visoka poslovna škola par

stručni studij poslovnoga upravljanja

Poslovna statistika

mr.sc. Bojan Kovačić

POSLOVNA STATISTIKA

interna skripta[1]

[1] Ova interna skripta prvi su put korištena u izvedbi nastave iz istoimena kolegija na stručnom studiju računovodstva i financija RRiF Visoke škole za financijski menadžment u Zagrebu u razdoblju od 01.03.2008. do 01.04.2010. Osim na Visokoj poslovnoj školi P.A.R., danas se jednim dijelom (poglavlja 1. – 4.) koriste i u nastavi iz kolegija *Vjerojatnost i statistika* na stručnom studiju elektrotehnike Tehničkoga veleučilišta u Zagrebu.

SADRŽAJ

PREDGOVOR 5
0. REKLI SU O STATISTICI… 8
1. UVOD U STATISTIKU 8
1.1. *Pojam statistike* 8
1.2. *Opisna i inferencijalna statistika* 8
1.3. *Statistička analiza, statistički skup i obilježja elemenata statističkog skupa* 9
1.3.1. *Skale obilježja* 11
1.3.2. *Vrste obilježja* 13
2. FAZE STATISTIČKOGA ISTRAŽIVANJA 17
2.1. Određivanje cilja i razrada plana istraživanja 17
2.2. Organizirano prikupljanje statističkih podataka 18
2.3. Grupiranje statističkih podataka 20
2.3.1. *Grupiranje kvalitativnih podataka* 20
2.3.2. *Grupiranje kvantitativnih podataka* 21
2.3.3. *Kumulativni nizovi frekvencija* 24
2.3.4. *Dijagram "stablo – list" (stem – leaf)* 28
2.4. Tablični prikaz statističkih podataka 30
2.5. Grafički prikaz statističkih podataka 35
2.5.1. Površinski grafikoni 36
2.5.2. Histogram 42
2.5.3. Linijski grafikoni 48
2.5.4. Ostale vrste grafikona 53
3. OSNOVNE KARAKTERISTIKE NUMERIČKIH NIZOVA 56
3.1. *Srednje vrijednosti numeričkih nizova* 56
3.1.1. Aritmetička sredina 57
3.1.2. Geometrijska sredina 66
3.1.3. Harmonijska sredina 70
3.1.4. Kvantili 74
3.1.5. Mod 81
3.2. *Mjere raspršenja (disperzije)* 82
3.2.1. Raspon varijacije 84

3.2.2. Interkvartil 85
3.2.3. Interpercentilni razmak....86
3.2.4. Varijanca i standardna devijacija 87
3.2.5. Srednje apsolutno odstupanje 92
3.2.6. Koeficijent varijacije 95
3.2.7. Koeficijent kvartilne devijacije....97
3.3. *Standardizirano obilježje*....101
3.4. *Vrste razdioba* (*distribucija*)....104
3.4.1. Mjere asimetrije 105
3.4.2. Mjere zaobljenosti 109
3.5. Programska potpora deskriptivnoj statističkoj analizi....110
4. KORELACIJSKA I REGRESIJSKA ANALIZA 111
4.1. *Korelacijska analiza*....112
4.1.1. Spearmanov koeficijent korelacije ranga 121
4.2. Regresijska analiza 126
4.2.1. Model jednostavne linearne regresije....128
4.2.2. Analiza varijance modela jednostavne linearne regresije 141
4.2.3. Mjere reprezentativnosti modela jednostavne linearne regresije....143
4.2.3.1. Varijanca regresijskoga modela 144
4.2.3.2. Standardna devijacija regresijskoga modela.... 144
4.2.3.3. Koeficijent varijacije regresijskoga modela 145
4.2.3.4 Koeficijent determinacije 145
4.3. Primjeri nelinearnih modela jednostavne regresije 148
4.3.1. Model jednostavne eksponencijalne regresije....148
4.3.2. Model jednostavne logaritamske regresije 156
4.3.3. Model jednostavne potencijske regresije (*POWER*)....160
4.4. Primjer modela višestruke (*multiple*) *regresije: model višestruke linearne regresije*....168
5. INDIVIDUALNI BROJČANI POKAZATELJI RAZVOJA VREMENSKO-GA NIZA 172
5.1. Definicija i vrste vremenskih nizova....174
5.2. *Pokazatelji pojedinačnih apsolutnih promjena*....176
5.3. *Modeli trenda*....178
5.4. *Osnovni brojčani pokazatelji relativnih promjena*....181
5.5. *Individualni indeksi* 183
5.5.1. *Verižni (lančani) indeksi* 184

5.5.2. *Indeksi na stalnoj bazi* (*bazni indeksi*) 188
5.5.3. *Pretvorba indeksa* 193
5.6. *Skupni indeksi* 198
5.6.1. *Skupni indeksi količina* 199
5.6.1.1. *Laspeyresov skupni indeks količina* 199
5.6.1.2. *Paascheov skupni indeks količina* 199
5.6.1.3. *Fisherov skupni indeks količina* 201
5.6.2. *Skupni indeksi cijena* 202
5.6.2.1. *Laspeyresov skupni indeks cijena* 203
5.6.2.2. *Paascheov skupni indeks cijena* 203
5.6.2.3. *Fisherov skupni indeks cijena* 205
5.6.3. *Skupni indeksi vrijednosti* 206
5.6.4. *Neki posebni oblici skupnih indeksa i njihova primjena* 210
5.6.4.1. *Izračunavanje realnih plaća na osnovu skupnih indeksa troškova života* 210
5.6.4.2. *Vrijednost u stalnim cijenama i indeks fizičkoga obujma* 215
6. LITERATURA 225

PREDGOVOR

Jedna od najprimjenjivanijih *matematičkih* disciplina u gotovo svim područjima čovjekova djelovanja danas je nesumnjivo statistika. Bez te su discipline današnja znanost i cjelokupan ljudski život praktički nezamislivi. Prikupljanje podataka, njihova obrada i analiza, te izvođenje zaključaka na temelju tih postupaka provode se u tolikoj mjeri da su razvojem moderne informatičke tehnologije usporedno razvijeni i "specijalizirani" računalni programi namijenjeni potrebama statističke analize i statistike općenito. S druge strane, zbog svojega raširenoga područja primjene, statistika je vrlo često potpuno neopravdano i neutemeljeno bila meta različitih kritika koje su je nastojale obezvrijediti, prikazati njezina područja i metode kao "nešto zanimljivo samo statističarima", a statističare kao neshvaćene asocijalne čudake koji, umjesto druženja s tzv. normalnim ljudima i uživanja u svim ljepotama i čarima ovozemaljskoga života, radije odabiru uživanje u hrpama i tonama podataka, tablicama, grafovima i sličnim statističkim objektima.

Još od samoga početka svojega bavljenja statistikom u studentskim danima nastojao sam tu matematičku disciplinu promatrati sa stajališta da je riječ o disciplini ništa manje vrijednoj i ništa manje važnoj od algebre, analize, geometrije i inih matematičkih disciplina. Kasnije, kao predavač statističkih kolegija na različitim visokim školama, s više ili manje uspjeha nastojao sam približiti statistiku studentima i ukazati im na svu pogrešnost uobičajenih zabluda o statistici. Kao rezultat takvoga pristupa razvile su se sljedeće temeljne postavke koje se provlače kroz ova skripta:

1.) *Statistika je sastavni dio matematike*, unatoč činjenicama da je znatan dio stručnjaka uporno pokušava ubrojiti u sastavnice ekonomske znanosti. Iako se statistika ponajviše primjenjuje u ekonomiji, pa se na ekonomskim fakultetima zbog toga izučava kao posebno područje, temelji statistike i statističkoga načina mišljenja su isključivo matematički.

2.) *Nitko ne može zamijeniti istraživača prigodom interpretacije rezultata statističke analize*. Računala i "specijalizirani" računalni programi mogu vrlo brzo i točno izračunati sve željene statističke pokazatelje, ali ne mogu dati njihovu univerzalnu interpretaciju jer ona ovisi o svakom pojedinom razmatranom slučaju. Statističar mora znati uz koje se temeljne pretpostavke izračunavaju pojedine statistički pokazatelji i provjeriti vrijede li te pretpostavke u slučaju koji razmatra. Ukoliko ne vrijede, dobiveni statistički pokazatelj bit će točno izračunat, ali neće biti *reprezentativan*, tj. neće dovoljno dobro opisivati podatke iz kojih je izračunat.

3.) *Ne postoje univerzalni statistički pokazatelji koji bi dovoljno reprezentativno opisivali bilo koji statistički niz podataka*. Ova zabluda naročito se odnosi na aritmetičku sredinu (popularni "prosjek") koju se, zbog jednostavnosti njezina izračuna, uporno nastoji "progurati" kao univerzalni statistički pokazatelj prikladan za opis bilo kojega statističkoga niza. Pojednostavljivanje ovdje uzrokuje nereprezentativnost dobivenoga statističkoga pokazatelja, što u svojim kritikama često iskorištavaju protivnici statistike. Ispravan pristup ovoj problematici je tipičan pristup problemima matematičkoga modeliranja: klasificirati

statističke probleme u tzv. *klase* dovoljno međusobno "sličnih" problema koji se mogu riješiti uz iste polazne pretpostavke i istu, zadovoljavajuću reprezentativnost dobivenih rezultata. Spomenuta klasifikacija nije nimalo jednostavna i predstavlja problem o kojemu se raspravlja i dan-danas, ali takav je slučaj s apsolutnom većinom matematičkih problema.
Stoga se na samom početku pisanja ovih skripta pojavio sljedeći problem: kako na dovoljno jednostavan i razumljiv način, ali uz uvažavanje svih gore navedenih postavki, napisati tekst koji će istodobno biti i podloga za predavanja i nastavni materijal prema kojemu će studenti moći pripremati ispit iz poslovne statistike? Navedeni je problem razlog što je ovaj tekst nastajao tijekom mojih svakogodišnjih predavanja statistike u gotovo svim godinama 21. stoljeća. Naime, odgovorno tvrdim da bilo koja skripta može biti nastavni materijal prema kojemu će studenti moći pripremati ispit ako i samo ako je skripta bila efektivno korištena u nastavi tijekom najmanje 5 godina. U tom su smislu i svi studenti – slušači statističkih kolegija koje sam predavao koautori ovih skripta jer su mi njihova pitanja, konstruktivne kritike i komentari pomogli uvidjeti na koje dijelove gradiva treba posebno obratiti pozornost i dodatno ih pojasniti jer nisu sami po sebi dovoljno razumljivi.

Iako u nazivu skripta stoji *Poslovna statistika*, ova bi se skripta mogla zvati i *Osnove statistike*. Gradivo koje je u njoj izloženo, naime, većina će statističara proglasiti upravo "osnovama osnovnih osnova" statistike. No, upravo te "osnove osnovnih osnova" danas tvore područje koje se najčešće primjenjuje u različitim poslovnim analizama i, kao takvo, sasvim je primjereno studentima *stručnih* studija. Stoga skripta mogu koristiti svi studenti stručnih studija ekonomije, ali i neekonomskih studija poput fizioterapije, upravnoga prava itd., te svi koji u svojim poslovima primjenjuju ma koji dio izloženoga gradiva.

Konceptualno je skripta podijeljena u pet poglavlja. U prva dva poglavlja obrađeni su osnovni statistički pojmovi definirani s matematičkoga stajališta, te osnovne faze statističkoga istraživanja. Posebna se pozornost posvećuje statističkim tablicama i grafikonima, kao i njihovim interpretacijama. U trećem se poglavlju obrađuju u praksi najčešće korištene srednje vrijednosti i mjere raspršenosti (disperzije) uz poseban naglasak na reprezentativnost svake pojedine srednje vrijednosti. Četvrto je poglavlje posvećeno osnovama korelacijske i regresijske analize, pri čemu se temeljne postavke i ideje uglavnom izlažu na problemima u kojima se razmatraju točno dvije varijable. U petom, posljednjem poglavlju izlažu se osnove analize vremenskih nizova i opisuju neke posebne vrste skupnih indeksa.

Kao dio osnovnoga teksta u svim je poglavljima naveden relativno velik broj detaljno riješenih primjera od kojih je veći dio riješen "klasično" (bez korištenja bilo kojega od "specijaliziranih" računalnih programa), a manji dio uz korištenje MS Excela. Budući da je riječ o nastavnom materijalu namijenjenom za predavanja, tehnike rješavanja statističkih problema u MS Excelu ovdje se ne opisuju zasebno, a zainteresirani ih čitatelj može pronaći u literaturi [1]. Uz gotovo sve izračunate statističke pokazatelje navedena je njihova interpretacija, a gdje god je to bilo primjereno, komentirana je i reprezentativnost konkretnoga pokazatelja.

Ugodna mi je dužnost da se na kraju ovoga predgovora iskreno zahvalim svima koji su na bilo koji način pomogli u nastajanju ovih skripta. To se ponajprije odnosi na već spomenute

studente koji su svojim pitanjima, konstruktivnim kritikama i komentarima utjecali na završnu verziju, odnosno kvalitetu skripta. U tom smislu unaprijed zahvaljujem i svima onima koji budu pročitali skripta i ukazali na neku nenamjerno napravljenu tiskarsku pogrešku, te dali konstruktivnu kritiku, primjedbu ili prijedlog za poboljšanje osnovnoga teksta.

U Zagrebu, početkom svibnja 2011.

Bojan Kovačić

0. REKLI SU O STATISTICI[2]...

"Statistika je poput bikinija: otkriva ono sugestivno, a skriva ono vitalno." (*Aaron Levenstein*)
"Kad bi netko stao jednom nogom u kipuću vodu, a drugom nogom u hladnjak, statističar bi rekao da se taj čovjek nalazi u prosječno ugodnoj temperaturi." (*Walter Heller*)
"Postoje tri vrste laži: laž, prokleta laž i statistika." (*Mark Twain*)
"Statistika je bajka razuma." (*Martin Kessel*)
"Statistika je skup točnih podataka koji daje pogrešan rezultat." (*Anonimus*)
"Statistika naša dika: štogod hoćeš, ona slika." (*Vladimir Bulatović Vib*)
"Ja jedem kupus, ti jedeš meso – u prosjeku jedemo sarmu." (*Anonimus*)

1. UVOD U STATISTIKU

1.1. *Pojam statistike*

Statistika je znanstvena disciplina koja na organiziran način pristupa *prikupljanju*, *odabiru*, *grupiranju*, *prezentaciji* i *analizi* podataka, te *interpretiranju rezultata* provedene analize u svrhu ostvaraja postavljenih istraživačkih ciljeva. Da bi se ostvarili svi prethodno navedeni zadaci i ciljevi, statistika je razvila svoje vlastite metode i tehnike. One se danas primjenjuju u vrlo različitim područjima: ekonomiji, zdravstvu, kulturi, politici, sportu, obrazovanju, znanosti itd.

1.2. *Opisna i inferencijalna statistika*

Kao zasebna znanstvena disciplina, statistika se grubo može podijeliti na dva posebna dijela: *opisnu* (*deskriptivnu*) i *inferencijalnu statistiku*. Potonji se često smatra dijelom tzv. *matematičke statistike*.

Opisna (deskriptivna) statistika temelji se na konkretnim rezultatima dobivenima nekim istraživanjima ili mjerenjima. U tim se istraživanjima skup svih promatranih objekata obuhvaća u cijelosti. Zadaća opisne statistike je "opisati" dobivene rezultate, tj. srediti ih i sažeti tako da budu što pregledniji, razumljiviji i pogodniji za interpretaciju, daljnju analizu i primjenu.

[2] Mnogobrojni slikovitiji komentari studenata koji su više-manje uspješno polagali statističke kolegije iz ćudorednih razloga ovdje nisu navedeni.

Inferencijalna statistika temelji se na djelomičnu obuhvatu skupa svih promatranih objekata. Rezultati se dobivaju ispitivanjem određenoga dijela (*uzorka*) promatranih objekata. Koristeći metodu "od pojedinačnoga ka općemu" (tzv. *induktivna metoda*) izvode se zaključci o karakteristikama (*obilježjima*) svih promatranih objekata. Pri izvedbi takvih zaključaka postoji određeni rizik pogrešno izvedenih zaključaka. Rizik i nesigurnost u primjeni rezultata dobivenih istraživanjem na temelju uzorka procjenjuje se na temelju načela *teorije vjerojatnosti*.
U ovom ćemo se kolegiju baviti *opisnom statistikom* jer se taj dio statistike praktično najviše primijenjuje u različitim poslovnim sustavima, procesima itd.

1.3. *Statistička analiza, statistički skup i obilježja elemenata statističkog skupa*

Statistička analiza bilo koje prirodne ili društvene pojave ili procesa nužno se temelji na statističkim informacijama, metodama i tehnikama.

Statističkom analizom dobiva se uvid u strukturu pojava ili procesa u vremenu i prostoru, te uvid u njihovu međusobnu povezanost. Predmet statističke analize različita su *kvalitativna* i *kvantitativna obilježja* promatranih pojava ili procesa vezanih za promatrani skup objekata.

Statistički skup je osnovni pojam u statistici (kao, uostalom, i skup u matematici). Tvore ga *statističke jedinice* ili *elementi* (osobe, poslovni subjekti, regije, države, predmeti itd.) koji imaju barem jedno zajedničko svojstvo (*obilježje* ili *varijabla*[3]) koje od elementa do elementa očituje statističku promjenjivost. Ukupan broj elemenata statističkoga skupa naziva se *opseg statističkoga skupa.*[4] Prema opsegu, statističke skupove (kao i sve druge skupove) dijelimo na *konačne* i beskonačne, a prema vrsti elemenata na *stvarne* (*realne*) i *zamišljene*[5] (*hipotetične*). Stvarni statistički skupovi sastoje se od elemenata koji postoje u tekućem vremenu i prostoru, a zamišljeni statistički skupovi sastoje se od elemenata generiranima različitim modelima procesa ili pokusa. Poslovna statistika se u pravilu bavi realnim konačnim statističkim skupovima.

Već je istaknuto da statistički skup tvore elementi koji imaju barem jedno zajedničko svojstvo. Stoga za svaki element statističkoga skupa znamo odgovarajući podatak o tome svojstvu. Skup svih takvih podataka naziva se *osnovni skup* ili *populacija*[6]. Poput statističkoga, i osnovni skup može biti *konačan* ili *beskonačan*, te *stvaran* ili *zamišljen*.

---

[3] U opisnoj se statistici izraz *varijabla* obično poistovjećuje s izrazima *obilježje* i *svojstvo*, dok se u matematičkoj statistici odnosi isključivo na numerička obilježja.
[4] U teoriji skupova taj se broj, inače, naziva *kardinalni broj skupa*, dok se termin *opseg* više koristi u geometrijskim razmatranjima.
[5] Pritom izraz *zamišljen* ne znači stanje statističkoga skupa u kojemu svi njegovi elementi promišljaju npr. o besmrtnosti komarca malaričara, nego nepostojanje u realnom svijetu.
[6] Izraz *populacija* obično se rabi isključivo u kontekstu vezanom za određeni skup živih bića (populacija ljudi, čudnovatih kljunaša, krokodila itd.).

Bilo koji neprazan podskup[7] statističkoga skupa i osnovnoga skupa naziva se *uzorak*. Dakle, pod terminom *uzorak* podrazumijevamo i neki dio elemenata statističkoga skupa i odgovarajuće podatke o tim elementima. U opisnoj se statistici nerijetko pojmovi *statističkoga skupa*, *osnovnoga skupa* i *uzorka* poistovjećuju jer se statistička analiza provodi na *isti način* neovisno o kojem od tih triju tipova skupa je riječ. Nasuprot tome, inferencijalna statistika strogo razdvaja navedene pojmove jer provedba statističke analize ovisi o tome iz kojega se od tih triju skupova uzimaju podaci koji se žele analizirati.

Statistički skup nužno mora biti precizno definiran kako bi se na temelju njegove definicije moglo jasno i jednoznačno odrediti pripada li neki element tome skupu ili ne. Statistički skup mora biti definiran *pojmovno, prostorno i vremenski*. *Pojmovna* definicija utvrđuje pripadnost bilo kojega objekta navedenom skupu, *prostorna* definicija utvrđuje prostor kojemu pripadaju svi elementi statističkoga skupa, a *vremenska* definicija utvđuje vremensko razdoblje ili vremenski trenutak za koji su vezani svi elementi statističkoga skupa.

Primjer 1. Promatra se statistički skup koji tvore svi studenti 1. godine stručnoga studija računovodstva i financija Visoke škole za financijski menadžment u Špičkovini na dan 01.03.2010. Definirajmo navedeni skup pojmovno, prostorno i vremenski:

- pojmovna definicija: *studenti* 1. *godine stručnoga studija računovodstva i financija*
- prostorna definicija: *Visoka škola za financijski menadžment u Špičkovini*[8]
- vremenska definicija: 01.03.2010.

Taj skup je zamišljen i konačan, a njegov opseg iznosi 100 [studenata][9]. Primjer osnovnoga skupa vezanoga za promatrani statistički skup jest skup svih zaključnih ocjena iz kolegija *Poslovna matematika*, pri čemu s 0 možemo označiti činjenicu da student dosad nije uspio položiti navedeni kolegij. Stoga osnovni skup tvore ukupno 100 zaključnih ocjena[10]. Jedan od uzoraka toga skupa je npr. *Top* 20, tj. skup svih studenata koji su navedeni kolegij uspjeli položiti sa zaključnom ocjenom *izvrstan*[11].

---

[7] Iz teorije skupova je poznato da je jedan od podskupova svakoga skupa tzv. *prazan skup* (skup koji nema niti jednoga jedinoga elementa). Prazni skupovi (pa, prema tome, i prazni uzorci) koriste se isključivo u matematičkoj statistici.

[8] Pri prostornom definiranju skupa obično se navodi i npr. naziv radne organizacije, visokoobrazovne ustanove itd. iako bi možda prirodnije pripadali pojmovnoj definiciji. Također, iako u pravilu ne treba pretjerivati s prostornim definiranjem, pa npr. reći: *Visoka škola za financijski menadžment u Špičkovini*, *država*: *Republika Hrvatska*, *kontinent*: *Europa*, *planet*: *Zemlja*, *galaksija*: *Milječna staza*, nerijetko ipak treba istaknuti neke dodatne prostorne pokazatelje (naročito ukoliko se npr. radi o radnicima na crno u gradu Ouagadougou).

[9] Skup je zamišljen jer dotična visokoobrazovna ustanova (nažalost) još ne postoji.

[10] Te ocjene, naravno, međusobno nisu u parovima različite jer ih ima svega 5 različitih: 0, 2, 3, 4 i 5. Koristeći npr. Dirichletovo načelo može se pokazati da sigurno postoji barem 20 studenata koji imaju istu zaključnu ocjenu iz *Poslovne matematike*.

[11] Pritom odmah treba naglasiti da taj uzorak *nije reprezentativan*, tj. na temelju njega ne možemo izvući istinit zaključak o uspjehu svih studenata na ispitima iz *Gospodarske matematike*.

1.3.1. *Skale obilježja*

Kako je istaknuto, obilježja (varijable) elemenata statističkog skupa su odgovarajuća svojstva po kojima uspoređujemo te elemente. Svako obilježje se javlja u više pojavnih oblika (*modaliteta*). Skup svih modaliteta nekoga obilježja naziva se *skala*[12]. Tako je npr. skala modaliteta obilježja "spol" *muški* i *ženski*, skala modaliteta obilježja "dan u tjednu" *ponedjeljak*, *utorak*, *srijeda*, *četvrtak*, *petak*, *subota* i *nedjelja*, skala modaliteta "mjesec u godini" *siječanj*, *veljača*, *ožujak*, *travanj*, *svibanj*, *lipanj*, *srpanj*, *kolovoz*, *rujan*, *listopad*, *studeni* i *prosinac* itd. Prilikom podjele elemenata statističkoga skupa prema barem jednom obilježju kad god je to moguće treba utvrditi skalu modaliteta. To se poglavito odnosi na obilježja čiji se modaliteti ne iskazuju numeričkim vrijednostima (tzv. *kvalitativna obilježja*).

Skale obilježja mogu se podijeliti sukladno određenim metričkim svojstvima. Na temelju toga razlikujemo *nominalnu*, *redosljednu* (*ordinalnu* ili *rang*), *intervalnu* i *omjernu* skalu.

Nominalna skala zadaje se u obliku skupa čiji su elementi nenumeričke vrijednosti: slovne oznake, pridjevi, kategorije itd. Ti se elementi prigodom definiranja skale mogu navesti *proizvoljno odabranim redoslijedom.* Nominalne skale dodatno se dijele na *atributivne* (varijabilitet promatranoga obilježja iskazuje se opisno) i *prostorne* (varijabilitet promatranoga obilježja iskazuje se zemljopisnim jedinicama). U nekim se slučajevima nenumeričke oznake zamjenjuju numeričkima[13], pri čemu navedene numeričke oznake služe isključivo kao identifikatori, pa s njima nema smisla provoditi nikakve računske operacije. Npr. zamijenimo li nenumerički modalitet *muški* numeričkim modalitetom 0, a nenumerički modalitet *ženski* modalitetom 1, onda nema smisla zapisati niti računati npr. $0 + 1$, $0 \cdot 1$ itd.[14] Istaknimo da nominalna skala ponajprije služi za *razvrstavanje* (kategorizaciju) elemenata osnovnoga skupa u kategorije, a ukoliko je njihov broj vrlo velik, skala se utvrđuje *dogovorno* i u takvim slučajevima govorimo o *nomenklaturi*. Tipični primjeri su različite nomenklature djelatnosti, industrijskih proizvoda, usluga itd.

Redosljedna (*ordinalna, rang*) *skala* elementima statističkoga skupa pridružuje slovne oznake, simbole ili brojeve prema intenzitetu mjerenoga svojstva, pa se elementi potom klasificiraju i uređuju sukladno stupnju promatranoga svojstva. Poredak modaliteta ovdje *nije proizvoljan*: modalitete obično navodimo od onoga koji označava najmanji stupanj svojstva do onoga koji označava najveći stupanj svojstva ili obrnuto. Tipični primjeri redosljedne skale su skala ocjena (1, 2, 3, 4, 5), skala stupnjeva stručne spreme, skala stupnjeva zadovoljstva kvalitetom nastave iz kolegija "*Zabavljačke vještine*" itd. Iako ova skala ima bolja metrička

---

[12] Unatoč upornom zagovaranju glazbeno nadarenih statističara, sintagma *ljestvica modaliteta* (još) nije ušla u svakodnevnu uporabu.

[13] Takva je "zamjena" zapravo bijekcija sa skale nenumeričkih modaliteta u jednakobrojnu skalu numeričkih modaliteta: svaki nenumerički modalitet zamjenjuje se točno jednim numeričkim modalitetom i nikoja dva nenumerička modaliteta nisu zamijenjena istim numeričkim modalitetom.

[14] Šovinisti (i šovinistice?) bi vjerojatno vrlo rado definirali operaciju *množenja*, a feministice (i feministi?) operaciju *zbrajanja* ovakvih brojeva sukladno uobičajenoj aritmetici.

svojstva od nominalne, osnovni nedostatak joj je nemogućnost *preciznoga* utvrđivanja razlika mjerenoga svojstva: npr. mi znamo da je ocjena 5 bolja od ocjene 2, ali ne možemo npr. reći da je ocjena 5 za 3 bolja od ocjene 2.[15]

Ukoliko elementima statističkoga skupa pridružujemo realne brojeve sukladno intenzitetu promatranoga svojstva, dobit ćemo *intervalnu skalu*. Ona uvijek ima *definiranu mjernu jedinicu* i *dogovorno utvrđenu nulu*, pa ima bolja metrička svojstva od prethodnih dviju skala jer omogućuje i razvrstavanje i rangiranje i brojčano utvrđivanje razlika mjerenoga svojstva. Naime, s podacima dobivenima mjerenjem na intervalnoj su skali suvislo definirane operacije zbrajanja i oduzimanja[16], pa tako jednake razlike brojeva na skali govore o jednakoj razlici mjerenih svojstava. Npr. ako je temperatura zraka na Zavižanu -20°C, u Zagrebu 0°C, a u Splitu 20°C, onda se može reći da je u Zagrebu za 20°C toplije nego na Zavižanu, a za 20°C hladnije nego u Splitu. Također, dogovorno utvrđena nula znači da element statističkoga skupa posjeduje određeno svojstvo i da je modalitet toga svojstva jednak nuli. Konkretno, činjenica da je temperatura zraka u Zagrebu 0°C znači da zrak u Zagrebu ima svoju temperaturu i da je ta temperatura jednaka 0°C.[17]

Skala s najboljim metričkim svojstvima je tzv. *omjerna skala*. Ona se dobije tako da se elementima statističkoga skupa pridruže realni brojevi sukladno intenzitetu mjerenoga svojstva, a, osim definirane mjerne jedinice, ima i nulu koja *označava nepostojanje svojstva*. Zbog toga se na takvoj skali mogu definirati sve četiri osnovne računske operacije jer razlike i omjeri vrijednosti imaju suvislo tumačenje. Npr. ako Blaženko na tekućem računu ima 10.000,00 kn, a Draženko 100.000,00 kn, onda Draženko na svojem tekućem računu ima 10 puta više novaca, odnosno za 90.000,00 kn više od Blaženka. Nadalje, ako Vlatka na svojem deviznom računu ima 0,00 €, znači da je njezin devizni račun potpuno prazan, tj. da na njemu (do daljnjega) nema novaca.

Poseban tip skale obilježja je *kronološka skala*. Ona obično sadrži vremenska razdoblja poredana kronološki, tj. počevši od razdoblja koje je nastupilo najranije i završavajući s razdobljem koje je nastupilo najkasnije. Modaliteti kronološke skale mogu se iskazivati nazivima (npr. skala dana u tjednu, skala mjeseci u godini itd.) ili brojevima (npr. skala minuta u pojedinom satu, skala sati u danu, skala datuma u pojedinom mjesecu, skala godina, desetljeća, stoljeća itd.). Na tako formiranoj skali ima smisla definirati operacije zbrajanja i oduzimanja, pri čemu se u nekim slučajevima uobičajene operacije zbrajanja i oduzimanja realnih brojeva zamijenjuju operacijama zbrajanja i oduzimanja modulo n[18], pri čemu je

---

[15] Tj., opet ne možemo definirati operacije zbrajanja i oduzimanja. Ovo se pravilo praktično zanemaruje prigodom računanja *prosječne* ocjene, o čemu će biti govora u posebnom poglavlju.

[16] Množenje i dijeljenje se ne definiraju zbog problema tipa: "Temperatura zraka u Zagrebu je 0°C. Ako je na Zavižanu 20 puta hladnije nego u Zagrebu, kolika je temperatura zraka na Zavižanu?".

[17] Matematički nadareni statističari odmah će se sjetiti da se na isti način u matematici definiraju skupovi površine nula: to su skupovi koji imaju površinu i ta površina je jednaka 0. Tipičan primjer takvoga skupa je točka u bilo kojem euklidskom prostoru.

[18] *Zbrajanje modulo n* (oznaka: $+_n$) je binarna operacija definirana formulom: $a +_n b := (a + b) \bmod n$ = ostatak pri dijeljenju zbroja $a + b$ (dobivenoga uobičajenim zbrajanjem cijelih brojeva) brojem n. Npr. na skali sati u danu je $17 +_{24} 8$ = ostatak pri dijeljenju broja 17 + 8 s 24 = ostatak pri dijeljenju broja 25 s 12 = 1. *Oduzimanje*

obično $n \in \{7, 12, 24, 28, 29, 30, 31, 60\}$. Tipični primjeri kronoloških skala su: skala dana u tjednu {*ponedjeljak*, *utorak*, *srijeda*, *četvrtak*, *petak*, *subota*, *nedjelja*}, skala godina u prvom desetljeću ovoga tisućljeća {2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010} itd.

1.3.2. *Vrste obilježja*

Sukladno navedenoj podjeli skala, opća podjela obilježja razlikuje *kvalitativna*, *kvantitativna* i *vremenska obilježja*.

Kvalitativna obilježja su obilježja kod kojih su svojstva elemenata statističkoga skupa iskazana određenim pridjevima ili atributima, kategorijama i sl. Ova obilježja mogu biti *nominalna* i *redosljedna* (*obilježja ranga*).

Nominalna obilježja vezana su uz nominalnu skalu izražavanja modaliteta promatranoga obilježja, a dodatno se dijele na *atributivna* i *prostorna* (*geografska*). *Atributivna obilježja* su obilježja čiji modaliteti iskazuju *svojstva* elemenata statističkoga skupa. Tipični primjeri takvih obilježja su: spol, boja kose, boja očiju, bračno stanje, tip osobnoga automobila, vrsta trgovačkoga društva, horoskopski znak, politička stranka itd. Atributivno obilježje čija se skala sastoji od točno dva modaliteta nazivamo *alternativno obilježje*. Tipični primjeri takvoga obilježja su: spol (npr. na statističkomu skupu svih ljudi) sa skalom {*muški*, *ženski*}[19]; način studiranja (npr. na statističkom skupu svih studenata Visoke škole za financijski menadžment u Bedekovčini na dan 01.03.2008.) sa skalom {*redovni*, *izvanredni*} itd.

Primjer 2. Primjer alternativnoga statističkoga obilježja:

Razdioba svih studenata Visoke škole za financijski menadžment u Špičkovini koji su položili kolegij *Poslovna matematika* putem kolokvija u akademskoj godini 2007/2008. prema spolu

spol	*broj studenata*	*struktura* [%]
muški	10	27.03
ženski	27	72.97
ukupno	**37**	**100.00**

Izvor: osobna evidencija predmetnoga nastavnika, veljača 2008.

Zadatak 1. **a)** Definirajte statistički skup u gornjem primjeru pojmovno, prostorno i vremenski, pa odredite njegov opseg.
b) Odredite osnovni skup i njegov opseg.

---

modulo n (oznaka: $-_n$) je binarna operacija na skali modaliteta definirana opisno s: $a -_n b$ = broj c takav da je $b +_n c = a$. Npr. na skali sati u danu je $1 -_{24} 8 = 17$ jer je, kako smo vidjeli, $17 +_{24} 8 = 1$.

[19]Novije sociološke analize u priznatim znanstvenim časopisima "*Glorija*", "*Tena*", "*Story*" itd. ozbiljno ugrožavaju ovu tradicionalističko-konzervativnu podjelu svih ljudi prema spolu.

Prostorno (geografsko) obilježje je nominalno obilježje čiji modaliteti odražavaju vezu elemenata statističkoga skupa s određenim prostorom. Primjeri takvoga obilježja su: mjesto rođenja, državljanstvo, prebivalište, boravište, sjedište tvrtke, država ili ekonomska unija država s kojom se vrši određena trgovinska razmjena itd.

Redosljedna obilježja ili *obilježja ranga* vezana su uz *ordinalnu skalu* iskazivanja modaliteta. Tipični primjeri obilježja ranga su ranije spomenuta ocjena, stručna kvalifikacija, stupanj akademskoga obrazovanja, stupanj zadovoljstva kvalitetom nastave itd.

Primjer 3. Primjer redosljednoga statističkoga obilježja:

Razdioba svih studenata Visoke poslovne škole "*Mutimir Škegro*" iz Špičkovine u akademskoj godini 2006/07. prema stupnju zadovoljstva radom studentske referade
(stanje na dan 30.09.2007.)

stupanj zadovoljstva	*broj studenata*	*struktura* [%]
vrlo nezadovoljan	15	9.32
nezadovoljan	32	19.88
uglavnom nezadovoljan	21	13.04
ni zadovoljan ni nezadovoljan	38	23.60
uglavnom zadovoljan	27	16.77
zadovoljan	17	10.56
vrlo zadovoljan	11	6.83
ukupno:	**161**	**100.00**

Izvor: Studentski zbor VPŠ "*Mutimir Škegro*"

Kvantitativna obilježja su obilježja kod kojih se svojstva elemenata statističkoga skupa izražavaju brojevima, a vezana su uz *intervalnu* i *omjernu skalu* iskazivanja modaliteta. Kvantitativna obilježja često se nazivaju i *numerička obilježja*, a dodatno se dijele na *diskretna* (*diskontinuirana*) i *kontinuirana*.

Diskretno (diskontinuirano) obilježje je obilježje čija je skala modaliteta konačan skup (općenito, realnih) brojeva takav da između svakih dvaju modaliteta postoji *razmak*. *Razmak* između dvaju modaliteta je, zapravo, apsolutna vrijednost[20] njihove razlike. Tipični primjeri obilježja su broj soba u stanu, broj djece u obitelji, iznos ukupnih mjesečnih primanja itd. Pritom modalitet diskretnoga obilježja ne mora biti nužno cjelobrojan. Postoje i četveroipolsobni stanovi, iznos mjesečnih primanja može biti necjelobrojan, veličina obuće može biti 48 i pol, ali neovisno o tome između bilo kojih dvaju različitih modaliteta uvijek postoji spomenuti razmak.

[20] Podsjetnik za nematematički nastrojene statističare: *apsolutna vrijednost* bilo kojega realnoga broja x definirana je formulom $|x| = \begin{cases} x, & \text{za } x \geq 0; \\ -x, & \text{za } x < 0. \end{cases}$

Primjer 4. Primjer kvantitativnoga diskretnoga obilježja:

Razdioba svih zaposlenika u tvrtki "*Brodolomac d.d.*" iz Muća prema broju djece
(stanje na dan 31.12.2007.)

broj djece	*broj zaposlenih*	*struktura* [%]
0	175	15.31
1	244	21.35
2	438	38.32
3	194	16.97
4 i više	92	8.05
ukupno	**1.143**	**100.00**

Izvor: Služba ljudskih potencijala tvrtke "*Brodolomac*"

Kontinuirano obilježje je obilježje čija je skala modaliteta (teoretski) beskonačan skup. Ta je skala najčešće poluotvoreni interval $\langle a, b]$[21] ili konačna unija takvih poluotvorenih intervala. Praktično, to znači da modalitet obilježja može biti *bilo koja* vrijednost unutar nekoga od intervala koji tvore omjernu skalu. Tipični primjeri ovoga obilježja su visina, masa itd.

Primjer 5. Primjer kvantitativnoga kontinuiranoga obilježja:

Razdioba svih učenika 4.b razreda gimnazije "*Ljilja Kokić*" iz Kašine u šk.god. 2007/2008.
prema visini (stanje na dan 01.03.2008.)

visina [cm]	*broj učenika*	*struktura* [%]
(160) – 165	3	11.11
165 – 170	6	22.22
170 – 175	7	25.93
175 – 180	4	14.81
180 – 185	5	18.52
185 – (195)	2	7.41
ukupno:	**27**	**100.00**

Izvor: osobna evidencija učitelja tjelesne i zdravstvene kulture Jože Kotrmelca

Zadatak 2. **a**) Definirajte statističke skupove u Primjerima 3. – 5. pojmovno, prostorno i vremenski, pa odredite njihove opsege.
b) Odredite osnovne skupove u Primjerima 3. – 5. i njihove opsege.
c) Odredite skale modaliteta obilježja promatranih u Primjerima 3. – 5.

Pri razlikovanju diskretna i kontinuirana obilježja u nekim je slučajevima vrlo važno precizno definirati promatrano kvantitativno obilježje. Npr. "starost (čovjeka, drveta, zgrade)" je kvantitativno kontinuirano obilježje, no, ukoliko to obilježje predefiniramo kao "navršene godine starosti", ono postaje kvantitativno diskretno obilježje (zašto?).

---

[21] Podsjetnik za nematematički nastrojene statističare: $\langle a, b]$ je skup svih realnih brojeva koji su strogo veći od realnoga broja a, a manji ili jednaki realnom broju b, tj. $\langle a, b] = \{x \in \mathbf{R}: a < x \leq b\}$.

Poseban tip obilježja je *vremensko obilježje*. To je obilježje vezano uz kronološku skalu modaliteta, odnosno odgovarajući kronološki uređaj. Sukladno njemu grupiramo elemente osnovnoga skupa i dobivamo *vremenski statistički niz*. Vremenski niz može biti *intervalan* i *trenutačan* ovisno o tome jesu li podaci iz osnovnoga skupa dobiveni neprekidnim praćenjem i kumuliranjem promatrane pojave u odabranom vremenskom intervalu ili pak određivanjem stanja pojave u odabranim vremenskim trenutcima.

Primjer 6. Primjer intervalnoga vremenskoga niza:

Razdioba ostvarenoga prometa u trgovini na malo u Kraljevini Dajdamdaš u razdoblju od 2002.-2006. prema godini

godina	*ostvareni promet* [000 000 USD]
2002.	9.45575824
2003.	11.79383472
2004.	18.51123944
2005.	25.55631827
2006.	30.10276687

Izvor: Statistički ljetopis Kraljevine Dajdamdaš za 2007. godinu

Primjer 7. Primjer trenutačnoga vremenskoga niza

Razdioba svih studenti koji su položili ispit iz *Poslovne statistike* na Veleučilištu u Brlog Gradu na ispitnim rokovima u razdoblju od 01.09.2007. do 01.03.2008. prema datumu polaganja ispita

datum	*broj studenata*
01.09.2007.	2
20.09.2007.	1
04.10.2007.	26
12.02.2008.	0
29.02.2008.	1

Izvor: Studentska služba Veleučilišta u Brlog Gradu

2. FAZE STATISTIČKOGA ISTRAŽIVANJA

Statističko istraživanje općenito možemo podijeliti na sljedeće faze:

1. Određivanje cilja i razrada plana istraživanja;
2. Organizirano prikupljanje statističkih podataka;
3. Grupiranje i sređivanje statističkih podataka;
4. Tablični i grafički prikaz statističkih sređenih podataka;
5. Statistička analiza i interpretacija njezinih rezultata.

U ovom ćemo se kolegiju detaljno baviti fazama 3., 4. i 5., pa ćemo prve dvije faze ovdje spomenuti samo okvirno.

2.1. *Određivanje cilja i razrada plana istraživanja*

Iako se vrlo često neopravdano zanemaruje, ova je faza zapravo jedna od najvažnijih faza cjelokupnoga statističkog istraživanja. Naime, prije negoli se krene na prikupljanje podataka, svakako treba jasno i precizno definirati što se želi postići statističkim istraživanjem jer već dobro definiran cilj istraživanja može nametnuti prirodan izbor pogodnih metoda za prikupljanje podataka. Ukoliko cilj istraživanja i plan istraživanja nisu dovoljno precizno definirani, moguće je uzalud prikupiti veliki broj podataka koje potom nije moguće dostatno kvalitetno podvrgnuti statističkoj analizi i interpretaciji.

Druga velika prednost preciznoga definiranja cilja i plana istraživanja je što znatno pojednostavljuje i skraćuje cjelokupni postupak. Npr. pretpostavimo da želimo ispitati zainteresiranost hrvatskih umirovljenika za aktivno sudjelovanje u tečaju osnova rada na osobnim računalima. Iako npr. odlazak u Hrvatsku stranku umirovljenika i razgovor s njezinim čelnicima može dati korisne podatke o cilju našega istraživanja, ta relativno jeftina i kratkotrajna metoda nije dobra za postizanje željenoga cilja[22] jer stav čelnika stranke ne mora održavati stav relativne većine hrvatskih umirovljenika. Teorijski najbolja metoda je postaviti odgovarajuće *anketno pitanje* svakom pojedinom umirovljeniku, ali ta je metoda npr. zbog svoje dugotrajnosti[23] uzrokovane znatnim brojem hrvatskih umirovljenika praktički potpuno neprimjerena. Praktično najpogodnija metoda je izabrati *dovoljno reprezentativan uzorak*[24] i

[22] Navedena metoda, međutim, može biti vrlo korisna za postavljanje *hipoteze* istraživanja. *Hipoteza* je temeljna *neprovjerena* pretpostavka koju nastojimo potvrditi ili oboriti statističkim istraživanjem. U navedenom bi primjeru (nakon obavljenih razgovora u HSU) postavljena hipoteza mogla glasiti: "Većina hrvatskih umirovljenika je zainteresirana za sudjelovanje u tečajevima osnova rada na osobnim računalima."

[23] Drugi prirodan razlog je cijena praktične provedbe takve metode koja nerijetko predstavlja ozbiljnu prepreku za kvalitetnu provedbu statističkoga istraživanja.

[24] Vrlo česta pogreška statističkih istraživanja upravo je "megalomanija" odgovarajućih planova istraživanja jer se istraživanje planira obaviti na prevelikom uzorku.

(npr. telefonski) anketirati sve članove toga uzorka. Metodama izbora uzoraka i procjenjivanja njihove reprezentativnosti bavi se posebna statistička disciplina *teorija uzoraka*. Detalje ovdje izostavljamo.

2.2. *Organizirano prikupljanje statističkih podataka*

Statistički podaci obično se prikupljaju:

a) brojanjem (npr. iznos mjesečne plaće svakoga od djelatnika tvtke "*Drp–commerce*")
b) mjerenjem (npr. masa svake od natjecateljica za Miss Mliječne staze)
c) ocjenjivanjem (npr. stupanj zadovoljstva kvalitetom nastave iz *Poslovne statistike*);
d) evidentiranjem (npr. zaključne ocjene iz *Poslovne statistike* svih studenata Visoke škole za financijski menadžment u Bedekovčini koji su položili taj kolegij u prošloj akademskoj godini);
e) anketiranjem (intervjuiranjem) (npr. politička stranka za koju na predstojećim mjesnim izborima namjerava glasovati svaki stanovnik Muća)

Rezultat organiziranoga prikupljanja statističkih podataka su tzv. "*sirovi*" (*negrupirani*, *neobrađeni*) podaci. Iz njih se najjasnije vidi koji je element osnovnoga skupa pridružen svakom pojedinom elementu statističkoga skupa. Uočimo da obrnuta tvrdnja nije točna jer bez grupiranja i sređivanja podataka općenito ne možemo "očitati" kojim je sve elementima statističkoga skupa pridružen svaki element osnovnoga skupa.

Najpoznatija, a možda i najprimjenjivanija metoda organiziranoga prikupljanja statističkih podataka je *anketiranje*. *Anketa* predstavlja skup različitih postupaka kojima se pobuđuju, prikupljaju i analiziraju izjave *anonimnih ljudi*[25] kako bi se dobili podaci o njihovu ponašanju, stavu prema određenom problemu, mišljenju, interesima, preferencijama itd. Ti se podaci dobivaju na temelju *anketnih upitnika* koji sadrže promišljeno odabrana pitanja koja trebaju biti jednostavna, jasna i nesugestivna. Ta se pitanja mogu postaviti *pismeno* i *usmeno*. Usmeno postavljanje anketnih pitanja obično se naziva *intervjuiranje*[26].

U nastavku navodimo primjer prvoga dobro strukturirane ankete za procjenu nastavnika i kolegija koja se sukladno načelima Bolonjskoga procesa provodi na svim znanstveno-nastavnim sastavnicama Sveučilišta u Zagrebu. Jedini možebitni nedostatak navedene ankete je pretpostavka da svaki pojedini anketirani student poznaje temeljni postotni račun od sto, što u praksi ne mora biti slučaj.

---

[25] Psihološka istraživanja pokazuju da je uvjet anonimnosti sudionika ankete iznimno važan jer čak može utjecati na iskrenost u davanju odgovora na anketna pitanja, a time i na realnost rezultata provedene ankete.
[26] Moderna praksa nerijetko pokazuje da se postavljanje pitanja (unaprijed) u pismenoj formi i prepisivanje njihovih odgovora također naziva *intervju*, napose ukoliko je riječ o medijski poznatim osobama (npr. predsjednici država, premijeri vlada itd.). Sa statističkoga je stanovišta takvo izjednačavanje potpuno pogrešno.

