ovom slučaju mogu se izračunati sve mjere varijabiliteta. Mi smo se odlučili za utvrđivanje interkvartilnog i poluinterkvartilnog raspona, standardne devijacije te varijance. U tu svrhu ponovo će nam poslužiti tabelarni prikaz distribucije visina učenika:

Visina učenika (cm)	f	rf(%)	cf	rcf(%)
131	1	4	1	4
132	2	8	3	12
133	1	4	4	16
134	4	16	8	32
135	7	28	15	60
136	5	20	20	80
137	3	12	23	92
138	1	4	24	96
139	1	4	25	100
Σ	25	100		

Iako se radio o negrupiranoj distribuciji rezultata, postupak računanja kvartila praktično je isti kao i kod grupiranih frekvencija rezultata.

Za određivanje poluinterkvartilnog raspršenja najprije ćemo odrediti Q1 i Q3, koristeći sljedeće izraze:

$$Q_1 = D + \left(\frac{25 \times N}{100} - f_D\right) \times \frac{i}{f_R} \qquad Q_3 = D + \left(\frac{75 \times N}{100} - f_D\right) \times \frac{i}{f_R}$$

$$Q_1 = 133{,}5 + \left(\frac{25 \times 25}{100} - 4\right) \times \frac{1}{4} \qquad Q_3 = 135{,}5 + \left(\frac{75 \times 25}{100} - 15\right) \times \frac{1}{5}$$

Q_1 = 134,06; Q_3 = 136,25.

Interkvartilni raspon iznosi:

$$IQR = Q_3 - Q_1$$

$$IQR = 136{,}25 - 134{,}06 = 2{,}19$$

Poluinterkvartilni raspon iznosi: SIQR= IQR/2=2,19/2=1,09.

Interkvartilni raspon obuhvata 50% središnjih rezultata. U našem primjeru, 50% učenika visoki su između 134,06 i 136,25.

Standardnu devijaciju distribucije visina učenika ćemo izračunati prema izrazu:

$$s = \sqrt{\frac{\sum f_i X^2}{\sum f_i} - M^2}$$

Od ranije nam je poznato da aritmetička sredina distribucije iznosi 135,08 cm. Tabelarnom prikazu distribucije podataka dodat ćemo stupce koji će nam poslužiti za određivanje standardne devijacije:

Visina učenika (cm)	f	X - M	$(X\text{-}M)^2$	$(X\text{-}M)^2xf$
131	1	-4,08	16,65	16,65
132	2	-3,08	9,49	18,97
133	1	-2,08	4,33	4,33
134	4	-1,08	1,17	4,67
135	7	-0,08	0,01	0,04
136	5	0,92	0,85	4,23
137	3	1,92	3,69	11,06
138	1	2,92	8,53	8,53
139	1	3,92	15,37	15,37
Σ	25			83,84

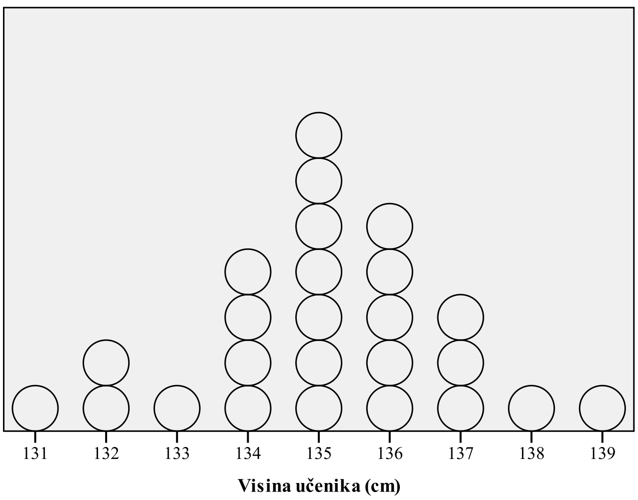
Standardna devijacija iznosi:

$$s = 1,83 \text{ cm.}$$

Varijanca visina učenika IIIc oko njihove aritmetičke sredine iznosi:

$$v = s^2 = 1,83^2 = 3,35 \text{ cm}$$

Jasniju „predstavu“ o varijanci rezultata oko njihove aritmetičke sredine, tj. o izraženosti variranja vrijednosti u distribuciji može nam pružiti i grafički prikaz. Na slici ispod svaki od učenika IIIc odjeljenja predstavljen je kružićem.



Na slici još jednom vidimo da se distribucija može smatrati simetričnom te da je variranje rezultata relativno malo – vidimo da je u distribuciji zastupljeno malo razl<u>ičitih</u> rezultata (samo 9 mogućih vrijednosti) koji se kreću u razmjerno malom rasponu.

PRIMJER 4.6

Učitelj u OŠ „Grbavica II" želi utvrditi varijabilitet (preciznije, interkvartilni i poluinterkvartilni raspon, standardnu devijaciju i varijancu) distribucije rezultata koje su učenici trećih razreda postigli na testu znanja iz matematike (vidi primjer 3.5). U tu svrhu nastavnik je načinio sljedeći tabelarni prikaz distribucije:

razred	f	cf	crf (%)	Sr. vr. razreda (X)	X^2	$f * X^2$
0-4	7	7	7	2	4	28
5-9	0	7	7	7	49	0
10-14	0	7	7	12	144	0
15-19	0	7	7	17	289	0
20-24	0	7	7	22	484	0
25-29	0	7	7	27	729	0
30-34	0	7	7	32	1024	0
35-39	0	7	7	37	1369	0
40-44	0	7	7	42	1764	0
45-49	0	7	7	47	2209	0
50-54	0	7	7	52	2704	0
55-59	0	7	7	57	3249	0
60-64	0	7	7	62	3844	0
65-69	9	16	16	67	4489	40401
70-74	13	29	30	72	5184	67392
75-79	61	90	92	77	5929	361669
80-84	7	97	99	82	6724	47068
85-89	1	98	100	87	7569	7569
Σ	98					524127

Za izračunavanje interkvartilnog i poluinterkvartilnog raspršenja koristit ćemo sljedeće izraze:

$$Q_1 = D + \left(\frac{25 \times N}{100} - f_D \right) \times \frac{i}{f_R} \qquad Q_3 = D + \left(\frac{75 \times N}{100} - f_D \right) \times \frac{i}{f_R}$$

Iz kolone rcf (%) očitavamo da se 25% distribucije „formiralo“ u razredu 70 – 74, a 75% distribucije odmah u narednom razredu 75 – 79 (iz čega je odmah jasno da nije riječ o simetričnoj distribuciji). Prema tome:

$$Q_1 = 69{,}5 + \left(\frac{25 \times 98}{100} - 16 \right) \times \frac{5}{13} = 72{,}77$$

$$Q_3 = 74{,}5 + \left(\frac{75 \times 98}{100} - 29 \right) \times \frac{5}{61} = 78{,}15$$

Interkvartilno raspršenje iznosi:

$$IQR = Q_3 - Q_1 = 78{,}15 - 72{,}77 = \underline{5{,}38}$$

Poluniterkvartilno raspršenje iznosi:

$$SQR = IQR/2 = 2{,}69$$

Za izračunavanje standardne devijacije može se upotrijebiti formula:

$$s = \sqrt{\frac{\sum f_i X^2}{\sum f_i} - M^2} = \sqrt{\frac{524127}{98} - 70{,}52^2} = 19{,}37 \text{ (bodova)}$$

Aritmetička sredina poznata je od ranije (vidi poglavlje 3) i iznosi 70,52 bodova.

Varijanca je jednaka kvadratu vrijednosti standardne devijacije, tj:

$$v = 19{,}37^2 = 375{,}16 \text{ (bodova)}$$

ZADACI

1. Izračunajte interkvartilni raspon i standardnu devijaciju za podatke iz zadatka 3.1.
2. Izračunajte interkvartilni raspon i standardnu devijaciju za podatke iz zadatka 3.2.
3. Izračunajte interkvartilni raspon i standardnu devijaciju za podatke iz zadatka 3.3.
4. Izračunajte interkvartilni raspon i standardnu devijaciju za podatke iz zadatka 3.4.
5. Na osnovu distribucije grupiranih rezultata iz zadatka 4. grafičkim putem odredite interkvartilni raspon.
6. Izračunajte interkvartilni raspon i standardnu devijaciju za podatke iz zadatka 2.1.
7. Izračunajte interkvartilni raspon i standardnu devijaciju za podatke iz zadatka 2.3.
8. Izračunajte interkvartilni raspon i standardnu devijaciju za podatke iz zadatka 2.6.
9. Ispod su dati rezultati testiranja znanja iz matematike grupe od 20 učenika.

A	7	1	14	1	36	6	2	16	4	2
	1	10	12	13	1	15	3	17	19	2
B	13	12	12	1	14	25	15	19	16	19
	17	20	18	15	16	19	17	25	14	25

Najprije odredite koju mjeru varijabiliteta treba izračunati (s obzirom na distribucije rezultata), a zatim ih izračunajte.

10. Na grupi od 20 studenata primjenjen je upitnik kojim se mjeri sramežljivost. Dobiveni rezultati prikazani su ispod.

1	1	1	1	2	2	2	2	4	6
7	10	10	13	13	15	16	16	17	40

Izračunajte aritmetičku sredinu i standardnu devijaciju.

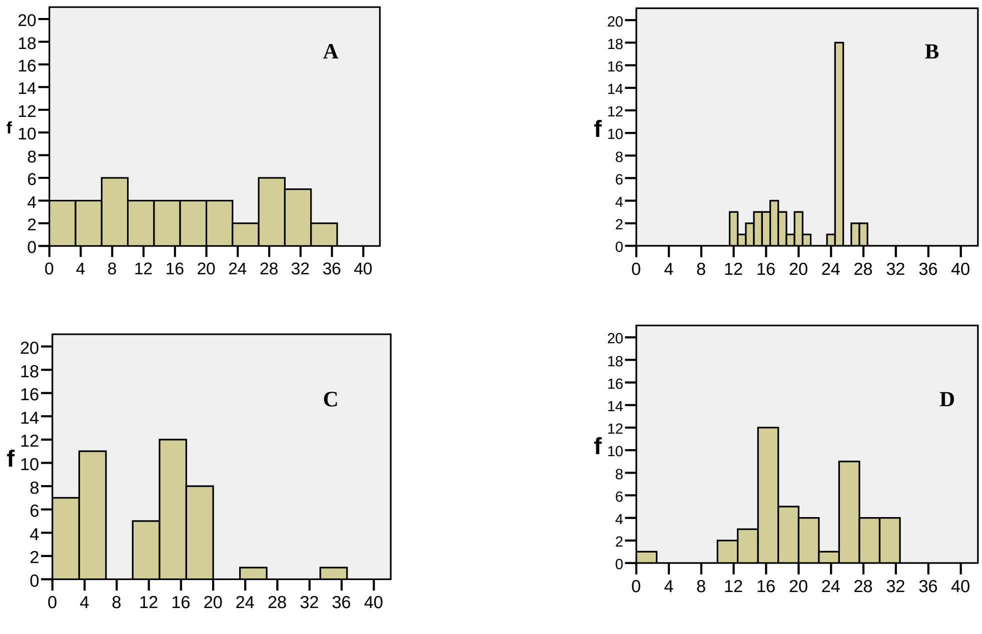
U skupu podataka rezultat 40 je ekstremna vrijednost. Izračunajte aritmetičku sredinu i standardnu devijaciju, ali nakon što ste iz skupa podataka isključili rezultat 40. Šta primjećujete? Na koji način ekstremne vrijednosti utječu na aritmetičku sredinu i standardnu devijaciju?

11. Za pet grupa podataka izračunate su vrijednosti prvog kvartila, centralne vrijednosti i trećeg kvartila. Na osnovu prikazanih vrijednosti, šta možete zaključiti o obliku distribucije? Da li su distribucije simetrične, ako nisu o kojoj se simetriji radi? U kojoj grupi podataka je najveća varijabilnost rezultata?

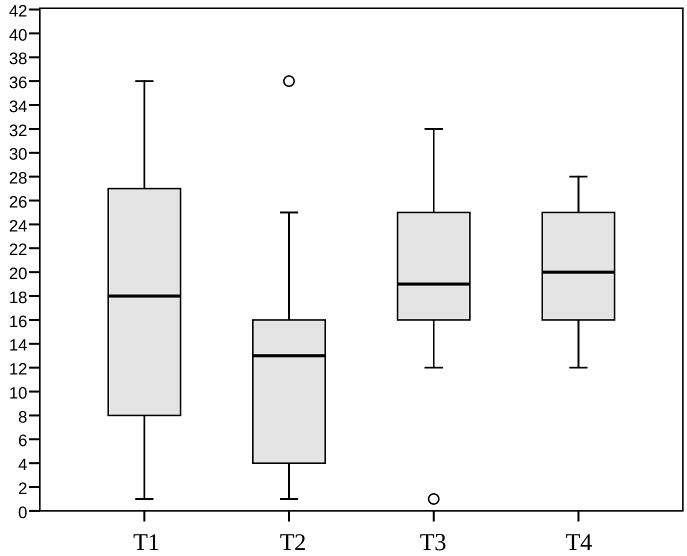
	I	II	III	IV	V
Q_1	3	4	5	7	1
C	10	6	6	8	10
Q_3	13	8	15	9	20

Pretpostavimo da se prva grupa podataka sastoji od deset rezultata. Jedan, maksimalni, je izgubljen, tako da znamo da se radi o sljedećih devet: 1, 2, 3, 3, 10, 10, 13, 13, 15. Da li na osnovu izračunatih vrijednosti (prvi, treći kvartili i C) možemo odrediti o kojem se rezultatu radi? Ako znamo da je za ovu grupu podataka aritmetička sredina M=8,6 te standardna devijacija s=5,8, da li možemo odrediti o kojem rezultatu se radi? Obrazložite odgovor i, ukoliko je moguće, odredite rezultat koji nedostaje.

12. Ispod su prikazani histogrami distribucije rezultata ispitanika na četiri testa (A, B, C, D).



Odredite kojem box-plot prikazu pripadaju gore prikazani histogrami.



13. Izračunajte interkvartilni raspon, varijancu i standardnu devijaciju za podatke iz zadatka 3.10 i 3.11.

14. Izračunajte interkvartilni raspon, standardnu devijaciju i varijancu za podatke iz zadatka 3.12. Imajući u vidu dobivene vrijednosti varijance i standardne devijacije, još jednom razmislite o najpogodnijoj mjeri centralne tendencije distribucije.

15. Izračunajte interkvartilni raspon, varijancu i standardnu devijaciju za podatke iz zadatka 3.13.

16. Izračunajte mjere varijabiliteta za podatke iz primjera 2.2 i 2.9. Koju mjeru varijabiliteta trebamo koristiti za adekvatnu deskripciju podataka svake od varijabli.

5. Osnovni koncepti vjerovatnoće

Teorija vjerovatnoće je grana matematike iz koje je nastala matematička statistika. **Zaključivanje u statistici temelji se na vjerovatnoći**. Zaključak kojim se prihvata ili odbacuje hipoteza donosi se uz vjerovatnoću da možda nismo u pravu. U tekstu ispod navest ćemo osnovne koncepte vjerovatnoće.

Postoje klasična, geometrijska, statistička (aksiomatska) i subjektivna definicija vjerovatnoće. **Subjektivna vjerovatnoća** je vjerovanje osobe u mogućnost (izvjesnost) pojave nekog događaja (npr. „Mislim da će sutra padati kiša“). Ova vjerovatnoća se ne temelji na matematičkim modelima ili definiciji već na osobnom iskustvu.

Klasična (Laplaceova) definicija vjerovatnoće

Slučajni eksperiment ima konačan broj (n) svih mogućih ishoda, pri čemu svaki od mogućih ishoda ima jednaku mogućnost pojavljivanja (tj. svi ishodi su jednako vjerovatni). Svi mogući ishodi slučajnog eksperimenta tvore potpuni skup Ω. Svaki događaj u slučajnom eksperimentu je podskup A skupa Ω svih mogućih ishoda ($A \subset \Omega$). Ako je skup Ω svih mogućih ishoda n-člani skup, a slučajni događaj m-člani podskup A, onda se općenito vjerovatnoća događaja A definira kao omjer broja slučajnih događaja (m) i broja svih mogućih ishoda (n).

Vjerovatnoća se definira kao omjer broja povoljnih i broja svih mogućih ishoda eksperimenta.

Vjerovatnoća slučajnog događaja izražava se realnim brojem između 0 i 1, uključujući 0 i 1, tj $0 \leq P(A) \leq 1$.

Vjerovatnoća se definira izrazom:

$$P(A)=\frac{m}{n}$$

Ako je slučajni eksperiment bacanje igraće kocke, skup svih mogućih ishoda bit će Ω = {1,2,3,4,5,6}. Događaj za kojeg želimo utvrditi vjerovatnoću je pojava parnog broja. Stoga je podskupa A={2,4,6}. Postoji n=6 svih mogućih ishoda, od kojih je m=3 povoljnih za događaj A. Tada je P(A) =3/6

$$P(A)=0{,}5$$

Vjerovatnoća događaja A iznosi 0,5

Kada je skup svih mogućih ishoda **beskonačan skup**, klasična definicija vjerovatnoće nije primjenljiva.

Klasični (kao i geometrijski) model vjerovatnoće temelje se na pretpostavci da su svi ishodi jednako mogući. U mnogim slučajnim pojavama u prirodi i društvu ova pretpostavka nije realna.

Statistička definicija vjerovatnoće

Statistička definicija vjerovatnoće temelji se na empirijskoj spoznaji o stabilnosti relativnih frekvencija slučajnih događaja kod ponavljanja eksperimenata. Frekvencija događaja A pri n ponavljanja slučajnog eksperimenta je broj m pojavljivanja događaja A. Omjer frekvencije (m) događaja A i broj (n) ponavljanja slučajnog eksperimenta zove se relativna frekvencija događaja A:

$$f_A = \frac{m}{n}$$

Ako se slučajni eksperiment sastoji od bacanja igrače kocke, onda se uzastopno bacanje smatra ponavljanjem slučajnog eksperimenta (u istim uvjetima). Neka je pojava parnog broja događaj A. Ako se nakon 100 uzastopnih bacanja kocke događaj A ostvari 45 puta, kažemo da je frekvencija m=45. Tada je $f_A=45/100$.

$$f_A=0{,}45$$

Relativna frekvencija zavisi, osim o događaju A, i o broju ponavljanja slučajnog eksperimenta. Za velike n, tj. pri velikom broju ponavljanja eksperimenta (ili opažanja), relativna frekvencija prestaje zavisiti od n i stabilizira se na određenu fiksnu vrijednost koja se interpretira kao statistička vjerovatnoća događaja A.

Praktično, relativna frekvencija se koristi kao vjerovatnoća događaja A, pri čemu se nastoji postići što je moguće veći broj (n) opažanja slučajne pojave.

Osnovne teoreme vjerovatnoće

Dvije osnovne teoreme vjerovatnoće, relevantne za uvod u statistiku su aditivna i multiplikaciona teorema.

1. Aditivna teorema

Za međusobno isključive događaje A_1 i A_2 vjerovatnoća da će se dogoditi ili A_1 ili A_2, jednaka je sumi vjerovatnoća za svaki događaj posebno:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

U kutiji se nalaze tri bijele, sedam crvenih i jedna crna kuglica. Kolika je vjerovatnoća da ćemo iz kutije izvaditi bijelu/ crvenu/ crnu kuglicu?Kolika je vjerovatnoća da ćemo iz kutije izvaditi ili bijelu ili crvenu kuglicu?

$P(A_1) = 3/11$

$P(A_2) = 7/11$

$P(A_3) = 1/11$

$P(A_1 \cup A_2) = P(A_1) + P(A_2) = 3/11 + 7/11 = 10/11$

2. Multiplikaciona teorema

Vjerovatnoća istovremenog događanja dva ili više nezavisnih događaja jednaka je produktu pojedinačnih vjerovatnoća tih događaja:

$$P(A_1 \cap A_2) = P(A_1) \times P(A_2)$$

Ako bacamo dvije kocke, kolika je vjerovatnoća da će i na jednoj i na drugoj kocki pasti broj 6?

$P(A_1 \cap A_2) = P(A_1) \times P(A_2) = 1/6 \times 1/6 = 1/36$

Slučajne varijable i distribucije vjerovatnoća

Slučajni događaj je događaj koji se pod određenim uvjetima može, ali i ne mora dogoditi (npr. pojava „pisma“ kod bacanja novčića, ili pojava rezultata većih od aritmetičke sredine na testu znanja iz Statistike). Za dva događaja kažemo da su **nezavisna** kada pojavljivanje ili nepojavljivanje jednog nema nikakvog efekta na pojavljivanje ili nepojavljivanje drugog. Za dva događaja kažemo da se **međusobno isključuju** ako istovremeno ne mogu nastupiti oba.

Slučajna varijabla je promjenljiva veličina koja poprima vrijednosti iz zadatog skupa S svih mogućih ishoda, tj. slučajnih događaja nekog slučajnog eksperimenta. Uz svaki element skupa S veže se određena vjerovatnoća njegovog ishoda. Slučajnoj varijabli X pripada određena **distribucija vjerovatnoća**.

Distribucije diskretnih i kontinuiranih varijabli tretiraju se na različit način u teoriji vjerovatnoće.

Diskretnoj slučajnoj varijabli pripada **diskretna distribucija vjerovatnoća**. Svaka vrijednost *i* (mogući ishod) opterećena je određenom vjerovatnoćom p_i.

Elementi skupa S={1,2,3,4,5,6} su ishodi slučajnog eksperimenta bacanja igrače kocke. Sa X označavamo slučajnu varijablu za koju je S skup mogućih vrijednosti. U slučaju idealne kocke, svakom ishodu slučajnog eksperimenta pripada vjerovatnoća 1/6, tj. P(X=1)=1/6. Također vrijedi P(X=2)=1/6, itd., P(X=6)=1/6.

Vjerovatnoće svakog ishoda možemo prikazati na sljedeći način:

Vrijednosti	***1***	***2***	***3***	***4***	***5***	***6***
Vjerovatnoće	*1/6*	*1/6*	*1/6*	*1/6*	*1/6*	*1/6*

Ako eksperiment radimo sa kockom koja nije idealna (jer je npr. pomaknuto težište), tada ishodi ne bi bili jednako vjerovatni, već bi im pripadale različite vjerovatnosti ($p_i \geq 0$, $\Sigma p_i = 1$).

Binomna distribucija (situacije u kojima svaki broj nezavisnih pokušaja rezultira jednim od dva moguća ishoda, pri čemu pojavljivanje jednog ishoda isključuje mogućnost pojavljivanja drugog ishoda – primjer bacanja novčića) i **Poissonova distribucija** (raspodjela rijetkih događaja) predstavljaju diskretne distribucije vjerovatnoća.

Kontinuiranoj slučajnoj varijabli pripada **distribucija vjerovatnoća vrijednosti (ishoda) unutar određenog intervala**. Distribucija vjerovatnoća kontinuirane slučajne varijable razlikuje se od distribucije vjerovatnoća diskretnih varijabli jer:

- ishod (događaj, rezultat) može biti bilo koja vrijednost unutar određenog opsega, s tim da ta vrijednost nije nužno cijeli broj,
- vjerovatnoća jedne specifične vrijednosti je nula, i
- vjerovatnoća se izražava u terminima površine pod krivom koja predstavlja kontinuiranu distribuciju.

Distribucija vjerovatnoća kontinuirane slučajne varijable može se prikazivati pomoću relativnih frekvencija (jer distribucija relativnih frekvencija ima slična svojstva kao i distribucija vjerovatnoća).

Statistikom „dominira“ jedna kontinuirana distribucija nazvana **normalna distribucija** jer **predstavlja model** mnogih kontinuiranih slučajnih varijabli, kao što su fizičke karakteristike (tjelesna težina, visina), rezultati na testovima ličnosti, inteligencije, itd. Pod određenim uvjetima, normalna distribucija predstavlja i **aproksimaciju** različitih diskretnih distiribucija (binomne i Poissonove).

PRIMJER 5.1

Neka su bačene dvije kocke. Kolika je vjerovatnoća da je zbir brojeva koji se pojave na kockama 7?

Svi mogući ishodi eksperimenta, tj. potpuni skup Ω sastoji se od svih uređenih parova (i, j). U svakom uređenom paru prvi element predstavlja broj koji se pojavljuje na gornjoj strani prve kocke, a drugi element u paru predstavlja broj koji se pojavljuje na gornjoj strani druge kocke. Skup Ω u ovom primjeru ima 36 elemenata.

Događaj A, da je zbir brojeva koji se pojave na kockama 7, jeste sljedeći podskup od Ω:

$$\Omega = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$$

Vjerovatnoća događaja A, jednaka je:

$$P(A) = 7 / 36=0,194$$

PRIMJER 5.2

U kutiji se nalaze pet bijelih, četiri žute i dvije crne kuglice. Kolika je vjerovatnoća da ćemo iz kutije izvaditi svijetlu (bijelu ili žutu) kuglicu?

Primjenit ćemo aditivnu teoremu, prema kojoj za međusobno isključive događaje A_1 i A_2 vjerovatnoća da će se dogoditi ili A_1 ili A_2 jednaka je sumi vjerovatnoća svakog pojedinačnog događaja. Događaj A_1 je izvlačenje bijele, a događaj A_2 je izvlačenje žute kuglice. Vjerovatnoće događaja A_1, A_2 i A_3 iznose:

$P(A_1) = 5/11$

$P(A_2) = 4/11$

$P(A_3) = 2/11$.

Vjerovatnoća izvlačenja svijetle kuglice (ili bijele ili žute) iznosi:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) = 5/11 + 4/11 = 9/11=0,818$$

PRIMJER 5.3

Ako bacamo dvije kocke, kolika je vjerovatnoća da će i na jednoj i na drugoj kocki pasti broj 1?

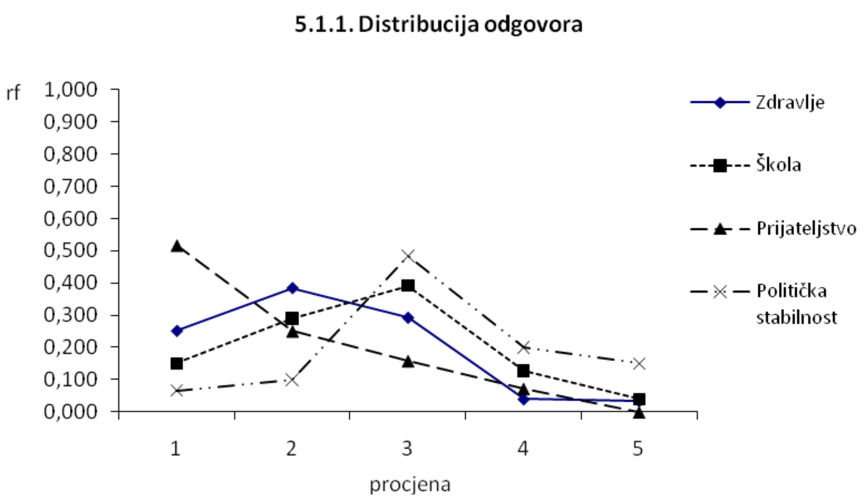
Prema multiplikacionoj teoremi, vjerovatnoća istovremenog događanja dva ili više nezavisnih događaja jednaka je produktu pojedinačnih vjerovatnoća tih događaja. Vjerovatnoća događaja A_1

jednaka je vjerovatnoći događaja A_2 i iznosi 1/6. Prema tome, vjerovatnoća da će i na jednoj i na drugoj kocki pasti broj 1 iznosi:

$$P(A_1 \cap A_2) = P(A_1) \text{ x } P(A_2) = 1/6 \text{ x } 1/6 = 1/36=0,028$$

PRIMJER 5.4

Pretpostavimo da veliku grupu srednjoškolaca (N=3000) pitamo da na skali od 1 (ekstremno važno) do 5 (uopće nije važno) procjene važnost različitih aspekata njihovog života: zdravlje, škola, prijatelji, politička stabilnost. Na slici 5.4.1 dat je grafički prikaz distribucije relativnih frekvencija odgovora ovog hipotetičkog primjera.



Na apscisi su nanesene procjene (od 1 do 5), a na ordinati relativne frekvencije odgovora svakog od procjenjivanog aspekta. Relativne frekvencije se praktično koriste kao vjerovatnoće događaja, u ovom slučaju procjene važnosti različitih aspekata života.

Iz grafičkog prikaza vidljive su različite distribucije odgovora za različite aspekte života. Nešto više od polovine ispitanika (rf=0,517) prijateljstvo procjenjuje kao ekstremno važan aspekt njihovog života, dok političku stabilnost na isti način procjenjuje jako mali broj ispitanika (rf=0,066). Drugim riječima, vjerovatnoća da će slučajno izabrana osoba prijateljstvo procjeniti kao ekstremno važno iznosi 0,517, dok je vjerovatnoća da će neka osoba (također odabrana po slučaju) političku stabilnost procjeniti kao ekstremno važnu samo 0,066.

PRIMJER 5.5

Studenti prve godine Odsjeka za psihologiju u Sarajevu su na vježbama iz predmeta Statistika u psihologiji I napravili mali eksperiment: željeli su empirijski provjeriti kako izgleda distribucija diskretne slučajne varijable u slučaju velikog broja rezultata. Da bi to postigli studenti su bacali parove igračih kocki i bilježili dobivene rezultate. Dakle, rezultat je operacionaliziran kao zbir brojeva koji se dobije bacanjem dvije igrače kocke.

Moguće kombinacije brojeva i odgovarajući rezultati koji se mogu dobiti bacanjem dvije kocke (kocke I i II) prikazani su u tabeli 5.5.1

Tabela 5.5.1. Kombinacije brojeva i odgovarajući rezultati koji se mogu dobiti bacanjem dvije kocke

Kocka I	Kocka II	Rezultat	Kocka I	Kocka II	Rezultat	Kocka I	Kocka II	Rezultat	Kocka I	Kocka II	Rezultat	Kocka I	Kocka II	Rezultat	Kocka I	Kocka II	Rezultat
1	1	2	2	1	3	3	1	4	4	1	5	5	1	6	6	1	7
1	2	3	2	2	4	3	2	5	4	2	6	5	2	7	6	2	8
1	3	4	2	3	5	3	3	6	4	3	7	5	3	8	6	3	9
1	4	5	2	4	6	3	4	7	4	4	8	5	4	9	6	4	10
1	5	6	2	5	7	3	5	8	4	5	9	5	5	10	6	5	11
1	6	7	2	6	8	3	6	9	4	6	10	5	6	11	6	6	12

Iz tabele se vidi da se mogući rezultati kreću u rasponu od 2 do 12. Mogućih kombinacija brojeva na dvije kocke imamo 36, dok se kombinacije koje daju pojedini rezultat pojavljuju različit broj puta (tako, npr., imamo samo jednu kombinaciju koja daje rezultat 2: 1+1=2, dok kombinacija koje daju rezultat 7 ima najviše: 1+6, 2+5, 3+4, 4+3, 5+2, 6+1). Na temelju ukupnog broja kombinacija (n=36) i broja kombinacija koje daju pojedini rezultat (m) moguće je izračunati vjerovatnoću javljanja svakog pojedinog rezultata (tabela 5.5.2.).

Tabela 5.5.2: Vjerovatnoća javljanja svakog pojedinog rezultata

Rezultat	m	n	p=m/n	%
2	1	36	0,0278	2,7778
3	2	36	0,0556	5,5556
4	3	36	0,0833	8,3333
5	4	36	0,1111	11,1111
6	5	36	0,1389	13,8889
7	6	36	0,1667	16,6667
8	5	36	0,1389	13,8889
9	4	36	0,1111	11,1111
10	3	36	0,0833	8,3333
11	2	36	0,0556	5,5556
12	1	36	0,0278	2,7778
total			1,0000	100

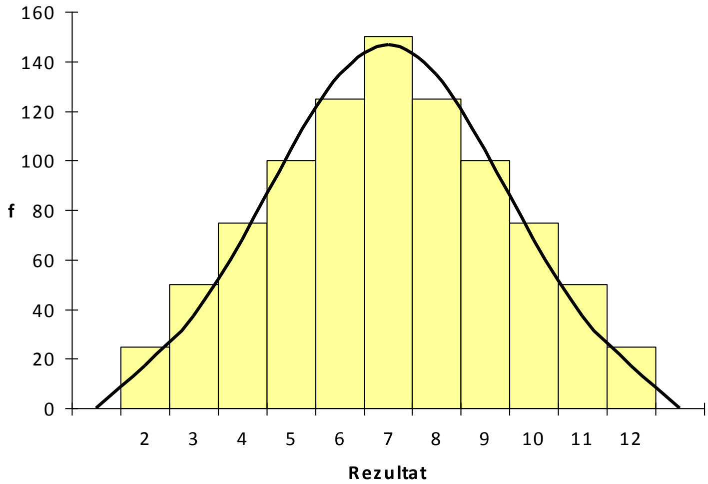
Studenti psihologije su na vježbama iz psihologije utvrdili ukupno N=900 rezultata bacanja dvije kocke (studenti su vježbe iz predmeta Statistika u psihologiji I pohađali u 3 grupe; u svakoj grupi, studenti su bacali 15 parova kocki po 20 puta; tako smo dobili skup od ukupno 900 rezultata: 3 grupe x 15 parova kocki x 20 bacanja = 900 rezultata). Na temelju gornje tabele sa teorijskim vjerovatnoćama pojedinih rezultata moguće je načiniti tabelu sa teorijskim (očekivanim) frekvencijama svakog pojedinog rezultata u skupu rezultata veličine N=900. Naime, vjerovatnoće pojedinih rezultata izražene kao proporcije i procenti u gornjoj tabeli nisu ništa drugo nego relativne frekvencije i relativne frekvencije u procentima pojedinih rezultata (tabela 5.5.3).

Table 5.5.3: Relativne frekvencije i relativne frekvencije u procentima pojedinih rezultata

Rezultat	p (rf)	% (rf%)	f
2	0,02778	2,778	25
3	0,05556	5,556	50
4	0,08333	8,333	75
5	0,11111	11,111	100
6	0,13889	13,889	125
7	0,16667	16,667	150
8	0,13889	13,889	125
9	0,11111	11,111	100
10	0,08333	8,333	75
11	0,05556	5,556	50
12	0,02778	2,778	25
total	1,00000	100,000	900

Gornja distribucija ima M=7 i s=2,42, te ima oblik normalne distribucije, što se može vidjeti i na slici 5.2.2.

Slika 5.2.2. Teorijska (očekivana) distribucija rezultata bacanja dvije kocke 900 puta



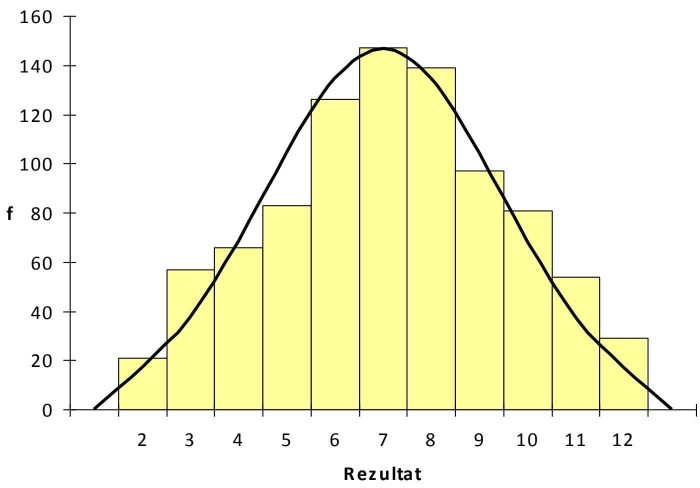
Sada možemo preći na empirijski dio eksperimenta. Studenti su prilikom 900 stvarnih bacanja kocki dobili frekvencije prikazane u tabeli 5.5.4.

Table 5.5.4: Empirijska distribucija rezultata

Rezultat	Opažena frekvencija
2	21
3	57
4	66
5	83
6	126
7	147
8	139
9	97
10	81
11	54
12	29
Σ	900

Vidimo da se opažene frekvencije relativno dobro podudaraju sa teorijski očekivanim. Empirijska distribucija prikazana je na slici 5.2.3.

Slika 5.2.3. Empirijska (opažena) distribucija rezultata bacanja dvije kocke 900 puta



Kao što se vidi, empirijska distribucija vjerno prati krivu normalne raspodjele.

Dobivena distribucija ima M=7,13 i s=2,43. Ove su vrijednosti vrlo bliske (praktično identične) teorijski očekivanim vrijednostima.

Prema tome, naš eksperiment je uspio: empirijski smo pokazali da se ishodi unutar slučajnih varijabli (uz uvjet da imamo dovoljan broj ishoda) distribuiraju u obliku normalne distribucije. Ovo je vrlo važna statistička činjenica jer u empirijskim istraživanjima vrlo često poredimo empirijski utvrđenu distribuciju rezultata sa normalnom distribucijom koju teorijski očekujemo kada imamo slučajnu varijablu. Ukoliko između empirijske i teorijske distribucije postoje značajna odstupanja zaključujemo da empirijska distribucija nije distribucija slučajne varijable. U tom slučaju smatramo da je na empirijsku distribuciju djelovao neki sistematski faktor koji je doveo do njenog odstupanja u jednom ili drugom smjeru. Taj sistematski faktor vrlo često nazivamo **nezavisnom varijablom**.

PRIMJER 5.6

Posmatrajmo eksperiment bacanja dvije kocke. Kolika je vjerovatnoća da je zbir brojeva koji se pojavljuju na kockama 6, uz uvjet da je drugi broj paran?

Neka je E događaj takav da je zbir brojeva koji se pojavljuju na kockama 6, tj. E = {(1,5), (2,4), (3,3), (4,2), (1,5)}. Neka je F događaj da je broj koji se pojavljuje na drugoj kocki paran, tj. F={(1,2), (2,2), (3,2), (4,2), (5,2), (6,2),(1,4), (2,4), (3,4), (4,4), (5,4), (6,4), (1,6), (2,6), (3,6), (4,6), (5,6), (6,6)}.

Kolika je vjerovatnoća događaja E uz uvjet da je zadovoljen događaj F? Ako je zadovoljen F, tj. na drugoj kocki se pojavljuje paran broj, onda je zbir brojeva na obje kocke 6, samo ako su ishodi sljedeći: (2,4), (4,2). Događaj F ima 18 elemenata. Stoga je vjerovatnoća događaja E uz uvjet da je dat F: 2 /18, tj.

$P(E \mid F) = 2/18$ (čitamo: vjerovatnoća od E dat F).

PRIMJER 5.7

Provedena su tri eksperimenta sa tri igraće kocke. U svakom eksperimentu igraću kocku bacali smo 600 puta. Ispod su prikazane učestalosti pojavljivanja brojeva 1, 2, 3, 4, 5, 6.

Strana kocke	Eksperiment 1	Eksperiment 2	Eksperiment 3
1	100	50	101
2	98	70	99
3	102	130	102
4	93	80	101
5	110	150	98
6	97	120	99
total	600	600	600

Šta možete sve zaključiti na osnovu dobivenih distribucija?

Vjerovatnoća pojavljivanja svakog od brojeva kocke iznosi p=1/6=0,167. Najprije, izračunat ćemo relativne frekvencije i uporediti ih sa teorijskim vjerovatnoćama (p=0,167 za svaki broj). Ukoliko je kocka ispravna ili je eksperiment proveden na pravilan način, relativne frekvencije trebale bi biti približno jednake teorijskim.

Strana kocke	Eksperiment 1		Eksperiment 2		Eksperiment 3	
	p_e	p_t	p_e	p_t	p_e	p_t
1	0,167	0,167	0,083	0,167	0,168	0,167
2	0,163	0,167	0,117	0,167	0,165	0,167
3	0,170	0,167	0,217	0,167	0,170	0,167
4	0,155	0,167	0,133	0,167	0,168	0,167
5	0,183	0,167	0,250	0,167	0,163	0,167
6	0,162	0,167	0,200	0,167	0,165	0,167
total	1	1	1	1	1	1

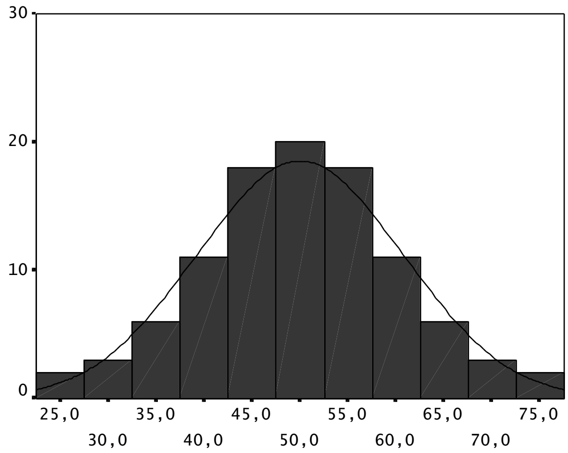
Očigledno je da su razlike između teorijskih vjerovatnoća i relativnih frekvencija najveće u eksperimentu 2. Npr., za broj pet, relativna frekvencija iznosi 0,250, dok je teorijska vjerovatnoća p=0,167. Izvjesne razlike utvrđene su u prvom eksperimentu, a najmanje su u trećem eksperimentu.

Na osnovu dobivenih rezultata možemo opravdano pretpostaviti da drugi eksperiment nije pravilno proveden ili da kocka nije bila ispravna.

6. Normalna raspodjela

Na slici 6.1 prikazan je histogram rezultata 100 ispitanika na testu X. Mnogi podaci, prikupljeni od relativno velikog broja ispitanika, raspoređuju se slično kao što je prikazano na slici. Možemo primjetiti da su krajevi histograma jednako udaljeni od jednog vrha pozicioniranog tačno u sredini.

Slika 6.1: Histogram rezultata testa X



Krivulja nacrtana kroz histogram predstavlja matematički model raspodjele rezultata i pruža kompaktnu sliku cjelokupne raspodjele rezultata (obzirom da je matematički model idealizirana slika raspodjele rezultata, zanemaruju se eventualna mala odstupanja).

Krivulja prikazana na slici je **normalna krivulja** i grafički opisuje kontinuiranu raspodjelu koju zovemo **normalna raspodjela**. Normalna raspodjela je zvonolikog oblika, simetrična i unimodalna.

Matematički je definirana izrazom:

$$p(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \quad e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

u kojem X označava rezultat, μ aritemetičku sredinu, a σ standardnu devijaciju (π i e su konstante). Prema tome, normalna raspodjela je u potpunosti determinirana vrijednostima aritmetičke sredine i standardne devijacije. Aritmetička sredina nalazi se u centru raspodjele, tj. tjemenu krive, i iste je vrijednosti kao i medijana.

Normalna raspodjela je važna u statistici jer:

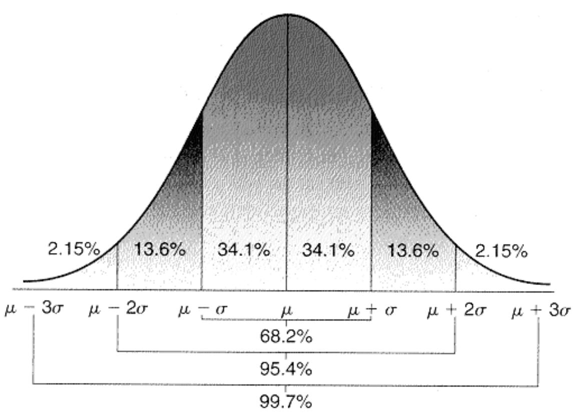
- vrijednosti mnogih psiholoških varijabli u populaciji se raspoređuju normalno,
- osnova je za mnoge statističke testove, i
- pod određenim uvjetima, predstavlja aproksimaciju različitih diskretnih raspodjela (binomne i Poissonove).

Ukoliko znamo vrijednosti aritmetičke sredine i standardne devijacije, tada možemo odrediti broj rezultata unutar nekog opsega[2]. Odnosno vrijedi da se u...

- intervalu $\mu \pm \sigma$ nalazi se 68,26% svih rezultata
- intervalu $\mu \pm 2\sigma$ nalazi se 95,44% svih rezultata
- intervalu $\mu \pm 3\sigma$ nalazi se 99,73% svih rezultata

Na slici ispod prikazana je normalna raspodjela sa navedenim intervalima i postocima rezultata koji se nalaze u datom intervalu.

Slika 6.2: Normalna raspodjela



Standardna normalna raspodjela

Empirijski dobivene normalne raspodjele možemo aproksimirati na jednu normalnu raspodjelu ako podatke izrazimo u jedinicama standardne devijacije, tj. transformiramo u z-vrijednosti.

---

[2] Određivanje broja podataka unutar zadatog intervala provodi se integriranjem funkcije: $p(X)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}$, u intervalu od X_1 do X_2. U statističkim tablicama koje se nalaze na kraju svakog udžbenika iz statistike, unijete su vrijednosti na osnovu kojih možemo odrediti broj podataka unutar nekog intervala.

Transformacija rezultata u z-vrijednosti naziva se **standardizacija rezultata**. Pomoću z-vrijednosti izražavamo koliko je neki rezultat udaljen od aritmetičke sredine i u kojem smjeru (desno ili lijevo od aritmetičke sredine), pri čemu se udaljenost izražava u jedinicama standardne devijacije. Z-vrijednost određujemo koristeći izraz:

$$Z = \frac{X - \mu}{\sigma}$$

Neka je M = 52, s = 12. Odredite z-vrijednost rezultata $X_1 = 70$

$z_1 = (70 – 56)/12 \; =2$

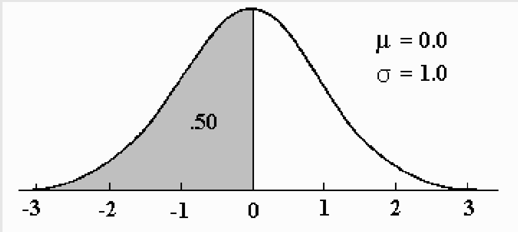
Rezultat 70 udaljen je od M za 2 standardne devijacije, u desnu stranu.

Raspodjela rezultata izraženih u z-vrijednostima naziva se **standardna normalna raspodjela**. Standardizacijom rezultata bilo koju normalnu raspodjelu svodimo na jednu, standardnu raspodjelu. Aritmetička sredina ove raspodjele iznosi $\mu=0$ a standardna devijacija $\sigma=1$. Površina pod standardnom normalnom raspodjelom iznosi p=1.

Površina pod krivom proporcionalna je broju podataka u raspodjeli, pa se određivanje broja podataka u određenom intervalu svodi na određivanje površine pod normalnom krivuljom. Za standardnu normalnu raspodjelu vrijednosti površina za pojedine opsege rezultata ispod normalne krivulje očitavamo iz tablice standardne normalne raspodjele, koje se nalaze u svakom udžbeniku statistike.

Postupak određivanja broja podataka u određenom intervalu sastoji se u tome da, najprije, granične vrijednosti intervala pretvorimo u z-vrijednosti, a zatim, koristeći se odgovarajućom tablicom, utvrdimo proporciju rezultata unutar datog intervala (tj. odredimo površinu krivulje koja odgovara datom intervalu).

U primjeru desno, označena je jedna strana pod krivuljom, čija površina iznosi P=0,5. U opsegu od -3 standardne deviacije do aritmetičke sredine (tj. 0 standardne devijacije) nalazi se 50 posto rezultata.



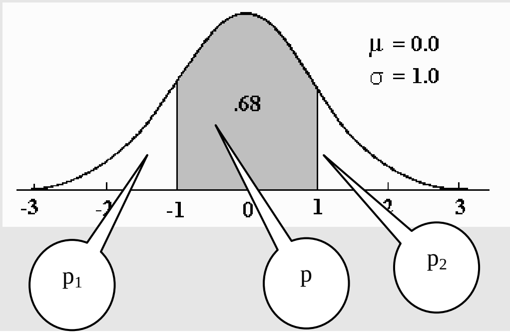
U primjeru desno, označena je površina pod normalnom raspodjelom između $z_1=-1$ *i* $z_2=1$.