Sveučilište u Zagrebu
Anketni list za procjenu nastavnika i kolegija (V1)

Nastavnik/nastavnica: _____ Kolegij: _____

Poštovane studentice i studenti,
u okviru projekta vrednovanja nastave na Sveučilištu putem ovog upitnika možete procijeniti svoje zadovoljstvo **kvalitetom nastave vaših nastavnika i asistenata na svakom pojedinom kolegiju, te izvedbu kolegija u cjelini.** Vaši iskreni odgovori mogu upozoriti na nedostatke i teškoće u nastavi, te utjecati na poboljšanje njezine kvalitete. Ako neku procjenu **ne možete dati ili nije primjenjiva** na određenog nastavnika ili kolegij, odaberite odgovor "ne mogu procijeniti". Molimo vas da sve procjene dajete zacrnjivanjem kružića uz odabrani odgovor (primjer: ●).

A Opći podaci o studentici/studentu:

1. Spol: ♀ ♂
2. Vaša prisutnost na nastavi ovoga nastavnika: **1 – rijetka (do 30%) 2 – povremena (30-70%) 3 – redovita (više od 70%)** ① ② ③
3. Kakav je, na početku nastave, bio vaš interes za sadržaje koje kolegij obrađuje: **1 – mali 2 – srednji 3 – veliki** ① ② ③
4. Koja vam je do sada najčešća ocjena u indeksu: ② ③ ④ ⑤
5. Koju ocjenu očekujete iz ovog kolegija: ② ③ ④ ⑤

B Procjena nastavnika/nastavnice na zadanom kolegiju:

Na sljedećoj ljestvici procijenite u kojoj mjeri navedene tvrdnje dobro opisuju rad nastavnika/nastavnice:

u potpunosti se ne slažem ① ② ③ ④ ⑤ *u potpunosti se slažem*

	1	2	3	4	5	ne mogu procijeniti
1. Kroz nastavu pokazuje dobro poznavanje sadržaja kolegija.	○	○	○	○	○	○
2. Na postavljena pitanja odgovara stručno i spremno.	○	○	○	○	○	○
3. Kvalitetnim primjerima i zadacima olakšava razumijevanje sadržaja.	○	○	○	○	○	○
4. Jasno i razumljivo izlaže/demonstrira nastavne sadržaje.	○	○	○	○	○	○
5. Nastava je dobro strukturirana i raspoloživo vrijeme je racionalno iskorišteno.	○	○	○	○	○	○
6. Jasno definira ciljeve nastave i ono što očekuje od studenata.	○	○	○	○	○	○
7. Nastava je zanimljiva i dinamična.	○	○	○	○	○	○
8. Uporabom nastavnih pomagala i suvremene tehnologije podiže kvalitetu nastave.	○	○	○	○	○	○
9. Ima dobre komunikacijske vještine i stvara ugodnu radnu atmosferu.	○	○	○	○	○	○
10. Prema studentima se odnosi korektno i s poštovanjem.	○	○	○	○	○	○
11. Dostupan je i susretljiv za konzultacije sa studentima.	○	○	○	○	○	○
12. Motiviran je za rad i savjesno izvršava svoje obaveze.	○	○	○	○	○	○
13. Nastavu održava redovito i na vrijeme.	○	○	○	○	○	○
14. Koju biste opću ocjenu dali ovom nastavniku/nastavnici u cjelini? *nedovoljno*	○	○	○	○	○	*izvrsno*

Procjene u nastavku koje se odnose na kolegij u cjelini procijenit ćete **samo ako ste u prethodnom dijelu "B" procjenjivali nastavnika koji je nositelj kolegija!**

C Procjena kolegija u cjelini (uključuje sudjelovanje svih nastavnika i asistenata):

Procjene uz pitanja 6. i 7. odnose se samo na kolegije koji uključuju vježbe i/ili seminare.	1	2	3	4	5	ne mogu procijeniti
1. Ciljevi i zahtjevi kolegija bili su jasno definirani.	○	○	○	○	○	○
2. Kolegij nudi uvid u ključne sadržaje i omogućuje njihovo razumijevanje.	○	○	○	○	○	○
3. Propisana literatura (udžbenici, skripta i ostali materijali) je korisna za razumijevanje sadržaja kolegija.	○	○	○	○	○	○
4 Količina nastavnog sadržaja odgovora predviđenoj satnici kolegija.	○	○	○	○	○	○
5. Kolegij vam je omogućio uvid u važnost područja i mogućnosti primjene.	○	○	○	○	○	○
6. Predavanja su na primjeren način bila popraćena vježbama, seminarima i drugim praktičnim oblicima nastave.	○	○	○	○	○	○
7. Vježbe i seminari omogućili su razvijanje vještina, te praktičnu primjenu znanja.	○	○	○	○	○	○
8. Organizacija kolegija potiče studente na aktivno sudjelovanje u nastavi.	○	○	○	○	○	○
9. Koju biste opću ocjenu dali ovom kolegiju u cjelini? *nedovoljno*	○	○	○	○	○	*izvrsno*

OKRENITE LIST !

2.3. Grupiranje statističkih podataka

Nakon što su prikupljanjem (i odgovarajućim unosom) dobiveni "sirovi" statistički podatci, vrlo često ih treba "srediti" tako da budu što pregledniji i lakši za daljnju statističku analizu[27]. U tu se svrhu koristi *grupiranje statističkih podataka*. Osnovna ideja grupiranja je podijeliti *statistički skup* na podskupove prema svim modalitetima koji tvore skalu obilježja, tj. za svaki pojedini modalitet utvrditi (prebrojati) kolikom broju elemenata statističkoga skupa je pridružen dotični modalitet. Pritom treba poštovati dva osnovna načela:

- *isključivost*: svaki element statističkoga skupa ne može istodobno biti u barem dva različita podskupa;
- *iscrpnost*: svaki element statističkoga skupa mora biti obuhvaćen grupiranjem podataka.[28]

2.3.1. *Grupiranje kvalitativnih podataka*

Kako je već istaknuto, grupiranjem podataka za svaki pojedini modalitet utvrđujemo kolikom je broju elemenata statističkoga skupa pridružen taj modalitet. Dobiveni broj obično se naziva *apsolutna frekvencija* (*učestalost*). Neformalno, *apsolutna frekvencija* nekoga modaliteta jednaka je ukupnom broju pojavljivanja toga modaliteta u osnovnom skupu[29]. Budući da je svakom elementu statističkoga skupa nužno pridružen točno jedan modalitet, zbroj apsolutnih frekvencija svih modaliteta mora biti jednak opsegu statističkoga skupa[30].

Relativna frekvencija modaliteta jednaka je količniku odgovarajuće apsolutne frekvencije toga modaliteta i opsega *statističkoga* skupa. Takav se račun provodi jer želimo vidjeti kolikom dijelu statističkoga skupa je pridružen određeni modalitet. Iako je relativna frekvencija općenito neki nenegativan realan broj iz segmenta [0, 1], radi praktičnih ga je potreba vrlo prikladno izraziti u proporcijama, postotcima ili promilima[31].

---

[27] Izuzetak su slučajevi kad osnovni skup sadrži relativno malo (obično ≤ 30) podataka.

[28] Budući da je preslikavanje f koje svakom elementu statističkoga skupa S pridružuje točno jedan element skale modaliteta M funkcija, za svaki modalitet $y \in M$ ima smisla definirati skup $f^{\leftarrow}(y) = \{x \in S: f(x) = y\}$. Navedeni skup se naziva *praslika* modaliteta y. Stoga strogo formalno možemo reći da je grupiranje podataka *particija statističkoga skupa* na međusobno *disjunktne praslike* pojedinih modaliteta.

[29] Strogo formalno, apsolutna frekvencija modaliteta y je ukupan broj elemenata skupa $f^{\leftarrow}(y)$, gdje je f funkcija iz Bilješke 28. Autori koji poistovjećuju pojmove *statistički skup* i *osnovni skup* definiraju apsolutnu frekvenciju kao ukupan broj pojavljivanja nekoga modaliteta u statističkom skupu. No, općenito je bolje i preciznije u definiciji rabiti pojam *osnovnoga skupa*.

[30] Aritmetika kardinalnih brojeva iz teorije skupova osigurava da navedeno svojstvo *ne ovisi* o tipu statističkoga skupa (konačan ili beskonačan). Dakle, neovisno o tome je li statistički skup konačan ili beskonačan, zbroj apsolutnih frekvencija svih modaliteta uvijek je jednak opsegu statističkoga skupa.

[31] *Proporcija* označava određeni dio jedinične veličine, a to je zapravo decimalan broj između 0 i 1. Kod beskonačnih se statističkih skupova relativne frekvencije modaliteta ne određuju jer aritmetika beskonačnih kardinalnih brojeva ne definira pojam *dijeljenja* kardinalnih brojeva. Stoga su relativne frekvencije praktično vezane isključivo za realne konačne statističke skupove. Budući da se obično iskazuju u postotcima ili promilima, za njihovo je računanje potrebno poznavati postotni račun od sto, odnosno promilni račun od tisuću.

Primjer 1. Promatramo statistički skup iz Primjera 1. u odjeljku 1.3. Ukupno 20 studenata (elemenata promatranoga skupa) iz kolegija *Poslovna matematika* ima zaključnu ocjenu *izvrstan*. Stoga je apsolutna frekvencija modaliteta *izvrstan* u odgovarajućem osnovnom skupu jednaka 20. Budući da je opseg promatranoga statističkoga skupa 100, relativna frekvencija modaliteta *izvrstan* jednaka je

$$\frac{20}{100} = 0.2 = 20\%.$$

Tako možemo zaključiti da *iz kolegija Poslovna matematika* 20% *svih promatranih studenata ima zaključnu ocjenu "izvrstan"*[32].

Primjer 2. Istraživanjem potrošnje inozemnih putnika u Kraljevini Eldorado na uzorku od ukupno 30 370 inozemnih putnika u razdoblju od lipnja do prosinca 2007. godine i odgovarajućim grupiranjem podataka, između ostalih, dobiveni su sljedeći podaci:

vrsta prijelaza	*broj putnika*	*struktura* [%]
cestovni	25 924	85.36%
zračni	1 354	4.46%
željeznički	2 075	6.83%
pomorski	1 017	3.35%
ukupno:	30 370	100.00%

Apsolutna frekvencija modaliteta *cestovni* je 25 924, što znači da je *ukupno* 25 924 *anketirana inozemna putnika anketirano na cestovnim graničnim prijelazima*. Odgovarajuća relativna frekvencija je 85.36%, što znači da je 85.36% *svih anketiranih inozemnih putnika anketirano na cestovnim graničnim prijelazima*. Za vježbu objasnite svaku od ostalih apsolutnih, odnosno relativnih frekvencija.

2.3.2. *Grupiranje kvantitativnih podataka*

Ukoliko se podaci koje grupiramo odnose na *kvantitativno* statističko obilježje, njihovim grupiranjem dobivamo poseban tip grupiranih statističkih podataka. Riječ je o tzv. *razdiobi* (*distribuciji*) *frekvencija*. U slučaju maloga broja modaliteta grupiranje ovakvih podataka praktično se ne razlikuje od grupiranja kvalitativnih podataka. Međutim, u slučaju velikoga broja modaliteta radi preglednosti ih je primjerenije grupirati u tzv. *razrede*. Definirajmo ukratko osnovne pojmove.

[32] Strogo formalno govoreći, interpretacija bi trebala glasiti: *Ukupan broj studenata koji su dobili ocjenu „izvrstan“ iznosi* 20% *od ukupnoga broja svih studenata*. Zbog svakodnevnoga govora mi ćemo dozvoljavati i navedenu interpretaciju.

<u>*Zatvoreni razred*</u>[33] je bilo koji poluzatvoreni podskup skupa realnih brojeva **R** oblika $\langle a, b]$, a <u>*otvoreni razred*</u> je bilo koji otvoreni podskup skupa realnih brojeva **R** oblika $\langle a, +\infty\rangle$ ili $\langle -\infty, a\rangle$[34]. Realan broj a nazivamo <u>*donja granica*</u> (zatvorenoga) <u>*razreda*</u>, a realan broj b <u>*gornja granica*</u> (zatvorenoga) <u>*razreda*</u>. Realan broj $h = b - a$ naziva se <u>*širina (veličina) razreda*</u> i taj je broj praktično uvijek strogo pozitivan (tj. strogo veći od nule). Istaknimo još da realan broj $s = \frac{1}{2} \cdot (a + b)$ uobičajeno nazivamo <u>*razredna sredina*</u>[35].

<u>Primjer 3.</u> Donja granica razreda $\langle 50, 100]$ jednaka je $a = 50$, dok je gornja granica jednaka $b = 100$. Širina razreda jednaka je $h = b - a = 100 - 50 = 50$, a razredna sredina je broj $s = \frac{1}{2}(a + b) = \frac{1}{2}(50 + 100) = 75$.

Ukupan broj (k) razreda u koje trebamo grupirati ukupno N različitih modaliteta kvantitativnoga obilježja općenito nije unaprijed određen. U većini statističkih istraživanja i izradama odgovarajućih računalnih programa primijenjuje se tzv. <u>*Sturgesovo*</u>[36] <u>*pravilo*</u>:

$$k \approx \lceil 1 + \log_2 N \rceil \approx \lceil 1 + 3.321928091 \cdot \log N \rceil \approx \lceil 1 + 1.442695 \cdot \ln N \rceil.$$ [37]

Pojedine širine razreda također se ne zadaju unaprijed, nego općenito ovise o tipu razdiobe podataka[38]. Najjednostavniji i praktično najprimjenjeniji slučaj je grupiranje podataka u razrede jednakih širina, pri čemu prešutno pretpostavljamo da je razdioba podataka barem približno simetrična. Tada izračunamo razliku najvećega i najmanjega modaliteta, pa tu razliku podijelimo ukupnim brojem razreda i tako dobijemo traženu širinu razreda. Ovaj se slučaj naročito primijenjuje kod tzv. *pravih razreda* (objašnjenje u nastavku teksta).

Ako razdioba podataka nije simetrična, primjereno je koristiti razrede manjih širina na području u kojemu je "gustoća" podataka veća, a razrede većih širina na području s manjom "gustoćom" podataka.

---

[33] Radi potreba daljnje statističke analize, mi ćemo pod pojmom *razred* podrazumijevati *zatvoreni razred*, te prešutno pretpostavljati da za realne brojeve a i b vrijedi stroga nejednakost $a < b$.

[34] Za matematički nenastrojene statističare: $\langle a, +\infty\rangle$ je skup svih realnih brojeva koji su strogo veći od broja a: $\langle a, +\infty\rangle = \{x \in \mathbf{R}: x > a\}$. $\langle -\infty, a\rangle$ je skup svih realnih brojeva koji su strogo manji od broja a, tj. $\langle -\infty, a\rangle = \{x \in \mathbf{R}: x < a\}$.

[35] Kasnije ćemo vidjeti da je razredna sredina zapravo aritmetička sredina donje i gornje granice razreda.

[36] Herbert Sturges, američki statističar.

[37] $\lceil x \rceil$ = najmanji cijeli broj veći ili jednak x. Npr. $\lceil 3{,}14 \rceil = 4$, $\lceil -0{,}05 \rceil = 0$ itd.

[38] O tipovima razdiobe podataka bit će više riječi kasnije u poglavlju o mjerama asimetrije. Ovdje istaknimo jedino da noviji statistički radovi predlažu da se širine razreda, umjesto izvedeno iz Sturgesova pravila, računaju prema Freedman – Diaconisovoj formuli: $h = \frac{2 \cdot IQ}{\sqrt[3]{n}}$, gdje je IQ interkvartil osnovnoga skupa. O interkvartilu ćemo nešto više reći kasnije u poglavlju o mjerama raspršenja.

Primjer 4. Mjerenjem visina (i iskazivanjem podataka u centimetrima) svih učenika 5., 6., 7. i 8. razreda Osnovne škole "*Milivoj Glavović*" iz Brckovljana na dan 29.02.2008. dobiveno je ukupno 506 podataka. Najmanji od njih je 151, najveći 193, a ukupan broj različitih modaliteta je 38. Prema Sturgesovu pravilu podaci se mogu grupirati u ukupno $k \approx \lceil 1 + \log_2 38 \rceil = \lceil 6.213 \rceil = 7$ razreda. Visine učenika su obično približno simetrično raspoređene, pa je dobivene podatke primjereno grupirati u jednake razrede čija je širina $h = \frac{193-151}{6} = 7.$

Uz grupiranje statističkih podataka često se javlja tzv. *problem nominalnih razreda*. Naime, grupiranjem modaliteta kvantitativnoga obilježja u razrede dobivamo točno jedan od sljedećih dvaju slučajeva:

1.) gornja granica svakoga pojedinoga razreda, osim posljednjega, jednaka je donjoj granici neposredno sljedećega razreda;
2.) gornja granica svakoga pojedinoga razreda, osim posljednjega, nije jednaka donjoj granici neposredno sljedećega razreda (tj. između dvaju uzastopnih razreda postoji *razmak*).

Prvi slučaj praktično primijenjujemo uvijek kad analiziramo modalitete *kvantitativnoga kontinuiranoga obilježja* kojih teorijski može biti beskonačno mnogo. Granice tako dobivenih razreda nazivamo *prave granice* i razredi s takvim granicama se najčešće koriste u statistici.

Drugi slučaj se obično primijenjuje pri analizi modaliteta *kvantitativnoga diskretnoga obilježja* kojih je obično konačno mnogo. Zbog toga između gornje granice jednoga razreda i donje granice njemu neposredno sljedećega razreda postoji *razmak* (obično jednak 1). Takve granice nazivaju se *nominalne granice*, a radi potreba daljnje statističke analize obično ih je potrebno "pretvoriti" u prave. Najjednostavniji postupak pokazat ćemo na sljedećem primjeru.

Primjer 5. Pretpostavimo da smo grupiranjem podataka o iznosu mjesečnih plaća svih zaposlenika knjigovodstvenoga servisa "*Knjiži kod mene*" iz Košara dobili sljedeću tablicu:

iznos mjesečne plaće [kn]	*broj zaposlenika*
1600 – 1900	2
2000 – 2400	4
2500 – 2900	6
3000 – 3400	2
3500 – 8000	1

Sve navedene granice su očito nominalne granice. Razmaci između dvaju uzastopnih razreda su jednaki i iznose 100. Pretvaranje u prave granice radimo tako da svaku pojedinu donju nominalnu granicu umanjimo za polovinu razmaka, a svaku pojedinu gornju nominalnu granicu uvećamo za polovicu razmaka. Budući da je polovina razmaka jednaka 50, dobivamo:

iznos mjesečne plaće [kn]	*broj zaposlenika*
1550 – 1950	2
1950 – 2450	4
2450 – 2950	6
2950 – 3450	2
3450 – 8050	1

2.3.3. *Kumulativni nizovi frekvencija*

Ponekad je radi kvalitetnije statističke analize apsolutne i relativne frekvencije potrebno grupirati u konačne nizove. To se poglavito radi ukoliko analiziramo kvalitativno redosljedno obilježje čiji su modaliteti iskazani numeričkim vrijednostima (npr. ocjena na ispitu) ili bilo koje kvantitativno obilježje. U praksi se najviše koriste dva tipa takvih nizova: *kumulativni niz* "*manje od*" i *kumulativni niz* "*veće od*". Objasnit ćemo ih na sljedećem primjeru.

Primjer 6. Na 1. godini stručnoga studija smjehologije na Višoj uzaludnoj školi u Špičkovini na dan 29.02.2008. bilo je ukupno 80 studenata. Svi oni su polagali pismeni ispit iz kolegija "*Zabavljačke vještine*". Rezultati pismenoga ispita dani su u sljedećoj tablici:

ocjena	*broj studenata*
1	0
2	10
3	15
4	25
5	30
ukupno:	80

Za svaki pojedini modalitet definiramo *kumulativnu apsolutnu frekvenciju* "*manje od*" kao *ukupan broj elemenata osnovnoga skupa koji su ili strogo manji od dotičnoga modaliteta ili jednaki tom modalitetu*. Tako je kumulativna apsolutna frekvencija "manje od" modaliteta 1 jednaka 0 jer niti jedan student nije ocijenjen ocjenom 1. Kumulativna apsolutna frekvencija "manje od" modaliteta 2 jednaka je 10 (nema studenata koji su dobili ocjenu strogo manju od 2, ali postoji 10 studenata koji su ocijenjeni ocjenom 2), kumulativna apsolutna frekvencija "manje od" modaliteta 3 jednaka je 25 (ukupno 10 studenata je ocijenjeno ocjenom strogo manjom od 3, a 15 studenata je ocijenjeno ocjenom 3), kumulativna apsolutna frekvencija "manje od" modaliteta 4 jednaka je 50 (25 studenata je ocijenjeno ocjenom strogo manjom od 4, a jednako toliko ih je ocijenjeno ocjenom 4) i kumulativna apsolutna frekvencija "manje od" modaliteta 5 jednaka je 80 (svi studenti su ocijenjeni ocjenom jednakom ili manjom od 5).

ocjena	*broj studenata*	*kumulativna apsolutna frekvencija "manje od"*
1	0	0
2	10	10
3	15	25
4	25	50
5	30	80
ukupno:	80	

Dobivene kumulativne apsolutne frekvencije objašnjavamo ovako: 10 *studenata je na ispitu ocijenjeno ocjenom* 2 (niti jedna ocjena nije strogo manja od 2), 25 *studenata je na ispitu ocijenjeno ocjenom* 2 *ili ocjenom* 3, 50 *studenata je na ispitu ocijenjeno ocjenom* 2, *ocjenom* 3 *ili ocjenom* 4, *a* 80 *studenata je na ispitu ocijenjeno ocjenom* 2, *ocjenom* 3, *ocjenom* 4 *ili ocjenom* 5. Kraće i matematičkije, tu rečenicu možemo izreći ovako: 10 *studenata je na ispitu ocijenjeno ocjenom najviše jednakom* 2, 25 *studenata je na ispitu ocijenjeno ocjenom najviše jednakom* 3, 50 *studenata je na ispitu ocijenjeno ocjenom najviše jednakom* 4, *a svi studenti su na ispitu ocijenjeni ocjenom najviše jednakom* 5.[39]

Označimo li s f_x apsolutnu frekvenciju, a s $c_x^<$ kumulativnu apsolutnu frekvenciju "manje od" modaliteta x, nije teško vidjeti da vrijedi jednakost:

$$c_x^< = \sum_{y \le x} f_y = \sum_{y < x} f_y + f_x$$

koja nam omogućuje relativno brzo računanje svake pojedine kumulativne apsolutne frekvencije. Računamo ih krećući se od *najslabijega modaliteta* prema *najboljemu* i za svaki modalitet (osim prvoga, čija je kumulativna apsolutna frekvencija "manje od" jednaka odgovarajućoj "običnoj" apsolutnoj frekvenciji) kumulativnu apsolutnu frekvenciju "manje od" izračunamo tako da izračunatoj kumulativnoj apsolutnoj frekvenciji "manje od" prethodnoga modaliteta pribrojimo apsolutnu frekvenciju modaliteta čiju kumulativnu frekvenciju "manje od" želimo izračunati. Kao provjera ispravnosti izračuna može nam poslužiti i činjenica da kumulativna apsolutna frekvencija "manje od" najboljega (praktično, posljednjega) modaliteta nužno mora biti jednaka opsegu statističkoga skupa.

Podijelimo li svaki član dobivenoga niza kumulativnih apsolutnih frekvencija s *opsegom promatranoga statističkoga skupa*, dobit ćemo <u>*kumulativni niz relativnih frekvencija*</u> "<u>*manje od*</u>". U promatranom primjeru taj je niz:

[39] Mini-tečaj hrvatskoga jezika, lekcija 1: *najviše jednak x* je istoznačnica izraza *manji ili jednak x*, odnosno izraza *jednak x ili strogo manji od x*.

ocjena	*broj studenata*	*kumulativna apsolutna frekvencija "manje od"*	*kumulativna relativna frekvencija "manje od" [%]*
1	0	0	0
2	10	10	12,5
3	15	25	31,25
4	25	50	62,5
5	30	80	100
ukupno:	80		

Dobiveni kumulativni niz interpretiramo na sljedeći način: *Niti jedan student na ispitu nije ocijenjen ocjenom* 1. 12,5% *studenata je na ispitu ocijenjeno ocjenom najviše jednakom* 2, 31,25% *studenata je na ispitu ocijenjeno ocjenom najviše jednakom* 3, 62,5% *studenata je na ispitu ocijenjeno ocjenom najviše jednakom* 4, *a svi studenti su na ispitu ocijenjeni ocjenom najviše jednakom* 5.

Na potpuno analogan način definiramo *kumulativnu apsolutnu frekvenciju* "*veće od*" svakoga pojedinoga modaliteta To je *ukupan broj elemenata osnovnoga skupa koji su ili strogo veći od dotičnoga modaliteta ili jednaki tom modalitetu*. Tako je kumulativna apsolutna frekvencija "veće od" modaliteta 1 jednaka 80 (svi studenti su ocijenjeni ocjenom 1 ili strogo većom od 1). Kumulativna apsolutna frekvencija "veće od" modaliteta 2 također je jednaka 80 jer su svi studenti ocijenjeni ocjenom 2 ili strogo većom od 2. Kumulativna apsolutna frekvencija "veće od" modaliteta 3 jednaka je 70 (15 studenata je ocijenjeno ocjenom 3, 25 studenata ocjenom 4, a 30 studenata ocjenom 5), kumulativna apsolutna frekvencija "veće od" modaliteta 4 jednaka je 55 (25 studenata ocijenjeno je ocjenom 4, a 30 studenata ocjenom 5) i kumulativna apsolutna frekvencija "veće od" modaliteta 5 jednaka 30 (30 studenata ocijenjeno je ocjenom 5, a niti jedan student nije ocijenjen ocjenom strogo većom od 5).

ocjena	*broj studenata*	*kumulativna frekvencija "veće od"*
1	0	80
2	10	80
3	15	70
4	25	55
5	30	30
ukupno:	80	

Dobivene apsolutne frekvencije objašnjavamo ovako: *Svi studenti su na ispitu ocijenjeni ocjenom* 2 *ili većom od* 2, 70 *studenata je na ispitu ocijenjeno ocjenom* 3 *ili većom od* 3, 55 *studenata je na ispitu ocijenjeno ocjenom* 4 *ili većom od* 4, *a* 30 *studenata je na ispitu ocijenjeno ocjenom* 5. Kraće i matematičkije, tu rečenicu možemo izreći ovako: *Svi studenti su na ispitu ocijenjeni ocjenom najmanje jednakom* 2, 70 *studenata je na ispitu ocijenjeno ocjenom najmanje jednakom* 3, 55 *studenata je na ispitu ocijenjeno ocjenom najmanje jednakom* 4, *a* 30 *studenata je na ispitu ocijenjeno ocjenom* 5.[40]

[40] Mini-tečaj hrvatskoga jezika, lekcija 2: *najmanje jednako x* ili *barem jednako x* je istoznačnica izraza *veći ili jednak x*, odnosno izraza *jednak x ili strogo veći od x.*

Kod izračuna pojedinih kumulativnih apsolutnih frekvencija "veće od" korisno je primijetiti da kumulativna apsolutna frekvencija "veće od" najslabijega (praktično, prvoga) modaliteta nužno mora biti jednaka opsegu statističkoga skupa i ta nam činjenica može biti svojevrsna provjera ispravnosti provedenih izračuna.

Označimo li s f_x apsolutnu frekvenciju, a s $c_x^>$ kumulativnu apsolutnu frekvenciju "veće od" modaliteta x, nije teško vidjeti da vrijedi jednakost:

$$c_x^> = \sum_{y \geq x} f_y = \sum_{y > x} f_y + f_x$$

koja nam omogućuje relativno brzo računanje svake pojedine kumulativne apsolutne frekvencije "veće od". Računamo ih krećući se od *najboljega modaliteta prema najlošijemu* i za svaki pojedini modalitet (osim najboljega čija je kumulativna apsolutna frekvencija "veće od" jednaka odgovarajućoj "običnoj" apsolutnoj frekvenciji) kumulativnu apsolutnu frekvenciju "veće od" izračunamo tako da izračunatoj kumulativnoj apsolutnoj frekvenciji "veće od" prethodnoga modaliteta pribrojimo apsolutnu frekvenciju modaliteta čiju kumulativnu frekvenciju "veće od" želimo izračunati.

Kao i u slučaju kumulativnih frekvencija "manje od", svaku od dobivenih kumulativnih frekvencija "veće od" možemo podijeliti s opsegom statističkoga skupa. Tako se dobije *kumulativni niz relativnih frekvencija* "*veće od*". U promatranom primjeru taj je niz:

ocjena	*broj studenata*	*kumulativna apsolutna frekvencija "veće od"*	*kumulativna relativna frekvencija "veće od"* [%]
1	0	80	100
2	10	80	100
3	15	70	87.5
4	25	55	68.75
5	30	30	37.5
ukupno:	80		

Dobivene kumulativne relativne frekvencije objašnjavamo ovako: *Svi studenti su na ispitu ocijenjeni ocjenom najmanje jednakom* 2, 87,5% *studenata je na ispitu ocijenjeno ocjenom najmanje jednakom* 3, 68,75% *studenata je na ispitu ocijenjeno ocjenom najmanje jednakom* 4, *a* 37,5% *studenata je na ispitu ocijenjeno ocjenom* 5.

2.3.4. *Dijagram "stablo – list"* (*stem – leaf*)

Jedan od načina sažetoga prikaza *kvantitativnih* statističkih podataka, osobito korišten u statističkim računalnim programima, jest *dijagram "stablo – list"*[41]. Ovaj način je osobito prikladan ukoliko želimo vidjeti oblik i razdiobu podataka, a iz njega se vrlo lako mogu izračunati *medijan* i *mod*[42]. Opišimo ukratko o čemu se radi.

Osnovna ideja ovakvoga tipa grupiranja podataka jest zapisati svaki numerički podatak u obliku *x.y*, gdje je *y* bilo koja znamenka dekadskoga brojevnoga sustava, . decimalna točka, a *x* ostatak numeričkoga podatka. U ovakvom zapisu *y* nazivamo *list*, a *x* *stablo*. Ukoliko numerički podatak uopće nema znamenaka iza decimalne točke ili ima barem dvije znamenke iza decimalne točke, list će biti posljednja znamenka broja, a stablo sve ostale znamenke broja poredane redoslijedom u kojemu se pojavljuju u zapisu broja. U takvim slučajevima obavezno moramo dati odgovarajuću "uputu za uporabu", tj. kako iz zapisa *x.y* pravilno očitati polaznu numeričku vrijednost. Ilustrirajmo to na primjerima.

Primjer 7. a) Stablo numeričkoga podatka 2008.7 je broj 2008, a list broj 7. Ovdje nam ne treba nikakva "uputa za uporabu" jer iz činjenica da je 2008 stablo, 7 list i izložene ideje grupiranja podataka odmah dobivamo originalni podatak: 2008.7.

b) Prirodan broj 2008 nema niti jedne znamenke iza decimalne točke. U tom slučaju list je posljednja znamenka toga broja, tj. 8, a stablo sve ostale znamenke broja u redoslijedu u kojemu se pojavljuju u samom broju, tj. 200. No, kad bismo iz ovakvoga zapisa htjeli dobiti originalni podatak, slijedeći osnovnu ideju dobili bismo broj 200.8, što je netočno. Stoga ovdje moramo dati "uputu za uporabu", odnosno reći kako iz podatka *x.y* dobiti originalni podatak. Ta je uputa ovdje jednostavna: *pomnožiti zapis x.y s* 10.

c) Decimalan broj 20.08 ima dvije znamenke iza decimalne točke. I u ovome je slučaju list znamenka 8, a stablo 200. Međutim, iz ovakvoga zapisa slijedi da je originalan podatak 200.8, što je netočno. Opet nedostaje "uputa za uporabu", a u ovom slučaju ona glasi: *podijeliti zapis x.y s* 10.

Nakon što smo svaki element osnovnoga skupa jedinstveno zapisali na gore navedeni način, preostaje nam grupirati elemente osnovnoga skupa prema stablima: u istu grupu svrstat ćemo sve elemente koji imaju isto stablo. Pritom ćemo stabla poredati prema uobičajenom uređaju u skupu realnih brojeva, tj. od najmanjega do najvećega. Pokažimo to na primjerima.

---

[41] Termini *stablo* (engl. *tree*) i *list* (engl. *leaf*) su dio osnovnih termina *teorije grafova*. Sukladno toj teoriji, ovaj je dijagram zapravo *unija* stabala kojima su listovi znamenke jedinica prirodnih brojeva, a korijeni sve ostale znamenke. Također, u duhu te teorije primjereniji prijevod naziva ovoga tipa dijagrama bio bi "*stabljika – list*".

[42] *Medijan* i *mod* su srednje vrijednosti ili mjere centralne tendencije o kojima će detaljnije biti govora u posebnom poglavlju.

Primjer 8. Svi studenti Visoke škole za financijski menadžment u Špičkovini koji su položili kolegij *Poslovna matematika* putem dvaju kolokvija ostvarili su sljedeće brojeve bodova: 64, 67, 66, 95, 72, 90, 94, 95, 76, 86, 76, 55, 80, 58, 64, 68, 81, 72, 76, 87, 80, 93, 92, 63, 93, 99, 59, 78, 63, 72, 61, 69, 58, 82, 51, 70 i 90. Budući da je riječ o dvoznamenkastim prirodnim brojevima, prva znamenka svakoga od tih brojeva bit će stablo, a druga list. Tako dobivamo sljedeću tablicu:

stablo	*list*
5	1,5,8,8,9
6	1,3,3,4,4,6,7,8,9
7	0,2,2,2,6,6,6,8
8	0,0,1,2,6,7
9	0,0,2,3,3,4,5,5,9
uputa za uporabu:	pomnožiti *stablo.list* s 10

Iz ovako navedene tablice se pomoću "upute za uporabu" vrlo lako dobivaju polazni podaci. Npr. u prvom retku tablice zapravo pišu decimalni brojevi 5.1, 5.5, 5.8, 5.8 i 5.9. "Uputa za uporabu" kaže da se originalni podaci dobiju množenjem svakoga od tih brojeva s 10. Dakle, originalni podaci su 51, 55, 58, 58 i 59. Na potpuno analogan način se iz ostalih redaka tablice dobiju ostali elementi osnovnoga skupa.

Primjer 9. Zaključne cijene dionica (u €) mobilnoga operatera "*Bla–bla–mobile* d.d." na bečkoj burzi u 30 slučajno odabranih dana u razdoblju između 31.12.2006. i 31.12.2007. iznosile su: 8.7, 6.9, 11.5, 13.6, 6.8, 11.7, 7.3, 7.4, 14.5, 11.9, 8.2, 12.2, 14.9, 8.1, 12.0, 10.6, 13.2, 13.5, 11.8, 13.4, 13.6, 12.2, 7.2, 13.9, 12.5, 8.8, 9.8, 13.3, 9.9 i 7.0. U ovome će slučaju stablo svakoga podatka biti znamenke ispred decimalne točke, a list znamenka iza decimalne točke. Tako dobivamo sljedeću tablicu:

stablo	*list*
6	8,9
7	0,2,3,4
8	1,2,7,8
9	8,9
10	6
11	5,7,8,9
12	0,2,2,5
13	2,3,4,5,6,6,9
14	5,9

"Uputu za uporabu" ovdje ne moramo navesti jer se originalni podaci dobivaju izravno bez ikakvih dodatnih aritmetičkih operacija. Npr. iz prvoga retka odmah dobivamo podatke 6.8 i 6.9, iz drugoga 7.0,7.2,7.3 i 7.4 itd.

Primjer 10. Zaključne cijene dionica (u kn) tvrtke "*Stara TV*" na frkljevačkoj burzi u veljači 2008. godine iznosile su: 59.52, 59.47, 59.51, 59.58, 59.57, 59.52, 59.49, 59.46, 59.43, 59.42, 59.39, 59.38, 59.37, 59.39, 59.4, 59.41, 59.43, 59.42, 59.44, 59.45, 59.48, 59.5, 59.53, 59.55, 59.58, 59.6, 59.62, 59.63 i 59.62. U ovome će slučaju list svakoga elementa osnovnoga skupa

biti znamenka na mjestu stotinki (druga znamenka iza decimalne točke), a stablo sve ostale znamenke. Tako dobivamo sljedeću tablicu:

stablo	*list*
593	7,8,9,9
594	0,1,2,2,3,3,4,5,6,7,8,9
595	0,1,2,2,3,5,7,8,8
596	0,2,2,3
uputa za uporabu:	podijeliti *stablo.list* s 10

I ovdje pomoću "upute za uporabu" vrlo lako dobivamo polazne podatke. Npr. u prvom retku tablice zapravo pišu decimalni brojevi 593.7, 593.8, 593.9 i 593.9. "Uputa za uporabu" kaže da odgovarajuće originalne podatke dobivamo dijeljenjem svakoga od navedenih decimalnih brojeva s 10, pa se tako dobiju podaci 59.37, 59.38, 59.39 i 59.39. Na potpuno analogan način se iz ostalih redaka dobiju ostali elementi osnovnoga skupa.

2.4. *Tablični prikaz statističkih podataka*

Tablični prikaz statističkih podataka je treća faza obrade "sirovih" statističkih podataka. Cilj te faze jest na jasan, cjelovit i pregledan način prezentirati rezultate prikupljanja i grupiranja podataka. U ovoj ćemo točki ukratko spomenuti neke najvažnije vrste tablica. Pritom istaknimo da svaka statistička tablica nužno mora imati sljedeće dijelove:

1.) naslov tablice;
2.) tekstualni dio tablice;
3.) numerički dio tablice;
4.) izvor podataka.

Točno jedan statistički niz nastao sređivanjem podataka prema modalitetima jednoga obilježja tabelarno se pregledno prikazuje *jednostavnom tablicom*. Više statističkih nizova nastalih sređivanjem podataka prema modalitetima jednoga obilježja pregledno prikazujemo *skupnom tablicom*. Jednostavne i skupne tablice vrlo je jednostavno izraditi i u MS Wordu i MS Excelu, ali se zbog potreba daljnje analize podataka preporučuje izraditi ih u MS Excelu. Pogledajmo po jedan primjer svake od spomenutih vrsta tablica.

Primjer 1. Primjer *jednostavne statističke tablice*:

Razdioba svih zaposlenika tvrtke "Lupeškić d.o.o." iz Zlobina na dan 10.03.2008. prema spolu

spol	*broj zaposlenika*
muški	7
ženski	3
ukupno:	10

Izvor: služba kadrovskih potencijala tvrtke "Lupeškić d.o.o."

Primjer 2. Primjer *skupne statističke tablice*:

Razdioba registriranih cestovnih motornih i priključnih vozila u Republici Hrvatskoj na dan 31.12.2007. prema vrsti vlasnika

vrsta vozila	*fizičke osobe*	*pravne osobe*	*ukupno*
autobusi	701	4 322	**5 023**
kombinirani automobili	6 316	4 592	**10 908**
mopedi	97 845	8 498	**106 343**
motocikli	52 104	4 191	**56 295**
osobna vozila	1 334 894	152 234	**1 487 128**
priključna vozila	19 792	15 118	**34 910**
vrsta vozila	*fizičke osobe*	*pravne osobe*	*ukupno*
radni strojevi	2 587	4 407	**6 994**
teretna i radna vozila	70 888	94 347	**165 235**
traktori	101 854	4 190	**106 044**

Izvor: mjesečno statističko izvješće broj 1/2008 Državnoga zavoda za statistiku Republike Hrvatske

Statistički skup moguće je podijeliti i prema najmanje dvama različitim obilježjima, pa se i tako dobiveni podaci mogu grupirati i sređivati. Dobiveni rezultati pregledno se prikazuju *kombiniranom* ili *kontingencijskom tablicom*. Ukupan broj obilježja prema kojemu je podijeljen skup naziva se *dimenzija tablice*, pa sukladno tome kombinirana tablica može biti dvodimenzionalna, trodimenzionalna, četverodimenzionalna itd. Svaka takva tablica uvijek se sastoji od predstupca, zaglavlja, polja tablice, marginalnoga retka (retka *ukupno*) i marginalnoga stupca (stupca *ukupno*). Ukoliko je statistički skup podijeljen prema točno dvama obilježjima, odgovarajuća kombinirana tablica je dvodimenzionalna, a njezin marginalni redak i marginalni stupac sadrže odgovarajuće apsolutne ili relativne frekvencije svakoga pojedinoga obilježja. Pogledajmo dva primjera tablica s apsolutnim frekvencijama.

Primjer 3. Primjer *dvodimenzionalne kombinirane statističke tablice*:

Razdioba svih studenata 1. godine stručnoga studija računovodstva i financija Veleučilišta u Špičkovini na dan 10.03.2008. prema spolu i ruci kojom pišu

	desna	lijeva	**ukupno:**
muški	12	5	**17**
ženski	17	6	**23**
ukupno:	**29**	**11**	**40**

izvor: studentska služba Veleučilišta u Špičkovini

Apsolutna frekvencija u retku *i* i stupcu *j* označava ukupan broj elemenata statističkoga skupa koje istodobno imaju oblik obilježja 1 naveden u retku *i* i oblik obilježja 2 naveden u retku *j*. Tako npr. apsolutna frekvencija 12 u presjeku drugoga retka i drugoga stupca znači da među svim promatranim studentima ima točno 12 studenata muškoga spola koji pišu desnom rukom

("dešnjaci"[43]), apsolutna frekvencija 6 u presjeku trećega retka i trećega stupca znači da među svim promatranim studentima ima točno 12 studentica koji pišu lijevom rukom ("ljevakinje"[44]) itd. Iz presjeka marginalnoga stupca i drugoga retka očitavamo da među svim promatranim studentima ima točno 17 studenata muškoga spola, a iz presjeka marginalnoga retka i trećega stupca očitavamo da među svim promatranim studentima ima točno 11 studenata koji pišu lijevom rukom. Za vježbu objasnite svaku od preostalih frekvencija navedenih u tablici.

Primjer 4. Primjer *četverodimenzionalne kombinirane statističke tablice*:

Razdioba svih zaposlenika tvrtke "*Viktor Stečaj*" iz Plomina na dan 31.12.2006. prema spolu, stupnju stručne spreme, navršenim godinama radnoga staža i iznosu prosječne mjesečne plaće

		3000-4999		5000-6999		7000-9000	*ukupno*
		srednja	viša	viša	visoka	visoka	
0-9	M	3	1	1	0	0	**5**
	Ž	3	1	0	2	1	**7**
10-19	M	3	1	3	3	3	**13**
	Ž	0	0	4	0	2	**6**
20-30	M	3	0	2	0	0	**5**
	Ž	3	0	0	0	1	**4**
ukupno		**15**	**3**	**10**	**5**	**7**	**40**

izvor: služba kadrovskih potencijala tvrtke "*Viktor Stečaj*"

Ovdje prva apsolutna frekvencija, tj. 3, znači da među svim promatranim zaposlenicima postoje točno 3 zaposlenika muškoga spola koji imaju srednju stručnu spremu, najviše 9 godina radnoga staža i prosječnu mjesečnu plaću između 3.000,00 kn i 5.000,00 (nećemo cjepidlačiti) kn. Posljednja apsolutna frekvencija, tj. 1, znači da među svim promatranim zaposlenicima postoji točno jedna zaposlenica (ženskoga spola) koja ima visoku stručnu spremu, između 20 i 30 godina radnoga staža, te prosječnu mjesečnu plaću između 7.000,00 i 9.000,00 kn. Frekvencija 13 iz marginalnoga stupca znači da među svim zaposlenicima *koji imaju između* 10 *i* 20 *godina radnoga staža* ima točno 13 zaposlenika muškoga spola, dok frekvencija 10 iz marginalnoga retka znači da među svim zaposlenicima *čija je prosječna mjesečna plaća između* 5.000,00 kn *i* 7.000,00 (opet ne cjepidlačimo) kn ima točno 10 zaposlenika koji imaju višu stručnu spremu itd. Iako te frekvencije nisu navedene, iz dobivene kombinirane tablice vrlo lako možemo izračunati koliko je žena među svim promatranim zaposlenicima, koliko ukupno zaposlenika ima visoku stručnu spremu i sl. Učinite to za vježbu i objasnite značenje svih preostalih apsolutnih frekvencija navedenih u tablici.

[43] Uobičajeni izraz za muškarca koji piše desnom rukom. Taj izraz treba razlikovati od izraza *desničar* koji se odnosi isključivo na političku orijentaciju. Također, muškarac koji piše lijevom rukom nije niti *ljevičar* niti *lijevak*, nego *ljevak*.

[44] Uobičajeni izraz za ženu koja piše lijevom rukom. Taj izraz treba razlikovati od izraza *ljevičarka* koji se odnosi isključivo na političku orijentaciju.

Kako je istaknuto, osim apsolutnih, kombinirana tablice može sadržavati i relativne frekvencije. Ovdje ćemo izdvojiti i na primjerima ilustrirati tri tipa takvih tablica: *tablica "kutno* 100", *tablica "vodoravno* 100" i *tablica "okomito* 100".

Tablica "kutno 100" dobije se iz kombinirane tablice koja sadrži apsolutne frekvencije tako da se svaka apsolutna frekvencija podijeli s *opsegom statističkoga skupa*. Budući da je zbroj svih apsolutnih frekvencija uvijek jednak opsegu statističkoga skupa, to zbroj svih relativnih frekvencija mora biti jednak 1 (ili 100%, izrazimo li relativne frekvencije u postotcima). Taj zbroj možemo izračunati na dva načina[45]:

1.) zbrajajući relativne frekvencije u svakom pojedinom retku, pa potom zbrojiti sve dobivene zbrojeve;
2.) zbrajajući relativne frekvencije u svakom pojedinom stupcu, pa potom zbrojiti sve dobivene zbrojeve.

Primjer 5. Primjer *tablice "kutno* 100":

Struktura svih umrlih nasilnom smrću u Republici Hrvatskoj u 2005. godini prema spolu i vrsti nasilne smrti

	žene	muškarci	*ukupno*
nesretni slučaj	23.56%	43.40%	66.96%
samoubojstvo	7.75%	22.65%	30.40%
ubojstvo	0.87%	1.18%	2.05%
posljedice ratnih operacija	0.00%	0.17%	0.17%
nepoznata nasilna smrt	0.07%	0.35%	0.42%
ukupno	32.24%	67.76%	**100,00%**

Izvor: "Žene i muškarci u Hrvatskoj", Državni zavod za statistiku, 2007.

Iz ove tablice, primjerice, možemo očitati da su *među svim umrlima nasilnom smrću u Republici Hrvatskoj u* 2005. *godini njih* 7.75% *žene koje su izvršile samoubojstvo*. Za vježbu objasnite svaku od ostalih relativnih frekvencija.

Tablica "vodoravno 100" dobije se iz kombinirane tablice s apsolutnim frekvencijama tako da *svaki redak* te tablice shvatimo kao tablični prikaz *uzorka* statističkoga skupa. Tada relativne frekvencije računamo tako da u svakom pojedinom retku svaku od apsolutnih frekvencija podijelimo s *opsegom* odgovarajućega *uzorka*.