$z_1 = -1 \Rightarrow p_1 = 0.1587$

$z_2 = 1 \Rightarrow p_2 = 0.1587$

$p = 1 - (p_1 + p_2) = 1 - 0,3174$

$\boldsymbol{p = 0,6826}$



Površina pod normalnom krivuljom (u datom intervalu) predstavlja **vjerovatnoću pojavljivanja rezultata datog intervala**. Vjerovatnoća da ćemo metodom slučajnog odabira izvući rezultat koji se nalazi u rasponu od -3 do 3 standardne devijacije iznosi 99,73%. Vjerovatnoća da ćemo izvući rezultat koji se nalazi u rasponu od -1 do 1 standardne devijacije iznosi 68,26%.

Iz grafičkog prikaza normalne krivulje očigledno je da rasponi iste veličine nemaju istu vjerovatnoću pojavljivanja.

PRIMJER 6.1

Raspodjela vrijednosti holesterola u krvi u populaciji osoba iste dobne grupe i spola približno je normalna. Za 14-godišnje dječake prosječna vrijednost u populaciji iznosi μ = 170 mg/dl, a standardna devijacija σ = 30 mg/dl. Vrijednosti iznad 240 mg/dl zahtijevaju medicinski tretman. Zanima nas postotak 14-godišnjih dječaka koji imaju vrijednost holesterola veću od 240 mg/dl?

Potrebno je odrediti postotak dječaka sa X>240 mg/dl. U raspodjeli rezultata određivanje postotka dječaka sa X>240 mg/dl podrazumijevalo bi određivanje proporcije, tj. površine pod normalnom krivuljom pomoću formule koja matematički definira normalnu krivulju. Umjesto toga, koristit ćemo standardnu normalnu raspodjelu za koju su vrijednosti površina pod krivuljom izračunate i sistematizirane u tablici. Stoga ćemo najprije izračunati z-vrijednost za X=240, a zatim iz tabele A očitati površinu koja korespondira izračunatoj z-vrijednosti.

μ = 170 mg/dl

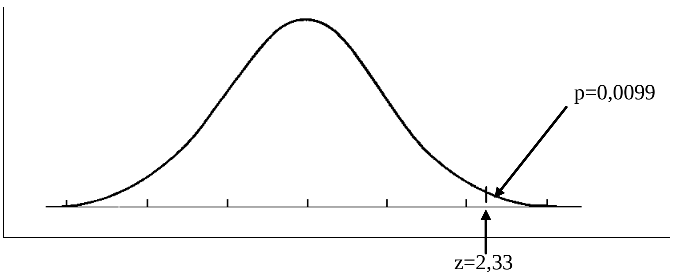
σ = 30 mg/dl

X = 240 mg/dl

$$z = \frac{X - M}{s} \qquad z = \frac{240 - 170}{30}$$

z = 2,33

Potrebno je odrediti površinu pod normalnom krivuljom za z > 2,33.



Iz tablice ćemo očitati da površina od z=2,33 do kraja krivulje iznosi p=0.0099, tj p≈0,01. Pretvoreno u procente, površina iznosi 1%.

Na kraju zaključujemo da u populaciji možemo očekivati 1% dječaka dobi od 14 godina koji imaju vrijednost holesterola u krvi veću od 240 mg/dl.

PRIMJER 6.2

Prosječno vrijeme trajanja trudnoće (od začeća do poroda) u populaciji iznosi μ=266 dana uz varijabilnost od σ=16 dana. Raspodjela je približno normalna.

a. Osoba A.B porodila se 282. dan. Koji postotak žena ima vrijeme trajanja trudnoće veće od osobe A.B.?

μ = 266 dana

σ = 16 dana

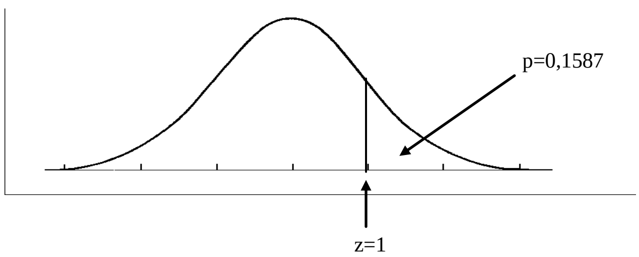
X=282 dana

z = ?

$$z = \frac{X - M}{s} \qquad z = \frac{282 - 266}{16}$$

z = 1

Potrebno je odrediti površinu pod normalnom krivuljom za z > 1.



Iz tablice ćemo očitati da površina od z=1 do kraja krivulje iznosi p=0,1587. Pretvoreno u procente, površina iznosi 15,87%. Na kraju, zaključujemo da u populaciji možemo očekivati 15,87% žena kod kojih trudnoća traje više od 282 dana.

b. Osoba C.D. porodila se 250-ti dan trudnoće. Koji postotak žena ima vrijeme trudnoće veće od osobe C.D.?

μ = 266 dana

σ = 16 dana

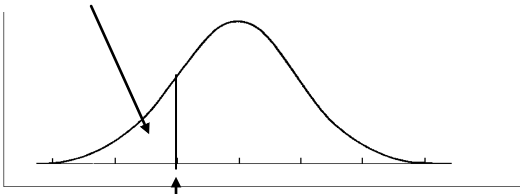
X = 250 dana

z = ?

$$z = \frac{X - M}{s} \qquad z = \frac{250 - 266}{16}$$

z = -1

Potrebno je odrediti površinu pod normalnog krivuljom za $z > -1$.



Iz tablice ćemo očitati da površina od z=-1 do bližeg kraja krivulje iznosi p=0,1587. Pretvoreno u procente površina iznosi 15,68%. Međutim, površina koja nas interesira nalazi se od z do desnog kraja krivulje. Stoga ćemo zaključiti da u populaciji možemo očekivati 100% - 15,87 % = 84,17 % žena kod kojih trudnoća traje više od 250 dana.

c. Koliko je žena kojima je vrijeme trudnoće između 250 i 282 dana.

$\mu = 266$ dana

$\sigma = 16$ dana

$X_1 = 250$ dana

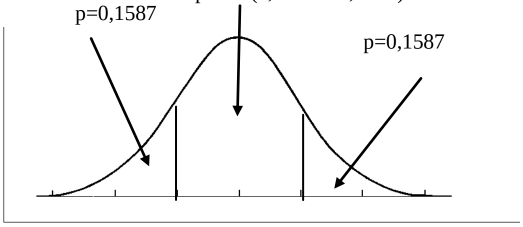
$X_2 = 282$ dana

$z_1 = ?$

$z_2 = ?$

$z_1 = -1$; $z_2 = 1$

Potrebno je odrediti površinu pod normalnom krivuljom između z=-1 i z=1.



$p = 1 - (0{,}1587 + 0{,}1587)$, $p = 0{,}6826$

$P = 68{,}26\ \%$

d. U kojem intervalu se nalazi 99,73% središnjih vrijednosti trajanja trudnoće?

U intervalu od -3z do 3z nalazi se 99,73% središnjih rezultata. Prema tome, potrebno je odrediti rezultate koji odgovaraju -3z i 3z.

Iz izraza:

$z = \frac{X - M}{s}$ slijedi da je: **X = M + zs,**

$X_1 = 266 + 3x16 = 314$,

$X_2 = 266 + (-3)x16 = 218$.

Dakle, u intervalu od 218 – 314 dana nalazi se 99,73% vrijednosti trajanja trudnoće.

e. U kojem intervalu se nalazi 95,44% središnjih vrijednosti trajanja trudnoće?

U intervalu od -2z do 2z nalazi se 95,44% rezultata. Prema tome, potrebno je odrediti rezultate koji odgovaraju -2z i 2z.

$X_1 = 266 + 2x16 = 298$

$X_2 = 266 + (-2)x16 = 234$

Dakle, u intervalu od 298 – 234 dana nalazi se 95,44% središnjih vrijednosti trajanja trudnoće.

f. Koliko traje trudnoća za žene koje se nalaze u prvom kvartilu distribucije?

Prvi kvartil sadržava prvih 25% vrijednosti. Znači da je gornja granica prvog kvartila 25. centil. Pod normalnom krivuljom prvih 25% rezultata odgovara površini p=0,25. Ako znamo p, iz tablice možemo očitati i njoj odgovarajuću z-vrijednost.

Za p = 0,25, z = 0,68

X = M + zs

X = 266 + (-0,68)x16 = 255,12

Zaključujemo da za žene koje se nalaze u prvom kvartilu vrijeme trajanja trudnoće iznosi maksimalno 255 dana.

PRIMJER 6.3

Primjenjujući test matematičkih kompetencija na velikom broju učenika prvog razreda srednjih škola dobivena je normalna raspodjela rezultata sa sljedećim vrijednostima M=500 i s=100.

a. Koji postotak učenika postiže rezultate veće od 600? Drugim riječima, koliko iznosi centilni rang rezultata 600?

$$z = \frac{X - M}{s}$$

z = (600 – 500)/100 = 1

p = 0,1587 →P = 15,87%.

15,87% učenika postiže rezultat veći od 600. Dakle, rezultat 600 leži na 84-om centilu (100-15,87=84,13≈84).

b. Koji postotak učenika postiže rezultate manje ili jednake 400? Drugim riječima, koliko iznosi centilni rang rezultata 400?

$$z = \frac{X - M}{s}$$

z = (400 – 500)/100 = -1

p = 0,1587

P = 15,87%,

Dakle, 15,87% učenika postiže rezultat manji ili jednak 400. Rezultat 400 leži na 16-om centilu (ispod ovog rezultata nalazi se oko 16% rezultata).

c. Koji rezultat je jednak manji od 75% rezultata postignutih na testu? Drugim riječima, koliko iznosi 25 centil?

p = 0,25

z = - 0,67

X = M + zs = 500 - 0,67x100 = 433

Rezultat 433 nalazi se na 25 centilu (ovaj rezultat je jednak ili manji od 75% postignutih rezultata).

U istom istraživanju primijenjen je test jezičkih kompetencija. Dobivene su sljedeće deskriptivne vrijednosti raspodjele: M = 550, s = 90.

Ispitanik A je na testu matematičkih kompetencija postigao rezultat 500, a na testu jezičkih kompetencija također rezultat 500. Na kojem testu je bio bolji?

$$z_1 = \frac{500 - 500}{100} = 0$$

$$z_2 = \frac{500 - 550}{90} = -0{,}55$$

Rezultat ispitanika A na testu matematičkih kompetencija jednak je vrijednosti aritmetičke sredine (z=0). Na testu jezičkih kompetencija nalazi se za nešto više od pola standardne devijacije ispod aritmetičke sredine (z=-0,55). Ispitanik A je postigao bolji rezultat na testu matematičih kompetencija.

PRIMJER 6.4

Kvocijent inteligencije (IQ) je standardizirani rezultat, a raspodjela vrijednosti IQ-a u populaciji približno je normalna, sa deskriptivnim populacijskim vrijednostima $\mu = 100$ i $\sigma = 16$.

a. Koliko iznosi vjerovatnoća slučajnog odabira rezultata vrijednosti 120 i više?

$z = (120-100)/16 = 1{,}25$

$p = 0{,}1056$

Dakle, vjerovatnoća slučajnog odabira rezultata vrijednosti 120 i više iznosi 10,56%.

b. Koliko iznosi vjerovatnoća slučajnog odabira rezultata vrijednosti 90 i manje?

$z = (90-100)/16 = -0{,}625$

$p = 0{,}2643$

Dakle, vjerovatnoća slučajnog odabira rezultata vrijednosti 90 i manje iznosi 26,43%.

c. Koliko iznosi vjerovatnoća slučajnog odabira rezultata vrijednosti od 90 do 120?

$z_1 = -0{,}625$; $z_2 = 1{,}25$

$p = 1-(p_1+p_2) = 1- (0{,}2643 + 0{,}1056) = 0{,}6301$

Dakle, vjerovatnoća slučajnog odabira rezultata između 90 i 120 iznosi 63,01%.

d. MENSA je organizacija koja okuplja ljude sa visokim IQ. Članovima ove organizacije može postati samo 2% osoba sa najvišim IQ rezultatima. Koliko iznosi najmanji rezultat koji pruža mogućnost učlanjenja u MENSA-u? Zapravo, interesira nas koja z vrijednost odgovara postotku od 2% najviših rezultata ispod standardne normalne krive. Iz tabele očitavamo da je to z=2,05.

X = M + zs

X = 100 + 2,05x16

X = 132,8

Dakle, najmanji IQ rezultat koji pruža mogućnost učlanjenja u MENSA-u iznosi 132,8.

PRIMJER 6.5

U primjeru 2.3 raspodjele visina učenika IIIc odjeljenja smo prikazivali tabelarno i grafički te smo ilustrovali vrste informacija koje možemo ekstrahirati iz ovih prikaza. Na ovom mjestu ćemo pokazati da do istih informacija možemo doći služeći se standardnom normalnom raspodjelom (obzirom da je i originalna raspodjela visina učenika također normalna) te pripadajućom tablicom *p* vrijednosti pod krivom standardne normalne raspodjele. U primjerima 3.4 i 4.5 utvrdili smo prosječnu vrijednost i pripadajuću standardnu devijaciju za raspodjelu visina 25 učenika IIIc odjeljenja: M = 135,08 cm i s = 1,83 cm.

a. Koliko učenika IIIc odjeljenja je visoko između 135,5 i 136,5 cm?

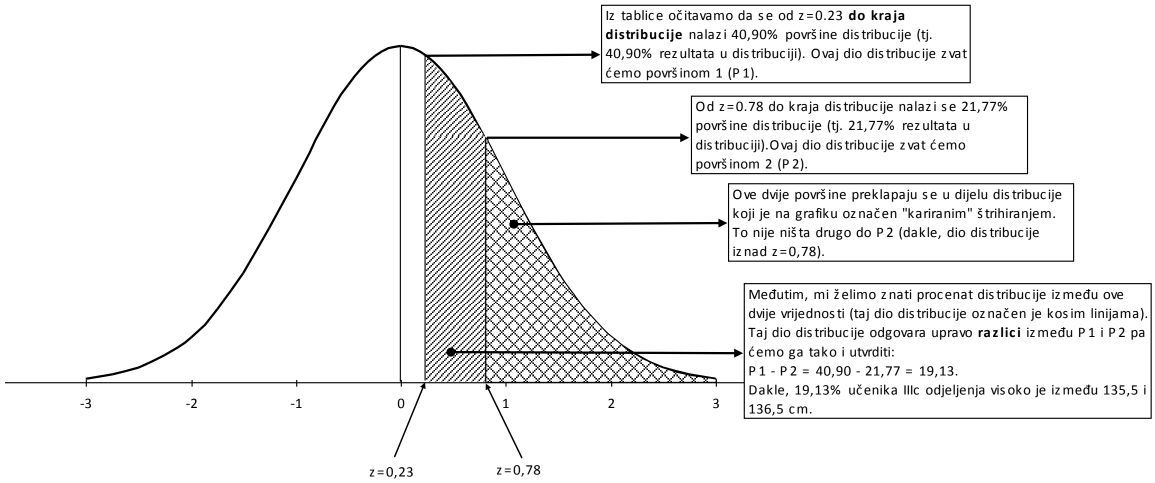
Da bismo utvrdili broj rezultata koji spadaju u određeni interval raspodjele moramo se koristiti tablicama *p* vrijednosti pod krivom standardne normalne raspodjele. Ove vrijednosti se iz tablica očitavaju samo u odnosu na standardne, z-vrijednosti. To znači da je neophodno utvrditi odgovarajuće z-vrijednosti za originalne rezultate (rezultate izražene na originalnoj skali mjerenja, tj. rezultate iz originalne raspodjele).

Pripadajuće z vrijednosti za originalne rezultate 135,5 i 136,5 cm su:

$$z = \frac{X - M}{s} = \frac{135{,}5 - 135{,}08}{1{,}83} = 0{,}23$$

$$z = \frac{X - M}{s} = \frac{136{,}5 - 135{,}08}{1{,}83} = 0{,}78$$

Na slici ispod prikazane su izračunate z-vrijednosti i korespondirajuće površine.



Dakle, koristeći se z-vrijednostima i odgovarajućim površinama distribucije došli smo do rezultata da je 19,13% učenika visoko između 135,5 i 136,5 cm.

Ukoliko se vratimo na primjer 2.3 iz tabele grupiranih rezultata možemo očitati da je 20% učenika visoko 136 cm (preciznije, između 135,5 i 136,5 cm), što je rezultat vrlo blizak ovom koji smo dobili pomoću z-vrijednosti.

20% od ukupno 25 učenika IIIc odjeljenja iznosi 5 učenika – što je upravo podatak koji možete utvrditi u tablici grupiranih rezultata u primjeru 2.3.

b. Koliko učenika je visoko najmanje 135,5 cm?

Podatak koliko učenika je visoko najmanje 135,5 cm već imamo utvrđen na grafikonu: njih 40,9 % (tj. od z=0,23 do bližeg kraja distribucije imamo 40,9% površine ispod krive). 40,9% od ukupno 25 učenika iznosi 10 učenika.

c. Koja je granična vrijednost koja odvaja 40% najnižih učenika od ostalih?

Granični z rezultat koji odvaja 40% najnižih učenika također možemo dobiti vrlo jednostavno – samo ćemo tablicu *p* vrijednosti ispod standardne normalne krive korsititi u obratnom smjeru: proporciji od 0,4000 površine ispod krive odgovara približna z-vrijednost od - 0,25 (z vrijednost je negativna obzirom da se nalazi ispod aritmetičke sredine).

Još jednom, izražavajući formulu za z rezultate preko X dolazimo do jednačine:

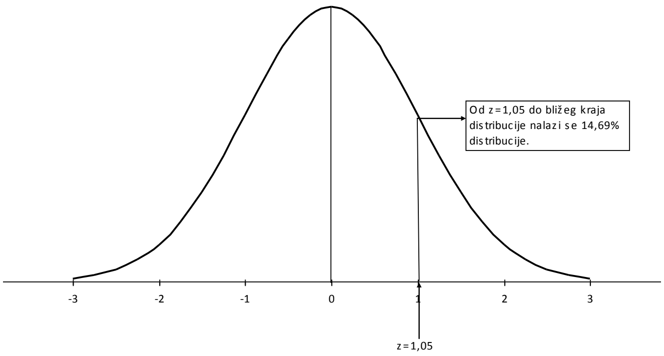
$X = M + zs = 135{,}08 + (-0{,}25)x1{,}83 = 134{,}62$

40% učenika IIIc odjeljenja niže je od 134,62 cm (na osnovu ogive u primjeru 2.3 mi smo ovu granicu postavili na visinu od oko 134,75 cm).

d. Koji procenat učenika je visok najmanje 137 cm?

U konkretnoj raspodjeli visina učenika IIIc odjeljenja, visini od 137 cm odgovara z rezultat od:

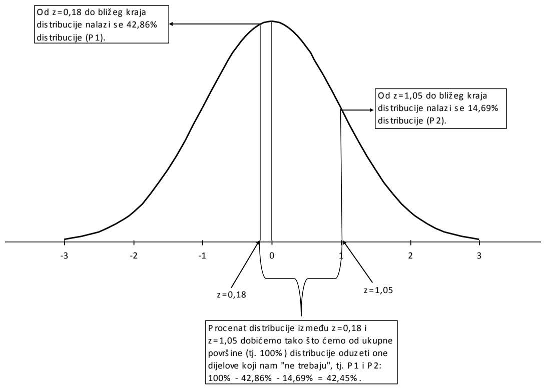
$$z = \frac{X - M}{s} = \frac{137 - 135{,}08}{1{,}83} = 1{,}05$$



Dakle, 14,69% učenika IIIc je više od 137 cm.

e. Koliko učenika je visoko između 135,5 i 136,5 cm?

Sa prethodnog grafika vidimo da je 14,69% učenika više od 137 cm. Ukoliko izračunamo odgovarajući z rezultat za visinu od 134,75 cm (a on iznosi -0,18) i pozicioniramo ga na grafu, dobit ćemo prikaz desno.



42,45% učenika IIIc odjeljenja (ili njih 11) visoko je između 134,75 cm i 137 cm.

ZADACI

1. U primjerima 2.3, 3.4 i 4.5 utvrdili smo karakteristike distribucije visina učenika IIIc odjeljenja (između ostalog i to da distribucija ima M = 135,08 cm i s = 1,83 cm). Utvrdite u kojem rasponu rezultata se u ovoj raspodjeli nalazi:

   a. 68,26% središnjih vrijednosti.
   b. 95,44% središnjih vrijednosti.
   c. 99,73% središnjih vrijednosti.

2. Iste raspone utvrdite i za distribucije iz zadataka: 2.1, 2.3, 3.1, 3.2, 3.3 i 3.4.

3. Distribucija rezultata ima M = 28 i s = 4. Utvrdite z-vrijednosti za sljedeće rezultate koji su izvučeni iz ove distribucije:

   a. X = 28.
   b. X = 32.
   c. X = 36.
   d. X = 24.
   e. X = 16.

   Objasnite zašto ste dobili upravo te z-vrijednosti.

4. Distribucija rezultata ima M = 95 i s = 12. Utvrdite vrijednosti sirovih rezultata kojima u ovoj distribuciji odgovaraju sljedeće z-vrijednosti:

   a. z = 2,5.
   b. z = -1,8.
   c. z = 0.
   d. z = 1.
   e. z = 2.
   f. z = -3.

   Obratite pažnju na rezultate koje ste dobili u zadacima c., d., e. i f. Objasnite zašto ste dobili upravo te sirove vrijednosti.

5. Distribucija rezultata ima s = 9. Ako rezultatu X = 36 u ovoj distribuciji odgovara z = -2,2, koliko iznosi aritmetička sredina distribucije?

6. Distribucija ima M = 41. Ako rezultatu X = 28 u ovoj distribuciji odgovara z = - 3,2, koliko iznosi standardna devijacija distribucije?

7. Normalna distribucija rezultata ima M = 49 i s = 7. Utvrdite z-vrijednosti za sljedeće rezultate koji su izvučeni iz ove distribucije:

   a. X = 26.
   b. X = 21.
   c. X = 40.
   d. X = 55.
   e. X = 63.
   f. X = 30.

   Za svaki od navedenih rezultata utvrdite procenat nižih i viših rezultata u distribuciji.

8. Normalna distribucija rezultata ima M = 55 i s = 6. Utvrdite z-vrijednosti za sljedeće rezultate koji su izvučeni iz ove distribucije:

   a. X = 48
   b. X = 41
   c. X = 54
   d. X = 57
   e. X = 62
   f. X = 65

   Za svaki od navedenih rezultata utvrdite procenat rezultata između datog rezultata i aritmetičke sredine distribucije.

9. Distribuciju visina učenika IIIc odjeljenja (iz primjera 2.3, 3.4 i 4.5) predstavite pomoću histograma. Zatim sve rezultate iz distribucije pretvorite u z-rezultate te dobivenu z-distribuciju prikažite na histogramu. Da li se oblik distribucije promijenio? Objasnite.

10. Obratite pažnju na predznak odgovarajućih z-rezultata za visine dva najniža i dva najviša učenika u IIIc odjeljenju. Šta vam govori predznak utvrđenih z-rezultata?

11. Za potrebe regrutiranja novih vojnika, Ministarstvo odbrane je testom inteligencije testiralo ukupno 1350 ispitanika. Dobivena je normalna distribucija rezultata sa sljedećim deskriptivnim vrijednostima: M = 202 i s = 38. Utvrdite z-vrijednosti za ispitanike sa sljedećim rezultatima:

   a. $X_1 = 115$
   b. $X_2 = 236$
   c. $X_3 = 302$
   d. $X_4 = 345$
   e. $X_5 = 98$
   f. $X_6 = 152$
   g. Za svakog od navedenih ispitanika utvrdite broj ispitanika koji su ostvarili bolji rezultat.
   h. Utvrdite broj ispitanika koji se po rezultatu na testu inteligencije nalaze između ispitanika 1 i 4; 2 i 5; 2 i 4; 1 i 6.
   i. Utvrdite koji (sirovi) rezultat dijeli distribuciju na 50% slabijih i 50% boljih ispitanika.

j. Ukoliko Ministarstvo odbrane želi regrutovati samo 250 najboljih ispitanika, koji (sirovi) granični rezultat će koristiti prilikom selekcije kandidata?
k. Ukoliko Ministarstvo odbrane želi regrutovati 35% najboljih kandidata, koji (sirovi) granični rezultat će koristiti prilikom selekcije kandidata?

12. Velika programerska kompanija želi zaposliti 25 novih radnika. Na konkurs se prijavilo ukupno 89 kandidata. Svi kandidati su testirani na Testu matematike i Testu informatike. Distribucije utvrđenih rezultata na ova dva testa imaju sljedeće deskriptivne pokazatelje:

    Test matematike: M = 125 i s = 17;

    Test informatike: M = 42 i s = 8.

a. Na intervju za posao biće pozvani svi kadidati koji su na Testu matematike ostvarili najmanje 145 bodova. Koliko kandidata će biti pozvano na intervju?
b. Kompanija je zaposlila sve intervjuirane kandidate. Obzirom da je ostao određeni broj nepopunjenih radnih mjesta, rukovodioci sektora za ljudske resurse odlučili su organizirati informatičku obuku za najbolje među preostalim (nezaposlenim) kandidatima. Na trening su odlučili pozvati 20% kandidata koji su na Testu matematike ostvarili najbolje rezultate kada se iz početne skupine isključe kandidati koji su već dobili posao. Koji rezultat na Testu matematike će biti korišten kao granični prilikom odlučivanja koga pozvati na trening?
c. Kandidat A je na Testu matematike ostvario rezultat 120, a na Testu informatike 48. Kandidat B je na Testu matematike ostvario 131 bodova, a na Testu informatike 42 boda. Ako kompanija oba testa smatra jednako važnim, koji od ova dva kandidata bi trebao imati prednost pri zapošljavanju?

13. U medicini se smatra da je normalna vrijednost sistoličkog krvnog pritiska kod odraslih muškaraca 120 mm/Hg. U velikom epidemiološkom istraživanju (provedenom na području cijele države) na reprezentativnom uzorku od N = 3.500 odraslih muškaraca utvrđeno je da se vrijednosti sistoličkog krvnog pritiska normalno distribuiraju sa M = 126 mm/Hg i s = 11 mm/Hg.

a. Koji procenat odraslog muškog stanovništva ima sistolički krvni pritisak veći od normalnog? Ako u datoj državi živi ukupno 4.250.827 odraslih muških stanovnika, koliko njih ima sistolički krvni pritisak veći od normalnog?
b. Ako je država u liječenju pacijenata sa sistoličkim krvnim pritiskom većim od 160 mm/Hg dužna učestvovati sa 3.80 Eur mjesečno po pacijentu, kolike mjesečne troškove liječenja ove bolesti može očekivati ministar zdravstva date države?

14. Prema novom zakonu o socijalnoj pomoći, domaćinstva u državi podijeljenja su u 6 kategorija prema visini mjesečnih primanja:

Kategorija I: do 120 KM (mjesečno u domaćinstvu);

Kategorija II: od 121 do 200 KM;

Kategorija III: od 201 KM do 350 KM;

Kategorija IV: od 351 KM do 600 KM;

Kategorija V: od 601 KM do 900 KM;

Kategorija VI: iznad 901 KM.

Prema istom zakonu, domaćinstva Kategorije I mjesečno će dobivati socijalnu pomoć u iznosu od 150 KM; domaćinstva Kategorije II mjesečno će dobivati 95 KM; domaćinstva kategorije III mjesečno će dobivati 50 KM socijalne pomoći.

Ako u državi ima ukupno 2.256.897 registriranih domaćinstava te ako se njihova mjesečna primanja raspoređuju u obliku normalne raspodjele sa M = 1.486 KM i s = 365 KM, koliko novca će država mjesečno plaćati socijalno ugroženim domaćinstvima?

15. Utvrdite vrijednosti skjunisa i kurtozisa za distribuciju visina učenika IIIc odjeljenja (primjeri 2.3, 3.4 i 4.5)? Šta možete zaključiti o obliku ove distribucije na temelju dobivenih vrijednosti?

16. Za sljedeće podatke odredite vrijednosti skjunisa i kurtozisa:

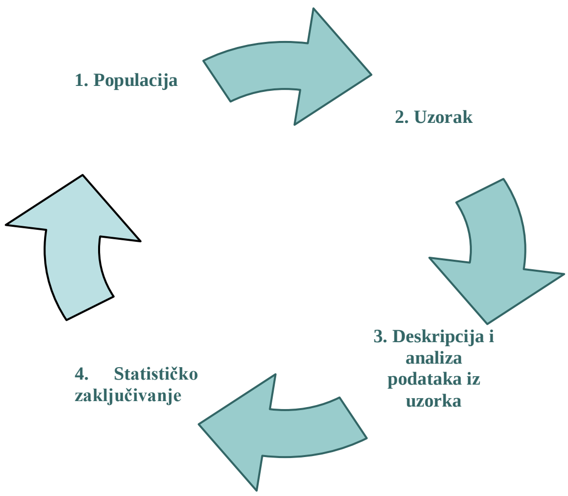
11	4	9	7	5	8
3	5	2	2	1	12
5	4	6	11	9	5
6	5	7	5	6	4
8	3	7	6	8	

Šta na osnovu dobivenih vrijednosti možete zaključiti o obliku raspodjele? Svaki rezultat iz prethodne distribucije uvećajte za 2 te ponovo izračunajte skjunis i kurtozis. Šta se dogodilo s oblikom distribucije? Objasnite.

7. Standardna pogreška aritmetičke sredine

Istraživanja u psihologiji u pravilu provodimo na **uzorcima** određenih populacija. Razlozi za to su ekonomičnost (ušteda novca i vremena) i praktičnost (npr. mjerenjem se ponekad uništava proizvod kao u slučaju kontrole kvalitete različitih industrijskih proizvoda). Međutim, iako istraživanja ne obuhvataju sve članove određene populacije koja je predmet našeg interesiranja (npr. svu djecu sa teškoćama u učenju, sve osobe sa depresivnim poremećajima, svu talentiranu djecu, sve građane sa pravom glasa itd.), možemo donositi sasvim valjane zaključke i na temelju vrijednosti koje smo dobili na uzorcima. Naime, na osnovu rezultata dobivenih na uzorcima izvode se zaključci o vrijednostima populacije.

Postupak donošenja zaključaka o populaciji na osnovu rezultata dobivenih na uzorku naziva se **statističko zaključivanje.** Dijagram ispod ilustrira proces statističkog zaključivanja. Iz populacije formiramo uzorak kojeg numeričkim i grafičkim metodama opisujemo. Određenim postupcima, o kojima će biti govora u nastavku Priručnika, donosimo zaključak o populaciji. U ovom poglavlju bavit ćemo se procjenom aritmetičke sredine populacije na osnovu vrijednosti koju smo dobili na uzorku.



Populaciju čine svi članovi neke grupe s određenom karakteristikom koju mjerimo. Deskriptivne vrijednosti populacije nazivamo **parametrima**: μ – aritmetička sredina populacije i σ – standardna devijacija populacije.

Uzorak je podskup populacije na kojem se vrši istraživanje. Da bi se rezultati dobiveni istraživanjem na uzorku mogli generalizirati na populaciju iz koje je uzorak izvučen, uzorak mora biti **reprezentativan**. Deskriptivne statističke vrijednosti (M i s) koje smo dobili na uzorku nazivamo «procjenama» parametara (procjene prave aritmetičke sredine i prave standardne devijacije) ili **statisticima**.

Obzirom da su statistici procjene parametara, prilikom statističkog zaključivanja izlažemo se pogrešci; u primjeru aritmetičke sredine ova pogreška naziva se **standardna pogreška aritmetičke sredine**.

Postupak donošenja zaključka o aritmetičkoj sredini populaciji na osnovu aritmetičke sredine uzorka, tj. statistika uzorka, kao i logika standardne pogreške aritmetičke sredine temelji se na nekoliko principa raspodjele aritmetičkih sredina velikog broja uzoraka (dakle, velikog broja statistika) oko jedne zajedničke (centralne) vrijednosti. U primjeru ispod demonstriran je postupak donošenja zaključka o aritmetičkoj sredini populacije i smisao standardne pogreške aritmetičke sredine.

Iz populacije veličine N=10000 sa deskriptivnim vrijednostim μ=49,84 i σ=9,88, metodom slučajnog odabira formirali smo određeni broj uzoraka veličine N=5, N=100 i N=5000. Svaki put kada bi formirali uzorak izračunali bi i njegovu aritmetičku sredinu. Teoretski, za svaku veličinu uzorka mogli smo formirati beskonačno mnogo uzoraka, odnosno aritmetičkih sredina. Za potrebe demonstracije dovoljno je da broj uzoraka, tj. aritmetičkih sredina bude 20. U tabeli ispod navedene su aritmetičke sredine uzoraka različitih veličina dobivenih u eksperimentu kao i aritmetičke sredine i standardne devijacije aritmetičkih sredina uzoraka (M i s).

Tabela 7.1: Aritmetičke sredine uzoraka veličina N=5, N=100 i N=5000

Redni broj uzorka	Veličina uzorka		
	N=5	N=100	N=5000
1	48,8	50,5	49,8
2	49,4	50,9	50,0
3	42,8	50,9	49,9
4	47,4	51,4	49,8
5	43,2	50,9	49,7
6	50,0	50,2	49,9
7	48,8	51,3	49,8
8	56,6	52,8	49,7
9	53,2	50,1	49,8
10	44,8	49,3	49,7
11	54,8	52,4	49,8
12	47,2	49,1	49,8
13	51,0	50,1	49,9
14	53,4	50,2	49,8
15	54,4	48,7	49,9
16	50,2	48,9	49,8
17	50,6	49,8	49,7
18	49,2	51,1	49,7
19	48,6	50,3	49,9
20	48,2	51,4	50,1
M	***49,6***	***50,5***	***49,8***
s	***3,7***	***1,1***	***0,1***

Na osnovu aritmetičkih sredina prikazanih u tabeli 7.1 možemo primjetiti da se u slučaju uzoraka veličine N=5 dobivaju aritmetičke sredine u najvećem rasponu (od 42,8 do 56,6), dok je u slučaju uzoraka veličine N=5000 raspon znatno manji (od 49,1 do 50,1). Uostalom, standardne devijacije aritmetičkih sredina potvrđuju ono što možemo primjetiti pregledom vrijednosti aritmetičkih sredina uzoraka. Najniža standardna devijacija dobivena je za uzorke veličine N=5000, a najveća za uzorke veličine N=5. Nadalje, ako uporedimo aritmetičke sredine uzoraka različitih veličina, možemo primjetiti da su aritmetičke sredine uzoraka veličine N=5000 u pravilu blizu pravoj aritmetičkoj sredini, dok se tek poneka aritmetička sredina za uzorke veličine N=5 približava pravoj aritmetičkoj sredini (npr. 50,0).

Ako bi nastavili eksperiment i formirali znatno veći broj aritmetičkih sredina uzoraka, mogli bi vidjeti da se one distribuiraju prema normalnoj raspodjeli. Aritmetička sredina aritmetičkih sredina bila bi jednaka pravoj aritmetičkoj sredini. Mogli bi se uvjeriti da vrijede sljedeća pravila:

1. Aritmetička sredina aritmetičkih sredina uzoraka (populacija aritmetičkih sredina uzoraka) iste veličine jednaka je «pravoj» aritmetičkoj sredini, tj. aritmetičkoj sredini populacije.

$$\mu_M = \mu$$

2. Varijanca populacije aritmetičkih sredina uzoraka jednaka je varijanci originalne populacije, podjeljenoj s veličinom uzorka.

$$\sigma^2_M = \sigma^2 / N$$

3. Varijance uzoraka čine takvu raspodjelu oko prave varijance da im aritmetička sredina odgovara pravoj varijanci:

$$\mu_{s^2} = \sigma^2$$

4. Standardna devijacija aritmetičkih sredina uzoraka oko prave aritmetičke sredine populacije je standardna pogreška aritmetičke sredine.

$$s_M = \sigma_M$$

Dakle, standardna pogreška aritmetičke sredine uzorka zapravo je standardna devijacija aritmetičkih sredina uzoraka oko prave aritmetičke sredine populacije.

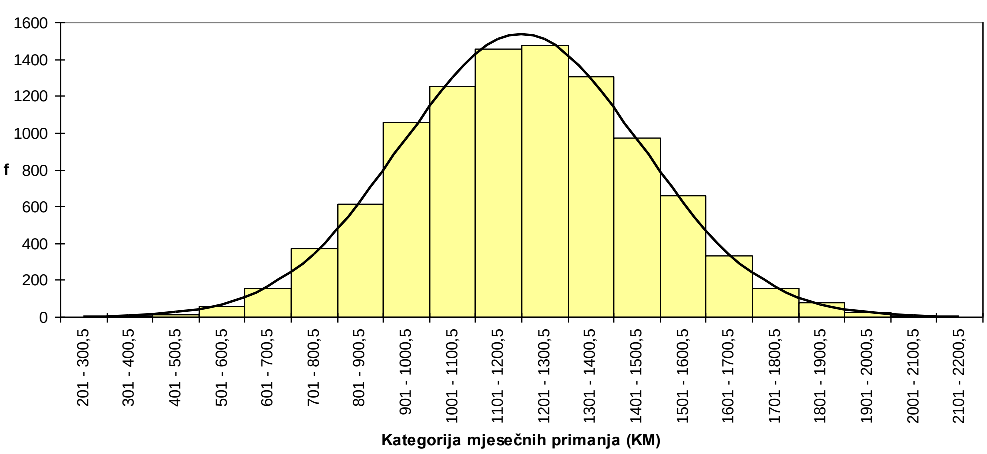
PRIMJER 7.1

Istraživač je u jednom manjem mjestu zaposlene građane pitao o visini njihove mjesečne zarade. Na taj način dobio je bazu podataka sa iznosima mjesečnih plata za N=10000 ispitanika [obzirom da se u ovom skupu nalaze sve zaposlene osobe koje žive u tom mjestu, ovaj skup nazivamo populacijom (zaposlenih osoba tog mjesta), a vrijednosti koje utvrdimo na ovim podacima – parametrima]. U tabeli 7.1.1 prezentirane su vrijednosti parametara ove populacije.

Tabela 7.1.1: Populacijske deskriptivne vrijednosti varijable *Mjesečna primanja* (N=10000)

μ	Medijan	σ	Skjunis	Kurtozis	Raspon	Minimum	Maximum
1200,47	1201,00	259,63	0,02	-0,08	1979,94	204,90	2184,84

Prema ovim pokazateljima možemo zaključiti da je riječ o normalno distribuiranoj varijabli, što nam potvrđuje i donja slika:



Slika 7.1.1.: Distribucija mjesečnih primanja u populaciji (N=10000)

Ovim primjerom želimo demonstrirati da do zadovoljavajuće **procjene populacijske aritmetičke sredine** (μ=1200,47) možemo doći i preko uzorka. Prednost rada sa uzorcima umjesto sa populacijama jasna je ako razmislimo o uštedama (u vremenu i novcu) koje možemo ostvariti ako anketiramo uzorak od npr. 300 zaposlenih građana, umjesto njih svih 10000.

Teoretski, iz naše populacije možemo **po slučaju** izvući jako **veliki broj različitih uzoraka iste veličine**, npr. veličine n=30 (ili npr. n=300 ili npr. n=800); kada kažemo "različiti uzorci iste veličine" mislimo na uzorke koji sadrže jednak broj ispitanika, ali koji se razlikuju u bar jednom od tih ispitanika. Kako bismo detaljnije demonstrirali principe izložene u uvodnom dijelu ovog poglavlja za početak ćemo iz ove populacije izvući 50 uzoraka veličine n=30. Za svaki od formiranih uzoraka izračunat ćemo aritmetičku sredinu čime ćemo dobiti distribuciju od ukupno 50 aritmetičkih sredina. Dobivene aritmetičke sredine i standardne devijacije za svaki formirani uzorak prezentirane su u tabeli 7.1.2.

Tabela 7.1. 2: Aritmetičke sredine i standardne devijacije uzoraka

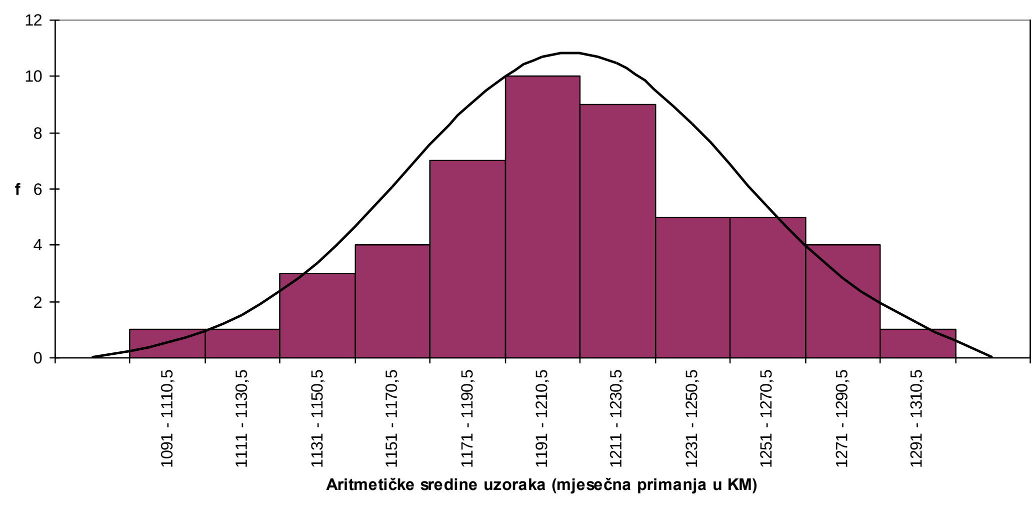
Redni broj uzorka	M	s
1	1159,21	237,79
2	1214,20	239,18
3	1195,01	268,70
4	1278,40	238,58
5	1161,32	217,88
6	1240,94	212,18
7	1215,48	247,92
8	1176,50	211,59
9	1221,17	214,70
10	1200,65	253,94
11	1216,56	262,68
12	1217,22	235,36
13	1278,56	273,00
14	1199,91	273,00
15	1155,46	300,92
16	1206,81	257,02
17	1183,36	269,74
18	1149,56	262,79
19	1193,08	216,72
20	1252,95	251,02
21	1228,54	246,04
22	1232,06	268,92
23	1164,10	269,81
24	1227,48	252,53
25	1193,44	237,11
26	1255,06	215,68
27	1171,73	211,25
28	1221,07	252,71
29	1296,07	240,02
30	1280,11	301,00
31	1191,20	228,79
32	1237,17	282,83
33	1281,10	165,10
34	1207,93	225,83
35	1174,78	252,89
36	1185,07	265,55
37	1187,82	285,36
38	1174,19	274,89
39	1264,19	234,01
40	1239,55	257,59
41	1126,24	288,01
42	1265,88	291,30
43	1201,25	324,63
44	1245,73	287,20
45	1202,71	284,12
46	1140,33	194,30
47	1093,32	223,49
48	1213,18	238,48
49	1149,27	266,60
50	1255,80	279,99
M	1208,45	251,81
s	44,10	

Prije svega, primjećujemo da niti jedna od utvrđenih aritmetičkih sredina formiranih uzoraka ne odgovara (u potpunosti) aritmetičkoj sredini populacije. Sve aritmetičke sredine uzoraka više ili manje variraju oko prave populacijske aritmetičke sredine. Ova pojava je sasvim očekivana – aritmetičku sredinu uzorka koja u potpunosti odgovara aritmetičkoj sredini populacije po slučaju možemo očekivati iznimno rijetko.

Ipak, ukoliko populacijska distribucija ima oblik normalne raspodjele te ukoliko iz te populacije po slučaju formiramo dovoljan broj uzoraka iste veličine i izračunamo njihove aritmetičke sredine, primjetićemo da se te aritmetičke sredine uzoraka grupiraju oko jedne centralne vrijednosti u obliku normalne raspodjele. Centralna vrijednost oko koje se te aritmetičke sredine grupiraju odgovara pravoj populacijskoj aritmetičkoj sredini (vidi *pravilo 1* u uvodnom dijelu).

Kako smo mi iz naše početne, “normalne” populacije od 10000 članova formirali relativno veliki broj uzoraka (50) iste veličine (n=30), možemo empirijski provjeriti gornju tvrdnju. Dakle, iz tabele 7.1.2 vidimo da zajednička aritmetička sredina aritmetičkih sredina svih uzoraka, tj. **aritmetička sredina distribucije aritmetičkih sredina uzoraka** iznosi M=1208,45[3]. Ova vrijednost bliska je populacijskoj aritmetičkoj sredini od μ=1200,47. Slika 7.1.2 ilustrira kako se aritmetičke sredine uzoraka distribuiraju oko svoje zajedničke aritmetičke sredine.

Slika 7.1.2: Distribucija aritmetičkih sredina 50 uzoraka veličine n=30



Kao što se vidi, gornja distribucija približno odgovara teoretskoj normalnoj distribuciji.

Prema *pravilu 2*, varijanca distribucije aritmetičkih sredina uzoraka jednaka je varijanci originalne populacije, podjeljenoj s veličinom uzorka, tj:

$$\mathbf{s^2_M = \sigma^2 / N}$$

[3] U uvodnom dijelu, za označavanje deskriptivnih vrijednosti distribucije aritmetičkih sredina uzoraka korišteni su simboli za označavanje populacijskih vrijednosti (μ i σ) obzirom da se misli na teoretsku, beskonačno veliku populaciju aritmetičkih sredina beskonačno velikog broja uzoraka iste veličine koji se mogu izvući iz beskonačno velike početne populacije. Obzirom da u ovom primjeru radimo sa konkretnom distribucijom aritmetičkih sredina 50 uzoraka (što je, u suštini, samo uzorak svih mogućih aritmetičkih sredina uzoraka koji se mogu izvući iz početne populacije), za označavanje deskriptivnih vrijednosti ove distribucije koristićemo se oznakama koje se i inače koriste za uzorke – M i s.

Obzirom da se standardna devijacija dobiva jednostavnim korjenovanjem vrijednosti varijance, gornja formula se može izraziti i za standardnu devijaciju:

$$\sqrt{\mathbf{s}^2_{\mathbf{M}}} = \sqrt{\boldsymbol{\sigma}^2} / \mathbf{N}$$

tj.

$$\mathbf{s_M} = \boldsymbol{\sigma} / \sqrt{\mathbf{N}}$$

Dakle, ukoliko standardnu devijaciju populacije podijelimo korjenom broja ispitanika u uzorku dobit ćemo vrijednost standardne devijacije distribucije aritmetičkih sredina uzoraka. Ovu tvrdnju provjerit ćemo na našem primjeru.

Standardna devijacija populacije iznosi σ=259,63 (vidi tabelu 7.1.1). Ako ovu vrijednost podjelimo korjenom broja ispitanika u uzorku (n=30) dobit ćemo:

$$s_M = 259{,}63 / \sqrt{30} = 47{,}40$$

Ova vrijednost je slična (mada ne identična) standardnoj devijaciji aritmetičkih sredina uzoraka oko njihove zajedničke standardne devijacije koja iznosi 44,10 (vidi tabelu 7.1.2).

Prema *pravilu 3* varijance uzoraka čine takvu raspodjelu oko prave varijance da im aritmetička sredina odgovara pravoj varijanci, odnosno:

$$\mathbf{M_{s^2}} = \boldsymbol{\sigma}^2$$

Ako i ovu formulu "prevedemo" u termine standardne devijacije dobit ćemo:

$$\mathbf{M_s} = \boldsymbol{\sigma},$$

tj. prosjek standardnih devijacija uzoraka jednak je standardnoj devijaciji populacije. Provjerimo to na našem primjeru – u tabeli 7.1.2 možemo vidjeti da je prosječna standardna devijacija 50 uzoraka M_s=251,81 što je vrijednost slična populacijskoj standardnoj devijaciji σ = 259,63.

Primjetili ste da u dosadašnjem dijelu ovog Proglavlja– kada govorimo o odnosu populacijskih vrijednosti i vrijednosti koje izračunavamo za distribuciju aritmetičkih sredina uzoraka – često koristimo odrednice „približno“, „blisko“, „slično“, tj. da nije demonstrirana apsolutna tačnost tri razmatrana pravila. Razlog za to je što ova pravila vrijede za teoretske, beskonačno velike populacije iz kojih po slučaju izvlačimo beskonačno veliki broj uzoraka iste veličine. Ipak, i kada radimo sa realnim populacijama i brojem uzoraka koji je relativno mali (u našem primjeru samo 50) vidimo da navedena pravila „prilično dobro funkcioniraju“.

Jedan od načina da se još više približimo populacijskim vrijednostima (pored izvlačenja većeg broja uzoraka iz populacije) jeste da iz populacije izvlačimo veće uzorke. Slično kao i u uvodnom dijelu, to ćemo demonstrirati tako što ćemo iz naše populacije izvući po 50 slučajnih uzoraka veličine n=300, odnosno n=800.

U tabeli 7.1.3 načinjena je komparacija triju formiranih distribucija od po 50 uzoraka (veličine n=30, n=300 i n=800); s ciljem usporedbe, prezentirani su i populacijski parametri.

Tabela 7.1.3.

Redni broj uzorka	n = 30		n = 300		n = 800	
	M	s	M	s	M	s
1	1159,21	237,79	1212,80	252,24	1205,77	258,32
2	1214,20	239,18	1171,57	246,16	1187,69	268,43
3	1195,01	268,70	1198,13	246,79	1209,88	248,12
4	1278,40	238,58	1213,14	259,72	1193,99	258,77
5	1161,32	217,88	1235,48	266,99	1197,00	253,27
6	1240,94	212,18	1214,32	270,36	1190,70	244,02
7	1215,48	247,92	1187,89	249,13	1201,27	252,40
8	1176,50	211,59	1186,08	263,07	1207,79	259,31
9	1221,17	214,70	1185,73	262,41	1190,56	264,03
10	1200,65	253,94	1183,70	256,35	1213,49	248,96
11	1216,56	262,68	1193,68	248,90	1216,45	248,39
12	1217,22	235,36	1216,49	254,39	1194,98	254,73
13	1278,56	273,00	1201,29	254,61	1208,70	256,15
14	1199,91	273,00	1191,34	260,68	1193,45	257,56
15	1155,46	300,92	1182,94	265,45	1201,16	256,61
16	1206,81	257,02	1182,29	275,51	1202,33	265,18
17	1183,36	269,74	1185,35	277,03	1209,66	264,35
18	1149,56	262,79	1187,75	277,65	1184,40	258,92
19	1193,08	216,72	1203,41	252,27	1212,43	252,60
20	1252,95	251,02	1192,18	269,66	1194,32	261,26
21	1228,54	246,04	1215,83	258,51	1191,88	262,74
22	1232,06	268,92	1200,76	241,56	1194,11	258,97
23	1164,10	269,81	1190,14	279,29	1206,70	256,39
24	1227,48	252,53	1228,05	251,95	1202,84	256,85
25	1193,44	237,11	1204,94	262,48	1191,34	253,98
26	1255,06	215,68	1190,77	254,68	1200,20	268,44
27	1171,73	211,25	1188,23	251,53	1195,80	245,25
28	1221,07	252,71	1205,05	264,23	1204,73	258,14
29	1296,07	240,02	1199,41	256,49	1194,88	266,50
30	1280,11	301,00	1223,49	265,27	1214,20	252,56
31	1191,20	228,79	1220,95	286,11	1211,56	261,39
32	1237,17	282,83	1192,60	283,42	1193,05	259,50
33	1281,10	165,10	1189,22	256,46	1202,08	263,06
34	1207,93	225,83	1206,08	255,90	1208,79	250,54
35	1174,78	252,89	1202,25	262,98	1212,08	257,34
36	1185,07	265,55	1190,49	264,05	1209,52	257,42
37	1187,82	285,36	1197,87	268,06	1202,61	266,85
38	1174,19	274,89	1214,96	243,33	1205,10	261,60
39	1264,19	234,01	1189,62	270,30	1209,66	257,23
40	1239,55	257,59	1221,46	250,39	1205,57	268,07
41	1126,24	288,01	1163,68	250,56	1192,17	253,91
42	1265,88	291,30	1202,90	255,26	1203,01	256,18
43	1201,25	324,63	1185,99	260,71	1198,76	261,86
44	1245,73	287,20	1209,04	253,01	1189,05	275,81
45	1202,71	284,12	1210,58	268,66	1204,23	260,74
46	1140,33	194,30	1207,22	256,21	1194,54	269,08
47	1093,32	223,49	1192,34	260,13	1189,54	253,00
48	1213,18	238,48	1226,84	270,23	1196,20	259,48
49	1149,27	266,60	1186,71	264,93	1210,77	258,50
50	1255,80	279,99	1200,06	260,13	1185,91	267,40
M	1208,45	251,81	1199,66	260,72	1200,74	258,60
s	44,10		15,03		8,45	

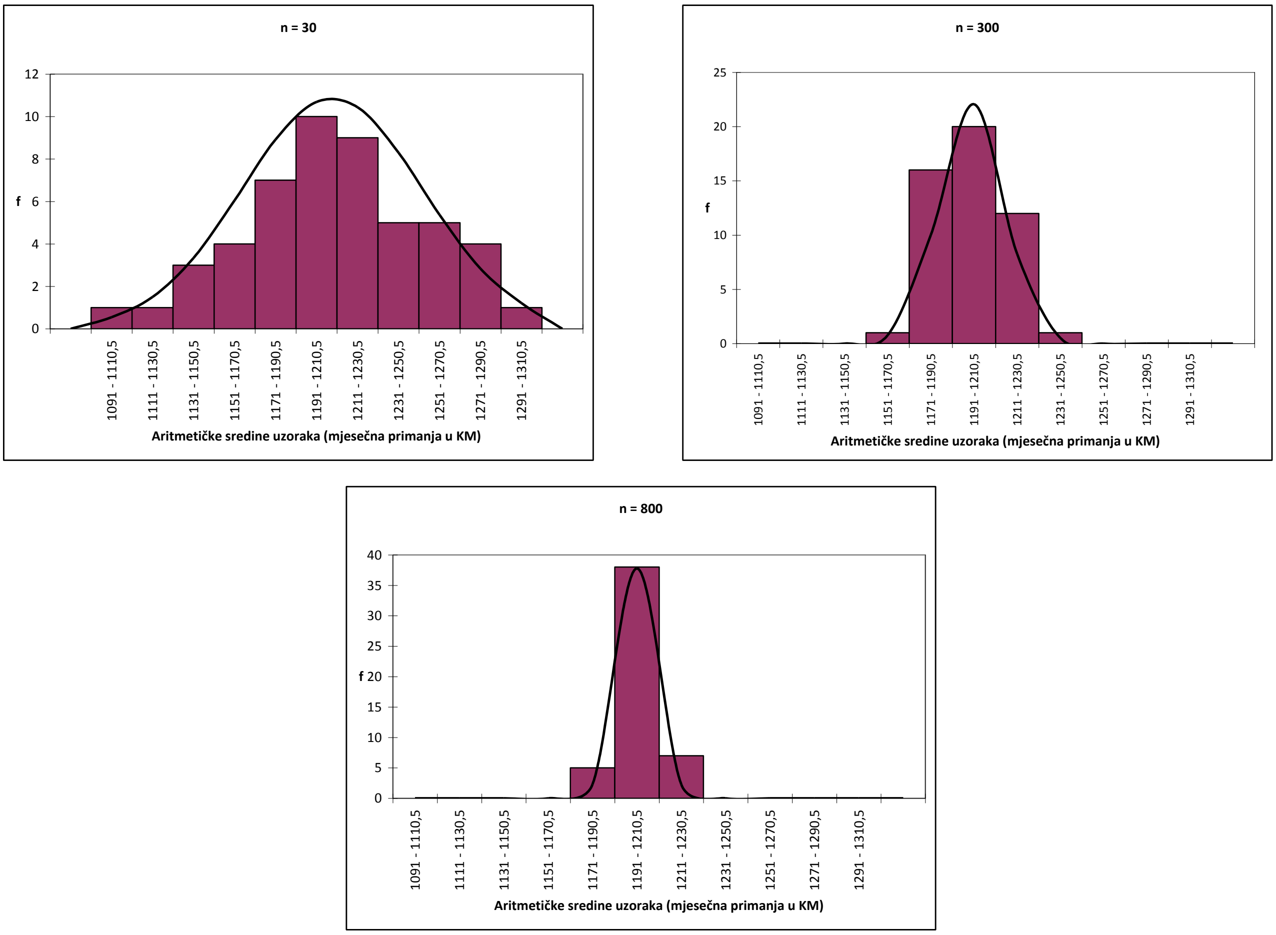
Populacijske vrijednosti:

μ	1200,47	
σ	259,63	

Kao i u primjeru u uvodnom dijelu, primjetno je da distribucija sa najvećim uzorcima (n=800) ima aritmetičku sredinu koja je najbliža pravoj aritmetičkoj sredini populacije. Također, ova distribucija pokazuje najmanje raspršenje (s=8,45), odnosno aritmetičke sredine ovih uzoraka se najuže grupiraju oko svoje zajedničke aritmetičke sredine (koja je, još jednom, vrlo bliska populacijskoj aritmetičkoj sredini – razlika je samo 0,27 KM).

Razlike između distribucija aritmetičkih sredina uzoraka različite veličine još su uočljivije na slici 7.1.3.

Slika 7.1.3.: Usporedba distribucija 50 aritmetičkih sredina uzoraka veličina n=30, n=300 i n=800



Za vježbu, provjerite koja od tri predstavljene distribucije najtačnije ilustriraju pravila 2 i 3 iz uvodnog dijela poglavlja.

Sada ćemo se koncentrirati na **pravilo 4** koji je u osnovi postupka statističkog zaključivanja i koji glasi: **Standardna devijacija aritmetičkih sredina uzoraka oko prave aritmetičke sredine populacije je standardna pogreška aritmetičke sredine:**

$$s_M = \sigma_M.$$

Podsjetimo se još jednom da je statističko zaključivanje postupak zaključivanja o populaciji na temelju vrijednosti utvrđenih na uzorku. Osnovu za ovakvo zaključivanje pruža nam upravo standardna **pogreška aritmetičke sredine.**

Standardna pogreška se računa prema izrazu:

$$s_M = s / \sqrt{N}$$

pri čemu je:

s – standardna devijacija uzorka,

N – veličina uzorka.

Primjećujete da se ova formula temelji na vrijednostima koje dobijamo na uzorku. To je sasvim razumljivo: iz gore navedenih razloga, istraživanja vrlo rijetko provodimo na cijelim populacijama. Ono što radimo mnogo češće jeste da iz populacije po slučaju izvučemo jedan uzorak i preko njegovih deskriptivnih vrijednosti pokušavamo zaključiti o pravim populacijskim vrijednostima.

U praksi bi to izgledalo ovako: iz naše početne populacije od 10000 zaposlenih građana (o čijim mjesečnim primanjima ne znamo ništa – što je i razlog provođenja istraživanja) bismo po slučaju izvukli samo jedan uzorak veličine, npr., n=800 ispitanika. Nakon što smo statistički obradili odgovore ovih 800 ispitanika na pitanje koliko mjesečno zarađuju, dobili smo aritmetičku sredinu mjesečnih primanja od M=1202,54 KM i standardnu devijaciju s=259,76 KM. Na temelju ovih vrijednosti želimo procijeniti prosječnu mjesečnu platu naših 10000 građana.

Ono što teoretski znamo jeste sljedeće:

1. naš uzorak samo je jedan od svih mogućih uzoraka veličine n=800 koji se mogu izvući iz naše populacije;

2. dakle, aritmetička sredina našeg uzorka je samo jedna od svih mogućih aritmetičkih sredina svih mogućih uzoraka veličine n=800 koji se mogu izvući iz naše populacije, a koje čine teoretsku distribuciju aritmetičkih sredina uzoraka;

3. obzirom da (pretpostavljamo) da se mjesečna primanja naših 10000 građana distribuiraju u obliku normalne distribucije, i teoretska distribucija aritmetičkih sredina uzoraka ima oblik normalne distribucije;

4. aritmetička sredina ove teoretske distribucije aritmetičkih sredina uzoraka jednaka je pravoj populacijskoj aritmetičkoj sredini (vidi pravilo 1 i gornju diskusiju)

5. standardna devijacija ove distribucije aritmetičkih sredina uzoraka jednaka je standardnoj devijaciji originalne populacije i računa se prema formuli:

   $$\sigma_M = \sigma / \sqrt{N}$$

   (vidi pravilo 2 i gornju diskusiju);

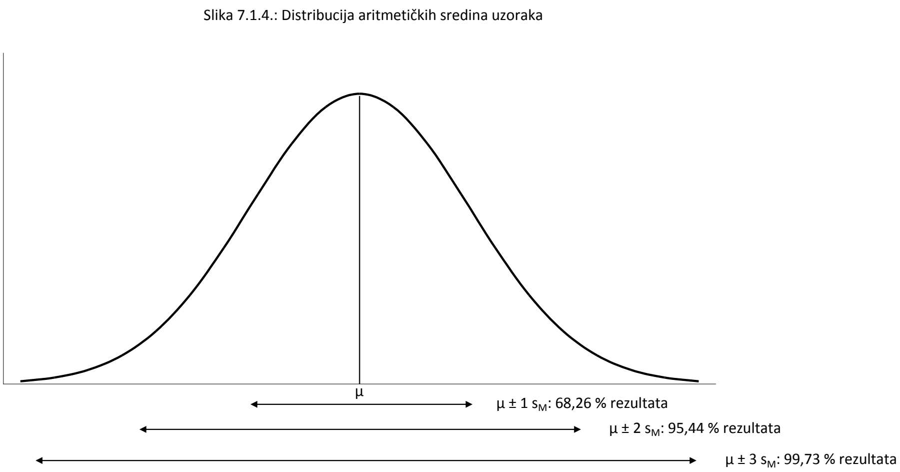
6. obzirom da ne znamo koliko iznosi populacijska standardna devijacija, najbolja aproksimacija standardne devijacije distribucije aritmetičkih sredina uzoraka do koje

možemo doći jeste ona preko standardne devijacije uzorka, a koja se naziva **standardna pogreška aritmetičke sredine** i koja se računa prema formuli:

$$s_M = s / \sqrt{N}$$

7. obzirom da je riječ o normalnoj distribuciji, i za distribuciju aritmetičkih sredina uzoraka vrijede principi po kojima kada aritmetičkoj sredini te distribucije dodamo jednu, dvije, odnosno tri standardne devijacije te distribucije, obuhvatamo raspon od 68,26, 95,44, odnosno 99,73% rezultata te distribucije. Još jednom, (a) rezultati koji čine distribuciju aritmetičkih sredina uzoraka nisu ništa drugo do aritmetičke sredine svih slučajnih uzoraka iste veličine koji se mogu izvući iz početne populacije; (b) aritmetička sredina te distribucije nije ništa drugo do prava aritmetička sredina populacije; (c) standardna devijacija te distribucije nije ništa drugo do standardna pogreška aritmetičke sredine. Grafički, distribucija aritmetičkih sredina uzoraka izgledala bi kao što je prikazano na slici 7.2.

Slika 7-1. Distribucija aritmetičkih sredina uzoraka

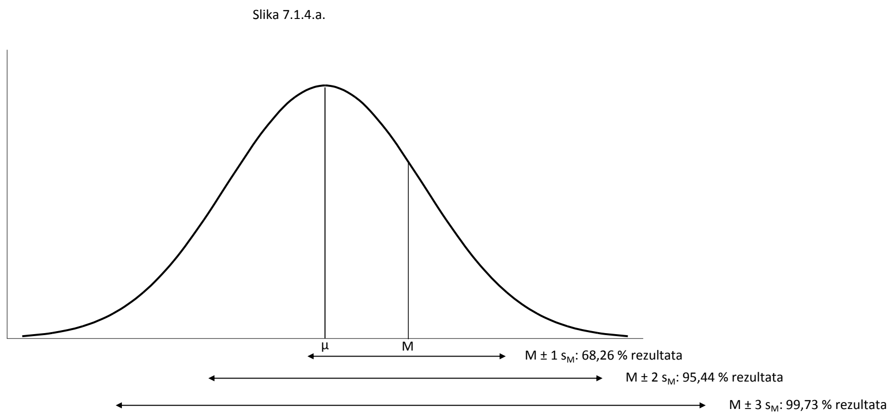


Postotak rezultata koji se nalaze u naznačenim intervalima oko aritmetičke sredine raspodjele (a to je populacijska aritmetička sredina) nije ništa drugo do vjerovatnoćama da ćemo kada po slučaju izvlačimo jedan rezultat iz ove raspodjele dobiti rezultat baš iz tog intervala.

Prilikom izvlačenja slučajnog uzorka iz populacije i računanja njegove aritmetičke sredine mi radimo upravo to – **iz distribucije aritmetičkih sredina uzoraka po slučaju izvlačimo jednu aritmetičku sredinu**. Pri tome imamo vjerovatnoću od 68,26, 95,44, odnosno 99,73% da ćemo izvući aritmetičku sredinu koja je od populacijske aritmetičke sredine (μ) udaljena manje od jedne, dvije, odnosno tri standardne pogreške aritmetičke sredine (s_M).

Koristeći se obratnom logikom – ako našoj aritmetičkoj sredini uzorka (M) dodamo i oduzmemo jednu, dvije, odnosno tri standardne pogreške aritmetičke sredine (s_M) imaćemo šansu od 68,26, 95,44, odnosno 99,73% da ćemo u istom tom intervalu obuhvatiti i pravu populacijsku aritmetičku sredinu. U svrhu ilustracije ta je situacija prikazana na slici 7.3.

Slika 7.3: Distribucija aritmetičkih sredina uzoraka-različiti intervali u kojima očekujemo pravu aritmetičku sredinu



Provjerimo da li ovaj princip vrijedi i za naš uzorak od n=800 i sa M=1202,54 KM i s=259,76 KM.

Standardna pogreška iznosi:

$$s_M = \frac{s}{\sqrt{N}} = \frac{259{,}76}{\sqrt{800}} = 9{,}18$$

Ako ovu vrijednost jednom, dva, odnosno tri puta dodamo i oduzmemo vrijednosti M dobit ćemo sljedeće intervale:

Interval I: 1193,36 - 1211,72

Interval II: 1184,17 - 1220,91

Interval III: 1174,99 - 1230,09

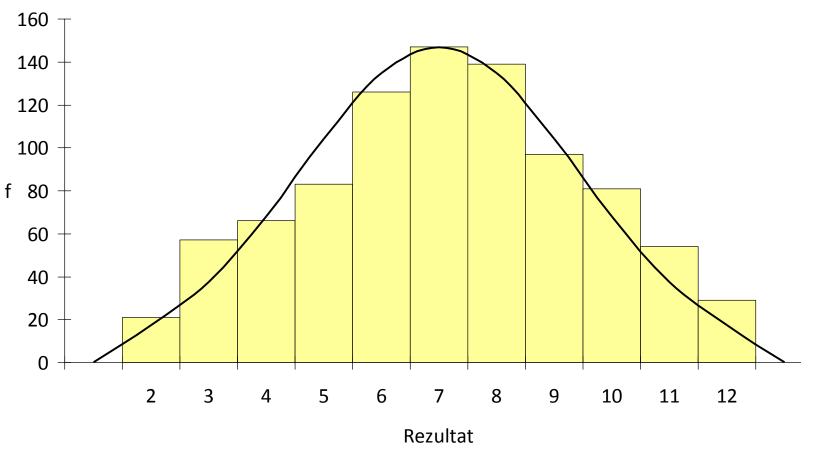
Kao što vidite, svi ovi intervali sadrže vrijednost populacijske aritmetičke sredine, μ=1200,47 KM. Dakle, u stvarnim istraživanjima provedenim na uzorcima mi nikada nećemo znati tačnu vrijednost populacijske aritmetičke sredine; najviše što ćemo moći učiniti jeste odrediti interval u kojem se sa određenim stepenom sigurnosti ta aritmetička sredina nalazi. Ti intervali nazivaju se **intervalima pouzdanosti** i vezuju se uz određeni **stupanj sigurnost**i da se u njima nalazi tražena vrijednost populacijske aritmetičke sredine. Tako, za interval I sa 68,26% sigurnosti tvrdimo da se u njemu nalazi populacijska aritmetička sredina; za interval II to tvrdimo sa 95,44% sigurnosti; za interval III sa 99,73% sigurnosti.

PRIMJER 7.3

U primjeru 5.6 opisali smo eksperiment studenata psihologije sa bacenjem para igraćih kockica 900 puta. Sada ćemo iste dobivene rezultate iskoristiti za ilustraciju principa uzorkovanja i distribuiranja aritmetičkih sredina uzoraka oko prave (populacijske) aritmetičke sredine μ.

Pri tome ćemo, naravno, krenuti od populacije. Našu **populaciju** u ovom slučaju čini 900 **rezultata** dobivenih bacanjem para kocaka (mogući rezultati kreću se u rasponu od 2 do 12). Prisjetimo se i izgleda raspodjele rezultata koje dobijemo bacajući par kocaka 900 puta (vidi sliku 5.2.3). Ta je raspodjela ponovo prikazana na slici 7.1.1 i ovaj put je nazvana populacijskom raspodjelom (obzirom da predstavlja našu populaciju od N=900 rezultata). Već znamo da ova distribucija ima μ=7,13 i σ=2,43.

Slika 7.1.1. Populacijska distribucija rezultata bacanja dvije kockice (N=900; μ= 7,127; σ=2,431)



Svaki par studenata dvije kockice je bacao po 20 puta. Ukupno je bilo 45 parova studenata, čime dobivamo našu **populaciju od ukupno 900 rezultata.** Obzirom da rezultati svakog pojedinog para studenata mogu predstavljati jedan od mogućih uzoraka koji se može izvući iz naše populacije, ove skupove od po 20 rezultata koje su dobili pojedinačni parovi studenata od sada ćemo nazivati **uzorcima**. Važno je napomenuti da su svi **uzorci iste veličine, n=20. U tabeli 7.1.1 prikazane su aritmetičke sredine za 45 uzoraka.**

Na osnovu aritmetičkih sredina pojedinačnih uzoraka možemo izračunati **zajedničku aritmetičku sredinu, odnosno aritmetičku sredinu aritmetičkih sredina uzoraka** te **standardnu devijaciju ove distribucije aritmetičkih sredina uzoraka**. Kako bismo ovu zajedničku aritmetičku sredinu razlikovali od aritmetičke sredine populacije (μ) i aritmetičkih sredina pojedinačnih uzoraka (M), označit ćemo je sa X'. Standardnu devijaciju aritmetičkih sredina uzoraka oko njihove zajedničke aritmetičke sredine označit ćemo sa s'.

Tabela 7.1.1: Aritmetičke sredine 45 uzoraka

R. br. uzorka	M	R. br. uzorka	M	R. br. uzorka	M
1	7,40	16	7,45	31	7,60
2	6,25	17	6,25	32	8,00
3	6,55	18	8,05	33	7,45
4	7,45	19	7,05	34	7,50
5	8,35	20	7,05	35	6,80
6	7,80	21	7,15	36	6,05
7	7,60	22	7,15	37	6,30
8	6,60	23	7,05	38	7,15
9	6,20	24	7,50	39	8,45
10	7,15	25	6,55	40	7,00
11	6,70	26	7,05	41	6,90
12	7,80	27	7,80	42	6,75
13	7,60	28	6,70	43	6,45
14	7,55	29	7,05	44	6,35
15	7,60	30	6,25	45	7,25

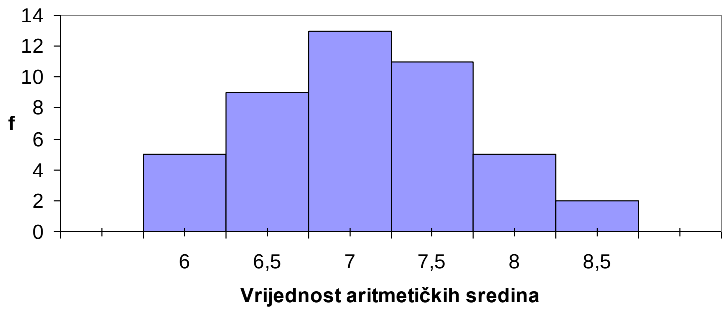
N	45
X'	7,127
s'	0,597

Dakle, aritmetička sredina pojedinačnih uzoraka iznosi X'=7,13. Kao što vidite, aritmetička sredina distribucije aritmetičkih sredina uzoraka iste veličine koji su po slučaju izvučeni iz populacije istovjetna je populacijskoj (pravoj) aritmetičkoj sredini.

Aritmetičke sredine pojedinih uzoraka distribuiraju se oko ove zajedničke (tj. populacijske) aritmetičke sredine u obliku normalne distribucije.

Iako gornja tvrdnja vrijedi za veliki broj uzoraka (teoretski, za beskonačno veliki broj uzoraka iste veličine koji se izvlače iz beskonačno velike populacije), tendencija normalnog **distribuiranja aritmetičkih sredina uzoraka** oko populacijske aritmetičke sredine primjetna je i na grafičkom prikazu distribucije naših 45 aritmetičkih sredina (slika 7.1.2). [Ukoliko bismo iz naše populacije od N=900 nastavili izvlačiti nove i nove uzorke veličine n=20, donja distribucija bi sve više nalikovala normalnoj distribuciji].

Slika 7.1.2 Distribucija aritmetičkih sredina uzoraka (broj uzoraka 45; veličina uzorka n=20; veličina populacije N=900); X'=7,127; s'=0,597



Iako je aritmetička sredina populacije istovjetna aritmetičkoj sredini aritmetičkih sredina uzoraka (7,13), obratite pažnju da se standardne devijacije distribucije rezultata u populaciji i distribucije aritmetičkih sredina značajno razlikuju (σ=2,43 naspram s'=0,60).

Standardna devijacija koju smo izračunali za distribuciju aritmetičkih sredina uzoraka (s'=0,60) približno odgovara vrijednosti standardne pogreške aritmetičke sredine koja se označava simbolom s_M. Standardna pogreška aritmetičke sredine računa se na temelju standardne devijacije uzorka i veličine uzorka, prema izrazu:

$$s_M = \frac{s}{\sqrt{N}}$$

Ova vrijednost je uvijek manja od standardne devijacije populacije. Razlog tome je što aritmetičke sredine uzoraka koje izvlačimo iz date populacije manje variraju oko svoje zajedničke aritmetičke sredine nego što originalni rezultati u toj populaciji variraju oko aritmetičke sredine populacije. Još jednom napominjemo da je aritmetička sredina distribucije aritmetičkih sredina uzoraka istovjetna populacijskoj aritmetičkoj sredini).

U istraživanjima rijetko unaprijed znamo vrijednost populacijske aritmetičke sredine (zapravo, za to i nema potrebe), ali ovu vrijednost ipak možemo procijeniti na temelju aritmetičke sredine jednog uzorka. Na koji način? Vratimo se na naš primjer. Zamislite da nismo znali koja je to aritmetička sredina naše populacije od 900 rezultata. Nadalje, zamislite da smo umjesto 45 različitih uzoraka iz ove populacije izvukli samo jedan slučajni uzorak veličine n=20. Pretpostavimo da smo na tako dobijenom uzorku utvrdili M=7,70 i s=2,28. Naš zadatak je da na temelju ovih vrijednosti procjenimo vrijednost aritmetičke sredine populacije!

Ono što prvo možemo primjetiti jeste da niti aritmetička sredina uzorka niti njegova standradna devijacija ne odgovaraju vrijednostima parametara.

No, ono što teoretski znamo jeste da u slučaju normalne distribucije (kakva je naša populacijska distribucija) interval vrijednosti definiran kao:

$\mu \pm \sigma$ obuhvata 68,26% svih rezultata, da interval

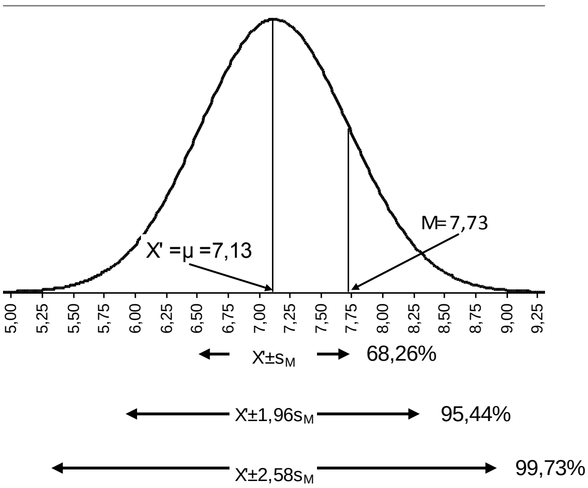
$\mu \pm 2\sigma$ obuhvata 95,44% svih rezultata, a da interval

$\mu \pm 3\sigma$ obuhvata 99,73% svih rezultata u populacijskoj distribuciji.

Ako ovu logiku primjenimo na distribuciju aritmetičkih sredina uzoraka, znaćemo da kada aritmetičkoj sredini te distribucije (koja nije ništa drugo do populacijska aritmetička sredina) dodamo i oduzmemu jednu, dvije, odnosno tri standardne devijacije te distribucije (a to je standardna pogreška aritmetičke sredine) imamo intervale koji obuhvataju 68,26%, 95,44%, odnosno 99,73% rezultata u distribuciji (a ti rezultati nisu ništa drugo do sve aritmetičke sredine uzoraka koje je moguće izvući iz populacije).

Jedna od tih aritmetičkih sredina koje čine distribuciju svih mogućih aritmetičkih sredina je i vrijednost koju smo izračunali na našem uzorku.

Slika 7.2: Distribucija aritmetičkih sredina uzoraka veličine n=20



Ako sada još jednom primjenimo istu logiku samo u obratnom smjeru možemo zaključiti sljedeće: ako aritmetičkoj sredini jednog uzorka dodamo i oduzmemo jednu, dvije, odnosno tri standardne pogreške aritmetičke sredine, imat ćemo 68,26%, 95,44%, odnosno 99,73% šanse da u dobivenim intervalima obuhvatimo i vrijednost prave aritmetičke sredine (vidi gornju sliku).

Razlog zašto aritmetičkoj sredini uzorka dodajemo i oduzimamo standardnu pogrešku aritmetičke sredine jeste taj što mi ne znamo koja je vrijednost standardne devijacije u distribuciji aritmetičkih sredina uzoraka (podsjećamo, iz populacije izvlačimo samo jedan uzorak). **Standardna pogreška aritmetičke sredine služi nam kao procijena vrijednosti standardne devijacije distribucije aritmetičkih sredina uzoraka!**

Prema tome, koristeći formulu za standardnu pogrešku aritmetičke sredine:

$$\mathbf{s_M = \frac{s}{\sqrt{N}}}$$

na našem uzorku dobivamo da je standardna pogreška aritmetičke sredine:

$$s_M = \frac{2,28}{\sqrt{20}} = 0,51$$

Kao što se može vidjeti, riječ je o vrlo bliskoj vrijednosti standardne pogreške koju smo izračunali kao standardnu devijaciju distribucije aritmetičkih sredina 45 uzoraka (s_M = 0,60).

Da zaključimo: na temelju vrijednosti M i s uzorka te njegove veličine, možemo (računajući standardnu pogrešku aritmetičke sredine uzorka) doći do procjene o vrijednosti aritmetičke sredine populacije!

PRIMJER 7.3

Ljekar opće prakse želi doći do informacije o prosječnoj visini populacije dječaka između 10 i 12 godina Kantona Sarajevo. Međutim zbog praktične neizvodljivosti, ljekar niti ne razmišlja o provođenju istraživanja na cjelokupnoj populaciji. Umjesto toga, svoje će istraživanje provesti na reprezentativnom (slučajnom) uzorku učenika sarajevskih osnovnih škola koji imaju između 8 i 10 godina. Uzorak broji 350 učenika. Vrijednosti utvrđene na uzorku su:

M = 139,36 cm

s = 26,22 cm

Na temelju ovih vrijednosti moguće je procijeniti prosječnu visinu muške djece ovog uzrasta u pripadajućoj populaciji (i uz svaku procjenu moguće je navesti i stupanj sigurnosti da je tačna).

Prvo što je ljekar uradio sa dobivenim statisticima jeste da je izračunao standardnu pogrešku aritmetičke sredine:

$$s_M = \frac{s}{\sqrt{N}} = \frac{26,22}{\sqrt{350}} = 1,40$$

U narednom koraku ljekar je izračunao 68, 95 i 99-postotne **intervale pouzdanosti aritmetičke sredine**:

139,36 ± 1,40:	137,96 – 140,76	68 %-tni interval pouzdanosti
139,36 ± 2 x 1,40:	136,56 – 142,16	95%-tni interval pouzdanosti
139,36 ± 3 x 1,40:	135,16 – 143,56	99%-tni interval pouzdanosti

U svom završnom izvještaju ljekar je naveo da sa **95% sigurnosti** tvrdi da se prava prosječna visina (svih) dječaka između 10 i 12 godina u Kantonu Sarajevo nalazi u intervalu od 136,56 do 142,16 cm.

Dakle, intervale koje dobijemo kada aritmetičkoj sredini uzorka na jednoj strani oduzmemo i na drugoj strani dodamo jednu, dvije, odnosno tri standardne pogreške aritmetičke sredine nazivamo **intervalima pouzdanosti aritmetičke sredine**. Uz svaki interval pouzdanosti vezuje se određeni **stepen sigurnosti** da se u njemu nalazi prava (populacijska) aritmetička sredina (ti odgovarajući stupanj sigurnosti su 68%, 95%, odnosno 99%).

ZADACI

1. Ukoliko je populacijska standardna devijacija (σ) za neku varijablu 22,5, koja je vrijednost standardne pogreške aritmetičke sredina kada je a) N = 2; b) N = 5; c) N = 25; d) N = 125; e) N = 545.

2. Za koliko bismo trebali promijeniti veličinu uzorka da bismo standardnu pogrešku aritmetičke sredine smanjili za pola; da bismo standardnu pogrešku aritmetičke sredine smanjili za 4 puta?

3. Istraživač je na uzorku N=1.350 utvrdio da je prosječna mjesečna zarada u Bosni i Hercegovini 987 KM, uz standardnu devijaciju 236 KM. Koliko iznosu prava prosječna mjesečna zarada stanovnika Bosne i Hercegovine? Sa kojim stepenom sigurnosti to tvrdite?

4. Izračunajte standardnu pogrešku aritmetičke sredine za zadatak 3.1.

5. Izračunajte standardnu pogrešku aritmetičke sredine za zadatak 3.2.

6. Izračunajte standardnu pogrešku aritmetičke sredine za zadatak 3.3.

7. Izračunajte standardnu pogrešku aritmetičke sredine za zadatak 4.10.

8. Iz tabele 7.1.3 po slučaju odaberite jedan uzorak veličine n=30. Na temelju tog uzorka procijenite populacijsku aritmetičku sredinu mjesečnih primanja u populaciji. Da li vrijede pravila navedena u uvodnom dijelu?

   Ponovite isti postupak kao u prethodnom zadatku, sa jednim uzorkom veličine n=300 koji ste ponovo po slučaju izvukli iz tabele 7.1.3 Objasnite kako veličina uzorka utiče na tačnost procjene populacijske aritmetičke sredine.

8. Testiranje hipoteza

Istraživanja u psihologiji provode se na uzorcima ispitanika. Međutim, cilj istraživanja je donošenje zaključka o populaciji. Na osnovu rezultata koje smo dobili na uzorku ispitanika, donosimo zaključke o stanju stvari u populaciji. U prethodnom poglavlju vidjeli smo kako nam standardna pogreška aritmetičke sredine omogućava utvrđivanje intervala u kojem se uz određenu pouzdanost (p) nalazi aritmetička sredina populacije. Na osnovu statistika uzorka (npr. M) zaključujemo o parametru populacije (μ). Dio statistike koja se bavi procjenama parametara populacije i testiranjem hipoteza naziva se **inferencijalna statistika**.

U svom općem značenju, hipoteze su misaone pretpostavke o nekoj pojavi, odnosima među pojavama ili među činiocima jedne pojave koja je predmet istraživanja. Hipoteza je sinonim za pretpostavku, odnosno tvrdnju o stanju stvari. U statistici, **testiranje hipoteza** je postupak kojim se donosi odluka o stanju stvari u populaciji na osnovu podataka prikupljenih na uzorku (npr. Da li se aritmetička sredina uzorka statistički značajno razlikuje od aritmetičke sredine populacije? Da li se aritmetičke sredine dva ili više uzoraka statistički značajno razlikuju?) Hipoteze se formiraju na osnovi prethodnih teorijskih saznanja, novootkrivenih empirijskih činjenica, svakodnevnog iskustva ili općih vjerovanja.

Statistička hipoteza uključuje **tvrdnju ili pretpostavku o parametru ili parametrima populacije** (npr. aritmetičkoj sredini ili varijanci populacije). Tvrdnja prema kojoj je populacijski parametar jednak određenoj vrijednosti ili da su populacijski parametri dvije ili više grupa jednaki (npr. *Prosječna visina muškaraca ista je kao i prosječna visina žena) naziva se* **nul-hipoteza (H_0).** Općenito, prema ovoj hipotezi ne postoji efekat, razlika. Tvrdnja prema kojoj populacijski parametar nije jednak određenoj vrijednosti ili da su populacijski parametri dvije ili više grupa različiti (npr. *Prosječna visina muškaraca nije ista kao i prosječna visina žena)* naziva se **alternativna hipoteza (H_1)**. Općenito, prema ovoj hipotezi se očekuje neki efekat, npr. postojanje razlika.

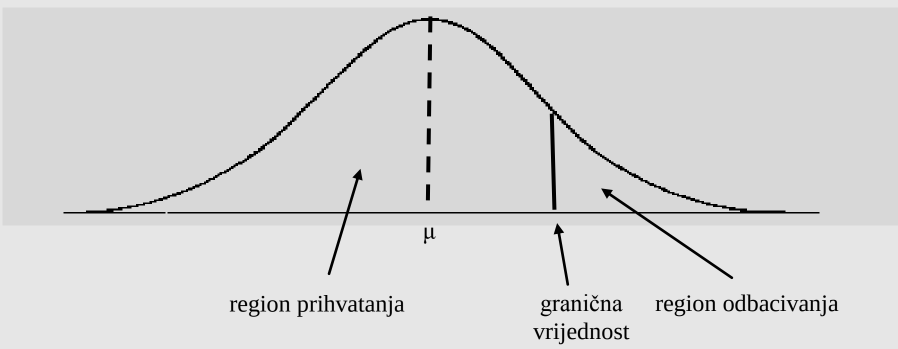
Statističko testiranje hipoteza zasniva se na metodi dokazivanja poznatoj kao *deductio ad absurdum*, odnosno, dovođenje do protivrječnosti ako se pretpostavi suprotna tvrdnja. Ako želimo dokazati neku tvrdnju, onda polazimo od suprotne tvrdnje. Stoga uvijek polazimo od nulte hipoteze. Postupak testiranja hipoteza sličan je sudskoj praksi u kojoj osoba koja je osumnjičena nije kriva dok se ne dokaže suprotno.

Npr., želimo dokazati da je novi lijek efikasan u tretiranju neke bolesti. Tvrdnja prema kojoj lijek ima efekta je alternativna hipoteza. Da bi dokazali ovu hipotezu, polazimo od suprotne, prema kojoj lijek nema efekta. Tvrdnja prema kojoj lijek nema efekta je nulta hipoteza. Da bi se dokazala početna tvrdnja, mora se oboriti suprotna (provjeravamo ima li dokaza protiv H_0, a u korist H_1).

Statistički test je postupak pomoću kojeg se dolazi do odluke o prihvatanju ili odbacivanje nulte hipoteze. Zasniva se na slučajnoj varijabli X kojom se matematički definira **distribucija statistika uzoraka** (npr. aritmetičkih sredina, razlika između dvije aritmetičke sredine, varijanci, medijana). Statistik uzorka (aritmetička sredina, razlika između aritmetičkih sredina, varijanca) je vrijednost koja se nalazi u određenom rasponu. Skup vrijednosti statistika za koje odbacujemo hipotezu H_0 nazivamo **oblast odbacivanja ili kritična oblast**. Veličina oblasti odbacivanja H_0 određena je vjerovatnoćom **(α)** pojavljivanja statistika uzorka u kritičnoj oblasti. Ova vjerovatnoća naziva se **nivo značajnosti testa** i određuje se kao vjerovatnoća da će vrijednost slučajne varijable X pasti u kritičnu oblast. Nivo značajnosti može biti određen arbitrarno, npr. 0,05 ili 0,01. Na osnovu unaprijed određenog nivoa značajnosti (tj. kritične oblasti) određujemo i **kritične vrijednosti** statistika, tj. granične vrijednosti kritične oblasti.

Na osnovu vrijednosti statistika i kritične oblasti, donosimo **zaključak**. Ako statistik testa pada u oblast odbacivanja, odbacujemo H_0. Ako statistik testa nije u oblasti odbacivanja, prihvatamo H_0 (drugim riječima, ako pada u oblast prihvatanja H_0). Praktično, kada je $p<\alpha$, test sugerira odbacivanje H_0 (“statistički značajno”).

Neka je nulta hipoteza da između aritmetičke sredine uzorka podataka i aritmetičke sredine populacije nema razlike. Zamislimo da provodimo sljedeći eksperiment: iz ciljne populacije metodom slučajnog odabira formiramo uzorak iste veličine kao i uzorak za čiju aritmetičku sredinu testiramo nultu hipotezu. Za ovaj uzorak izračunamo aritmetičku sredinu. Zatim podatke uzorka „vratimo“ u populaciju i ponovimo isti postupak: formiramo novi uzorak, izračunamo njegovu aritmetičku sredinu i podatke „vratimo“ u populaciju. Opisani postupak ponovimo veliki broj puta. Na ovaj način dobit ćemo veliki broj aritmetičkih sredina uzoraka. Distribucija aritmetičkih sredina uzoraka opisana je slučajnom varijablom X čije se vrijednosti normalno distribuiraju. Ova distribucija zapravo je distribucija statistika uzoraka. Aritmetička sredina statistika uzoraka jednaka je aritmetičkoj sredini populacije. Standardna devijacija statistika uzoraka zapravo je standardna pogreška aritmetičkih sredina uzoraka. Region odbacivanja H_0, granične vrijednosti i region prihvatanja H_0 određujemo arbitrarno. Distribucija statistika uzoraka, granične vrijednosti i regioni prihvatanja i odbacivanja H_0 prikazani su ispod.

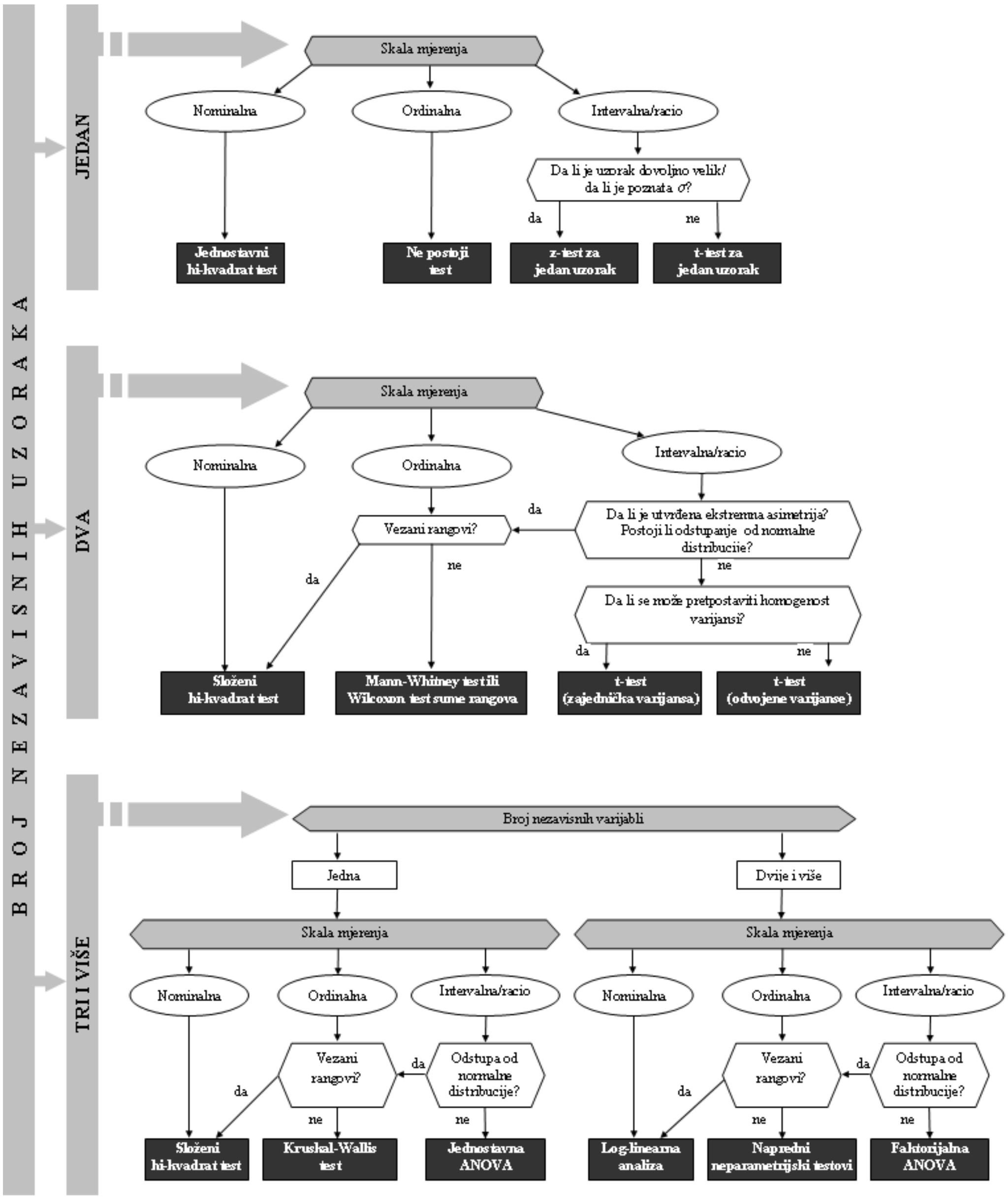


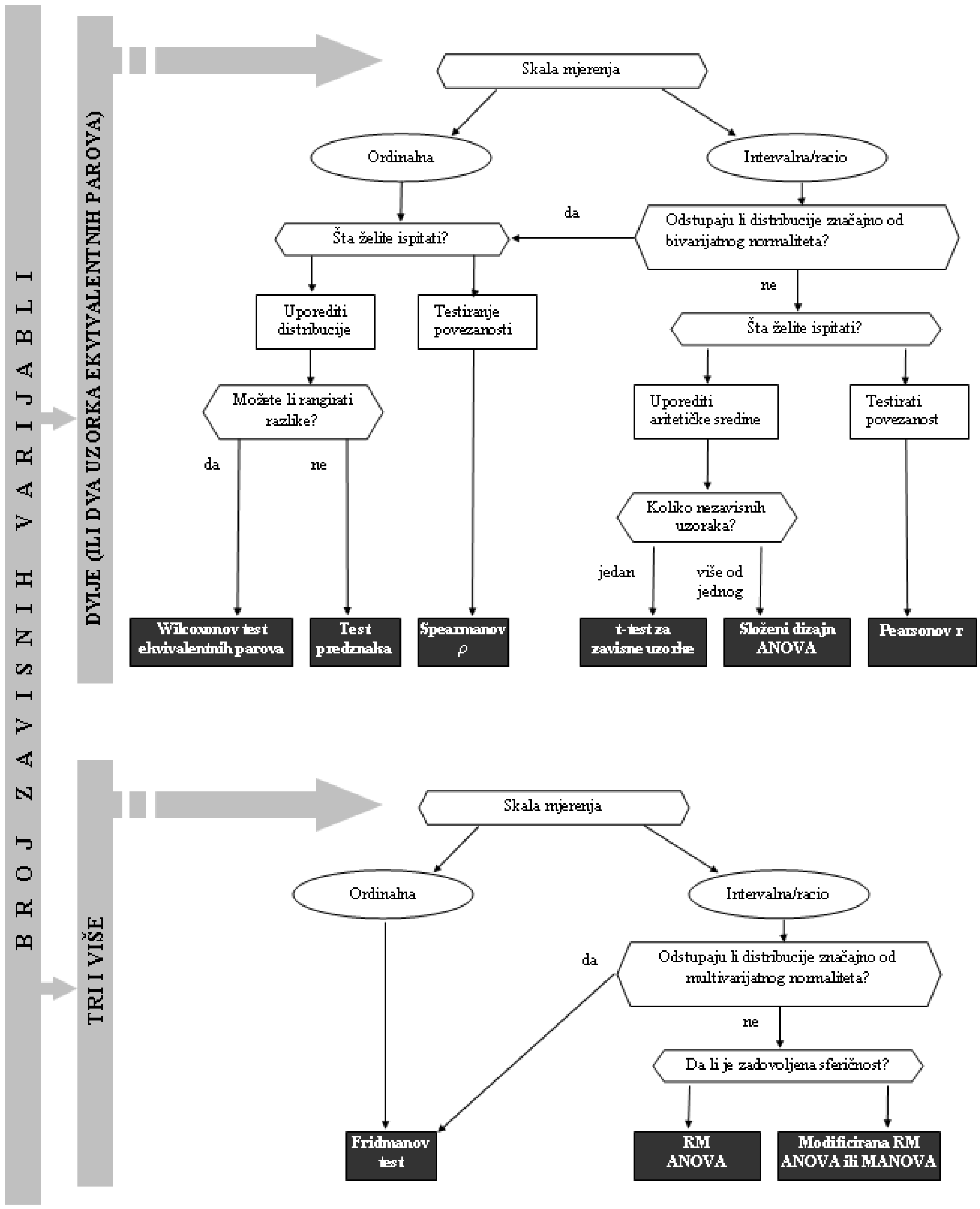
Izbor statističkog testa prvenstveno zavisi od problema istraživanja, odnosno istraživačke hipoteze. Ako, npr. želimo ispitati hipotezu prema kojoj se muškarci i žene ne razlikuju u verbalnim sposobnostima, to znači da ćemo imati dva nezavisna uzorka (muškarci i žene) čije deskriptivne vrijednosti treba uporediti i, pomoću odgovarajućeg statističkog testa, donijeti statistički zaključak.