Primjer 6. Primjer *tablice "vodoravno* 100":

Struktura stanovništva Republike Hrvatske (popis iz 2001. godine) prema spolu, razini obrazovanja i navršenim godinama starosti

[45] Navedeni načini posljedica su svojstva komutativnosti i asocijativnosti zbrajanja realnih brojeva koja zajednički kažu da zbroj proizvoljno (ali konačno) mnogo realnih brojeva ne ovisi o poretku tih brojeva, te da unutar toga zbroja možemo postaviti parove zagrada na koja god mjesta želimo, pa posebno izračunati zbroj unutar svakoga pojedinoga para zagrada.

		osnovno ili manje	srednje	više ili visoko	ukupno:
15 - 24	žene	41.00%	56.50%	2.50%	**100.00%**
	muškarci	42.20%	56.30%	1.50%	**100.00%**
25 - 34	žene	18.30%	61.90%	19.80%	**100.00%**
	muškarci	17.70%	68.60%	13.70%	**100.00%**
35 - 44	žene	27.40%	55.10%	17.50%	**100.00%**
	muškarci	20.10%	64.00%	15.90%	**100.00%**
45 - 54	žene	41.60%	42.40%	16.00%	**100.00%**
	muškarci	25.30%	57.60%	17.10%	**100.00%**
55 - 64	žene	65.70%	24.10%	10.20%	**100.00%**
	muškarci	37.30%	46.20%	16.50%	**100.00%**
65 - 74	žene	81.70%	13.60%	4.70%	**100.00%**
	muškarci	52.50%	34.90%	12.60%	**100.00%**
75 - 84	žene	86.30%	10.90%	2.80%	**100.00%**
	muškarci	61.40%	25.20%	13.40%	**100.00%**
85 +	žene	88.60%	9.00%	2.40%	**100.00%**
	muškarci	68.50%	22.70%	8.80%	**100.00%**

izvor: "Žene i muškarci u Hrvatskoj", Državni zavod za statistiku, 2007.

Uočimo da ovdje zapravo imamo 16 *uzoraka*: prvi uzorak tvore sve žene koje su navršile između 15 i 24 godine, drugi uzorak tvore svi muškarci koji su navršili između 15 i 24 godine, treći uzorak tvore sve žene koje su navršile između 25 i 34 godine itd. Granice razreda u ovom su primjeru nominalne, tj. riječ je o segmentima [15, 24], [25, 34] itd.

Tako, primjerice, relativna frekvencija 27.40% u drugom stupcu znači da *među svim ženama koje su navršile između* 35 *i* 44 *godine njih* 27,4% *ima osnovno obrazovanje ili manje*, a relativna frekvencija 1,50% u četvrtom stupcu znači da *među svim muškarcima koji su navršili između* 15 *i* 24 *godine njih* 1,50% *ima više ili visoko obrazovanje*. Za vježbu objasnite svaku od ostalih relativnih frekvencija.

Tablica "okomito 100" dobije se iz kombinirane tablice s apsolutnim frekvencijama tako da *svaki stupac* te tablice shvatimo kao tablični prikaz *uzorka* statističkoga skupa. Tada relativne frekvencije računamo tako da u svakom pojedinom stupcu svaku od apsolutnih frekvencija podijelimo s *opsegom* odgovarajućega *uzorka*.

Primjer 7. Primjer *tablice "okomito* 100":

Struktura svih državljana Republike Hrvatske koji su titulu *doktor znanosti* stekli 2005. godine prema spolu i području znanosti iz kojega su obranili doktorsku disertaciju

	žene	muškarci
prirodne znanosti	34.48%	10.90%
tehničke znanosti	8.62%	27.96%
biomedicina i zdravstvo	20.69%	24.64%
biotehničke znanosti	9.77%	9.48%
društvene znanosti	12.64%	12.32%
humanističke znanosti	13.79%	14.69%
ukupno:	**100.00%**	**100.00%**

izvor: "Žene i muškarci u Republici Hrvatskoj", Državni zavod za statistiku, 2007.

Uočimo da ovdje zapravo imamo dva *uzorka*: prvi uzorak tvore sve žene koje su titulu dr.sc. stekle 2005. godine, a drugi uzorak tvore svi muškarci koji su titulu dr.sc. stekli 2005. godine. Tako, primjerice, relativna frekvencija 20,69% navedena u drugom stupcu znači da je *među svim ženama koje su titulu dr.sc. stekle* 2005. *godine njih* 20.69% *tu titulu steklo na području biomedicine i zdravstva*, a relativna frekvencija 10.90% navedena u trećem stupcu znači da je *među svim muškarcima koji su titulu dr.sc. stekli* 2005. *godine njih* 10.9% *tu titulu steklo na području prirodnih znanosti*. Za vježbu objasnite svaku od ostalih relativnih frekvencija.

2.5. *Grafički prikaz statističkih podataka*

Grafički prikaz statističkih podataka je jedan od najvažnijih dijelova *prezentacije* rezultata statističkoga istraživanja. Naime, kako je istaknuto u uvodu, statistika se, između ostaloga, bavi i organiziranim načinima prezentacije rezultata statističkoga istraživanja. Psihologija ljudskoga ponašanja pokazuje da se u gotovo svim vrstama tiskovina, javnih medija (osim radija), izlaganja na različitim vrstama znanstvenih/stručnih skupova, ali i izlaganja na svim vrstama nastave u različitim tipovima škola (osnovne, srednje, visoke, fakulteti, …) pozornost slušatelja najprije i najvećim dijelom usmjerava na grafičke prikaze podataka. Stoga ćemo u ovoj točki ukratko navesti osnovne i najčešće korištene statističke prikaze podataka ilustrirajući ih na primjerima. Ipak, odmah istaknimo da tablice i grafički prikazi pružaju isključivo *prve informacije* o promatranoj pojavi, a ni u kojemu slučaju nisu (potpuno) dovoljne za statističku analizu i interpretaciju njezinih rezultata (5. faza statističkoga istraživanja).

Grafičke prikaze ili, kraće, *grafikone* obično dijelimo u četiri osnovne skupine:

1.) *<u>površinski grafikoni</u>*;
2.) *<u>linijski grafikoni</u>*;
3.) *<u>točkasti grafikoni</u>*;
4.) *<u>kartogrami</u>*

Kako bi grafički prikaz imao svoj potpuni smisao, osobito treba istaknuti da uz svaki grafikon obavezno treba navesti *odgovarajuće oznake*: naslov grafikona (najčešće: "Razdioba…", "Struktura…" itd.), nazive kategorija na koordinatnim osima (npr. *godina*, *broj zaposlenika*,

vrsta stručne spreme itd.), legendu (dio grafikona koji dodatno pojašnjava osobitosti grafikona) itd. Bez spomenutih oznaka grafikon je nepotpun i "nečitljiv", pa ni u kojemu slučaju ne može predstavljati dobar prikaz rezultata statističke analize.

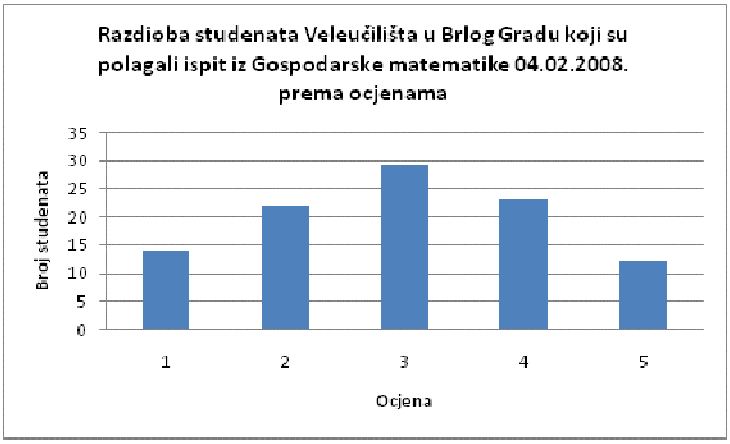
Ukratko i uz odgovarajuće primjere opisat ćemo najčešće predstavnike svake pojedine vrste grafikona.

2.5.1. *Površinski grafikoni*

Osnovna karakteristika ove vrste grafikona jest da se frekvencije podudaraju s *površinama geometrijskih likova* (pravokutnik, krug). Drugim riječima, apsolutna (ili relativna) frekvencija svakoga modaliteta jednaka je *površini* tom modalitetu odgovarajućega pravokutnika ili dijela kruga. Budući da je površina pravokutnika jednaka umnošku njegove duljine a i širine b, za širinu se najčešće uzima $b = 1$ jed., pa je u tom slučaju apsolutna frekvencija identički jednaka duljini pravokutnika. U slučaju strukturnoga kruga i prikaza relativnih frekvencija površina *kružnoga isječka* identički je jednaka relativnoj frekvenciji tom isječku odgovarajućega modaliteta, a krug može imati *bilo koji* polumjer (a ne nužno polumjer $r = 1$ jed.). Krug – kojega možemo zamišljati i kao kut od 360° – dijelimo na kružne isječke čiji su središnji kutovi $360° \cdot r_i$, gdje je r_i relativna frekvencija i – toga modaliteta iskazana kao proporcija (tj. kao decimalan broj). Npr. relativnoj frekvenciji 33.3333% odgovara središnji kut $360° \cdot 0{,}333333 = 120°$ itd.

Najčešći i praktično najviše korišteni tipovi površinskih grafikona su različite vrste stupaca (jednostavni, položeni, dvostruki, višestruki, strukturni), strukturni krug, Paretov dijagram i histogram. O histogramu ćemo nešto više reći u posebnoj podtočki, a ovdje samo istaknimo da se nabrojene vrste stupaca, te strukturni krug najčešće koriste za grafički prikaz *kvalitativnih* obilježja.

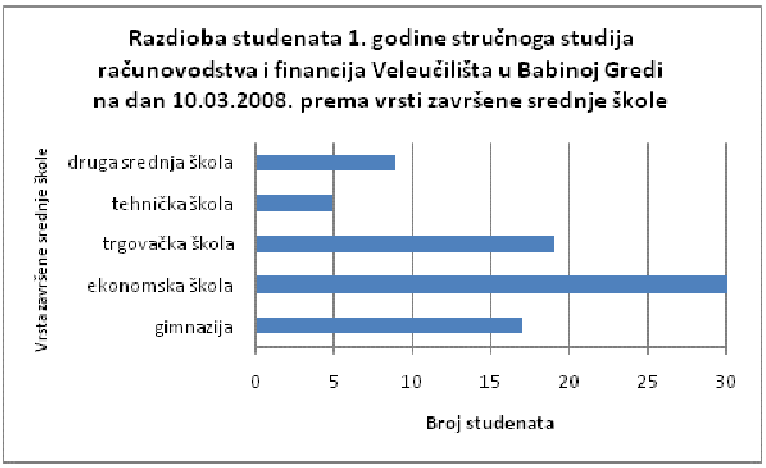
<u>Primjer 1.</u> Primjer *<u>jednostavnih stupaca</u>*:



Uočimo odmah da svi stupci imaju jednaku širinu i ta je širina *jedinična* (tj. b = 1 jed.). Stoga je površina svakoga pravokutnika identički jednaka duljini pravokutnika, što znači da je i apsolutna frekvencija svakoga od pet modaliteta identički jednaka duljini pripadnoga pravokutnika.

Iz zadanoga grafikona ne možemo očitati točne apsolutne frekvencije svakoga pojedinoga modaliteta, ali možemo zaključiti da je najveći broj promatranih studenata na ispitu ocijenjen ocjenom 3, a najmanji broj promatranih studenata ocijenjen ocjenom 5.

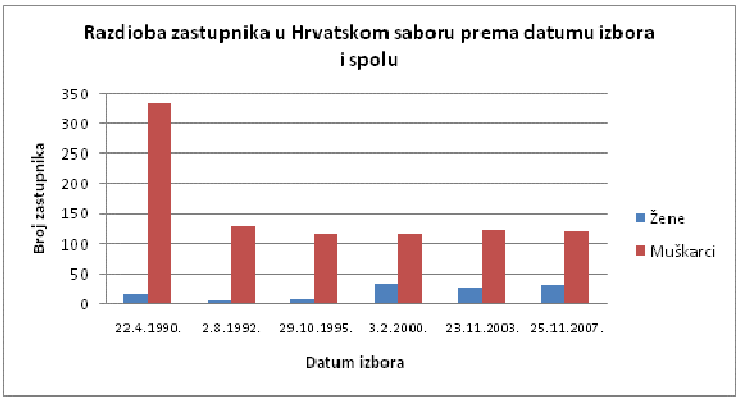
Primjer 2. Primjer *položenih stupaca*:



Položene stupce možemo shvatiti kao jednostavne stupce zarotirane za 90° ulijevo, pa potom zrcaljene s obzirom na os modaliteta. (Na ovom načelu se konstruiraju položeni stupci u MS Excelu.) Oni imaju potpuno ista svojstva kao i jednostavni stupci, a koriste se najčešće kod obilježja s nomenklaturama.

Niti iz ovoga grafikona ne možemo "očitati" točnu apsolutnu frekvenciju *svakoga* pojedinoga modaliteta (za dva modaliteta to možemo učiniti, ali za preostala tri ne), ali možemo zaključiti da je najviše promatranih studenata završilo ekonomsku školu, a najmanje tehničku školu.

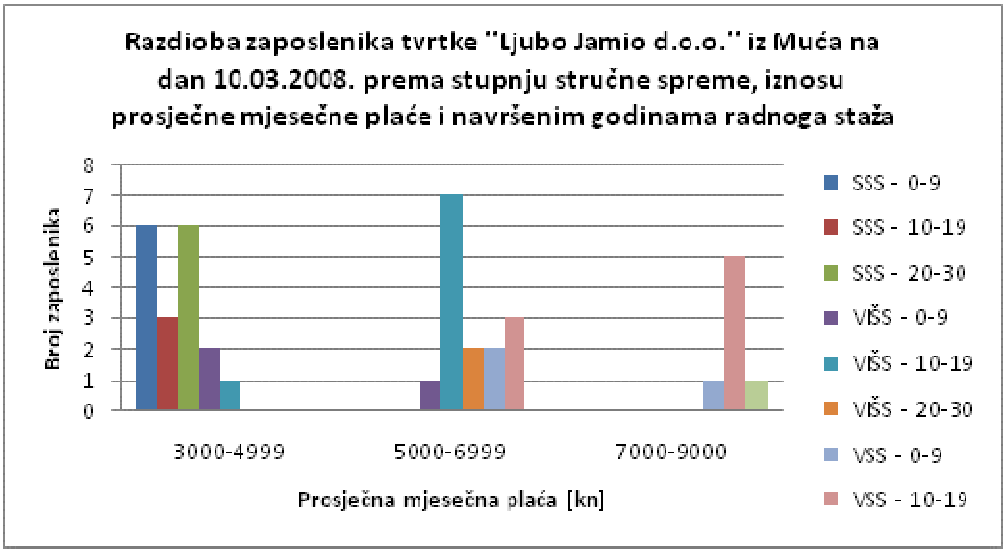
Primjer 3. Primjer *dvostrukih stupaca*:



Uočimo da se ovdje zapravo radi o podjeli statističkoga skupa (svi dosadašnji zastupnici u Hrvatskom saboru) na 6 međusobno nedisjunktnih uzoraka (dio zastupnika je biran u barem dva saziva Hrvatskoga sabora), pa je potom svaki od uzoraka podijeljen prema obilježju *spol*. U ovom je slučaju tumač (*legenda*) prikazana sa strane neizostavan dio grafikona jer bez nje ne bismo mogli točno znati (već samo prilično točno naslutiti) površine kojih pravokutnika označavaju apsolutne frekvencije modaliteta *muškarci*, a koji modaliteta *žene*.[46]

Iz zadanoga grafikona, primjerice, možemo zaključiti jedino da je u svim dosadašnjim sazivima Hrvatskoga sabora ukupan broj saborskih zastupnika bio znatno veći od ukupnoga broja saborskih zastupnica, ali ne možemo reći npr. za koliko posto veći, koliko puta veći itd. Zbog toga su za daljnju analizu primjereniji strukturni stupci (vidjeti Primjer 6.)

<u>Primjer 4.</u> Primjer <u>*višestrukih stupaca*</u>:



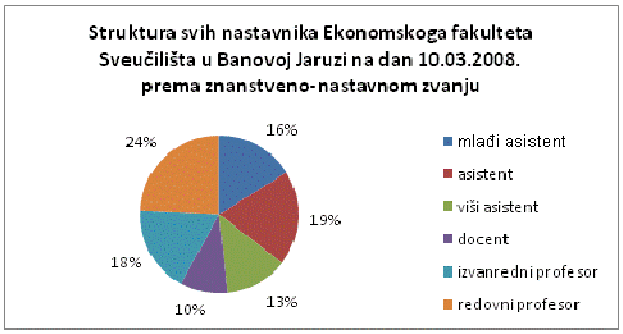
Ovaj grafikon prikazuje podjelu istoga statističkoga skupa (svi zaposlenici tvrtke "Ljubo Jamio" iz Mića na dan 10.03.2008.) prema trima različitim obilježjima: stupnju stručne spreme, iznosu prosječne mjesečne plaće i navršenim godinama radnoga staža. Potonja dva obilježja su kvantitativna, te su grupirana u razrede s nominalnim granicama (neprikladno za daljnju statističku analizu, ali prikladno za grafički prikaz). I ovdje je tumač sa strane nužan kako bismo znali koja se boja odnosi na svaku kategoriju. Svi stupci imaju širinu jednaku 1, pa je, kao i kod jednostavnih stupaca, duljina pravokutnika jednaka apsolutnoj frekvenciji.

Iz prikazanoga grafikona možemo izvući razne zaključke, npr. da postoji 6 zaposlenika koji imaju srednju stručnu spremu, najviše 9 godina radnoga staža i prosječnu mjesečnu plaću između 3.000,00 i 4.999,00 kn, da od svih zaposlenika koji imaju visoku stručnu spremu, te između 10 i 19 godina radnoga staža najviše njih ima plaću između 7.000,00 i 9.000,00 kn itd. Za vježbu objasnite svaki prikazani stupac, te pokušajte izvući još neke zaključke iz ovoga

[46] Strogo precizno, u legendi bi trebali pisati točni nazivi modaliteta obilježja *spol*, tj. nazivi *muški* i *ženski*. No, ponekad se radi lakše interpretacije grafikona ti modaliteti zamjenjuju imenicama *muškarci* i *žene*.

prikaza (npr. što sve možete reći o zaposlenicima koji imaju srednju stručnu spremu, što o zaposlenicima čija je plaća između 5.000,00 i 6.999,00 kn itd.).

<u>Primjer 5.</u> Primjer *<u>strukturnoga kruga</u>*:



Na ovom grafikonu površina svakoga pojedinoga kružnoga isječka odgovara relativnoj frekvenciji modaliteta kojega predstavlja taj isječak. Kako je već istaknuto, duljina polumjera kruga može biti bilo koji strogo pozitivan realan broj, a središnji kutovi odgovarajućih isječaka dobiveni su množenjem broja 360 redom s brojevima 0.24, 0.16, 0.19, 0.13, 0.1 i 0.18, pa je tako površina narančastoga kružnoga isječka jednaka 24% površine cijeloga kruga, površina tamnoplavoga kružnoga isječka jednaka 16% površine cijeloga kruga itd. Napomenimo da neki programi (MS Excel) omogućuju dobivanje strukturnoga kruga na osnovu jednostavne tablice s apsolutnim frekvencijama, pa se time izbjegava efektivno računanje relativnih frekvencija.

Iz navedenoga strukturnoga kruga možemo "očitati" da *među svim nastavnicima Ekonomskoga fakulteta Sveučilišta u Banovoj Jaruzi ima* 16% *mlađih asistenata*, 24% *redovnih profesora* itd. Za vježbu objasnite značenje svake preostale relativne frekvencije.

<u>Primjer 6.</u> Primjer *<u>razdijeljenih stupaca</u>* (dobivenih iz odgovarajuće tablice "okomito 100"):

Širina svakoga stupca jednaka je 1 jed., a duljina 100 jed., pa je njegova površina 100 kv. jed. Ta je površina jednostavnim računom diobe podijeljena *u omjeru* relativnih frekvencija, pa se za svaku pojedinu relativnu frekvenciju određuje pripadni dio površine pravokutnika. Pritom poredak modaliteta *nije bitan*, tj. ako imamo ukupno n relativnih frekvencija, jedan razdijeljeni stupac možemo "nacrtati" na $n! = 1 \cdot 2 \cdot \ldots \cdot n$ načina. Radi lakše usporedbe, u svakomu je razdijeljenom stupcu jednak *redoslijed* pojedinih modaliteta.

Iz ovako prikazanih stupaca možemo izvući razne zaključke, npr. da *među svim zaposlenim ženama u Republici Hrvatskoj u* 2006. *godini najviše* (točnije, 55.9%) *njih je sa srednjim obrazovanjem*, *a najmanje* (točnije, 21.8%) *s osnovnim ili manjim obrazovanjem*, da *među svim nezaposlenim muškarcima u Republici Hrvatskoj u* 2006. *godini najviše* (točnije, 62.7%) *njih je sa srednjim obrazovanjem*, *a najmanje* (točnije, 6.4%) *s višim ili visokim obrazovanjem* itd. Za vježbu, objasnite svaku pojedinu relativnu frekvenciju, pa iz navedenoga grafikona pokušajte izvući još neke zaključke.

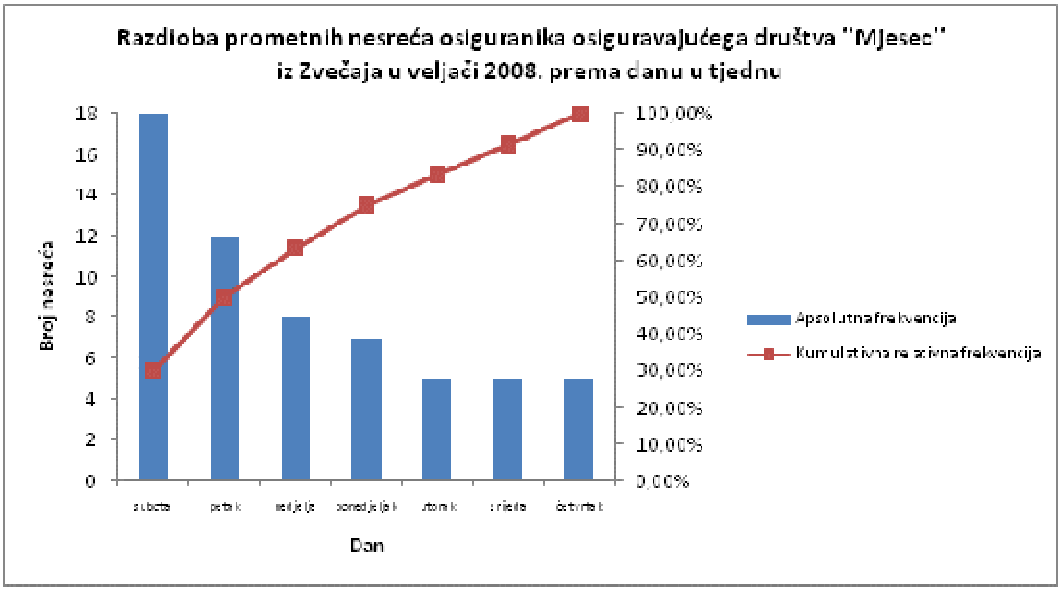
<u>Primjer 7.</u> Primjer *<u>položenih razdijeljenih stupaca</u>* (dobivenih iz odgovarajuće tablice "vodoravno 100"):

Položeni razdijeljeni stupci su vrlo prikladni za prikaz strukture statističkoga skupa pri podjeli prema kvalitativnom obilježju s nomenklaturom. Strogo formalno rečeno, dobivaju se rotiranjem "običnih" razdijeljenih stupaca za 90° ulijevo, pa zrcaljenjem s obzirom na os modaliteta. Širina svakoga pojedinoga pravokutnika jednaka je 1 jed., a duljina 100 jed., pa je površina svakoga pravokutnika 100 kv. jed., a dijelovi koji odgovaraju pojedinim relativnim frekvencijama dobivaju se pomoću dijeljenjem te površine jednostavnim računom diobe u omjeru svih zadanih relativnih frekvencija.

Ovakav se grafikon često konstruira tako da se odabere jedan modalitet (u ovom je slučaju to modalitet *žene*), te se njegove relativne frekvencije u svakoj pojedinoj kategoriji tablice "vodoravno 100" poredaju od najveće do najmanje. U tome se poretku slažu i odgovarajući stupci. Naravno, u slučajevima kad se u stupcima tablice "vodoravno 100" nalaze točno dva modaliteta, takvo je slaganje stupaca ekvivalentno slaganju stupaca prema rastućem poretku (od najmanje do najveće) relativnih frekvencija drugoga modaliteta. Odmah treba napomenuti da se ovakva konstrukcija provodi isključivo radi lakše mogućnosti usporedbe relativnih frekvencija istoga modaliteta u različitim kategorijama, ali ni u kojemu slučaju ne predstavlja pravilo prema kojemu treba slagati stupce.

Iz zadanoga grafikona mogu se izvući razni zaključci, npr. da *u privatnim kućanstvima sa zaposlenim osobljem ima* 14 *puta više zaposlenih žena nego muškaraca*, *među svim zaposlenicima u financijskom posredovanju njih* 70% *su žene* itd. Za vježbu, objasnite značenje svake pojedine relativne frekvencije, te pokušajte izvući još nekoliko zaključaka iz prikazanoga grafikona.

<u>Primjer 8.</u> Primjer *<u>Paretova[47] dijagrama</u>*:



[47] Vilfredo Federico Damaso Pareto, francuski sociolog, ekonomist i filozof, porijeklom Talijan.

Za crtanje Paretova dijagrama treba poredati apsolutne frekvencije modaliteta u silaznom poretku, tj. od najveće do najmanje, te izračunati relativne frekvencije i formirati odgovarajući kumulativni niz relativnih frekvencija. Formiranje toga niza potpuno je jednako formiranju kumulativnoga niza "manje od", ali zbog kvalitativnoga obilježja *dan* riječi "manje od" ovdje treba izostaviti. Ovaj je dijagram svojevrsna "kombinacija" površinskoga grafikona (jednostavnih stupaca) i linijskoga grafikona (kojim se prikazuje kumulativni niz relativnih frekvencija), a specifičnost mu je što se na istom grafikonu primjenjuju dva tipa mjerila: prvo mjerilo odnosi se na prikaz apsolutnih frekvencija jednostavnim stupcima, a drugo na prikaz kumulativnih relativnih frekvencija. Istaknimo odmah da primjena dvaju ili više tipova mjerila na istom grafikonu u pravilu nije dozvoljena.

Iz Paretova je dijagrama vrlo lako odrediti *mod*[48] kvalitativnoga niza, odnosno modalitet koji se najčešće pojavljuje u osnovnom skupu. To je jednostavno modalitet koji odgovara prvom stupcu u Paretovu grafikonu, a lako je vidjeti i postoji li još koji modalitet s istom apsolutnom frekvencijom kao i prvi (prema načelima konstrukcije Paretova dijagrama, on se nužno mora nalaziti neposredno iza prvoga modaliteta). U ovome je slučaju najčešći modalitet *subota*, pa možemo zaključiti da se *najviše promatranih prometnih nesreća dogodilo subotom*, te da se *najmanje promatranih prometnih nesreća dogodilo sredinom tjedna* (tj. utorkom, srijedom ili četvrtkom) Također, iz linijskoga grafikona kumulativnih relativnih frekvencija možemo "očitati" da se 30% *svih promatranih prometnih nesreća dogodilo subotom*, *polovica svih prometnih nesreća petkom ili subotom* itd.

Istaknimo da se Paretov dijagram u MS Excelu vrlo lako može dobiti primjenom procedure *Analiza podataka* (*Data Analysis*), ali je nužan preduvjet za primjenu te procedure da modaliteti obilježja budu numerički. Stoga, ukoliko je obilježje kvalitativno, a modaliteti iskazani riječima, slovima itd., najprije treba provesti "kodiranje" modaliteta[49] (zamjenu riječi brojevima), a nakon provedbe cijele procedure "dekodirati" brojeve, tj. vratiti originalne modalitete.

2.5.2. *Histogram*

Histogram je posebna vrsta površinskoga stupčastoga grafikona koja se najčešće koristi za *prikaz modalitetâ kvantitativnih obilježja grupiranih u razrede*. Osnovna ideja konstruiranja histograma je potpuno jednaka ideji izloženoj kod opisa stupčastih grafikona navedenih u točki 2.5.1.: apsolutne frekvencije modaliteta trebaju biti jednake površini tom modalitetu odgovarajućega pravokutnika. Kod kvalitativnih i negrupiranih kvantitativnih obilježja taj je postupak bio relativno jednostavan: pretpostavili smo da su širine svih pravokutnika jednake 1 jed., pa je apsolutna frekvencija zbog toga bila jednaka duljini pravokutnika. No, kod grupiranja kvantitativnih podataka u razrede cijela se priča malo komplicira. Radi jednostavnosti, pretpostavit ćemo da histogramom želimo prikazati isključivo apsolutne

[48] O modu će više riječi biti u 3. poglavlju.
[49] Usporediti s bilješkom br. 13.

frekvencije jer za prikaz relativnih frekvencija obično koristimo strukturne stupce opisane u prethodnoj točki.

Prvo ćemo opisati najlakši slučaj, a taj je kad su modaliteti kvantitativnoga obilježja grupirani u *prave razrede jednakih širina*. Označimo navedenu jednaku širinu s d. To znači da će širina svakoga pravokutnika čija će površina predstavljati apsolutnu frekvenciju razreda biti jednaka d. Da bismo nacrtali pravokutnik, potreban nam je podatak o njegovoj duljini. Njegovu duljinu a (za svaki pojedini razred) jednostavno dobijemo tako da *apsolutnu frekvenciju* dotičnoga *razreda podijelimo brojem d*. To ujedno znači da nam duljina pravokutnika a općenito više neće biti jednaka apsolutnoj frekvenciji (osim u slučaju $d = 1$), nego d puta manja od nje. Zbog toga budite oprezni: kod histograma **duljina pravokutnika općenito ne mora biti jednaka apsolutnoj frekvenciji tom pravokutniku odgovarajućega razreda**, nego može biti nekoliko puta veća ili manja od nje (ovisno o tome je li $0 < d < 1$ ili $d > 1$). Istaknimo da se duljina a izračunata na opisani način naziva *korigirana apsolutna frekvencija*. Primijetimo još jednu važnu činjenicu: budući da apsolutnu frekvenciju dijelimo strogo pozitivnim realnim brojem, razred s najvećom apsolutnom frekvencijom imat će najveću korigiranu apsolutnu frekvenciju. Drugim riječima, možemo *uspoređivati* korigirane apsolutne frekvencije pojedinih razreda (ali ne i "*očitavati*" apsolutne frekvencije) kako smo to činili dosad s apsolutnim frekvencijama.

Malo složeniji je slučaj kad su podaci grupirani u prave razrede različitih širina. Tada za svaki pojedini razred moramo izračunati odgovarajuću širinu i taj će strogo pozitivan realan broj ujedno biti širina pripadnoga pravokutnika. Potom računamo korigiranu apsolutnu frekvenciju kao količnik apsolutne frekvencije i upravo izračunate širine pravokutnika. Iako se naizgled čini da je ovaj slučaj samo poopćenje prethodnoga, to nije tako: naime, ovdje **nema smisla** uspoređivati korigirane apsolutne frekvencije jer pravokutnici nemaju jednaku širinu. Drugim riječima, *ne vrijedi* pravilo da najveća korigirana apsolutna frekvencija odgovara razredu koji ima najveću apsolutnu frekvenciju. Stoga oprez: jedino što možemo napraviti jest uspoređivati površine pravokutnika, a to već zahtijeva točno "očitanje" duljina stranica svakoga pravokutnika, što uz nedovoljno "gusto" mjerilo na koordinatnim osima ne mora biti moguće izvesti.

Najsloženiji je slučaj kad su podaci grupirani u nominalne razrede (što se općenito događa kod diskretnih obilježja) ili ako je barem jedan od razreda otvoren. Da bismo dobili histogram, nominalne razrede najprije moramo pretvoriti u prave[50], definirati nedostajuće donje ili gornje granice, pa potom primijeniti postupak ili iz prvoga ili iz drugoga opisanoga slučaja. Niti u ovom slučaju nema smisla uspoređivati korigirane apsolutne frekvencije jer pravokutnici općenito ne moraju imati jednaku širinu.

U nastavku ćemo dati nekoliko primjera histograma dobivenih pomoću MS Excela, te ukazati na posebnosti izradbe histograma pomoću ovoga programa.

---

[50] Vidjeti Primjer 5. u točki 2.3.2.

Primjer 9. Primjer *histograma* (pravi razredi jednakih širina):

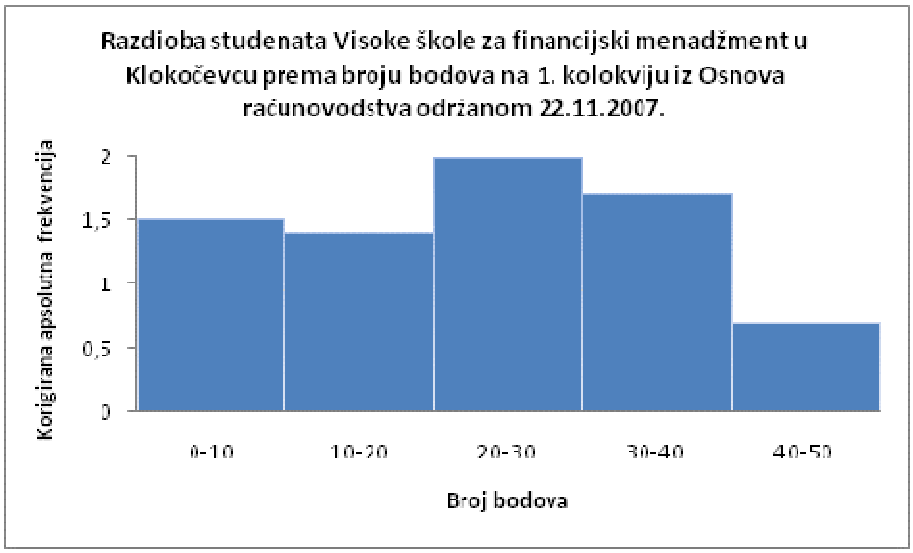
Pretpostavimo da smo brojeve bodova na 1. kolokviju iz kolegija *Osnove računovodstva* održanom 22.11.2007. koje su postigli studenti 1. godine stručnoga studija računovodstva i financija Visoke škole za financijski menadžment u Klokočevcu grupirali u sljedeće razrede:

broj bodova	*broj studenata*
0-10	15
10-20	14
20-30	20
30-40	17
40-50	7

Odmah vidimo da svaki razred ima širinu 10, što znači da je riječ o pravim razredima jednakih širina. Izračunajmo korigirane apsolutne frekvencije podijelivši svaku pojedinu apsolutnu frekvenciju s 10:

broj bodova	*korigirana apsolutna frekvencija*
0-10	1,5
10-20	1,4
20-30	2
30-40	1,7
40-50	0,7

Tako histogram u ovom primjeru izgleda ovako:



Uočimo odmah da natpis podataka na osi *y* nije *Broj studenata* (kao u slučaju jednostavnih stupaca), nego *Korigirana apsolutna frekvencija*. Istaknimo još jednom da je broj studenata, odnosno apsolutna frekvencija pojedinoga razreda jednaka *površini* tom razredu odgovarajućega pravokutnika, pa nema nikakvoga smisla podatke na osi frekvencija nazvati *apsolutne frekvencije* (to bi značilo da je 1.5 studenata na kolokviju ostvarilo najviše 10 bodova, što je besmisleno). Iako mjerilo na osi frekvencija nije dovoljno precizno za izračun apsolutnih frekvencija svakoga pojedinoga razreda, ipak možemo zaključiti da je *najviše studenata* (njih ukupno $2 \cdot 10 = 20$) *na kolokviju ostvarilo između* 20 *i* 30 *bodova*, *a najmanje studenata ostvarilo je između* 40 *i* 50 *bodova*, što se potpuno slaže s podacima iz originalne tablice.

Primjer 10. Primjer *histograma* (pravi razredi različitih širina):

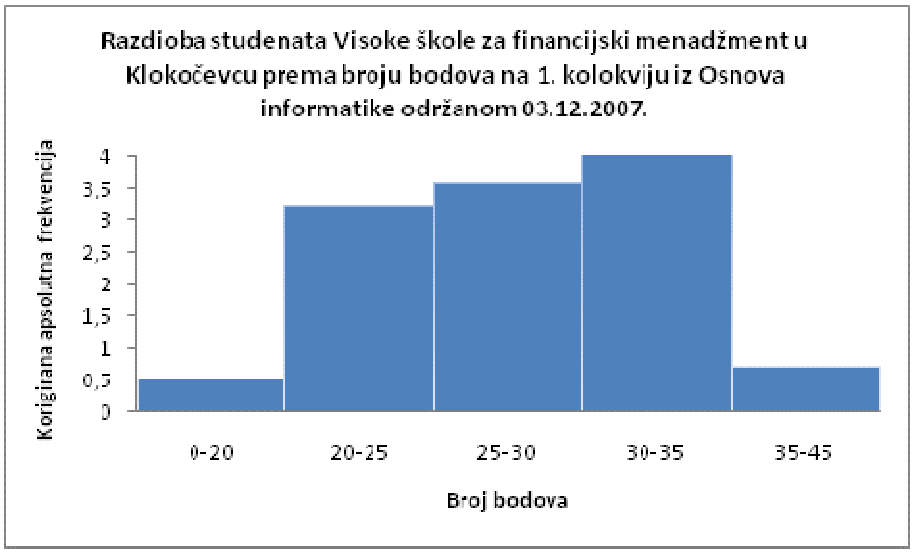
Pretpostavimo da smo brojeve bodova na 1. kolokviju iz kolegija *Osnove informatike* održanom 03.12.2007. koje su postigli studenti 1. godine stručnoga studija računovodstva i financija Visoke škole za financijski menadžment u Klokočevcu grupirali u sljedeće razrede:

broj bodova	*broj studenata*
0-20	10
20-25	16
25-30	18
30-35	22
35-45	7

Iako je riječ o pravim razredima, oni nisu jednakih širina. Širina prvoga razreda jednaka je 20, širine triju srednjih razreda jednake su 5, dok je širina posljednjega razreda jednaka 10. Računamo korigirane apsolutne frekvencije:

broj bodova	*korigirana apsolutna frekvencija*
0-20	0.5
20-25	3.2
25-30	3.6
30-35	4.4
35-45	0.7

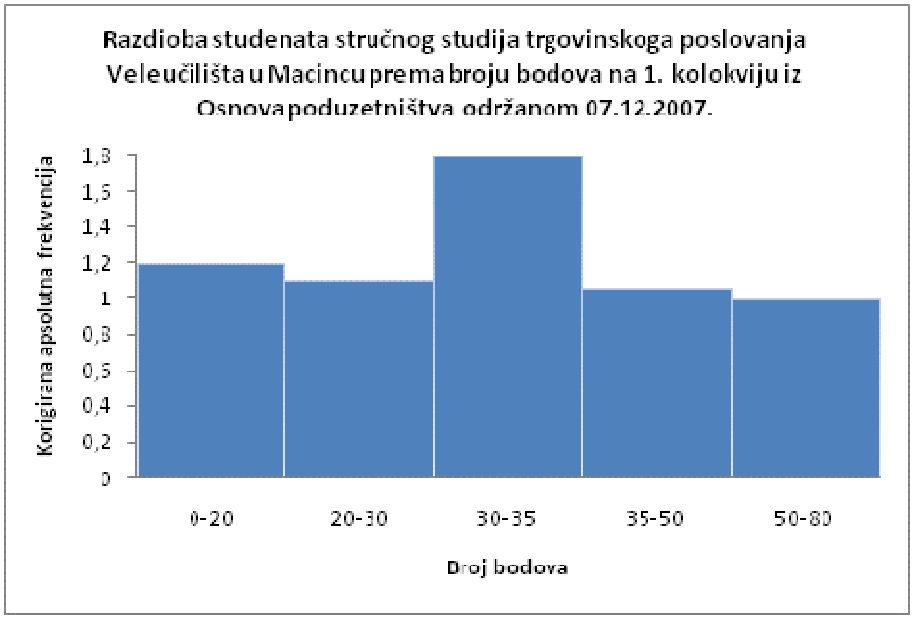
pa pripadni histogram izgleda ovako:



Uočimo odmah da usporedba korigiranih apsolutnih frekvencija ovdje navodi na **pogrešan** zaključak da je najmanje studenata na kolokviju ostvarilo najviše 20 bodova. Zbog nedovoljno preciznoga mjerila ovdje ne možemo točno očitati korigirane frekvencije, pa je jedan od zaključaka koji možemo izvesti na temelju ovoga grafikona npr. da je *među svim studentima koji su na kolokviju postigli između* 20 *i* 35 *bodova najviše onih koji su postigli između* 30 *i* 35 *bodova*. (Uspoređivali smo razrede istih širina za koje vrijedi pravilo da razred s najvećom apsolutnom frekvencijom ima najveću korigiranu frekvenciju.)

<u>Primjer 11.</u> Primjer *<u>histograma</u>* (pravi razredi različitih širina):

Iz sljedećega histograma



usporedbom korigiranih apsolutnih frekvencija zaključujemo da je najviše promatranih studenata na kolokviju ostvarilo između 30 i 35 bodova, a najmanje studenata između 50 i 80 bodova. Činjenično stanje, međutim, potpuno je drugačije jer je gornji histogram dobiven iz sljedeće tablice:

broj bodova	*broj studenata*
0-20	24
20-30	11
30-35	9
35-50	16
50-80	30

Iako se ovdje prirodno nameće zaključak da je *uvijek* najbolje grupirati podatke u razrede jednakih širina jer je tada usporedba najlakša, takav pristup općenito nije uvijek primjeren, a naročito ne u slučaju izrazito asimetričnih razdioba. Detalje ovdje izostavljamo.

Primjer 12. Primjer *histograma* (nepravi razredi različitih širina):

Zadana je razdioba rastavljenih brakova prema navršenim godinama života supruge u Kraljevini Dajdamdaš 2006. godine:

navršene godine života	do 23	24 – 32	33 - 39	40 - 44	45 - 49	50 - 59	60 i više
broj brakova	107	374	475	456	372	212	182

izvor: Statistički ljetopis Kraljevine Dajdamdaš za 2006. godinu

Odmah uočimo da su prvi i posljednji razred otvoreni, tj. prvi razred nema definiranu donju, a posljednji razred gornju granicu. Procjena tih granica ovisi o svakom slučaju posebno. Ovdje ćemo pretpostaviti da je donja granica prvoga razreda 16, a gornja granica posljednjega razreda 80 godina. Tako dobivamo tablicu:

navršene godine života	16 - 23	24 - 32	33 – 39	40 - 44	45 - 49	50 - 59	60 – 80
broj brakova	107	374	475	456	372	212	182

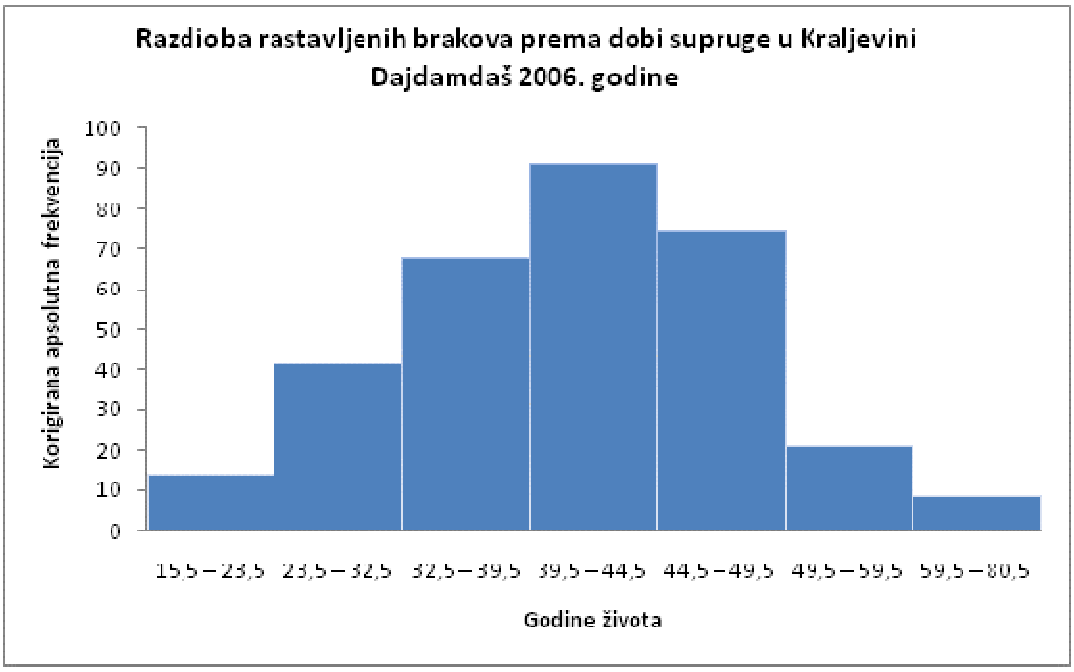
Razmaci između donje granice svakoga razreda (osim prvoga) i gornje granice tom razredu neposredno prethodnoga razreda su jednaki i iznose 1. Stoga nominalne granice pretvaramo u prave tako da svaku donju granicu smanjimo za polovicu razmaka, tj. za 0.5, a svaku gornju uvećamo za polovicu razmaka (opet za 0.5). Tako dobivamo sljedeću tablicu:

godine života	15.5 – 23.5	23.5 – 32.5	32.5 – 39.5	39.5 – 44.5	44.5 – 49.5	49.5 – 59.5	59.5 – 80.5
broj brakova	107	374	475	456	372	212	182

Dalje postupamo kao i u prethodnom primjeru. Razredi nemaju jednake širine, pa je potrebno izračunati korigirane apsolutne frekvencije za svaki pojedini razred. Dobivamo:

godine života	15.5 – 23.5	23.5 – 32.5	32.5 – 39.5	39.5 – 44.5	44.5 – 49.5	49.5 – 59.5	59.5 – 80.5
korigirana apsolutna frekvencija	13.375	41.555556	67.85714	91.2	74.4	21.2	8.666666667

pa je odgovarajući histogram



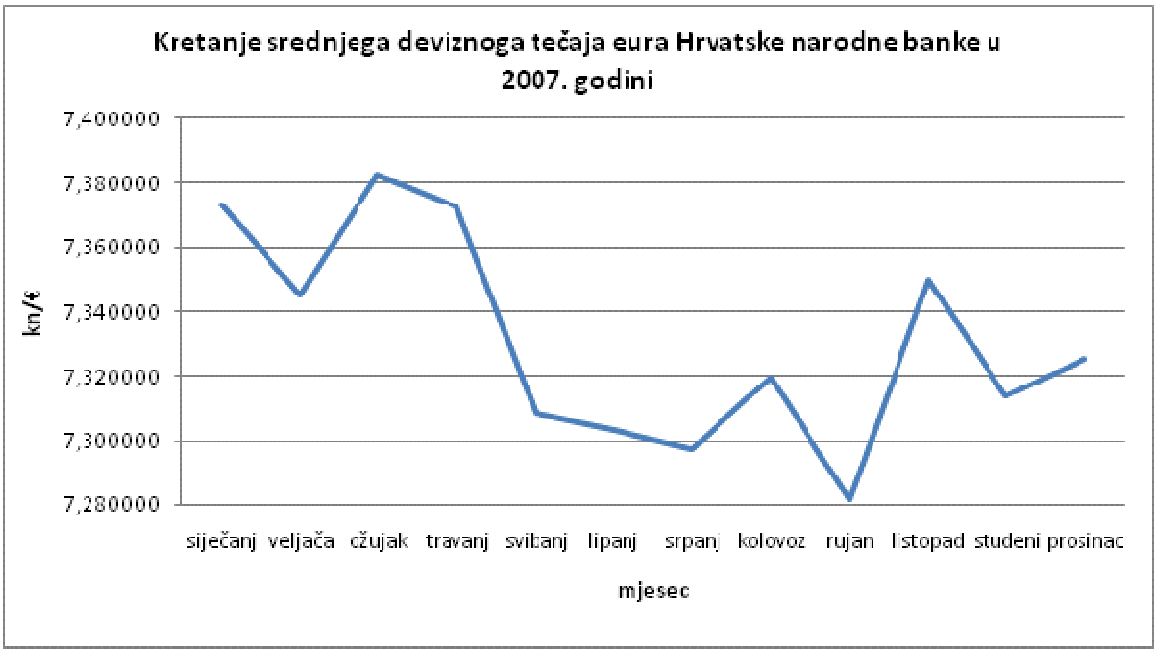
Primijetimo ponovno da razred s najvećom/najmanjom apsolutnom frekvencijom nema najveću/najmanju korigiranu apsolutnu frekvenciju, pa bismo samo na temelju grafikona mogli donijeti pogrešan zaključak o najčešćoj dobi supruge. Za vježbu navedite najmanje jedan *točan* zaključak koji možemo izvesti na temelju gornjega grafikona.

2.5.3. *Linijski grafikoni*

Linijski grafikoni pripadaju u tipove grafikona naročito pogodne za grafički prikaz *vremenskih nizova*, *poligona frekvencija* i *kumulante* određene razdiobe (najčešće prema nekom kvantitativnom obilježju), te za usporedbu najmanje dvaju nizova kvantitativnih podataka. Po jednim primjerom ilustrirat ćemo svaku pojedinu podvrstu linijskih grafikona.

Primjer 13. Primjer *linijskoga grafikona* – grafički prikaz vremenskoga niza[51]

Srednji devizni tečaj eura Hrvatske narodne banke u 2007. godini (stanje potkraj mjeseca)



Međusobno spojene su točke A = (siječanj, 7.3734) i B = (veljača, 7.345292), B = (veljača, 7.345292) i C = (ožujak, 7.382466) itd. Iz grafikona možemo zaključiti da je najniži srednji devizni tečaj eura bio potkraj mjeseca rujna, a najviši potkraj mjeseca ožujka, te da je tijekom godine srednji devizni tečaj pokazivao tendenciju rasta u razdobljima veljača – ožujak, srpanj – kolovoz, rujan – listopad i studeni – prosinac, dok je u ostalim vremenskim razdobljima tečaj pokazivao tendenciju pada.

Primjer 14. Primjer *linijskoga grafikona* – usporedba dvaju vremenskih nizova

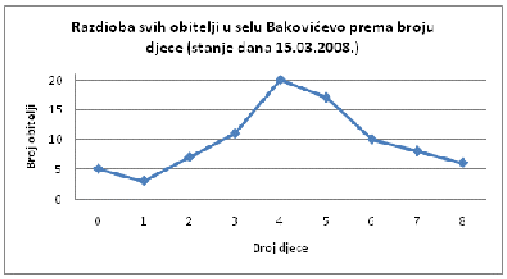
Na sljedećem grafikonu prikazana je grafička usporedba srednjih deviznih tečajeva dviju inozemnih valuta (euro i američki dolar) Hrvatske narodne banke tijekom 2007. godine (stanje potkraj razdoblja). Radi lakše usporedbe vrijednosti za pojedini mjesec, odgovarajuće su točke spojene okomitom spojnicom. Iz grafikona zaključujemo da je tečaj američkoga dolara gotovo tijekom cijele godine pokazivao tendenciju pada i bio bitno nestabilniji od tečaja eura.[52]

---

[51] Osnovama analize vremenskih nizova detaljnije ćemo se baviti kasnije.
[52] Nestabilnost tečaja dolara u odnosu na tečaj eura bila bi još izraženija kad bi se mjerilo na okomitoj osi sljedećega grafikona podudaralo s mjerilom na okomitoj osi gornjega grafikona.

Poseban primjer linijskoga grafikona je *poligon*[53] *frekvencija*. Riječ je o linijskom grafikonu namijenjenu grafičkom prikazu razdiobe različitih vrsta frekvencija (dakle, riječ je o linijskom analogonu histograma). Ukoliko modaliteti kvantitativnoga obilježja nisu grupirani u razrede, poligon frekvencija dobiva se spajanjem točaka koje predstavljaju uređeni par modaliteta i njegove apsolutne frekvencije. Ukoliko su ti modaliteti grupirani u razrede, poligon frekvencija dobiva se tako da se, sukladno redoslijedu razreda na osi apscisa, spoje točke koje predstavljaju uređeni par razredne sredine i apsolutne ili korigirane apsolutne frekvencije dotičnoga razreda. Oblik poligona frekvencija ukazuje na način rasporeda podataka, a razlike ordinata (tj. drugih koordinata) točaka o razlikama frekventiranosti oblika obilježja.

Primjer 15. Primjer *poligona frekvencija* (kvantitativno obilježje s negrupiranim modalitetima)

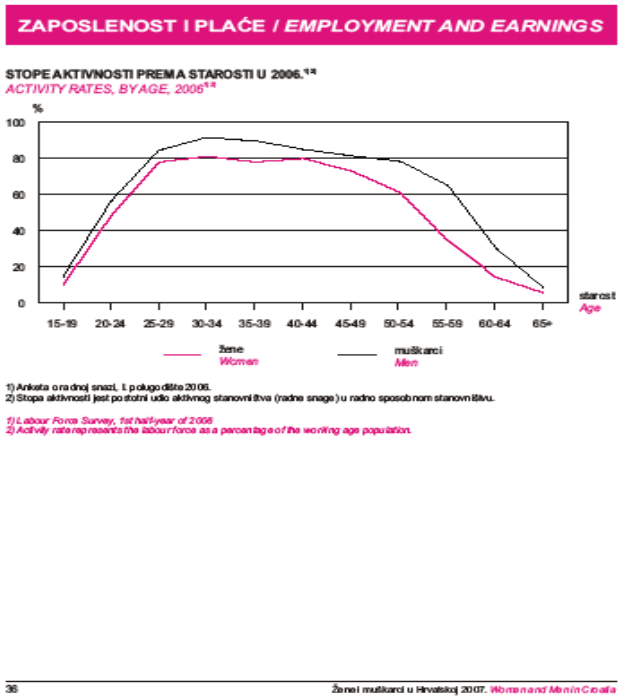


Prikazani grafikon predstavlja svojevrstan linijski analogon jednostavnih (ili položenih) stupaca. Budući da ne možemo točno očitati apsolutnu frekvenciju svakoga modaliteta, možemo zaključiti npr. da najviše obitelji ima dvoje djece, a najmanje obitelji ima sedmero djece. Razlike u frekventiranosti nisu odveć značajne jer krivulja nije zamjetno "izbočena".

---

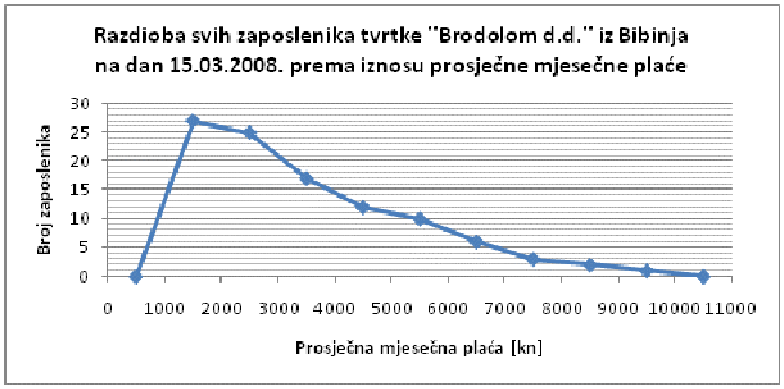
[53] Prema strogoj matematičkoj definiciji, poligon frekvencija je zapravo *otvorena poligonalna crta*. Da bi se dobio zatvoreni poligon, treba spojiti prvu i posljednju ucrtanu točku. Iako matematički netočan, naziv *poligon* (mnogokut) zadržao se kao dio uobičajenoga statističkoga nazivlja.

<u>Primjer 16.</u> Primjer *<u>poligona frekvencija</u>* (kvantitativno obilježje s modalitetima grupiranim u razrede s nominalnim granicama)



Na gornjoj su slici prikazana dva poligona *relativnih* frekvencija koja predočuju stope aktivnosti pojedinih dobnih razreda, i to posebno za žene, a posebno za muškarce. Ovaj je grafikon zapravo svojerstan linijski analogon strukturnih stupaca. Iz njega, primjerice, možemo očitati da za oba spola nominalni razred [30, 34] ima najveću stopu aktivnosti (tj. postotni udio radne snage u odnosu na radno sposobno stanovništvo je najveći za promatrani razred), a otvoreni razred [65, +∞⟩ najmanju stopu aktivnosti. Pokušajte navesti još neke zaključke koji se mogu izvesti iz navedenoga grafikona.

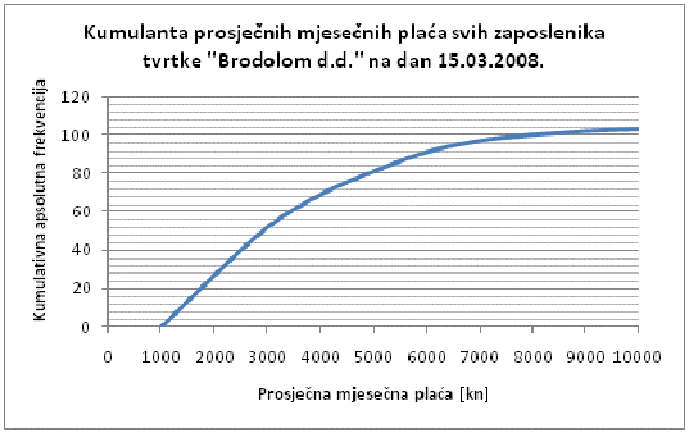
Primjer 17. Primjer *poligona frekvencija* (kvantitativno obilježje čiji su modaliteti grupirani u prave razrede)



Razredi u koje su grupirani iznosi prosječne mjesečne plaće su ⟨1000, 2000], ⟨2000, 3000], ⟨3000, 4000] itd., a odgovarajuće razredne sredine su 1500, 2500, 3500 itd. Da se (zajedno osi apscisa) dobije zatvorena poligonalna crta, "umjetno" su dodani razredi ⟨0, 1000] i ⟨10000, 11000] čija je apsolutna frekvencija 0. Možemo zaključiti da najviše zaposlenika promatrane tvrtke ima plaću između 1.000,00 i 2.000,00 kn, a najmanje između 9.000,00 i 10.000,00 kn. Uočimo da bi najprimjereniji površinski analogon prikazanoga linijskoga grafikona bio Paretov dijagram (apsolutne frekvencije su poredane od najveće do najmanje) iako bi se promatrana razdioba mogla prikazati i histogramom.

Kumulativni niz apsolutnih (ili relativnih) frekvencija "manje od" uobičajeno se grafički prikazuje linijskim grafikonom tzv. *kumulantom*. Praktično crtanje kumulante u pravokutnom koordinatnom sustavu u ravnini ovisi o tipu obilježja (diskretno ili kontinuirano), a mi ćemo ga ilustrirati na razdiobi iz Primjera 16.

Primjer 18. Primjer *kumulante* (kvantitativno obilježje čiji su modaliteti grupirani u prave razrede)

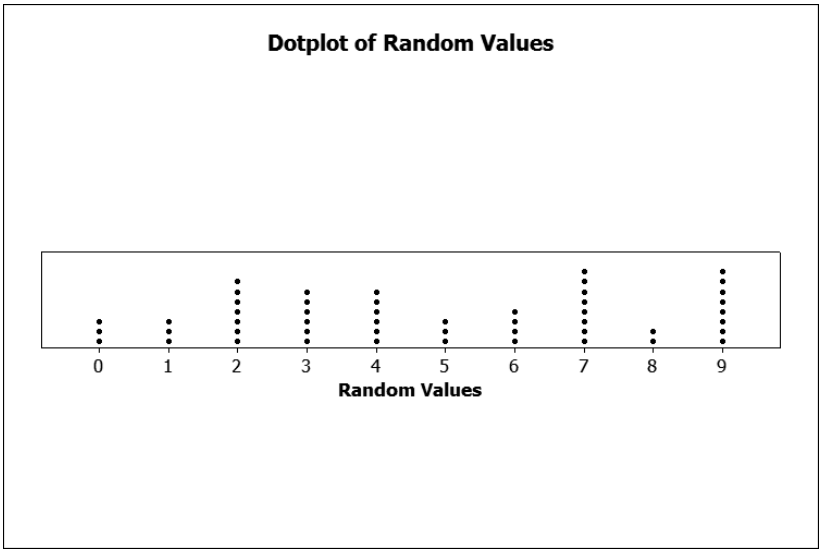


Kumulanta se crta tako da se *donjoj granici prvoga razreda* (u našem je slučaju to 1000) pridruži frekvencija 0 (na taj smo način dobili točku (1000, 0)). Potom se *svakoj* od *gornjih granica* razreda (to su 2000, 3000, 4000, …, 10000) pridruži odgovarajuća kumulativna apsolutna frekvencija "manje od", te se sve tako dobivene točke spoje glatkom[54] krivuljom. Napomenimo da se, iako teorijski ima smisla, kumulanta niza "veće od" u praksi vrlo rijetko koristi, pa je iz tih razloga ovdje i ne navodimo.

2.5.4. *Ostale vrste grafikona*

Budući da ćemo se u ovom kolegiju susretati isključivo s površinskim i linijskim grafikonima, ostale vrste grafikona dajemo samo pregledno bez detaljnijih opisa i analiza.

Primjer 19. Primjer *točkastoga grafikona* – dijagram točaka (razdioba statističkoga skupa prema jednom obilježju)



Iznad svakoga pojedinoga modaliteta broj točaka označava odgovarajuću apsolutnu frekvenciju. Ovdje prikazani dijagram točaka označava razdiobu rezultata ukupno 50 slučajnih pokusa[55] *izbor jedne dekadske znamenke*, a iz njega je lako očitati *mod* dobivenoga statističkoga niza, tj. najčešću znamenku koja se pojavila kao rezultat slučajnoga pokusa. U promatranom slučaju imamo dvije takve znamenke: 7 i 9.[56]

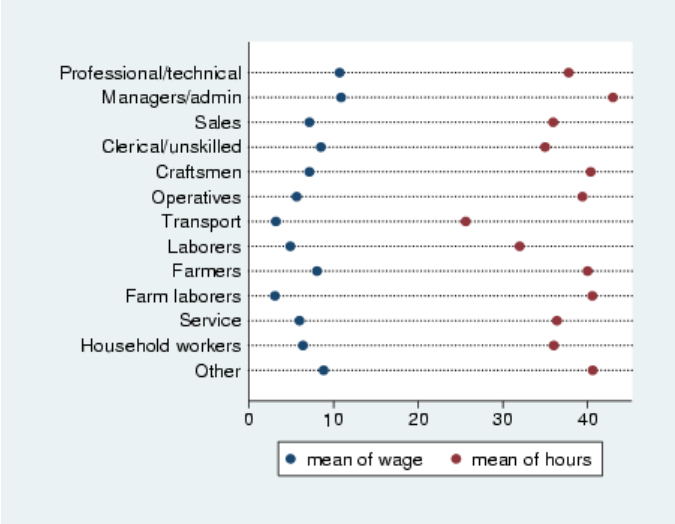
---

[54] U matematici se pod pojmom *glatka krivulja* podrazumijeva graf bilo koje neprekidno diferencijabilne funkcije. Grubo govoreći, riječ je o grafu koji ne sadrži niti jedan "šiljak". Primjer grafa koji sadrži "šiljke" je bilo koji poligon frekvencija.

[55] *Slučajni pokus* je jedan od osnovnih pojmova teorije vjerojatnosti, a time i matematičke statistike.

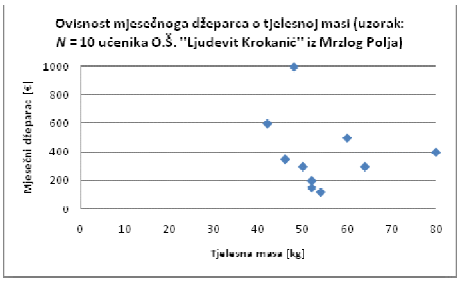
[56] Razdiobu u kojoj imamo točno dvije najčešće vrijednosti nazivamo *bimodalna razdioba*. Iako ovaj primjer to ne daje naslutiti, *zakon velikih brojeva* (jedan od najznačajnijih poučaka teorije vjerojatnosti) tvrdi da će u

Primjer 20. Primjer *točkastoga grafikona* – dijagram točaka (razdioba statističkoga skupa prema trima obilježjima)



Na istom grafikonu prikazana je podjela svog radno aktivnog stanovništva određene regije prema trima obilježjima: vrsta posla kojim se bave, prosječna plaća i prosječan broj tjednih radnih sati. Uočimo da dva kvantitativna obilježja *nemaju* istu mjernu jedinicu, no, ta se činjenica kod točkastih grafikona može zanemariti (kod površinskih i linijskih se ta činjenica ne smije zanemariti). Iz prikazanoga grafikona se može zaključiti da su menadžeri najbolje plaćeni, ali i da najviše rade, dok su težaci najslabije plaćeni za svoj rad unatoč većem broju tjednih radnih sati. Također, može se zaključiti da je *varijabilitet* prosječnih plaća manji od *varijabiliteta* prosječnih tjednih radnih sati, tj. promatrano stanovništvo se više razlikuje prema tjednom radnom vremenu, nego prema prosječnim plaćama.[57]

Primjer 21. Primjer *točkastoga grafikona* – dijagram rasipanja (dijagram raspršenja, oblak raspršenja)



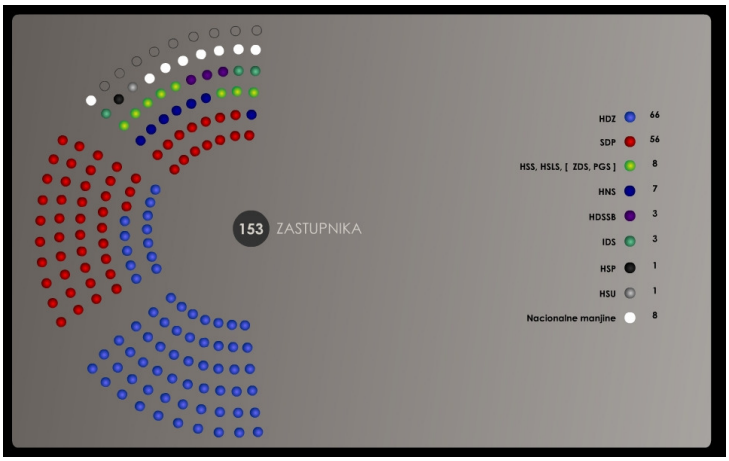
---

slučaju *vrlo velikoga* broja provedenih slučajnih pokusa (pod istim uvjetima) dobivene apsolutne frekvencije znamenaka biti približno jednake, tj. u dobivenom statističkom nizu svaka od znamenaka pojavit će se (približno) jednako često.

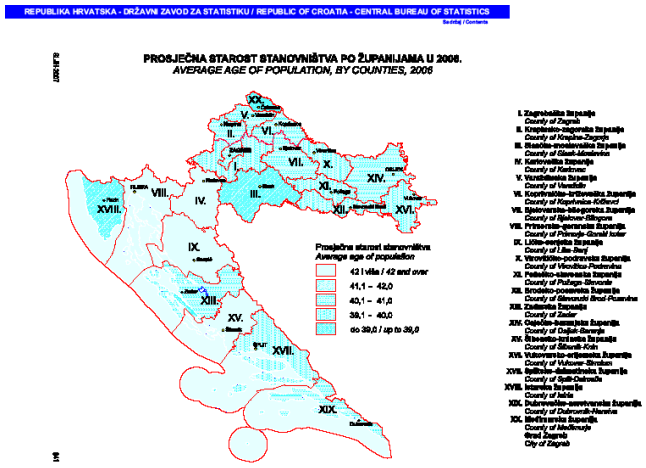
[57] Navedena usporedba varijabiliteta obično se utvrđuje na temelju *koeficijenata varijacije* o kojima ćemo više reći u 4. poglavlju.

O dijagramu rasipanja više ćemo govoriti u poglavlju 4. Ovdje istaknimo samo da se ovakva vrsta grafikona uobičajeno koristi prigodom utvrđivanja *korelacije* dvaju *kvantitativnih* obilježja, tj. prigodom ispitivanja postoji li statistički značajna veza među promatranim kvantitativnim obilježjima (definiranih na istom statističkom skupu, odnosno uzorku toga skupa). Točke se *ne* spajaju nikakvim krivuljama, nego se iz njihova položaja i oblika točastoga grafikona nastoje izvući zaključci o jačini i smjeru veze među promatranim obilježjima. Ovdje se već na prvi pogled zbog "raspršenosti" točaka po grafikonu može naslutiti da *na promatranom uzorku* ne postoji statistički značajna veza između iznosa mjesečnoga džeparca učenika i njihove tjelesne mase.

Primjer 22. Primjer *dijagrama točaka* – struktura 6. saziva Hrvatskoga sabora



Primjer 23. Primjer *kartograma*[58] – statistička karta



[58] Najpoznatiji tipovi kartograma su *piktogram* i *statistička karta*, no, oni se koriste isključivo u zemljopisu za grafički prikaz nominalnih prostornih obilježja, pa detalje ovdje izostavljamo.

3. OSNOVNE KARAKTERISTIKE NUMERIČKIH NIZOVA

3.1. *Srednje vrijednosti numeričkih nizova*

U praksi se vrlo često javlja problem kvalitetnoga reprezentiranja velikih nizova podataka kvalitativnih ili kvantitativnih obilježja vezanih za relativno velike statističke skupove (npr. cjelokupno stanovništvo nekoga grada/županije/države). I dok je u slučaju kvalitativnih obilježja najbolji mogući rezultat tablični ili grafički prikaz podataka, za kvantitativna se obilježja uvode posebni numerički pokazatelji. Ti pokazatelji su tzv. *srednje vrijednosti*.

Srednja vrijednost[59] je općenito neka realna konstanta čiji je cilj na što reprezentativniji način predočiti niz varijabilnih podataka numeričkog niza. U pravilu je riječ o vrijednosti oko koje se "gomila" većina podataka numeričkog niza, pa se zato naziva i *mjera središnje (centralne) tendencije*. Što je više podataka "nagomilano" oko pojedine srednje vrijednosti, njezina će reprezentativnost biti bolja.

Osnovne vrste srednjih vrijednosti su *aritmetička sredina*, *geometrijska sredina*, *harmonijska sredina*, *medijan* i *mod*. Vrlo je važno istaknuti da je primjena svake pojedine vrste srednje vrijednosti određena *karakterom* (*svojstvima*) promatranoga kvantitativnoga obilježja, pa ne postoji "univerzalna" srednja vrijednost koja će dovoljno reprezentativno opisati *bilo koji* numerički niz podataka. U praksi se ovo pravilo često vrlo neopravdano zanemaruje, pa se za spomenutu "univerzalnu" srednju vrijednost, zbog relativno jednostavnoga načina izračuna, uzima aritmetička sredina. Takav je pristup potpuno pogrešan jer, kako ćemo vidjeti, postoje brojni primjeri numeričkih nizova takvi da aritmetička sredina nije dovoljno dobar reprezentant svakoga od tih nizova.[60]

U izračunu pojedine srednje vrijednosti mogu se pojaviti svi članovi numeričkoga niza ili samo dio tih članova. Sukladno tome, srednje vrijednosti dijelimo na *potpune* i *položajne srednje vrijednosti*.

U *potpune srednje vrijednosti* ubrajamo aritmetičku sredinu, geometrijsku sredinu i harmonijsku sredinu. U izračunu tih vrijednosti koriste se *svi* elementi numeričkoga niza. U *položajne srednje vrijednosti* ubrajamo medijan i mod. Njihova je vrijednost određena *položajem* unutar numeričkog niza, pa u njihovu izračunu ne sudjeluju svi elementi numeričkoga niza.

Zaključno istaknimo da svaku od navedenih srednjih vrijednosti možemo računati iz negrupiranih ("sirovih") i grupiranih podataka. Zajednička karakteristika svih takvih izračuna jest da su vrijednosti dobivene izračunom iz grupiranih podataka nepreciznije od vrijednosti dobivenih izračunom iz negrupiranih podataka, naročito ako su podaci grupirani u razrede relativno velikih širina.

---

[59] U svakodnevnom se životu pojam *srednje vrijednosti* vrlo često pogrešno zamjenjuje s pojmom *prosjeka*, odnosno *prosječne vrijednosti*. Pod pojmom *prosjek* zapravo se podrazumijeva aritmetička sredina.

[60] Tipičan primjer takvoga niza je niz mjesečnih plaća svih zaposlenika na (ne)određeno vrijeme u našoj zemlji.

3.1.1. *Aritmetička sredina*

Aritmetička sredina (oznake: $\overline{X}$, $\overline{x}$, $E(x)$) je najraširenija i najčešće korištena srednja vrijednost, a prilično je upitno proglašavati je i najvažnijom srednjom vrijednošću. Definiramo li *total* kao zbroj svih elemenata *konačnoga*[61] numeričkoga niza, onda se aritmetička sredina *definira* kao *jednaki dio totala po jednom elementu niza*. To praktično znači da se aritmetička sredina dobije tako da se total podijeli s ukupnim *brojem* elemenata numeričkoga niza.

Postoji nekoliko različitih vrsta aritmetičkih sredina. Najpoznatija i najraširenija je *jednostavna aritmetička sredina*. Za njezino je izračunavanje potrebno zadati *konačan* niz *negrupiranih* numeričkih podataka x_1, x_2, …, x_n. Tada se jednostavna aritmetička sredina računa pomoću formule:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Primjer 1. Zadan je numerički niz Smiljkovih mjesečnih neto–plaća (iskazanih u kn) u prošloj (2007.) godini:

1.810,25; 1.810,25; 1.810.25; 1.850,5; 1.850,5; 1.850,5; 1.862,4; 1.862,4; 1.862,4; 1.875,8; 1.875,8; 1.875,8

Izračunajmo Smiljkovu prosječnu mjesečnu neto-plaću u prošloj godini:

$$\overline{X} = \frac{1.810,25 + 1.810,25 + 1.810.25 + 1.850,5 + 1.850,5 + 1.850,5 + 1.862,4 + 1.862,4 + 1.862,4}{12} +$$
$$+ \frac{1.875,8 + 1.875,8 + 1.875,8}{12} = 1.849,7375 \approx 1.849,74$$

Dakle, Smiljkova prosječna mjesečna neto–plaća u prošloj (2007.) godini iznosi približno 1.849,74 kn. Uporabljujući ovakvu rečenicu mi niz od ukupno 12 numeričkih vrijednosti *zamjenjujemo jednom jedinom vrijednošću*: 1.849,74 kn. Stoga se prirodno postavlja pitanje "uspješnosti" takve zamjene, odnosno, preciznije, opisuje li dobivena vrijednost dovoljno dobro niz podataka kojega zamjenjuje. Odgovor ćemo moći djelomično dati nakon

[61] Ovdje namjerno naglašavamo riječ *konačan* jer se u matematici niz (u bilo kojem skupu *S*) definira kao funkcija sa skupa prirodnih brojeva **N** u skup *S*. Sukladno tome, niz može biti *konačan* (ako je prirodno područje definicije odgovarajuće funkcije konačan podskup skupa **N**) i *beskonačan* (ako je prirodno područje definicije odgovarajuće funkcije cijeli skup **N**). U deskriptivnoj se statistici promatraju isključivo konačni nizovi, dok se beskonačni nizovi promatraju u matematičkoj statistici.

iskazivanja općih svojstava bilo koje aritmetičke sredine, a potpuno nakon definiranja mjera raspršenja (disperzije) u sljedećem poglavlju.

Ukoliko se aritmetička sredina računa iz grupiranih podataka, govorimo o *vaganoj* ili *ponderiranoj aritmetičkoj sredini*. Pri izračunu vagane aritmetičke sredine razlikujemo slučajeve kad modaliteti kvantitativnoga obilježja nisu grupirani u razrede i kad su modaliteti kvantitativnoga obilježja grupirani u razrede.
U prvom slučaju zapravo se radi o "skrivenoj" jednostavnoj aritmetičkoj sredini jer grupiranje podataka isključivo prema modalitetima nije ništa drugo nego "skraćeni" (tablični) zapis numeričkoga niza. Formalno, ako su x_1, x_2, …, x_k međusobno *različiti* modaliteti kvantitativnoga obilježja, a f_1, f_2, …, f_k njima odgovarajuće *apsolutne* frekvencije, onda se vagana aritmetička sredina računa pomoću izraza

$$\overline{X} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_k x_k}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

Često se kaže da u izrazima ovoga oblika apsolutne frekvencije imaju ulogu *pondera* ili *težine* pojedinoga modaliteta, pa otuda i naziv odgovarajuće aritmetičke sredine.

Primjer 2. Zadana je razdioba svih studenata 1. godine studija smjehologije na Višoj uzaludnoj školi u Špičkovini (akademska godina 2007/2008.) prema broju položenih jednosemestralnih kolegija:

broj položenih jednosemestralnih kolegija	*broj studenata*
0	7
1	12
2	17
3	25
4	21
5	18
ukupno:	100

izvor: studentska služba Više uzaludne škole u Špičkovini

Navedena tablica zapravo zamjenjuje niz od 100 numeričkih podataka koji se sastoji od 7 nula, 12 jedinica, 17 dvojki, 25 trojki, 21 četvorke i 18 petica. Budući da nam je bitno lakše umjesto ukupno 99 operacija zbrajanja izvršiti 6 operacija množenja i 4 operacije zbrajanja, prosječan broj položenih jednosemestralnih ispita računamo kao vaganu aritmetičku sredinu:

$$\overline{X} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i} = \frac{7 \cdot 0 + 12 \cdot 1 + 17 \cdot 2 + 25 \cdot 3 + 21 \cdot 4 + 18 \cdot 5}{100} = \frac{295}{100} = 2.95 .$$

Dobiveni rezultat obično interpretiramo ovako: *Prosječan broj položenih jednosemestralnih kolegija po jednom studentu približno iznosi* 3.[62] Ovdje treba pripaziti i na red riječi u rečenici: prilog *prosječno* uvijek dolazi neposredno ispred veličine na koju se odnosi. Stoga je pogrešno reći: *Svi promatrani studenti prosječno su položili* 3 *jednosemestralna kolegija* ili *Prosječan promatrani student položio je* 3 *jednosemestralna kolegija* jer sintagma *prosječno položiti* obično znači *položiti s ocjenom dobar*(3), a sintagma *prosječan student* obično znači *student čiji je prosjek ocjena približno* 3,00. Odmah primijetimo da je točno 3 jednosemestralna kolegija *uistinu* položilo točno 25 studenata, a ne svih 100, pa je u ovakvim slučajevima nužno izračunati dodatne pokazatelje reprezentativnosti aritmetičke sredine.

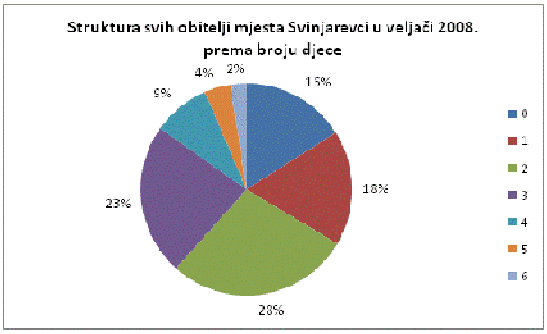
<u>Zadatak 1.</u> a) Definirajte statistički skup iz prethodnoga primjera pojmovno, prostorno i vremenski, pa odredite njegov opseg i odgovarajući osnovni skup.
b) Prema kojemu je obilježju podijeljen zadani statistički skup? U koju grupu obilježja pripada i uz koju je skalu izražavanja modaliteta vezano to obilježje? Objasnite svoje odgovore.

Aritmetička sredina može se računati i ukoliko su, umjesto, apsolutnih zadane odgovarajuće ("obične") relativne frekvencije. Radi jednostavnosti, pretpostavimo da su relativne frekvencije iskazane u postotcima. Formalno, ako su x_1, x_2, …, x_k međusobno *različiti* modaliteti kvantitativnoga obilježja, a p_1, p_2, …, p_k njima odgovarajuće *relativne* frekvencije *iskazane u postotcima*[63], onda se vagana aritmetička sredina računa pomoću izraza

$$\overline{X} = p_1x_1 + p_2x_2 + ... + p_kx_k = \sum_{i=1}^{k} p_ix_i \,.^{64}$$

Ilustrirajmo primjenu ove formule na sljedećem primjeru.

<u>Primjer 3.</u> Podaci o broju djece u svakoj obitelji iz mjesta Svinjarevci u veljači 2008. godine obrađeni su i grafički prikazani strukturnim krugom.



[62] Strogo formalna interpretacija, sukladno definiciji aritmetičke sredine, glasi: *Svaki od promatranih studenata položio je prosječno* 3 *jednosemestralna kolegija*. No, u interpretacijama se obično primjenjuje zamjena izraza "*svaki element nekoga skupa X ima svojstvo S*" izrazom "*svi elementi skupa X imaju svojstvo S*".
[63] Relativne frekvencije p_1, …, p_k ovdje su razlomci s nazivnikom 100, ali mogu biti iskazane i kao *proporcije* (decimalni zapisi omjera apsolutne frekvencije modaliteta i totala).
[64] Čitatelji upoznati s osnovama diskretne teorije vjerojatnosti odmah će uočiti da je navedeni izraz zapravo definicija *matematičkoga očekivanja* diskretne slučajne varijable koja poprima vrijednosti x_1, x_2, …, x_n.

Na temelju navedenih podataka možemo izračunati prosječan broj djece po jednoj obitelji iz promatranoga mjesta:

$$\bar{X} = \frac{\sum_{i=1}^{7} p_i x_i}{100} = \frac{16 \cdot 0 + 18 \cdot 1 + 28 \cdot 2 + 23 \cdot 3 + 9 \cdot 4 + 4 \cdot 5 + 2 \cdot 6}{100} = \frac{211}{100} = 2.11$$

Dobiveni rezultat možemo objasniti ovako: *Prosječan broj djece po jednoj* (*promatranoj*) *obitelji jednak je* 2. Opet treba pripaziti i na red riječi u rečenici jer je pogrešno reći npr. *Prosječna* (*promatrana*) *obitelj ima dvoje djece* (kako to često čine u javnim medijima!).

Zadatak 2. **a)** Definirajte statistički skup u prethodnom primjeru pojmovno, prostorno i vremenski, pa odredite pripadni osnovni skup.
b) Prema kojemu je obilježju podijeljen zadani statistički skup? U koju grupu obilježja pripada i uz koju je skalu izražavanja modaliteta vezano to obilježje? Objasnite svoje odgovore.

Spomenimo još da vagana aritmetička sredina dobivena iz relativnih frekvencija teorijski mora biti identički jednaka jednostavnoj aritmetičkoj sredini, ali da praktično dolazi do više-manje značajnih odstupanja nastalih uslijed *približnih izračuna* relativnih frekvencija. Što je približni izračun relativnih frekvencija točniji, spomenuto odstupanje bit će manje, pa se u ovakvim slučajevima preporučuje izračunavati relativne frekvencije na najmanje 5 decimalnih mjesta.

U slučajevima poput netom razmotrenih, dakle, jednostavna aritmetička sredina identički mora biti jednaka vaganoj aritmetičkoj sredini. No, ukoliko podatke grupiramo u razrede, dolazi do razlika koje, u pravilu, ovise o *veličini* (*širini*) pojedinih razreda, a ne o ukupnom broju razreda. Objasnimo nastajanje tih razlika na konkretnom primjeru.

Primjer 4. Zadana je razdioba svih nezaposlenih državljana Republike Ljenčarije u veljači 2008. godine prema dobi

navršene godine starosti	*broj nezaposlenih* [000]
15 – 24	25.2
25 – 49	75.8
50 – 64	19.2
≥ 65	0.6

izvor: mjesečno statističko izvješće Državnoga zavoda za statistiku Republike Ljenčarije, ožujak 2008.

Da bismo mogli izračunati aritmetičku sredinu i u ovom slučaju, zadane nominalne granice razreda trebamo pretvoriti u prave, te procijeniti gornju granicu posljednjega (otvorenoga) razreda. Procijenit ćemo da spomenuta gornja granica jednaka 70. U slučaju obilježja *navršene godine starosti* prave granice se dobiju izjednačavanjem gornje granice prethodnoga razreda s donjom granicom tekućega razreda :

navršene godine starosti	*broj nezaposlenih* [000]
15 – 25	25.2
25 – 50	75.8
50 – 65	19.2
65 – (70)	0.6

Prigodom računanja aritmetičke sredine iz podataka grupiranih u prave razrede, ulogu ponderâ (težinâ) i dalje imaju apsolutne frekvencije, ali "uloge" modaliteta x_1, x_2, …, x_k preuzimaju *razredne sredine*. Točnije, za svaki pojedini razred trebamo izračunati njegovu razredu sredinu, pa dobivenoj vrijednosti kao ponder pridružiti odgovarajujuću apsolutnu frekvenciju. Upravo na ovom mjestu nastaju spomenute razlike između jednostavne i vagane aritmetičke sredine: različite podatke grupirane u isti razred zamjenjujemo *istim* podatkom, tj. razrednom sredinom. Konkretno, razredna sredina prvoga razreda je

$$s_1 = \frac{15+25}{2} = 20,$$

pa u daljnjem izračunu niz od 25 200 podataka iz prvoga razreda među kojima sigurno ima međusobno različitih (jer inače grupiranje u razrede nema smisla) zamjenjujemo nizom od 25 200 *međusobno jednakih podataka* i svaki od tih novih podataka jednak je razrednoj sredini prvoga razreda, tj. 20. Budući da takvu zamjenu *ne* provodimo kod izračuna jednostavne aritmetičke sredine, nego pri njezinu izračunu rabimo originalne podatke, vagana aritmetička sredina izračunata iz podataka grupiranih u razrede je u pravilu lošiji reprezentant promatranoga statističkoga skupa od jednostavne aritmetičke sredine.[65]
Izračunom razrednih sredina dobivamo:

navršene godine starosti	*razredna sredina*	*broj nezaposlenih* [000]
15 – 25	20	25.2
25 – 50	37.5	75.8
50 – 65	57.5	19.2
65 – (70)	67.5	0.6

pa je vagana aritmetička sredina izračunata iz podataka grupiranih u razrede[66]:

$$\overline{X} = \frac{\sum_{i=1}^{4} f_i \cdot s_i}{\sum_{i=1}^{4} f_i} = \frac{25.2 \cdot 20 + 75.8 \cdot 37.5 + 19.2 \cdot 57.5 + 0.6 \cdot 67.5}{25.2 + 75.8 + 19.2 + 0.6} = \frac{4491}{120.8} \approx 37.177$$

[65] Iznimka"koja potvrđuje pravilo" su nizovi podataka s ekstremno velikim ili ekstremno malim vrijednostima. Grupiranjem takvih nizova u razrede i zamjenom originalnih podataka razrednom sredinom moguće je znatno smanjiti razliku između pojedinih ekstremno velikih vrijednosti i ostalih vrijednosti skupa, a time i donekle poboljšati reprezentativnost aritmetičke sredine.
[66] Prigodom izračuna aritmetičke sredine "mjernu jedinicu" apsolutnih frekvencija (000) možemo zanemariti, tj. ne moramo računati s tisućama. Razlog tome je što se ista jedinica pojavljuje i u brojniku i u nazivniku odgovarajućega razlomka, pa prigodom dijeljenja dolazi do njezinoga kraćenja.

Dobiveni rezultat možemo interpretirati ovako: *Procijenjeni prosjek navršenih godina starosti promatranih državljana približno je jednak* 37 (pogrešno je reći npr. *Prosječan nezaposleni državljanin Republike Ljenčarije ima navršenih* 37 *godina* ili neku sličnu "mudroliju"). Ovdje **obavezno** moramo navesti riječ *procijenjena* jer smo navedeni rezultat dobili koristeći *procjenu* gornje granice posljednjega razreda, te zamijenivši originalne podatke razrednim sredinama (u svakom od tih slučajeva smo napravili određenu pogrešku, pa ne možemo govoriti o stvarnoj, nego isključivo o procijenjenoj aritmetičkoj sredini).

Zadatak 3. **a)** Definirajte statistički skup iz prethodnoga primjera pojmovno, prostorno i vremenski, pa odredite njegov opseg.
b) Prema kojemu je obilježju podijeljen zadani statistički skup? U koju grupu obilježja pripada i uz koju je skalu izražavanja modaliteta vezano to obilježje? Objasnite svoje odgovore.
c) U koliko su razreda podijeljeni dobiveni podaci? Za svaki od njih odredite donju granicu, gornju granicu i veličinu.

Za izračunavanje aritmetičke sredine iz grupiranih podataka, umjesto apsolutnih, možemo koristiti i relativne frekvencije. Radi jednostavnosti pretpostavit ćemo da su relativne frekvencije već iskazane u postotcima. U takvom slučaju razredne sredine imaju ulogu modaliteta, a relativne frekvencije ulogu pondera (težina). Njihov je zbroj jednak 100, pa će se u nazivniku odgovarajućega razlomka naći broj 100. Ilustrirajmo to primjerom.

Primjer 5. Zadana je struktura žena – ovisnica o drogama liječenih u zdravstvenim ustanovama u Republici Hrvatskoj tijekom 2005. godine prema navršenim godinama starosti:

navršene godine starosti	*struktura* [%]
≤ 14	0.5
15 – 19	12.1
20 – 24	27.0
25 – 29	25.8
30 – 34	14.4
35 – 39	18.0
≥ 40	2.2
ukupno:	**100**

izvor: "Žene i muškarci u Republici Hrvatskoj u 2007. godini“, Državni zavod za statistiku, 2008.

Procijenimo prosječnu navršenu dob svih promatranih žena. Kao donju granicu prvoga razreda uzmimo 10, a kao gornju granicu posljednjega razreda uzmimo 60. Nominalne granice pretvorimo u prave izjednačavanjem gornje granice prethodnoga razreda s donjom granicom tekućega razreda:

navršene godine starosti	*struktura* [%]
(10) – 15	0.5
15 – 20	12.1
20 – 25	27.0
25 – 30	25.8
30 – 35	14.4
35 – 40	18.0
40 – (60)	2.2
ukupno:	**100**

Kako je već istaknuto, u ovakvim slučajevima razredne sredine shvaćamo kao modalitete, a relativne frekvencije (iskazane u postotcima) kao pondere (težine). Zbroj svih relativnih frekvencija uvijek je jednak 100, pa prije izračuna procijenjene aritmetičke sredine najprije moramo izračunati odgovarajuće razredne sredine (s_i):

navršene godine starosti	*razredna sredina*	*struktura* [%]
(10) – 15	12.5	0.5
15 – 20	17.5	12.1
20 – 25	22.5	27.0
25 – 30	27.5	25.8
30 – 35	32.5	14.4
35 – 40	37.5	18.0
40 – (60)	50	2.2
ukupno:		**100**

Procijenjena aritmetička sredina izračunata iz grupiranih podataka jednaka je:

$$\bar{X} = \frac{\sum_{i=1}^{7} p_i \cdot s_i}{100} = \frac{0.5 \cdot 12.5 + 12.1 \cdot 17.5 + 27 \cdot 22.5 + 25.8 \cdot 27.5 + 14.4 \cdot 32.5 + 18 \cdot 37.5 + 2.2 \cdot 50}{100} =$$
$$= \frac{2788}{100} = 27.88$$

Dobiveni rezultat možemo interpretirati ovako: *Procijenjena prosječna dob svih promatranih žena je približno jednaka* 28 *godina* (pogrešno je npr. reći: *Prosječna žena – ovisnica o drogama ima navršenu dob od* 28 *godina* i sl.).

Naposljetku, u praksi se često javlja slučaj kad treba izračunati "prosjek prosjeka", tj. aritmetičku sredinu aritmetičkih sredina. U takvim slučajevima treba poznavati apsolutne ili relativne frekvencije odgovarajućih modaliteta[67]. Ulogu pondera (težina) u ovom slučaju imaju zadane aritmetičke sredine. Pokažimo to na primjeru:

Primjer 6. Zadani su podaci o ukupnom broju zaposlenih u Republici Hrvatskoj u području financijskoga posredovanja na dan 31.3.2006. i njihovim prosječnim mjesečnim neto–plaćama u razdoblju od 01.01.2006. do 31.3.2006.

[67] Zadavanje navedenih vrijednosti je nužan uvjet za izračun prosječne vrijednosti svih podataka. Bez poznavanja navedenih vrijednosti izračun aritmetičke sredine aritmetičkih sredina je praktički besmislen.

vrsta djelatnosti	*broj zaposlenih*	*prosječna mjesečna neto – plaća* [kn]
financijsko posredovanje osim osiguranja i mirovinskih fondova	25 310	6.616,00
osiguranje i mirovinski fondovi osim obveznoga osiguranja	6 077	6.505,00
pomoćne djelatnosti u financijskom posredovanju	1 763	4.444,00
ukupno:	33 150	

izvor: Statistički ljetopis Republike Hrvatske za 2006. godinu, Državni zavod za statistiku, Zagreb, 2007.

Procijenimo prosječnu mjesečnu neto–plaću svih promatranih zaposlenika u navedenom razdoblju. Trebamo, dakle, izračunati aritmetičku sredinu aritmetičkih sredina neto–plaća. Brojevi zaposlenih su apsolutne frekvencije i one imaju uobičajenu ulogu pondera (težine), dok iznosi prosječnih neto–plaća imaju ulogu modaliteta. Stoga je aritmetička sredina aritmetičkih sredina neto–plaća jednaka

$$\overline{X} = \frac{\sum_{i=1}^{3} f_i \cdot x_i}{\sum_{i=1}^{3} f_i} = \frac{25310 \cdot 6.616,00 + 6077 \cdot 6.505,00 + 1763 \cdot 4.444,00}{25310 + 6077 + 1763} = \frac{214.816.617,00}{33150} \approx 6.480,14$$

Prema tome, *procijenjena prosječna mjesečna neto – plaća svih promatranih zaposlenika u navedenom razdoblju iznosi* 6.480,14 kn. Riječ *procijenjena* ovdje navodimo zato jer ne znamo metodologiju izračuna zadanih prosječnih mjesečnih neto–plaća: moguće je da su one izračunate iz originalnih podataka, ali je isto tako moguće da su izračunate i iz podataka grupiranih u razrede. Stoga je u interpretaciji primjereno koristiti riječ *procijenjena* neovisno o tome što je moguće da je dobivena vrijednost ujedno i stvarna prosječna mjesečna neto-plaća promatranih zaposlenika.

<u>Zadatak 4.</u> a) Definirajte statistički skup iz prethodnoga primjera pojmovno, prostorno i vremenski, pa odredite njegov opseg.
b) Prema kojemu je obilježju podijeljen zadani statistički skup? U koju grupu obilježja pripada i uz koju je skalu izražavanja modaliteta vezano to obilježje? Objasnite svoje odgovore.

<u>Primjer 7.</u> (*problem školskih ocjena*) Kako smo već ranije istakli, školske ocjene su primjer *kvalitativnoga redosljednoga obilježja* i vezane su uz *redosljednu* skalu izražavanja modaliteta, pa se nad njima ne mogu provoditi osnovne računske operacije. No, u praksi se u brojnim slučajevima određuje *prosječna ocjena* ili *prosjek ocjena* (kao zbroj svih ocjena podijeljen s njihovim ukupnim brojem) iako ta veličina, prema dosad iskazanoj statističkoj teoriji, uopće nema smisla. U čemu se zapravo sastoji ovaj prividni paradoks?