Izbor statističkog testa zavisi i od skale mjerenja i oblika raspodjele podataka. Zapravo, statističke testove dijelimo na **parametrijske i neparametrijske**. Parametrijski testovi koriste se za provjeru hipoteza o nepoznatoj vrijednosti parametara populacije; preduvjet za njihovo korištenje je intervalni/ racio nivo mjerenja. Parametrijski testovi se vrše na osnovu nekih od teorijskih raspodjela: normalne, Studentove t-raspodjele, F-raspodjele, binomne raspodjele, itd. Kada su podaci prikupljeni korištenjem nominalne ili rang skale mjerenja i kada podaci prikupljeni intervalnom ili racio skalom mjerenja ne zadovoljavaju određene karakteristike distribucije, koristimo neparametrijske testove.

Ispod je dat shematski prikaz izbora statističkog testa zavisno od istraživačkog problema, skale mjerenja i određenih preduvjeta potrebnih za pojedine testove, prema Barry Cohenu[4].

[4] Cohen, B. (2011). *Explaining Psychological Statistics (1 edition).* Wiley, New York

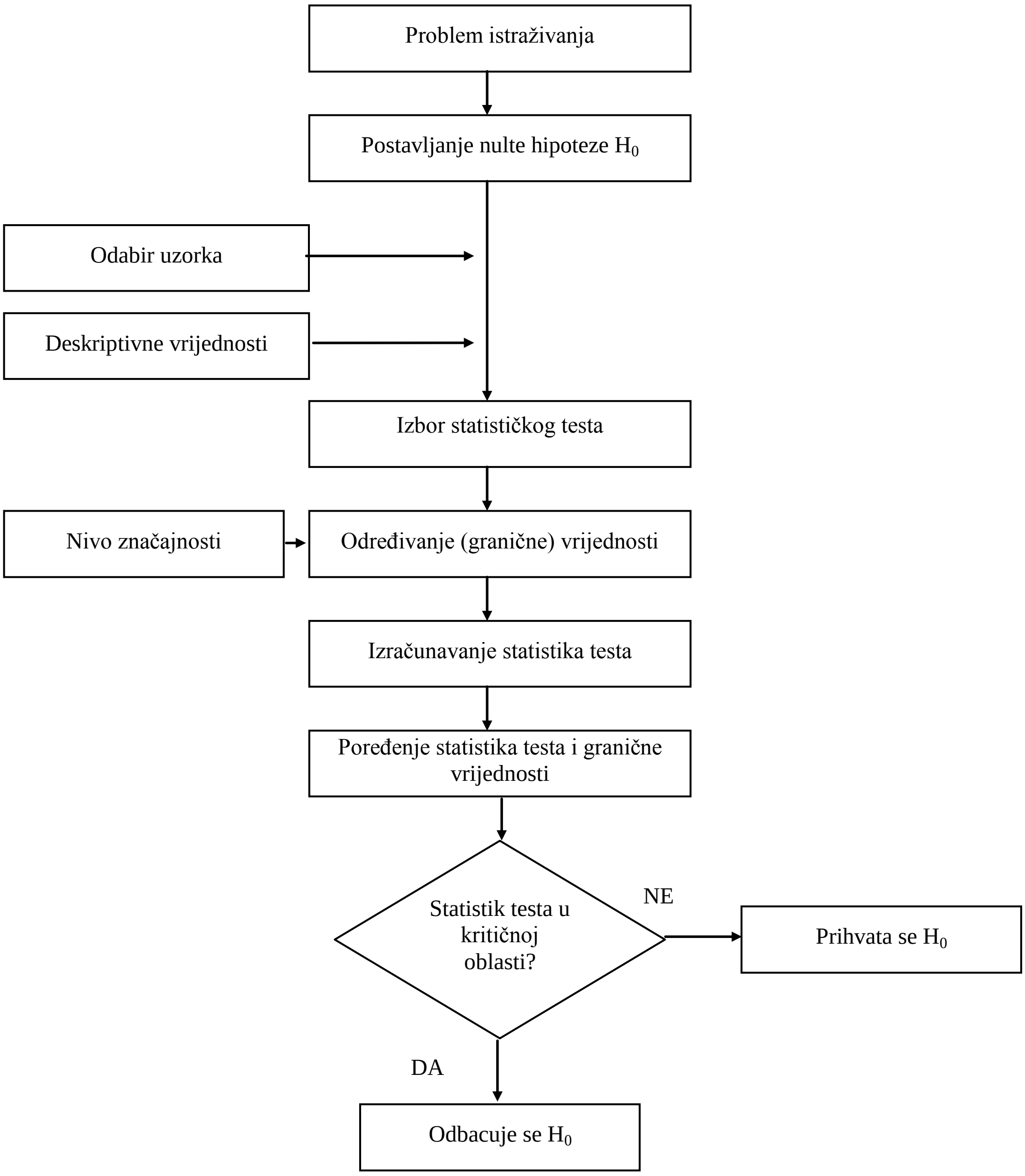

B R O J N E Z A V I S N I H U Z O R A K A
JEDAN
Skala mjerenja
Nominalna
Ordinalna
Intervalna/racio
Da li je uzorak dovoljno velik/ da li je poznata σ?
da
ne
Jednostavni hi-kvadrat test
Ne postoji test
z-test za jedan uzorak
t-test za jedan uzorak
DVA
Skala mjerenja
Nominalna
Ordinalna
Intervalna/racio
Vezani rangovi?
da
ne
Da li je utvrđena ekstremna asimetrija? Postoji li odstupanje od normalne distribucije?
Da li se može pretpostaviti homogenost varijansi?
Složeni hi-kvadrat test
Mann-Whitney test ili Wilcoxon test sume rangova
t-test (zajednička varijansa)
t-test (odvojene varijanse)
TRI I VIŠE
Broj nezavisnih varijabli
Jedna
Dvije i više
Skala mjerenja
Nominalna
Ordinalna
Intervalna/racio
Vezani rangovi?
Odstupa od normalne distribucije?
Složeni hi-kvadrat test
Kruskal-Wallis test
Jednostavna ANOVA
Log-linearna analiza
Napredni neparametrijski testovi
Faktorijalna ANOVA



Ukoliko nam istraživački nacrt nalaže da ispitamo razlike u verbalnim sposobnostima između muškaraca i žena (tj. između dva nezavisna uzorka), a da smo pri tom podatke prikupili koristeći intervalni nivo mjerenja, i da su podaci normalno raspoređeni, onda ćemo koristiti t-test. Međutim, ukoliko su podaci izrazito asimetrične raspodjele, onda će odgovarajući statistički test biti Mann-Whitneyev test, a ne t-test.

Kada testiramo hipotezu prolazimo određene korake. Zavisno od problema istraživanja postavljamo statističku nultu hipotezu, te biramo odgovarajući test. Nadalje, na osnovu prihvaćenog nivoa značajnosti određujemo granične vrijednost (definiramo oblast prihvatanja H_0), određujemo statistik testa i poredimo izračunati statsitik s graničnom vrijednosšću. Na kraju donosimo odluku.

Ako se statistik testa nalazi u kritičnoj oblasti odbacujemo H_0, ako ne, prihvatamo H_0. Shematski prikaz koraka u testiranju hipoteza dat je ispod.



Testiranje hipoteze s jednim uzorkom

U nastavku poglavlja prikazat ćemo postupak testiranje hipoteze s jednim uzorkom. Istraživački nacrt kojim u kojem testiramo hipotezu s jednim uzorkom podrazumijeva da poredimo aritmetičku sredinu uzorka s aritmetičkom sredinom populacije. Kada provjeravamo hipotezu o razlici između poznate populacijske vrijednosti aritmetičke sredine i aritmetičke sredine uzorka koristimo z-test ili t-test za jedan uzorak. Prvi test koristimo kada je poznata populacijska vrijednost varijance, a t-test kada nam ova vrijednost nije poznata, pa o njoj zaključujemo na osnovu varijance uzorka.

Postupak testiranja hipoteze s jednim uzorkom temelji se na **central-limit teoremi** (eng. *central limit theorem*), jednoj od najpoznatijih teorema u statistici. Prema ovoj teoremi, raspodjela aritmetičkih sredina uzoraka približava se normalnoj distribuciji s povećanjem veličine uzorka. Nadalje, u populaciji sa aritmetičkom sredinom μ i varijancom σ^2, distribucija aritmetičkih sredina uzoraka imat će aritmetičku sredinu $\mu_M=\mu$ i varijancu jednaku $\sigma^2_M= \sigma^2/N$. Na osnovu central-limit teoreme, poznate su nam sve važne karakteristike raspodjele (oblik, aritmetička sredina i varijanca), što nam omogućava testiranje hipoteza o aritmetičkim sredinama.

z-test za jedan uzorak

Kao smo već naveli, z-test koristimo kada nam je poznata populacijska vrijednost varijance. Na primjeru navedenom ispod objasnit ćemo logiku z-testa i postupak testiranja hipoteze s jednim uzokom kada znamo varijancu populacije.

Na grupi učenika (N=100) primjenjen je test inteligencije. Dobivena je aritmetička sredina M=105. Da li se ova grupa učenika prema mjerenom svojstvnu razlikuje od populacije? Pretpostavimo da nam je poznata aritmetička sredina populacije, μ=100.

- Nulta hipoteza: Grupa učenika ne razlikuje se statistički značajno od populacije u intelektualnim *sposobnostima.*

- Alternativna hipoteza: Grupa učenika razlikuje se statistički značajno od populacije u intelektualnim *sposobnostima.*

Kako bi utvrdili lokaciju na koju pada statistik uzorka (aritmetička sredina grupe) dobiveni rezultat, tj. aritmetičku sredinu treba pretvoriti u z-vrijednost. Kada smo odredili z-vrijednost, koristeći tabelu za standardnu normalnu distribuciju, možemo jednostavno odrediti oblast ispod ili iznad z-vrijednosti.

Dakle, koristit ćemo izraz:

$$z = \frac{M - \mu}{\sigma_M}$$

gdje je M aritmetička sredina uzorka, μ aritmetička sredina populacije, σ_M standardna devijacija distribucije uzoraka.

Kako bi odredili standardnu devijaciju distribucije uzoraka trebamo poznavati standardnu devijaciju populacije. Ovaj parametar obično ne poznajemo, ali za neke varijable (kao što je IQ) standardna devijacija je poznata (jer se rezultati standardiziraju na velikim uzorcima). Pretpostavimo da je σ=20. Odredit ćemo z-vrijednost:

$$z = \frac{M - \mu}{s_M} = \frac{105 - 100}{\frac{20}{\sqrt{100}}} 2,5$$

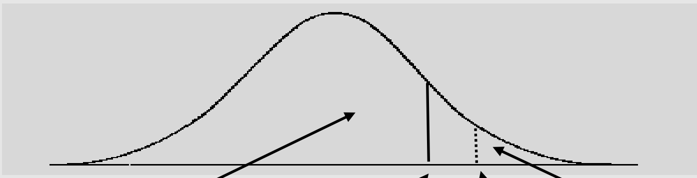
z-vrijednost iznosi 2,5. Sada možemo odrediti oblasti ispod i iznad izračunate z-vrijednosti. Iz tablice možemo odrediti da je površina od z=2,5 do kraja krivulje p=0,0062. Prema tome, vjerovatnoća da ćemo dobiti vrijednost veću od M=105 (uz uvjet da je uzorak veličine N=100) je veoma mala. Ako bi zamišljeni eksperiment formiranja velikog broja uzoraka i određivanja distribucije statistika uzoraka ponovili 1000 puta, *tek bi u šest slučajeva dobili aritmetičku sredinu uzorka veću od 105. Da li nam je ovo dovoljno za tvrdnju da je razlika* ***statistički značajna****? Odgovor na ovo pitanje zavisi od unaprijed definirane kritične oblasti. Uz uvjet da je nivo značajnosti α=0,05 statistik pada u oblast odbacivanja nulte hipoteze. Odnosno, trebamo uporediti p-vrijednost i nivo značajnosti α=0,05. Obzirom da je $p<\alpha$,, test sugerira odbacivanje H_0.*

Statistički zaključak mogli smo izvući i na osnovu upoređivanja dobivenog statistika sa kritičnom vrijednosti. Uz nivo značajnosti od α=0,05, granična vrijednost iznosi z_{gr}=1,66. Graničnu vrijednost očitavamo iz tablice za standardnu normalnu distribuciju.

Obzirom da je:

$$z > z_{gr},$$

zaključujemo da statistik pada u oblast odbacivanja nulte hipoteze. Ispod je ilustrirana distribucija statistika i položaj statistika grupe učenika.



Na kraju, rezultat testiranja hipoteza formuliramo na sljedeći način:

$$z\,(100)=2{,}5;\ p=0{,}006 \text{ ili } z\,(100)=2{,}5;\ p<0{,}001,$$

Vrijednost u zagradi predstavlja korigirani broj rezultata ili stepene slobode (SS).

t-test za jedan uzorak

U praksi, varijanca populacije (σ^2) najčešće nije poznata, nego je procjenjujemo na osnovu varijance uzorka (s^2). Stoga se i postupak testiranja hipoteze mijenja: ne koristimo z-vrijednost niti tablice sa z-vrijednostima. Za testiranje hipoteza s jednim uzorkom, u slučaju kada nam varijanca populacije nije poznata te je procjenjujemo na osnovu varijance uzorka, koristimo t-test i tablice t-vrijednosti. Statistici uzoraka raspoređuju se po **Studentovoj t-distribuciji**.

t-vrijednost (statistik) izračunava se prema izrazu:

$$t = \frac{M - \mu}{s_M}$$

Distribucija t-vrijednosti je zvonolika, unimodalna, simetrična, dok je aritmetička sredina jednaka nuli. Definirana je korigiranim brojem podataka – vrijednošću koju nazivamo stupnjevi slobode (SS); jedna t-distribucija ne korespondira svim mogućim veličinama uzoraka. Stoga postoji porodica t-distribucija. Stupnjevi slobode za t-test za jedan uzorak određuju se prema izrazu: SS=N-

1. Što je veličina uzorka veća, to je t-raspodjela sličnija normalnoj. Kažemo da kada N→∞, t-distribucija postaje ekvivalentna z-distribuciji.

Na grupi učenika (N=25) primjenjen je test inteligencije. Dobivena je aritmetička sredina M=105. Da li se ova grupa učenika prema mjerenom svojstvu razlikuje od populacije? Pretpostavimo da nam je poznata aritmetička sredina populacije, μ=100.

Nulta hipoteza: Grupa učenika ne razlikuje se od populacije u intelektualnim sposobnostima.

Alternativna hipoteza: Grupa učenika razlikuje se od populacije u intelektualnim sposobnostima.

Izračunat ćemo t-vrijednost:

$$t = \frac{M - \mu}{s_M} = \frac{105 - 100}{\frac{15}{\sqrt{25}}} = \frac{5}{3} = 1{,}66$$

Iz tabele t-vrijednosti očitat ćemo graničnu t-vrijednost za SS=24 uz nivo rizika od α=0,05: t_{gr}=1,711.Obzirom da je:

$$t < t_{gr},$$

zaključujemo da naš statistik pada u oblast prihvatanja H_0.

Pogreške u statističkom zaključivanju

Niti jedan statistički zaključak o populaciji nije stopostotno siguran (jer se temelji na uzorku); stoga i prihvaćanje neke hipoteze ne znači da je ta hipoteza apsolutno tačna. Umjesto "hipotezu prihvaćamo" ispravnije je reći "na osnovi uzorka ne postoji razlog za odbacivanje hipoteze".

Prilikom statističkog zaključivanja moguće je napraviti dvije vrste pogrešaka (tabela 10.1):

- **pogreška tipa I** – odbacivanje nulte hipoteze ako je ona istinita (vjerojatnost njenog pojavljivanja je nivo značajnosti α);
- **pogreška tipa II** – prihvaćanje nulte hipoteze ako je ona lažna.

Tabela 8.1: Pogreške u statističkom zaključivanju

STATISTIČKA ODLUKA	STANJE U POPULACIJI	
	Nema razlike između dvije aritmetičke sredine	*Postoji razlika između dvije aritmetičke sredine*
Odbacujemo nul-hipotezu	Pogreška tipa 1($p=\alpha$) lažno pozitivan	Ispravna odluka ($p=1-\beta$)
Prihvaćamo nul-hipotezu	Ispravna odluka ($p=1-\alpha$)	Pogreška tipa 2 ($p=\beta$) lažno negativan

Uz blaži kriterij (niži nivo značajnosti) izlažemo se riziku da proglasimo da se dvije aritmetičke sredine razlikuju, a zapravo među AS populacije nema razlike. Najčešće vrijednosti α su 0,10; 0,05; 0,01. Rizik od pogreške tipa 1 je pod našom kontrolom jer postavljamo nivo značajnosti prije testiranja hipoteze. Obično je to 0,05 ili manje.

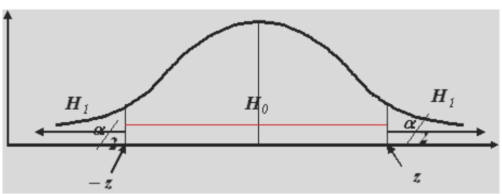
Uz strožiji kriterij (viši nivo značajnosti) izlažemo se riziku da ne proglasimo da se dvije aritmetičke sredine razlikuju, a zapravo među AS populacije postoji razlika. Što je uzorak manji veća je mogućnost pogreške tipa II. Vrijednost $1-\beta$ predstavlja snagu statističkog testa.

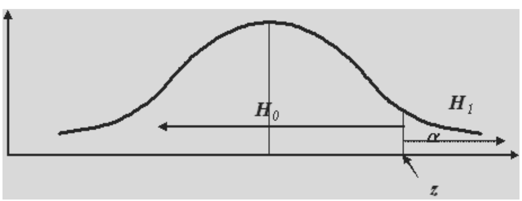
Vjerovatnoće pogrešaka tipa I i tipa II su u inverznom odnosu: sa smanjenjem α povećava se β, a smanjenjem β povećava se α.

Jednosmjerno i dvosmjerno testiranja

Ako unaprijed pretpostavimo smjer razlike (npr. $M>\mu$, ili obratno, $M<\mu$) region odbacivanja lociran je na jednom kraju distribucije statistika uzoraka; stoga ovakav test nazivamo **jednosmjerni (direktni) test**. Ukoliko pretpostavimo oba smjera razlike ($M>\mu$, i obratno, $M<\mu$), region odbacivanja lociran je na oba kraja distribucije; ovakav test nazivamo **dvosmjerni (nedirektni) test**.

Razlika između jednosmjernog i dvosmjernog testa je u graničnim vrijednostima regiona odbacivanja H_0. S obzirom da se nivo značajnosti kod dvosmjernog testa dijeli na dva dijela (jedna polovina s jedne strane, druga s druge strane raspodjele) bit će i **različite granične vrijednosti**. Međutim, ako posmatramo samo jednu stranu raspodjele, tada je vjerovatnoća da će vrijednost slučajne varijable X pasti u kritičnoj oblasti dva puta veća kod jednosmjernog u odnosu na dvosmjerni test. Razlike između jednosmjernog i dvosmjernog testa ilustrirane su ispod.





PRIMJER 8.1

Pođimo od hipoteze da je kod bacanja ispravnog novčića vjerovatnoća pojavljivanja pisma p=0,5. Pretpostavimo da smo kod 20 bacanja novčića u 17 slučajeva dobili pismo. Rezultat ovog eksperimenta nije apsolutni dokaz da je novčić neispravan (ili da je eksperiment proveden na neispravan način) jer nije nemoguće da se ovakav rezultat dobije (čak i kod ispravnog novčića). Međutim, iz iskustva znamo da se ovo rijetko može desiti sa ispravnim novčićem. Dobivanje 10 ili 11 pisama ne bi izazvalo sumnju uz hipotezu da je p=0,5, dok 18 ili 19 slučajeva pisma pruža osnovu za odbacivanje ove hipoteze kao malo vjerovatne. Drugim riječima, rezultat našeg eksperimenta dovoljan je dokaz protiv hipoteze p=0,5, a u korist hipoteze p>0,5. Gdje je granica između prihvatanja i odbacivanja hipoteze? Da li je to pojavljivanje pisma 13, 14 ili 15 puta? **Statistička teorija testiranja hipoteza** omogućava kvantifikaciju stupnja sumnje u neku hipotezu.

PRIMJER 8.2

Pretpostavimo da smo proveli istraživanje u kojem smo pitali studente koliko sati spavaju. Obzirom na obaveze prema studiju, te njihov stil života, očekujemo da u prosjeku studenti spavaju manje od prosjeka populacije. Kako bi ispitali ovu tvrdnju provedeno je istraživanje, prikupljeni su podaci i testirana je hipoteza.

Nulta hipoteza: Studenti spavaju isto kao i prosjek populacije:

$$M_{grupe}=\mu_{populacije}$$

Alternativna hipoteza: Studenti spavaju manje od prosjeka populacije:

$$M_{grupe}>\mu_{populacije}$$

Istraživanje je provedeno na uzorku od 100 studenata. Nakon prikupljanja podataka, utvrđeno je da studenti u prosjeku spavaju M=6,5 sati. Aritmetička sredina populacije iznosi μ=8 (znamo da u prosjeku čovjek tokom 24 sata provede 8 sati spavajući), dok standardna devijacija iznosi σ=2,5. Provjeravamo razliku između aritmetičke sredine jedne grupe i aritmetičke sredine populacije. Obzirom da nam je poznata standardna devijacija populacije, hipotezu ćemo testirati z-testom za jedan uzorak.

$$z=\frac{M-\mu}{s_M}=\frac{6,5-8}{\frac{2,5}{\sqrt{100}}}=\frac{-1,5}{0,25}=-6$$

Uz nivo značajnosti od p=0,05 i dvosmjerno testiranje razlike, granična z-vrijednost iznosi z_{gr}=1,96. Poređenjem izračunate z-vrijednosti i granične z-vrijednosti možemo zaključiti da je $z>z_{gr}$, odnosno da z-vrijednost pada u oblast odbacivanja H_0. Prema tome, grupa studentata koja je učestvovala u istraživanju u prosjeku spava manje od prosjeka populacije, što izražavamo i na sljedeći način:

$$z\ (100)=-6;\ p<0{,}05$$

PRIMJER 8.3

Da li djeca koja su bila izložena traumatskim događajima ispoljavaju statistički značajno više problema u ponašanju u odnosu na djecu koja nisu bila izložena traumatskim događajima? Na grupi od 120 djece koja su bila izložena traumatskim događajima primjenjen je upitnik pomoću kojeg se ispituju problemi u ponašanju djece i mladih. Aritmetička sredina ove grupe djece na ovom upitniku iznosi M = 55,0 uz s=10. Da li je aritmetička sredina grupe djece koja su bila izložena velikom broju traumatskih iskustava statistički značajno veća od aritmetičke sredine populacije? Poznat nam je parametar populacije: $\mu = 50$.

Nulta hipoteza: Broj problema u ponašanju djece koja su bila izložena traumatskim iskustvima ne razlikuje se statistički značajno od broja problema u populaciji njihovih vršnjaka:

$$M_{grupe}=\mu_{populacije}.$$

Nultu hipotezu mogli smo formulirati i kao: Grupa djece koja su učestvovala u istraživanju pripada populaciji čija je aritmetička sredina μ=50.

Alternativna hipoteza: Djeca koja su doživjela traumatska iskustva iskazuju statistički značajno veći broj problema u ponašanju od djece iz populacije njihovih vršnjaka bez traumatskih iskustava:

$$M_{grupe}>\mu_{populacije}.$$

Obzirom da nam nije poznata standardna devijacija populacije, hipotezu ćemo testirati t-testom za jedan uzorak. Neka je nivo značajnosti p=0,05, a testiranje dvosmjerno.

$$t=\frac{M-\mu}{s_M}=\frac{55-50}{\frac{10}{\sqrt{120}}}=\frac{5}{0{,}91}=5{,}49$$

Granična t-vrijednost za SS=119 iznosi z_{gr}=1,96. Poređenjem izračunate t-vrijednosti i granične t-vrijednosti možemo zaključiti da je $z>z_{gr}$, odnosno da t-vrijednost pada u oblast odbacivanja H_0. Prema tome, grupa djece koja je učestvovala u istraživanju u prosjeku postiže statistički značajno više vrijednosti od prosjeka populacije.

Hipotezu možemo testirati i na nivou značajnosti od p=0,01. Granična vrijednost iznosi t_{gr}=2,576. Prema tome, H_0 se odbacuje i na nivou od p=0,01. Nadalje, hipotezu testiramo i na nivou p=0,001. Granična vrijednost iznosi t_{gr}=3,291. Zaključujemo da se H_0 odbacuje i na nivou značajnosti od p=0,001. Dakle, tvrdnju statistički izražavamo na sljedeći način:

$$t(100)=5{,}49;\ p<0{,}001.$$

ZADACI

1. Pretpostavimo da direktor jedne osnovne škole tvrdi da učenici te škole u prosjeku dnevno uče 6 sati. Opišite postupak kojim bi smo provjerili tvrdnju direktora.

2. U hipotetičkom istraživanju na grupi od 20 studenata izmjeren je krvni pritisak i dobivena prosječna vrijednost M=124 mm Hg. Prosječni krvni pritisak u populaciji je $\mu=120$, a standardna devijacija $\sigma=10$. Da li se aritmetička sredina krvnog pritiska grupe studenata značajno razlikuje od aritmetičke sredine krvnog pritiska populacije?

3. Na grupi od 25 učenika IV razreda utvrđena je prosječna vrijednost inteligencije M=106.

   a. Testirajte hipotezu da se ova grupa djece statistički značajno ne razlikuje od opće populacije učenika IV razreda ($\mu=100$, $\sigma=15$) uz nivo značajnosti p=0,05, uz dvosmjerno testiranje.
   b. Testirajte hipotezu da se ova grupa djece statistički značajno ne razlikuje od opće populacije učenika IV razreda ($\mu=100$, $\sigma=15$) uz nivo značajnosti p=0,01, uz dvosmjerno testiranje.

4. Pretpostavimo da je uzorak iz zadatka 1 povećan za pet puta, a da je aritmetička sredina uzorka ostala ista.

   a. Testirajte hipotezu uz nivo značajnosti od p=0,01.
   b. Uporedite vrijednosti izračunate u zadatku 8.1 i 8.3.a. Šta možete zaključiti?

5. Pretpostavimo da je grupa studenata rješavala zadatke matematičkog rezoniranja iz standardiziranog testa za koji vrijedi da je $\mu=50$ i $\sigma=10$. Studenti su postigli sljedeće rezultate: 70, 45, 56, 59, 63, 50, 54, 67, 51, 48.

   a. Testirajte nultu hipotezu prema kojoj se ova grupa studenata statistički značajno ne razlikuje od populacije (nivo značajnosti p=0,05).
   b. U ovom slučaju, kojem tipu pogreške se izlažete kod zaključivanja?

6. Pretpostavimo da povećanje supstance TRY u krvi dovodi do ozbiljnih zdravstvenih problema. U hipotetičkom istraživanju testiran je novi lijek koji dovodi do smanjenja TRY-a u krvi. Na 100 pacijenata izmjerena je količina TRY supstance nakon konzumiranja lijeka te je utvrđena prosječna vrijednost M=58,02. Pretpostavimo da za populaciju vrijedi da je prosječna količina TRY supstance u krvi $\mu=60$, a standardna devijacija $\sigma=10$.

   a. Testirajte nultu hipotezu na nivou značajnosti p=0,05, dvosmjerno.
   b. Testirajte nultu hipotezu na nivou značajnosti p=0,01, dvosmjerno.
   c. U ovom slučaju, kojem tipu pogreške se izlažete kod zaključivanja?
   d. Kakva je praktična posljedica pogrešaka tipa I i II?

7. Jednu osnovnu školu pohađa 500 učenika. Direktor te škole smatra da su učenici te škole iznadprosječnih intelektualnih sposobnosti. Prema njegovom mišljenju, prosječni kvocijent inteligencije (IQ) iznosi najmanje 110. U cilju provjere ove tvrdnje, provedeno je ispitivanje inteligencije na slučajno odabranom uzorku od 40 učenika. Prosječna vrijednost na testu inteligencije iznosila je M=107, a standardna devijacija 10. Na osnovu dobivenih rezultata izvedite zaključak o prihvatanju ili odbacivanju tvrdnje direktora. Hipotezu testirajte uz nivo značajnosti od 0,01.

8. Grupa od 25 učenika rješavala je zadatke matematičkog rezoniranja iz standardiziranog testa za koji vrijedi da je μ=100. Standardna devijacija za populaciju nam nije poznata. Prosječna vrijednost grupe učenika iznosi M=95, a standardna devijacija s=10.

    a. Testirajte nultu hipotezu prema kojoj se ova grupa studenata ne razlikuje od populacije (p=0,05).
    b. Testirajte nultu hipotezu prema kojoj se ova grupa studenata ne razlikuje od populacije (p=0,01).

9. Nastavnik statistike provjeravao je znanje studenata iz matematike. Prethodne generacije studenata na istom testu tačno su rješavale u prosjeku 50 zadataka. Deset studenata, odabranih metodom slučajnog odabira, rješavali su test znanja iz matematike i postigli sljedeće rezultate: 64, 54, 48, 39, 62, 58, 46, 45, 50, 51. Može li nastavnik biti barem 90% siguran da će prosječan broj tačnih riješena biti najmanje 50?

10. Trener lokalnog košarkaškog kluba želi znati da li se prosječan broj koševa koje igrači njegovog kluba postignu tokom sezone takmičenja značajno razlikuje od državnog prosjeka. Pretpostavimo da na nivou države prosječan broj ubačenih koševa iznosi μ=65, a standardna devijacija σ=8. Prosječan broj koševa njegovog tima je M=68. Da li ovaj tim postiže više koševa od državnog prosjeka? Testirajte hipotezu uz nivo značajnosti od 0,05.

9. Testiranje razlika između dvije aritmetičke sredine

U znanstvenim istraživanjima vrlo često postavljamo pitanje „Da li se aritmetičke sredine dva uzorka statistički značajno razlikuju?“. Da bismo odgovorili na ovo pitanje, prije svega moramo razumjeti izraz „statistički značajna razlika“. Statistički rečeno, razlika između dvije aritmetičke sredine je **statistički značajna** ako je **po slučaju** očekujemo u manje od 5% slučajeva distribucije svih mogućih razlika (naravno, ako prihvatimo stupanj rizika od 5%; ako govorimo o stupnju rizika od 1%, onda tu razliku po slučaju očekujemo u manje od 1% slučajeva; drugim riječima, razliku koju proglašavamo „statistički značajnom“ po slučaju očekujemo u proporciji manjoj od postavljenog α nivoa).

Pored statističkog određenja „statističke značajnosti razlike između dvije aritmetičke sredine“, možemo govoriti i o „znanstveno – istraživačkom“ određenju: ako „statistički značajnu razliku“ po slučaju očekujemo u veoma malom procentu slučajeva (dakle, manjem od α), a u našem istraživanju dobijemo upravo tu razliku, onda možemo zaključiti da ona najvjerovatnije nije rezultat slučaja, već je rezultat djelovanja nekog **sistematskog faktora**. Naime, u znanstvenim istraživanjima mi se **sistematski trudimo** da razliku između dvije aritmetičke sredine učinimo što većom, tj. značajnijom, ili provjeravamo razlike između dvije grupe u slučajevima kada pretpostavljamo postojanje nekog sistematskog faktora koji dovodi do tih razlika. Npr., time što u okviru kliničke studije jednoj grupi depresivnih pacijenata dajemo novi (i, nadamo se, efikasniji) lijek protiv depresije, mi sistematskim naporom (davanjem novog lijeka) nastojimo ovu grupu, po **prosječnom** broju njihovih depresivnih simptoma, učiniti što je moguće više različitom (tj. manje depresivnom) u odnosu na drugu skupinu depresivnih pacijenata koji koriste klasični lijek protiv depresije. Ili, time što jednu grupu učenika podučavamo fiziku po novoj (i, nadamo, se boljoj) metodi, mi sistematski nastojimo dovesti do toga da prosječni uspjeh ovih učenika iz fizike bude „statistički značajno“ veći od prosječnog uspjeha druge skupine učenika koja je gradivo iz fizike savladavala po klasičnoj metodi podučavanja. Ili, pak, provjeravamo hipotezu o statističkoj značajnosti razlike između dječaka i djevojčica u agresivnom ponašanju, koju očekujemo obzirom na sistematsko djelovanje različitih faktora u procesu socijalizacije.

Koliko velika razlika između dvije aritmetičke sredine treba biti da bismo je proglasili statistički značajnom? Distribucija aritmetičkih sredina uzoraka o kojoj smo govorili u prethodna dva poglavlja nam je demonstrirala da prilikom izvlačenja slučajnih uzoraka iz populacije, aritmetičke sredine tih uzoraka obavezno variraju. Dakle, mi uvijek unaprijed očekujemo da ćemo prilikom formiranja dva različita slučajna uzorka iz iste populacije dobiti manje ili više različite vrijednosti njihovih aritmetičkih sredina. Još jednom – da li je ta razlika rezultat slučajnih variranja uzoraka ili je, pak, rezultat djelovanja nekog sistematskog faktora?

Kako je objašnjeno u prethodnom poglavlju, donijeti odluku o statističkoj značajnosti razlike između dvije aritmetičke sredine zapravo znači testirati tu razliku (stoga ovo poglavlje nosi naslov „Testiranje razlika između dvije aritmetičke sredine“), odnosno, provjeriti da li statistik testa pada u kritičnu oblast. Logika testiranja razlika između dvije aritmetičke sredine ista je kao i kod testiranje hipoteze s jednim uzorkom.

Obzirom da se kriterij za proglašavanje statističke (ne)značajnosti temelji na vjerovatnoći, tj. α vrijednosti, ovu razliku moramo pozicionirati u neku distribuciju koja bi nam omogućila „očitanje“ vjerovatnoće njenog javljanja po slučaju (dakle, uz važenje H_0). Ta teorijska distribucija, u situacijama kada testiramo značajnost razlika između dvije aritmetičke sredine, se naziva **t-raspodjela**. Zapravo, govorimo o porodici t-raspodjela jer je oblik ove raspodjele matematički određen veličinom uzorka; t-raspodjela, kao i normalna raspodjela, je unimodalna, zvonolikog oblika i simetrična. Što je uzorak veći, t-raspodjela sličnija je normalnoj.

Logiku t-testa kojeg koristimo kod testiranja razlike između dvije aritmetičke sredine možemo objasniti pomoću misaonog eksperimenta kojeg smo koristili u prethodnom poglavlju.

Zamislimo da imamo dvije identične populacije sa identičnim parametrima (ili da smo jednu populaciju klonirali pa njen klon posmatrali kao drugu populaciju). Iz prve populacije metodom slučajnog odabira formiramo uzorak i izračunamo aritmetičku sredinu. Zatim iz druge populacije metodom slučajnog odabira formiramo uzorak, iste veličine kao i prvi uzorak, i izračunamo aritmetičku sredinu. Zatim podatke uzoraka „vratimo“ u populacije. Opisani postupak ponovimo veliki broj puta, i svaki put izračunamo i razliku između dvije aritmetičke sredine. Na ovaj način dobit ćemo veliki broj razlika između aritmetičkih sredina uzoraka. Distribucija razlika aritmetičkih sredina uzoraka opisana je slučajnom varijablom X čije se vrijednosti raspoređuju prema t-raspodjeli. Aritmetička sredina razlika između aritmetičkih sredina parova uzoraka jednaka je razlici između aritmetičkih sredina populacija, odnosno jednaka je 0. Standardna devijacija razlika aritmetičkih sredina zapravo je standardna pogreška razlike aritmetičkih sredina uzoraka.

Praktično, razlikujemo testiranje razlika između aritmetičkih sredina za velike nezavisne i zavisne, te za male nezavisne i zavisne uzorke.

t-test za velike nezavisne uzorke

Kod velikih uzoraka, t-raspodjela je normalna raspodjela razlika između svih mogućih parova aritmetičkih sredina svih mogućih uzoraka iste veličine koji se mogu izvući iz početne (teorijski – beskonačno velike) populacije. Teorijski, aritmetička sredina t-raspodjele iznosi:

$$\boldsymbol{\mu_{M1-M2}=0},$$

a njena standardna devijacija, koja se naziva **standardna pogreška razlike između aritmetičkih sredina**, se računa prema formuli:

$$s_{M1-M2}=\sqrt{\frac{s_1^{\,2}}{N_1}+\frac{s_2^{\,2}}{N_2}}$$

Prema H_0 razlika između dvije aritmetičke sredine neće biti statistički značajna, tj. u raspodjeli će se pozicionirati oko aritmetičke sredine (preciznije, oko vrijednosti $\mu_{M1-M2}=0$) (u tom slučaju zaključujemo, npr., da novi antidepresiv nije značajno efikasniji od klasičnog; nova metoda podučavanja iz fizike nije bolja od stare).

Nasuprot tome, prema alternativnoj hipotezi, razlika između dvije aritmetičke sredine će biti dovoljno velika (u apsolutnim vrijednostima) da će u t-distribuciji izlaziti izvan intervala $\boldsymbol{\mu_{M1-M2} \pm 1{,}96s_{M1\text{-}M2}}$, uz nivo značajnosti od 5% ili intervala $\boldsymbol{\mu_{M1-M2} \pm 2{,}58s_{M1\text{-}M2}}$**, uz nivo značajnosti od 1%** (ovo vrijedi za velike uzorke, tj. n>40). Dakle, da bismo je proglasili statistički značajnom na nivou značajnosti od 5%, odnosno od 1%, razlika između dvije aritmetičke sredine mora biti 1,96, odnosno 2,58 puta veća od svoje pogreške. U tom slučaju iz naših primjera zaključujemo: da je novi antidepresiv u suzbijanju simptoma depresije efikasniji od klasičnog; učenici koji su učili fiziku po novoj metodi postižu bolje rezultate od učenika koji su učili po staroj metodi itd.

Statistički postupak koji nam omogućava da utvrdimo odnos razlike između dvije aritmetičke sredine i standardne pogreške te razlike se naziva **t-test** i računa se prema formuli:

$$\mathbf{t} = \frac{\textbf{razlika između aritmetičkih sredina}}{\textbf{pogreška razlike}}$$

tj.:

$$t = \frac{M_1 - M_2}{s_{M1\text{-}M2}}$$

Formiranje t-distribucije i logiku testiranja statističke značajnosti razlika između aritmetičkih sredina dva velika uzorka demonstrirat ćemo u primjeru 9.1.

t-test za velike zavisne uzorke

U odnosu na situaciju sa nezavisnim uzorcima, kada radimo sa zavisnim uzorcima, tj. kada se mjerenja u obje situacije koje želimo porediti vrše na isti ispitanicima, standardna pogreška razlike između aritmetičkih sredina se smanjuje (tj. smanjuje se greška mjerenja). To se događa zbog toga što se ovakvim istraživačkim nacrtima iz greške mjerenja isključuje varijabilitet između ispitanika. Obzirom na to, t-test za velike zavisne uzorke se računa prema sljedećoj formuli (pri tome, logika testiranja značajnosti razlike između dvije aritmetičke sredine ostaje ista kao i u slučaju velikih nezavisnih uzoraka):

$$t = \frac{M_1 - M_2}{\sqrt{s_{M1}{}^2 + s_{M2}{}^2 - 2r_{1,2}s_{M1}s_{M2}}}$$

gdje je $r_{1.2}$ korelacija između dvije varijable.

t-test za male nezavisne uzorke

Naveli smo da za velike uzorke t-distribucija ima oblik normalne distribucije (preciznije rečeno, kako veličina uzorka teži beskonačnosti t-raspodjela sve više poprima oblik normalne raspodjele). S druge strane, što je uzorak manji t-raspodjela postaje šira na krajevima, a tjeme niže. Stoga je prilikom razmatranja da li je razlika između aritmetičkih sredina koje su dobivene na malim uzorcima značajna, obavezno konzultiranje tablica sa graničnim t-vrijednostima.

t-test za male nezavisne uzorke temelji se na zajedničkoj standardnoj devijaciji za oba uzorka. Ovu vrijednost smijemo koristiti samo ako smo sigurni da se standardne devijacije uzoraka značajno ne razlikuju. Kako bismo to provjerili moramo izračunati F-test prema sljedećoj formuli:

$$F = \frac{\text{veća } s^2}{\text{manja } s^2}$$

Ako je gornji F-omjer statistički neznačajan (što opet utvrđujemo pomoću tablice graničnih F vrijednosti za testiranje razlika među varijancama) izračunavanje t-testa ćemo nastaviti prema sljedećoj proceduri:

$$\text{Zajednička } s = \sqrt{\frac{\sum(X - M_1)^2 + \sum(X - M_2)^2}{(N_1 - 1) + (N_2 - 1)}}$$

$$s_{M1\text{-}M2} = \text{Zajednička } s \sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

$$t = \frac{M_1 - M_2}{s_{M1\text{-}M2}}$$

$$SS = (N_1 - 1) + (N_2 - 1)$$

t-test za male zavisne uzorke

Kada radimo sa malim zavisnim uzorcima t-test računamo prema metodi diferencijacije koja uključuje i korelaciju između rezultata dva mjerenja. U ovom slučaju t-test se temelji na računanju aritmetičke sredine, standardne devijacije i standardne pogreške razlika parova rezultata. Testiranja statističke značajnosti razlika između aritmetičkih sredina dva mala zavisna uzorka demonstrirat ćemo u primjeru 9.4.

PRIMJER 9.1

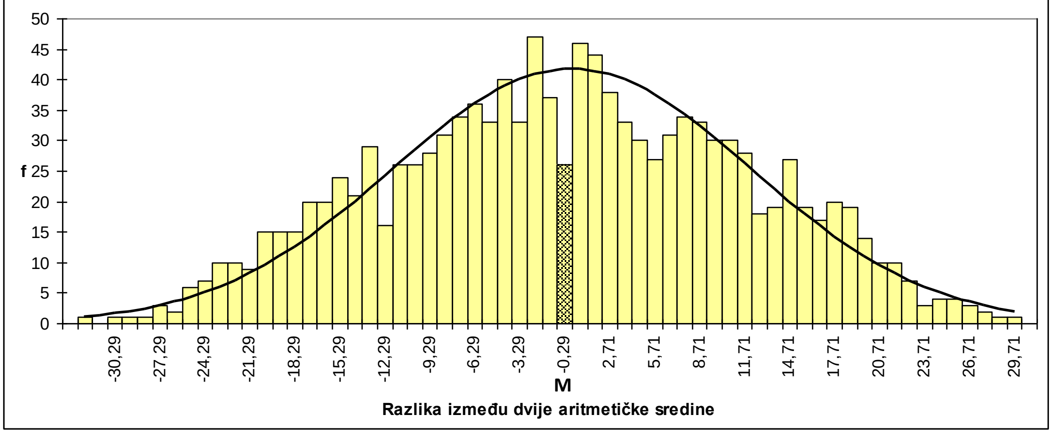
U primjeru 7.1 smo iz početne populacije od 10000 zaposlenih građana zamišljenog malog mjesta formirali distribuciju aritmetičkih sredina 50 uzoraka veličine n=800. Na osnovu ovih aritmetičkih sredina možemo vrlo jednostavno formirati distribuciju razlika aritmetičkih sredina ovih uzoraka: izračunat ćemo razlike za sve moguće parove ovih aritmetičkih sredina (M_1-M_2, M_1-M_3, M_1-M_4,... M_2-M_3, M_2-M_4, M_2-M_5... M_{49}-M_{50}). Kako se ukupno može načiniti 1225[5] ovakvih parova, tako ćemo dobiti i distribuciju od ukupno 1225 razlika između aritmetičkih sredina.

U tabeli 9.1.1 i na slici 9.1.1 predstavljene su deskriptivne vrijednosti i izgled distribucije aritmetičkih sredina 50 uzoraka:

Tabela 9.1.1: Deskriptivne vrijednosti distribucije razlika aritmetičkih sredina 50 uzoraka veličine n=800 (ukupno 1225 razlika)

M	Medijan	s	Skjunis	Kurtozis	Raspon	Minimum	Maximum
-0,29	-0,22	11,95	-0,02	-0,61	61,84	-32,05	29,79

Slika 9.1.1.: Distribucija razlika između aritmetičkoh sredina 50 uzoraka veličine n=800 (ukupno 1225 razlika)



Zaključujemo da dobivena raspodjela odgovara normalnoj raspodjeli. Mala odstupanja od očekivanih vrijednosti koja smo dobili (npr. aritmetička sredina distribucije iznosi M=-0,29 umjesto očekivanih 0 – objašnjavamo odstupanjem ovog primjera od početnih teorijskih pretpostavki). Da bismo dobili savršeno normalnu distribuciju koja ima aritmetičku sredinu tačno u tački 0 potrebna

---

[5] Broj mogućih parova računa se prema formuli k(k-1)/2, pri čemu je k broj aritmetičkih sredina.

nam je beskonačno velika početna distribucija iz koje formiramo beskonačno veliki broj različitih uzoraka iste veličine; za potrebe naše demonstracije i ovaj primjer će biti sasvim zadovoljavajući.

Ono što u suštini radimo kada testiramo značajnost razlike između aritmetičkih sredina dva uzorka jeste da **iz teorijske distribucije razlika između aritmetičkih sredina** (koja ima $\mu_{M1-M2}=0$ i čija je standardna devijacija – standardna pogreška razlike između aritmetičkih sredina) **izvlačimo jednu razliku između dvije aritmetičke sredine (naših konkretnih uzorka) i procjenjujemo vjerovatnoću javljanja te razlike po slučaju.** Ako je ta vjerovatnoća veća od 5%, odnosno od 1% - razliku nećemo proglasiti statistički značajnom na nivou značajnosti od 5%, odnosno 1%. Ako je, pak, ta vjerovatnoća mala, tj. manja od 5%, odnosno 1%, onda tu razliku proglašavamo statistički značajnom na nivou od 5%, odnosno od 1%.