Prigodom *određivanja* (pogrešno: računanja!) prosječne ocjene[68] mi zapravo *kodiramo* svaku pojedinu ocjenu, tj. svakoj ocjeni bijektivno pridružujemo točno jedan prirodan *broj*: ocjeni "nedovoljan" pridružujemo *prirodan broj* 1, ocjeni "dovoljan" prirodan broj 2, ocjeni "dobar" prirodan broj 3, ocjeni "vrlo dobar" prirodan broj 4 i ocjeni "izvrstan" prirodan broj 5. Na taj način kvalitativni niz ocjena bijektivno preslikavamo u numerički niz prirodnih brojeva, pa – budući da je na skupu prirodnih brojeva definirana operacija zbrajanja – možemo računati aritmetičku sredinu dobivenoga numeričkoga niza i zaokružiti je na najbliži prirodan broj. Jedno od osnovnih svojstava aritmetičke sredine[69] je da se ona nalazi između najmanje i najveće vrijednosti u numeričkom nizu, a to u ovom slučaju znači da je riječ o jednom od prirodnih brojeva između 1 i 5. *Dekodiranjem* dobivenoga prirodnoga broja, tj. "očitavanjem" kojoj je ocjeni pridružen taj prirodan broj, dobivamo spomenutu prosječnu ocjenu.

Opisani postupak *kodiranja* i *dekodiranja* je vrlo čest u statistici, a primjenjuje se u slučajevima kad je niz *kvalitativnih* podataka potrebno (dovoljno reprezentativno) opisati *točno jednim* podatkom. Alternativni načini opisivanja takvih podataka u slučaju redosljednih obilježja mogu biti i određivanje moda ili medijana[70].
Na kraju ove točke navodimo osnovna svojstva aritmetičke sredine koja uvelike mogu koristiti pri gruboj procjeni[71] je li izračunana aritmetička sredina dobar reprezentant numeričkoga niza.

Svojstvo 1. *Svaka* razdioba u kojoj su vrijednosti obilježja pridružene elementima skupa na temelju *intervalne* ili *omjerne* skale ima aritmetičku sredinu.
Svojstvo 2. (*potpunost*) Pri izračunu aritmetičke sredine niza numeričkih podataka obuhvaćene su sve vrijednosti numeričkoga niza.
Svojstvo 3. (*jedinstvenost*) Svaki numerički niz ima *točno jednu* aritmetičku sredinu.
Svojstvo 4. Aritmetička sredina može poslužiti za usporedbu dvaju ili više numeričkih nizova nastalih grupiranjem prema istom kvantitativnom obilježju.
Svojstvo 5. Aritmetička sredina se uvijek nalazi između najmanje (x_{min}) i najveće (x_{max}) vrijednosti numeričkoga niza, tj. vrijede nejednakosti:

$$x_{min} \leq \overline{X} \leq x_{max}.$$

Svojstvo 6. Aritmetička je sredina *jedina* srednja vrijednost takva da je *zbroj odstupanja svih elemenata numeričkoga niza* od nje *uvijek jednak nuli*.[72] Točnije, vrijede sljedeće jednakosti:

---

[68] Napomenimo da se prosječna ocjena praktično *ne* računa ukoliko je barem jedan element odgovarajućega kvalitativnoga niza jednak "nedovoljan". To osobito znaju učenici koji muče muku s "ispravljanjem ocjena".
[69] Vidjeti svojstvo 5. na sljedećoj stranici.
[70] Navedeni se načini u školstvu ipak ne primjenjuju unatoč činjenici da bi cjelokupna učenička populacija nesumnjivo zdušno podržala da zaključna ocjena iz nekoga predmeta bude mod niza pozitivnih ocjena pojedinih područja iz dotičnoga predmeta.
[71] Statističke pokazatelje "finijih" procjena reprezentativnosti aritmetičke sredine navodimo u sljedećem poglavlju.
[72] Iz samoga iskaza svojstva vidljivo je da se ono *ne* odnosi na aritmetičku sredinu izračunatu iz podataka grupiranih u razrede, nego isključivo na *jednostavnu* aritmetičku sredinu ili aritmetičku sredinu izračunatu iz podataka grupiranih prema modalitetima (za koju smo već utvrdili da je zapravo identički jednaka jednostavnoj aritmetičkoj sredini).

$$\sum_{i=1}^{n}(x_i - \overline{X}) = 0 \qquad \sum_{i=1}^{n} f_i \cdot (x_i - \overline{X}) = 0$$

<u>Svojstvo 7.</u> (*svojstvo minimuma zbroja kvadrata*) Neka je $x_1, x_2, \ldots, x_k$ zadani numerički niz podataka. Tada realna funkcija $f(x) = \sum_{i=1}^{n}(x_i - x)^2$ poprima minimalnu vrijednost za $x = \overline{X}$, tj. navedeni zbroj kvadrata poprima minimalnu vrijednost[73] kad je vrijednost varijable x jednaka aritmetičkoj sredini elemenata zadanoga niza.
<u>Svojstvo 8.</u> Aritmetička sredina *<u>nije dobar reprezentant numeričkoga niza</u>* ukoliko u numeričkom nizu *<u>postoje ekstremno male ili velike vrijednosti</u>* promatranoga kvantitativnoga obilježja. (Ovo se svojstvo vrlo neopravdano zanemaruje pri izračunu prosječne (mjesečne, godišnje) plaće svih registriranih djelatnika neke zemlje.)
<u>Svojstvo 9.</u> Aritmetička sredina izračunata na temelju razdiobe frekvencija u kojoj su modaliteti obilježja predstavljeni razredima *uvijek* sadrži pogrešku budući da *izračunata* razredna sredina, korištena za izračun vagane aritmetičke sredine, predstavlja samo približnu zamjenu *stvarne* razredne sredine[74] nekoga razreda.
<u>Svojstvo 10.</u> Problem reprezentativnosti aritmetičke sredine dodatno je izražen u slučaju kada razdioba frekvencija sadrži barem jedan otvoren razred, a osobito kada nije moguće objektivno procijeniti nepoznate granice otvorenih razreda.

3.1.2. *Geometrijska sredina*

Neka je $x_1, x_2, \ldots, x_n$ konačan numerički niz takav da za svaki $i = 1, 2, \ldots, n$ vrijedi $x_i > 0$. *<u>Geometrijska sredina</u>* (oznaka: G) je srednja vrijednost koja se dobije kao n-ti korijen iz umnoška svih članova niza. Ukoliko je barem jedan član numeričkoga niza najviše jednak nuli, geometrijska se sredina takvoga niza *ne definira*.

Za negrupirane podatke geometrijsku sredinu računamo prema formuli:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots\ldots \cdot x_n}\ .$$

<u>Primjer 1.</u> Geometrijska sredina niza 1, 2, 3, 4, 5 jednaka je

$$G = \sqrt[5]{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} \approx 2.605171085.$$

---

[73] Dijeljenjem navedene minimalne vrijednosti ukupnim brojem elemenata numeričkoga niza dobiva se *varijanca* niza i ona se često neprecizno objašnjava kao *prosječno kvadratno odstupanje od aritmetičke sredine*. O varijanci ćemo više reći u sljedećem poglavlju.
[74] *Stvarna* razredna sredina je aritmetička sredina svih numeričkih podataka koji pripadaju dotičnom razredu.

Pretpostavimo sada da su podaci grupirani, tj. da imamo ukupno n različitih modaliteta x_1, x_2, …, x_n. Za svaki $i = 1, 2, \ldots, n$ označimo s f_i apsolutnu frekvenciju modaliteta x_i. Tada je <u>*vagana*</u> (<u>*ponderirana*</u>) <u>*geometrijska sredina*</u> izračunata iz grupiranih podataka dana izrazom

$$G = \sqrt[\sum_{i=1}^{n} f_i]{x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots\ldots \cdot x_n^{f_n}} = \sqrt[\sum_{i=1}^{n} f_i]{\prod_{i=1}^{n} x_i^{f_i}} .$$

<u>Primjer 2.</u> Zadana je razdioba studenata 1. godine stručnoga studija računovodstva i financija na Veleučilištu u Piškorevcima dana 20.3.2008. prema broju članova domaćinstva u kojemu žive.

broj članova domaćinstva	*broj studenata*
2	25
3	20
4	30
5	15
6	10
ukupno:	100

Geometrijska sredina zadane razdiobe jednaka je

$$G = \sqrt[100]{2^{25} \cdot 3^{20} \cdot 4^{30} \cdot 5^{15} \cdot 6^{10}} \approx 3.419482401.$$

U promatranim primjerima osnovni je problem bio što dobiveni rezultat nismo mogli uobičajeno interpretirati. Drugim riječima, mogli smo reći da je prosječna vrijednost svih članova niza u Primjeru 1. približno jednaka 2.6, a prosječan broj članova domaćinstva po jednom studentu iz Primjera 2. približno jednak 3, ali te bismo interpretacije uobičajeno shvaćali kao interpretacije aritmetičke sredine zbog već istaknutoga uobičajenoga poistovjećivanja pojma *prosjek* s aritmetičkom sredinom. Stoga se nameće pitanje u kojim situacijama geometrijsku sredinu možemo točno interpretirati kao prosječnu vrijednost neke pojave. Tipična takva situacija je *računanje prosječne promjene cijena* u nekom promatranom razdoblju[75]. Točnije, promatramo sljedeći opći problem:

<u>*Problem*</u>: Neka se pojava (npr. cijena, mjesečna plaća, broj stanovnika) tijekom razdoblja od ukupno n vremenskih jedinica promijenila ukupno m puta. Označimo te promjene s p_1, p_2, …, p_m i pretpostavimo da su iskazane u postotcima. Odredimo *prosječnu promjenu* $\overline{p}$ promatrane pojave u jednoj vremenskoj jedinici.

<u>*Odgovor*</u>: Neka je C_0 početna vrijednost pojave, a C_n njezina vrijednost na kraju razdoblja od

[75] Takva situacija pripada u analizu vremenskih nizova, o čemu ćemo više reći u posebnom poglavlju.

n vremenskih jedinica. Koristeći formulu za sukcesivnu promjenu osnovne svote[76] dobivamo da istodobno moraju vrijediti sljedeće jednakosti:

$$C_n = C_0 \cdot \left(1 + \frac{\overline{p}}{100}\right)^n$$

$$C_n = C_0 \cdot \left(1 + \frac{p_1}{100}\right) \cdot \left(1 + \frac{p_2}{100}\right) \cdot \ldots \cdot \left(1 + \frac{p_m}{100}\right)$$

Lijeve strane tih jednakosti su jednake, pa takve moraju biti i desne strane. Njihovim izjednačavanjem i sređivanjem dobivenoga izraza dobiva se

$$\overline{p} = 100 \cdot \left[\sqrt[n]{\left(1 + \frac{p_1}{100}\right) \cdot \left(1 + \frac{p_2}{100}\right) \cdot \ldots \cdot \left(1 + \frac{p_m}{100}\right)} - 1\right].$$

U slučaju kad je ukupan broj promjena (m) jednak ukupnom broju razdoblja (n) gornju jednakost možemo zapisati u obliku

$$\overline{p} = \sqrt[n]{(100 + p_1) \cdot (100 + p_2) \cdot \ldots \cdot (100 + p_n)} - 100,$$

pa je u takvim slučajevima *prosječna promjena* pojave u jedinici vremena jednaka razlici *geometrijske sredine* niza $100 + p_1$, $100 + p_2$, …, $100 + p_n$ i broja 100.[77]
Ilustrirajmo netom provedeno razmatranje na primjerima.

<u>Primjer 3.</u> Cijena loživoga ulja se u minula dva mjeseca mijenjala točno četiri puta: najprije je povećana za 0.5%, potom je povećana za daljnjih 0.7%, pa je snižena za 0.2% i naposljetku povećana za još 1.5%. Odredimo *prosječnu mjesečnu promjenu* cijene loživoga ulja. Jedinično vremensko razdoblje je 1 mjesec, pa je ukupan broj jediničnih razdoblja $n = 2$, a ukupan broj promjena cijena $m = 4$. Te promjene iznose redom $p_1 = +0.5$, $p_2 = +0.7$, $p_3 = -0.2$ i $p_4 = +1.5$. (Predznak + označava da je riječ o povećanju cijene, a predznak – označava da je riječ o sniženju cijene.) Tražena prosječna promjena jednaka je

$$\overline{p} = 100 \cdot \left[\sqrt{\left(1 + \frac{p_1}{100}\right) \cdot \left(1 + \frac{p_2}{100}\right) \cdot \left(1 + \frac{p_3}{100}\right) \cdot \left(1 + \frac{p_4}{100}\right)} - 1\right] =$$

$$= 100 \cdot \left[\sqrt{\left(1 + \frac{0.5}{100}\right) \cdot \left(1 + \frac{0.7}{100}\right) \cdot \left(1 + \frac{-0.2}{100}\right) \cdot \left(1 + \frac{1.5}{100}\right)} - 1\right]$$

$$\approx 1.250239$$

---

[76] Preporučuje se ponoviti gradivo postotnoga računa od sto iz kolegija *Gospodarska matematika*.
[77] Brojevi $100 + p_1$, …, $100 + p_n$ u praksi se nazivaju <u>*verižni*</u> (<u>*lančani*</u>) <u>*indeksi*</u>. O njima ćemo nešto više reći u točki 5.5.1.

Uočimo da je, zbog relativno male razlike između navedenih numeričkih vrijednosti, navedena prosječna promjena *približno jednaka*

$$\overline{p} \approx \frac{\sum_{i=1}^{m} p_i}{n} = \frac{p_1 + p_2 + p_3 + p_4}{2} = \frac{0.5 + 0.7 + (-0.2) + 1.5}{2} = 1.25^{78},$$

ali *ne* i aritmetičkoj sredini svih četiriju promjena koja iznosi

$$\overline{p}_a = \frac{0.5 + 0.7 + (-0.2) + 1.5}{4} = 0.625.$$

No, u slučajevima kad između numeričkih vrijednosti promjena postoje veće razlike, razlika između aritmetičke i geometrijske sredine je bitno izraženija, kako pokazuje sljedeći primjer.

Primjer 4. Cijena četvornoga metra ("kvadrata") zemljišta u Lokvama u protekle je dvije godine najprije udvostručena, a potom povećana za još 20%. Odredimo prosječnu godišnju promjenu cijene četvornoga metra toga zemljišta. Iz navedenih podataka vidimo da su numeričke vrijednosti promjena p_1 = +100 i p_2 = +20. Ukoliko bismo traženu prosječnu godišnju promjenu računali kao jednostavnu aritmetičku sredinu (što je i formalno ispravno jer imamo dvije promjene u dvije godine), dobili bismo:

$$\overline{p}_a = \frac{100 + 20}{2} = 60,$$

pa bismo zaključili da se promatrana cijena *povećavala za prosječno* 60% *godišnje*. No, izračunom pomoću navedene formule za prosječnu promjenu dobivamo:

$$\overline{p} = \sqrt{(100 + 100) \cdot (100 + 20)} - 100 \approx 54.91933,$$

pa zaključujemo da se *promatrana cijena povećavala za prosječno* (*približno*) 55% *godišnje*. Provjerimo dobivena razmatranja uzimajući da je početna cijena četvornoga metra zemljišta 100,00 €. Tada je cijena nakon prvoga povećanja 200,00 €, a nakon drugoga 240,00 €. Ukoliko cijenu od 100,00 € povećavamo dva puta po 60%, dobit ćemo najprije 160,00 €, a potom 256,00 €, što je za 16 € (ili 6.67%) više od stvarne krajnje cijene zemljišta. No, ukoliko cijenu od 100,00 € povećavamo dva puta po 54.91933% dobit ćemo najprije 154,92 €, a potom cijenu od 240,00 €, što je i stvarna krajnja cijena.

Netom provedena razmatranja u praksi treba provoditi vrlo oprezno jer je njihov temeljni nedostatak činjenica da prigodom izračuna prosječne promjene zapravo uzimaju u obzir

[78] Navedeni približni izraz, naravno, *nije* aritmetička sredina promatranih promjena jer zbroj svih članova numeričkoga niza ne dijelimo ukupnim brojem tih članova.

jedino početnu i konačnu vrijednost pojave, ali ne i njezine međuvrijednosti, što ponekad može dovesti do potpuno pogrešnih zaključaka. Pogledajmo to na primjeru.

Primjer 5. Cijena automobila "*Ficho*" u protekle se dvije godine mijenjala dva puta: najprije je udvostručena, a potom "prepolovljena" (snižena za 50%). Početna cijena je tada očito jednaka krajnjoj, pa uvrštavanjem $p_1 = +100$, $p_2 = -50$ u formulu za prosječnu promjenu cijene dobivamo da je ta promjena jednaka

$$\overline{p} = \sqrt{(100+100)\cdot(100-50)} - 100 = 0,$$

otkuda slijedi da se *u prosjeku cijena automobila godišnje nije mijenjala* (pogrešno je reći: *prosječna cijena automobila se godišnje nije mijenjala*!). Taj zaključak, međutim, očito nedovoljno reprezentativno opisuje navedeno kretanje cijena automobila jer smo u promatrane dvije godine imali ukupno dvije različite promjene: jedno povećanje i jedno sniženje. Aritmetička sredina je još lošiji pokazatelj:

$$\overline{p}_a = \frac{100+(-50)}{2} = 25,$$

pa bismo zaključili da se *cijena automobila prosječno povećavala za* 25% *godišnje*, što je u suprotnosti s točnom činjenicom da je početna cijena automobila jednaka njegovoj krajnjoj cijeni. Stoga sve vrste sredina treba primjenjivati oprezno i uz dodatne statističke analize, a nikako "šablonski" i prema nekom automatizmu.

3.1.3. *Harmonijska sredina*

Neka je $x_1, x_2, \ldots, x_n$ konačan numerički niz podataka takav da za svaki $i = 1, 2, \ldots, n$ vrijedi nejednakost $x_i > 0$. <u>*Harmonijska sredina*</u> (oznaka: H) navedenoga numeričkoga niza definira se kao *recipročna vrijednost aritmetičke sredine recipročnih vrijednosti svih elemenata niza*.[79]. Za numeričke nizove kojima je barem jedan član najviše jednak nuli harmonijska sredina se *ne definira*.

Sukladno navedenoj definiciji, harmonijska sredina *negrupiranih numeričkih podataka* računa se iz izraza:

$$H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}.$$

[79] Iako se od čitanja naglas ove definicije može zapetljati jezik, navedena je definicija uistinu jedina potpuno ispravna definicija harmonijske sredine.

Primjer 1. Odredimo harmonijsku sredinu niza 2, 3, 6. Ona je jednaka

$$H = \frac{3}{\frac{1}{2}+\frac{1}{3}+\frac{1}{6}} = 3 .$$

Pretpostavimo sada da su podaci grupirani, tj. da imamo ukupno n različitih modaliteta x_1, x_2, …, x_n. Za svaki i = 1, 2, …, n označimo s f_i apsolutnu frekvenciju modaliteta x_i. Tada je *vagana* (*ponderirana*) *harmonijska sredina* izračunata iz grupiranih podataka dana izrazom

$$H = \frac{f_1 + f_2 + \ldots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_n}{x_n}} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}} .$$

Primjer 2. Odredimo harmonijsku sredinu za razdiobu iz Primjera 2. u točki 3.1.2.:

broj članova domaćinstva	*broj studenata*
2	25
3	20
4	30
5	15
6	10
ukupno:	100

Ona je jednaka

$$H = \frac{\sum_{i=1}^{5} f_i}{\sum_{i=1}^{5} \frac{f_i}{x_i}} = \frac{100}{\frac{25}{2}+\frac{20}{3}+\frac{30}{4}+\frac{15}{5}+\frac{10}{6}} = \frac{150}{47} \approx 3.19149 .$$

Kao ni u slučaju geometrijske sredine, ni ovdje interpretacije *Prosječna vrijednost svih članova niza* 2, 3, 6 *jednaka je* 3 i *Prosječan broj članova domaćinstva po jednom studentu približno je jednak* 3 ne zvuče kao prirodne interpretacije harmonijske sredine.[80] Zbog toga hamonijsku sredinu koristimo prigodom rješavanja problema u kojima se pojavljuju *obrnuto razmjerne veličine*. U nastavku dajemo dva primjera takvih problema.

[80] Prilikom čitanja ovih dviju interpretacija, ali bez naznake o kojoj se srednjoj vrijednosti radi, svaki nestatistički nastrojen *homo sapiens* odmah bi pomislio da je riječ o interpretaciji aritmetičke sredine.

Primjer 1. Četiri radnika rade istovrstan posao, ali različitim brzinama. U četiri sata prvi je radnik izradio 48 komada proizvoda, drugi 80 komada, treći 96 komada, a četvrti 120 komada. Odredimo prosječno vrijeme izrade jednoga proizvoda *za sve radnike zajedno*, te prosječan broj izrađenih proizvoda po jednoj osmosatnoj radnoj smjeni.

Za izradu *jednoga* proizvoda prvi radnik utroši $t_1 = \frac{4}{48} = \frac{1}{12}$ sati, drugi $t_2 = \frac{4}{80} = \frac{1}{20}$ sati, treći $t_3 = \frac{4}{96} = \frac{1}{24}$ sati, a četvrti $t_4 = \frac{4}{120} = \frac{1}{30}$ sati. Mi tražimo vrijeme t za koje će četvorica radnika proizvesti ukupno 4 proizvoda. Za navedeno vrijeme t prvi će radnik proizvesti ukupno $\frac{t}{\frac{1}{12}}$ proizvoda, drugi $\frac{t}{\frac{1}{20}}$ proizvoda, treći $\frac{t}{\frac{1}{24}}$ proizvoda, a četvrti $\frac{t}{\frac{1}{30}}$ proizvoda, pa će sva četvorica u vremenu t proizvesti ukupno $\frac{t}{\frac{1}{12}} + \frac{t}{\frac{1}{20}} + \frac{t}{\frac{1}{24}} + \frac{t}{\frac{1}{30}} = t \cdot \left(\frac{1}{\frac{1}{12}} + \frac{1}{\frac{1}{20}} + \frac{1}{\frac{1}{24}} + \frac{1}{\frac{1}{30}} \right)$ proizvoda. Taj broj treba biti jednak 4, pa dobivamo jednadžbu

$$t \cdot \left(\frac{1}{\frac{1}{12}} + \frac{1}{\frac{1}{20}} + \frac{1}{\frac{1}{24}} + \frac{1}{\frac{1}{30}} \right) = 4$$

iz koje je

$$t = \frac{4}{\frac{1}{\frac{1}{12}} + \frac{1}{\frac{1}{20}} + \frac{1}{\frac{1}{24}} + \frac{1}{\frac{1}{30}}}.$$

Prema definiciji harmonijske sredine, odavde izravno slijedi da je traženo prosječno vrijeme izrade jednoga proizvoda za sva četiri radnika zajedno jednako *harmonijskoj sredini* niza t_1, t_2, t_3, t_4, tj. niza $\frac{1}{12}, \frac{1}{20}, \frac{1}{24}, \frac{1}{30}$:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{t_i}} = \frac{4}{\frac{1}{\frac{1}{12}} + \frac{1}{\frac{1}{20}} + \frac{1}{\frac{1}{24}} + \frac{1}{\frac{1}{30}}} = \frac{4}{12+20+24+30} = \frac{4}{86} = \frac{2}{43} \text{ radna sata.}$$

Provjerimo dobiveni rezultat. Za $\frac{2}{43}$ sata prvi će radnik proizvesti prosječno $\frac{\frac{2}{43}}{\frac{1}{12}}=\frac{24}{43}$ proizvoda, drugi prosječno $\frac{\frac{2}{43}}{\frac{1}{20}}=\frac{40}{43}$ proizvoda, treći prosječno $\frac{\frac{2}{43}}{\frac{1}{24}}=\frac{48}{43}$ proizvoda, a četvrti prosječno $\frac{\frac{2}{43}}{\frac{1}{30}}=\frac{60}{43}$ proizvoda. Stoga će sva četvorica radnika za $\frac{2}{43}$ sata proizvesti prosječno $\frac{24}{43}+\frac{40}{43}+\frac{48}{43}+\frac{60}{43}=\frac{172}{43}=4$ proizvoda. Odatle izravno slijedi da će za 8 radnih sati promatrana četvorica radnika proizvesti prosječno $\frac{8}{\frac{2}{43}}$ = 172 proizvoda.

<u>Primjer 2.</u> U donjoj su tablici navedeni podaci o prosječnoj prodajnoj cijeni udžbenika iz *Poslovne statistike* autora Statha Isticka tijekom 2007. godine, te strukturi vrijednosti prodaje u trima prodajnim područjima

prodajno područje	*prosječna prodajna cijena* [kn]	*struktura vrijednosti prodaje* [%]
sjever	290,00	25
središnja regija	300,00	35
jug	295,00	40

Odredimo *prosječnu* prodajnu cijenu za sva tri prodajna područja zajedno. Podsjetimo se, *vrijednost prodaje* jednaka je umnošku ukupnoga broja prodanih proizvoda i cijene jednoga proizvoda, dok je *prosječna prodajna cijena* jednaka količniku vrijednosti prodaje i ukupnoga broja prodanih proizvoda. Budući da su postoci vrijednosti prodaje upravno razmjerni vrijednostima prodaje svakoga pojedinoga područja, traženu prosječnu prodajnu cijenu izračunat ćemo kao *vaganu harmonijsku sredinu* shvaćajući prosječne prodajne cijene kao modalitete, a strukture vrijednosti prodaje kao pondere (težine):

$$H=\frac{\sum_{i=1}^{3} p_i}{\sum_{i=1}^{3}\frac{p_i}{x_i}}=\frac{25+35+40}{\frac{25}{290}+\frac{35}{300}+\frac{40}{295}}\approx 295,45\text{ kn.}$$

Zaključno istaknimo da među trima sredinama numeričkoga niza vrijedi poznata <u>*A – G – H nejednakost*</u>, tj. ako su *A*, *G* i *H* redom aritmetička, geometrijska i harmonijska sredina *istoga*

numeričkoga niza, onda vrijedi nejednakost:

$$H \leq G \leq A.$$

Znakovi jednakosti vrijede ako i samo ako je odgovarajući numerički niz *konstantan*, tj. ako je $a_k = c$, za neki strogo pozitivan realan broj $c \in \langle 0, +\infty \rangle$ i za svaki $k = 1, 2, \ldots, n$.

3.1.4. *Kvantili*

Neka je $x_1, x_2, \ldots, x_m$ niz modaliteta *kvantitativnoga* ili *kvalitativnoga redosljednoga* obilježja. Svaki takav niz možemo *urediti* tako da podatke ili poredamo po veličini (od najmanjega do najvećega ili obrnuto) ako je riječ o modalitetima kvantitativnoga obilježja ili poredamo prema stupnjevima intenziteta mjerenoga svojstva ako je riječ o modalitetima kvalitativnoga redosljednoga obilježja. Na taj način dobivamo *uređeni niz podataka*.

Primjer 1. a) Niz kvantitativnih modaliteta 10, 7, 14, 5, 20, 8 možemo zapisati kao uređeni niz 5, 7, 8, 10, 14, 20 ili kao uređeni niz 20, 14, 10, 8, 7, 5.
b) Niz kvalitativnih redosljednih modaliteta SSS, VSS, VŠS, VSS, SSS, VŠS, VŠS, VSS, SSS možemo zapisati kao uređeni niz SSS, SSS, SSS, VŠS, VŠS, VŠS, VSS. VSS, VSS ili kao uređeni niz VSS, VSS, VSS, VŠS, VŠS, VŠS, SSS, SSS, SSS.

U daljnjem, radi određenosti, pretpostavljamo da je *konačan* niz kvantitativnih, odnosno kvalitativnih redosljednih modaliteta uređen *uzlazno*, tj. od najmanjega (najlošijega) prema najvećemu (najboljemu), te promatramo isključivo tako uređene *konačne* nizove. Problem koji ćemo razmatrati jest:

Problem 1: Zadani uređeni *konačan* niz podijeliti na *jednakobrojne dijelove* zadržavajući pritom svojstvo uređenosti niza. Formalnije i preciznije, *uređeni* skup $S = (x_1, x_2, \ldots, x_m)$ podijeliti na n *uređenih* podskupova $S_1, S_2, \ldots, S_n$ tako da vrijede sljedeća svojstva:

1.) $\mathrm{card}(S_1) = \mathrm{card}(S_2) = \ldots = \mathrm{card}(S_n)$, tj. svi podskupovi $S_1, S_2, \ldots, S_n$ imaju jednako mnogo elemenata;

2.) za svaki $i = 2, \ldots, n$ svaki element skupova $S_1, \ldots S_{i-1}$ nije veći od svakoga elementa skupa S_i ili, ekvivalentno, za svaki $x \in \bigcup_{k=1}^{i-1} S_k$ i svaki $y \in S_i$ vrijedi nejednakost $x \leq y$;

3.) svaka dva uzastopna člana niza $S_1, S_2, \ldots, S_n$ imaju najviše jedan zajednički element, dok nesusjedni članovi niza nemaju niti jedan zajednički element;

4.) $\bigcup_{i=1}^{n} S_i = S_1 \cup S_2 \cup \ldots \cup S_n = S$, tj. unija[81] svih uređenih podskupova treba biti jednaka polaznom uređenom skupu.

[81] Navedenu uniju dobijemo tako da svaki od podskupova najprije shvatimo kao "običan" skup, napravimo uobičajenu operaciju unije skupova uz dozvoljeno ponavljanje elemenata, a potom uredimo dobiveni skup.

Odmah primijetimo da u postavci navedenoga problema **nema** uobičajenoga zahtjeva da skupovi na koje dijelimo polazni uređeni skup nemaju niti jedan zajednički element, tj. da su međusobno disjunktni. Taj zahtjev je namjerno izostavljen jer dozvoljavamo mogućnost da se isti član niza pojavi u dva uzastopna podskupa. Formalno se ne navodi niti bilo kakva veza između brojeva m i n, no, može se pokazati da iz uvjeta 1.), 3.) i 4.) slijedi nejednakost $n \leq m$. Ona se intuitivno može objasniti kao činjenica da uređeni skup od m elemenata možemo podijeliti na najviše m dijelova tako da svi dijelovi imaju jednako mnogo elemenata.

Primjer 2. Sukladno gore navedenim uvjetima, podijelimo uređeni niz 1, 2, 3, 4 na dva jednakobrojna dijela. Tražimo dva uređena podskupa S_1 i S_2 uređenoga skupa $S = (1, 2, 3, 4)$ takva da svaki element skupa S_1 nije veći od svakoga elementa skupa S_2, da skupovi S_1 i S_2 imaju najviše jedan zajednički element, te da njihova unija daje skup S. Lako se vidi da je jedina mogućnost takve podjele $S_1 = (1, 2)$ i $S_2 = (3, 4)$.

Primjer 3. Sukladno uvjetima problema, uređeni niz iz Primjera 1. podijelimo na tri jednakobrojna dijela. Sada tražimo tri uređena podskupa S_1, S_2 i S_3 takva da svaki element skupa S_1 nije veći od svakoga elementa skupa S_2 i svaki element skupa S_2 nije veći od svakoga elementa skupa S_3, te da unija svih triju uređenih skupova daje skup S. Nije teško vidjeti da je jedina mogućnost takve podjele $S_1 = (1, 2)$, $S_2 = (2, 3)$ i $S_3 = (3, 4)$. Uočimo da skupovi S_1 i S_2 imaju zajednički element 2, skupovi S_2 i S_3 zajednički element 3, dok skupovi S_1 i S_3 nemaju zajedničkih elemenata.

Primjer 4. Sukladno uvjetima problema, uređeni niz 1, 2, 3 podijelimo na dva jednakobrojna dijela. U ovom se slučaju dobiva $S_1 = (1, 2)$ i $S_2 = (2, 3)$. Uočimo da skupovi S_1 i S_2 imaju zajednički element 2.

Netom razmatrani problem poopćuje se u matematičkoj statistici, i to prigodom proučavanja svojstava *kumulativne funkcije distribucije slučajne varijable*[82]. Ondje se Problem 1. svodi na to da se za zadanu slučajnu varijablu X i zadani realan broj $p \in \langle 0, 1 \rangle$ odredi vrijednost x takva da je *vjerojatnost da je vrijednost slučajne varijable X strogo manja od x jednaka najviše p*, *a vjerojatnost da je vrijednost slučajne varijable X strogo veća od x jednaka najviše* $1 - p$. Može se pokazati da tražena vrijednost x uistinu postoji i da je jedinstvena, te se ona naziva *p–kvantil*. *Kvantili* su vrlo značajni i u deskriptivnoj statistici, ali njihov osnovni nedostatak je nemogućnost točne interpretacije: deskriptivna statistika "ne poznaje" termin *slučajna varijabla*, pa se gornja, točna definicija p – kvantila mora preinačiti, čime se gubi na preciznosti (i istinitosti) interpretacije. Približni deskriptivni "ekvivalent" p – kvantila je tzv. (*per*)*centil* i njega opisujemo u nastavku.

Percentili ili *centili* su *položajne* vrijednosti koje uređeni niz dijele na ukupno 100 jednakobrojnih dijelova. Ima ih ukupno $100 - 1 = 99$ i označavaju se s $P_1, P_2, P_3, \ldots, P_{98}, P_{99}$. Približna formula za izračun k – toga percentila P_k za uređeni niz $x_1, x_2, \ldots, x_m$ dana je s:

---

[82]U matematičkoj je statistici *slučajna varijabla* bilo koja realna funkcija čije je područje definicije određeni vjerojatnosni prostor. Ovdje je spominjemo isključivo u kontekstu definiranja kvantila.

$$P_k = x_{c+1} + d \cdot (x_{c+2} - x_{c+1}), \text{ za } k = 1, 2, \ldots, 98, 99,$$

gdje je c cijeli, a d decimalni dio broja $\frac{k}{100} \cdot (m - 1)$.[83] (Približna i matematički neprecizna!)

Interpretacija k – toga percentila P_k obično je jedna od sljedećih interpretacija:

- *prvih k% članova niza ima vrijednost manju ili jednaku P_k.*
- *(100 – k)% članova niza ima vrijednost veću ili jednaku P_k.*

Nerijetko se (opet približno točno i matematički neprecizno) kaže i da k – ti percentil P_k *odvaja k% najmanjih članova niza*, odnosno *(100 – k)% najvećih članova niza*.

Primjer 5. Prikupljanjem podataka o navršenim godinama radnoga staža za ukupno 60 zaposlenika tvrtke "*Mućkalić – commerce* d.o.o" iz Klokočevca i odgovarajućim izračunom dobivena je vrijednost

$$P_5 = 7.25$$

Dobivenu vrijednost možemo interpretirati na jedan od sljedećih dvaju načina:

- 5% *promatranih zaposlenika ima najviše* 7.25 *godina navršenoga radnoga staža*;
- 95% *promatranih zaposlenika ima barem* 7.25 *godina navršenoga radnoga staža*.

Budući da znamo ukupan broj zaposlenih (60), navedene interpretacije zapravo znače da:

- *približno* 3 zaposlenika promatrane tvrtke imaju najviše 7 godina navršenoga radnoga staža;
- *približno* 57 zaposlenika ima barem 8 godina navršenoga radnoga staža.

Odmah istaknimo da ovdje prilog "približno" *ne* znači da npr. točno 3.14 zaposlenika ima najviše 7 godina navršenoga staža (pa smo taj broj zaokružili na najbliži cijeli broj). Greška je bitno veća: moguće je da točno 2 zaposlenika ili točno 4 zaposlenika promatrane tvrtke imaju najviše 7 godina navršenoga radnoga staža. Zbog toga je u ovakvim interpretacijama bolje izbjegavati navođenje konkretnih apsolutnih frekvencija.

Primjer 6. Prikupljanjem podataka o visinama (iskazanima kao cijeli broj cm) za ukupno 30 učenika 8.a razreda Osnovne škole "*Štefica Šniclek*" iz Careva Sela i odgovarajućim izračunom dobivena je vrijednost

$$P_{30} = 170.66$$

Dobivenu vrijednost možemo interpretirati na jedan od sljedećih dvaju načina:

[83] Upravo navedenu formulu koristi MS Excel prigodom izračuna vrijednosti pojedinih percentila.

- 30% *promatranih učenika visoko je najviše* 170.66 cm;
- 70% *promatranih zaposlenika visoko je barem* 170.66 *cm*.

U ovom je slučaju obilježje "*visina*" *kvantitativno diskretno obilježje* (jer zanemarujemo milimetre, mikrometre itd.), pa navedene interpretacije možemo preformulirati ovako:

- 30% *promatranih učenika visoko je najviše* 170 cm;
- 70% *promatranih zaposlenika visoko je barem* 171 *cm*.

Kao i u Primjeru 5., budući da znamo ukupan broj učenika, navedene interpretacije zapravo znače da je *približno* 9 promatranih učenika visoko najviše 170 cm, dok visina svakoga od preostalih 21 učenika iznosi barem 171 cm. Međutim, moguće je i da točno 10 promatranih učenika bude visoko najviše 170 cm, pa je opet primjerenije izostaviti interpretaciju s apsolutnim frekvencijama.

U praksi se vrlo često koriste dvije posebne vrste percentila: <u>*decili*</u> i <u>*kvartili*</u>. <u>*Decili*</u> su vrijednosti koje uređeni niz modaliteta dijele na 10 jednakobrojnih dijelova. Ima ih ukupno 10 – 1 = 9 i označavaju se s D_1, D_2, …, D_9. Nije teško vidjeti da za svaki i = 1, 2, …, 8, 9 vrijedi jednakost

$$D_i = P_{10 \cdot i}$$

koja, zapravo, tvrdi da je i – ti decil identički jednak (10 · i) – tom percentilu. Ovo vrlo korisno svojstvo omogućuje da se interpretacija svakoga decila poistovjeti s interpretacijom njemu odgovarajućega percentila.

<u>Primjer 7.</u> Prikupljanjem podataka o ostvarenom uspjehu (1 – nedovoljan, 2 – dovoljan, …, 5 – izvrstan) svih studenata 1. godine stručnoga studija računovodstva i financija Veleučilišta u Špičkovini akademske godine 2006/2007. koji su kolegij *Poslovna statistika* polagali putem dvaju kolokvija i odgovarajućim izračunom dobivena je vrijednost

$$D_4 = 1.25$$

Riječ je o 4. decilu, a toj vrijednosti odgovara 10 · 4 = 40. percentil. Stoga dobivenu vrijednost možemo teorijski interpretirati na jedan od sljedećih dvaju načina:

- 40% promatranih studenata dobilo je ocjenu najviše 1.25;
- 60% promatranih studenata dobilo je ocjenu barem 1.25.

Budući da da dobivanje ocjene 1 znači da kolokviji nisu položeni, navedene interpretacije možemo preformulirati u sljedeće:

- 40% *promatranih studenata nije položilo kolegij Poslovna statistika putem kolokvija*;
- 60% *promatranih studenata je položilo kolegij Poslovna statistika putem kolokvija*.

Primjer 8. Prikupljanjem podataka o stupnju stručne spreme (uz uobičajeno rangiranje stupnjeva stručnih sprema) za svakoga zaposlenika tvrtke "*Pijetlović* d.o.o." i odgovarajućim izračunom dobivena je vrijednost

$$D_8 = \text{VŠS}.$$

Riječ je o 8. decilu, a toj vrijednosti odgovara $10 \cdot 8 = 80$. percentil. Stoga dobivenu vrijednost možemo teorijski interpretirati na jedan od sljedećih dvaju načina:

- 80% promatranih zaposlenika ima višu stručnu spremu ili stručnu spremu manju od više stručne spreme;
- 20% promatranih zaposlenika ima višu stručnu spremu ili stručnu spremu veću od više stručne spreme.

Budući da je jedino visoka stručna sprema veća od više stručne spreme, drugu interpretaciju možemo pojednostavniti ovako: 20% *promatranih zaposlenika ima višu ili visoku stručnu spremu*. U ovom je slučaju ta interpretacija jednostavnija i razumljivija od prve.

Uočimo da se u Primjerima 3. i 4. radilo o određivanju percentila za niz modaliteta *kvalitativnoga redosljednoga obilježja*.

Još jedna od spomenutih posebnih vrsta percentila su *kvartili*. *Kvartili* su *položajne* vrijednosti koje uređeni numerički niz dijele na ukupno 4 jednakobrojna dijela. Ima ih ukupno $4 - 1 = 3$, a imaju svoja posebna imena i oznake:

- *prvi* ili *donji kvartil* (oznaka: Q_1);
- *drugi kvartil* ili *medijan* (oznake: Q_2, M_e);
- *treći* ili *gornji kvartil* (oznaka: Q_3)

Nije teško vidjeti da za svaki $i = 1, 2, 3$ vrijedi jednakost

$$Q_i = P_{25 \cdot i}$$

koja, zapravo, tvrdi da je i – ti kvartil identički jednak $(25 \cdot i)$ – tom percentilu. Ovo vrlo korisno svojstvo omogućuje da se interpretacija svakoga kvartila poistovjeti s interpretacijom njemu odgovarajućega percentila. Istaknimo da se u slučajevima u kojima, zbog postojanja ekstremno malih ili ekstremno velikih vrijednosti, aritmetička sredina nije reprezentativan pokazatelj statističkoga niza kao "zamjenski" pokazatelji rabe upravo kvartili, a poglavito medijan.

Primjer 9. Zadan je uređeni niz prosječnih mjesečnih plaća (iskazanima u kn) svih stalno zaposlenih djelatnika poduzeća "*Schwindl–trade d.o.o.*" iz Šuplje Lipe u razdoblju od 01.01.2007. – 31.12.2007.:

1.852,95; 1.884,76; 2.248,62; 2.417,25; 2.646,27; 2.718,12; 2.947,83; 3.085,45; 3.090,31; 50.000,64; 60.000,75

Odgovarajućim izračunima se dobiva:

- prvi ili donji kvartil: Q_1 = 2.332,935;
- medijan: Q_2 = 2.718,12;
- treći ili gornji kvartil: Q_3 = 3.087,88;
- jednostavna aritmetička sredina: $\overline{X}$ = 12.081,18 (kn)
-

Te vrijednosti možemo redom interpretirati na sljedeći način:

- 25% *promatranih djelatnika ima prosječnu mjesečnu plaću manju od* 2.332,94 kn;
- 50% *promatranih djelatnika ima prosječnu mjesečnu plaću najviše* 2.718,12 kn;
- 25% *promatranih djelatnika ima prosječnu mjesečnu plaću veću od* 3.087,88 kn;
- *Prosječna mjesečna plaća svakoga djelatnika navedene tvrtke iznosi* 12.081,18 kn.

Iz zadanoga uređenoga niza lako možemo vidjeti da točno 6 promatranih djelatnika ima prosječnu mjesečnu plaću najviše 2.718,12 kn, dok 6 promatranih djelatnika ima prosječnu mjesečnu plaću barem 2.718,12 kn. (1 djelatnik pripada objema kategorijama.) To znači da je iznos 2.718,12 *točno* rješenje Problema 1. za navedeni niz plaća jer uistinu dijeli niz na dva jednakobrojna dijela:

(1.852,95; 1.884,76; 2.248,62; 2.417,25; 2.646,27; 2.718,12) i (2.718,12; 2.947,83; 3.085,45; 3.090,31; 50.000,64; 60.000,75).

I sva tri kvartila su *točno* rješenje Problema 1. za navedeni niz plaća jer uistinu dijele niz na četiri jednakobrojna dijela:

(1.852,95; 1.884,76; 2.248,62), (2.417,25; 2.646,27; 2.718,12), (2.718,12; 2.947,83; 3.085,45) i (3.090,31; 50.000,64; 60.000,75).

Zbog postojanja ekstremno velikih vrijednosti (50.000,64 i 60.000,75) u navedenom nizu plaća, očito je da su u ovom slučaju sva tri kvartila bitno reprezentativnije vrijednosti od aritmetičke sredine. Stoga je u statističkoj analizi takvih nizova, kad god je to moguće, primjerenije odrediti barem jedan kvartil (poželjno: sva tri), negoli aritmetičku sredinu.

Svi percentili, pa posebno i sva tri kvartila, mogu se određivati iz negrupiranih i grupiranih podataka. Određivanje pomoću grupiranih podataka zasniva se deskriptivnom analogonu funkcije distribucije, a to je kumulativni niz "manje od" kojega smo već upoznali. Na primjeru ćemo pokazati kako se iz grupiranih podataka određuje medijan.

Može se dogoditi da kod parnoga broja elemenata niza prvi središnji element pripada jednoj, a drugi središnji element drugoj frekvenciji kumulativnoga niza. U tom slučaju medijan predstavlja aritmetičku sredinu dviju njima odgovarajućih vrijednosti diskretnoga obilježja.

Primjer 10. Zadana je razdioba svih studenata 1. godine studija smjehologije na Višoj uzaludnoj školi u Špičkovini (akademska godina 2007/2008.) prema broju položenih jednosemestralnih kolegija[84]:

broj položenih jednosemestralnih kolegija	*broj studenata*
0	7
1	12
2	17
3	25
4	21
5	18
ukupno:	100

izvor: studentska služba Više uzaludne škole u Špičkovini

Odredimo medijan ovako grupiranoga niza podataka i objasnimo dobiveni rezultat. Na ranije opisani način formiramo kumulativni niz apsolutnih frekvencija "manje od":

broj položenih jednosemestralnih kolegija (x_i)	*broj studenata* (f_i)	*kumulativni niz apsolutnih frekvencija "manje od"*
0	7	7
1	12	19
2	17	36
3	25	61
4	21	82
5	18	100
ukupno:	100	

Promatrani statistički niz je već uređen uzlazno. On ima ukupno 100 članova. Stoga se njegovi središnji članovi nalaze na mjestima $\frac{100}{2}$ i $\frac{100}{2}+1$, tj. na 50. i 51. mjestu. Iz stupca u kojemu su navedene kumulativne frekvencije "očitavamo" da su na mjestima 37., 38., 39., …, 49., 50., 51., 52., …, 59., 60. i 61. uređenoga niza napisane "trojke". Prema tome, $x_{50} = x_{51} = \ = 3$, pa je medijan jednak 3, što bi značilo da je 50% *studenata položilo najviše* 3 *ispita*, *a* 50% *studenata položilo je najmanje* 3 *ispita*. Ta je interpretacija potpuno točna: na prvih 50 mjesta imamo najprije 7 nula, pa 12 jedinica, pa 17 dvojki i 14 trojki, dok na preostalih 50 mjesta imamo 11 trojki, 21 četvorku i 18 petica.

Na sličan se način iz grupiranih podataka određuju i preostala dva kvartila. Detalje, kao i način određivanja kvartila iz podataka grupiranih u razrede ovdje izostavljamo.

Na kraju ove točke dajemo pregled najvažnijih svojstava medijana kao svojevrsnoj položajnoj alternativi aritmetičke sredine:

---

[84] Usporediti s Primjer 2., točka 3.1.1. Aritmetička sredina.

Svojstvo 1. (*jedinstvenost medijana*) U svakoj razdiobi prema kvantitativnomu ili kvalitativnom redosljednom obilježju postoji *točno jedan* medijan.
Svojstvo 2. Medijan se nalazi *između najmanje i najveće vrijednosti* obilježja u razdiobi.
Svojstvo 3. Na vrijednost medijana *ne utječu ekstremne vrijednosti* obilježja u razdiobi.
Svojstvo 4. Medijan se bez teškoća može odrediti i u razdiobama kod kojih se pojavljuju otvoreni razredi (osim ako je razred kojem pripada medijan otvoren).
Svojstvo 5. Medijan je srednja vrijednost osobito primjerena za opis izrazito *asimetričnih* razdioba.

3.1.5. *Mod*

Pri analizi statističkih nizova vrlo često se najprije određuje najčešći modalitet koji se pojavljuje u nizu (tj. modalitet s najvećom apsolutnom ili relativnom frekvencijom). Takav modalitet naziva se *mod*, označava s M_o i pripada u *položajne srednje vijednosti*. Već iz same definicije moda očito je da se on može odrediti *neovisno o tipu obilježja*, tj. mod je jedina srednja vrijednost koja se može odrediti i za kvalitativna i za kvantitativna obilježja. Dakako, nužan uvjet za postojanje moda jest pojavljivanje barem dva jednaka modaliteta u statističkom nizu.

Oprez: Klasična pogreška prigodom interpretacije modova jest interpretiranje *najveće apsolutne* ili *relativne frekvencije* kao moda pripadnoga osnovnoga skupa. To nije točno: *mod* je isključivo *najčešći modalitet* statističkoga obilježja, a određuje se pomoću najveće apsolutne ili relativne frekvencije (za negrupirana obilježja ili obilježja grupirana prema modalitetima).

Primjer 1. Promatramo skup svih automobila parkiranih u Ratkajevu prolazu 01.04.2008. u 16:30 sati i dijelimo njegove elemente prema obilježju *marka automobila*. Analizom dobivenoga osnovnoga skupa utvrdili smo da je mod promatranoga obilježja BMW. Prema tome, među svim promatranim automobilima ima najviše automobila marke BMW. Iako je takva interpretacija potpuno točna, ona ne mora dobro opisivati pripadni osnovni skup. Npr. ako promatrani statistički skup tvore 2 BMW-a, 1 "Mercedes", 1 "Škoda Octavia", 1 "Opel Corsa", 1 "Grand Cherokee", 1 "Yugo" i 1 "Fićo", onda opis "*Među promatranim automobilima ima najviše automobila marke BMW*" očito nedovoljno dobro opisuje "šarolikost" skupa promatranih automobila. No, ako promatrani statistički skup tvore 6 BMW-a i dvije "Opel Corse", onda navedena interpretacija prilično dobro opisuje taj skup.

Primjer 2. Neto-plaće (iskazane u kn) svih djelatnika knjigovodstvenoga servisa "*Knjiži kod mene*" iz Košara isplaćene u siječnju ove godine tvore niz 1.800,00; 1.800,00; 1.800,00; 1.800,64; 1.800,64; 1.820,25; 1820,25; 30.000,75. Mod promatranoga niza je 1.800,00, što znači da je najčešća plaća promatranih djelatnika 1.800,00 kn. Ta vrijednost dobro opisuje pripadni osnovni skup jer plaće 7 od ukupno 8 djelatnika odstupaju od moda za najviše 1,125%.

Primjer 3. Neto–plaće (iskazane u kn) svih djelatnica knjigovodstvenoga servisa "*Đ. & J.* d.o.o." iz Grudnjaka isplaćene u veljači ove godine tvore niz 1.800,00; 1.800,00; 4.928,50; 5.103,75; 5.224,47; 7.842,86; 7.915,92; 10.165,78 i 10.201,54. I u ovome je slučaju mod promatranoga niza jednak 1.800,00 (interpretacija je ista kao i u Primjeru 2.), no, u ovom slučaju on loše opisuje pripadni osnovni skup jer plaće 7 od ukupno 9 djelatnika odstupaju od moda za barem 173,806%.

Jedno od osnovnih svojstava moda je da *ne mora nužno biti jedinstven*. Naime, može se dogoditi da barem dva modaliteta imaju jednake apsolutne frekvencije, a da su apsolutne frekvencije svih ostalih modaliteta strogo manje od njih. Stoga razlikujemo *unimodalne razdiobe* (razdiobe koja imaju točno jedan mod), *bimodalne razdiobe* (razdiobe koja imaju točno dva moda) i *multimodalne razdiobe* (razdiobe koje imaju barem tri moda).

Primjer 4. **a)** Uređeni niz 1, 2, 2, 3, 3, 4, 5 ima točno dva moda: 2 i 3. Stoga je taj niz bimodalan.
b) Uređeni niz 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5 ima točno tri moda: 1, 2 i 3. Stoga je taj niz multimodalan.
Uočimo da modovi dobro opisuju oba navedena niza jer je u svakom od njih više od polovice elemenata niza jednako nekomu od modova.

Kao i sve ostale srednje vrijednosti, i modovi se mogu određivati iz negrupiranih i grupiranih podataka, ali i iz grafičkih prikaza razdioba, poglavito Paretova grafikona koji kao prvu kategoriju na osi *Ox* navodi upravo modalnu vrijednost. Iako se ova vrijednost vjerojatno određuje najjednostavnije od svih srednjih vrijednosti, treba istaknuti da su slučajevi u kojima mod dobro opisuje pripadni osnovni skup relativno rijetki. Također, jedan od ozbiljnijih nedostataka većine računalnih statističkih potpora (npr. MS Excel) je nemogućnost "ispisa" svih modova pojedine razdiobe. Ukoliko uopće postoji mogućnost određivanja moda primjenom neke funkcije, njezina izlazna vrijednost je prvi mod po veličini, tj. najmanji ili najveći mod, pa time zapravo gubimo mogućnost utvrđivanja je li neki statistički niz unimodalan, bimodalan ili multimodalan.

Sve navedeno samo potvrđuje da problem opisa osnovnoga skupa ili statističkoga niza pomoću bilo koje od potpunih ili položajnih vrijednosti nije nimalo trivijalan i da obavezno zahtijeva dodatnu statističku analizu reprezentativnosti određenih vrijednosti. O nekim pokazateljima reprezentativnosti govorimo u sljedećoj točki.

3.2. *Mjere raspršenja (disperzije)*

Izračun neke od potpunih ili položajnih srednjih vrijednosti predstavlja tek jedan korak u statističkoj analizi nekoga obilježja. Sljedeći je korak analiza *reprezentativnosti* izračunate srednje vrijednosti, odnosno, pojednostavljeno rečeno, određivanje *pokazatelja* koliko izračunata srednja vrijednost dobro opisuje elemente osnovnoga skupa. Ti pokazatelji gotovo su jednako važni kao i izračunate srednje vrijednosti. Ilustrirajmo to na primjeru.

Primjer 1. Tvrtka "*Super Silva* d.o.o." iz Duhova zapošljava ukupno 10 stalno zaposlenih djelatnika. Troje djelatnika ima prosječnu mjesečnu neto–plaću 7.800,00 kn, četvero prosječnu mjesečnu neto-plaću 8.000,00 kn, a troje prosječnu mjesečnu neto–plaću 8.200,00 kn. Stoga prosječna mjesečna neto–plaća po zaposleniku u toj tvrtki iznosi 8.000,00 kn.

Tvrtka "*Super Silvio* d.o.o." iz Duhića također zapošljava ukupno 10 stalno zaposlenih djelatnika. 9 djelatnika ima prosječnu mjesečnu neto–plaću 3.000,00 kn, a jedan djelatnik prosječnu mjesečnu neto–plaću 53.000,00 kn. Prosječna mjesečna neto–plaća po zaposleniku u promatranoj tvrtki također iznosi 8.000,00 kn.

Isključivo na temelju *izračunatih prosječnih mjesečnih neto–plaća*, tj. *bez poznavanja originalnih podataka o prosječnim mjesečnim neto–plaćama* nameće se zaključak da među promatranim tvrtkama nema razlika u modalitetima obilježja "*prosječna mjesečna plaća*", što je očito netočno.

Kako bi se moglo što bolje opisati *stupanj varijabilnosti* statističkih podataka, definiraju se odgovarajući statistički pokazatelji pod zajedničkim imenom *mjere raspršenja* (stupnja varijabilnosti; disperzije). Mjere raspršenja općenito mogu biti *apsolutne* i *relativne*.

Apsolutne mjere raspršenja su *raspon varijacije, interkvartil, interpercentilni razmak, srednje apsolutno odstupanje* i *varijanca* (odnosno, iz nje izvedena *standardna devijacija*). Zajedničko obilježje svih navedenih mjera jest iskazivanje u jedinicama mjere obilježja statističkog skupa (kn, €, kg, cm itd.)

Relativne mjere raspršenja obično se izražavaju u postocima, a dvije najčešće su *koeficijent kvartilne devijacije i koeficijent varijacije*.

Odmah napomenimo da *izbor mjere* raspršenja u pravilu ovisi o mjernim svojstvima statističkih varijabli, tj. *ne postoji* "univerzalna" mjera raspršenja koja bi se mogla primijeniti u svim slučajevima. Ukoliko je promatrano obilježje *kvantitativno*, primjereno je izračunati i interpretirati sve navedene mjere. Ukoliko je promatrano obilježje *kvalitativno redosljedno*, za analizu raspršenja mogu se koristiti raspon varijacije, interkvartil i koeficijent kvartilne devijacije. Grubu sliku ("prvi dojam") o raspršenosti podataka u osnovnom skupu zorno može predočiti i odgovarajući grafikon, a riječ je o *dijagramu rasipanja* ili *dijagramu raspršenja* spomenutom u 2. poglavlju.

Navedene mjere raspršenja, dakle, shvaćamo kao svojevrsne mjere reprezentativnosti srednjih vrijednosti. Manja mjera raspršenja znači bolju reprezentativnost srednje vrijednosti i obrnuto[85]. Pomoću mjera raspršenja uspoređuju se razlike u varijablitetu barem dviju razdioba, pri čemu vrijede sljedeća praktična pravila:

Pravilo 1. *Ukoliko se radi o razdiobama prema istim obilježjima* (*s istom skalom modaliteta*), *za usporedbu varijabiliteta koriste se apsolutne mjere raspršenja*.

Pravilo 2. *Ako se radi o razdiobama različitih obilježja ili istih obilježja s različitim skalama modaliteta*, *za usporedbu varijabliteta koriste se relativne mjere raspršenja.*

[85] Grubo možemo reći da su te dvije veličine *obrnuto razmjerne*, premda to matematički nije točno.

3.2.1. *Raspon varijacije*

Raspon varijacije (oznaka: R) je najjednostavnija mjera raspršenja definirana kao razlika između najveće (M) i najmanje (m) vrijednosti *kvantitativnoga* (ili, eventualno, kvalitativnoga redosljednoga) obilježja. Dakle,

$$R = M - m.$$

Najmanja moguća vrijednost raspona varijacije jednaka je nuli i nastupa u slučajevima kad je odgovarajući niz vrijednosti kvantitativnoga obilježja *konstantan*. Najveća vrijednost raspona varijacije ne može se općenito odrediti.

Primjer 2. Navršene godine radnoga staža svih 10 zaposlenika tvrtke "*Super Silva* d.o.o." tvore niz: 20, 12, 7, 29, 32, 10, 18, 16, 8, 17. Najveća vrijednost u tome nizu je $M = 32$, a najmanja $m = 7$. Stoga je raspon varijacije vrijednosti promatranoga obilježja

$$R = M - m = 32 - 7 = 25 \text{ (godina)}.$$

Primjer 3. Prosječne mjesečne neto–plaće (iskazane u kn) u razdoblju od 01.01.2008. do 31.03.2008. svih 6 zaposlenika knjigovodstvenoga servisa "*Knjiži kod mene*" iz Košara na dan 01.04.2008. tvore niz 2.528,42; 2.527,75; 25.000,64; 2.584,12; 2.598,02; 2.576,25. Najveća vrijednost u tome nizu je $M = 25.000,64$, a najmanja $m = 2.527,75$. Stoga je raspon varijacije promatranoga obilježja

$$R = M - m = 25.000,64 - 2.527,75 = 22.472,89 \text{ (kn)}.$$

Primjer 4. Zadana je razdioba svih nezaposlenih državljana Republike Ljenčarije u veljači 2008. godine prema navršenim godinama starosti:

navršene godine starosti	*broj nezaposlenih* [000]
15 – 24	25,2
25 – 49	75,8
50 – 64	19,2
≥ 65	0,6

izvor: Mjesečno statističko izvješće Državnoga zavoda za statistiku Republike Ljenčarije, ožujak 2008.

Da bismo mogli odrediti raspon varijacije ovoga obilježja, za najmanju vrijednost (m) uzimamo *donju granicu prvoga razreda*, a za najveću (M) *gornju granicu posljednjega razreda*. Uočimo da je posljednji razred otvoren, pa njegovu gornju granicu moramo procijeniti sami. Uzmemo li da je procijenjena gornja granica posljednjega razreda 70, slijedi da je $M = 70$ i $m = 15$, pa je *procijenjeni* raspon varijacije promatranoga obilježja

$$R = M - m = 70 - 15 = 55 \text{ (godina)}.$$

Osnovna prednost raspona varijacije u praktičnoj primjeni je *jednostavnost* njegova izračuna, odnosno interpretacije. Njegov osnovni nedostatak je korištenje točno dvaju elemenata osnovnoga skupa (u slučaju negrupiranih podataka) ili razdiobe (u slučaju grupiranih podataka) u njegovu izračunu. Stoga ga, kao reprezentativnu mjeru raspršenja, ima smisla primijeniti u slučajevima kad promatrani niz podataka ne sadrži *ekstremno velike* ili *ekstremno male* vrijednosti jer je u suprotnim slučajevima vrlo nepouzdan (npr. u Primjeru 3. bi se moglo pogrešno zaključiti da promatrane prosječne mjesečne plaće jako variraju jer odgovarajući niz sadrži ekstremno veliku vrijednost 25.000,64). Još općenitije govoreći, raspon varijacije može se iskoristiti za "grubi" opis osnovnoga skupa ili statističkoga niza, ali nije dovoljno precizna mjera za daljnju detaljnu analizu tih objekata.

3.2.2. *Interkvartil*

Interkvartil (oznaka: I_q) je apsolutna mjera raspršenja definirana kao *raspon varijacije središnjih 50% elemenata uređenoga niza* (kvantitativnih ili kvalitativnih redosljednih) *statističkih podataka*. Ekvivalentno rečeno, interkvartil je, zapravo, jednak razlici trećega (gornjega) i prvoga (donjega) kvartila:

$$I_q = Q_3 - Q_1.$$

Primjer 5. Prvi (donji) kvartil niza navršenih godina radnoga staža iz Primjera 2. jednak je Q_1 = 10.5 (godina), a treći (gornji) kvartil istoga niza Q_3 = 19.5. Odatle slijedi da je interkvartil jednak

$$I_q = Q_3 - Q_1 = 19{,}5 - 10{,}5 = 9 \text{ (godina)}.$$

Dakle, raspon varijacije središnjih 50% navršenih godina radnoga staža iznosi 9 godina. Iz dobivenih vrijednosti slijedi da 25% promatranih zaposlenika ima manje od 10,5 godina radnoga staža, 25% promatranih zaposlenika ima više od 19,5 godina radnoga staža, a navršene godine radnoga staža preostalih 50% promatranih zaposlenika su u rasponu od 9 godina.

Primjer 6. Prvi (donji) kvartil niza prosječnih mjesečnih neto–plaća iz Primjera 3. jednak je Q_1 = 2.540,38 (kn), a treći (gornji) kvartil istoga niza Q_3 = 2.594,55 (kn). Odatle slijedi da je interkvartil jednak

$$I_q = Q_3 - Q_1 = 2.594{,}55 - 2.540{,}38 = 54{,}17 \text{ (kn)}.$$

Dakle, raspon varijacije središnjih 50% prosječnih mjesečnih neto–plaća iznosi 54,17 kn. Iz dobivenih vrijednosti slijedi da 25% promatranih zaposlenika ima prosječnu mjesečnu neto–plaću manju od 2.540,38 kn, 25% promatranih zaposlenika ima prosječnu mjesečnu neto–plaću veću od 2.594,55 kn, a preostalih 50% promatranih zaposlenika ima prosječnu mjesečnu neto–plaću u rasponu od 54,17 kn.

Osnovna prednost interkvartila u odnosu na raspon varijacije je zanemarivanje ekstremno malih ili ekstremno velikih vrijednosti u statističkom nizu, čime se dobiva točniji opis većine elemenata niza. No, kao i za raspon varijacije, za izračun interkvartila koriste se točno dvije položajne vrijednosti u statističkom nizu, pa u izračunu interkvartila ne sudjeluju svi elementi niza. Zbog toga se ta mjera raspršenja smatra nepotpunom.

3.2.3. *Interpercentilni razmak*

Interkvartil je poseban slučaj tzv. *interpercentila*. Za bilo koja dva prirodna broja k, $l \in \mathbf{N}$ takva da vrijedi nejednakost $0 < k \leq l < 100$ *interpercentil* je razlika l – toga i k – toga percentila:

$$I_{P_l - P_k} = P_l - P_k.$$

Prema definiciji percentila vrijedi ekvivalencija

$$(k \leq l) \Leftrightarrow (P_k \leq P_l),$$

pa je bilo koji interpercentil uvijek nenegativan realan broj. Obično se interpretira kao *raspon varijacije središnjih*[86] *$(l - k)$% elemenata uređenoga statističkoga niza*. Posebno, za $k = 25$ i $l = 75$ dobivamo interkvartil, a ako su oba broja k i l djeljiva s 10, dobivamo tzv. *interdecilni razmak* (razlika bilo kojih dvaju decila). Praktično se brojevi k i l obično biraju tako da vrijedi jednakost

$$k + l = 100$$

jer se na taj način "odvaja" jednak postotak "najlošijih" i "najboljih" elemenata niza.

Primjer 7. U tvrtki "*Kamen Stečaj* d.d." iz Brckovljana 10% svih zaposlenika ima prosječnu mjesečnu neto–plaću manju od 1.900,00 kn, dok drugih 10% svih zaposlenika ima prosječnu mjesečnu neto–plaću veću od 2.200,00 kn. Iz interpretacije navedenih vrijednosti slijedi:

$$P_{10} = 1.900{,}00 \text{ i } P_{90} = 2.200{,}00,$$

pa je odgovarajući interpercentilni razmak

[86] Izraz *središnjih* posljedica je činjenice da percentili P_k i P_l, zanemarujući ostalih 97 percentila, dijele uređeni statistički niz na 3, općenito nejednaka, dijela: prvi ("donji") dio tvori k% podataka manji ili jednak P_k, drugi ("središnji") dio tvore podaci smješteni između P_k i P_l, a treći ("gornji") dio tvori $(100 - l)$% podataka većih ili jednakih P_l.

$$I_{P_{90}-P_{10}} = P_{90} - P_{10} = 2.200{,}00 - 1.900{,}00 = 300{,}00 \text{ (kn)}.$$

Dakle, raspon varijacije središnjih 90% – 10% = 80% prosječnih mjesečnih neto–plaća iznosi 300,00 kn. Primijetimo da je taj raspon ujedno i razlika 9. i 1. decila.

Primjer 8. U zaštitarskoj tvrtki "*Orlović* d.o.o" iz Donje Vrbe 15% svih zaposlenika nije više od 185 cm, a 15% svih zaposlenika nije niže od 197 cm. Iz interpretacije navedenih percentila slijedi

$$P_{15} = 185 \text{ i } P_{85} = 197,$$

pa je odgovarajući interpercentilni razmak

$$I_{P_{85}-P_{15}} = P_{85} - P_{15} = 197 - 185 = 12 \text{ (cm)}.$$

Dakle, raspon varijacije središnjih 85% – 15% = 70% visina iznosi 12 cm.

Osnovna prednost interpercentila u odnosu na raspon varijacije je zanemarivanje obično jednakoga postotka ekstremno malih i ekstremno velikih vrijednosti u statističkom nizu, čime se dobiva točniji opis većine elemenata niza. No, kao i za raspon varijacije, za izračun interpercentila koriste se točno dvije položajne vrijednosti u statističkom nizu, pa u njegovu izračunu ne sudjeluju svi elementi niza. Zbog toga se ta mjera raspršenja smatra nepotpunom.

3.2.4. *Varijanca i standardna devijacija*

Varijanca i iz nje izvedena standardna devijacija obično se svrstava u red najvažnijih pokazatelja varijabiliteta ili raspršenosti modaliteta kvantitativnih obilježja. Osnovni razlog je činjenica da da je zbroj odstupanja svih vrijednosti kvantitativnoga obilježja od njihove aritmetičke sredine *uvijek* jednak nuli, pa taj pokazatelj ne možemo koristiti za opis varijabiliteta statističkoga niza. Drugi je razlog tome što u izračunu varijance, a samim tim i standardne devijacije, koristimo sve elemente statističkoga niza, pa je možemo kategorizirati kao *potpunu* mjerom raspršenja. Drugim riječima, varijanca i standardna devijacija mogu se računati i iz "sirovih" (negrupiranih ili neuređenih) i iz grupiranih podataka.

Varijanca (oznaka: σ^2 (grčko slovo σ čitajte: *sigma*)) se definira kao *aritmetička sredina kvadrata odstupanja vrijednosti kvantitativnoga obilježja od aritmetičke sredine svih vrijednosti*. Nerijetko se varijanca – matematički netočno, ali intuitivno prihvatljivije – definira kao *prosječno kvadratno odstupanje* vrijednosti kvantitativnoga obilježja od aritmetičke sredine tih vrijednosti. Sukladno takvoj definiciji, formule za izračun varijance su:

a) za niz negrupiranih podataka $x_1, x_2, \ldots, x_n$ čija je aritmetička sredina $\bar{x}$:

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n} \quad \text{ili, ekvivalentno,} \quad \sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2,$$

b) za podatke čija je aritmetička sredina $\overline{x}$ grupirane prema točno k različitih modaliteta tako da za svaki $i = 1, 2, \ldots, k$ modalitet x_i ima apsolutnu ili relativnu frekvenciju f_i :

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i \cdot \left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{k} f_i} \quad \text{ili, ekvivalentno,} \quad \sigma^2 = \frac{\sum_{i=1}^{k} f_i \cdot x_i^2}{\sum_{i=1}^{k} f_i} - \overline{x}^2$$

c) za podatke grupirane u točno k razreda takvih da je za svaki $i = 1, 2, \ldots, k$ x_i razredna sredina, f_i odgovarajuća apsolutna ili relativna frekvencija, a $\overline{x}$ aritmetička sredina svih podataka:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i \cdot \left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{k} f_i} \quad \text{ili, ekvivalentno,} \quad \sigma^2 = \frac{\sum_{i=1}^{k} f_i \cdot x_i^2}{\sum_{i=1}^{k} f_i} - \overline{x}^2.$$

Primijetimo da je varijanca jednaka nuli ako i samo ako je pripadni statistički niz *konstantan*, tj. ako postoji realan broj $c \in \mathbf{R}$ takav da za svaki $k = 1, 2, \ldots, n$ vrijedi $x_k = c$. U tom je slučaju i aritmetička sredina jednaka broju c. No, što promatrani niz sadrži više međusobno različitih članova, tj. što je veći stupanj varijabilnosti niza, to su veća i odstupanja pojedinih modaliteta kvantitativnoga obilježja od njihove aritmetičke sredine. Stoga možemo zaključiti da je varijanca *upravno razmjerna* stupnju raspršenja statističkoga niza.

Budući je varijanca "kvadratna" mjera raspršenja, relativno teško ju je interpretirati. Pomoću drugoga korijena iz varijance dolazimo do praktično najprimjenjenije potpune mjere raspršenja: *standardne devijacije***.** *Standardna devijacija* (oznaka: σ) je, po definiciji, *drugi korijen iz varijance*. Takva formalna definicija obično se zamjenjuje matematički netočnom, ali intuitivno prihvatljivijom definicijom prema kojoj je standardna devijacija *prosječno odstupanje vrijednosti kvantitativnoga obilježja od aritmetičke sredine tih vrijednosti* ili, još kraće i još nepreciznije, *prosječno odstupanje od prosjeka*. Istaknimo da je "mjerna jedinica" za standardnu devijaciju jednaka "mjernoj jedinici" odgovarajućega kvantitativnoga obilježja (kn, €, kg, cm itd.).

Radi potpunosti, navodimo formule za izračun standardne devijacije:

a) za negrupirane podatke:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}} \quad \text{ili} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x^2}}$$

b) za podatke grupirane prema modalitetima ili u razrede:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i \cdot \left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{k} f_i}} \quad \text{ili} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i \cdot x_i^2}{\sum_{i=1}^{k} f_i} - \overline{x^2}}$$

<u>Primjer 9.</u> Odredimo varijancu i standardnu devijaciju niza navršenih godina radnoga staža iz Primjera 2. Aritmetička sredina vrijednosti toga obilježja jednaka je:

$$\overline{x} = \frac{20+12+7+29+32+10+18+16+8+17}{10} = 16.9 \text{ (godina)},$$

pa je varijanca tih podataka:

$$\sigma^2 = \frac{\sum_{i=1}^{10} x_i^2}{10} - \overline{x}^2 =$$

$$\sigma^2 = \frac{20^2+12^2+7^2+29^2+32^2+10^2+18^2+16^2+8^2+17^2}{10} - 16.9^2$$

$$\sigma^2 = 63.49$$

Odavde slijedi da je standardna devijacija jednaka

$$\sigma = \sqrt{\sigma^2} = \sqrt{63.49} = 7.968 \approx 8 \text{ godina.}$$

Možemo, dakle, reći da je *prosječno odstupanje pojedinačnih navršenih godina radnoga staža od njihove aritmetičke sredine* približno 8 *godina*.

<u>Primjer 10.</u> Promatramo statistički skup svih stanova u onim stambenim zgradama (novogradnja i dogradnja) koje su 2006. godine dobile građevinsku dozvolu u Republici Hrvatskoj. Promatrani skup dijelimo prema broju soba u stanu:

broj soba (x_i)	*broj stanova* (f_i)
1	3 256
2	7 428
3	7 244
4	4 414
5	1 794
6	997
7	266
8	118
ukupno:	25 517

izvor: Statistički ljetopis Republike Hrvatske za 2006. godinu, Državni zavod za statistiku, 2007.

Odredimo prosječan broj soba po jednom stanu, te varijancu i standardnu devijaciju broja soba. Prosječan broj soba po jednom stanu je vagana aritmetička sredina promatrane razdiobe:

$$\bar{x} = \frac{\sum_{i=1}^{8} f_i x_i}{\sum_{i=1}^{8} f_i} = \frac{3256 \cdot 1 + 7428 \cdot 2 + 7244 \cdot 3 + 4414 \cdot 4 + 1794 \cdot 5 + 997 \cdot 6 + 266 \cdot 7 + 118 \cdot 8}{25517}$$

$$\bar{x} = 2.949$$

Dakle, *prosječan broj soba po jednom stanu iznosi* približno 3. Varijanca broja soba jednaka je:

$$\sigma^2 = \frac{\sum_{i=1}^{8} f_i \cdot x_i^2}{\sum_{i=1}^{8} f_i} - \bar{x}^2$$

$$\sigma^2 = \frac{3256 \cdot 1^2 + 7428 \cdot 2^2 + 7244 \cdot 3^2 + 4414 \cdot 4^2}{25517} +$$

$$+ \frac{1794 \cdot 5^2 + 997 \cdot 6^2 + 266 \cdot 7^2 + 118 \cdot 8^2}{25517} - 2.949328^2$$

$$\sigma^2 = 1.887192$$

otkuda slijedi da je odgovarajuća standardna devijacija broja soba

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.887192} = 1.373751.$$

Dakle, možemo reći da *prosječno odstupanje pojedinačnih brojeva soba od njihove aritmetičke sredine iznosi* približno 1,374.

Primjer 11. U donjoj je tablici prikazana razdioba svih počinitelja kaznenih djela u Republici Hrvatskoj osuđenih na bezuvjetni zatvor u 2006. godini prema trajanju zatvorske kazne:

trajanje kazne [godina]	*broj počinitelja*
0 – 1	2 535
1 – 2	658
2 – 3	232
3 – 5	164
5 – 10	102
10 – 15	29
15	11
20	3
20 – 40	8
ukupno:	3 742

izvor: Statistički ljetopis Republike Hrvatske za 2006. godinu, Državni zavod za statistiku, 2007.

Odredimo prosječno trajanje kazne po jednom počinitelju, te varijancu i standardnu devijaciju trajanja kazni. U tablici imamo ukupno 7 pravih razreda i dva modaliteta koja nisu grupirana u razrede. Za svaki pojedini razred računamo *razrednu sredinu*, a potom i vaganu aritmetičku sredinu:

trajanje kazne [godina]	*broj počinitelja* (f_i)	*razredna sredina* (x_i)	$f_i \cdot x_i$	$f_i \cdot x_i^2$
0 – 1	2 535	0.5	1267.5	633.75
1 – 2	658	1.5	987	1480.5
2 – 3	232	2.5	580	1.450
3 – 5	164	4	656	2.624
5 – 10	102	7.5	765	5737.5
10 – 15	29	12.5	362.5	4531.25
15	11	15	165	2.475
20	3	20	60	1.200
20 – 40	8	30	240	7.200
ukupno:	3 742		5083	27.332

Vagana aritmetička sredina jednaka je:

$$\bar{x} = \frac{\sum_{i=1}^{9} f_i x_i}{\sum_{i=1}^{9} f_i} = \frac{5083}{3742} = 1.358365,$$

pa *prosječno trajanje kazne po jednom počinitelju iznosi* približno 1.358365 godina, odnosno 1 godinu 4 mjeseca i 9 dana. Varijanca trajanja kazni iznosi:

$$\sigma^2 = \frac{\sum_{i=1}^{9} f_i \cdot x_i^2}{\sum_{i=1}^{9} f_i} - \bar{x}^2 = \frac{27332}{3742} - 1.358365^2 = 5.458961,$$

otkuda slijedi da je standardna devijacija trajanja kazni

$$\sigma = \sqrt{\sigma^2} = \sqrt{5.458961} = 2.336442 \text{ godina},$$

tj. *prosječno odstupanje pojedinačnih trajanja kazni od njihove aritmetičke sredine iznosi* približno 2.336442 godina, odnosno 2 godine 4 mjeseca i 1 dan. Primijetimo da u promatranom primjeru nešto više od 85% svih podataka pripada prvim dvama razredima, dok svi ostali razredi obuhvaćaju nešto manje od 15% svih podataka. Na temelju aritmetičke sredine i standardne devijacije ne možemo zaključiti npr. da postoji razred 20 – 40 kojemu pripada 8 elemenata osnovnoga skupa, pa statistička analiza promatranoga obilježja mora obuhvatiti i izračun dodatnih pokazatelja (npr. koeficijent varijacije o kojemu govorimo u podtočki 3.2.6.).

3.2.5. *Srednje apsolutno odstupanje*

Već je istaknuto da je zbroj svih odstupanja pojedinih vrijednosti kvantitativnoga obilježja od aritmetičke sredine jednak nuli. No, uzmemo li *apsolutnu vrijednost* svakoga odstupanja (izračunamo odstupanje i zanemarimo njegov predznak), dobit ćemo nenegativne realne brojeve. Njihov je zbroj biti jednak nuli ako i samo ako je niz vrijednosti kvantitativnoga obilježja *konstantan*, a to općenito ne vrijedi (kvantitativna obilježja imaju najmanje dva različita modaliteta). Stoga ima smisla definirati *srednje apsolutno odstupanje* (oznaka: $MAD_{\bar{x}}$, pri čemu kratica: *MAD* od engleskih riječi ***M**ean **A**bsolute **D**eviation*) kao *prosječno apsolutno odstupanje od aritmetičke sredine*. $MAD_{\bar{x}}$ se izračunava prema sljedećim formulama:

a) za niz negrupiranih modaliteta kvantitativnoga obilježja x_1, x_2, …, x_n čija je aritmetička sredina $\bar{x}$:

$$MAD_{\bar{x}} = \frac{\left|x_1 - \bar{x}\right| + \left|x_2 - \bar{x}\right| + \ldots + \left|x_n - \bar{x}\right|}{n} = \sum_{k=1}^{n} \frac{\left|x_k - \bar{x}\right|}{n};$$

b) za podatke čija je aritmetička sredina $\bar{x}$ grupirane prema točno k različitih modaliteta tako da za svaki $i = 1, 2, \ldots, k$ modalitet x_i ima apsolutnu ili relativnu frekvenciju f_i :

$$MAD_{\bar{x}} = \frac{f_1 \cdot \left|x_1 - \bar{x}\right| + f_2 \cdot \left|x_2 - \bar{x}\right| + \ldots + f_n \cdot \left|x_n - \bar{x}\right|}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{k=1}^{n} f_k \cdot \left|x_k - \bar{x}\right|}{\sum_{k=1}^{n} f_k};$$

c) za podatke grupirane u točno k razreda takvih da je za svaki i = 1, 2, …, k x_i razredna sredina, f_i odgovarajuća apsolutna ili relativna frekvencija, a $\bar{x}$ aritmetička sredina svih podataka:

$$MAD_{\bar{x}} = \frac{f_1 \cdot \left|x_1 - \bar{x}\right| + f_2 \cdot \left|x_2 - \bar{x}\right| + \ldots + f_n \cdot \left|x_n - \bar{x}\right|}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{k=1}^{n} f_k \cdot \left|x_k - \bar{x}\right|}{\sum_{k=1}^{n} f_k}.$$

<u>Primjer 12.</u> Odredimo srednje apsolutno odstupanje od aritmetičke sredine za niz navršenih godina radnoga staža iz Primjera 2. U Primjeru 9. izračunali smo da je aritmetička sredina toga niza $\bar{x}$ = 16.9 (godina), pa je srednje apsolutno odstupanje od aritmetičke sredine:

$$MAD_{\bar{x}} = \sum_{k=1}^{10} \frac{\left|x_k - \bar{x}\right|}{10} = \frac{|20-16.9| + |12-16.9| + |7-16.9| + |29-16.9| + |32-16.9| + |10-16.9|}{10} +$$
$$+ \frac{|18-16.9| + |16-16.9| + |8-16.9| + |17-16.9|}{10}$$

$$MAD_{\bar{x}} = \frac{|3.1| + |-4.9| + |-9.9| + |12.1| + |15.1| + |-6.9| + |1.1| + |-0.9| + |-8.9| + |0.1|}{10}$$

$$MAD_{\bar{x}} = \frac{3.1 + 4.9 + 9.9 + 12.1 + 15.1 + 6.9 + 1.1 + 0.9 + 8.9 + 0.1}{10} = 6.3$$

Dakle, *prosječno apsolutno odstupanje od aritmetičke sredine* iznosi 6.3 godina. Dobivena vrijednost je strogo manja od standardne devijacije istoga niza ($\sigma \approx 8$ godina).

<u>Primjer 13.</u> Odredimo srednje apsolutno odstupanje za vrijednosti obilježja iz Primjera 11:

trajanje kazne [godina]	*broj počinitelja*	*razredna sredina* (x_i)
0 – 1	2 535	0.5
1 – 2	658	1.5
2 – 3	232	2.5
3 – 5	164	4
5 – 10	102	7.5
10 – 15	29	12.5
15	11	15
20	3	20
20 – 40	8	30
ukupno:	3 742	

U tom smo primjeru izračunali vaganu aritmetičku sredinu:

$$\bar{x} = 1.358365 \text{ (godina)},$$

pa je traženo srednje apsolutno odstupanje

$$MAD_{\bar{x}} = \frac{\sum_{k=1}^{9} f_k \cdot \left|x_k - \bar{x}\right|}{\sum_{k=1}^{9} f_k}$$

$$MAD_{\bar{x}} = \frac{2535 \cdot |0.5 - 1.358365| + 658 \cdot |1.5 - 1.358365| + 232 \cdot |2.5 - 1.358365|}{3742} +$$

$$+ \frac{164 \cdot |4 - 1.358365| + 102 \cdot |7.5 - 1.358365| + 29 \cdot |12.5 - 1.358365| +}{3742}$$

$$+ \frac{11 \cdot |15 - 1.358365| + 3 \cdot |20 - 1.358365| + 8 \cdot |30 - 1.358365|}{3742}$$

$$MAD_{\bar{x}} = 1.163$$

Dakle, *prosječno apsolutno odstupanje od aritmetičke sredine iznosi* približno 1.63 godina.

Istaknimo da se prosječno apsolutno odstupanje može računati i u odnosu na medijan, pri čemu su odgovarajuće formule potpuno analogne navedenima (aritmetička sredina se zamijeni medijanom, a sve ostale veličine ostaju nepromijenjene). Može se pokazati i da je prosječno apsolutno odstupanje od aritmetičke sredine ili medijana uvijek manje ili jednako standardnoj devijaciji.

Zaključno istaknimo da dio novijih radova iz statistike preporučuje primjenu srednjih apsolutnih odstupanja (od aritmetičke sredine ili medijana) umjesto "tradicionalne" standardne

devijacije jer je praksa pokazala da pogreške u mjerenju, neidealizirani uvjeti provedbe *slučajnih pokusa* ili *opažanja* općenito itd. imaju manji utjecaj na prosječno apsolutno odstupanje, nego na standardnu devijaciju. Drugim riječima, sve se više zagovara da se u statističkoj analizi realnih procesa (koji se odvijaju u realnim, a ne idealnim uvjetima) primjenjuje prosječno apsolutno odstupanje, tim više što se utvrdilo da primjena takvoga odstupanja na različite vrste razdioba daje približno jednake ili čak bolje rezultate od standardne devijacije.

3.2.6. *Koeficijent varijacije*

Već smo istakli da je standardna devijacija potpuna mjera raspršenja iskazana u jedinicama promatranoga kvantitativnoga obilježja. Kao takva, standardna devijacija nije prikladna za usporedbu *varijabiliteta* najmanje dvaju različitih tipova kvantitativnih obilježja prema kojima se može podijeliti isti statistički skup. Promatramo li npr. skup svih natjecatelja u izboru za ovogodišnjeg Mistera Svrzigaća i podijelimo li sve njegove elemente prema obilježjima *masa* i *visina*, onda na temelju standardne devijacije ne možemo reći razlikuju li se promatrani natjecatelji više prema masi ili prema visini. U takvim se slučajevima *obavezno* primjenjuju *relativne mjere raspršenja* jer one ne ovise o "mjernim jedinicama" promatranih obilježja. Dvije najčešće primjenjivane mjere su *koeficijent varijacije* i *koeficijent kvartilne devijacije*, a koriste se ponajprije za usporedbu varijabiliteta dvaju razdioba.

Koeficijent varijacije (oznaka: V) definira se kao omjer standardne devijacije i aritmetičke sredine iskazan u postotcima. Slobodno govoreći, možemo reći da je koeficijent varijacije zapravo prosječno odstupanje vrijednosti kvantitativnoga obilježja od aritmetičke sredine iskazano u postotcima. Sukladno netom navedenoj definiciji, koeficijent varijacije neovisno o tome jesu li podaci grupirani ili nisu računamo prema formuli:

$$V = \frac{\sigma}{\bar{x}} \cdot 100 .$$

Istaknimo da se koeficijent varijacije primijenjuje gotovo isključivo u statističkoj analizi kvantitativnih obilježja vezanih za *ordinalnu* skalu (zbog dogovorno definirane nule nije primjenjiv u obilježjima vezanima za *intervalnu* skalu). Njegovi osnovni nedostaci su loša reprezentativnost u slučaju ekstremnih vrijednosti ili otvorenih razreda, te osjetljivost na male promjene aritmetičke sredine u slučajevima kad je aritmetička sredina vrlo blizu nuli.

Iako ne postoji opći kriterij procjene varijabiliteta prema vrijednostima koeficijenta varijacije, jedan od najčešće primjenjivanih je kriterij naveden u sljedećoj tablici. Naglasimo i činjenicu da je najmanja moguća vrijednost koeficijenta varijacije jednaka nuli (i postiže se isključivo za konstantne nizove podataka), dok najveća vrijednosti općenito nije određena. Praktične primjene pokazuju da koeficijent varijacije relativno rijetko prelazi 100%.

V [%]	*varijabilitet*
0 – 10	vrlo slab
10 - 30	relativno slab
30 - 50	umjeren
50 - 70	relativno jak
≥ 70	vrlo jak

Primjer 16. Za svih 11 nogometaša prve postave NK "*Akumulator*" iz Babine Grede na dan 01.04.2008. prikupljeni su podaci o masi i prosječnim mjesečnim neto–plaćama. Podaci su prikazani u donjoj tablici.

redni br. nogometaša	*masa* [kg]	*prosječna mjesečna neto–plaća* [€]
1	79	1.190,00
2	89	1.195,00
3	82	1.196,00
4	78	1.185,00
5	91	1.183,00
6	84	1.198,00
7	83	1.197,00
8	86	1.191,00
9	90	1.189,00
10	78	1.300,00
11	81	1.400,00

Utvrdimo razlikuju li se promatrani nogometaši više prema masi ili prema prosječnim mjesečnim neto-plaćama. U nizu masa i nizu prosječnih mjesečnih neto–plaća nema ekstremno velikih vrijednosti, niti vrijednosti blizu nule, pa koeficijent varijacije obaju obilježja računamo koristeći formule za jednostavnu aritmetičku sredinu i standardnu devijaciju za negrupirani niz podataka. Dobiva se:

obilježje	*aritmetička sredina*	*standardna devijacija*	*koeficijent varijacije*
masa	83.72727 (kg)	4.51444 (kg)	5.391839%
prosječna neto–plaća	1.220,36 (€)	64,88 (€)	5.316734%

Budući da je koeficijent varijacije masa strogo veći od koeficijenta varijacije prosječnih neto–plaća, zaključujemo da se promatrani nogometaši više razlikuju prema masama. Uočimo da su varijabiliteti obaju promatranih obilježja vrlo slabi, pa praktično možemo zaključiti da nema zamjetnih razlika između varijabiliteta promatranih obilježja.

Primjer 17. Utvrdimo varijabilitet navršenih godina starosti svih nezaposlenih osoba u Republici Ljenčariji u veljači 2008. godine (Primjer 4. u podtočki 3.1.1.):

navršene godine starosti	*broj nezaposlenih* [000]
15 – 25	25.2
25 – 50	75.8
50 – 65	19.2
65 – (70)	0.6

U navedenom smo primjeru izračunali:

$$\bar{x} \approx 37.177 \text{ godina,}$$

pa je standardna devijacija računata iz podataka grupiranih u razrede

$$\sigma = 11.48167 \text{ godina,}$$

što znači da je koeficijent varijacije promatranoga obilježja

$$V = \frac{\sigma}{\bar{x}} \cdot 100 = \frac{11.48167}{37.17715} \cdot 100 = 30.88\%.$$

U ovome je slučaju, dakle, varijabilitet "na granici" relativno slaboga i umjerenoga, a budući da je aritmetička sredina računata iz grupiranih podataka nepreciznija od aritmetičke sredine računate iz negrupiranih podataka i (obično) manja od nje, možemo zaključiti da je varijabilitet vrijednosti promatranoga obilježja umjeren.

3.2.7. *Koeficijent kvartilne devijacije*

Kako je već istaknuto, koeficijent varijacije je loš pokazatelj varijabiliteta vrijednosti promatranoga obilježja ukoliko promatrani niz vrijednosti obilježja sadrži ekstremno male ili ekstremno velike vrijednosti. U takvim se slučajevima raspršenost vrijednosti obilježja obično iskazuje interkvartilom (kao apsolutnom mjerom raspršenja), odnosno *koeficijentom kvartilne devijacije* kao pripadnom relativnom mjerom raspršenja.

Koeficijent kvartilne devijacije (oznaka: V_q) je, prema definiciji, omjer interkvartila i zbroja prvoga i trećega kvartila:

$$V_q = \frac{I_q}{Q_3 + Q_1} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

V_q se interpretira kao *intenzitet varijabiliteta središnje polovice* (središnjih 50%) *elemenata uređenoga statističkoga niza*, a može poprimiti točno sve vrijednosti iz segmenta [0, 1]. Što je varijablitet središnje polovice niza manji, vrijednost V_q je bliža 0, a što je varijablitet središnje polovice veći, vrijednost V_q je bliža 1. Iako ne postoji opći kritetij za utvrđivanje varijabiliteta

središnje polovice vrijednosti kvantitativnoga obilježja na temelju koeficijenta kvartilne devijacije, jedan od najčešće primijenjenih kriterija naveden je u tablici.

V_q	*varijabilitet*
0.0 - 0.1	vrlo slab
0.1 - 0.2	relativno slab
0.2 - 0.3	umjeren
0.3 - 0.5	relativno jak
0.5 – 1.0	vrlo jak

Izračun koeficijenta kvartilne devijacije iz obiju vrsta podataka pokazat ćemo na sljedećim primjerima.

Primjer 18. Odredimo varijabilitet središnjih 50% prosječnih mjesečnih plaća (iskazanima u kn) svih stalno zaposlenih djelatnika poduzeća "*Schwindl–trade d.o.o.*" iz Šuplje Lipe u razdoblju od 01.01.2007. – 31.12.2007. navedenih u Primjeru 9. U ovom slučaju statistički niz sadrži dvije ekstremno velike vrijednosti (50.000,64 i 60.000,75), pa je za dobivanje slike o varijabilitetu primjereno uporabiti koeficijent kvartilne devijacije.

U Primjeru 9. izračunali smo:

- prvi ili donji kvartil: Q_1 = 2.332,935;
- treći ili gornji kvartil: Q_3 = 3.087,88;

Stoga je koeficijent kvartilne devijacije promatranoga niza jednak

$$V_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{3.087,88 - 2.332,935}{3.087,88 + 2.332,935} = 0.13927 .$$

Dakle, *varijabilitet središnjih* 50% *prosječnih mjesečnih plaća je relativno slab*.

Primjer 19. Zadana je razdioba svih honorarnih nastavnika Visoke škole za financijski menadžment u Špičkovini u zimskom semestru akademske godine 2007/2008. prema prosječnom mjesečnom neto–honoraru za sve održane oblike nastave:

prosječan mjesečni neto–honorar (€)	*broj nastavnika*
100,00	5
200,00	8
400,00	21
500,00	12
600,00	6
ukupno:	52

Odredimo koeficijent kvartilne devijacije i objasnimo njegovo značenje[87]. U ovom slučaju prvi i treći kvartil određujemo iz grupiranih podataka, pri čemu smo podatke već zapisali u obliku uređenoga niza. Ukupan broj podataka jednak je n = 52 i djeljiv je s 4, pa je približna pozicija prvoga kvartila jednaka

$$k_1 = \frac{n+1}{4} = \frac{52+1}{4} = 13.25\,.$$

Vrijednost prvoga kvartila iz grupiranih podataka određuje se iz izraza:

$$Q_1 = 13.\text{ vrijednost u nizu} + \frac{1}{4}\cdot(14.\text{ vrijednost u nizu} - 13.\text{ vrijednost u nizu}) =$$

$$= 200{,}00 + \frac{1}{4}\cdot(400{,}00 - 200{,}00) = 250{,}00.$$

Analogno, približna pozicija trećega kvartila jednaka je

$$k_2 = \frac{3}{4}\cdot(n+1) = \frac{3}{4}\cdot(52+1) = 39.75\,,$$

pa je vrijednost trećega kvartila:

$$Q_3 = 39.\text{ vrijednost u nizu} + \frac{3}{4}\cdot(40.\text{ vrijednost u nizu} - 39.\text{ vrijednost u nizu}) =$$

$$= 500{,}00 + \frac{3}{4}\cdot(500{,}00 - 500{,}00) = 500{,}00.$$

Tako je koeficijent kvartilne devijacije jednak

$$V_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{500{,}00 - 250{,}00}{500{,}00 + 250{,}00} = 0.33333$$

i zaključujemo da je *varijabilitet središnjih* 50% *prosječnih mjesečnih neto–honorara relativno jak*.