Zamislimo da smo iz naše početne populacije izvukli dva slučajna uzorka, npr. uzorke pod rednim brojem 7 i 12 sa sljedećim deskriptivnim vrijednostima:

Tabela 9.1.2: M i s prosječnih mjesečnih primanja (u KM) u dva slučajna uzorka

Redni broj uzorka	N	M	s
7	800	1201,27	252,40
17	800	1209,66	264,35

Prema $\mathbf{H_0}$ smatramo da je razlika između ova dva uzorka rezultat slučajnih varijacija, tj. da ona **nije statistički značajna**; drugim riječima, prema H_0 smatramo da **ova dva uzorka pripradaju istoj populaciji.**

Prema alternativnoj hipotezi (H_1), razlika između ova dva uzorka je statistički značajna, tj. posljedica je djelovanja nekog sistematskog (sistematskih) faktora. Drugim riječima, ova dva uzorka reprezentiraju dvije različite populacije.

Obzirom da znamo da su oba uzorka izvučena iz iste populacije (zaposlenih osoba koje žive u istom mjestu) pretpostavljamo da razlika neće biti statistički značajna. Ipak, da bismo bili sigurni da uočena razlika između aritmetičkih sredina nije statistički značajna nužno je provesti t-test:

$$t = \frac{M_1 - M_2}{s_{M1-M2}} = \frac{M_1 - M_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$t = \frac{-8{,}39}{12{,}92} = -0{,}65$$

Dakle, i statistički smo potvrdili da naša dva uzorka dolaze iz iste početne populacije: t-test (t-omjer) nam pokazuje da razlika između dvije aritmetičke sredine nije dva puta veća od svoje pogreške, tako da je nećemo proglasiti statistički značajnom na nivou od 5% značajnosti.

Obzirom da su oba uzorka relativno velika i iste veličine, u ovom primjeru ne moramo konsultirati tablice za očitanje značajnosti rezultata t-testova. Ipak, standardna procedura za zaključivanje o statističkoj značajnosti nekog t-testa podrazumijeva konsultiranje graničnih vrijednosti t uz zadani broj stupnjeva slobode. Stupnjevi slobode u slučaju velikih **nezavnisnih uzoraka** računaju se prema formuli:

$$SS = (n_1 - 1) + (n_2 - 1)$$

$$SS = 799 + 799 = 1598$$

Granična vrijednost za utvrđeni SS na nivou značajnosti od 5% iznosi $t_{gr} = 1,96$.

Obzirom da je $t<t_{gr}$, prihvatamo H_0 i to statistički izražavamo kao:

$$t(1598) = -0,65;\ p>0,05.$$

Zamislimo sada malo realniju situaciju. Istraživača interesira da li postoji statistički značajna razlika u prosječnim mjesečnim primanjima muškaraca i žena. Kako bi to utvrdio, istraživač je u istoj anketi u kojoj je ispitanike pitao o njihovim primanjima zabilježio i spol ispitanika. Anketirano je ukupno 436 žena i 312 muškaraca iz početne populacije od 10000 zaposlenih građana. Deskriptivne vrijednosti prikazane su u tabeli 9.1.3.

Tabela 9.1.3: M i s prosječnih mjesečnih primanja (u KM) u skupini žena i muškaraca

Spol	n	M	s
Ž	436	1367,68	232,40
M	312	1051,89	198,73
total		1235,96	

Na osnovu ovih vrijednosti dobivamo sljedeći t-omjer:

$$t = \frac{315,79}{15,83} = 19,95$$

$$SS = 746$$

Obzirom da je granična t-vrijednost za zadate stupnjeve slobode na nivou rizika od 5% $t_{gr}=1,96$, odnosno za 1% rizika $t_{gr}=2,58$, zaključujemo da je utvrđena razlika između prosječnih primanja muškaraca i žena statistički značajna (tj. žene u prosjeku mjesečno zarađuju statistički značajno više novca). Drugim riječima, po svojim mjesečnim primanjima žene i muškarci čine dvije različite populacije, što izražavamo na sljedeći način:

$$t(746) = 19{,}95;\ p<0{,}01$$

PRIMJER 9.2

Istraživača interesira da li program stručnog usavršavanja ima utjecaja na visinu mjesečnih primanja uposlenika. U tu svrhu je utvrdio prosječna mjesečna primanja 260 uposlenika nekoliko firmi prije i nakon pohađanja 6-mjesečnog kursa menadžmenta. Deskriptivne vrijednosti prikazane su u tabeli 9.2.1.

Tabela 9.2.1: M i s mjesečnih primanja (u KM) prije i nakon treninga menadžmenta

Mjerenje	M	s
Prije treninga	1150,00	181,21
Nakon treninga	1320,00	236,43

Korelacija mjesečnih primanja u dvije mjerne tačke iznosila je r=0,53.

Prema formuli, t-test iznosi:

$$t = \frac{1150 - 1320}{\sqrt{11{,}24^2 + 14{,}66^2 - 2*0{,}53*11{,}24*14{,}66}}$$

$$t = \frac{-170{,}00}{12{,}91} = -13{,}17$$

$$SS = n - 1 = 259$$

Odgovarajuća granična vrijednost za nivo značajnosti od 5%, odnosno za 1% koju očitavamo iz tablice iznosi t_{gr}=1,97, odnosno t_{gr}=2,58. U skladu s tim zaključujemo da obuka uposlenika iz oblasti menadžmenta statistički značajno povećava iznos njihovih primanja te navodimo:

$$t(259) = -13{,}17;\ p<0{,}01$$

PRIMJER 9.3

Prema jednoj od teorija koje objašnjavaju poremećaj deficita pažnje (PDP) kod djece, djeca sa ovim poremećajem imaju očuvanu sposobnost selektivne pažnje – sposobna su usmjeravati pažnju na ciljne podražaje uz istovremeno zanemarivanje drugih, irelevantnih podražaja, ali imaju poteškoća u održavanju pažnje na ciljnom podražaju duži vremenski period koji je potreban za uspješno rješavanje određenog zadatka (dakle, oslabljena im je sposobnost tzv. održavane pažnje). Kako bi testirao ovu hipotezu, psiholog je testirao grupu od 6 djece sa PDP i 9 djece bez ove dijagnoze na testu održavane pažnje. [Ispitanici na testu održavane pažnje imaju zadatak da reagiraju na svaku promjenu prezentiranog stimulusa na ekranu; mjeri se broj pogrešaka, tj. broj propuštenih reakcija]. Rezultati za ove dvije skupine djece predstavljeni su u tabeli ispod:

Tabela 9.3.1: Broj pogrešaka i deskriptivne vrijednosti djece sa i bez PDP na testu održavane pažnje

Djeca bez PDP		Djeca sa PDP	
R.br.	Broj grešaka	R.br.	Broj grešaka
1	7	1	18
2	5	2	12
3	9	3	13
4	11	4	17
5	10	5	15
6	6	6	21
7	6		
8	9		
9	8		
M	7,89		16,00
s	2,03		3,35

Da bismo provjerili možemo li u ovom slučaju računati nekorigirani t-test, u prvom koraku moramo utvrditi da li su varijance ova dva uzorka homogene:

$$F = \frac{3{,}35^2}{2{,}03^2} = 0{,}37$$

Kako je granična F-vrijednost koju očitavamo iz tablice F_{gr}=4,82, zaključujemo da među varijancama nema statistički značajne razlike tako da možemo nastaviti sa računanjem t-testa:

$$\text{Zajednička } s = \sqrt{\frac{32{,}89 + 56}{8 + 5}}$$

$$\text{Zajednička } s = 0{,}73$$

$$s_{M1\text{-}M2} = 0{,}73 \sqrt{\frac{9 + 6}{9x6}}$$

$$s_{M1\text{-}M2} = 0{,}38$$

$$t = \frac{7{,}89 - 16{,}00}{0{,}38}$$

$$t = -21{,}08$$

$$SS = 8 + 5 = 13$$

Granične vrijednosti koje očitavamo iz tablice t-vrijednosti za zadate stupnjeve slobode i nivo rizika od 5%, odnosno 1% iznose: 2,16, odnosno 3,01. Dakle, navodimo;

$$t(13) = -21{,}08;\ p<0{,}01$$

Na osnovu dobivenog t-omjera možemo zaključiti da djeca sa PDP, u odnosu na djecu bez ovog poremećaja, imaju statistički značajno slabiju sposobnost održavanja pažnje kroz duže vremenske intervale.

PRIMJER 9.4

Grupa ispitanika učila je seriju besmislenih slogova. Poslije dva ponavljanja serije, broj slogova koji su članovi grupe tačno reproducirali naveden je u koloni “*Prije*” u tabeli ispod. Nakon toga, ispitanici su, tri dana po jedan sat, učili serije besmislenih slogova iste dužine kao i u prvoj seriji, ali su dobili nove instrukcije o metodama uspješnog pamćenja. Poslije tri dana zadata je nova serija besmislenih slogova iste dužine i težine kao i prva serija i ponovljena je dva puta. Ispitanicu su tada postigli rezultate predstavljene u koloni “*Poslije*” .

ISP.	Prije	Poslije
1	3	5
2	5	6
3	4	6
4	7	9
5	8	9
6	9	8
7	6	9
8	5	8
9	7	8

Potrebno je ustanoviti da li je trodnevno vježbanje u učenju besmislenih slogova i korištenje mnemotehnike dovelo do statistički značajnog povećanja broja tačno reproduciranih slogova posljednjeg dana.

Najprije ćemo izračunati razlike između prvog i drugog rezultata za svakog ispitanika te aritmetičku sredinu i standardnu devijaciju utvrđenih razlika:

ISP.	Prije	Poslije	Diferencijacija (D)	d ($D-M_D$)	d^2
1	3	5	-2	-0,44	0,20
2	5	6	-1	0,56	0,31
3	4	6	-2	-0,44	0,20
4	7	9	-2	-0,44	0,20
5	8	9	-1	0,56	0,31
6	9	8	1	2,56	6,53
7	6	9	-3	-1,44	2,09
8	5	8	-3	-1,44	2,09
9	7	8	-1	0,56	0,31

$\sum D = -14$

$\sum d_2 = 12,22$

$M_D = -1,56$

$s_D = 1,24$

Na osnovu s_D izračunat ćemo standardnu pogrešku aritmetičke sredine razlika:

$$s_{MD} = \frac{s_D}{\sqrt{N}}$$

$$s_{MD} = \frac{1,24}{\sqrt{9}}$$

$$s_{MD} = 0,41$$

Konačno, t-test računamo prema formuli:

$$t = \frac{M_D}{s_{MD}}$$

$$t = \frac{-1,56}{0,41}$$

$$t = -3,78$$

Stupnjeve slobode računamo prema formuli: SS=N-1, što u našem primjeru iznosi 8. Kako je t_{gr}=2,36 (za nivo od 5% rizika), odnosno t_{gr}=3,36 (za nivo od 1% rizika), izračunati t-test proglašavamo statistički značajnim:

$$t(8) = -3,78;\ p<0,01,$$

i zaključujemo da trodnevna vježba i korištenje mnemotehnika značajno unapređuje pamćenje (besmislenih slogova).

ZADACI

1. Pacijenti koji se liječe od depresije podijeljeni su u dvije skupine. Prvu skupinu čini 50 pacijenata koji su tokom dvije sedmice uzimali novi lijek za koji se smatra da umanjuje broj simptoma. Druga skupina od 43 pacijenta za to vrijeme je bila bez medikamentoznog tretmana. Nakon dvije sedmice ponovo je izmjeren broj simptoma kod obje skupine pri čemu su dobiveni sljedeći rezultati:

1. skupina (podvrgnuta tretmanu)	2. skupina (bez tretmana)
M = 14	M = 18
s = 3,2	s = 4,4

Da li je utvrđena razlika između aritmetičkih sredina statistički značajna?

2. Istraživača je interesiralo da li djeca koja žive u gradu čitaju brže od djece koja žive na selu. U tu svrhu po slučaju je formirao dvije skupine djece. Prvu skupinu od 150 djece činila su djeca iz grada. Drugu skupinu od 139 djece činila su djeca sa sela. Na temelju zadatka brzog čitanja, istraživač je dobio sljedeće rezultate: djeca iz grada zadati tekst su u prosjeku čitala za 93 sek., uz standardnu devijaciju od 6,5 sek.; djeci iz sela za čitanje istog odlomka teksta u prosjeku je trebala 91 sek., uz standardnu devijaciju od 7,1 sek. Da li je utvrđena razlika u brzini čitanja između dvije navedene grupe statistički značajna?

3. Psiholog je želio testirati učinkovitost novog programa edukacije vozača koji prave veliki broj prekršaja u saobraćaju. Da bi to učinio, iz generalne populacije vozača selektirao je (na temelju podataka u MUP-u) 120 vozača koji su u periodu od proteklih 12 mjeseci napravili više od 20 prekršaja. Statistički pokazatelji za ovu grupu vozača su sljedeći:

$M_1 = 49$ (prosječan broj prekršaja u poroteklih 12 mjeseci)

$s_1 = 10$ (standardna devijacija broja prekršaja u uzorku)

Ovi vozači su upućeni na novi program edukacije. Nakon završene edukacije, psiholog je u periodu od 12 mjeseci pratio navedenu grupu vozača i registrirao broj prekršaja koji je svaki od njih načinio. Utvrdio je sljedeće statističke pokazatelje:

$M_2 = 35$ (prosječan broj prekršaja u 12 mjeseci nakon edukacije)

$s_2 = 7$ (standardna devijacija broja prekršaja u uzorku)

Korelacija između broja prekršaja načinjenih u dva perioda (prije i poslije edukacije) je r=0,69. Da li je novi program edukacije učinkovit u smanjenju broja saobraćajnih prekršaja?

4. Test ličnosti ABC primjenjen je na dvije grupe učenika: prvoj grupi učenika koji se u školi ponašaju neupadljivo, i drugoj grupi učenika sa određenim problemima u ponašanju. Dobivene su sljedeće deskriptivne vrijednosti:

	I grupa	II grupa
N	8	5
M	105,5	113
s	6,87	3,08

Da li se ove dvije grupe učenika značajno razlikuju po prosječnim rezultatima na testu ličnosti?

5. Na testu verbalnih sposobnosti, grupa ispitanika čiji roditelji imaju visoko ili više obrazovanje i grupa ispitanika čiji roditelji imaju osnovno ili srednje obrazovanje postigle su sljedeće rezultate:

	OBRAZOVANJE RODITELJA	
	Osnovno i srednje	Visoko i više
N	28	24
M	16,92	17,90
s	4,94	4,21

Utvrdite da li ove dvije grupe ispitanika imaju statistički značajno različit prosječan uspjeh na testu verbalnih sposobnosti.

6. Psihologa interesuje kako alkohol utiče na tačnost prepoznavanja saobraćajnih znakova. Da bi odgovorio na to pitanje selektirao je dvanaest ispitanika. Na početku eksperimenta svaki ispitanik imao je zadatak da prepozna (tj. da imenuje) 46 saobraćajnih znakova. Nakon toga, pošto su u periodu od 45 minuta ispitanicu popili po 1,5 dl crvenog vina, psiholog im je ponovo dao zadatak prepoznavanja 46 znakova. I u jednom i u drugom mjerenju registriran je broj pogrešnih odgovora ispitanika:

ISP.	BROJ POGREŠNIH PREPOZNAVANJA SAOBRAĆAJNIH ZNAKOVA I mjerenje	II mjerenje
1	6	9
2	8	15
3	5	7
4	12	15
5	8	8
6	2	6
7	10	13
8	8	14
9	4	8
10	9	15
11	10	11
12	7	9

Da li alkohol smanjuje sposobnost prepoznavanja saobraćajnih znakova?

7. U jednom eksperimentu ispitivalo se da li novi lijek dovodi do smanjenja koncentracije tiroksina u krvi kod bolesnika koji boluju od Bazadovljeve bolesti. 134 pacijenata koji boluju od navedene bolesti po slučaju su podijeljene u dvije grupe – eksperimentalnu i kontrolnu. Pacijenti iz eksperimentalne grupe dobivali su tokom deset dana novi lijek, kod su pacijenti iz kontrolne dobivali inaktivnu fiziološku tvar (placebo). Jedanaesti dan izmjerena je koncentracija hormona kod svakog pacijenta eksperimentalne i kontrolne grupe. Dobivene su sljedeće prosječne vrijednosti i raspršenja:

	GRUPA kontrolna	eksperimentalna
N	77	57
M	17,4	24
s	3,7	4,5

Da li je novi lijek bio efikasan u smanjenju koncentracije tiroksina? Obrazložite odgovor.

8. U jednom istraživanju testirana je hipoteza prema kojoj stariji ispitanici, s obzirom da slabije procesiraju verbalne informacije, imaju slabije pamćenje za riječi od mlađih ispitanika. Zadatak ispitanka bio je da zapamte što više riječi. Nakon neutralnog zadatka utvrđen je broj

riječi kojih su se ispitanici mlađe i starije dobi mogli dosjetiti. Dobivene su sljedeće prosječne vrijednosti i raspršenja:

	GRUPA	
	mlađi	stariji
N	10	10
M	19,3	12
s	7,1	14

Da li je hipoteza potvrđena?

9. Dvije grupe učenika rješavale su test znanja iz matematike. Dobivene su sljedeće deskriptivne statističke vrijednosti:

	GRUPA	
	I	II
N	100	100
M	100	110
s	20	40

a. Koja grupa je bolje rješavala test znanja iz matematike? Obrazložite odgovor!
b. Kakav će biti odgovor na pitanje pod 4.a (Koja grupa je bolje rješavala test znanja iz matematike?) ako se standardna pogreška aritmetičke sredine prve grupe poveća dva puta? (aritmetičke sredine i broj ispitanika ostaju isti!)

10. Pretpostavimo da će dešnjaci značajno brže prepoznati predmete koji se nalaze u njihovoj desnoj ruci, od predmeta koji se nalaze u njihovoj lijevoj ruci (ispitivanje se izvodi s povezom preko očiju ili s nepropusnim naočalima). U sljedećoj tabeli prikazan je broj predmeta koje ispitanici prepoznaju za lijevu i desnu ruku u vremenu od 2 minute.

ISP.	lijeva ruka	desna ruka
1	8	10
2	5	9
3	11	14
4	9	7
5	7	10
6	8	5
7	10	15
8	7	7
9	12	11
10	6	12
11	11	11
12	9	10

Testirajte nul-hipotezu!

10. Analiza varijance

Analiza varijance (eng. analysis of variance) ili skraćeno ANOVA, je postupak koji se koristi za ispitivanje statističke značajnosti razlika između aritmetičkih sredina više grupa. Za razliku od t-testa, kojeg koristimo kada testiramo statističku značajnost razlike između dvije aritmetičke sredine, ANOVA-om možemo testirati razlike između bilo kojeg broja aritmetičkih sredina.

Naziv postupka ne ukazuje da se ispituju razlike između aritmetičkih sredina. Međutim, neka vas naziv postupka ne navodi na pogrešan zaključak.Zaista, analizom varijance varijabilitet rezultata se razlaže na određene dijelove jer se totalni varijabilitet zavisne varijable razlaže na manje dijelove, i to na dio varijance koji se pripisuje nezavisnoj varijabli i dio koji predstavlja ostatak, tj. rezidual, ili varijancu pogreške. No, premda se analizira varijabilitet, ipak nas procedura vodi ka zaključku o razlikama između aritmetičkih sredina. Razlog zbog kojeg se koristi naziv "analiza varijance", a ne multigrupna analiza aritmetičkih sredina, je taj da se ovim postupkom zaista upoređuju aritmetičke sredine, ali analiziranjem i upoređivanjem varijabiliteta, tj. varijanci.

Razumno pitanje je zašto ne koristiti t-test za svaki par AS? Nekoliko je razloga zbog kojih se ne koriste t-testovi. Najprije, korištenje većeg broja t-testova nije ekonomično. Sa povećanjem broja grupa značajno se povećava posao! Ukoliko imamo tri aritmetičke sredine broj parova za koje treba primjeniti t-test je 3, za četiri aritmetičke sredine 6, a za npr. šest potrebno je primjeniti 15 t-testova. Nadalje, pitamo se da li su razlike između više AS statistički značajne, a ne da li je razlika između dvije AS statistički značajna. Najvažniji razlog je da se s povećanjem broja t-testova, povećava i vjerovarnoća javljanja pogreške tipa I. Na kraju, u slučajevima kada imamo dvije ili više nezavisnih varijabli istovremeno, želimo znati ne samo o efektima pojedine varijable već i o efektu interakcije dvije ili više varijabli.

Najjednostavniji primjer analize varijance je jednostavna ili **jednosmjerna** (*one-way*) analiza varijance, kod koje imamo jednu nezavisnu varijablu, tj. **faktor** i jednu zavisnu varijablu. Faktor je kategorijalna varijabla, a vrijednosti varijable nazivaju se nivoi. Ukoliko imamo dva ili tri faktora, govorimo o **dvo- ili tro-smjernoj analizi varijance**. Ako su isti ispitanici uključeni u sve nivoe nezavisne varijable, koristimo **analizu varijance za zavisne uzorke** (RM ANOVA, od engl. *repeated measures* ANOVA). Analiza varijance sa jednom zavisnom varijablom naziva se **univarijatna**, a s dvije ili više zavisnih varijabli, **multivarijatna** tj. MANOVA-a.

Jednostavna (jednosmjerna) analiza varijanca - ANOVA

Logiku jednostavne analize varijance objasnit ćemo na jednom primjeru. Pretpostavimo da eksperimentalnom metodom želimo ispitati efekat distraktora na sposobnost rješavanja matematičkih zadataka. Nezavisna varijabla je nivo distraktora (nizak, srednji i visoki), a zavisna je broj tačno rješenih zadataka u određenom vremenskom periodu. Formirane su tri grupe ispitanika. Tokom rješavanja matematičkih zadataka ispitanici su bili izloženi distraktoru različitog intenziteta. Ispitanici prve grupe bili su izloženi distraktoru niskog intenziteta, druge srednjeg, a ispitanici treće grupe distraktoru visokog intenziteta. U svakoj grupi bilo je pet ispitanika. Broj tačno riješenih zadataka ispitanika grupa A, B i C prikazan je u tabeli ispod.

A	**B**	**C**
10	17	20
14	19	19
12	18	23
13	20	26
11	16	22

Da li je razlika između tri aritmetičke sredine statistički značajna?

Polazimo od pretpostavke da između aritmetičkih sredina populacija ne postoji razlika, tj. pretpostavljamo da se radi o identičnim populacijama. Prema tome, postavit ćemo nultu hipotezu, prema kojoj vrijedi:

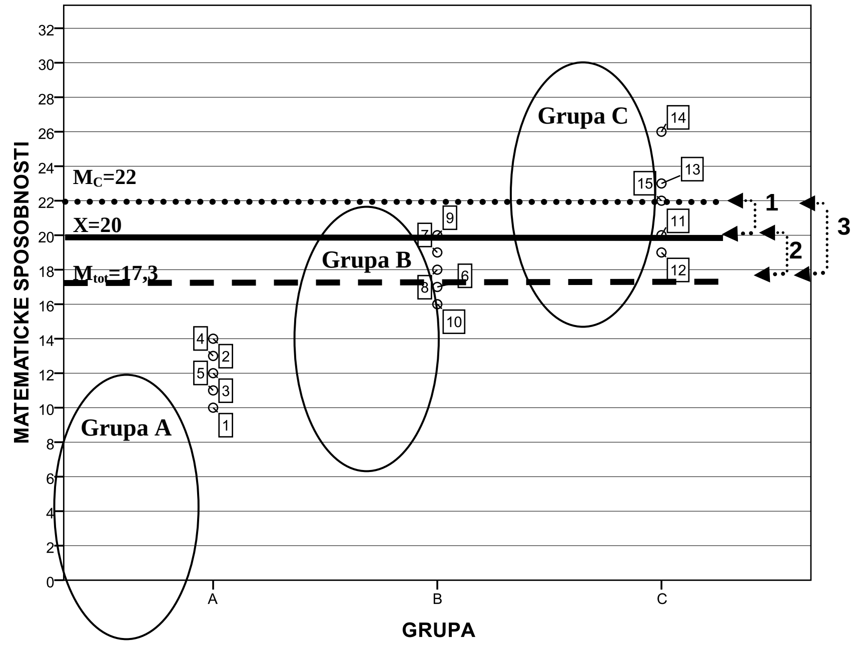
$$\mathbf{H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_n}$$

Pokušajmo o podacima razmišljati na nešto drugačiji način. Neka $X_{i,j}$ označava rezultat ispitanika *i* u grupi *j* (npr. X_{1C} se odnosi na rezultat prvog ispitanika u grupi C). Rezultat bilo kojeg ispitanika u testu matematičkih zadataka u funkciji je tri komponente: prosječne vrijednosti svih ispitanika koji bi teoretski mogli učestvovati u eksperimentu bez obzira na grupnu pripadnost (μ), odstupanja prosječne vrijednosti rezultata grupe (μ_j) od μ ($\tau = \mu_j - \mu$), i odstupanja rezultata ispitanika *i* u grupi *j* od prosječne vrijednosti svoje grupe ($e_{ij} = X_{ij} - \mu_j$). Odnosno:

$$\mathbf{X_{ij} = \mu + (\mu_j - \mu) + e_{ij} = \mu + \tau_j + e_{ij}.}$$

Svaki rezultat odstupa od aritmetičke sredine svih podataka (bez obzira na grupu), aritmetičke sredine podataka grupe kojoj pripada i odstupanja aritmetičke sredine grupe kojoj podatak pripada od sredine svih podataka (bez obzira na grupu). Na slici 2 prikazana su odstupanja rezultata prvog ispitanika grupe C (X_{1C}) od aritmetičke sredine svih podataka (X-M_{tot}), zatim od aritmetičke sredine podataka grupe kojoj pripada (X-M_j) i odstupanje aritmetičke sredine grupe kojoj ispitanik pripada od aritmetičke sredine svih podataka (M_j-M_{tot}).

Tabela 10.1: Grafički prikaz odstupanja rezultata u funkciji tri komponente



Odstupanja možemo prikazati na sljedeći način:

Total	Unutar grupe	Između grupa
$X - M_{tot}$	$X - M_j$	$M_j - M_{tot}$

Za svaki rezultat mogu se izračunati navedena odstupanja. Kada kvadriramo ova odstupanja dobit ćemo sume kvadrata. Tako imamo:

1. Sumu kvadrata totala (SS_{tot})
2. Sumu kvadrata unutar grupa (SS_{wg}), i
3. Sumu kvadrata između grupa (SS_{bg}).

Na osnovu sume kvadrata izračunat ćemo varijance:

$$s^2 = v = MS = \frac{SS}{df} = \frac{\sum (x - M)^2}{N - 1}$$

S obzirom da se varijanca zasniva na prosječnoj sumi kvadrata, u analizi varijance koristi se termin «**prosječni kvadrat**» i označava se sa MS (engl. mean-square).

Varijabilitet svih dobivenih rezultata rastavlja se na dijelove od kojih je sastavljen, tj. na interni varijabilitet unutar svake pojedine grupe rezultata i na varijabilitet između pojedinih grupa. Iz odnosa tih dvaju varijabiliteta može se zaključiti radi li se o grupama koje ne pripadaju istoj populaciji ili su njihove razlike samo slučajne pa sve grupe potječu iz iste populacije.

Varijanca unutar svake pojedine grupe zasniva se na varijabilitetima **unutar grupe** i označava sa $\mathbf{MS_{wg}}$ (*mean-square whitin groups*). Kada su veličine grupa jednake, varijanca unutar grupa jednaka je:

$$MS_{wg} = \Sigma s^2_j / k,$$

s^2_j – varijance rezultata pojedinih grupa

k – broj grupa

Varijanca između pojedinih grupa zasniva se na varijabilitetu **između** aritmetičkih sredina pojedinih **grupa** i označava se sa $\mathbf{MS_{bg}}$ (*mean-square between groups*):

$$MS_{bg} = n \times s^2_M,$$

s^2_M – varijanca aritmetičkih sredina grupa

n – broj ispitanika u jednoj grupi ($n_1 = n_2 = ... = n_n$)

Odnos između varijabiliteta između grupa i varijabiliteta unutar grupa je **F-omjer**:

$$F = \frac{MS_{bg}}{MS_{wg}}$$

Kao što smo koristili t-test u cilju donošenja odluke o prihvatanju ili odbacivanju nulte hipoteze, kod ANOVA-e koristimo F-test. Postoji **porodica F-distribucija**, zavisno od vrijednosti stupnjeva slobode brojnika i nazivnika. Obzirom da je F-omjer omjer varijanci, njegova vrijednost ne može biti manja od nula. F-distribucija je **pozitivno asimetrična**; samo u slučaju ekstremno velikih uzoraka, oblik F-distribucije približava se normalnoj (tačnije kada df_{bg} i df_{wg} teže ka beskonačnoj vrijednosti).

Kritičnu vrijednost F-omjera očitavamo iz tablica na osnovu stupnjeva slobode brojnika i nazivnika. Ako je izračunati F veći od kritične vrijednosti F (uz određene stupnjeva slobode), onda je varijabilitet između grupa statistički značajno veći od varijabiliteta unutar grupa, tj. utvrđena je statistički značajna razlika između aritmetičkih sredina.

Preduvjeti za korištenje ANOVA-e

Homogenost varijanci, normalnost distribucije i nezavisni uzorci osnovni su preduvjeti za korištenje analize varijance.

1. Homogenost varijanci

Populacije svake grupe imaju jednake varijance:

$$\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4 = \sigma^2_5$$

Za testiranje statističke značajnosti razlika između varijanci koristi se npr. Levenov F-test homogenosti varijanci. U praksi se često koristi princip prema kojem najveća standardna devijacija ne smije biti dva puta veća od najmanje standardne devijacije, te smatramo da su varijance najvjerovatnije homogene. Osim toga, kada su grupe jednake ili podjednake veličine, najvjerovatnije su varijance homogene.

2. Normalna distribucija

Distribucije rezultata zavisne varijable u svakoj grupi trebaju biti približno normalne.

3. Nezavisni uzorci

Uzorci (grupe) trebaju biti formirani metodom slučajnog odabira, tj. rezultati jednog uzorka su nezavisni od rezultata drugog uzorka.

S ciljem provjere preduvjeta za ANOVA-u, provode se deskriptivni statistički postupci (distribucije frekvencija, aritmetička sredina i standardna devijacija, koeficijenti simetričnosti i spljoštenosti, testiranje normalnosti distribucija). U slučajevima kada se pokaže da distribucije rezultata nisu normalno distribuirane ili da su varijance heterogene, provode se različiti postupci kako bi se zadovoljili preduvjeti. Npr. ukoliko u distribuciji postoje ekstremne vrijednosti, a postoji logička opravdanosti, takve podatke treba isključiti. Ponekad je potrebno provesti nelinearnu transformaciju podataka kako bi se postigla normalnost distribucije. Ukoliko niti jedan postupak ne da očekivani ishod, treba koristiti alternativne, neparametrijske postupke (Kruskall-Wallisov test).

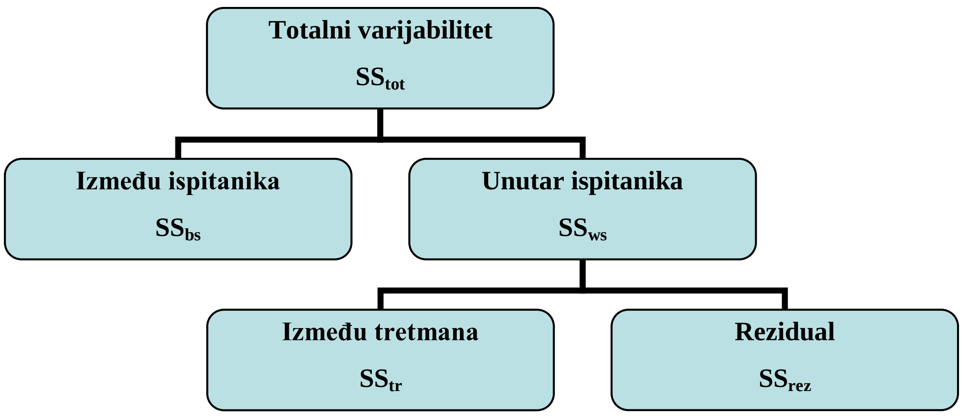
Općenito vrijedi da ukoliko su grupe podjednake veličine i varijance podjednake vrijednosti, onda možemo koristiti ANOVA-u.

Analiza varijance za zavisne uzorke

Kod analize varijance za zavisne uzorke totalni varijabilitet dijeli se na varijabilitet između ispitanika i varijabilitet unutar ispitanika. Varijabilitet unutar ispitanika čine dvije komponente: varijabilitet između tretmana (ili tačaka mjerenja) i rezidual, tj. ostatak kojeg ne možemo objasniti efektom tretmana (tj. tačaka mjerenja). Vrijedi:

$$SS_{tot} = SS_{bs} + (SS_{tr} + SS_{rez})$$

Struktura varijabiliteta može se prikazati na sljedeći način:



Pri čemu je:

SS_{tot} – Totalna suma kvadrata

SS_{bs} – Suma kvardata između ispitanika

SS_{ws} – Suma kvadrata unutar ispitanika

SS_{tr} – Suma kvadrata između tretmana

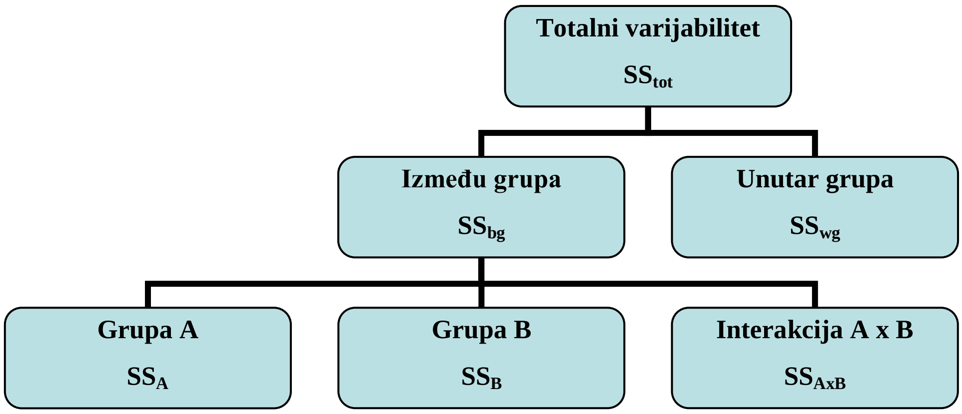
SS_{rez} – Suma kvadrata reziduala.

Faktorijalna analiza varijance

U terminima faktorijalne analize varijance nezavisne varijable nazivaju se još i faktorima. Stoga ovaj tip analize varijance nazivamo faktorijalna analiza varijance. Faktorijalna analiza varijance sa dvije nezavisne varijable zove se dvosmjerna analiza varijance, sa tri trosmjerna, itd.

Uz naziv „faktorijalna analiza varijance”, navode se informacije o broju grupa, tj. nivoa nezavisnih varijabli. Tako, faktorijalna analiza varijance 2 x 3 znači da prvi faktor ima dva, a treći tri nivoa. Faktorijalna analiza 3 x 2 x 4 znači da imamo tri nezavisne varijable, od kojih prva ima tri, druga dva, a treća četiri nivoa.

Kao i kod jednosmjerne analize varijance, kod faktorijalne analize varijace totalni varijabilitet dijeli se na varijabilitet između grupa i varijabilitet untar grupa. Varijabilitet između grupa dijeli se na varijabilitet grupe A, varijabilitet grupe B i varijabilitet interakcije grupe A i grupe B.



Pri čemu je:

SS_{tot} – Totalna suma kvadrata

SS_{bg} – Suma kvardata između grupa

SS_{wg} – Suma kvadrata unutar grupa

SS_{A} – Suma kvadrata grupe A

SS_{B} – Suma kvadrata grupe B

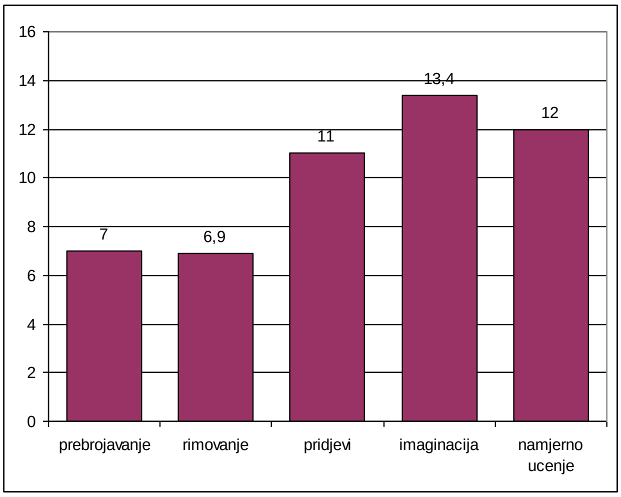
SS_{AxB} – Suma kvadrata interkacije AxB

PRIMJER 10.1

Pretpostavimo da želimo istražiti efekte različitih načina procesiranja riječi na njihovo pamćenje. U zamišljenom eksperimentu formirano je pet grupa ispitanika. U svakoj grupi bilo je deset ispitanika. Ispitanici prve grupe prebrojavali su slova zadate riječi, druge određivali riječ koja se rimuje sa zadatom riječi, ispitanici treća grupe su određivali pojam koji opisuje zadatu riječ, četvrte zamišljali predmet koji imenuje zadata riječ, dok su ispitanici pete grupe dobili eksplicitan zadatak da što bolje upamte zadatu riječ. Nakon neutralnog zadatka ispitanici su pitani da se dosjete riječi. U tabeli ispod navedeni su rezultati dobiveni u istraživanju. Broj označava količinu zapamćenih riječi.

	Prebrojavanje	**Rimovanje**	**Određivanje pridjeva**	**Imaginacija**	**Namjerno učenje**	**Total**
	9	7	11	12	10	
	8	9	13	11	19	
	6	6	8	16	14	
	8	6	6	11	5	
	10	6	14	9	10	
	4	11	11	23	11	
	6	6	13	12	14	
	5	3	13	10	15	
	7	8	10	19	11	
	7	7	11	11	11	
Total (T_j)	70	69	110	134	120	503
M	7,00	6,90	11,00	13,40	12,00	10,06
s	1,83	2,13	2,49	4,50	3,74	4,01
Varijanca	3,33	4,54	6,22	20,27	14,00	16,06

Da li postoji statistički značajna razlika između aritmetičkih sredina pet grupa? Ako postoji, koja grupa je najbolje rješavala zadatak upamćivanja? Koji način procesiranja je najefikasniji? Najprije ćemo grafički prikazati aritmetičke sredine. Najvišu vrijednost AS postigli su ispitanici koji su tokom nenamjernog učenja koristili imaginaciju (M=13,5), zatim ispitanici koji su dobili eksplicitnu instrukciju da uče material (M=12) i ispitanici koji su određivali pridjeve (M=11).



Da li su razlike statistički značajne?

Računski postupak provodi se u nekoliko koraka i logički odgovara smislu analize varijance. Najprije izračunamo pomoćne vrijednosti – veliki total (GT) i sumu kvadriranih vrijednosti svih rezultata (X^2), a zatim sume kvadrata totala, između grupa i unutar grupa. U sljedećim koracima potrebno je odrediti stupnjeve slobode, a zatim izračunati prosječne kvadrate (MS_{bg} i MS_{wg}). Na kraju, izačunat ćemo F-omjer i uporediti izračunatu vrijednost s kritičnom, koju smo očitali iz tablice.

- Izračunati GRAND TOTAL (GT): $\mathbf{GT = (\Sigma X)^2 / N}$

  ΣX : suma svih rezultata u svim grupama

  $GT = (9 + 8 + 6 + \ldots + 7 + 9\ldots + 15 + 11 + 11)^2 / 50 = 5060{,}18$

  $\boxed{\mathbf{GT = 5060{,}18}}$

- Izračunati TOTALNU SUMU KVADRATA (SS_{tot}): $\mathbf{SS_{tot} = \Sigma X^2 - GT}$

  ΣX^2: totalna suma kvadriranih rezultata u svim grupama

  GT: grand total

  $SS_{tot} = (9^2 + 8^2 + 6^2 + \ldots + 7^2 + 9^2 + \ldots + 15^2 + 11^2 + 11^2) – 5060{,}18$

  $SS_{tot} = 5847 – 5060{,}18 = 786{,}82$

  $\boxed{\mathbf{SS_{tot} = 786{,}82}}$

- Izračunati sumu kvadrata IZMEĐU GRUPA (SS_{bg}): $\mathbf{SS_{bg} = \Sigma T^2_j / n – GT}$

  T^2_j: kvadrirana suma rezultata u pojedinim grupama (j)

  n: broj rezultata u pojedinim grupama

  GT: grand total

$SS_{bg} = (70^2 + 69^2 + 110^2 + 134^2 + 120^2) / 10 – 5060{,}18$

$SS_{bg} = 5411{,}7 – 5060{,}18 = 351{,}52$

$\boxed{\mathbf{SS_{bg} = 351{,}52}}$

- Izračunati sumu kvadrata UNUTAR GRUPA (SS_{wg}): $\mathbf{SS_{wg} = \Sigma X^2 - \Sigma(T^2_j / n_j)}$

  ΣX^2: totalna suma kvadriranih rezultata u svim grupama

  T^2_j: kvadrirana suma rezultata u pojedinim grupama (j)

  n_j: broj rezultata u pojedinim grupama

  $SS_{wg} = (9^2 + 8^2 + 6^2 + … + 7^2 + 9^2 + … + 15^2 + 11^2 + 11^2) – (70^2 / 10 + 69^2 / 10 +110^2 / 10 + 134^2 / 10 + 120^2 / 10)$

  $SS_{wg} = 5847 – 5411{,}7 = 435{,}3$

  $\boxed{\mathbf{SS_{wg} = 435{,}3}}$

PROVJERA:

Ako su tačno izračunate vrijednosti SS_{tot}, SS_{bg} i SS_{wg}, tada mora biti zadovoljena jednakost:

$SS_{tot} = SS_{bg} + SS_{wg}$

$786{,}82 = 351{,}52 + 435{,}3$

- Odrediti STUPNJEVE SLOBODE za svaku sumu kvadrata:

$df_{bg} = (k-1)$	k: broj grupa	$df_{bg} = 5 – 1 = 4$	$df_{bg} = 4$
$df_{wg} = (N-k)$		$df_{wg} = 50 – 5 = 45$	$df_{wg} = 45$
$df_{tot} = df_{bg} + df_{wg}$		$df_{tot} = 4 + 45 = 49$	$df_{tot} = 49$

- Odrediti VARIJANCE (prosječne kvadrate):

  MS = suma kvadrata/df

  $MS_{bg} = SS_{bg} / df_{bg}$ $MS_{bg} = 351{,}52 / 4 = 87{,}88$

  $MS_{wg} = SS_{wg} / df_{wg}$ $MS_{wg} = 435{,}3 / 45 = 9{,}67$

- Izračunati F:

  $F = MS_{bg} / MS_{wg}$

  $F = 87{,}88 / 9{,}67 = 9{,}08$

Iz tablice L očitati graničnu F vrijednost za određene stupnjeve slobode. F–tablica se čita tako da se stupnjevi slobode brojnika čitaju na gornjem rubu tablice, a stupnjevi slobode nazivnika na njenom lijevom rubu.

$F_{0,05}\,(4{,}45) = 2{,}58$; $F_{0,01}\,(4{,}45) = 3{,}78$

Unijeti rezultate u TABLICU ANALIZE VARIJANCE

Izvor varijabiliteta	Suma kvadrata (SS)	Stupnjevi slobode (df)	Varijanca (MS)	F
između grupa	351,52	4	87,88	9,08
unutar grupa	435,30	45	9,67	
Total	786,82	49		

Odbacujemo H_0 i zaključujemo da postoji značajna razlika između aritmetičkih sredina grupa!

Možemo zaključiti da postoji statistički značajna razlika između ispitanika pet grupa. Međutim, još uvijek ne znamo između kojih grupa postoji statistički značajna razlika. Kako bi odgovorili na ovo pitanje, provodimo tzv. **post-hock postupak**. Prilikom izračunavanja razlika između aritmetičkih sredina nakon završenog F-testa (a posteriori) može se koristiti neki od testova: Scheffeov test, LSD, Bonferroni. U nastavku prikazat ćemo postupak u kojem se koristi Scheffeov test.

Scheffeov test

Nakon izračunatog F-omjera u analizi varijance, za svaki par aritmetičkih sredina koje želimo usporediti primjeniti sljedeću formulu:

$$F = (M_a - M_b)^2 / [MS_{wg} (n_a + n_b) / n_a n_b]$$

- Iz F-tablice očitamo granični F uz željeni nivo značajnosti, za (k – 1) i (N – k) stupnjeve slobode.
- Očitana granična vrijednost F pomnoži se sa (k – 1) i tako dobijemo novu graničnu vrijednost F´.
- Izračunati F uporedimo sa F´. Ako je F>F´ razliku možemo smatrati značajnom.
- Postupak od 1 do 4 ponovit ćemo za svaki par aritmetičkih sredina.

Za utvrđivanje parova AS za koje postoji statistički značajna razlika upotrijebit ćemo Scheffeov test.

$F = (M_a - M_b)^2 / [MS_{wg} (n_a + n_b) / n_a n_b]$

$F = (M_1 - M_2)^2 / [MS_{wg} (n_1 + n_2) / n_1 n_2]$

$F = (7 – 6{,}90)^2 / [9{,}67 \times (10 + 10) / 10 \times 10]$

$F = 0{,}01^2 / 9{,}67 \times 0{,}2$

$F = 0{,}005$

$F_{0,05}\ (4{,}45) = 2{,}58$; $F_{0,01}\ (4{,}45) = 3{,}78$

$F_{gr} \times (k – 1) = 2{,}58 \times 4 = 10{,}32$

$0{,}005 < 10{,}32$

Zaključujemo da između aritmetičkih sredina prve i druge grupe ne postoji značajna razlika!

Postupak ćemo ponoviti za svaki par AS. U tabeli ispod prikazane su izračunate F-vrijednosti i nivo značajnosti.

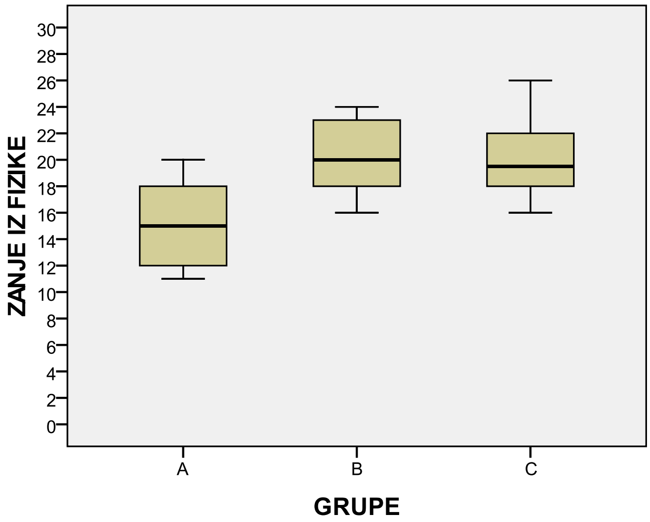
Razlike između AS	F	p
M1 - M2	0,005	> 0,05
M1 - M3	8,27	> 0,05
M1 - M4	**21,18**	< 0,01
M1 - M5	**12,93**	< 0,05
M2 - M3	8,69	>0,05
M2 - M4	**21,85**	<0,01
M2 - M5	**13,45**	>0,05
M3 - M4	2,98	>0,05
M3 - M5	0,52	>0,05
M4 - M5	1,01	>0,05

Najefikasiji načini procesiranja su četvrti i peti, tj. imaginacija i namjerno učenje.