<u>Primjer 20.</u> Zadana je razdioba svih djelatnika Sveleučilišta u Gaćelezima na dan 22.3.2008. prema iznosu primljene "uskrsnice":

---

[87] U ovome se primjeru isključivo radi ilustracije računanja koeficijenta kvartilne devijacije iz podataka grupiranih prema modalitetima koriste relacije za izračun kvartila koje se mogu naći u knjizi I. Šošića "*Primijenjena statistika*".

iznos "uskrsnice" (€)	*broj djelatnika*
351,00 – 500,00	48
501,00 – 650,00	189
651,00 – 800,00	88
801,00 – 950,00	47
951,00 – 1.100,00	28
ukupno:	400

Odredimo koeficijent kvartilne devijacije i objasnimo dobiveni rezultat.[88] Neprave razrede najprije pretvorimo u prave i izračunajmo pripadne kumulativne apsolutne frekvencije "manje od". Dobivamo:

iznos "uskrsnice" (€)	*broj djelatnika*	*kumulativna apsolutna frekvencija "manje od"*
350,50 – 500,50	48	48
500,50 – 650,50	189	237
650,50 – 800,50	88	325
800,50 – 950,50	47	372
950,50 – 1.100,50	28	400
ukupno:	400	

Četvrtina ukupnoga broja djelatnika jednaka je 100, a 100. element uređenoga niza nalazi se u razredu 500,50 – 650,50. Donja granica toga razreda je g_1 = 500,50, odgovarajuća apsolutna frekvencija f_1 = 189, kumulativna apsolutna frekvencija "manje od" *prethodnoga* razreda jednaka je m_1 = 48, a širina uočenoga razreda je h_1 = 650,50 – 500,50 = 150. Stoga je prvi kvartil jednak

$$Q_1 = g_1 + \frac{\frac{n}{4} - m_1}{f_1} \cdot h_1 = 500{,}50 + \frac{\frac{400}{4} - 48}{189} \cdot 150{,}00 \approx 541{,}77.$$

Nadalje, tri četvrtine ukupnoga broja djelatnika jednake su 300, a 300. element niza nalazi se u rauzredu 650,50 – 800,50. Donja granica toga razreda je g_3 = 650,50, odgovarajuća apsolutna frekvencija f_3 = 88, kumulativna apsolutna frekvencija "manje od" *prethodnoga* razreda jednaka je m_2 = 237, a širina uočenoga razreda je h_3 = 850,50 – 650,50 = 15,00. Stoga je treći kvartil jednak

$$Q_3 = g_3 + \frac{\frac{3}{4} \cdot n - m_2}{f_3} \cdot h_3 = 650{,}50 + \frac{\frac{3}{4} \cdot 400 - 237}{88} \cdot 150{,}00 \approx 757{,}89.$$

[88] U ovome se primjeru isključivo radi ilustracije računanja koeficijenta kvartilne devijacije iz podataka grupiranih u razrede koriste relacije za izračun kvartila koje se mogu naći u knjizi I. Šošića "*Primijenjena statistika*".

Stoga je koeficijent kvartilne devijacije jednak

$$V_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{757,89 - 541,77}{757,89 + 541,77} = 0.1663,$$

pa zaključujemo da je *varijabilitet središnjih* 50% *iznosa* "*uskrsnica*" *relativno slab*.

3.3. *Standardizirano obilježje*

U praksi se vrlo često postavljaju problemi *usporedbe* različitih nizova vrijednosti kvantitativnih obilježja iskazanima u istim mjernicm jedinicama, ali s različitim stupnjem varijabilnosti, problemi usporedbe raznorodnih numeričkih nizova, problemi usporedbe relativnoga položaja barem dvaju podataka u istom numeričkom nizu itd. Tipični primjeri takvih problema su:

***Problem* 1.** Prosječna godišnja neto–plaća zaposlenika s visokom stručnom spremom u nekoj tvrtki iznosi 62.000,00 kn uz standardnu devijaciju od 7.100,00 kn. Može li se reći da je djelatnik koji ima visoku stručnu spremu i ukupnu godišnju neto–plaću 40.000,00 kn diskriminiran?

***Problem* 2.** Prosječan broj bodova na pismenom ispitu iz *Gospodarske matematike* održanom 25.02.2008. iznosi 55 uz standardnu devijaciju od 8 bodova. Prosječan broj bodova na pismenom ispitu iz *Osnova računovodstva* održanom 21.02.2008. iznosi 65 uz standardnu devijaciju 12 bodova. Ćiro je na pismenom ispitu iz *Gospodarske matematike* ostvario 62 boda, a na pismenom ispitu iz *Osnova računovodstva* 75 boda. Na kojemu je ispitu Ćiro ostvario bolji uspjeh?

Ovakvi problemi rješavaju se izračunom vrijednosti tzv. *standardiziranoga obilježja* (oznaka: *z*). Riječ je o linearnoj transformaciji vrijednosti kvantitativnoga obilježja definiranoj formulom:

$$z = \frac{1}{\sigma} \cdot x - \frac{\bar{x}}{\sigma} = \frac{x - \bar{x}}{\sigma},$$

gdje je uobičajeno $\bar{x}$ aritmetička sredina, a σ standardna devijacija vrijednosti varijable x. Može se pokazati da je aritmetička sredina *svakoga* standardiziranoga obilježja jednaka nuli, a odgovarajuća standardna devijacija jednaka 1. Kako bi se mogli donijeti zaključci o raspršenosti elemenata polaznoga osnovnoga skupa, koristi se sljedeće

<u>Pravilo 1.</u> (*Čebiševljevo*[89] *pravilo*) Neka je X niz vrijednosti kvantitativnoga obilježja čija je aritmetička sredina $\bar{x}$, a standardna devijacija σ. Tada se za svaki prirodan broj $k > 1$ u segmentu $[\bar{x} - k \cdot \sigma, \bar{x} + k \cdot \sigma]$ nalazi najmanje $100 \cdot \left(1 - \frac{1}{k^2}\right)$ % članova niza X.

<u>Posljedice</u>: U segmentu $[\bar{x} - 2 \cdot \sigma, \bar{x} + 2 \cdot \sigma]$ nalazi se najmanje 75% podataka, a u segmentu $[\bar{x} - 3 \cdot \sigma, \bar{x} + 3 \cdot \sigma]$ nalazi se najmanje 89% podataka bilo kojega niza numeričkih podataka.

Pokažimo na nekoliko primjera primjenu standardiziranoga obilježja i Čebiševljeva pravila.

<u>Primjer 1.</u> Riješimo Problem 1. Izračunajmo vrijednost standardizirana obilježja za $x = 40.000{,}00$, $\bar{x} = 62.000{,}00$ i $\sigma = 7.100{,}00$:

$$z = \frac{x - \bar{x}}{\sigma} = \frac{40.000{,}00 - 62.000{,}00}{7.100{,}00} = -3.0986,$$

pa zaključujemo da ukupna godišnja plaća od 40.000,00 kn odstupa od prosječne godišnje neto–plaće naniže za više od tri standardne devijacije. Prema Čebiševljevu pravilu, najmanje 89% elemenata niza odstupa od prosječne vrijednosti niza za najviše tri standardne devijacije. To znači da najviše 100% – 89% = 11% elemenata niza odstupa od prosjeka za više od tri standardne devijacije. Promatrani se djelatnik nalazi upravo u tih 11% elemenata, pa se može reći da je diskriminiran.

<u>Primjer 2.</u> Riješimo Problem 2. Prosudba uspjeha na ispitima mora se temeljiti isključivo na standardiziranom obilježju jer je ono neovisno o mjernim jedinicama, a eliminira i problem različitih raspona varijacije.
Za prvi je pismeni ispit $\bar{x}_1 = 55$ i $\sigma_1 = 8$, pa je vrijednost standardizirana obilježja Ćirina uspjeha na tom ispitu

$$z_1 = \frac{x_1 - \bar{x}_1}{\sigma_1} = \frac{62 - 55}{8} = 0.875.$$

Za drugi je pismeni ispit $\bar{x}_2 = 65$ i $\sigma_2 = 12$, pa je vrijednost standardizirana obilježja Ćirina uspjeha na tom ispitu

$$z_2 = \frac{x_2 - \bar{x}_2}{\sigma_2} = \frac{75 - 65}{12} = 0.83333.$$

[89] Pafnutij Lavovič Čebišev (1821. – 1894.), ruski matematičar.

Budući da je vrijednost standardizirana obilježja za prvi pismeni ispit strogo veća od vrijednosti standardizirana obilježja za drugi pismeni ispit, zaključujemo da je na pismenom ispitu iz *Gospodarske matematike* Ćiro postigao nešto bolji uspjeh (odstupanje naviše iznosi 87,5% standardne devijacije).

Primjer 3. Tečaj dionica (u kn) portfolia zabilježen na Frkljevačkoj burzi u razdoblju od 01.03.2008. do 25.03.2008. bio je sljedeći:

126	128	133	128	123	130	122	125	129	133	127	126	130
129	133	123	122	121	126	134	128	126	128	135	135	

Dva trgovca dionicama procjenjuju tečaj dionice u vremenu predviđenom za prodaju. Prvi trgovac procjenjuje da će cijena biti 140 kn, a drugi da će cijena biti 139 kn. Pretpostavimo da je kretanje tečaja dionice "normalno". Odredimo koji se trgovac izlaže većem riziku donošenja pogrešne poslovne odluke o prodaji dionice. Najprije treba izračunati aritmetičku sredinu (prosječni tečaj dionice) i standardnu devijaciju iz negrupiranih podataka. Dobiva se:

$$\bar{x} = 128{,}00 \text{ kn}, \ \sigma = 4{,}08 \text{ kn}.$$

Izračunajmo vrijednost standardiziranoga obilježja za svaku pojedinu procjenu i odrediti koja od njih više odstupa od prosjeka:

$$z_1 = \frac{x_1 - \bar{x}}{\sigma} = \frac{140 - 128}{4{,}08} = 2.94\,,$$

$$z_2 = \frac{x_1 - \bar{x}}{\sigma} = \frac{139 - 128}{4{,}08} = 2.697\,.$$

Prema tome, prvi trgovac se izlaže većem riziku donošenja pogrešne poslovne odluke o prodaji dionice.

Primjer 4. 20 stručnjaka iz područja managementa i gospodarstva procjenjivalo je godišnju stopu inflacije u Republici Niškoristiji za 2007. godinu:

3,52	3,73	4,05	4,12	4,25	3,85	3,68	3,74	4,19	4,42	4,28	4,36	4,09
3,92	3,67	4,26	4,35	3,59	3,97	4,32						

Primjenom Čebiševljeva pravila za $k = 2$ odredimo prognostički interval stope inflacije. Jednostavna aritmetička sredina procjenjenih vrijednosti godišnje stope inflacije iznosi 4.018, a odgovarajuća standardna devijacija 0.280546. Prema tome, traženi prognostički interval stope inflacije je $[4.018 - 2 \cdot 0.280546, 4.018 + 2 \cdot 0.280546] = [3.457, 4.579]$.

Primjer 5. (*problem skaliranja*) Prijamnom ispitu za upis na Ekonomski fakultet Sveučilišta u Banovoj Jaruzi pristupilo je ukupno 2000 kandidata. Prosječan broj bodova iz matematike

iznosio je 150 sa standardnom devijacijom 30 bodova. Povjerenstvo za provedbu razredbenoga ispita odlučilo je skalirati rezultate tako da im aritmetička sredina bude 100, a standardna devijacija 12. Odredimo broj bodova (na novoj skali) kandidata koji je na ispitu osvojio točno 130 bodova.

Da bismo preračunali rezultate ispita u skladu s izmijenjenim parametrima, najprije ćemo odrediti relativan položaj postignutoga rezultata od 130 bodova u prvotnom numeričkom nizu. Taj se položaj svodi na izračun vrijednosti standardiziranoga obilježja za $x = 130$, $\overline{x}_1 = 150$ i $\sigma_1 = 30$, pa se dobije:

$$z_1 = \frac{x_1 - \overline{x}_1}{\sigma_1} = \frac{130 - 150}{30} = -0.66666\,.$$

Dakle, rezultat kandidata odstupa od prvotnoga prosjeka za 0.66666 prvotnih standardnih devijacija naniže. Stoga i novi broj bodova mora odstupati od predefiniranoga prosjeka za 0.66666 predefiniranih standardnih devijacija naniže, pa je

$$x_2 = \overline{x}_2 + z_1 \cdot \sigma_2 = 100 + (-0.66666) \cdot 12 = 92.$$

Prema tome, kandidat će na novoj skali imati ukupno 92 boda.

3.4. *Vrste razdioba* (*distribucija*)

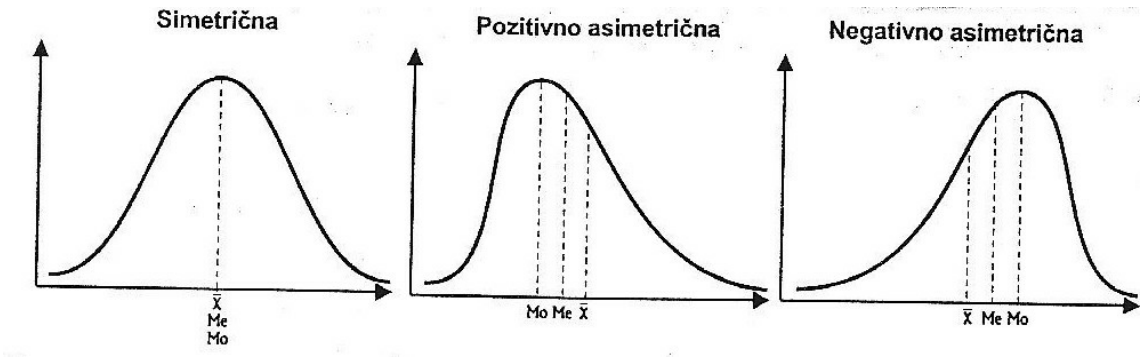
Razdiobe statističkih skupova prema *kvantitativnim* obilježjima grupiramo prema tome kako su vrijednosti obilježja elemenata toga skupa raspoređene oko najvažnije srednje vrijednosti: aritmetičke sredine. U skladu s tim razlikujemo tri osnovne vrste razdioba statističkoga skupa:

Ako su vrijednosti obilježja elemenata statističkoga skupa ravnomjerno raspoređene oko aritmetičke sredine, govorimo o <u>*simetričnoj razdiobi* (*distribuciji*)</u>**.** Kod simetrične razdiobe sve tri srednje vrijednosti su jednake ($\overline{X} = M_e = M_o$)**.**

Ako skup vrijednosti obilježja elemenata statističkoga skupa sadrži barem jednu *ekstremno veliku* vrijednost, govorimo o <u>*pozitivno asimetričnoj razdiobi* (*distribuciji*)</u>. Kod pozitivno asimetrične razdiobe vrijede nejednakosti: $\overline{X} > M_e > M_o$.

Ako skup vrijednosti obilježja elemenata statističkoga skupa sadrži barem jednu *ekstremno malu* vrijednost, govorimo o <u>*negativno asimetričnoj razdiobi* (*distribuciji*)</u>. Kod negativno asimetrične razdiobe vrijede nejednakosti: $\overline{X} < M_e < M_o$.

Gornje definicije zgodno je popratiti i grafički:



3.4.1. *Mjere asimetrije*

Pri opisu razdiobe vrijednosti obilježja elemenata određenoga statističkoga skupa, osim srednjih vrijednosti i mjera raspršenja, koristimo i *mjere asimetrije*. Tim mjerama nastoji se jednim numeričkim pokazateljem opisati *način rasporeda* kvantitativnih *podataka* prema aritmetičkoj sredini ili nekoj drugoj srednjoj vrijednosti. U praksi se najčešće koriste tri mjere asimetrije:

1.) <u>*Pearsonova*[90] *mjera asimetrije*</u> (oznaka: S_k) je mjera asimetrije definirana kao *omjer aritmetičke sredine* $\bar{x}$ *i medijana* M_e, odnosno *moda* M_o, *izražen u jedinicama standardne devijacije* σ:

$$S_k = \frac{3 \cdot (\bar{x} - M_e)}{\sigma}; \quad S_k = \frac{\bar{x} - M_o}{\sigma}$$

Ova je mjera zapravo standardizirano odstupanje medijana ili moda od aritmetičke sredine. Razlika aritmetičke sredine i medijana, odnosno moda podijeljena je sa standardnom devijacijom radi uklanjanja utjecaja mjernih jedinica kvantitativnoga obilježja na vrijednost pokazatelja.

Koeficijent S_k[91] može poprimiti točno sve vrijednosti iz segmenta [–3, 3]. Za *simetričnu* razdiobu njegova je vrijednost $S_k = 0$, za *pozitivno asimetričnu* $S_k > 0$, a za *negativno asimetričnu* $S_k < 0$. Što je apsolutna vrijednost od S_k bliža 3, asimetričnost je izraženija. Ovaj koeficijent u pravilu se računa za razdiobe prema kvantitativnim kontinuiranim varijablama.

2.) <u>*Bowleyeva*[92] *mjera asimetrije*</u> (oznaka: S_{kQ}) računa se iz međusobnoga odnosa *svih triju kvartila*: prvoga (Q_1), trećega (Q_3) i medijana (M_e). Na temelju toga odnosa određuje se

---

[90] Karl Pearson (1857. – 1936.), britanski statističar, jedan od najzaslužnijih za razvoj statistike kao zasebne znanstvene discipline.
[91] Oznaka S_k potječe od engleske riječi *skewness* koja znači *nagnutost*.
[92] Arthur Lyon Bowley (1869. – 1957.), engleski ekonomist i statističar, zaslužan za razvoj teorije uzoraka.

asimetričnost središnje polovice (središnjih 50%) *elemenata uređenoga statističkoga niza*. Računamo je prema formuli:

$$S_{kQ} = \frac{Q_1 + Q_3 - 2 \cdot M_e}{Q_3 - Q_1}.$$

Koeficijent S_{kq} poprima točno sve vrijednosti iz segmenta [-1, 1]. Za *simetričnu razdiobu* je $S_{kQ} = 0$, za *pozitivno asimetričnu* $S_{kQ} > 0$, a *za negativno asimetričnu* $S_{kQ} < 0$. Što je apsolutna vrijednost koeficijenta S_{kQ} bliža 1, asimetričnost je izraženija.

3.) <u>*Koeficijent asimetrije*</u> (oznaka: α_3) je *potpuna* i *najčešće korištena* mjera asimetrije definirana formulom:

$$\alpha_3 = \frac{\sum_{k=1}^{n} \frac{(x - \bar{x})^3}{n}}{\sigma}$$

gdje je n ukupan broj elemenata statističkoga niza (odnosno podataka u odgovarajućem osnovnom skupu), $\bar{x}$ aritmetička sredina tih podataka, a σ njihova standardna devijacija.

Koeficijent α_3 poprima točno sve vrijednosti iz segmenta [-2, 2]. Za *simetričnu razdiobu* vrijednost toga koeficijenta je $\alpha_3 = 0$, za pozitivno asimetričnu $\alpha_3 > 0$, a za negativno asimetričnu $\alpha_3 < 0$. Što je apsolutna vrijednost koeficijenta α_3 bliža 2, asimetričnost je izraženija.

Valjda istaknuti da koeficijent asimetrije – kao potpuna mjera asimetrije – uvijek pruža najvjerniju sliku o asimetriji, dok s ostalim dvama mjerama to nije slučaj. Sve tri mjere obično se (ali ne i uvijek) podudaraju jedino u predznaku, a njihova izravna usporedba *nije moguća* jer se odgovarajući izračuni temelje na različitim principima.

<u>Primjer 1.</u> Radi provjere deklarirane mase paketića bombona "*Kikić*" izabran je *uzorak* od ukupno 25 različitih paketića. Mjerenjem mase (iskazane u gramima) dobiveni su sljedeći rezultati:

101	103	98	99	100	102	96	102	101	105	93	98	98
94	96	101	102	97	97	100	93	104	102	91	104	

Odredimo vrijednost svih triju pokazatelja asimetrije, objasnimo njihova značenja i prikažimo dobivenu razdiobu linijskim grafikonom. Prikažemo li dobivene podatke tablično, dobivamo:

masa (g)	*broj paketića*
91	1
93	2
94	1
96	2
97	2
98	3
99	1
100	2
101	3
102	4
103	1
104	2
105	1
ukupno:	25

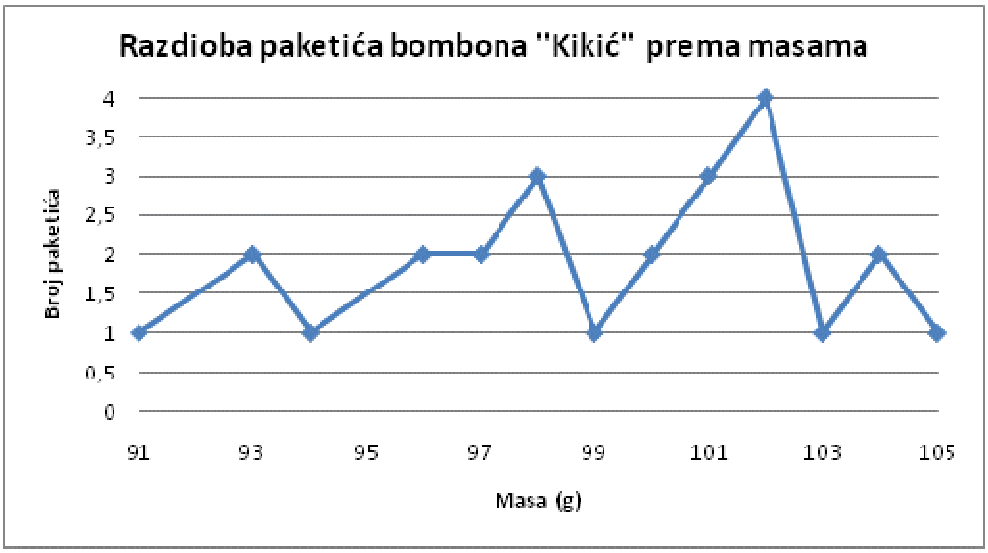
Odgovarajućim izračunima lagano se dobiva:

- aritmetička sredina: $\bar{x}$ = 99.08 (grama);
- medijan: M_e = 100 (grama);
- mod: M_o = 102 (grama);
- 1. ili donji kvartil: Q_1 = 97 (grama);
- 3. ili gornji kvartil: Q_3 = 102 (grama);
- standardna devijacija: σ = 3.69778312 (grama).

Za vježbu objasnite svaku od neposredno izračunatih vrijednosti. Prema tome, tražene vrijednosti mjera asimetrije su:

- Pearsonova mjera asimetrije (odstupanje medijana): S_k = -0.746393153;
- Pearsonova mjera asimetrije (odstupanje moda): S_k = -0.789662321;
- Bowleyeva mjera asimetrije: $S_{kQ} = -0.2$;
- koeficijent asimetrije: α_3 = -0.471860569.

Pearsonove mjere asimetrije upućuju na zaključak da je *promatrana razdioba srednje negativno asimetrična* (prevladavaju vrijednosti veće od aritmetičke sredine), Bowleyeva mjera asimetrije upućuje na zaključak da je *razdioba središnjih* 50% *paketića blago negativno asimetrična*, dok koeficijent asimetrije upućuje na zaključak da je *promatrana razdioba srednje negativno asimetrična*. To se vidi i iz dobivene tablice jer postoji ukupno 13 vrijednosti strogo većih od aritmetičke sredine i 12 vrijednosti manjih od nje. Odgovarajući prikaz linijskim grafikonom dan je na sljedećoj slici.



<u>Primjer 2.</u> Prema podacima Hrvatske elektroprivrede za 2006. godinu[93], prosječna godišnja potrošnja električne energije po potrošaču iznosi 2422.168207 kWh uz standardnu devijaciju 2383.869183 kWh. Pritom 25% svih potrošača godišnje potroši najviše 683.319 kWh, polovica svih potrošača najviše 1766.895 kWh, a 25% svih potrošača barem 3370.143 kWh. Odredimo vrijednosti odgovarajućih relativnih mjera raspršenja i mjera asimetrije, te objasnimo dobivene rezultate. Iz zadanih podataka slijedi:

- aritmetička sredina: $\bar{x}$ = 2422.168207 (kWh);
- medijan: M_e = 1766.895 (kWh);
- 1. ili donji kvartil: Q_1 = 683.319 (kWh);
- 3. ili gornji kvartil: Q_3 = 3370.143 (kWh);
- standardna devijacija: σ = 2383.869183 (kWh);

pa su tražene vrijednosti:

- koeficijent varijacije: V = 98.4188124%, što znači da je *stupanj raspršenja* (*varijabilitet*) *godišnje potrošnje električne energije vrlo jak*;
- koeficijent kvartilne devijacije: V_q = 66.2846722%, što znači da je *stupanj raspršenja* (*varijablitet*) *središnjih* 50% *potrošača električne energije vrlo jak*;
- Pearsonova mjera asimetrije (odstupanje medijana): S_k = 0.824634017, što znači da je *promatrana razdioba potrošača prema godišnjoj potrošnji električne energije srednje pozitivno asimetrična*;
- Bowleyeva mjera asimetrije: S_{kQ} = 0.193414976, što znači da je *razdioba središnjih* 50% *potrošača blago pozitivno asimetrična*;

[93] Podaci su preuzeti iz Statističkoga ljetopisa Republike Hrvatske za 2006. godinu.

Koeficijent asimetrije *ne* možemo izračunati na temelju zadanih podataka. Spomenimo tek da iz originalnih podataka navedenih u referenciranoj literaturi slijedi $\alpha_3 = 2.133$, što znači da je *razdioba potrošača prema godišnjoj potrošnji električne energije vrlo pozitivno asimetrična*, tj. vrlo izraženo prevladavaju manje vrijednosti potrošnje električne energije.

3.4.2. Mjere zaobljenosti

Većina pojava u prirodi, raznim gospodarskim i tehničkim procesima i sl. raspoređena je u skladu s tzv. *normalnom razdiobom*. Ova se razdioba vrlo precizno definira u matematičkoj statistici i za to je potreban nešto složeniji aparat teorije vjerojatnosti. Ovdje ćemo samo napomenuti da je riječ o potpuno simetričnoj unimodalnoj razdiobi takvoj da se otprilike svi podaci (točnije, 99.73% njih) promatranoga obilježja nalaze u rasponu $[\bar{x} - 3 \cdot \sigma, \bar{x} + 3 \cdot \sigma]$, gdje je $\bar{x}$ aritmetička sredina, a σ standardna devijacija razdiobe. Ova je činjenica nerijetko vrlo korisna prigodom procjenjivanja je li neka razdioba (približno) normalna ili nije.

S obzirom na veliki značaj normalne razdiobe u statistici, definirane su posebne mjere kojima se pojedine razdiobe uspoređuju s normalnom. To su tzv. *mjere zaobljenosti* i ima ih točno dvije: *koeficijent zaobljenosti* (*kurtosis*[94]) i *eksces*[95]. *Koeficijent zaobljenosti* (*kurtosis*) (oznaka: α_4) je mjera zaobljenosti modalnoga vrha neke razdiobe, dok je *eksces* (oznaka: κ) jednak razlici koeficijenta zaobljenosti i broja 3:

$$\kappa = \alpha_4 - 3.$$

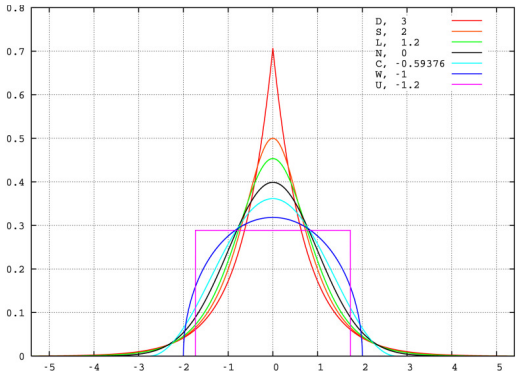
Od navedene dvije mjere jednostavnija je *eksces* jer eksces normalne razdiobe iznosi 0, pa je na temelju te činjenice lako izvršiti usporedbu. Prema predznaku ekscesa razlikujemo tri vrste razdioba:

- *platikurtične* čiji je eksces strogo manji od nule, tj. $\kappa < 0$;
- *mezokurtične* čiji je eksces jednak nuli, tj. $\kappa = 0$;
- *leptikurtične* čiji je eksces strogo veći od nule, tj. $\kappa > 0$.

Slobodnije govoreći, možemo reći da su platikurtične razdiobe "plosnatije", a leptikurtične "šiljastije" u odnosu na normalnu razdiobu. Usporedba nekih značajnijih vrsta razdioba s normalnom razdiobom (krivulja nacrtana crnom bojom) navedena je na donjoj slici.

---

[94] Riječ *kurtosis* na grčkom znači *izbočenost*.
[95] Riječ *eksces* na latinskom znači *prekoračenje* ili *pretjeranost*.



3.5. Programska potpora deskriptivnoj statističkoj analizi

Danas je u uporabi više različitih računalnih programa namijenjenih statističkim analizama. Najpoznatiji od njih su *SAS*, *SPSS* i *STATISTICA*. U *MS EXCEL*-u potpunu statističku analizu vrijednosti *kvantitativnoga* obilježja elemenata statističkoga skupa možemo dobiti rabeći alat *Descriptive Statistics* smještenu u izborniku *Alati* (*Tools*), opcija: *Analiza podataka* (*Data Analysis*). Taj je alat dio *Analysis ToolPak* alata za analizu, a stvara statistički izvještaj za podatke u ulaznom rasponu i daje podatke o tendenciji gomilanja podataka oko aritmetičke sredine i njihovoj promjenljivosti (varijabilitetu). Jedini potrebni ulazni podatak jest raspon ćelija u kojima su smješteni *negrupirani* numerički podaci. Tzv. *izlazni izvještaj* (*Summary Statistics*) obavljene statističke analize sadrži, između ostalih, i podatke o aritmetičkoj sredini, medijanu, modu, standardnoj devijaciji, varijanci, koeficijentu asimetrije, najmanjoj i najvećoj vrijednosti, opsegu i ukupnom zbroju svih članova promatranoga numeričkoga niza. Primjer takvog izvještaja prikazan je u donjoj tablici.

Ocjene iz matematike	
Mean	3.46666667
Standard Error	0.21298823
Median	3
Mode	2
Standard Deviation	1.16658456
Sample Variance	1.36091954
Kurtosis	-1.4568278
Skewness	0.08605128
Range	3
Minimum	2
Maximum	5
Sum	104
Count	30

4. KORELACIJSKA I REGRESIJSKA ANALIZA

U prethodnim smo se poglavljima bavili tzv. *jednodimenzionalnom analizom* određenih pojava. Pri tome nismo vodili računa o uzrocima i posljedicama tih pojava, odnosno o međusobnoj povezanosti dviju ili više pojava. U ovom ćemo poglavlju obratiti pozornost upravo na takve slučajeve. Pri tome ćemo (radi analogije s matematikom) termin *obilježja* (koji se obično rabi u kvalitativnim slučajevima) zamijeniti ekvivalentnim nazivom *varijabla* (jer ćemo promatrati kvantitativna obilježja).

Veza između dviju varijabli općenito može biti *funkcijska* (*deterministička*) ili *statistička* (*slučajna*, *stohastička*). U slučaju *funkcijske* povezanosti *svakoj vrijednosti* jedne varijable odgovara *točno određena i jedinstvena* vrijednost druge varijable, dok u slučaju *statističke* povezanosti određenoj vrijednosti jedne varijable mogu odgovarati barem dvije vrijednosti druge varijable.

Primjer 1. Akcijska cijena jednoga automobila marke "*Ficho*" je 50.000,00 €. Neka nam n označava broj kupljenih automobila, a C_n ukupan iznos koji treba izdvojiti za kupnju tih automobila. Tada je veza između varijabli n i C_n funkcijska i može se opisati analitičkim izrazom (*zatvorenom formulom*):

$$C_n = 50.000{,}00 \cdot n.$$

To znači da će *svaki* kupac koji kupi točno n automobila za njih morati izdvojiti točno $50.000{,}00 \cdot n$ €. I u slučaju uvođenja dodatnih varijabli (npr. dodatni popust na gotovinsko plaćanje) veza će i dalje ostati funkcijska.

Primjer 2. *"Čitlučka banka"* odobrava stambene kredite uz 5.99% složenih dekurzivnih godišnjih kamata na 30 godina uz otplatu jednakim godišnjim postnumerando anuitetima. Neka je C_0 iznos zajma koji se odobrava zajmoprimcu, a C_n ukupan iznos koji u tih 30 godina zajmoprimac treba vratiti banci. Tada je veza između varijabli C_0 i C_n funkcijska i može se opisati analitičkim izrazom (zatvorenom formulom):

$$C_n = 2.17713578 \cdot C_0.$$

To znači da *svaki* zajmoprimac koji dobije zajam u iznosu od C_0 kn banci treba vratiti ukupno $2.17713578 \cdot C_0$ kn.

Pojave koje istražuje statistika obično nisu tako "čvrsto" povezane kao varijable u prethodnim primjerima. Tada govorimo o *statističkoj* ili *stohastičkoj* vezi među varijablama.

Primjer 3. Ekonomskim je analizama utvrđeno da postoji veza između izdvajanja za marketing nekoga prehrambenoga proizvoda i ukupne vrijednosti prodanih komada toga

proizvoda. No, ne može se reći da npr. sve paštete za čiju je promidžbu utrošeno 100.000,00 kn imaju jednake ukupne vrijednosti prodanih komada.

Primjer 4. Ekonomskim je analizama utvrđeno da postoji veza između ukupnih prosječnih mjesečnih primanja i prosječnih mjesečnih izdvajanja za prehranu. No, ne može se reći da npr. svi ljudi koji mjesečno zarađuju prosječno 3.000,00 € imaju jednaka prosječna mjesečna izdvajanja za prehranu.

Korelacijska analiza primjenom posebnih metoda i tehnika ispituje stupanj povezanosti između barem dviju varijabli. Ukoliko se utvrdi određeni stupanj povezanosti između dviju varijabli, *regresijskom analizom* razvija se odgovarajući *analitički izraz* ili *algebarski model* te veze kako bi se moglo *prognozirati očekivanu* vrijednost jedne varijable za unaprijed zadanu vrijednost druge varijable.

4.1. *Korelacijska analiza*

Korelacijska analiza povezanosti dviju varijabli obično se provodi u tri koraka:

1.) konstrukcija odgovarajućega *grafičkog prikaza* odnosa među varijablama (tzv. *dijagram rasipanja*, *dijagram raspršenja* ili *oblak raspršenja*);
2.) određivanje *brojčanoga pokazatelja jakosti* i *smjera veze* među varijablama (*koeficijent korelacije*);
3.) određivanje *brojčanoga pokazatelja statističke značajnosti koeficijenta korelacije* (tzv. p – vrijednost).

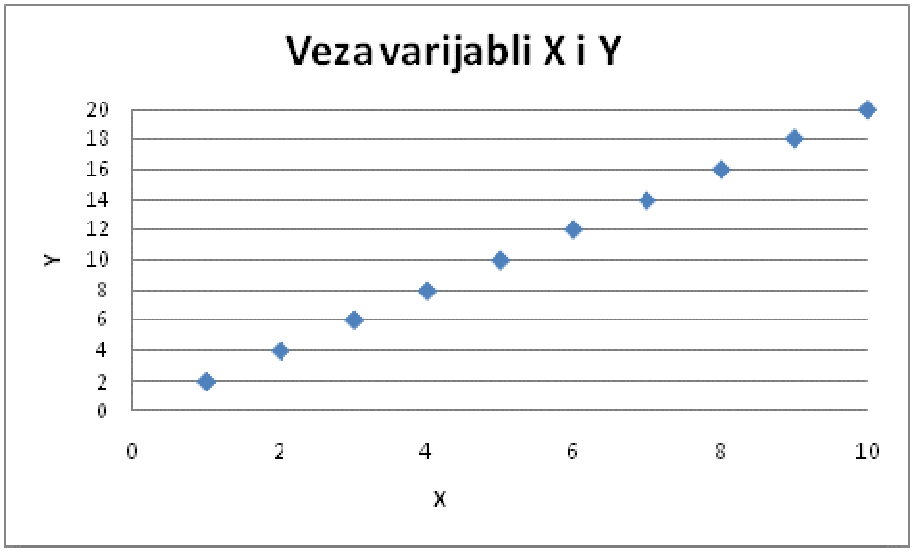
Ukoliko se utvrđuje povezanost najmanje triju varijabli, prvi korak se vrlo često izostavlja jer je prostorne analogone dijagrama rasipanja vrlo teško nacrtati čak i uz pomoć suvremenih računalnih programa. Stoga se dijagram rasipanja obično crta isključivo kod utvrđivanja povezanosti točno dviju varijabli.

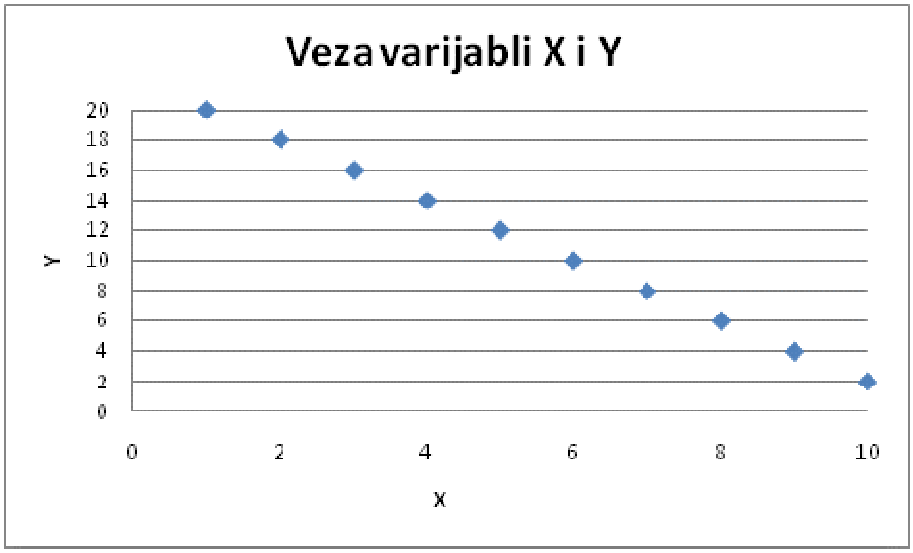
Govorimo li o odnosu između dviju varijabli, obično razlikujemo *nezavisnu varijablu* (X) i *zavisnu varijablu* (Y). Kako sugeriraju sami nazivi, *nezavisna varijabla* (X) je "varijabla koja utječe" na drugu varijablu, a moguće utjecaje na nju samu zanemarujemo, dok je *zavisna varijabla* (Y) "varijabla na koju se utječe", tj. varijabla čije vrijednosti zavise o vrijednostima nezavisne varijable X. U praksi, međutim, nije uvijek točno određeno koja varijabla je nezavisna, a koja zavisna. Npr. promatramo li varijable *ocjena iz matematike* i *ocjena iz statistike*, onda ocjena iz statistike može zavisiti o ocjeni iz matematike (dobar matematičar obično je i dobar statističar), ali i ocjena iz matematike može zavisiti i o ocjeni iz statistike (ako netko na ispitu iz statistike dobije 5, to ga može motivirati da sjedne i nauči matematiku barem za 4). U takvim slučajevima, dakle, ne možemo potpuno precizno utvrditi koja varijabla je zavisna, a koja nezavisna. No, u slučajevima kada imamo varijable poput *iznos mjesečne plaće* i *iznos mjesečnoga izdvajanja za prehranu* jasno je da iznos mjesečnoga izdvajanja za prehranu ovisi o iznosu primljene mjesečne plaće, a da iznos primljene

mjesečne plaće ne ovisi o iznosu mjesečnoga izdvajanja za prehranu (osim ako je posao djelatnika usko vezan za prehranu, kao što je npr. degustator).

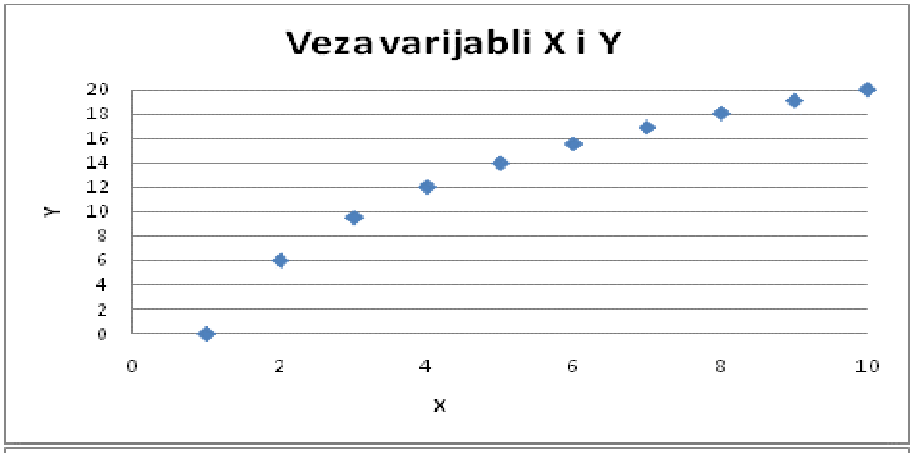
Kako je već istaknuto, kad god je to moguće, odnos između zavisne i nezavisne varijable najprije je zgodno predočiti grafički. U tu se svrhu rabi već spomenuti *dijagram rasipanja*. Taj se dijagram konstruira tako da da točke s odgovarajućim vrijednostima varijabli *X* i *Y* nanesemo u *pravokutni koordinatni sustav u ravnini*. Iz oblika dijagrama rasipanja može se približno procijeniti *postoji* li uopće veza između dviju varijabli i, ako postoji, kojega je *oblika* (pravocrtna (pravolinijska) ili krivocrtna (krivolinijska)).

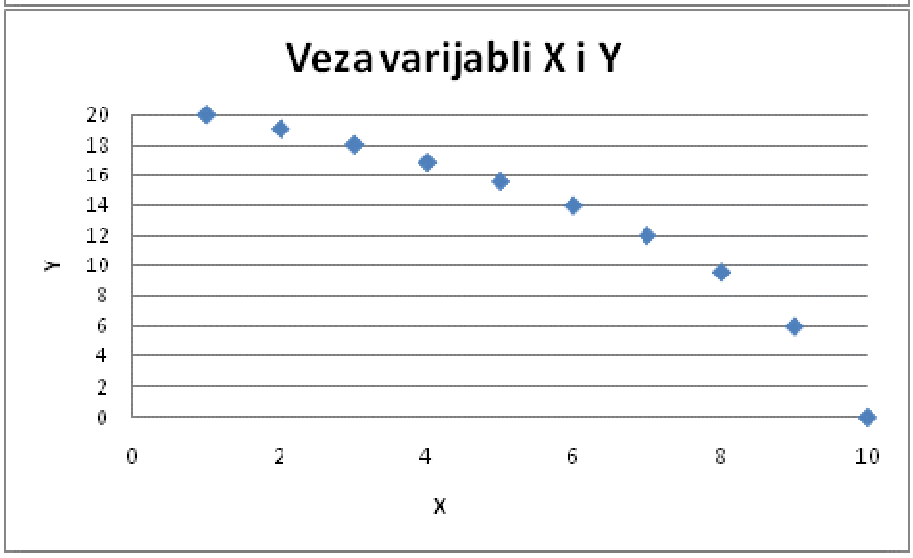
Primjer 1. Primjeri potpune linearne (pravocrtne) veze između dviju varijabli:



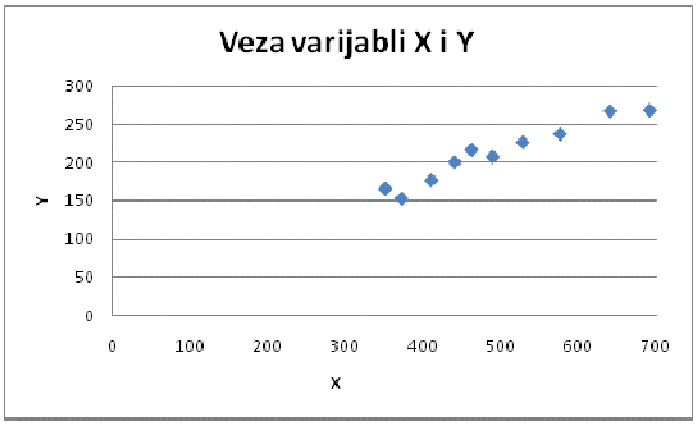


Primjer 2. Primjeri potpune nelinearne (krivocrtne) veze između dviju varijabli:

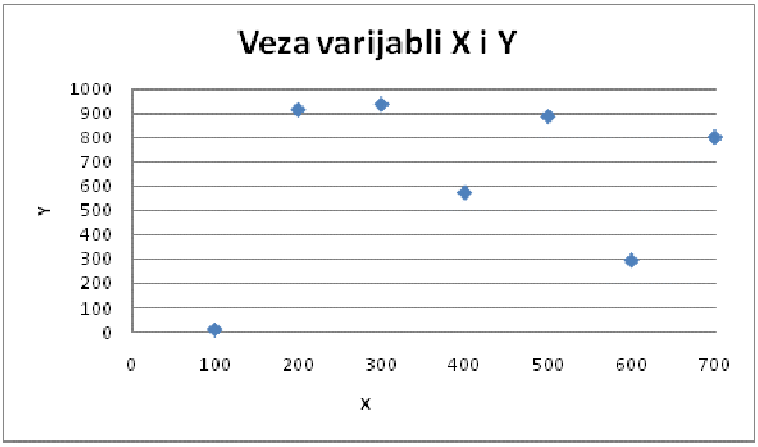




Primjer 3. Primjer postojanja statističke veze između dviju varijabli:



<u>Primjer 4.</u> Primjer slučajne (statistički beznačajne) veze između dviju varijabli:



Ukoliko se nakon konstrukcije dijagrama rasipanja utvrđi postojanje veze među promatranim varijablama, prelazi se na izračun odgovarajućega numeričkoga pokazatelja *<u>jakosti</u>* i *<u>smjera</u>* utvrđene veze. Pritom određivanje *<u>smjera veze</u>* znači utvrđivanje uzrokuje li *povećanje* vrijednosti jedne varijable *povećanje* ili *smanjenje* vrijednosti druge varijable. Ako *povećanje* vrijednosti jedne varijable uzrokuje *povećanje* vrijednosti druge varijable, kažemo da je veza među varijablama *<u>pozitivna</u>*. Uzrokuje li, pak, *povećanje* vrijednosti jedne varijable *smanjenje* vrijednosti druge varijable, kažemo da je veza *<u>negativna</u>*. Treba napomenuti da je teorijski veza među varijablama pozitivna (negativna) i ako smanjenje vrijednosti jedne varijable uzrokuje smanjenje (povećanje) vrijednosti druge varijable, ali se taj kriterij praktično ne koristi jer su vrijednosti na svakoj od osi u dijagramu rasipanja poredane uzlazno (od najmanje do najveće), pa je praktičnije promatrati povećanje bilo koje vrijednosti umjesto njezina smanjenja.

Najjednostavniji slučaj analitičkoga opisivanja veze među dvjema varijablama jest ukoliko iz dijagrama rasipanja utvrdimo da je ta veza linearna. U takvim se slučajevima računa tzv. *<u>Pearsonov[96] koeficijent jednostavne linearne korelacije</u>* (oznaka: r). Uz pretpostavku da imamo ukupno n uređenih parova (x_i, y_i) vrijednosti varijabli X i Y, koeficijent r računamo prema formuli:

$$r = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i \right)^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i \right)^2}}$$

[96] Karl Pearson (1857. – 1936.), britanski statističar, jedan od najzaslužnijih za razvoj statistike kao zasebne znanstvene discipline.