PRIMJER 10.2

Pretpostavimo da je edukacijski psiholog želio ispitati koja je metoda podučavanja fizike najefikasnija, metoda A, B ili C. Formirane su tri grupe ispitanika koje su podučavane trima metodama, nakon čega je utvrđen nivo znanja primjenom Testa znanja iz fizike. Rezultati su prikazani u tabeli ispod:

A	B	C
12	17	20
14	24	19
12	18	23
13	20	26
11	24	22
17	18	17
19	23	19
18	20	18
20	20	20
16	16	16



Da li se aritmetičke sredine ispitanika grupa A, B i C statistički značajno razlikuju? Koja metoda je najefikasnija?

Najprije ćemo provjeriti da li distribucije rezultata tri grupe značajno odstupaju od normalne i da li su varijance homogene. Box-plot prikaz distribucije rezultata može nam poslužiti za brzu provjeru normalnosti distribucija rezultata (bez zaključka), ali i ukazati na moguće razlike u varijancama. Na osnovu box-plot prikaza možemo zaključiti da su distribucije približno simetrične, i da se varijance značajno ne razliku. Upoređujući standardne devijacije također možemo zaključiti da je zadovoljen preduvjet homogenosti varijanci.

	A	B	C
Ti	152	200	200
M	15,2	20	20
s	3,2	2,9	3,0

1.	$\mathbf{GT = (\Sigma X)^2 / N}$	$GT = 10156{,}80$
2.	$\mathbf{\Sigma X^2}$	$SumX^2 = 10558$
3.	$\mathbf{SS_{tot} = \Sigma X^2 - GT}$	$SS_{tot} = 401{,}20$
4.	$\mathbf{SS_{bg} = \Sigma T^2_j / n - GT}$	$SS_{bg} = 153{,}60$
5.	$\mathbf{SS_{wg} = \Sigma X^2 - \Sigma(T^2_j / n_j)}$	$SS_{wg} = 247{,}6$
provjera	$\mathbf{SS_{tot} = SS_{bg} + SS_{wg}}$	$SS_{tot} = 401{,}20$
6.	$\mathbf{df_{bg} = (k-1)}$	$df_{bg} = 2$
	$\mathbf{df_{wg} = (N-k)}$	$df_{wg} = 27$
	$\mathbf{df_{tot} = df_{bg} + df_{wg}}$	$df_{tot} = 29$
7.	$\mathbf{MS_{bg} = SS_{bg} / df_{bg}}$	$MS_{bg} = 76{,}80$
	$\mathbf{MS_{wg} = SS_{wg} / df_{wg}}$	$MS_{wg} = 9{,}17$
8.	$\mathbf{F = MS_{bg} / MS_{wg}}$	$F = 8{,}375$
	$\mathbf{p}$	$p = 0{,}001$
	$\mathbf{F_{0,05}\ (2,27)}$	

9. Tabela ANOVA

Izvor varijabiliteta	*Suma kvadrata (SS)*	*Stepeni slobode (df)*	*Varijanca (MS)*	*F*	*p*
između grupa	153,60	2	76,80	8,375	0,001
unutar grupa	247,60	27	9,17		
Total	**401,20**	**29**			

Razlike između aritmetičkih sredina su statistički značajne!

S obzirom da su aritmetičke sredine grupa B i C jednake (i da su standardne devijacije podjednake), zaključujemo da su ove dvije metode podjednako efikasne, i da su efikasnije od metode A.

PRIMJER 10.3

U zadatku 12 iz Poglavlja 3 (*Mjere centralne tendencije*) bilo je riječi o anketi koju je nastavno vijeće OŠ „Sigmund Freud“ uputilo roditeljima učenika šestih razreda. Jedno od anketnih pitanja odnosilo se i na novčana primanja majki učenika (anketirano je ukupno 98 porodica). Na temelju ovih podataka nastavno vijeće želi dobiti informaciju da li se visina novčanih primanja majki statistički značajno razlikuje obzirom na njihovu stručnu spremu. U tu svrhu, na temelju podataka o stručnoj spremi prikupljenih anketnim upitnikom, majke učenika podijeljenje su u četiri kategorije: 1 – niža stručna sprema (NSS), 2 – srednja stručna sprema (SSS), 3 – viša stručna sprema (VŠS) i 4 – visoka stručna sprema (VSS). Obzirom da u dva anketna upitnika nije bila navedena stručna sprema ispitanica, podaci za ove dvije majke su isključeni iz daljnje analize, čime je broj ispitanika u konačnom uzorku iznosio 96. Podaci organizirani na ovaj način prezentirani su u tabeli ispod.

	Stručna sprema majke			
	NSS	SSS	VŠS	VSS
Mjesečni novčani prihodi majke (KM)	450	750	900	1.150
	450	750	900	1.150
	450	750	900	1.150
	500	800	950	1.200
	500	800	950	1.200
	500	850	950	1.200
	500	850	950	1.200
	550	850	950	1.200
	600	850	950	1.250
	600	850	950	1.250
	600	850	1.000	1.250
	650	850	1.000	1.250
	650	850	1.000	1.250
	650	850	1.000	1.250
	650	900	1.050	1.250
	700	900	1.050	1.350
	700	900	1.050	1.350
	700	900	1.050	1.350
	700	900	1.050	1.400
	700	900	1.100	1.400
	700	900	1.100	1.400
	750	900	1.100	7.750
	750	900	1.150	8.000
	750	900	1.150	10.600
N	24	24	24	24
M	614,58	854,17	1.008,33	2.200,00
s	102,66	50,90	76,14	2.585,20
ΣX	14.750,00	20.500,00	24.200,00	52.800,00
$(\Sigma X)^2$	217.562.500,00	420.250.000,00	585.640.000,00	2.787.840.000,00
ΣX^2	9.307.500,00	17.570.000,00	24.535.000,00	269.875.000,00

U okviru računskog postupka za jednosmjernu analizu varijance dobivene su sljedeće vrijednosti:

1.	$\mathbf{GT = (SX)^2 / N}$	$GT = 131.250.651,04$
2.	$\mathbf{SumX^2}$	$SumX^2 = 321.287.500,00$
3.	$\mathbf{SS_{tot} = SX^2 - GT}$	$SS_{tot} = 190.036.848,96$
4.	$\mathbf{SS_{bg} = ST^2_j / n - GT}$	$SS_{bg} = 35.886.536,46$
5.	$\mathbf{SS_{wg} = SX^2 - S(T^2_j / n_j)}$	$SS_{wg} = 154.150.312,50$
provjera	$\mathbf{SS_{tot} = SS_{bg} + SS_{wg}}$	$SS_{tot} = 190.036.848,96$
6.	$\mathbf{df_{bg} = (k - 1)}$	$df_{bg} = 3$
	$\mathbf{df_{wg} = (N - k)}$	$df_{wg} = 92$
	$\mathbf{df_{tot} = df_{bg} + df_{wg}}$	$df_{tot} = 95$
7.	$\mathbf{MS_{bg} = SS_{bg} / df_{bg}}$	$MS_{bg} = 11.962.178,82$
	$\mathbf{MS_{wg} = SS_{wg} / df_{wg}}$	$MS_{wg} = 1.675.546,88$
8.	$\mathbf{F = MS_{bg} / MS_{wg}}$	$F = 7,14$
	p	$p = 0,00$
	$\mathbf{F_{0,05}\ (3,92)}$	$\approx 2,72$

9. Tabela ANOVA

Izvor varijabiliteta	*Suma kvadrata (SS)*	*Stepeni slobode (df)*	*Varijanca (MS)*	*F*	*p*
između grupa	35.886.536,46	3	11.962.178,82	7,14	0,00
unutar grupa	154.150.312,50	92	1.675.546,88		
Total	**190.036.848,96**	**95**			

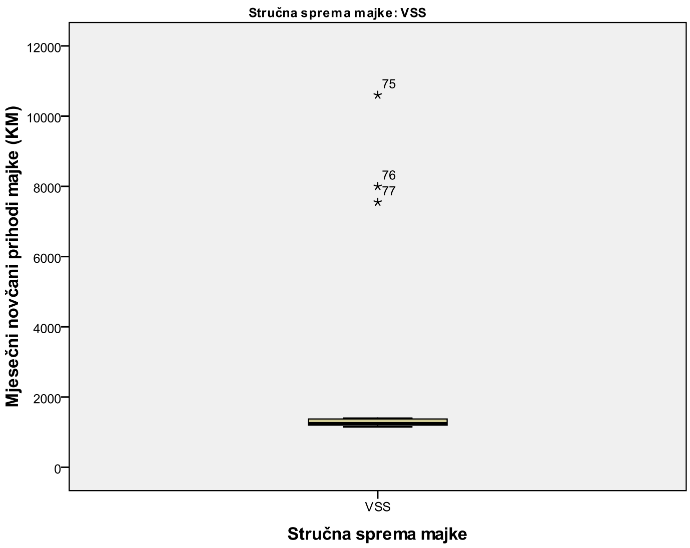
Analiza varijance pokazuje da među grupama (definiranim na temelju stručne spreme) postoji statistički značajna razlika u visini mjesečnih primanja (što znači da nul-hipotezu treba odbaciti). Međutim, naknadnom provjerom početne distribucije rezultata, nastavnik matematike (koji je provodio analizu) uvidio je da postoji osnova za sumnju da četiri grupne distribucije nisu normalno distribuirane (čime ne bi bio ispunjen jedan od osnovnih uvjeta za provođenje analize varijance –

normalnost distribucija). Osim toga, varijabilitet rezultata grupa značajno se razlikuje (najniža s=50,9, najviša s=2.585,20). Kako bi provjerio svoju pretpostavku, nastavnik je za svaku grupu izračunao veličinu odstupanja distribucije mjesečnih primanja od normalne distribucije (drugim riječima, izračunao je skjunis za svaku grupu). Dobio je sljedeće rezultate:

	Stručna sprema			
	NSS	SSS	VŠS	VSS
Skjunis	-,364	-,988	,339	2,600
Std. greška skjunisa	,472	,472	,472	,472

Kako se iz tabele vidi, sve distribucije više ili manje odstupaju od normalne (distribucije za NSS i SSS su negativno asimetrične, a distribucije za VŠS i VSS pozitivno asimetrične). Međutim, samo za VSS distribucija mjesečnih primanja *značajno* odstupa od normalne distribucije. To je moguće zaključiti na osnovu toga što je skjunis za ovu distribuciju veći od svoje standardne pogreške za više od tri puta.

Razlog asimetričnosti distribucije za grupu VSS jesu tri ekstremna rezultata, kako je to ilustrirano u dolje priloženom box-plotu. Ovi rezultati (kako se vidi na grafikonu) pripadaju ispitanicama pod rednim brojem 75, 76 i 77 u bazi podataka, odnosno to su rezultati od 7.750, 8.000 i 10.600 KM i svojom veličinom distribuciju „pomjeraju" u desnu stranu.



Nastavnik matematike imao je nekoliko opcija na raspolaganju kako bi riješio problem esktremnih rezultata. Prva opcija bila je da pronađe anketne upitnike za ove tri osobe i da provjeri da li je navedena vrijednost mjesečnih primanja tačno unesena u bazu podataka. Ispostavilo se da jeste. Druga opcija koju je nastavnik imao na raspolaganju jeste da telefonski kontaktira navedene ispitanice kako bi provjerio da li je do eventualne greške došlo prilikom ispunjavanja ankete. Nakon što je ispitanicama detaljno objasnio zašto ih ponovo kontaktira vezano za njihova (visoka) primanja (a na njihovo insistiranje, kako su sve tri bile vrlo sumnjičave) te nakon što se još jednom obavezao da se ovi podaci nigdje neće povezati sa njihovim ličnim informacijama, sve su ispitanice potvrdile da su to zaista njihova tačna primanja. Međutim, kako su ovakvi slučajevi zaista ekstremni u populaciji, nastavnik je još uvijek želio biti siguran da su utvrđene razlike među grupama značajne i bez ovako rijetkih vrijednosti. Kako bi to učinio, nastavnik je odlučio isključiti ova tri ekstremna rezultata iz analize varijance. Na taj način grupa VSS bi imala 21 ispitanicu, čime veličina grupa uključenih u analizu ne bi bila jednaka (što može narušiti pretpostavku o homogenosti varijanci u grupama). Zbog toga je nastavnik odlučio iz svake od tri preostale grupe po slučaju isključiti po tri ispitanice. Konačno, dobio je sljedeće distribucije rezultata:

	Stručna sprema majke			
	NSS	SSS	VŠS	VSS
Mjesečni novčani prihodi majke (KM)	450	750	900	1150
	450	750	900	1150
	500	750	900	1150
	500	800	950	1200
	500	800	950	1200
	500	850	950	1200
	550	850	950	1200
	600	850	950	1200
	600	850	1000	1250
	600	850	1000	1250
	650	850	1000	1250
	650	850	1000	1250
	650	850	1050	1250
	650	900	1050	1250
	700	900	1050	1250
	700	900	1050	1350
	700	900	1050	1350
	700	900	1100	1350
	750	900	1100	1400
	750	900	1150	1400
	750	900	1150	1400
N	21	21	21	21
M	614,29	850,00	1.009,52	1.259,52
s	100,18	52,44	76,84	83,09
ΣX	12.900,00	17.850,00	21.200,00	26.450,00
(ΣX)2	166.410.000,00	318.622.500,00	449.440.000,00	699.602.500,00
ΣX2	8.125.000,00	15.227.500,00	21.520.000,00	33.452.500,00

Ponovnim uvidom u mjere (a)simetričnosti pojedinačnih distribucija, nastavnik je zaključio da se sve četiri distribucije sada mogu smatrati normalnim (niti u jednoj distribuciji vrijednost skjunisa nije tri puta veća od pripadajuće standardne pogreške skjunisa):

	Stručna sprema majke			
	NSS	SSS	VŠS	VSS
Skjunis	-,270	-,862	,287	,533
Std. greška skjunis	,501	,501	,501	,501

U okviru računskog postupka analize varijance, nastavnik je dobio sljedeće vrijednosti:

1.	$\mathbf{GT = (SX)^2 / N}$	$GT = 73173333{,}33$
2.	$\mathbf{SumX^2}$	$SumX^2 = 78325000$
3.	$\mathbf{SS_{tot} = SX^2 - GT}$	$SS_{tot} = 5151666{,}667$
4.	$\mathbf{SS_{bg} = S\ T2j / n - GT}$	$SS_{bg} = 4639761{,}905$
5.	$\mathbf{SS_{wg} = SX^2 - S(T2j / nj)}$	$SS_{wg} = 511904{,}7619$
provjera	$\mathbf{SS_{tot} = SS_{bg} + SS_{wg}}$	$SS_{tot} = 5151666{,}667$
6.	$\mathbf{df_{bg} = (k - 1)}$	$df_{bg} = 3$
	$\mathbf{df_{wg} = (N - k)}$	$df_{wg} = 80$
	$\mathbf{df_{tot} = df_{bg} + df_{wg}}$	$df_{tot} = 83$
7.	$\mathbf{MS_{bg} = SS_{bg} / df_{bg}}$	$MS_{bg} = 1546587{,}302$
	$\mathbf{MS_{wg} = SS_{wg} / df_{wg}}$	$MS_{wg} = 6398{,}809524$
8.	$\mathbf{F = MS_{bg} / MS_{wg}}$	$F = 241{,}70$
	$\mathbf{p}$	$p = 0{,}0000$
	$\mathbf{F_{0,05}\ (3, 80)}$	$\approx 2{,}72$

9. Tabela ANOVA

Izvor varijabiliteta	*Suma kvadrata (SS)*	*Stupnjevi slobode (df)*	*Varijanca (MS)*	*F*	*p*
između grupa	4.639.761,90	3	1.546.587,30	241,70	0,00
unutar grupa	511.904,76	80	6.398,81		
Total	**5.151.666,67**				

Kao što vidimo, dobiveni F omjer je ponovo statistički značajan; međutim u ovoj analizi je mnogo veći od onog dobivenog u prvom postupku (7,14 prema 241,70).

Nakon zaključka da se grupe različitih nivoa obrazovanja statistički značajno razlikuju prema visini mjesečnih primanja, nastavniku matematike ostalo je da utvrdi koje tačno grupe se međusobno razlikuju (odnosno u kojim parovima aritmetičkih sredina se pojavljuje statistički značajna razlika). Da bi to utvrdio, nastavnik je proveo post-hoc analizu razlika među aritmetičkim sredinama, služeći se Scheffe-ovim postupkom:

$$F = (M_a - M_b)^2 / [MS_{wg} (n_a + n_b) / n_a n_b]$$

$$F_{0,05}\ (3,80) \approx 2,72$$

$$F_{gr}\ (k - 1) = 2,72 \times 3 = 8,16$$

Dobiveni su sljedeći rezultati:

Razlike između AS	**F**	**p**
M1 - M2	91,17	<0,05
M1 - M3	256,33	<0,05
M1 - M4	683,17	<0,05
M2 - M3	41,76	<0,05
M2 - M4	275,20	<0,05
M3 - M4	102,56	<0,05

Kako se sve aritmetičke sredine statistički značajno razlikuju jedna od druge, možemo zaključiti da stručna sprema (na svim nivoima) utiče na mjesečna primanja ispitanica.

PRIMJER 10.4

Jedna grupa ispitanika (N=6) učestvovala je u istraživanju zapamćivanja riječi različitog emocionalnog značenja (neutralne, pozitivne i negativne riječi). U tabeli ispod prikazan je broj zapamćenih riječi s obzirom na njihov emocionalni ton.

Ispitanik	**Neutralne riječi**	**Pozitivne riječi**	**Negativne riječi**	T_s
1	12	16	15	*43*
2	10	11	13	*34*
3	14	14	18	*46*
4	8	9	11	*28*
5	12	13	15	*40*
6	16	15	18	*49*
T_t	*72*	*78*	*90*	

Izračunavanje pomoćnih vrijednosti (ΣX^2 i $(\Sigma X)^2 / N$)

$\Sigma X^2 = 12^2 + 10^2 + 14^2 + ... + 11^2 + 15^2 + 18^2 = 3340$

$(\Sigma X)^2 / N = (12 + 10 + 14 + ... + 11 + 15 + 18)^2 / 18 = 3200$

Totalna suma kvadrata

$SS_{tot} = \Sigma X^2 - (\Sigma X)^2 / N$

$SS_{tot} = 3340 - 3200 = 140$

Suma kvadrata između ispitanika

$SS_{bs} = \Sigma T_s^2 / t - (\Sigma X)^2 / N$

$SS_{bs} = (43^2 + 34^2 + 46^2 + ... + 49^2) / 3 - 3200 = 3302 - 3200 = 102$

Suma kvadrata između tretmana

$SS_t = \Sigma T_t^2 / n - (\Sigma X)^2 / N$

$SS_{tr} = (72^2 + 78^2 + 90^2) / 6 - 3200 = 3228 - 3200 = 28$

Suma kvadrata reziduala

$SS_{rez} = SS_{tot} - SS_{bs} - SS_{tr}$

$SS_{rez} = 140 - 102 - 28 = 10$

Određivanje stupnjeva slobode

$df_{bs} = n – 1 = 6 – 1 = 5$

$df_{tr} = t – 1 = 3 – 1 = 2$

$df_{rez} = (n – 1)(t – 1) = 5 \times 2 = 10$

Određivanje prosječnog kvadrata

$MS_{tr} = SS_{tr} / df_{tr} = 28 / 2 = 14$

$MS_{rez} = SS_{rez} / df_{rez} = 10 / 10 = 1$

Određivanje F-vrijednosti

$F = MS_{tr} / MS_{rez} = 14 / 1 = 14$

Granična F-vrijednosti: $F_{0,05}$ (2 / 10) = 4,10

Tabela analize varijance

Izvor varijabiliteta	**Suma kvadrata (SS)**	**Stupnjevi slobode (df)**	**Varijanca (MS)**	**F**	**p**
Između ispitanika	102	5			
Unutar ispitanika					
Između tretmana	28	2	14	14	<0,05
Rezidual	10	10	1		
Total	140	17			

PRIMJER 10.5

Tri grupe ispitanika muškog i ženskog spola učestvovali su u eksperimentalnom ispitivanju efekata različitih doza lijeka na simptome depresivnosti. U tabeli ispod prikazani su rezultati koje su ispitanici postigli na skali depresivnosti nakon tretmana odgovarajućom dozom lijeka (veći rezultat znači izraženije simptome depresivnosti).

	Placebo	Srednja doza lijeka	Velika doza lijeka	T_{Ai}
žene	38	33	23	
	35	32	26	
	33	26	21	
	T_{11} = 106	T_{12} = 91	T_{13} = 70	267
muškarci	33	34	34	
	31	36	31	
	28	34	32	
	T_{21} = 92	T_{22} = 104	T_{23} = 97	293
T_{Bj}	198	195	167	560

Faktorijalnom analizom varijance provjeriti ćemo **glavne efekte** i **efekat interakcije** Spol x Doza.

Izračunavanje pomoćnih vrijednosti (ΣX^2 i GT)

$\Sigma X^2 = 38^2 + 35^2 + 33^2 + 33^2 + ... + 33^2 + ... + 32^2 = 17776$

$(\Sigma X)^2 / N = (38 + 35 + 33 + 33 + ... + 32)^2 / 18 = 560^2 / 18 = 17422,22$

Totalna suma kvadrata

$SS_{tot} = \Sigma X^2 – (\Sigma X)^2/N = 17776 – 17422,22 = 353,78$

Suma kvadrata između grupa

$SS_{bg} = \Sigma T_{ij}^2 / n – (\Sigma X)^2/N = (106^2 + 91^2 + 70^2 + 92^2 + 104^2 + 97^2) / 3 – 17422,22$

$= 53106 / 3 – 17422,22 = 17702 – 17422 = 279,78$

Suma kvadrata za prvi faktor

$SS_A = \Sigma T_A^2 / nb – (\Sigma X)^2/N = (267^2 + 293^2) / 3 x 3 – 17422,22$

$= 17459,78 – 17422,22 = 37,56$

Suma kvadrata za drugi faktor

$SS_B = \Sigma T_B^2 / na - (\Sigma X)^2/N = (198^2 + 195^2 + 167^2) / 3 \times 2 - 17422{,}22$

$= 17519{,}67 - 17422{,}22 = 97{,}45$

Suma kvadrata za interakciju

$SS_{AB} = SS_{bg} - SS_A - SS_B = 279{,}78 - 37{,}56 - 97{,}45 = 144{,}77$

Suma kvadrata unutar grupa

$SS_{wg} = \Sigma X^2 - \Sigma T_{ij}^2 / n = 17776 - 17702 = 74$

Provjera: $SS_{tot} = SS_{bg} + SS_{wg}$

353,78 = 279,78 + 74

Određivanje stupnjeva slobode

$df_{tot} = N - 1 = 18 - 1 = 17$

$df_A = a - 1 = 2 - 1 = 1$

$df_B = b - 1 = 3 - 1 = 2$

$df_{AB} = (a - 1)(b - 1) = 1 \times 2 = 2$

$df_{wg} = N - a \times b = 18 - 2 \times 3 = 18 - 6 = 12$

$df_{tot} = df_{wg} + df_A + df_B + df_{AB}$

Određivanje prosječnog kvadrata

$MS_{wg} = SS_{wg} / df_{wg} = 74 / 12 = 6{,}17$

$MS_A = SS_A / df_A = 37{,}56 / 1 = 37{,}56$

$MS_B = SS_B / df_B = 97{,}45 / 2 = 48{,}72$

$MS_{AB} = SS_{AB} / df_{AB} = 144{,}77 / 2 = 72{,}38$

Određivanje F-vrijednosti

$F_A = MS_A / MS_{wg} = 37{,}56 / 6{,}17 = 6{,}09$

$F_B = MS_B / MS_{wg} = 48{,}72 / 6{,}17 = 7{,}90$

$F_{AB} = MS_{AB} / MS_{wg} = 72{,}38 / 6{,}17 = 11{,}73$

Granične F-vrijednosti:

A: $F_{0,05}$ (1/12) =4,75

B: $F_{0,05}$ (2/12) = 3,89

A x B: $F_{0,05}$ (2/12) = 3,89

Tabela analize varijance

Izvor varijabiliteta	**Suma kvadrata (SS)**	**Stepeni slobode (df)**	**Varijanca (MS)**	**F**	**p**
Između grupa	279,78	5			
A (spol)	37,56	1	37,56	6,09	p<0,05
B (tretman)	97,45	2	48,72	7,90	p<0,05
A x B	144,77	2	72,38	11,73	p<0,05
Unutar grupa	74	12	6,17		
Total	353,78	17			

Glavni efekti

Ako posmatramo razliku između muškaraca i žena, bez obzira na tretman, govorimo o glavnom efektu varijable Spol.

$M_m = 32,56$, $M_ž = 29,67$

$F = 6,09$; $p < 0,05$

Također, ako posmatramo razliku između tretmana, bez obzira na spol, govorimo o glavnom efektu varijable Tretman.

$M_p = 33$, $M_{sd} = 32,5$, $M_{vd} = 27,83$

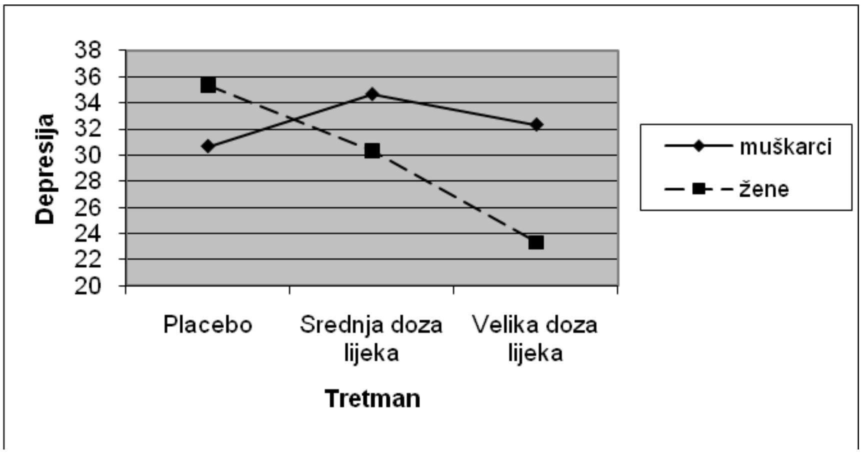
$F = 7,90$; $p < 0,05$

Interakcija

Kako bi se razumjela interakcija, korisno je odrediti aritmetičke sredine za svaku ćeliju.

	Placebo	Srednja doza lijeka	Velika doza lijeka	M_A
žene	35,33	30,33	23,33	**29,67**
muškarci	30,67	34,67	32,33	**32,56**
M_B	**33**	**32,5**	**27,83**	$M_G = 31,11$

Nakon određivanja AS, potrebno je napraviti grafički prikaz i interpretirati interakciju. Zavisnu varijablu (Depresija) nanosimo na Y osu, jednu od nezavisnih varijabli na X osu (Tretman), a posebnim linijama označavamo dva nivoa druge nezavisne varijable (Spol).



Iz grafičkog prikaza očigledno je da se nivo depresivnosti na različite načine mijenja kod muškaraca i žena s obzirom na dozu lijeka. Kod žena velika doza lijeka dovodi do znatnog smanjenja simptoma, dok dok muškaraca to nije slučaj – čak je nivo depresivnosti podjednak kao i kod placeba!

ZADACI

1. Testirajte značajnost razlika između aritmetičkih sredina.

A	B	C
90	100	86
92	79	91
87	114	120
110	89	115
115	93	79

2. Testirajte značajnost razlika između aritmetičkih sredina.

A	B	C
9	10	14
10	10	13
12	11	17
13	17	22
11	9	18
17	17	17
19	11	19
14	19	15
20	18	20
16	17	24

3. Bez računanja odredite F i p.

A	B	C	D
10	12	10	14
13	14	14	10
12	10	12	12
14	13	11	13
11	11	13	11

4. U jednom istraživanju ispitivano je da li postoje razlike u uspješnosti podučavanja biologije s obzirom na korištene metode. Eksperimentalni dizajn podrazumijeva formiranje tri grupe (A, B i C metoda) i razvrstavanje učenika u jednu od tri grupe metodom slučajnog odabira. Na kraju školske godine svi učenici su rješavali standardizirani test znanja iz biologije (raspon rezultata od 0 do 100 bodova). Rezultati ANOVA-e su prikazani u tabeli ispod.

	SS	**df**	**MS**	**F**	**p**
Između grupa (bg)	300				
Unutar grupa (wg)		20	20		
Total	700	22			

a. Na osnovu rezultata prikazanih u tabeli, odredite vrijednosti koje nedostaju (SS_{wg}, df_{bg} i MS_{bg}) i upišite ih u tabelu u odgovarajuća polja.
b. Koliko je ispitanika učestvovalo u istraživanju?
c. Šta možemo zaključiti na osnovu dobivene F-vrijednosti?
d. Da li se aritmetičke sredine statistički značajno razlikuju? DA NE
e. Šta biste uradili nakon što odgovorite na pitanje da li se aritmetičke sredine statistički značajno razlikuju?

5. Dopunite tabelu i odgovorite na pitanja ispod.

Izvor varijabiliteta	***Suma kvadrata (SS)***	***Stepeni slobode (df)***	***Varijanca (MS)***	***F***	***p***
između grupa		3			
unutar grupa	180,80				
Total	**276,95**	**19**			

Koliko je ukupno ispitanika učestvovalo u istraživanju? Da li su razlike između AS statistički značajne?

6. Tri grupe ispitanika muškog i ženskog spola učestvovale su u tri različita eksperimentalna uvjeta zapamćivanja besmislenih riječi. Rezultati za svaku situaciju prikazani su u donjoj tabeli kao i broj upamćenih riječi za svakog ispitanika.

		I	**II**	**III**
SPOL		10	11	19
		11	13	17
	m	14	15	18
		13	10	15
		18	14	14
		11	10	20
		11	14	16
	ž	14	15	18
		14	10	15
		17	15	14

Ispitati statističku značajnost efekata faktora SPOL, GRUPA te interakcije SPOL X GRUPA.

7. U tabeli ispod navedite vrijednosti koje nedostaju te grafički prikažite glavne efekte i efekat interakcije varijabli A i B (nije potrebno računski provoditi analizu varijance) (na graficima na odgovarajući način označite X ose te, ukoliko je potrebno formirajte legendu).

	B1	B2	B3	M_A
A1	2	4	14	
A2	6	8	9	
A3	12	2	6	
M_B				

8. Tri grupe od po 9 ispitanika su učili gradivo različitim metodama (metode A, B i C). Uspješnost učenja izmjerena je testom znanja. Vrijednosti u bodovima navedene su za svakog ispitanika u svakoj grupi u tabeli ispod.

GRUPA A	**GRUPA B**	**GRUPA C**
49	47	32
31	25	20
28	32	30
35	35	18
38	33	22
50	41	37
55	44	34
30	34	32
42	40	35

a. Provjerite značajnost razlika u uspješnosti učenja između tri grupe ispitanika.
b. Koja metoda je najmanje efikasna u učenju (A, B ili C)?

9. Dopunite tabelu sa podacima za grupe B i C, tako da vrijede rezultati prikazani u tabeli ANOVA-e.

A	**B**	**C**
10		
11		
12		
11		
10		

Izvor varijabiliteta	***Suma kvadrata (SS)***	***Stepeni slobode (df)***	***Varijanca (MS)***	***F***	***p***
između grupa	0,00	2	0,00	0,000	1,000
unutar grupa	8,40	12	0,70		
Total	**8,4**	**14**			

10. U sljedećoj tabeli prikazani su rezultati analize varijance nekog hipotetičkog eksperimenta u kojem su ispitanici rješavali matematičke zadatke na četiri nivoa težine i bili kažnjeni za pogrešna rješenja sa pet različitih tipova kazne. Neke vrijednosti u tabeli su izbrisane, ali sve što je izbrisano može se izračunati na osnovu datih rezultata. Kompletirajte tabelu tako što ćete upisati odgovarajuće vrijednosti!

Izvor variranja	**SS**	**df**	**MS**	**F**	**p**
Između grupa					
Težina zadatka (A)	100				
Tip kazne (B)	150				
A X B			5		
Unutar grupa		90			
Total	1190				

11. Korelacija i regresija

Korelacija je statistička mjera povezanosti između dvije ili više varijabli. Korelacijom opisujemo odnos između dvije ili više varijable: kakvog je oblika (npr.: linearan, kvadratičan, logaritamski, obrnuto U) i koliko iznosi. Npr., konzumiranje cigareta povezano je sa različitim oboljenjima; sa povećanjem konzumiranja cigareta povećava se učestalost različitih oboljenja. Sposobnost odgađanja zadovoljenja određenih potreba u ranom djetinjstvu povezano je sa socijalnom kompetencijom u odrasloj dobi ili – s povećanjem sposobnosti odgađanja zadovoljenja potreba u ranoj dobi, povećava se socijalna kompetencija u odrasloj dobi. U medicinskim istraživanjima utvrđena je povezanost između tjelesne visine i učestalosti srčanog udara kod žena: što je visina veća, učestalost srčanog udara je manja. U svakodnevnom životu često koristimo koncept povezanosti. Npr., ako na parkingu ispred tržnog centra vidite veliki broj parkiranih automobila, očekujete da će u centru biti gužva; ako je na parkingu mali broj automobila, ne očekujete gužvu. Na osnovu ovih očekivanja (predviđanja) donosite odgovarajuću odluku.

Korelacija može biti pozitivna (porastu jedne odgovara porast druge varijable), negativna (porastu jedne odgovara opadanje druge varijable), maksimalna, perfektna (veće slaganje ne može postojati) i nulta (između dvije varijable ne postoji povezanost).

Ako su dvije varijable u korelaciji, to znači da na osnovu jedne varijable možemo, manje ili više precizno, predvidjeti rezultate druge varijable. Konceptu korelacije, blizak je **koncept regresije** (tačnije, regresijske analize). Regresijskom analizom **predviđamo** vrijednosti jedne varijable, koju nazivamo **kriterijska**, na osnovu informacija koje imamo o drugoj varijabli, koju nazivamo **prediktorska** varijabla.

U tekstu koji slijedi najprije ćemo opisati koncept korelacije, a zatim koncept regresije.

Korelacija

U opisivanju povezanosti koristimo **grafičke** i **numeričke** metode.

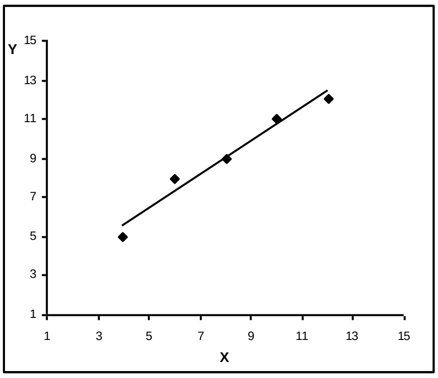
Grafička metoda podrazumijeva grafički prikaz povezanosti izmedu dvije varijable preko **skater – dijagrama** (eng. *scatter – diagram*). Vrijednosti jedne varijable nanose se na X osu, a druge na Y osu. Svaki ispitanik se prikazuje kao tačka u koordinatnom sistemu [uređeni par, (x, y)]. Prediktorska varijabla nanosi se na apscisu, a kriterijska na ordinatu. Skater dijagram omogućava da: odredimo smjer povezanosti; procjenimo da li je povezanost visoka, umjerena ili niska; utvrdimo kakvog je oblika povezanost (linearna, inverzna U, ili neka druga); detektiramo ekstremne rezultate.

U skater dijagram unosi se **pravac** kojeg nazivamo "pravac regresija" Y na X. Predstavlja najbolju predikciju Y_i za datu vrijednost X_i. Stupanj u kojem se tačke "okupljaju" oko pravca odražava veličinu povezanosti tj. korelacije između X i Y. Ukoliko se sve tačke nalaze na pravcu regresije, radi se o perfektnoj povezanosti, tj. maksimalno mogućoj povezanosti.

Pretpostavimo da se u istraživanju ispitivala povezanost između inteligencije (X) i školskog uspjeha (Y) i da smo prikupili podatke (školski uspjeh i rezultat na testu inteligencije) za pet ispitanika. Možemo pretpostaviti sljedeća četiri opisa povezanosti između ove dvije varijable: 1) učenici koji su inteligentniji, postižu viši školski uspjeh, 2) učenici koji su inteligentniji, postizat će niži školski uspjeh, 3) učenici koji su visoko inteligentni ili manje inteligentni, postizat će niži školski uspjeh u poređenju sa prosječno inteligentnim učenicima, koji će postizati visok školski uspjeh, i 4) ne postoji povezanost između inteligencije i školskog uspjeha Za svaki opis navedeni su podaci, a skater-dijagramom grafički je prikazana povezanost.

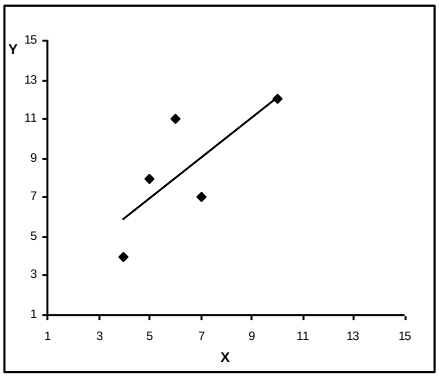
1. *Učenici koji su inteligentniji, postižu viši školski uspjeh*. Povezanost je pozitivna. Ispitanici koji na jednoj varijabli postižu više rezultate, također postižu više rezultate i na drugoj varijabli, i obratno, oni koji postižu niže rezultate na jednoj varijabli, postižu niže rezultate i na drugoj varijabli.

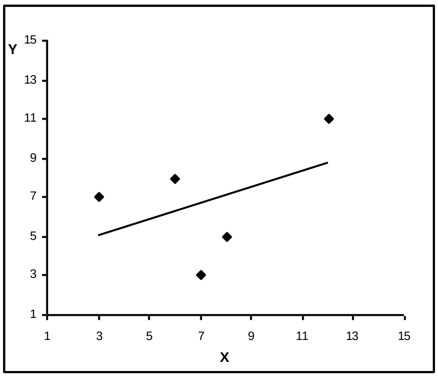
ispitanik	x	y
A	4	5
B	6	8
C	8	9
D	10	11
E	12	12



Skater dijagram u ovom slučaju opisuje pozitivnu povezanost između dvije varijable. Povezanost nije perfektna (tačke ne leže tačno na pravcu regresije). S obzirom da su tačke blizu pravca, povezanost je visoka.

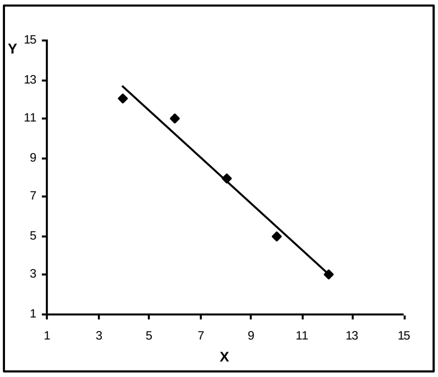
U skater-dijagramu ispod, opisani su slučajevi kada tačke nisu toliko blizu pravca regresije. Povezanost nije toliko visoka.





2. *Učenici koji su inteligentniji, postizat će niži školski uspjeh (neće se dovoljno truditi jer gradivo smatraju jednostavnim, možda i dosadnim).* Povezanost je negativna. Ispitanici koji na jednoj varijabli postižu više rezultate, na drugoj varijabli postižu niže rezultate, i obratno, oni koji postižu niže rezultate na jednoj varijabli, postižu više rezultate na drugoj varijabli.

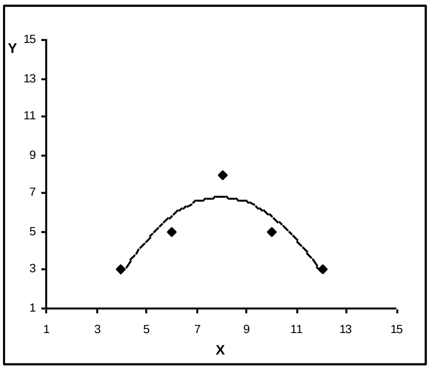
ispitanik	x	y
A	4	12
B	6	11
C	8	8
D	10	5
E	12	3



Skater dijagram opisuje negativnu povezanost između dvije varijable. Kao i u prethodnom slučaju, povezanost nije perfektna (tačke ne leže tačno na pravcu regresije), ali je visoka.

3. *Učenici koji su visoko inteligentni ili manje inteligentni, postizat će niži školski uspjeh u poređenju sa prosječno inteligentnim učenicima, koji će postizati visok školski uspjeh.* Povezanost između dvije varijable je nelinearna (zakrivljena). Ispitanici koji na jednoj varijabli postižu više i niže rezultate, postižu niže rezultate na drugoj varijabli, za razliku od ispitanika koji na prvoj varijabli postižu prosječne vrijednosti, a na drugoj visoke.

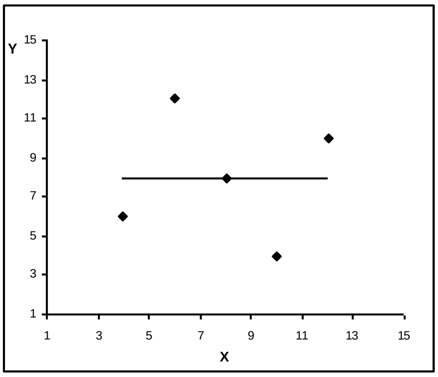
ispitanik	x	y
A	4	3
B	6	5
C	8	8
D	10	5
E	12	3



Povezanost između dvije varijable je nelinearna (zakrivljena). Pravac ne opisuje najbolje odnos između ove dvije varijable!

4. *Ne postoji povezanost između inteligencije i školskog uspjeha.* Povezanost između dvije varijable ne postoji (nulta povezanost). Bez obzira koji rezultat ostvarili na jednoj varijabli, na drugoj varijabli ispitanici mogu postići bilo koji rezultat.

ispitanik	x	y
A	4	6
B	6	12
C	8	8
D	10	4
E	12	10



Povezanost između dvije varijable ne postoji (nulta povezanost). Bilo gdje da ucrtamo pravac, nećemo adekvatno opisati nultu povezanost. Zapravo, ne postoji pravac koji najbolje opisuje ovakvu povezanost.

Pearsonov koeficijent korelacije

Stupanj povezanosti izražava se **koeficijentom korelacije**, r. Vrijednost koeficijenta korelacije kreće se u granicama od -1 (potpuno negativna povezanost) do 1 (potpuno pozitivna povezanost). Karl Pearson razradio je **računski postupak** za izračunavanje stupnja povezanosti (Pearsonov produkt-moment koeficijent korelacije).

Izračunavanje Pearsonovog koeficijenta korelacije preko kovarijanci

Kovarijanca je stepen u kojem dvije varijable zajedno variraju (ko-variraju). Kovarijanca ukazuje na „dijeljenu“ varijancu varijabli. Kovarijanca se izračunava preko izraza:

$$cov_{xy} = \frac{\sum(X - Mx)(Y - My)}{N - 1}$$

Pearsonov koeficijent korelacije izračunava se preko izraza:

$$r_{xy} = \frac{cov_{xy}}{s_x \times s_y}$$

Na grupi od deset ispitanika prikupljeni su podaci za varijable X i Y. Ispod je prikazan postupak izračunavanja Pearsonovog koeficijenta korelacije.

ISPITANICI	X	Y	X - Mx	Y - My	(X - Mx)(Y - My)
1	3	10	-3,60	-3,60	12,96
2	4	11	-2,60	-2,60	6,76
3	5	12	-1,60	-1,60	2,56
4	5	12	-1,60	-1,60	2,56
5	6	13	-0,60	-0,60	0,36
6	7	14	0,40	0,40	0,16
7	8	15	1,40	1,40	1,96
8	9	16	2,40	2,40	5,76
9	9	16	2,40	2,40	5,76
10	10	17	3,40	3,40	11,56
M=	6,6	13,6			Σ=50,4
s=	2,4	2,4			

cov_{xy} =5,6
s_x x s_y =5,6

Koeficijent korelacija iznosi:

$$r_{xy} = \frac{cov_{xy}}{s_x \times s_y} = \frac{5,6}{5,6} = 1$$

Izračunavanje Pearsonovog koeficijenta korelacije preko z-vrijednosti

Pearsonov koeficijent korelacije izračunava se preko izraza:

$$r_{xy} = \frac{\sum z_x \times z_y}{N-1}$$

Postupak izračunavanja Pearsonovog koeficijenta korelacije prikazan je ispod.

ISPITANICI	X	Y	z_x	z_y	z_x x z_y
1	3	10	-1,52	-1,52	2,31
2	4	11	-1,10	-1,10	1,21
3	5	12	-0,68	-0,68	0,46
4	5	12	-0,68	-0,68	0,46
5	6	13	-0,25	-0,25	0,06
6	7	14	0,17	0,17	0,03
7	8	15	0,59	0,59	0,35
8	9	16	1,01	1,01	1,03
9	9	16	1,01	1,01	1,03
10	10	17	1,44	1,44	2,06
M=	6,6	13,6			Σ=9
S=	2,4	2,4			

Koeficijent korelacija iznosi:

$$r_{xy} = \frac{\sum z_x \times z_y}{N-1} = \frac{9}{9} = 1$$

Važno je primjetiti da su z-vrijednosti varijabli X i Y za svakog ispitanika identične (npr. za prvog ispitanika, z_x=-1,52; z_y=-1,52), što je slučaj samo kada je povezanost između dvije varijable maksimalna.

Izračunavanje Pearsonovog koeficijenta korelacije iz sirovih rezultata

Za izračunavanje Pearsonovog koeficijenta korelacije iz sirovih rezultata koristimo sljedeći izraz:

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left\{\left[\sum X^2 - \left(\sum X\right)^2\right] \times \left[\sum Y^2 - \left(\sum Y\right)^2\right]\right\}}}$$

Postupak izračunavanja Pearsonovog koeficijenta korelacije prikazan je ispod.

ISPITANICI	X	Y	X^2	Y^2	XY
1	3	10	9	100	30
2	4	11	16	121	44
3	5	12	25	144	60
4	5	12	25	144	60
5	6	13	36	169	78
6	7	14	49	196	98
7	8	15	64	225	120
8	9	16	81	256	144
9	9	16	81	256	144
10	10	17	100	289	170
Σ=	66	136	486	1900	948

$(\Sigma X)^2 = 4356$
$(\Sigma Y)^2 = 18496$

Koeficijent korelacije iznosi:

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}} =$$

$$= \frac{10 \times 948 - 66 \times 136}{\sqrt{(10 \times 486 - 4356) \times (10 \times 1900 - 18496)}} = 1$$

Testiranje značajnosti r

Vrijednost dobivenog koeficijenta korelacije mogla je biti rezultat slučaja. Stoga je potrebno testirati **statističku značajnost** dobivenog koeficijenta korelacije. Matematički model (uz nultu hipotezu, tj. da ne postoji povezanost između dvije varijable) uključuje distribuciju svih koeficijenata korelacije od -1 do +1, sa aritmetičkom sredinom M=0 (r=0).

Statističku značajnost koeficijenta korelacije možemo provesti na dva načina: preko t-vrijednosti i uporeďivanjem dobivenog r sa graničnom vrijednošću r očitanom iz tablice. t-vrijednost se izračunava korištenjem izraza:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Preduvjeti za korištenje Pearsonovog koeficijenta korelacije

Pearsonov koeficijent korelacije računa se ako su ispunjeni sljedeći preduvjeti:

- Intervalna ili omjerna skala mjerenja
- Normalna, ili barem simetrična distribucija
- Linearan odnos između varijabli

Korelacija i uzročna veza

Sama činjenica da između dvije pojave postoji povezanost ne daje nam za pravo da te pojave povežemo uzročnom vezom.

Moguća objašnjenja odnosa između varijabli X i Y:

a. varijabla X utječe na varijablu Y ($X \rightarrow Y$)
b. varijabla Y utječe na varijablu X ($Y \rightarrow X$)
c. varijabla X utječe na varijablu Y i varijabla Y utječe na varijablu X ($X \leftrightarrow Y$)
d. varijabla Z utječe na varijable X i Y ($X \leftarrow Z \rightarrow Y$)

Korelacijskim istraživačkim nacrtom ne možemo odrediti prirodu odnosa između X i Y (a,b,c ili d). Uzročno-posljedična veza između dvije pojave može se dokazati samo eksperimentom.

Pravac regresije

Pravac ucrtan u skater-dijagram je pravac koji "najbolje odgovara podacima" i predstavlja najbolju moguću predikciju vrijednosti Y_i za datu vrijednost X_i. Uz pomoć pravca regresije možemo za bilo koju vrijednost X najtačnije prognozirati vrijednost varijable Y.

Prognoziranu vrijednost Y označavamo sa **Y′**. Pravac regresije definiran je nagibom i odsječkom na osi Y (kada je X=0), tj. izrazom:

$$\mathbf{Y' = a + bX}$$

gdje je:

Y′– prognozirani rezultat

a – odsječak na osi Y kada je X=0

b – nagib pravca

X – vrijednost prediktora

Nagib pravca govori nam koliko se mijenja vrijednost Y varijable uz jediničnu promjenu vrijednosti X varijable. Odsječak na Y osi je prognozirani rezultata kada je X=0.

Vrijednosti a i b određuju se izrazima:

$$a = M_y - bM_x$$

$$b = \frac{cov_{xy}}{s_x^{\,2}} \quad \text{ili} \quad b = r_{xy}\frac{s_y}{s_x}$$

Pretpostavimo da smo za grupu od pet ispitanika dobili sljedeće vrijednosti varijabli X i Y:

ispitanik	x	y
A	4	5
B	10	8
C	8	9
D	6	11
E	12	12

Za određivanje jednačine pravca regresije potrebno je izračunati vrijednosti *a* i *b*. Najprije ćemo izračunati nagib pravca, *b*. Za određivanje *b*, potrebno je izračunati kovarijancu (uz pomoć vrijednosti izračunatih ispod).

$(x-M_x)$	$(y-M_y)$	$(x-M_x)(y-M_y)$
-4	-4	16
2	-1	-2
0	0	0
-2	2	-4
4	3	12

$\Sigma(x-M_x)(y-M_y) = 22$

$$cov_{xy} = \frac{\sum(X-Mx)(Y-My)}{N-1} = \frac{22}{4} = 5{,}5$$

$$s_x^{\ 2} = 10$$

$$b = \frac{cov_{xy}}{s_x^{\ 2}} = \frac{5{,}5}{10} = 0{,}55$$

Odsječak na Y osi, kada je X=0, odredit ćemo preko izraza:

$$a = M_y - bM_x \qquad a = 9 - 0{,}55 \text{ x } 8 = 4{,}6$$

Uvrštavanjem izračunatih vrijednosti dolazimo do definiranog pravca regresije:

$$Y' = 4{,}6 + 0{,}55X$$

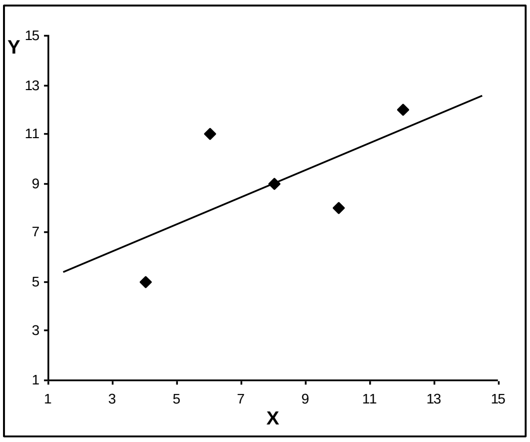
Kada znamo jednačinu pravca regresije, koristeći samo dvije vrijednosti X, u skater dijagramu možemo nacrtati pravac regresije. Odredimo Y´ za, npr., X_1=5 i X_2=11.

$Y'_1 = 4{,}6 + 0{,}55 \times 5$; $Y'_1 = 7{,}35$; Prva tačka A (X_1=5 ; Y'_1=7,35)

$Y'_2 = 4{,}6 + 0{,}55 \times 11$; $Y'_1 = 10{,}65$; Druga tačka B (X_2=11; Y'_2=10,65)

Ako kroz dvije tačke, A i B, povučemo liniju, dobit ćemo pravac regresije (slika ispod).

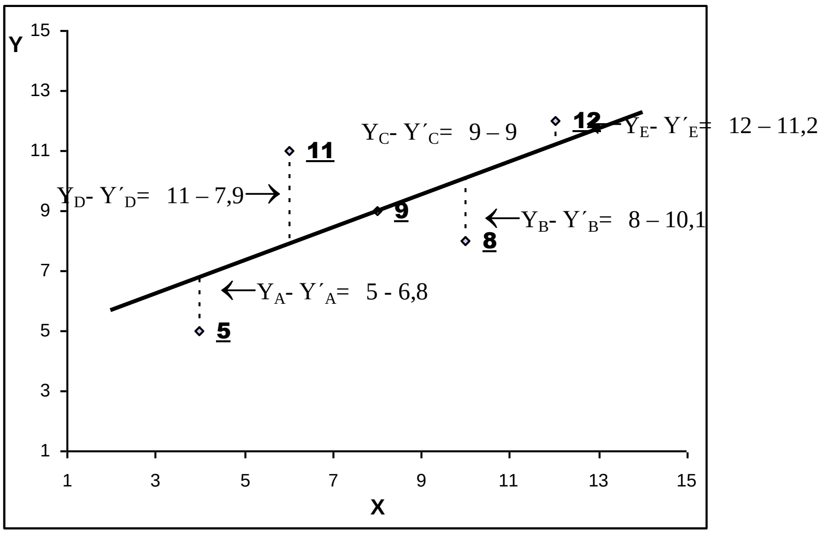
Slika 11.1: Pravac regresije



Pored pravca regresije Y na X, možemo odrediti i regresijski pravac X na Y. Ova dva pravca sjeku se u tački M_x, M_y. Pravci su identični samo u slučaju perfektne korelacije (r=±1).

Na istom primjeru pokazat ćemo i logiku određivanja pravca regresije. U koordinatnom sistemu možemo ucrtati beskonačno mnogo pravaca, ali samo jedan pravac je optimalan, i to onaj koji je tako definiran vrijednostima *a* i *b* da minimizira sumu kvadriranih Y–Y´ odstupanja. Ova metoda određivanja pravca naziva se **metoda najmanje sume kvadrata odstupanja**. U skater-dijagramu (slika 11.2) označena su Y–Y´ odstupanja za svaku Y vrijednost.

Slika 11.2: Odstupanja Y od prognozironog rezultata Y´



Vrijednosti *a* i *b* su takve da definiraju „najpošteniju” linearnu fukciju, tj. takve da pravac regresije prolazi što bliže vrijednostima Y varijable. Drugim riječima, potrebno je naći najoptimalnije vrijednosti *a* i *b*. Stoga se *a* i *b* određuje u terminima **pogreške predikcije**, tj. preko odstupanja rezultata Y od prognoziranog rezultata (koji se nalazi na pravcu regresije). Obzirom da je Y rezultat kojeg smo dobili, a Y´ rezultat kojeg očekujemo jednačinom pravca regresije, odstupanja Y–Y´ nazivamo **rezidualom**. Pravac koji prolazi skater-dijagramom treba da minimizira sumu kvadriranih odstupanja, tj. minimizira $\Sigma(Y–Y´)^2$. Na taj način dobiju se optimalne vrijednosti *a* i *b*.

Koristeći jednačinu pravca regresije možemo **prognozirati rezultat Y** na osnovu bilo kojeg rezultata X. Prognoza rezultata može se provesti grafičkom (koristeći skater-dijagram i pravac regresije) ili numeričkom metodom (koristeći jednačinu pravca regresije).

Koji je najvjerovatniji rezultat na varijabli Y ispitanika koji je na varijabli X postigao rezultat x=15?

$Y´ = 4{,}6 + 0{,}55X,$

$Y´ = 4{,}6 + 0{,}55 \times 15$

Dakle, najvjerovatniji rezultat, tj. prognozirani rezultat je:

$$Y´ = 12{,}85.$$

Drugi izraz (praktičniji) za određivanje prognoziranog rezultata je:

$$Y' = r\left(\frac{s_y}{s_x}\right)(X - Mx) + My$$

Iako bi se na osnovu pojma „prognoza” moglo zaključiti da varijabla X ima efekat na varijablu Y, to je pogrešno. Prognoza rezultata Y ne znači da smo utvrdili efekat jedne varijable na drugu! Povezanost između dvije varijable ne znači da jedna varijabla objašnjava drugu!

Pogreška prognoze

U samo jednom slučaju, prognoza rezultata Y je maksimalno precizna: kada je povezanost između dvije varijable maksimalna moguća. Kada je r=±1, sve tačke u skater-dijagramu nalaze se na pravcu regresije, što znači da su odstupanja Y–Y´ jednaka nuli, tj. rezidual je 0. U svim drugim slučajevima postoji određena odstupanja između utvrđenog i prognoziranog rezultata.

Izraz $\Sigma(Y-Y')^2$ nazivamo **suma kvadrata reziduala** i označavamo sa SS_{rez}. Pogreška prognoze je to veća što je suma kvadrata reziduala veća. SS_{rez} predstavlja varijabilitet koji „ostaje” kada na osnovu X prognoziramo Y (kažemo da se radi o varijabilitetu koji se ne može objasniti prognoziranjem Y na osnovu X).

U našem primjeru, odredit ćemo $\Sigma(Y-Y')^2$.

ispitanik	x	y	y´	y- y´	(y–y´)^2
A	4	5	6,8	-2	4
B	10	8	10,1	-2	4
C	8	9	9	0	0
D	6	11	7,9	3	9
E	12	12	11,2	1	1
				$\Sigma(Y-Y')^2=$	18

Suma kvadrata reziduala iznosi 18.

Suma kvadrata reziduala samo po sebi ne govori o pogrešci koja postoji prilikom prognoziranja rezultata. Za utvrđivanje pogreške prognoze koristimo tzv. **standardnu pogrešku prognoze**.

Standardna pogreška prognoze definirana je izrazom:

$$s_{y,x} = \sqrt{\frac{SS_{rez}}{N-2}} \quad \text{ili} \quad s_{y,x} = \sqrt{\frac{\sum\left(Y-Y'\right)^2}{N-2}}$$

Kvadrirana vrijednost $s_{y,x}$ naziva se **rezidualna varijanca** ili **varijanca pogreške**. U našem primjeru, standardna pogreška prognoze iznosi:

$$s_{y,x} = \sqrt{\frac{\Sigma\left(Y-Y'\right)^2}{N-2}} = \sqrt{\frac{18}{3}} = 2,45$$

Za određivanje **standardne pogreške prognoze koristimo i sljedeći izraz:**

$$s_{y,x} = s_y\sqrt{(1-r^2)\frac{N-1}{N-2}}$$

Za velike uzorke (N-1)/(N-2) praktično je 1, pa se koristi i izraz:

$$s_{y,x} = s_y\sqrt{(1-r^2)}$$

Standardna pogreška prognoze interpretira se kao standardna devijacija! Na osnovu standardne pogreške prognoze određujemo **interval pouzdanosti** prognoziranog rezultata:

Interval pouzdanosti = $Y' \pm s_{y,x} \text{ x } t_{\alpha/2}$

Izračunat ćemo interval pouzdanosti prognoziranog rezultata na varijabli Y, ispitanika koji je na varijabli X postigao rezultat x=15. Interval pouzdanosti iznosi (već smo odredili da je prognozirani rezultat Y´=12,85):

$IP = Y' \pm s_{y,x}\ t_{\alpha/2} = 12{,}85 \pm 2{,}45 \text{ x } 3{,}182 = 12{,}85 \pm 7{,}80$

$5{,}05 \leq Y \geq 20{,}65$

($t_{\alpha/2}$ očitavamo iz tablice graničnih vrijednosti t, za npr. 95% pouzdanost)

Dakle, 95% granice pouzdanosti su od 5,05 do 20,65.

Međutim, kada pravac regresije određujemo na osnovu relativno malog broja podataka, koristimo sljedeći izraz:

$$Y' \pm s_{y,x}\, t_{\alpha/2}\sqrt{1+\frac{1}{N}+\frac{(X-M_x)^2}{(N-1)s_x^{\,2}}}$$

Veličina intervala pouzdanosti prognoziranog rezultata određena je **veličinom koeficijenta korelacije** i, donekle, **veličinom uzorka**. Što je koeficijent korelacije veći, pogreška prognoze se smanjuje (ako je r=1, tada je $s_{y,x}=0$, dok kod r=0, $s_{y,x}=s_y$). Veličina uzorka ne utječe u značajnoj mjeri na veličinu intervala pouzdanosti.

Preduvjeti za korištenje regresije

Preduvjeti za pravac regresije isti su kao i za Pearsonov koeficijent korelacije. Osim toga, varijance Y za svaki X trebaju biti podjednake duž pravca regresije – homogenost varijanci, a vrijednosti Y′ za svaki X trebaju se normalno distribuirati oko prognozirane vrijednosti Y′(duž pravca regresije).

Koeficijent determinacije

Iz dva izraza za određivanje standardne pogreške prognoze slijedi:

$$s_{Y,X} = s_Y\sqrt{1-r^2}, s_{y,x} = \sqrt{\frac{SS_{rez}}{N-2}} \Rightarrow SS_{rez} = SS_y(1-r^2)$$

odnosno:

$$SS_{rez} = SS_Y - SS_Y\, r^2,$$

što nas dovodi do izraza:

$$r^2 = \frac{SS_y - SS_{rez}}{SS_y}$$

pri čemu je: $SS_y = \Sigma(Y - M_y)^2$

U gornjem izrazu SS_y je totalna suma kvadrata rezultata Y (totalno variranje rezultata Y) i sadržava totale:

1. sume kvadrata Y objašnjen sa X - $SS_y(r^2)$, i
2. sume kvadrata Y koji je nezavisan od X - SS_{rez}.

U našem primjeru, totalna suma kvadrata rezultata školskog uspjeha dijeli se na dio koji se može objasniti inteligencijom i dio (ostatak) koji objašnjavaju druge varijable (ne znamo koje). SS_{rez} je suma kvadrata Y koja je nezavisna od X i predstavlja mjeru pogreške nakon što preko X prognoziramo Y.

Nekoliko izvora varijabiliteta mogu se sumirati na sljedeći način:

1. $SS_X = \Sigma(X - M_X)^2$; varijabilitet inteligencije;
2. $SS_Y = \Sigma(Y - M_Y)^2$; varijabilitet školskog uspjeha;
3. $SS_{Y'} = \Sigma(Y' - M_y)^2$; varijabilitet školskog uspjeha objašnjen varijabilitetom inteligencije i
4. $SS_{rez} = \Sigma(Y - Y')^2 = SS_Y - SS_{Y'}$; varijabilitet školskog uspjeha koji se ne može objasniti varijabilitetom inteligencije.

Kako je već rečeno, jedan dio varijabiliteta školskog uspjeha objašnjen je inteligencijom ($SS_{Y'}$), a jedan nije (SS_{rez}). Bilo bi korisno da odredimo procenat ukupnog varijabiliteta školskog uspjeha koji se može objasniti varijabilitetom inteligencije, tj. potrebna nam je mjera koja predstavlja odnos:

$$\frac{SS_{y'}}{SS_y} = \frac{SS_y - SS_{rez}}{SS_y}$$

Ta mjera je **koeficijent determinacije, r^2**:

$$r^2 = \frac{SS_y{'}}{SS_y}$$

Koeficijent determinacije koristimo kako bi odredili postotak prognoziranog varijabiliteta. Koeficijent determinacije govori o proporciji varijance jedne varijable koja se može objasniti varijancom druge varijable.

Na primjeru prognoze školskog usjeha na osnovu inteligencije, koeficijent determinacije iznosi:

$$D = r^2 = 0{,}493.$$

Možemo zaključiti da je 49,3% varijabiliteta školskog uspjeha objašnjeno varijabilitetom inteligencije. Ostalo je 50,7% varijabiliteta školskog uspjeha koji se ne može objasniti varijabilitetom inteligencije.

PRIMJER 11.1

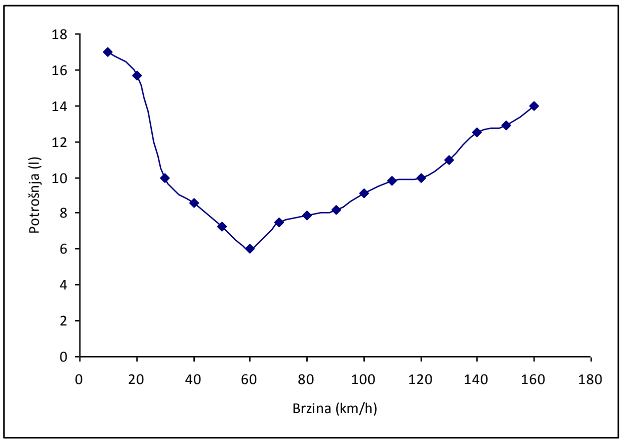
Na koji način se mjenja potrošnja goriva s povećanjem brzine automobila? Ispod su prikazani fiktivni podaci (koji oslikavaju realnu situaciju) količine potrošenog goriva na rastojanju od 100 km, pri različitim brzinama za automobil marke ASD.

Brzina (km/h)	Potrošnja (l)
10	20,0
20	15,7
30	10,0
40	8,6
50	7,3
60	6,0
70	7,5
80	7,9

Brzina (km/h)	Potrošnja (l)
90	8,2
100	9,1
110	9,8
120	10,0
130	11,0
140	12,5
150	12,9
160	13,3

Prikazat ćemo grafički povezanost između brzine automobile i količine potrošenog goriva, a zatim opisati oblik povezanosti.

Slika 7



Skater dijagram prikazan je na slici 3. Jasno se vidi da povezanost nije linearna. Do vrijednosti brzine x=60 km/h potrošnja goriva opada, ali pri brzinama većim od ove, potrošnja goriva raste. Pri tome je tempo opadanja potrošnje goriva veći od tempa porasta potrošnje goriva. Obzirom da povezanost nije linearna, nema smisla govoriti o pozitivnoj ili negativnoj povezanosti. Snagu povezanosti ne možemo izračunati preko Pearsonovog koeficijenta korelacije jer nije zadovoljen uvjet linearne povezanosti između varijabli. Međutim, na osnovu oblika krivulje, mogli bi opravdano pretpostaviti da je povezanost snažna (postupci izračunavanja koeficijenata povezanosti u slučajevima kada nije zadovoljen uvjet linearnosti, bit će objašnjeni u sljedećem poglavlju).

PRIMJER 11. 2

Da li su verbalne sposobnosti povezane sa općom informiranošću? U cilju ispitivanja ove povezanosti grupa od 30 studenata psihologije rješavala je test verbalnih sposobnosti (X) i test opće informiranosti (Y). Rezultati su prikazani u tabeli ispod.

1. *Koristeći kovarijancu odrediti povezanost između inteligencije i opće informiranosti.*

ISPITANIK	X	Y	$X–M_X$	$Y–M_Y$	$(X–M_X)(Y–M_Y)$
1	38,5	18	-0,13	2,13	-0,284
2	39,5	13	0,87	-2,87	-2,4969
3	47,5	15	8,87	-0,87	-7,7169
4	50,5	17	11,87	1,13	13,4131
5	38	14	-0,63	-1,87	1,1781
6	48	22	9,37	6,13	57,4381
7	43	16	4,37	0,13	0,5681
8	37	19	-1,63	3,13	-5,1019
9	45,5	24	6,87	8,13	55,8531
10	35	15	-3,63	-0,87	3,1581
11	25	11	-13,63	-4,87	66,3781
12	41	12	2,37	-3,87	-9,1719
13	28,5	8	-10,13	-7,87	79,7231
14	41,5	13	2,87	-2,87	-8,2369
15	39,5	18	0,87	2,13	1,8531
16	40,5	19	1,87	3,13	5,8531
17	34	19	-4,63	3,13	-14,4919
18	43,5	13	4,87	-2,87	-13,9769
19	29	14	-9,63	-1,87	18,0081
20	27	5	-11,63	-10,87	126,4181
21	35,5	20	-3,13	4,13	-12,9269
22	40	16	1,37	0,13	0,1781
23	29	20	-9,63	4,13	-39,7719
24	39	19	0,37	3,13	1,1581
25	40	21	1,37	5,13	7,0281
26	40	14	1,37	-1,87	-2,5619
27	39,5	12	0,87	-3,87	-3,3669
28	31	8	-7,63	-7,87	60,0481
29	49	25	10,37	9,13	94,6781
30	44	16	5,37	0,13	0,6981

$\Sigma(X–M_X)(Y–M_Y)=$ 473,5259

Deskriptivne vrijednosti iznose:

$M_X = 38{,}63$; $s_X = 6{,}63$

$M_Y = 15{,}87$; $s_Y = 4{,}66$

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{s_x \times s_y} = \frac{16{,}33}{6{,}63 \times 4{,}66} = \frac{16{,}33}{30{,}90} = 0{,}528$$

Pearsonov koeficijent korelacije iznosi:

$$r = 0{,}528.$$

U narednom koraku trebamo odrediti da li je utvrđeni koeficijent korelacije statistički značajan.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = 0{,}528\frac{\sqrt{30-2}}{\sqrt{1-0{,}528^2}} = 0{,}528\frac{5{,}29}{0{,}849} = 3{,}29$$

Granična vrijednost t za df=28 iznosi t_{gr}= 2,048. Dakle, možemo zaključiti da je izračunati koeficijent korelacije statistički značajan na nivou p=0,05. tj:

$$r = 0{,}528,\ p < 0{,}05$$

2. Odrediti pravac regresije

Pravac regresije definiran je jednačinom:

$$Y' = a + bX$$

Za izračunavanje b i a koristit ćemo sljedeće izraze:

$$a = M_y - b\,M_x \quad \text{i} \quad b = \frac{\mathrm{cov}_{xy}}{s_x^{\,2}}$$

$$b = \frac{\mathrm{cov}_{xy}}{s_x^{\,2}} = \frac{16{,}33}{6{,}63^2} = \frac{16{,}33}{43{,}96} = 0{,}37$$

$a = M_y - bM_x = 15{,}87 - 0{,}37 \text{ x } 38{,}63 = 1{,}58$

Pravac regresije određen je sljedećom jednačinom:

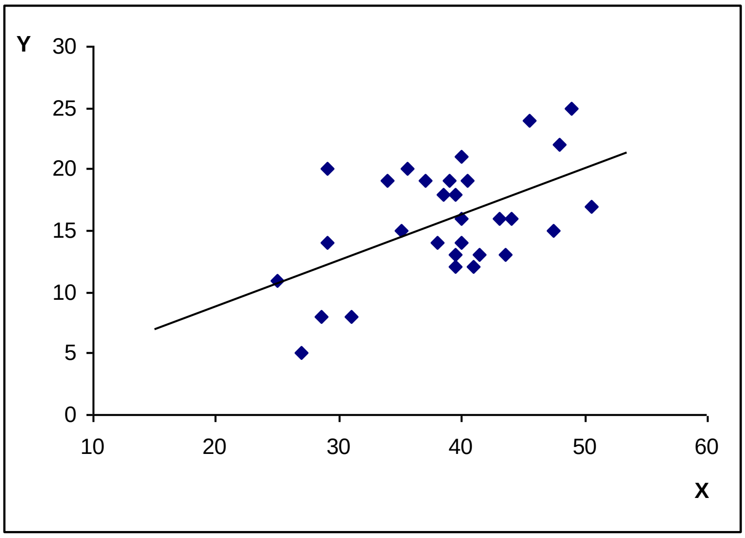
$$Y' = 1{,}58 + 0{,}37X$$

Za crtanje pravca potrebne su dvije tačke. Odredit ćemo ih preko jednačine pravca regresije.
Za X = 35, $Y' = 1{,}58 + 0{,}37 \times 35 = 14{,}53$
Za X = 45, $Y' = 1{,}58 + 0{,}37 \times 45 = 18{,}23$

Koristeći jednačinu pravca regresije odredili smo dvije tačke, A i B:

A(35;14,53), i B (45;18,23)

Skater-dijagran i pravac regresije prikazani su ispod.



3. *Odrediti koeficijent determinacije*

Za određivanje koeficijenta determinacije potrebno je kvadrirati koeficijent korelacije:

$$r^2 = 0{,}528^2 = 0{,}2789$$

Zaključit ćemo da se 27,89% proporcije varijance opće informiranosti može objasniti verbalnom inteligencijom.

4. *Odredit ćemo najvjerovatniji rezultat na testu opće informiranosti ispitanika koji na testu verbalne inteligencije postiže rezultat X=42. Odredit ćemo i 95% interval pouzdanosti prognoziranog rezultata.*

$$Y' = r\left(\frac{s_y}{s_x}\right)(X - Mx) + My = 0{,}528\,\frac{4{,}66}{6{,}63}(42 - 38{,}63) + 15{,}87 = 17{,}12$$

Za određivanje intervala pouzdanosti, potrebno je izračunati standardnu pogrešku prognoze:

$$s_{y,\,x} = s_y\sqrt{(1 - r^2)} = 4{,}66\sqrt{1 - 0{,}528^2} = 4{,}66 \times 0{,}85 = 3{,}96$$

Interval pouzdanosti iznosi:

$$IP = Y' \pm s_{y,x}\; t_{\alpha/2} = 17{,}12 \pm 3{,}96 \text{ x } 2{,}048 = 17{,}12 \pm 8{,}11$$

9,00 – 25,23

Uz 95% sigurnost možemo tvrditi da se prognozirani rezultat nalazi u intervalu od 9 do 25,23.

PRIMJER 11.3

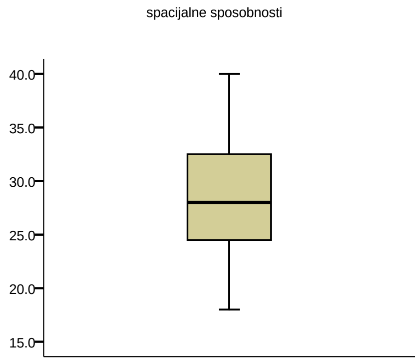
Psihologa je interesovalo da li učenici sa višim spacijalnim sposobnostima (npr. sposobnost mentalne rotacije objekata) prave manje pogrešaka na testu matematičkih sposobnosti? Grupa od 30 učenika rješavala je test spacijalnih sposobnosti (X) i utvrđen je broj pogrešaka na test matematičkih sposobnosti (Y). Rezultati su prikazani ispod.

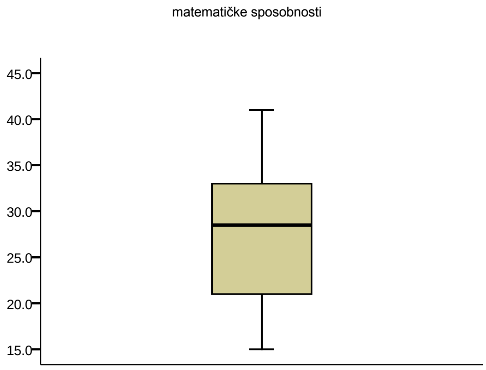
ISPITANIK	X	Y
1	29	15
2	34,5	15
3	30	19
4	35	20
5	30,5	20
6	37	20,5
7	31	21
8	37	21
9	37,5	22
10	39	22
11	19	23
12	27	23
13	40	25
14	22	27
15	22	28

ISPITANIK	X	Y
16	26	29
17	23	29
18	24,5	29
19	25,5	30
20	20	30
21	28	31
22	25	32
23	28	33
24	28,5	33,5
25	32,5	37,5
26	24,5	39
27	18	40
28	27	40,5
29	21	41
30	28	41

1. *Ispitati preduvjet za korištenje Perasonovog koeficijenta korelacije*

- Za mjerenje spacijalnih i matematičkih sposobnosti korištene su intervalne skale mjerenja.
- Oblik distribucije rezultata obje varijble ispitat ćemo grafičkom i numeričkom metodom, preko box-plot prikaza i skjunisa.





Box-plot prikaz ukazuje da se rezultati varijable „spacijalne sposobnosti“ vjerovatno simetrično distribuiraju, a da je distribucija rezultata varijable „matematičke sposobnosti“ negativno asimetrična. Kako bi bili sigurni da li distribucije rezultata odstupaju statistički značajno od simetričnosti, izračunat ćemo skjunis i odrediti standardnu pogršku skjunisa. Ispod su navedene izračunate vrijednosti.

Varijabla	skjunis	st.pog. skjunisa
Spacijalne sposobnosti	,287	,427
Matematičke	,250	,427

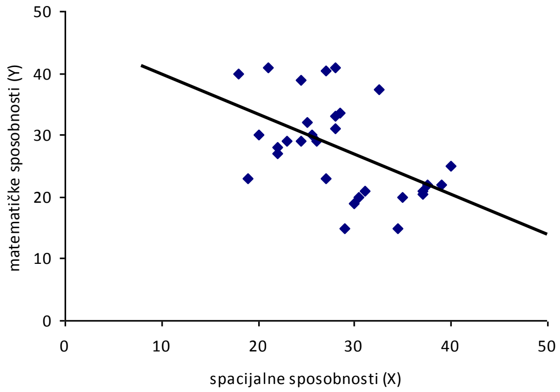
Možemo zaključiti da distribucija rezultata ne odstupa statistički značajno od simetričnosti.

Normalnost distribucije možemo ispitati koristeći Kolmogorov-Smirnov Z test. Izračunate vrijednosti KS-Z i p prikazani su ispod.

Varijabla	KS-Z	p
Spacijalne sposobnosti	,492	,969
Matematičke	,737	,649

Na osnovu dobivenih vrijednosti možemo tvrditi da distribucije rezultata na varijablama ne odstupaju statistički značajno od normalne distribucije.

- Linearnost odnosa između varijabli utvrdit ćemo grafičkim putem, preko skater-dijagrama i pravca regresije.



Na osnovu skater dijagrama i pravca regresije možemo tvrditi da su varijable u linearnom odnosu.

2. *Izračunati Pearsonov koeficijent korelacije i testirati statističku značajnost*

Pearsonov koeficijent korelacije izračunat ćemo preko izraza za sirove podatke:

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left\{\left[\sum X^2 - \left(\sum X\right)^2\right] \times \left[\sum Y^2 - \left(\sum Y\right)^2\right]\right\}}}$$

ISPITANIK	X	Y	X²	Y²	XY
1	29	15	841	225	435
2	34,5	15	1190,25	225	517,5
3	30	19	900	361	570
4	35	20	1225	400	700
5	30,5	20	930,25	400	610
6	37	20,5	1369	420,25	758,5
7	31	21	961	441	651
8	37	21	1369	441	777
9	37,5	22	1406,25	484	825
10	39	22	1521	484	858
11	19	23	361	529	437
12	27	23	729	529	621
13	40	25	1600	625	1000
14	22	27	484	729	594
15	22	28	484	784	616
16	26	29	676	841	754
17	23	29	529	841	667
18	24,5	29	600,25	841	710,5
19	25,5	30	650,25	900	765
20	20	30	400	900	600
21	28	31	784	961	868
22	25	32	625	1024	800
23	28	33	784	1089	924
24	28,5	33,5	812,25	1122,25	954,75
25	32,5	37,5	1056,25	1406,25	1218,75
26	24,5	39	600,25	1521	955,5
27	18	40	324	1600	720
28	27	40,5	729	1640,25	1093,5
29	21	41	441	1681	861
30	28	41	784	1681	1148
Σ=	850	837	25166	25126	23010

$(\Sigma X)^2 = 722500$

$(\Sigma Y)^2 = 700569$

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left\{\left[\sum X^2 - \left(\sum X\right)^2\right] \times \left[\sum Y^2 - \left(\sum Y\right)^2\right]\right\}}} = \frac{30 \times 23010 - 850 \times 837}{\sqrt{\left[\left(25166 - 722500\right) \times \left(25126 - 700569\right)\right]}}$$

$r_{xy} = -0,509$

Statističku značajnost dobivenog Pearsonovog koeficijenta korelacije ispitat ćemo preko t-testa:

Uz df=28, za p=0,01 tgr=2,763. Prema tome utvrđeni koeficijent korelacije statistički je značajan na nivou od 0,01.

3. *Odrediti koeficijent determinacije*

$$r^2 = -0{,}509^2 = 0{,}259$$

Možemo zaključiti da je 25,9% varijabiliteta matematičkih sposobnosti objašnjeno varijabilitetom spacijalnih sposobnosti.

4. *Odrediti pravac regresije za predikciju rezultata na testu matematičkih sposobnosti na osnovu rezultata na testu spacijalnih sposobnosti*

Pravac regresije definiran je jednačinom:

$$Y' = a + bX$$

Za izračunavanje b i a koristit ćemo sljedeće izraze:

$$a = M_y - bM_x \qquad b = \frac{cov_{xy}}{s_x^{\,2}}$$

Najprije ćemo izračunati cov_{xy} i $s_x^{\,2}$.

$cov_{xy} = -24{,}31$; $s_x^{\,2} = 37{,}33$

$$b = \frac{cov_{xy}}{s_x^{\,2}} = \frac{-24{,}31}{37{,}33} = -0{,}651$$

($M_x = 28{,}33$; $M_y = 27{,}9$; $s_x = 6{,}11$; $s_y = 7{,}82$)

$a = M_y - bM_x = 27{,}9 - (-0{,}651) \times 28{,}33 = 46{,}35$

Pravac regresije određen je sljedećom jednačinom:

$$Y' = 46{,}35 - 0{,}651X$$

5. *Koji je najvjerovatniji rezultat na testu matematičkih sposobnosti ispitanika koji na testu spacijalnih sposobnosti postiže rezultat 20? Odrediti 95% interval pouzdanosti.*

Za prognozu rezultata koristit ćemo jednačinu pravca regresije.

$$Y' = 46{,}35 - 0{,}651X = 46{,}35 - 0{,}651 \times 20 = 33{,}33$$

Najvjerovatniji rezultat na testu matematičkih sposobnosti ispitanika koji na testu spacijalnih sposobnosti postiže rezultat 20 iznosi $Y' = 33{,}33$.

Za određivanje intervala pouzdanosti, potrebno je izračunati standardnu pogrešku prognoze:

$$s_{y,x} = s_y\sqrt{(1-r^2)} = 7{,}82\sqrt{1-0{,}509^2} = 7{,}82 \times 0{,}86 = 6{,}73$$

Interval pouzdanosti iznosi:

$$IP = Y' \pm s_{y,x}\ t_{\alpha/2} = 33{,}33 \pm 6{,}73 \times 2{,}048 = 33{,}33 \pm 13{,}78, \text{ tj.}$$

$$19{,}55 - 47{,}11$$

Uz 95% sigurnost možemo tvrditi da se prognozirani rezultata nalazi u intervalu od 19,55 do 47,11. Interval je širok, ali ne koliko i 95% interval pouzdanosti od 11,88 do 43,91, kojeg bismo imali na osnovu dobivenih rezultata Y (tj. $M_y \pm 2{,}048 \times s_y$,), odnosno kada prognozu ne bi temeljili na X.

6. *Odrediti pravac regresije za predikciju rezultata na testu spacijalnih sposobnosti na osnovu rezultata na testu matematičkih sposobnosti*

U slučaju da prognoziramo rezultat na testu spacijalnih sposobnosti na osnovu rezultata na testu matematičkih sposobnosti, ponovit ćemo postupak opisan u 10.3.4 s tim da će nam X varijabla biti matematičke sposobnosti, a Y varijabla spacijalne sposobnosti. Na kraju, dobili bi da je pravac regresije definiran sljedećim izrazom:

$$Y' = 39{,}42 - 0{,}3975X$$

7. *Koji je najvjerovatniji rezultat na testu spacijalnih sposobnosti ispitanika koji na testu matematičkih sposobnosti postiže rezultat 20? Odrediti pogrešku prognoze.*

Za prognozu rezultata koristit ćemo jednačinu pravca regresije.

$$Y' = 39{,}423 - 0{,}397X = 39{,}42 - 0{,}397 \times 20 = 31{,}48$$

Najvjerovatniji rezultat na testu spacijalnih sposobnosti ispitanika koji na testu matematičkih sposobnosti postiže rezultat 20 iznosi Y´ = 31,48.

Pogrešku prognoze izračunat ćemo preko izraza:

$$s_{y,x} = s_y\sqrt{(1-r^2)} = 6{,}11\sqrt{1-0{,}509^2} = 6{,}11 \times 0{,}86 = 5{,}25$$

Obzirom da prognoziramo rezultat u varijabli spacijalne sposobnosti, u gornji izraz uvrštena je vrijednost standardne devijacije rezultata varijable matematičke sposobnosti.

8. *Koliki bi bio najvjerovatniji prognozirani rezultat u testu matematičkih sposobnosti ispitanika koji u testu spacijalnih sposobnosti postiže rezultat 35 u slučajevima*:

   a. Kada bi pogreška prognoze bila nula.

   b. Kada bi pogreška prognoze bila maksimalna.

a. U slučaju kada je pogreška prognoze jednaka nuli, korelacija bi bila maksimalno moguća, tj. r=±1. Tada vrijedi:

$$Y' = r\left(\frac{s_y}{s_x}\right)(X - Mx) + My = \pm 1\left(\frac{7{,}82}{6{,}11}\right)(35 - 28{,}33) + 27{,}9 = 27{,}9 \pm 8{,}54$$

Y´ = 19,36 i Y´ = 36,44

b. Kod maksimalno moguće pogreške prognoze korelacija je jednaka nuli, tj. r=0. Tada vrijedi:

$$Y' = r\left(\frac{s_y}{s_x}\right)(X - Mx) + My = 0\left(\frac{s_y}{s_x}\right)(X - Mx) + My = My$$

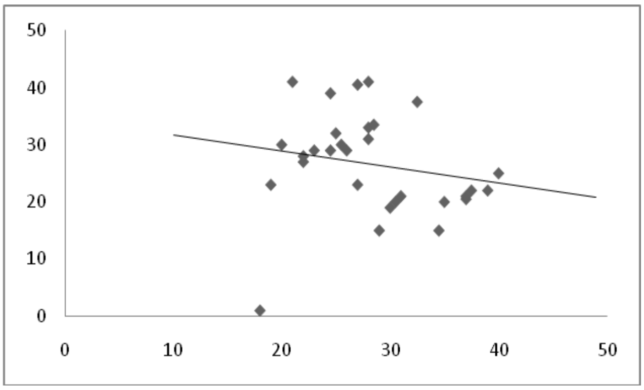
Y´ = My = 27,9

PRIMJER 11.4

Pretpostavimo da je u prethodnom primjeru (10.3) ispitanik pod rednim brojem 27. na testu iz matematike napravio samo jednu pogrešku (y=1). Izračunajte koeficijent korelacije.

Koeficijent korelacije iznosi r=-0,191, što je znatno manje u odnosu na vrijednost izračunatu u prethodnom primjeru (r=-0,509). Vidimo kako jedan podatak, koji značajno odstupa od prosjeka, značajno mijenja vrijednost koeficijenta korelacije. **Ekstremna vrijednost** je podatak koji u značajnoj mjeri odstupa od drugih. Koeficijent korelacije osjetljiv je na ekstremne vrijednosti i stoga njihova pojava može značajno izmjeniti realnu sliku povezanosti između varijabli.

Na skater dijagramu (slika ispod) možemo uočiti u kojoj mjeri ovaj podatak odstupa od ostalih.



Za y=1 kažemo da predstavlja ekstremnu vrijednost. Prisjetimo se deskriptivnih vrijednosti iz primjera 10.3:

$$M_Y = 15,87;\ s_Y = 4,66,$$

na osnovu kojih možemo izračunati standardnu vrijednost za y=1, tj:

$$z = -3,19,$$

što nam pokazuje da ova vrijednost značajno odstupa od aritmetičke sredine (za 3,19 SD).

U praksi se koristi nekoliko postupaka za eliminaciju efekata ekstremnih vrijednosti (npr. ekstremna vrijednost se isključi ili zamijeni aritmetičkom sredinom).

PRIMJER 11.5

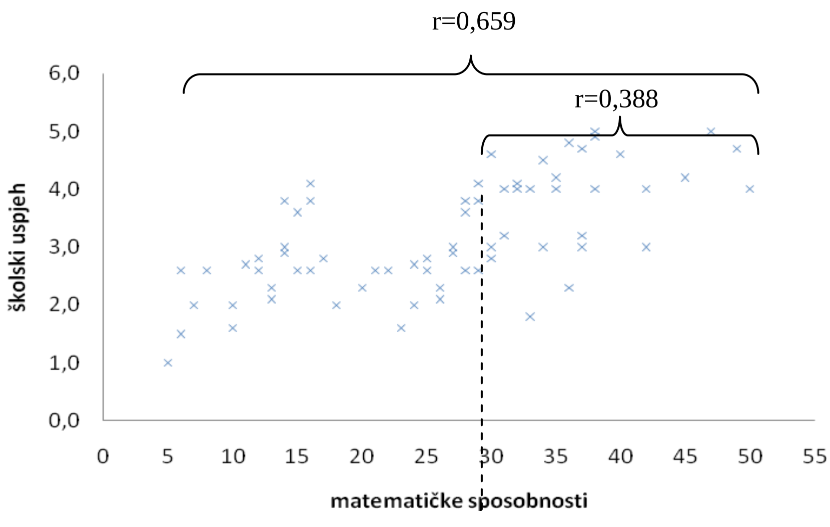
Pretpostavimo da su na grupi učenika prvog razreda nekoliko srednjih škola prikupljeni podaci o školskom uspjehu na kraju polugodišta i da je primjenjen test matematičkih sposobnosti. Stupanj povezanost između matematičkih sposobnosti i školskog uspjeha utvrđen je preko Pearsonovog koeficijenta korelacije koji je iznosio r=0,659.

Nadalje, pretpostavimo da su iz grupe podataka izdvojeni oni koji pripadaju učenicima matematičke gimnazije. Ponovo je izračunat Pearsonov koeficijent korelacije. Ovaj put iznosio je r=0,388. Kako možemo objasniti razliku u dobivenim koeficijentima korelacije? Zašto je u drugom slučaju dobiven gotovo dva puta manji koeficijent korelacije?

Za odgovor na postavljena pitanja trebamo znati da su učenici matematičke gimnazije selekcionirana grupa. Naime, jedan od kriterija za pohađanje matematičke gimnazije su razvijene matematičke sposobnosti. Tako je kriterij za upis u ovu školu bio rezultat jednak ili veći medijani na testu matematičkih sposobnosti (pretpostavimo da u ovom hipotetskom primjeru iznosi C=28). Stoga u grupi učenika iz matematičke gimnazije uopće nema učenika ispodprosječnih matematičkih sposobnosti. Prema tome, opseg rezultata na testu matematičkih sposobnosti učenika matematičke gimnazije manji je od opsega rezultata svih učenika koji su učestvovali u istraživanju. Kao posljedica **restrikcije opsega rezultata** dobivena je niža vrijednost koeficjenta korealcije na grupi učenika iz matematičke gimnazije.

Na slici 10.5 vidimo kako je povezanost između dvije varijable veća za cjelokupnu grupu podataka u odnosu na podatke učenika matematičke gimnazije. Isprekidanom linijom označena je donja granica rezultata na testu matematičkih sposobnosti učenika matematičke gimnazije (X=28). Ako prekrijemo tačke koje se nalaze ispod X=28, vidjet ćemo da je grupiranje oko zamišljenog pravca manje u odnosu na grupiranje tačkaka svih učenika.

Najčešće, restrikcija opsega uzrokuje smanjenje koeficijenta korelacije, ali je moguće i da dovede do njegovog povećanja.



PRIMJER 11.6

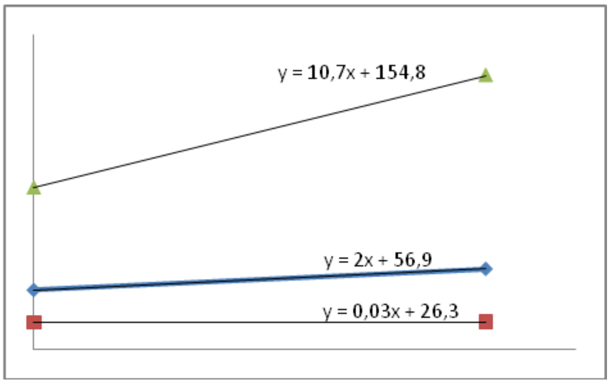
Jedna od mjera za povećanje profita koju kompanija M planira poduzeti je povećati ulaganja u reklamiranje svojih proizvoda. Kako bi ispitala isplativost ovog poslovnog poteza, prikupljeni su podaci o ulaganju u reklame i ostvarenoj dobiti 20 kompanija koje se bave prodajom različitih proizvoda. Određene su regresijske jednačine za proizvode A, B i C. Podaci su prikazani u tabeli ispod.

Proizvod	Regresijska jednačina
A	Y´ = 56,9 + 2X
B	Y´ = 26,3 + 0,03X
C	Y´ = 154,8 + 10,7X

Šta možemo zaključiti na osnovu regresijskih jednačina? Da li reklamiranje proizvoda doprinosi većoj prodaji, a time i većoj dobiti? Za koji proizvod se najviše isplati ulagati u njegovo reklamiranje?

Već na osnovu nagiba pravaca (b vrijednosti) možemo zaključiti da je dobit najveća od reklamiranja proizvoda C. Najmanji nagib pravca (b) je za proizvod B (b=0,03), a najveći za prozivod C (b=10,7). Stopa rasta dobiti u zavisnosti od ulaganja u reklame veća je za proizvod A (b=2) nego za prozvod B, ali znatno manja nego za proizvod C. Ako kompanija ne ulaže niti jednu KM u reklame (tj. ako je X=0), najveća dobit je za proizvod C (Y´=154,8), a najmanja za proizvod B (Y´=26,3).

Pravci regresije prikazani na slici 10.6 ilustriraju povezanost ostvarene dobiti i ulaganja u reklamiranje tri proizvoda.



Na osnovu navedenog, možemo zaključiti da se najveću dobit može očekivati ako se ulaže u reklamiranje proizvoda C.

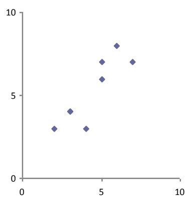
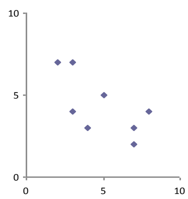
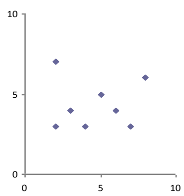
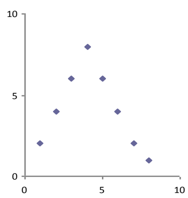
ZADACI

1. Ispod su data četiri skupa podataka (A, B, C i D).

A		B		C		D	
x	y	x	y	x	y	x	y
2	3	9	3	4	3	1	2
5	6	5	2	5	5	3	4
3	4	3	4	3	4	5	6
7	7	7	7	7	3	7	4
6	8	6	9	2	7	9	2

   a. Na osnovu podataka odredite, bez računanja i crtanja skater dijagrama, smjer povezanosti između varijabli x i y.
   b. Nacrtajte skater-dijagram i odredite kakvog smjera su povezanosti između varijabli.
   c. Izračunajte Pearsonov koeficijent korelacije za svaki skup podataka.

2. Ispod su prikazani skater-dijagrami za četiri skupa podataka (A, B, C i D).

   a. Na osnovu skater-dijagrama odredite smjer i procjenite stupanj povezanosti.
   b. Za svaki skater-dijagram procijenite položaj pravca regresije.
   c. Na osnovu skater dijagrama odredite vrijednosti varijabli x i y i izračunajte Pearsonov koeficijent korelacije.

3. Bez računanja odredite vrijednost Pearsonovog koeficijenta korelacije.

A		B		C	
x	y	x	y	x	y
2	3	15	10	2	6
5	6	5	0	3	5
3	4	8	3	4	4
7	8	10	5	5	3
6	7	6	1	6	2

4. Ispod su navedene z vrijednosti. Koliko iznosi r?

A		B	
x	y	x	y
1,55	1,55	-1,25	1,25
-0,95	-0,95	-0,625	0,625
-0,2	-0,2	0	0
0,3	0,3	0,625	-0,625
-0,7	-0,7	1,25	-1,25

5. Za podatke ispod izračunajte Pearsonov koeficijent korelacije

A		B		C		D	
x	y	x	y	x	y	x	y
2	3	4	3	4	1	3	2
5	6	5	5	5	2	5	4
3	4	3	4	6	3	4	6
7	7	7	3	7	4	3	8
4	3	2	7	2	5	3	6
5	7	8	4	8	6	6	4
3	4	7	2	3	7	4	2
6	8	3	7	2	8	7	1
4	3	2	3	4	5	3	2
5	5	5	6	5	3	5	5
3	4	3	4	3	7	4	3
7	3	7	7	7	4	3	7
2	7	4	3	2	5	7	4
8	4	5	7	8	3	4	5
7	2	3	4	7	6	2	3
3	7	6	8	3	8	7	6

6. Na grupi od 29 učenika primjenjena je skala samopoštovanja. Za svakog učenika izračunat je prosjek ocjena iz svih predmeta na kraju polugodišta. Ispod su dati rezultati mjerenja samopoštovanja (X) i prosječni školski uspjeh (Y) učenika.

	X	Y		X	Y
1	31	4	*16*	33	4
2	25	2,8	*17*	28	3,6
3	20	2,3	*18*	29	3,8
4	21	2,6	*19*	23	1,6
5	22	2,6	*20*	29	2,6
6	29	4,1	*21*	30	2,8
7	24	2	*22*	18	2
8	32	4,1	*23*	30	4,6
9	25	2,6	*24*	31	3,2
10	30	3	*25*	17	2,8
11	26	2,1	*26*	32	4
12	26	2,3	*27*	27	2,9
13	24	2,7	*28*	32	4
14	27	3	*29*	33	1,8
15	28	3,8			

   a. Povezanost grafički predstavite u skater-dijagramu.
   b. Odredite jednačinu pravca regresije i ucrtajte pravac.
   c. Izračunajte Pearsonov koeficijent korelacije između samopoštovanja i školskog uspjeha.

7. Ako je korelacija između tjelesne težine (mjerene u kilogramima) i visine (mjerene u metrima) r=0,60, koliko će iznositi korelacija između težine (mjerene u kN) i visine (mjerene u inčima)?

8. Na grupi od 20 učenika prikupljeni su podaci o broju sati provedenih u pripremi za test znanja iz fizike (X) i rezultatu kojeg su postigli na testu (Y).

	X	Y
1	5	35
2	14	58
3	10	60
4	9	65
5	4	41
6	7	53
7	12	60
8	15	79
9	4	35
10	9	50
11	8	59
12	11	64
13	5	45
14	6	52
15	10	65
16	11	69
17	16	69
18	13	72
19	15	73
20	12	50

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Na osnovu skater-dijagrama odredite smjer povezanosti. Procijenite stupanj povezanosti.
c. Izračunajte Pearsonov koeficijent korelacije. Odredite statističku značajnost.
d. Odredite jednačinu pravca regresije. Kakav je smisao a i b u jednačini pravca regresije?
e. Ako učenik provede 15 sati u učenju, koji je najvjerovatniji rezultat na testu iz fizike?
f. Ako učenik provede 10 sati u učenju, koji je najvjerovatniji rezultat na testu iz fizike?
g. Koliko sati provede u učenju učenik koji na testu iz fizike postiže 25 bodova?

9. Nastavnik je ispitivao povezanost između vremena rješavanja testa (X) i broja tačnih odgovora koje učenik postigne na testu (Y). Dobiveni podaci prikazani su ispod.

	X	Y
1	30	75
2	45	100
3	60	54
4	36	81
5	39	94
6	54	65
7	56	79
8	46	79
9	37	95
10	51	82
11	62	54
12	47	82

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Odredite jednačinu pravca regresije.
c. Izračunajte Pearsonov koeficijent korelacije. Odredite statističku značajnost dobivenog koeficijenta korelacije.

10. Da li je depresivnost povezana sa boljim pamćenjem uznemirujućih događaja, slika, scena? U istraživanju je na grupi od 20 ispitanika izmjerena depresivnost (X) i broj zapamćenih uznemirujućih sadržaja iz filma (Y). Podaci su prikazani ispod.

	X	Y
1	11	8
2	4	4
3	19	9
4	8	10
5	9	6
6	9	11
7	10	7
8	2	4
9	13	9
10	15	6
11	7	7
12	15	9
13	16	5
14	4	6
15	17	8
16	18	11
17	14	12
18	20	12
19	15	9
20	18	7

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Odredite jednačinu pravca regresije.
c. Izračunajte Pearsonov koeficijent korelacije. Odredite statističku značajnost dobivenog koeficijenta korelacije.

11. U istraživanju djetetove percepcije roditeljskog ponašanja primjenjen je instrument koji između ostalog mjeri psihološku kontrolu. Ispod su prikazane procjene koje su djeca dala o ponašanju svojih očeva (X) i majki (Y) na subskali psihološke kontrole.

	x	y
1	1,3	0,7
2	2,4	2,1
3	0,8	1,5
4	2,1	2
5	1,8	1
6	0,9	0,9
7	1,8	1,5
8	2,8	2,1
9	1,1	1,8
10	1,3	0,9
11	0,7	1
12	1,7	1,8

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Odredite jednačinu pravca regresije.
c. Izračunajte Pearsonov koeficijent korelacije. Odredite statističku značajnost dobivenog koeficijenta korelacije.

12. Da li je dob pacijenta povezana sa vremenom oporavka nakon operacije? U istraživanju su prikupljeni podaci o dobi pacijenta (X) i vremena provedenog u bolnici tokom postoperativnog perioda (Y).

	x	y
1	27	13
2	36	13
3	30	7
4	27	15
5	24	7
6	22	10
7	20	9
8	27	5
9	29	11
10	32	14
11	25	8
12	37	6

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Odredite jednačinu pravca regresije.
c. Izračunajte Pearsonov koeficijent korelacije. Odredite statističku značajnost dobivenog koeficijenta korelacije.

13. Provedeno je istraživanje s ciljem ispitivanja povezanosti između količine uzimanja vitamina B-12 (X) i kognitivnih funkcija kod starijih osoba (Y). Ispitanici su vitamin B-12 konzumirali u različitim dozama (1-preporučena dnevna doza; 1,5-jedna i pol preporučena dnevna doza; 2- dva puta veća; 2,5- dva i pol puta veća; 3- tri puta veća; 3,5- tri i pol puta veća; 4- četiri puta veća doza od preporučene dnevne). Dobiveni podaci prikazani su ispod.

	x	y
1	1	79
2	2,5	81
3	2	82
4	4	89
5	3	92
6	3,5	94
7	4	95
8	1	95
9	2	95
10	1,5	110
11	3	110
12	2,5	115

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Izračunajte Pearsonov koeficijent korelacije.

14. Na grupi učenika izmjerena je visina i težina.

	x	y
1	160	75
2	175	65
3	176	67
4	177	75
5	180	72
6	182	74
7	183	79
8	186	85
9	189	80
10	193	86

a. Izračunajte Pearsonov koeficijent korelacije.
b. U skater-dijagramu prikažite povezanost između varijabli x i y. Šta možete zaključiti na osnovu skater dijagrama?
c. Izračunajte Pearsonov koeficijent korelacije nakon što ste analizirali skater dijagram i proveli odgovarajući postupak.

15. Na početku nastave iz statistike studenti su rješavali test matematičkih sposobnosti (X). Na kraju nastave utvrđen je uspjeh iz statistike (Y).

	x	y
1	50	90
2	23	85
3	52	95
4	40	90
5	33	87
6	21	77
7	18	74
8	31	79
9	38	80
10	45	92

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Izračunajte Pearsonov koeficijent korelacije. Odredite statističku značajnost.
c. Odredite jednačinu pravca regresije. Kakav je smisao a i b u jednačini pravca regresije?
d. Koji je najvjerovatniji uspjeh iz statistike studenta koji na testu iz matematičkih sposobnosti postiže rezultat 20?
e. Koji je najvjerovatniji uspjeh iz statistike studenta koji je na testu iz matematičkih sposobnosti bolji od 50% ostalih studenata?

16. Ispravne kočnice na automobilu važne su za sigurnu vožnju. S vremenom snaga kočnica slabi. Kako bi ispitali povezanost između starosti automobila i snage kočnice provedeno je istraživanje u kojem je automobilima različite starosti (X) mjerena distanca zaustavljanja (y) pri brzini od 120 km/h. Starost automobila izražena je u mjesecima a distanca zaustavljanja u metrima. Podaci su prikazani u tabeli ispod.

	x	y
1	12	25,40
2	17	30,30
3	25	39,60
4	31	37,20
5	38	36,50
6	45	35,30
7	56	36,20
8	61	45,10
9	65	44,80
10	74	50,20

a. U skater-dijagramu prikažite povezanost između varijabli x i y.
b. Izračunajte Pearsonov koeficijent korelacije. Odredite statističku značajnost.
c. Odredite jednačinu pravca regresije. Kakav je smisao a i b u jednačini pravca regresije?
d. Koja je najvjerovatnija distanca zaustavljanja novog automobila?
e. Kolika je najvjerovatnija distanca zaustavljanja automobila starog 10 godina?

17. Pretpostavimo da želimo ispitati povezanost između tjelesne težine (X) i tjelesnog selfa (Y). U istraživanju je učestvovalo deset mlađih adolescenata i deset starijih osoba. Tjelesna težina izražena je u kilogramima. Tjelesni self izmeren je skalom tjelesnog selfa. Prikupljeni su podaci od ispitanika muškog spola. Podaci su prikazani ispod.

Mlađi adolescenti			Starije osobe		
	X	Y		X	Y
1	75	2	*1*	61	2,6
2	65	2,3	*2*	70	3
3	67	2,6	*3*	67	3,2
4	75	2,6	*4*	72	3,4
5	72	2,6	*5*	72	3,6
6	74	2,8	*6*	70	3,8
7	79	3,1	*7*	79	3,9
8	85	3,1	*8*	63	4,5
9	80	5	*9*	71	5

a. Izračunajte Pearsonov koeficijent korelacije za sve podatke (bez obzira na dob ispitanika)
b. Izračunajte Pearsonove koeficijente korelacije za mlađe adolescente i starije osobe.
c. Uporedite izračunate koeficijente korelacije. Šta možete zaključiti?

18. Grupa od 14 ispitanika rješavala je dva testa, iz geografije (X) i matematike (Y). U tabeli ispod navedeni su rezultati (bodovi na testu) za svakog učenika u oba testa.

	X	Y
1	30	15
2	35	12
3	37	8
4	28	7
5	33	12
6	37	14
7	39	6
8	40	15
9	42	16
10	44	13
11	50	17
12	31	7
13	18	10
14	15	9

a. Izračunajte Pearsonov koeficijent korelacije.
b. Odredite pravac regresije.
c. Kolika je pogreška prognoze ako na osnovu rezultata u testu geografije prognoziramo neki rezultat u testu matematike?
d. Koji je najvjerovatniji rezultat učenika u testu matematike koji u testu geografije postiže rezultat 50?
e. Koliki bi bio najvjerovatniji prognozirani rezultat u testu znanja iz matematike ispitanika koji u testu znanja iz geografije postiže rezultata 40 u slučajevima:
   - Kada je pogreška prognoze nula?
   - Kada je pogreška prognoze maksimalna za ovu situaciju?

19. Grupa od 100 ispitanika rješavala je dva testa, iz hemije i biologije. Povezanost između ova dva testa iznosi r=0,70. Dobivene su sljedeće deskriptivne statističke vrijednosti:

$$M_h = 50 \quad M_b = 60$$

$$S_h = 10 \quad s_b = 10$$

a. Kolika je pogreška prognoze ako na osnovu rezultata u testu hemije prognoziramo neki rezultat u testu biologije?
b. Koliko treba iznositi koeficijent korelacije da se standardna pogreška prognoze smanji za 50%?
c. Koji je najvjerovatniji rezultat učenika u testu biologije koji u testu hemije postiže rezultat 50?

20. Grupa od 15 stonotenisača pripremala se za državno takmičenje. Tokom priprema bilježeno je vrijeme koje su igrači proveli trenirajući (x) (u satima) sedam dana prije takmičenja. Također, za svakog takmičara registrirana je broj pobjeda (y) na takmičenju. Rezultati (vrijeme provedeno u treningu i broj pobjeda na takmičenju) za svakog takmičara prikazani su u donjoj tabeli.

	x	y
1	40	10
2	20	3
3	10	1
4	15	2
5	18	4
6	35	6
7	27	9
8	16	8
9	4	5
10	33	7
11	35	8
12	15	4
13	5	2
14	10	5
15	13	6

a. Da li možemo zaključiti da je broj pobjeda povezan sa vremenom provedenim u treningu? Obrazložite odgovor.
b. Kolika je pogreška prognoze broja pobjeda na takmičenju?
c. Koliki je najvjerovatniji broj pobjeda takmičara koji prije takmičenja uopšte nije trenirao?
d. Koliki je najvjerovatniji broj pobjeda takmičara koji je prije takmičenja trenirao 11 sati?
e. Koliko bi iznosila standardna pogreška prognoze u slučaju kada je r=0 i r=1?

12. Rješenja

1. OSNOVNI STATISTIČKI KONCEPTI

1.

Varijabla	Tip varijable	Vrijednosti
Pacijent	KATEGORIJALNA/ NOMINALNA	AB, CR, NT, SQ, TW
Spol	KATEGORIJALNA/ NOMINALNA	M, Ž
Dob	KATEGORIJALNA/ NOMINALNA	MLAĐA, SREDNJA, STARIJA
Dijagnoza	KATEGORIJALNA/ NOMINALNA	ANKSIOZNI POREMEĆAJ, FOBIJA, DEPRESIVNOST
Vrsta terapije	KATEGORIJALNA/ NOMINALNA	KBT, GEŠTALT, TAA
Trajanje terapije	KVANTITATIVNA/ RACIO-OMJERNA	3, 4, 9, 10, 15

2.

a. KATEGORIJALNA/ NOMINALNA
b. KONTINUIRANA/ RACIO-OMJERNA
c. KATEGORIJALNA/ NOMINALNA
d. KONTINUIRANA/ RACIO
e. KONTINUIRANA/ RACIO
f. KONTINUIRANA/ RACIO
g. KATEGORIJALNA / NOMINALNA
h. KONTINUIRANA/ INTERVALNA
i. KONTINUIRANA/ RACIO
j. KATEGORIJALNA/ NOMINALNA

3.

a. Broj novorođenčadi - RACIO; spol – NOM.; porođajna težina – RAC.; datum i vrijeme rođenja – RAC.
b. Dužine (u mm) zadnjeg donjeg molara – RACIO; broj pacijenata - RACIO.
c. Brok kontrolisanih vozila – RACIO; vrsta prekršaja – NOM.; registarske tablice – NOM.; brizna kretanja vozla – RACIO; spol vozača – NOM.
d. Visina djece –racio; težina djece – RACIO; stanje pluća – NOM.; stanje abdomena – NOM.; stanje ekstremiteta – NOM.; stanje očiju – NOM.; visina sedimentacije – RACIO; broj eritrocita – RACIO; broj leukocita – RACIO; broj roditelja u pratnji djeteta - RACIO.
e. Broj mjerenja – RACIO; dimenzije obrisa kocke ucrtanih na kartonu - RACIO.
f. Broj listova na biljci – RACIO; broj biljaka - RACIO.
g. Broj mjerenja krvnog pritiska – RACIO; Visina krvnog pritiska – RACIO.
h. Broj razreda; broj odjeljenja; broj učenika - RACIO.
i. Broj opažane djece – RACIO; vremensko trajanje igre - RACIO.
j. Broj atletičara – RACIO; Poredak atletičara na cilju – RANG.

7.

a. INTERVALNA SKALA

a.1.

1 - NIJE POLOŽIO IPIT: Ren

2 – POLOŽIO ISPIT: Ben, Den, Ken, Jen, Wen.

Rezultati izraženi na nominalnoj skali: 1, 2, 2, 2, 2, 2

a.2. Rezultati izraženi na rang skali:

1. Wen, 2. Ken, 3. Ben, 4. Jen, 5. Den, 6. Ren

b. RACIO SKALA

b.1.

1 – NIJE ZAVRŠIO TRKU: Asaffa Powell

2 – ZAVRŠIO TRKU: Carl Lewis, Tyson Gay, Donovan Bailey, Usain Bolt, Leroy Burrell, Maurice Green

Rezultati izraženi na nominalnoj skali: 1, 2, 2, 2, 2, 2, 2

b.2. Rezultati izraženi na rang skali:

1. Usain Bolt, 2. Tyson Gay, 3. Maurice Green**,** 4. Donovan Bailey, 5. Leroy Burrell, 6. Carl Lewis, 7. Asaffa Powell

2. GRAFIČKO I TABELARNO PREDSTAVLJANJE PODATAKA

2.

a. Distribucija frekvencija rezultata grupiranih u 8 razreda.

Razredi	f
67-76	7
77-86	5
87-96	6
97-106	4
107-116	5
117-126	2
127-136	0
137-146	1
total	30

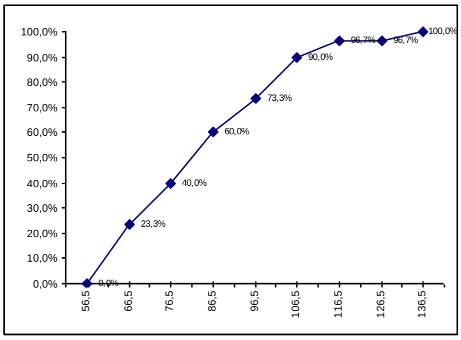
b. 136,5 – 146,5

c. *Stem-and-leaf* prikaz

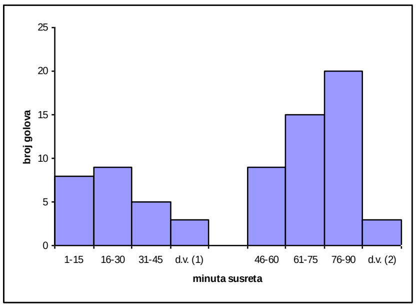
6	8	9					(2)
7	0	0	3	4	6	9	(6)
8	2	3	4	5	9		(5)
9	1	3	3	5	6	7	(6)
10	0	0	4	9			(4)
11	0	3	4	6			(4)
12	1	6					(2)
14	2						(1)

8 | 2 znači 8,2

e.



5.



6.

žene	61	61	61	62	63	63	72	72	73	73	74	76	
	81	83	83	89	92	93	95	95					
muškarci	64	68	72	75	75	79	81	81	81	81	83	84	85
	85	86	86	86	88	89	89	92	92	93	96	98	

10.Histogram

11. Torta ili stupčasti dijagram. Histogram.

12. Histogram ili „stablo i list“ prikaz.

14.Stupčasti dijagram.

15.Histogram ili „stablo i list“ prikaz.

3. MJERE CENTRALNE TENDENCIJE

1. Mod= 15, 20; C=18; M=17,57

2. Mod=2; C=2; M=3

3. C=10; M=10,03

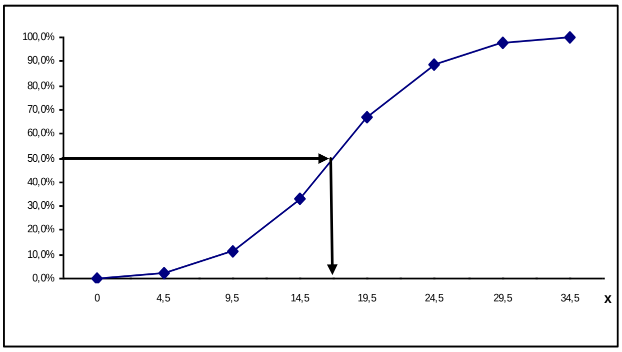
4. Tabela grupiranih rezultata:

Razredi	x	f	fx	cf	rcf%
0-4	2	1	2	1	2,2%
5-9	7	4	28	5	11,1%
10-14	12	10	120	15	33,3%
15-19	17	15	255	30	66,7%
20-24	22	10	220	40	88,9%
25-29	27	4	108	44	97,8%
30-34	32	1	32	45	100,0%
Σ		45	765		

C=17

M=17

3.5.



6. Mod=70; C=93; M=94,07 (vrijednosti C i M utvrđene iz negrupiranih rezultata).

Distribucija podataka nije simetrična. Kao mjeru centralne tendencije možemo uzeti medijanu.

7. C=13,9; M=13,7 (vrijednosti C i M utvrđene iz negrupiranih rezultata).

8. $Mod_ž$=61; $C_ž$=73,5

Mod_m=81 C_m=85

9. C=77; M=125,26

Distribucija rezultata je izrazito asimetrična. Stoga je medijana reprezentativnija mjera centralne tendencije.

10. C=12,4; M=14,07

U nizu podataka jedna vrijednost je ekstremna (55,8). Stoga je medijana reprezentativnija mjera centralne tendencije.

4. MJERE VARIJABILITETA

1. Q_1=15; C=18; Q_3=20; IQR=5; s=3,59
2. Q_1=1; C=2; Q_3=4; IQR=3; s=2,64
4. Q_1=12,63; Q_3=21,38; IQR=8,75; s=6,24
7. Q_1=11,75; Q_3=15,525; IQR=3,775; s=2,25
10. M_1=8,95; s_1= 9,41

M_2= 7,32; s_2=6,09

Vrijednost aritemtičke sredine i standardne devijacije manja je nakon što je isključena ekstremna vrijednost.

11. Na osnovu kvartila i centralne vrijednosti ne možemo odrediti o kojem rezultatu se radi.

Koristeći M i s, odredit ćemo da je u pitanju rezultat 16.

12. A-T1; B-T4;C-T2; D-T3.

13. Interkvarilni raspon:

Q_1 = 32,33

Q_3 = 39,34

IQR = 7,01

v = 89,68

s = 9,47

4.14.

Q_1 = 749,75

Q_3 = 1.149,80

IQR = 400,05

s = 1.387,535

v = 1.925.254

6. NORMALNA RASPODJELA

1.

68,26% središnjih vrijednosti: M ± s = 135,08 ± 1,83: 133,25 - 136,91

95,44% središnjih vrijednosti: M ± 2s = 135,08 ± 2*1,83: 131,42 - 138,74

99,73% središnjih vrijednosti: M ± 3*s =135,08 ± 3*1,83:129,59 - 140,57

3.

a. $z = 0$
b. $z = 1$
c. $z = 2$
d. $z = -1$
e. $z = -3$

4.

a. $X = 125$
b. $X = 73{,}4$
c. $X = 95$
d. $X = 107$
e. $X = 119$
f. $X = 59$

5.

$M = 55{,}8$

7.

a. $z = -3{,}29$
% nižih rez.: 0,06 % viših rez.: 99,94

b. $z = -4{,}00$
% nižih rez.: 0 % viših rez.: 100

c. $z = -1{,}29$
% nižih rez.: 9,85 % viših rez.: 90,15

d. $z = 0{,}86$
% nižih rez.: 80,51 % viših rez.: 19,49

e. $z = 2{,}00$
% nižih rez.: 97,72 % viših rez.: 2,28

f. $z = -2{,}71$
% nižih rez.: 0,34 % viših rez.: 99,66

11.

a. $z_1 = -2{,}29$
b. $z_2 = 0{,}89$
c. $z_3 = 2{,}63$
d. $z_4 = 3{,}76$
e. $z_5 = -2{,}74$
f. $z_6 = -1{,}32$
g. 1335; 252; 6; 0; 1346; 1224
h. 1334; 1094; 252; 111
i. $X = 202$
j. $X \approx 236{,}20$
k. $X \approx 216{,}82$

12.

a. 11 kandidata.
b. $X \approx 133$
c. Kandidat A.

14. Oko 179.423,3 KM.

15. Skjunis: 0,22

Kurtozis: 13,56

8. TESTIRANJE HIPOTEZA

2. $z(19)=1,79$; $p>0,05$ ($t_{gr}=1,96$)

3 $z(24)=2$; a) $p<0,05$; b) $p>0,01$

4. $z(124)=6,74$; $p<0,01$

7. $t(39)=-1,90$ $t_{0,05}=1,684$; $t_{0,01}=2,423$ $p<0,05$; $p>0,01$ (jednosmjerno testiranje)

10. ANOVA

1

Izvor varijabiliteta	*Suma kvadrata (SS)*	*Stepeni slobode (df)*	*Varijanca (MS)*	*F*	*p*
između grupa	41,73	2	20,87	0,094	0,911
unutar grupa	2659,60	12	221,63		
Total	**2701,33**	**14**			

2

Izvor varijabiliteta	*Suma kvadrata (SS)*	*Stepeni slobode (df)*	*Varijanca (MS)*	*F*	*p*
između grupa	101,60	2	50,80	3,603	0,041
unutar grupa	380,70	27	14,10		
Total	**482,30**	**29**			

3 F=0; p=1 Grupe su identične, stoga je variranje između grupa nula!

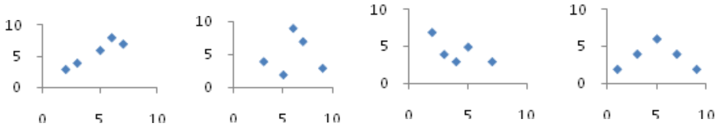
4

Izvor varijabiliteta	*Suma kvadrata (SS)*	*Stepeni slobode (df)*	*Varijanca (MS)*	*F*	*p*
između grupa	96,15	3	32,05	2,836	0,071
unutar grupa	180,80	16	11,30		
Total	**276,95**	**19**			

$F=2,87$, $p>0,05$. Razlike između aritmetičkih sredina nisu statistički značajne!

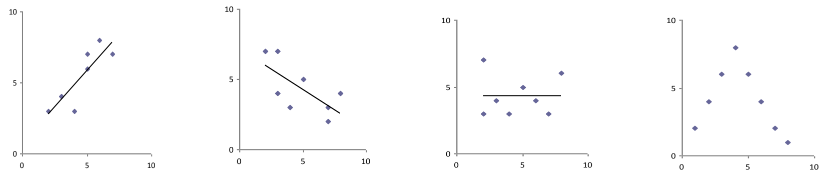
11. KORELACIJA I REGRESIJA

1.



d. r= 0,942; r= 0,077; r= -0,652; r=0

2.



c.

x	y	x	y	x	y	x	y
2	3	4	3	4	3	1	2
5	6	5	5	5	5	2	4
3	4	3	4	6	4	3	6
7	7	7	3	7	3	4	8
4	3	2	7	2	3	5	6
5	7	8	4	8	6	6	4
3	4	7	2	3	4	7	2
6	8	3	7	2	7	8	1

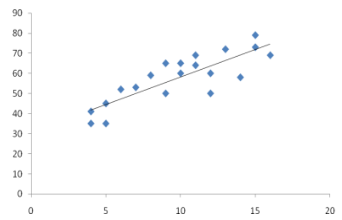
r=0,867; r=-0,680; r=0,005; povezanost nije linearna

3. r=1; r=1; r=-1
4. r=1; r=-1
5. r=-0,020; r=-0,020; r=-0,389;r=-0,132
6.
b. y = 0,1087x + 0,093

c. r=0,585

7. Korelacija će biti identična jer su mjere korištene u drugom primjeru linearne transformacije mjera košištenih u prvom primjeru

8.a.b.



c. r=0,835

d. y = 2,7445x + 30,804

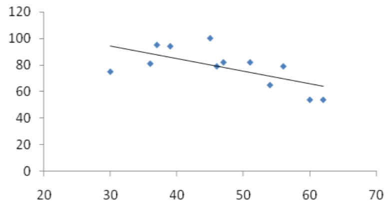
e. y´=72 sata

f. y´=58 sata

g. y = 0,2542x - 4,8646; y´=1,5 sati

9.

a.

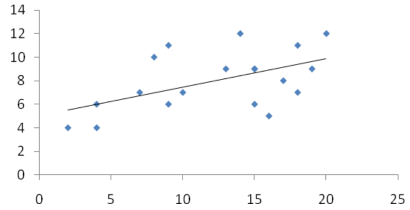


b. y = -0,9497x + 122,89

c. r=-0,64838

10.

a.

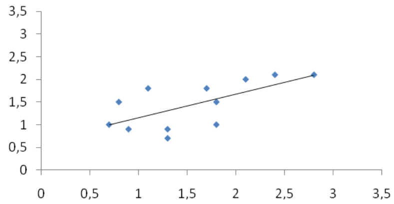


b. y = 0,241x + 5,0444

c. r= 0,528271

11.

a.

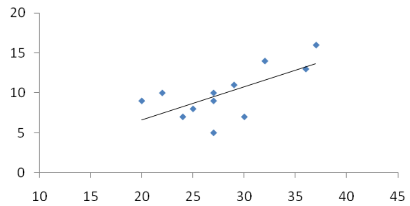


b. y = 0,5223x + 0,6278

c. r=0,661921

12.

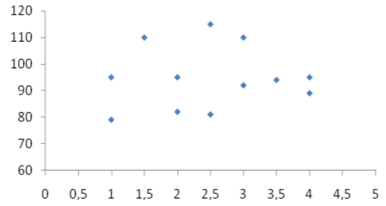
a.



b. y = 0,4116x - 1,6071

c. r=0,670059

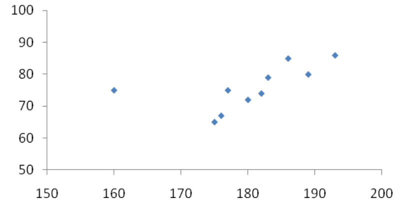
13.
a.



b. r=0,085

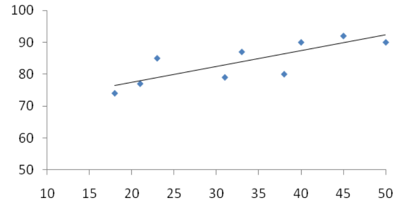
14.

a. r=0,598; koeficijent korelacije trebao bi biti veći
b. Na skater dijagramu jasno je uočljiv ekstremni rezultata.



c. Nakon eliminacije ekstremnog rezultata r=0,895824

15.
a.



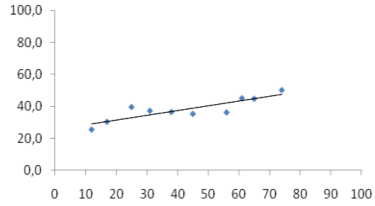
b. r=0,844

c. y = 0,4966x + 67,468

d. y´=77,4

e. z=0, X=M+zs=35,1; y´=84,9

16.



b. r=0,864

c. y = 0,2965x + 25,489

d. x=0; y´=25,489 m

e. x=120; y´=61,1 m

a. r=0,136
b. za mlađe adolescente r=0,526; za starije ispitanike r=0,251

c. U istraživanju su sudjelovali ispitanici različite dobi (heterogene grupe ispitanika)

17.

a.r=0,500

b. y = 0,1931x + 4,8944

c. s_y=3,67; $s_{y,x}$=3,18

d. y´=14,55

e.i. Ako je s_y=0, onda je r=+/- 1. y´=+/-1 x (3,67/9,52) x (40-34,21) + 11,5 = +/-2,23 + 11,5= 9,27 i 13,73

e.ii. Ako je s_y=max, onda je r=0; y´=M_y=11,5

18.
a. $s_{y,x}$=7
b. $s_{y,x}$´=3,5; $s_{y,x}^2=s_y^2(1-r^2)$ □ $r^2=(s_y^2-s_{y,x}^2)/s_y^2$=0,8775: r=0,937

red. prof. dr. Valentin Bucik
Oddelek za psihologijo
Filozofska fakulteta
Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana, Slovenija
Tel. +386 1 4213 595
e-mail: tine.bucik@ff.uni-lj.si

Recenzija priručnika *„Statistika u psihologiji"* autora Nermina Đape i Ratka Đokića, koje izdavaju u oviru Filozofskog fakulteta Univerziteta u Sarajevu

Autori Nermin Đapo i Ratko Đokić, nastavnik i asistent na predmetima sa područja psihološke metodologije i statističnih analiza u studijskom programu psihologije na Odsjeku za psihologiju Filozofskog fakulteta Unvierziteta u Sarajevu, predložili su priručnik za studente (prije svega za studente početne godine prvog ciklusa studija psihologije), koji su nazvali *„Statistika u psihologiji"*. Nakon pregleda teksta dajem sljedeću recenziju.

Autori su priručnik pripremili s osnovnim ciljem da studentima pruži kratka i jasna objašnjenje osnovnih statističkih pojmova, čije poznavanje je preduslov za naprednije nastavne programe iz statistike, kao i naučno-istraživačku praksu. Autori su pri tome koristili pristup interpretacije koncepata kojim se daju definicija i objašnjenja bez komplikovanih matematičkih izraza. Ovakav pristup u edukaciji iz statistike ustaljen je na studijama iz društvenih i humanističkih nauka i kod studenata je jako popularan a ima i motivacijsku ulogu za studij i rad. Statistika je grana matematike i predstavlja skup postupaka koje koristimo za prikupljanje, prezentaciju, analizu i interpretaciju podataka. Desi se, da neki studenti odaberu studij psihologije jer žele izbjeći matematiku... Greška, jer ih na početku studija čekaju metodološke naučne teme, među kojima je i statistika, gdje treba da uče o metodama i tehnikama rada sa podacima, prikupljenim u empirijskim naučnim istraživanjima u psihologiji. A nastavnici i asistenti na studiju su tu da im taj posao što više olakšamo. Na studiju psihologije izučava se primijenjena statistika, odnosno metode za deskripciju i analizu podataka izvedene iz osnovnih matematičkih principa. Nastavnici na studiju psihologije žele te principe studentima, koji imaju dobar a ne opsežan arsenal formalnih, teoretskih matematičkih znanja, približiti i prikazati na što jednostavniji, strukturiraniji i jasniji način. Čini mi se, da su autori u predloženom priručniku napravili baš to. Uspjeli su izgraditi kratak, jasan, dobro strukturiran i studijski vrlo aplikabilan material, koji će biti od velike pomoći studentima tekom studija, pogotovo pri praktičnom radu sa podatcima.

Priručnik se sastoji iz 11 poglavlja koja svojim sadržajem obuhvaćaju desktiptivnu i osnove inferencijalne statistike. Svako poglavlje započinje uvodom u kojem su data osnovna i kratka objašnjenja, koja su kroz primjere u nastavku poglavlja dodatno interpretirana i u kojima je ukazano na njihovu praktičnu primjenu. Na kraju svakog poglavlja nalaze se zadaci koji služe za vježbanje i ponavljanje gradiva određene teme. Struktura priručnika je u tom smislu didaktički veoma dobra. Znači da u svakom poglavlju imamo najprije kratka i jednostavna objašnjenja fenomena (bez komplikacija, što je dobro, jer elaborirana objašnjenja student može i mora naći u statističkim udžbenicima – ovaj priručnik nema ambiciju biti cjelokupan statistički udžbenik, nego više praktično oruđe za razumijevanje i konkretan operativan rad studenta na početku studija psihološke nauke i njezine metodologije), onda više primjera kroz koje autori žele ilustrirati fenomen i njegovu upotrebu, a na kraju bogata serija zadataka, namijenjenih studentima, da preko njih usvoje osnovna znanja o prezentiranom fenomenu.

Autori u Priručniku studenta vode za ruku, polako, sigurno i suvereno kroz sve osnovne („introductory") teme iz psihološke statistike: kroz postupke koje koristimo u opisu podataka (npr. određivanje broj kategorija, centralne vrijednosti, aritmetičke sredine, itd.) a koje su dio takozvane deskriptivne statistike, kao i kroz postupke donošenja zaključaka o populaciji na osnovu podataka dobivenih na uzorku, koji su dio takozvane inferencijalne statistike. Na kraju autori pružaju osnovne informacije o postupcima zaključivanja, kada nas zanimaju razlike između grupa (preko postupaka analize varijance) i kada nas zanima stupanj povezanosti među variablama ili grupama (korelacija i regresija). Vrlo je dragocjeno, da su autori odlučili neko vrijeme u priručniku (a poslijedično u radu sa studentima) posvjetiti osnovnim principima vjerovatnoće, jer se u toj temi sakriva niz važnih činjenica, koje omogućavaju razumjevanje osnovnih statističnih principa, pogotovo inferencijalnih. Dužan sam isto tako pohvaliti pristup autora u objašnjavanju i ilustriranju (preko primjera) očitovanja i izračunavanja *z*-vrijednosti u poglavlju o normalnoj distribuciji; nije baš mnogo knjiga i priručnika sa tako kompletnim i elaboriranim objašnjenjem jednog principa, bez kojeg studenti ne mogu razumjeti cijelog niza zahtijevnijih statističkih principa (kao što je korelacija).

Dodajem nekoliko primjedbi ili sugestija za sljedeće izdanje priručnika, do kojeg, vjerujem, će doći jer znamo, da su priručnici uvijek „živ" organizam, koji raste, širi se, dopunjuje i poboljšava sa svakim slijedećim izdanjem. Najprije, ne bi bilo loše da autori u slijedećem izdanju u poglavlju o testiranju hipoteza uvrste i objašnjenja o veličini efekta, pogotovo jer znamo, da klasičan pristup testiranju hipoteza (pa i zaključivanje na osnovu alfa i beta pogreške) pati od niza metodoloških i tehničkih poteškoća, kao što je veličina uzorka i/ili broj variabli (i njihovih međusobnih uspoređivanja) u analizi. Čini se, da bi tekst mogao biti još jasniji i bilo bi ga jednostavnije slijediti, ako bi autori sve formule i algoritme opisali (imenovali) sa brojevima. U poglavlju o korelaciji i regresiji autori bi mogli više elaborirati uslov homoscedasciteta za korektnu upotrebu Pearsonovog koeficienta korelacije, a kod primjera 11.4 mogli bi studenta utješiti, da postoji mogučnost vršenja korekcije korelacije za restrikciju ranga (što znači, da je problem moguče riješiti), ako imamo pored informacije o variansi restriktiranog uzorka i informaciju o variansi nerestriktiranog uzorka odnosno osnovne populacije.

Bilo bi odlično, ako bi autori (i studenti sa njima) praktičan rad na primjerima (pa i rad sa zadacima) mogli obavljati putem nekog konkretnog računarskog oruđa, recimo statističkog programskog paketa za obradu podataka u društvenim naukama SPSS (ili pak nekog drugog, sličnog), a potpuno smo svjesni činjenice, da dok studentima nisu pružene sve institucionalne mogučnosti jednostavne (i jeftine) upotrebe programskih oruđa u tu svrhu (licencije, pa i računarska oprema), nema smisla upozoravati na taj ili onaj proizvod, nego je pametnije „držati se sa strane" i prepuštati konkretne odluke oko statističnih programskih paketa studentima, a to je točno ono što su autori napravili u primjeru ovog priručnika.

Uvod u statistiku u psihologiji je priručnik namijenjen prvenstveno studentima koji započinju izučavati metodologiju psiholoških istraživanja i mislim, da će priručnik studentima jako olakšati rad i studij. Premda se statistika u nastavnim programima psihologije izučava kao zaseban predmet, statističke metode zapravo su sastavni dio istraživačkog procesa, što su autori u pisanju ovog priručnika nastojali posebno naglasiti i mislim, da su u toj intenciji bili potpuno uspješni. Ovaj priručnik poslužiti će prvo studentima ali bit će koristan i drugima koji se iz različitih razloga interesuju za statistiku u psihologiji ili pak imaju potrebu za osvježavanjem znanja o osnovnim statističkim principima u svom profesionalnom aplikativnom ili istraživačkom radu. Pridružujem se autorima u apelu, da priručnik nema namjeru zamijeniti udžbenik iz statistike. Njegovu svrha treba prepoznati u samom nazivu, dakle, da bude pri ruci studentu koji izučava određene teme iz statistike. Priručniku želim uspješan život.

dr. Valentin Bucik
redovni profesor za psihološku metodologiju

Recenzija rukopisa dr. Đape Nermina „Statistika u psihologiji: Priručnik za studente“

Recenziju napisala: dr. Dženana Husremović

Statistika u društvenim znanostima, pa i u psihologiji, predstavlja jedno od najizazovnijih područja za podučavanje. Nastavnici na društvenim znanostima koji podučavaju predmet Statistika često se susreću sa studentima koji nisu očekivali da će na odabranom studiju provesti značajan dio vremena izučavajući metode istraživanja i obrada podataka koje društvene znanosti i čine znanostima. Studenti društvenih znanosti uglavnom dolaze sa očekivanjima da na studiju više neće morati baratati brojevima i matematičkim konceptima. Njihovo odbijanje da se uhvate u koštac sa metodološkom grupom predmeta nastavnike dovodi u poziciju da osim podučavanja, moraju puno raditi na motiviranju studenata i objašnjavanju važnosti Statistike za samu društvenu znanost. Stoga uspješan nastavnik statistike u društvenim naukama nije onaj koji dobro poznaje primijenjenu statistiku, nego je to onaj koji statističke koncepte zna približiti studentima i motivirati ih da znanstveno razmišljaju, zaključuju i predviđaju. U podučavanju studenata, nastavnici imaju veliki broj knjiga i udžbenika na raspolaganju. Međutim, studentima su ti udžbenici prečesto napredna literatura s obzirom na nivo znanja i vještina matematičke pismenosti sa kojim dođu iz srednje škole. I tako je dugo postojao prazan prostor u kojem je nedostajao priručnik koji bi na jednostavan način objašnjavao najvažnije koncepte iz statistike i omogućio studentima da napreduju prema udžbenicima koji jesu preporučena literatura. Nermin Đapo i Ratko Đokić su sa rukopisom „Statistika u psihologiji: Priručnik za studente“ vrlo uspješno popunio ovaj prostor.

Priručnik se sastoji od dvanaest cjelina. Prvih jedanaest cjelina daju pregled najvažnijih koncepata u deskriptivnoj i inferencijalnoj statistici, dok su u dvanaestoj cjelini ponuđena rješenja za sve zadatke. Svaka cjelina organizirana je na sličan način: prvo je detaljno objašnjen koncept, a zatim se na praktičnim primjerima studentima metodom korak-po-korak demonstrira provođenje određenog postupka. Potom su studentima dati tipski zadaci za vježbanje i utvrđivanje gradiva.

U prvoj cjelini date su definicije ključnih termina u statistici, te objašnjenje varijabli što studentima značajno olakšava savladavanje ostalih elemenata statistike. U drugoj cjelini prikazani su grafički i tabelarni postupci organiziranja i prikazivanja podataka na tako jednostavan način da svaki prosječan student može, nakon čitanja ovog poglavlja, naučiti očitavati i prirpemati grafikone i tabele. Treća i četvrta cjelina se bave objašnjavanjem dva glavna deskriptivna parametra – mjerama centralne tendencije i mjerama varijabiliteta, dok je u petoj cjelini dat pregled glavnih pojmova teorije vjerovatnoće. Upravo čitanje ovog petog poglavlja omogućava studentima da razumiju izazovnost zaključivanja i predviđanja u društvenim znanostima koje su bazirane na teoriji vjerovatnoće. U ovom logičkom slijedu, šesto poglavlje objašanjava studentima što je normalna raspodjela i kako se ova distribucija povezuje za teorijom vjerovatnoće. Sa sedmim poglavljem autor uvodi studente u područje inferencijalne statistike, odnosno u područje statističkog zaključivanja. Kroz praktične primjere studenti jednosatvno uočavaju razlike u rezultatima dobivenim na uzorcima i populaciji, te uviđaju važnost pravilnog uzorkovanja u svrhu generalizacije rezultata na populacije.

Osmo poglavlje otvara područje testiranja hipoteza. Termin hipoteze nije jako poznat studentima na početku njihovog studija jer se u srednjoj školi ne susreću često sa testiranjima određenih pretpostavki. Autor na početku poglavlja vrlo stručno i za studente prihvatljivo objašnjava zašto su hipoteze važne i kako se one definiraju, te u poglavljima devet i deset daje prikaz osnovnih testova za ispitivanje značajnosti razlika artimetičkih sredina (t-testova i analizi varijanci). U posljednjem jedanaestom poglavlju studenti imaju priliku pročitati i naučiti što je to povezanost između dvije varijable, te naučiti kako se testira značajnost korelacije.

Ovaj priručnik predstavlja jedinstveno djelo po pristupu u objašnjavanju statističkih koncepata. Na prvo čitanje je jasno da je napisan od strane autora koja ima višegodišnje iskustvo u podučavanju studenata statistici. Autor piše tako da prati misaoni tok studenta i daje mu upravo onoliko informacija koliko je potrebno da se shvati suština. Autor poštuje sve principe logičkog izlaganja i ide od jednostavnijih ka težim stvarima, a apstraktne koncepte objašnjava kroz razumljive praktične primjere i slikovne prikaze koji omogućavaju studentima da u potpunosti razumiju što se, u stvari krije iza formula. Na kraju svakog poglavlja autor daje zadatke za vježbanje kako bi studenti izgradili znanje i vještinu obrade podataka, te tako postali samostalni i kompetentni za budući rad.

Ovaj priručnik je usmjeren prema studentu i napisan je za studenta i može služiti kao primjer metodički adekvatnog i psihološki motivirajućeg materijala. Autor je uspio pokazati da je kao stručnjak i naučnik vrsni poznavalac statistike, a da je kao nastavnik okrenut prema studentu i sposoban vrlo složene stvari objasniti tako da ga razumiju i oni koji ne vole matematiku.

Korištenje ovog priručnika kod studenata će zasigurno umanjiti anksioznost od pripreme ispita, te razbiti predrasude o statistici kao matematici rezerviranoj samo za „odabrane". Ovaj priručnik će omogućiti svim zainteresiranim studentima da dožive statistiku kao temelj istraživanja, zaključivanja i predviđanja u psihologiji.

Husremović

Dr. Dženana Husremović

Biografija

Nermin Đapo rođen je 14. 1. 1970. godine u Mrkonjić Gradu. Diplomirao je na Filozofskom fakultetu u Sarajevu, Odsjek za psihologiju. Od 1997. godine zaposlen je na Odsjeku za psihologiju, naprije kao asistenta na predmetima Statistika u psihologiji i Metodologija istraživanja u psihologiji, zatim kao viši asistenta na predmetima Statistika u psihologiji i Opća psihologija II, a od 2007. godine kao docenta na predmetima Statistika u psihologiji i Kognitivna psihologija. Na Filozofskom fakultetu u Zagrebu 2001. godine odbranio je magistarsku radnju pod naslovom *Kompozitno pamćenje i kohezija memorijskih tragova*. Doktorsku disertaciju *Interna i eksterna validacija dinamičkog testiranja inteligencije* odbranio je 2006. na Odsjeku za psihologiju Filozofskog fakulteta u Sarajevu. U okviru stručnog i naučnog usavršavanja boravio je na LMU u Minhenu, Njemačka i Institutu za psihologiju u Geteborgu, Švedska. Područje stručnog i naučnog rad su primijenjena statistika, inteligencija i nadarenost. Objavio je više naučnih i stručnih članaka, u autorstvu i koautorstvu, u domaćim i internacionalnim časopisima (*Naša škola, Didaktički putokazi, Psihologijske teme, Group Dynamics: Theory, Research and Practice, Journal of the American Academy of Child and Adolescent Psychiatry, Mankind Quarterly, Personality and Individual Differences, Temas em Psicologia)*. Sudjelovao je kao voditelj ili konsultant na više istraživačkih i aplikativnih projekata. Učestvovao je na domaćim i međunarodnim naučnim i stručnim skupovima i seminarima.

Ratko Đokić rođen je 3. 1. 1979. godine u Sarajevu. 2002. godine diplomirao je na Odsjeku za psihologiju Filozofskog fakulteta u Sarajevu. Od 2001. do danas učestvovao u ili vodio niz projekata socijalnih, tržišnih i medijskih istraživanja te istraživanja javnih politika (za klijente kao što su UNICEF, BBC World Service Trust, Intermedia Washington, Unilever, UNDP, Save the Children UK, Vijeće za štampu BiH...). Od 2007. godine radi na Odsjeku za psihologiju Filozofskog fakulteta u Sarajevu gdje je kao asistent angažovan na predmetima Statistika u psihologiji I i II, Metodologija eksperimentalne psihologije i Metodologija neeksperimentalne psihologije. Trenutno pohađa Doktorski studij psihologije u Zagrebu.