Može se pokazati da koeficijent r može poprimiti sve vrijednosti iz segmenta [–1, 1], pa je, dakle, njegova najmanja vrijednost jednaka –1, a najveća –1. Korelacijska analiza zasebno interpretira *predznak* Pearsonova koeficijenta jednostavne linearne korelacije, a zasebno njegovu *apsolutnu vrijednost*.

<u>*Pozitivan koeficijent jednostavne linearne korelacije*</u> (onaj koji je strogo veći od 0, odnosno onaj kojemu je predznak +) ukazuje da *povećanje vrijednosti jedne varijable uzrokuje povećanje vrijednosti druge varijable*, ali i obrnuto: *smanjenje vrijednosti jedne varijable uzrokuje smanjenje vrijednosti druge varijable*. Grubo bismo mogli reći da su promatrane varijable upravno razmjerne, no, to ne bi bilo matematički korektno: upravna razmjernost vrijednosti varijabli znači da ako se, za neki $k \in \langle 0, +\infty \rangle$, vrijednost jedne varijable poveća (smanji) za k%, onda se i vrijednost druge varijable poveća (smanji) za k%. Kod linearne ovisnosti, međutim, to općenito nije točno: ako se vrijednost jedne varijable poveća (smanji) za k%, vrijednost druge varijable će se svakako povećati (smanjiti), ali ne nužno za k% (nego najčešće za strogo manje od k%).

<u>*Negativan koeficijent jednostavne linearne korelacije*</u> ukazuje da *povećanje vrijednosti jedne varijable uzrokuje smanjenje vrijednosti druge varijable*, ali i obrnuto: *smanjenje vrijednosti jedne varijable uzrokuje povećanje vrijednosti druge varijable.* Kao i u slučaju pozitivnoga koeficijenta, grubo bismo mogli reći da su promatrane varijable obrnuto razmjerne, no, to ne bi bilo matematički korektno: obrnuta razmjernost vrijednosti varijabli znači da ako se, za neki $k \in \mathbf{R}$, vrijednost jedne varijable poveća (smanji) za k%, onda se vrijednost druge varijable smanji (poveća) za k%. Kod linearne ovisnosti, međutim, to općenito nije točno: ako se vrijednost jedne varijable poveća (smanji) za k%, vrijednost druge varijable će se smanjiti (povećati), ali ne nužno za k% (nego najčešće za strogo više od k%).

<u>*Apsolutna vrijednost koeficijenta korelacije*</u> (oznaka: $|r|$) ukazuje na <u>*jakost linearne*</u> <u>*veze*</u> među varijablama. Što je $|r|$ bliže nuli, veza je slabija, a što je bliže jedinici, veza je jača. Jedan od mogućih kriterija jakosti veze naveden je u donjoj tablici.

koeficijent korelacije	*jakost veze među varijablama*
1	potpuna
$0.80 \le \lvert r \rvert < 1$	jaka
$0.50 \le \lvert r \rvert < 0.8$	srednje jaka
$0.20 \le \lvert r \rvert < 0.5$	slaba
$0.0 < \lvert r \rvert < 0.2$	neznatna
0	potpuna odsutnost

Ovdje odmah treba istaknuti najčešću pogrešku koja se javlja prigodom uporabe Pearsonova koeficijenta jednostavne linearne korelacije. Naime, u slučajevima kad je $|r| \approx 0$, često se *pogrešno* zaključuje da nema *nikakve* statističke veze među promatranim varijablama. Veza među varijablama može biti i eksponencijalna, logaritamska, potencijska, polinomna (takva da je stupanj odgovarajućega polinoma najmanje 2) itd., a ne samo linearna. Zbog toga prigodom interpretacije ovoga koeficijenta uvijek treba imati na umu da on opisuje *isključivo* smjer i jakost *linearne* veze među varijablama.

Posljednji korak korelacijske analize je *procjenjivanje statističke značajnosti* izračunatoga koeficijenta korelacije. Grubo govoreći, takvom procjenom želimo utvrditi je li promjena vrijednosti zavisne varijable doista uzrokovana utjecajem, odnosno promjenom vrijednosti nezavisne varijable ili je do promjene vrijednosti zavisne varijable došlo slučajno stjecajem drugih okolnosti (pod djelovanjem tzv. *nekontroliranih uvjeta* ili *činitelja*). U tu svrhu s p označavamo *vjerojatnost* da je do promjene zavisne varijable došlo slučajno stjecajem drugih okolnosti i taj broj nazivamo *p – vrijednost*. Broj p je neki element segmenta [0, 1] i za njega određujemo *gornju granicu* do koje ćemo smatrati da je do promjene vrijednosti zavisne varijable došlo slučajno. Obično se razlikuju dva kriterija:

- *blaži kriterij*: smatramo da je do promjene vrijednosti zavisne varijable došlo slučajno ako i samo ako je $p \geq 0.05$ i tada kažemo da smo procjenu izvršili uz *razinu pouzdanosti od* 95%;
- *stroži kriterij*: smatramo da je do promjene vrijednosti zavisne varijable došlo slučajno ako i samo ako je $p \geq 0.01$ i tada kažemo da smo procjenu izvršili uz *razinu pouzdanosti od* 99%.

Ekvivalentno, prema blažem kriteriju smatramo da je promjena vrijednosti zavisne varijable doista uzrokovana promjenom vrijednosti nezavisne varijable, tj. da je *veza promatranih varijabli statistički značajna* ako i samo ako je $p < 0.05$. U slučaju strožega kriterija odgovarajući uvjet je $p < 0.01$.

Budući da ćemo raditi uglavnom s "malim" uzorcima (koji imaju najviše 30 elemenata), p – vrijednost koja odgovara Pearsonovu koeficijentu jednostavne linearne korelacije računat ćemo koristeći tzv. *Studentovu t– razdiobu*[97] ili, kako se često naziva, "razdiobu malih uzoraka" jer se u slučaju "velikih“ uzoraka (koji imaju npr. 300 elemenata) praktički podudara s normalnom razdiobom. Jedan od modificiranih algoritama za izračun p – vrijednosti Pearsonova koeficijenta jednostavne linearne korelacije na temelju Studentove t – razdiobe u MS Excelu može se formulirati ovako:

Korak 1. Izračunati Pearsonov koeficijent jednostavne linearne korelacije (r).
Korak 2. Izračunati vrijednost broja $d := N - 2$, gdje je N *opseg uzorka* (tj. ukupan broj elemenata u uzorku). Vrijednost d obično se naziva *broj stupnjeva slobode*[98].
Korak 3. Izračunati pripadnu t – *vrijednost* prema formuli

$$t = r \cdot \sqrt{\frac{d}{1-r^2}}\,.$$

[97] Studentovu t-razdiobu je 1908. prvi otkrio engleski statističar William Sealy Gosset (1876. – 1937.), suradnik Karla Pearsona. On je tada radio za pivovaru „*Guiness*“ koja je zahtijevala od njega da svoje radove objavljuje pod pseudonimom, pa je Gosset objavio članak o novoj razdiobi u Pearsonovu časopisu "*Biometrics*" pod pseudonimom *Student*. Taj se pseudonim i danas koristi u nazivu razdiobe.
[98] Grubo govoreći, broj stupnjeva slobode može se definirati kao ukupan broj međusobno različitih "slobodnih" (nezavisnih) parametara čije vrijednosti treba zadati kako bi se potpuno odredila vrijednost neke veličine.

Korak 4. Izračunati p – vrijednost kao vrijednost Studentove t – razdiobe s parametrima $|t|$, d i 2. (Broj 2 treba upisati jer kod korelacije uvijek promatramo *dvosmjerne* veze, tj. ne treba unaprijed zadati koja varijabla je zavisna, a koja nezavisna. Ukoliko se to ipak učini, govori se o *jednosmjernoj* vezi.) U MS Excelu ta se vrijednost dobiva primjenom funkcije TDIST.
Korak 5. Usporediti p – vrijednost ili s 0.01 (stroži kriterij) ili s 0.05 (blaži kriterij). Pearsonov koeficijent jednostavne linearne korelacije je statistički značajan ako i samo ako je $p < 0.01$ (ili $p < 0.05$).

Pokažimo primjenu navedenih postupaka na primjerima.

Primjer 5. (preuzeto iz *The World Almanac and The Book of Fact*, 1975.) Ispituje se odnos između visine i mase žena u dobi od 30 do 40 godina. Odabrano je ukupno 15 uzoraka žena navedene dobi i dobiveni su sljedeći podaci:

uzorak	*prosječna visina* [m]	*prosječna masa* [kg]
1	1.47	52.21
2	1.5	53.12
3	1.52	54.48
4	1.55	55.84
5	1.57	57.2
6	1.60	58.57
7	1.63	59.93
8	1.65	61.29
9	1.68	63.11
10	1.7	64.47
11	1.73	66.28
12	1.75	68.1
13	1.78	69.92
14	1.8	72.19
15	1.83	74.46
ukupno:	24.76	931.17

Odredimo Pearsonov koeficijent jednostavne linearne korelacije i procijenimo njegovu statističku značajnost uz razinu pouzdanosti od 95%. Radi određenosti, neka nezavisna varijabla (X) bude varijabla *prosječna visina*, a zavisna varijabla (Y) varijabla *prosječna masa*.[99] Izračunajmo najprije sve potrebne pomoćne numeričke podatke:

[99] Suprotna bi pretpostavka (ovisnost visine o masi) bila daleko više motivirajuća npr. za liječenje od anoreksije i bulimije.

uzorak	*prosječna visina* [m] (*X*)	*prosječna masa* [kg] (*Y*)	$x_i \cdot y_i$	x_i^2	y_i^2
1	1.47	52.21	76.7487	2.1609	2725.8841
2	1.5	53.12	79.68	2.25	2821.7344
3	1.52	54.48	82.8096	2.3104	2968.0704
4	1.55	55.84	86.552	2.4025	3118.1056
5	1.57	57.2	89.804	2.4649	3271.84
6	1.6	58.57	93.712	2.56	3430.4449
7	1.63	59.93	97.6859	2.6569	3591.6049
8	1.65	61.29	101.1285	2.7225	3756.4641
9	1.68	63.11	106.0248	2.8224	3982.8721
10	1.7	64.47	109.599	2.89	4156.3809
11	1.73	66.28	114.6644	2.9929	4393.0384
12	1.75	68.1	119.175	3.0625	4637.61
13	1.78	69.92	124.4576	3.1684	4888.8064
14	1.8	72.19	129.942	3.24	5211.3961
15	1.83	74.46	136.2618	3.3489	5544.2916
ukupno:	24.76	931.17	1548.245	41.053	58498.5439

Stoga je traženi Pearsonov koeficijent jednostavne linearne korelacije:

$$r = \frac{\sum_{i=1}^{15} x_i \cdot y_i - \frac{1}{15} \cdot \sum_{i=1}^{15} x_i \cdot \sum_{i=1}^{15} y_i}{\sqrt{\sum_{i=1}^{15} x_i^2 - \frac{1}{15} \cdot \left(\sum_{i=1}^{15} x_i\right)^2} \cdot \sqrt{\sum_{i=1}^{15} y_i^2 - \frac{1}{15} \cdot \left(\sum_{i=1}^{15} y_i\right)^2}},$$

$$r = \frac{1548.245 - \frac{1}{15} \cdot 24.76 \cdot 931.17}{\sqrt{41.053 - \frac{1}{15} \cdot 24.76^2} \cdot \sqrt{58498.5439 - \frac{1}{15} \cdot 931.17^2}}$$

$$r = 0{,}994583794$$

Prije interpretacije dobivenoga koeficijenta procijenimo njegovu statističku značajnost. Izračunajmo najprije pripadnu t – vrijednost:

$$t = r \cdot \sqrt{\frac{d}{1 - r^2}} = 0.994583794 \cdot \sqrt{\frac{15 - 2}{0.994583794}} \approx 34.5016057\,.$$

Pripadnu p – vrijednost izračunamo koristeći MS Excel. Dobivamo:

$$p \approx 3.60352\text{E-}14 = 0.00000000000000360352.$$

Očito je $p < 0{,}05$, pa *postoji statistički značajna linearna veza između prosječne visine i prosječne mase* na promatranim uzorcima žena. Iz vrijednosti Pearsonova koeficijenta jednostavne linearne korelacije zaključujemo da je ta veza *pozitivna i jaka*, tj. da će povećanje prosječne visine vrlo vjerojatno uzrokovati i povećanje prosječne mase (i obrnuto).

Primjer 6. U sljedećoj je tablici navedeno ukupno 10 parova vrijednosti nezavisne varijable X i zavisne varijable Y.

I	x_i	y_i
1	10	0.666654428469398
2	11,8	0.666664817791255
3	16	0.666666644193195
4	20	0.666666666329664
5	30	0.666666666666657
6	50	0.666666666666667
7	70	0.666666666666667
8	90	0.666666666666667
9	120	0.666666666666667
10	600	0.666666666666667

Odgovarajućim se izračunom dobiva da je vrijednost Pearsonova koeficijenta jednostavne linearne korelacije

$$r = 0.208423,$$

pripadna t – vrijednost

$$t \approx 0.602745504,$$

a toj vrijednosti odgovarajuća p – vrijednost

$$p \approx 0.563369789.$$

Uz razinu pouzdanosti od 95%, može se zaključiti da *ne postoji statistički značajna linearna korelacija* promatranih varijabli. No, zaključak da ne postoji *bilo kakva* korelacija promatranih varijabli potpuno je pogrešan: veza promatranih varijabli, naime, zapravo je potpuna, tj. funkcijska i dana je analitičkim izrazom:

$$y = \frac{1}{1.5 + e^{-1.05 \cdot x}} \text{ (primjer tzv. } \textit{logističke funkcije}\text{).}$$

Ovaj primjer još jednom ukazuje na činjenicu da se u primjeni Pearsonova koeficijenta jednostavne linearne korelacije uporno mora naglašavati riječ *linearne* jer se radi o pokazatelju isključivo linearne, a ne bilo kakve veze između promatranih varijabli.

4.1.1. *Spearmanov*[100] *koeficijent korelacije ranga*

Pearsonov se koeficijent jednostavne linearne korelacije u pravilu primjenjuje za utvrđivanje stupnja *linearne* povezanosti između numeričkih varijabli čije jedinice odgovaraju *intervalnoj* ili *omjernoj* skali. No, ukoliko imamo dvije varijable od kojih je barem jedna *redosljedna* (*varijabla ranga*), ne moramo zahtijevati da određivanje stupnja linearne povezanosti, niti da vrijednosti varijable (modaliteti) pripadaju intervalnoj ili omjernoj skali. U takvim slučajevima za određivanje stupnja linearne veze obično se koristi *Spearmanov koeficijent korelacije ranga* (oznaka: r_s)

Spearmanov koeficijent korelacije ranga pripada u tzv. *neparametarske mjere* korelacije jer njegovo izračunavanje ovisi o *poretku* vrijednosti modaliteta promatranih varijabli, a ne izravno o samim vrijednostima varijabli. Pritom se *ne* pretpostavlja da vrijednosti varijable pripadaju nekoj od uobičajenih razdioba (npr. normalnoj razdiobi). Već smo istakli da se ne zahtijeva utvrđivanje linearne veze i pripadnost modaliteta intervalnoj ili omjernoj skali. Napomenimo da se obično *ne* računa za pojave u *bilo kakvom monotonom* (rastućem ili padajućem) odnosu.

Dakle, pretpostavljamo da modaliteti pripadaju *ordinalnoj* skali, pa modalitete možemo *rangirati*, odnosno poredati od najslabijega prema najboljemu (*uzlazno*) ili od najboljega prema najslabijemu (*silazno*). *Dogovorno* ćemo modalitete rangirati *silazno* od najboljega prema najslabijemu tako da najbolji modalitet uvijek ima rang 1. Rang pojedinoga modaliteta zapravo možemo shvatiti kao broj položaja (mjesta, pozicije) toga modaliteta u "rang-listi".

Oprez: Dogovorno modalitete rangiramo silazno, tj. najbolji modalitet uvijek dobiva rang 1. No, ovisno o promatranom obilježju, povećanje ranga ne mora nužno značiti smanjenje odgovarajuće vrijednosti modaliteta. Npr. rangiramo li neki uzorak studenata prema trenutnom prosjeku ocjena, onda će student s najboljim prosjekom imati rang 1. Povećanje ranga (s 1 na 2) u ovom slučaju znači uobičajeno smanjenje vrijednosti modaliteta. No, rangiramo li isti uzorak studenata prema broju izostanaka s redovne nastave, onda će student s najmanjim brojem izostanaka imati najbolji rang, tj. u ovome slučaju povećanje ranga povlači povećanje vrijednosti modaliteta. Ovo obavezno treba imati na umu prigodom *interpretacije* Spearmanova koeficijenta korelacije ranga.

Pri ovakvom rangiranju se prirodno pojavljuje problem određivanja ranga barem dvaju jednakih modaliteta. U takvim slučajevima govorimo o tzv. *vezanim rangovima*. Npr. u nizu ocjena 5, 4, 2, 2, 2 imamo 3 jednaka modaliteta (2), pa treba odrediti rang svakoga od njih. Budući da su modaliteti međusobno jednaki, nema razloga preferirati bilo kojega od njih, pa se utvrđuje tzv. *zajednički rang* kao aritmetička sredina trenutnih pozicija istih modaliteta. Konkretno, vrijednost 5 ima rang 1 (najbolja ocjena), vrijednost 4 ima rang 2, a tri iste vrijednosti 2 imaju zajednički rang $\frac{3+4+5}{3}=4$ jer se u *trenutnom* nizu nalaze na 3., 4. i 5. mjestu.

---

[100] Charles Edward Spearman (1863. – 1945.), engleski statističar i psiholog, jedan od začetnika ideje faktorske analize varijance.

Unatoč ovakvom načinu "eliminiranja" vezanih rangova, treba istaknuti da se u slučajevima pojave nezanemarivoga broja vezanih rangova (u odnosu na ukupan broj parova vrijednosti varijabli), umjesto Spearmanova, preporučuje računanje Pearsonova koeficijenta jednostavne linearne korelacije. Pritom kao vrijednosti nezavisne, odnosno zavisne varijable treba uzeti izračunate rangove modaliteta.

Nakon rangiranja modaliteta, Spearmanov koeficijent r_s računa se prema formuli :

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n},$$

gdje su:

n – ukupan broj elemenata niza vrijednosti varijable X (ili varijable Y)

$d_i = r(x_i) - r(y_i)$ = *razlika rangova* odgovarajućega para vrijednosti varijable X i varijable Y , za svaki $i = 1, 2, \ldots, n$.

Navedimo neka korisna svojstva vezana uz Spearmanov koeficijent korelacije ranga:

<u>*Svojstvo* 1.</u> Zbroj svih razlika rangova jednak je 0.
<u>*Svojstvo* 2.</u> Ako su u svakom paru rangovi jednaki, njihove su razlike jednake nuli. U tom je slučaju vrijednost Spearmanova koeficijenta korelacije ranga jednaka $r_s = 1$ i tada govorimo o *potpunoj* (*savršenoj*) *pozitivnoj korelaciji ranga*.
<u>*Svojstvo* 3.</u> Ako je redoslijed rangova jedne varijable suprotan redoslijedu rangova druge varijable, vrijednost Spearmanova koeficijenta korelacije ranga jednaka je $r_s = -1$ i tada govorimo o *potpunoj* (*savršenoj*) *negativnoj korelaciji ranga*.
<u>*Svojstvo* 4</u>. Kao kriterij jakosti veze među varijablama ranga koristi se kriterij naveden kod Pearsonova koeficijenta jednostavne linearne korelacije.

<u>Primjer 1.</u> Na slučajno odabranom uzorku od 10 ljudi ispituje se veza kvocijenta inteligencije i ukupnoga tjednoga vremena provedenoga u gledanju TV programa. Pritom se, kao radna pretpostavka, uzima da je poželjno što manje vremena tjedno provoditi pred televizijskim prijamnikom. Dobiveni podaci prikazani su u donjoj tablici.

količnik inteligencije (IQ)	*vrijeme provedeno u gledanju TV* [sati/tjedan]
106	7
86	0
100	27
101	50
99	28
103	29

količnik inteligencije (IQ)	*vrijeme provedeno u gledanju TV* [sati/tjedan]
97	20
113	12
112	6
110	17

Ispitajmo vezu promatranih varijabli pomoću Spearmanova koeficijenta korelacije ranga. Najprije rangirajmo modalitete. Kod varijable *IQ* logično je da osoba (iz promatranoga uzorka) s najvećim količnikom inteligencije ima rang 1, a osoba s najmanjim koeficijentom inteligencije rang 10. Ovdje, dakle, povećanje ranga znači smanjenje vrijednosti varijable, odnosno smanjenje količnika inteligencije. Kod varijable *vrijeme provedeno u gledanju TV* osobi koja tjedno najmanje vremena provede u gledanju televizijskih programa, u skladu s polaznom temeljnom pretpostavkom, dodijelit ćemo rang 1, a osobi koja tjedno najviše vremena provede u gledanju televizijskih programa rang 10. Tako dobivamo sljedeću tablicu:

količnik inteligencije (IQ) (x_i)	*vrijeme provedeno u gledanju TV* [sati/tjedan] (y_i)	$r(x_i)$	$r(y_i)$	d_i	d_i^2
106	7	4	3	1	1
86	0	10	1	9	81
100	27	7	7	0	0
101	50	6	10	-4	16
99	28	8	8	0	0
103	29	5	9	-4	16
97	20	9	6	3	9
113	12	1	4	-3	9
112	6	2	2	0	0
110	17	3	5	-2	4
ukupno:				0	136

Prema tome, Spearmanov koeficijent korelacije ranga jednak je:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{10} d_i^2}{10^3 - 10} = 1 - \frac{6 \cdot 136}{990} \approx 0.175757576.$$

Dakle, na temelju promatranoga uzorka može se zaključiti da postoji *neznatna pozitivna korelacija* količnika inteligencije i vremena provedenim u gledanju televizijskih programa. Napomenimo da se može pokazati da je pripadna p – vrijednost strogo veća od 0.05, tj. da izračunati Spearmanov koeficijent korelacije ranga *nije* statistički značajan pokazatelj.

Primjer 2. U donjoj su tablici navedeni ukupni bodovi na pismenim ispitima iz *Gospodarske matematike*, odnosno *Poslovne statistike* za svakoga od 8 slučajno izabranih studenata stručnoga studija računovodstva i financija Veleučilišta u Frkljevcima u akademskoj godini 2006/2007.

bodovi na ispitu iz gospodarske matematike	*bodovi na ispitu iz poslovne statistike*
45	50
20	10
50	50
0	20
100	90
50	50
30	40
80	100

Odredimo Spearmanov koeficijent korelacije ranga i objasnimo njegovo značenje. U ovom se zadatku radi o slučaju gdje nije potpuno jasno koja je varijabla zavisna, a koja nezavisna. No, kako ispitujemo isključivo *stupanj veze* među promatranim varijablama, ta činjenica nije bitna. Stoga ćemo pretpostaviti da su bodovi na ispitu iz *Gospodarske matematike* nezavisna, a bodovi na ispitu iz *Poslovne statistike* zavisna varijabla. Navedene su varijable očito redosljedne, tj. student koji je na ispitu iz *Gospodarske matematike* postigao 100 bodova bolji je od studenta koji je na istom ispitu postigao 50 bodova itd. U oba slučaja modalitete obilježja poredat ćemo silazno jer student s najvećim brojem bodova treba ima najveći rang. Tako dobivamo sljedeću tablicu:

bodovi na ispitu iz Gospodarske matematike (x_i)	*bodovi na ispitu iz Poslovne statistike* (y_i)	$r(x_i)$	$r(y_i)$	d_i	d_i^2
45	50	5	4	1	1
20	10	7	8	-1	1
50	50	3.5	4	-0.5	0.25
0	20	8	7	1	1
100	90	1	2	-1	1
50	50	3.5	4	-0.5	0.25
30	40	6	6	0	0
80	100	2	1	1	1
ukupno:				0	5.50

Uočimo da u vrijednostima varijable X imamo dvije međusobno jednake (50) koje dijele 3. i 4. mjesto. Stoga je njihov zajednički rang jednak $\frac{3+4}{2} = 3.5$. Analogno, u varijabli Y imamo 3 takve vrijednosti (ponovno 50) koje dijele 3., 4. i 5. mjesto. Stoga je njihov zajednički rang

jednak $\frac{3+4+5}{3} = 4$. Tako je traženi Spearmanov koeficijent korelacije ranga jednak:

$$r_s = 1 - \frac{6 \cdot 5.50}{8^3 - 8} \approx 0.93$$

Zaključujemo da postoji *jaka pozitivna* korelacija između bodova na ispitu iz matematike i bodova na ispitu iz statistike. Iz dobivena koeficijenta zaključujemo da *povećanje ranga* broja bodova iz *Gospodarske matematike* vrlo vjerojatno uzrokuje *povećanje ranga* broja bodova iz *Poslovne statistike*. To znači da će student koji ostvari visoki rezultat u jednom ispitu *vrlo vjerojatno* postići visoki rezultat i u drugom, i obrnuto. Može se pokazati da je pripadna p – vrijednost strogo manja od 0,01, pa je dobiveni koeficijent statistički značajan.

Primjer 3. Mužjaci smeđokriloga brzana (*Fregata maginificens*) imaju crvenu grlenu vrećicu koju u doba svadbenoga plesa[101] nadimaju kao svadbeno ruho (vidjeti sliku) i pomoću nje proizvode poseban zvuk kojim privlače ženke. Proučavajući opisano ponašanje, grupa ornitologa utvrdila je da ženke odabiru mužjake upravo na temelju veličine njihove grlene vrećice, pa je provedeno dodatno istraživanje postoji li statistički značajna korelacija između obujma grlene vrećice i frekvencije zvuka kojim mužjaci privlače ženke. U donjoj su tablici navedeni prikupljeni podaci.

obujam grlene vrećice [cm^3]	*frekvencija zvuka* [Hz]
1760	529
2040	566
2440	473
2550	461
2730	465
2740	532
3010	484
3080	527
3370	488
3740	485
4910	478
5090	434
5090	468
5380	449
5850	425
6730	389
6990	421
7960	416

mužjak smeđokriloga brzana

[101] *Svadbeni ples* u biologiji je skupni naziv za sva ponašanja koja predstavljaju uvod u parenje životinja.

Korelacija promatranih varijabli utvrđena je na temelju Spearmanova koeficijenta korelacije ranga: mužjak s najvećim obujmom grlene vrećice ima rang 1, a mužjak s najmanjim obujmom rang 18, pri čemu mužjaci s obujmom grlene vrećice 5090 cm^3 imaju rang 6.5. Za varijablu *frekvencija zvuka* primijenjen je suprotan kriterij: mužjak s najdubljom frekvencijom zvuka ima rang 1, a mužjak s najvišom frekvencijom rang 18. Izračunom Spearmanova koeficijenta korelacije ranga dobiva se:

$$r_s = 0.763158.$$

Dakle, postoji srednje jaka pozitivna korelacija između obujma grlene vrećice smeđokriloga brzana i frekvencije zvuka kojim vabi ženke. To znači da povećanje ranga obujma grlene resice vjerojatno uzrokuje povećanje ranga frekvencije zvuka, odnosno da će mužjak s većim obujmom grlene vrećice vrlo vjerojatno imati dublju frekvenciju zvuka kojim privlači ženke. Napomenimo da je na temelju odgovarajućega statističkoga pokazatelja utvrđeno da je dobiveni rezultat statistički značajan, tj. zaključak izveden iz promatranoga uzorka od 18 mužjaka može se primijeniti i na cjelokupnu populaciju smeđokriloga brzana, te da se ovisnost frekvencije zvuka (zavisne varijable Y) o obujmu grlene vrećice (nezavisnoj varijabli X) može relativno dobro opisati tzv. regresijskom jednadžbom:

$$Y = 563.95 \cdot e^{-0,0005 \cdot X}.$$

<u>*Zadatak za vježbu*</u>: Unesite podatke iz gornje tablice u radni list MS Excela, pa provjerite valjanost izračuna Spearmanova koeficijenta korelacije ranga. *Oprez*: Rangove mužjaka s obujmom grlene vrećice 5090 cm^3 morate upisati sami.

4.2. *Regresijska analiza*

Korelacijska je analiza vrlo često tek prvi korak u utvrđivanju međusobne ovisnosti dviju ili više promatranih varijabli. Ukoliko se pomoću odgovarajućih statističkih pokazatelja utvrdi da postoji statistički značajna veza među promatranim varijablama, sljedeći korak u istraživanju te veze je tzv. <u>*regresijska*[102] *analiza.*</u>

Regresijsku su analizu u statistici prvi koristili matematičari Legendre[103] i Gauss koji su početkom 19. stoljeća objavili radove o *metodi najmanjih kvadrata*. Međutim, sâm pojam *regresijska analiza* u statistiku je uveo Sir Francis Galton[104]. On je 1886. u *Glasniku antropološkoga instituta* u Londonu objavio rad u kojemu je promatrao ovisnost visine djece o prosječnoj visini njihovih roditelja. U tom je istraživanju došao do zanimljiva zaključka:

---

[102] Riječ *regresija* potječe od latinske riječi *regressio* = povratak, uzmak, odstupanje, nazadovanje. Objašnjenje uporabe ovoga naziva u statistici slijedi u daljnjem tekstu.
[103] Adrien-Marie Legendre (1752. – 1833.), francuski matematičar, bavio se matematičkom analizom, teorijom brojeva i algebrom. Jedan od Mjesečevih kratera dobio je ime po njemu.
[104] Francis Galton (1822. – 1911.), engleski znanstvenik, nećak Charlesa Darwina, začetnik moderne korelacijske i statističke analize, mentor Karla Pearsona i Charlesa Edwarda Spearmana.

ukoliko je prosječna visina obaju roditelja veća od *medijana* uređenoga niza svih prosječnih visina roditelja, visina njihova djeteta će očekivano biti *manja* od prosječne visine obaju roditelja (i obrnuto: ukoliko je prosječna visina obaju roditelja manja od medijana, visina njihova djeteta očekivano će biti veća od prosječne visine od prosječne visine obaju roditelja). Ovakvu je pojavu Galton nazivao *efekt nazadovanja* ili *regresijski efekt*. Zaslugom Karla Pearsona, koji je poopćio Galtonova razmatranja na opće statističke modele, riječ *regresija* u statistici je dobila današnje značenje iako, praktično, s tim značenjem nema gotovo nikakvih dodirnih točaka.

Regresijska analiza ili, kraće, *regresija* općenito obuhvaća primjenu različitih metoda ispitivanja ovisnosti *zavisne* varijable (Y) o barem jednoj *nezavisnoj* varijabli (X). Temeljni je cilj provedbe takvih metoda vezu između promatranih varijabli izraziti ili opisati odgovarajućim analitičkim izrazom (*regresijskim modelom*). Takav izraz omogućuje ne samo objašnjavanje ovisnosti promatranih pojava, nego i *procjenjivanje* vrijednosti zavisne varijable Y za određene vrijednosti barem jedne nezavisne varijable X.

Ovisno o broju zavisnih i nezavisnih varijabli koje se pojavljuju u odgovarajućem modelu, regresija općenito može biti:

1.) *jednostavna* (u modelu se pojavljuje točno jedna zavisna i točno jedna nezavisna varijabla);
2.) *višestruka* ili *multipla* (u modelu se pojavljuju točno jedna zavisna i barem dvije nezavisne varijable).

U ovom ćemo se kolegiju baviti modelima jednostavne regresije, a od primjera modela višestruke regresije kratko ćemo spomenuti samo najjednostavniji i najprimjenjivaniji model višestruke linearne regresije**.**

Opći oblik modela jednostavne regresije jest:

$$Y = f(X) + e \text{ ili } Y = \hat{Y} + e.$$

U ovakvom modelu varijabla e je tzv. *stohastička* (*slučajna*) *varijabla* koja, slobodno govoreći, "mjeri" stupanj nesustavnih utjecaja na zavisnu varijablu Y. To je potpuno u skladu s ranijom podjelom veza među varijablama na funkcijske i statističke: kod funkcijskih su veza vrijednosti varijable u jednake nuli, dok su kod statističkih veza one različite od nule. Opisna statistika – kojom se bavimo u ovom kolegiju – ne bavi se posebnom analizom stohastičke varijable e, nego je smatra isključivo odstupanjima od funkcijskoga dijela regresijskoga modela[105].

O čemu se, zapravo, radi? Kad istražujemo ovisnosti pojedinih varijabli, statističkim istraživanjem dobivamo određene podatke o vrijednostima tih varijabli. Npr. želimo li utvrditi

[105] Otuda dolazi i oznaka e: engl. *error* = greška.

kako mjesečno izdvajanje za odjeću zaposlenika tvrtke "*Drpislavić* d.o.o." iz Babine Grede ovisi o njihovim mjesečnim primanjima, statističkim istraživanjem (npr. anketiranjem) dobivamo podatke o mjesečnim primanjima i izdvajanja za odjeću svakoga od zaposlenika. Te podatke sredimo i pregledno prikažemo tablično, kako smo to naučili u 2. poglavlju. Označimo li s X nezavisnu varijablu *mjesečna primanja*, a s Y zavisnu varijablu *izdvajanja za odjeću*, cilj nam je odrediti *funkciju* koja će svakoj vrijednosti varijable X pridružiti *statističkim istraživanjima dobivenu* vrijednost varijable Y. Najčešće takva funkcija *ne postoji* jer je veza među varijablama stohastička. Stoga želimo naći funkciju $f(X)$ koja će što točnije vrijednostima varijable X pridruživati vrijednosti varijable Y. U svakom takvom pridruživanju nužno će se javiti određena greška (tzv. *rezidualno*[106] *odstupanje*). Iako bi teorijski bilo najbolje da svako pojedino rezidualno odstupanje bude što manje, to praktično nije moguće: minimiziranje odstupanja nastaloga jednim pridruživanjem u pravilu uzrokuje povećanje odstupanja nastaloga drugim pridruživanjem. Može se pokazati da je prosječna vrijednost svih rezidualnih odstupanja jednaka nuli, što znači da – kao i kod uvođenja pojma standardne devijacije (vidjeti poglavlje 3.) – nije moguće minimizirati prosječno rezidualno odstupanje. Stoga je potrebno analizirati i minimizirati *varijancu* svih rezidualnih odstupanja (koja je uvijek strogo veća od nule). Takvo razmatranje dovodi do definiranja pojmova *varijanca regresije* i *standardna devijacija regresije* o kojima ćemo nešto više reći u točki 4.2.3.

Prema tome, dva su osnovna cilja koja se žele postići stvaranjem regresijskoga modela:

1) odrediti *tip* realne funkcije (linearna, kvadratna, eksponencijalna, logaritamska, …) koja najbolje opisuje vezu između promatranih varijabli;
2) procijeniti osnovne parametre te funkcije tako da zbroj kvadrata svih rezidualnih odstupanja bude što manji.

U nastavku ćemo dati kratak opis nekoliko takvih regresijskih modela.

4.2.1. *Model jednostavne linearne regresije*

Ukoliko iz dijagrama rasipanja uvidimo da ravnomjerno povećanje vrijednosti nezavisne varijable X uzrokuje približno ravnomjerno povećanje (ili smanjenje) zavisne varijable Y, regresijski model koji najbolje opisuje promatranu vezu varijabli X i Y jest *model jednostavne linearne regresije.*

Opći oblik toga modela je:

$$Y = aX + b + e,$$

gdje su a, $b \in \mathbf{R}$ *parametri modela*, a e, kako je istaknuto u uvodu, stohastička (slučajna) varijabla. Cilj nam je odrediti parametre a i b tako da varijanca svih rezidualnih odstupanja

[106] Lat. *residuum* = ostatak; razlika

bude minimalna. Budući da je graf svake linearne funkcije *pravac*, to praktično znači da tražimo jednadžbu pravca koji će najbolje aproksimirati skup točaka dobiven konstrukcijom dijagrama rasipanja.

Svakim statističkim istraživanjem dobivamo ukupno n različitih uređenih parova vrijednosti varijable X i vrijednosti varijable Y, tj. dobivamo uređene parove (x_1, y_1), (x_2, y_2), … (x_n, y_n). Za svaki od tih parova vrijedi jednakost:

$$y_i = a \cdot x_i + b + e_i,\ i = 1, 2, \ldots, n,$$

pri čemu je e_i rezidualno odstupanje, odnosno "promašaj" koji činimo zamijenimo li statističkim istraživanjem dobivenu vrijednost y_i računski dobivenom vrijednošću $a \cdot x_i + b$.[107] Time, zapravo, dobivamo sustav od n linearnih jednadžbi s $n + 2$ nepoznanice: a, b, e_1, e_2, …, e_n. Taj sustav ima beskonačno mnogo realnih rješenja, ali točno jedno od njih zadovoljava uvjet da varijanca svih rezidualnih odstupanja bude minimalna. Takvo se rješenje određuje tzv. *metodom najmanjih kvadrata*.

Osnovna ideja metode najmanjih kvadrata je postići da *zbroj kvadrata* odstupanja empirijskih vrijednosti zavisne varijable y_i od očekivanih vrijednosti (oznaka: $\hat{Y}_i$) te varijable bude *minimalan*. Budući da se pri izvodu parametara koristi tzv. diferencijalni račun, u detalje cijeloga postupka nećemo nulaziti, nego ćemo samo navesti rezultat:

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2}, \quad b = \overline{y} - a \cdot \overline{x},$$

pri čemu su $\overline{x}$ i $\overline{y}$ redom *aritmetička sredina empirijskih vrijednosti* varijable x, odnosno varijable y. Tako smo, dakle, dobili model jednostavne linearne regresije:

$$\hat{y} = a \cdot x + b.$$

(Oznaka $\hat{y}$ sugerira da je riječ o vrijednosti koja se izračunava na temelju jednadžbe regresijskoga modela, a ne o vrijednosti dobivenoj statističkim istraživanjem.) Navedimo ukratko nazive i značenje dobivenih parametara.

Vrijednost b naziva se naziva se <u>*konstantni član*</u>[108]. Njegova teorijska interpretacija je sljede-

---

[107] Vrijednosti dobivene statističkim istraživanjem u nastavku ćemo zvati *originalne* ili *empirijske vrijednosti*, a vrijednosti dobivene računski (pomoću regresijskoga modela) zvat ćemo *očekivane*, *procijenjene* ili *regresijske vrijednosti*.
[108] U osnovnoj i srednjoj školi tu vrijednost nerijetko (neprecizno!) nazivaju *odsječak na osi y*.

ća: b je <u>*očekivana vrijednost zavisne varijable y kada je vrijednost nezavisne varijable x jednaka nuli*</u>. To se lako vidi uvrštavanjem $x = 0$ u jednadžbu $\hat{y} = a \cdot x + b$. Ovaj parametar nema veliko praktično značenje jer nerijetko nema praktičnoga smisla (npr. u modelima koji promaraju ovisnost izdvajanja za prehranu o ukupnim mjesečnim primanjima jer je u takvim slučajevima praktično besmisleno procjenjivati iznos izdvajanja za prehranu osobe bez ikakvih ukupnih mjesečnih primanja).

Vrijednost a naziva se <u>*regresijski koeficijent*</u> i to je ujedno najvažniji parametar dobivenoga modela. Njegova teorijska interepretacija je sljedeća: vrijednost a je <u>*očekivana prosječna promjena zavisne varijable y kada se nezavisna varijabla x poveća za jednu jedinicu mjere*</u>. Strogo pozitivna vrijednost regresijskoga koeficijenta upućuje na to da jedinični porast vrijednosti nezavisne varijable X uzrokuje porast vrijednosti zavisne varijable Y (ili, kako se matematički netočno često ističe, da su varijable X i Y upravno razmjerne), a strogo negativna vrijednost regresijskoga koeficijenta upućuje na to da jedinični porast vrijednosti nezavisne varijable X uzrokuje smanjenje vrijednosti zavisne varijable Y (ili, kako se matematički netočno često ističe, da su varijable X i Y obrnuto razmjerne).

Značenje regresijskoga koeficijenta nije isključivo u promjeni vrijednosti nezavisne varijable za jednu jedinicu mjere, nego je ono puno veće. Naime, vrijede sljedeća svojstva:

<u>Svojstvo 1.</u> Ukoliko se vrijednost nezavisne varijable X *promijeni* (poveća ili smanji) za k jedinica mjere, očekivana *apsolutna* prosječna promjena vrijednosti zavisne varijable Y iznosi $k \cdot a$ (jedinica mjere).

<u>Dokaz</u>: Neka je x_1 početna vrijednost nezavisne varijable X. Tada je pripadna očekivana vrijednost zavisne varijable Y jednaka

$$\hat{y}_1 = a \cdot x_1 + b.$$

Ukoliko se vrijednost x_1 promijeni za k jedinica mjere, nova vrijednost varijable X je:

$$x_2 = x_1 + k,$$

pri čemu za $k > 0$ imamo povećanje vrijednosti nezavisne varijable X, a za $k < 0$ smanjenje vrijednosti te varijable. Pripadna vrijednost zavisne varijable Y jednaka je

$$\hat{y}_2 = a \cdot x_2 + b = a \cdot (x_1 + k) + b = a \cdot x_1 + b + a \cdot k = \hat{y}_1 + a \cdot k,$$

a odatle slijedi:

$$\hat{y}_2 - \hat{y}_1 = a \cdot k.$$

Dakle, apsolutna prosječna promjena vrijednosti zavisne varijable Y jednaka je $a \cdot k$, što je i trebalo pokazati. ■

Svojstvo 2. (česta zabluda) Ukoliko se vrijednost nezavisne varijable x *promijeni* (poveća ili smanji) za $k\%$, očekivana *relativna* prosječna promjena vrijednosti zavisne varijable y iznosi $k\%$ ako i samo ako je $b = 0$. Drugim riječima, promatrane varijable su upravno ili obrnuto razmjerne ako i samo ako je $b = 0$.

Dokaz: Neka je x_1 početna vrijednost zavisne varijable X. Tada je pripadna očekivana vrijednost zavisne varijable Y jednaka

$$\hat{y}_1 = a \cdot x_1 + b.$$

Ukoliko se vrijednost x_1 promijeni za $k\%$, pri čemu je k realan broj takav da je $k \neq 0$ (za $k = 0$ početna vrijednost varijable X se ne mijenja), nova vrijednost varijable X je:

$$x_2 = x_1 + \frac{k}{100} \cdot x_1 = (1 + \frac{k}{100}) \cdot x_1,$$

pri čemu za $k > 0$ imamo povećanje vrijednosti nezavisne varijable X, a za $k < 0$ smanjenje vrijednosti te varijable. Pripadna vrijednost zavisne varijable Y jednaka je:

$$\hat{y}_2 = a \cdot x_2 + b = a \cdot (1 + \frac{k}{100}) \cdot x_1 + b = a \cdot x_1 + b + a \cdot \frac{k}{100} \cdot x_1 = \hat{y}_1 + a \cdot \frac{k}{100} \cdot x_1.$$

Stoga je pripadna relativna prosječna promjena vrijednosti zavisne varijable Y jednaka:

$$\frac{\hat{y}_2 - \hat{y}_1}{\hat{y}_1} = \frac{\hat{y}_1 + a \cdot \frac{k}{100} \cdot x_1 - \hat{y}_1}{\hat{y}_1} = \frac{a \cdot \frac{k}{100} \cdot x_1}{\hat{y}_1} = \frac{a \cdot \frac{k}{100} \cdot x_1}{a \cdot x_1 + b},$$

Ta će vrijednost biti jednaka $k\%$ ako i samo ako vrijedi jednakost:

$$\frac{\hat{y}_2 - \hat{y}_1}{\hat{y}_1} = \frac{k}{100},$$

odnosno jednakost

$$\frac{a \cdot \frac{k}{100} \cdot x_1}{a \cdot x_1 + b} = \frac{k}{100}.$$

Prema definiciji jednakosti dvaju razlomaka, gornja je jednakost ekvivalentna jednakosti

$$a \cdot \frac{k}{100} \cdot x_1 \cdot 100 = (a \cdot x_1 + b) \cdot k \ ,$$

tj. jednakosti

$$a \cdot k \cdot x_1 = a \cdot x_1 \cdot k + b \cdot k,$$

a odavde je

$$b \cdot k = 0.$$

Prema pretpostavci je $k \neq 0$, pa dijeljenjem s k dobivamo

$$b = 0,$$

što je i trebalo pokazati. ■

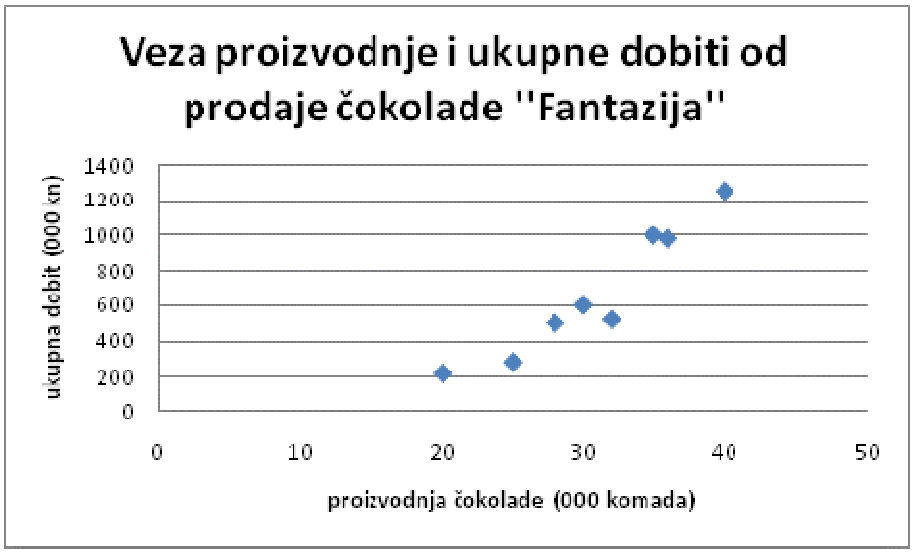
Radi potpunosti, spomenimo da se *grafički prikaz* bilo kojega modela *jednostavne* regresije naziva *regresijska krivulja*. Regresijske krivulje mogu biti pravci, grafovi eksponencijalne i logaritamske funkcije, grafovi opće potencije itd. Grafički prikaz modela jednostavne linearne regresije uvijek je *pravac* (jer je grafički prikaz bilo koje linearne funkcije pravac), i taj se pravac naziva *regresijski pravac*. Njegova jednadžba u pravokutnom koordinatnom sustavu u ravnini podudara se s jednadžbom regresijskoga modela, tj. ona glasi: $\hat{y} = a \cdot x + b$.

<u>Primjer 1.</u> Menadžment tvornice prehrambenih proizvoda "*Mljac-mljac* d.d." iz Zvečaja razmatra vezu između opsega proizvodnje čokolade "*Fantazija*" i ukupne zarade dobivene prodajom te čokolade u proteklih 8 mjeseci. Dobiveni podaci prikazani su u donjoj tablici.

proizvodnja čokolade [000 kom.]	*ukupna dobit* [000 kn]
20	220
25	280
32	520
28	500
40	1250
35	1000
36	980
30	600

Prikažimo dobivene podatke dijagramom rasipanja, izračunajmo Pearsonov koeficijent jednostavne linearne korelacije i procijenimo njegovu statističku značajnost na razini pouzdanosti od 95%. Potom procijenimo parametre modela jednostavne linearne regresije koji najbolje opisuje ovisnost ukupne dobiti o količini proizvedene čokolade, te, na temelju njega, ukupnu dobit ukoliko se proizvede 30 000 komada čokolade i ukupnu proizvodnju čokolade u mjesecu u kojemu je ostvarena dobit od 800.000,00 kn.

Budući da ćemo u drugom dijelu zadatka promatrati ovisnost ukupne dobiti o količini proizvedene čokolade, količina proizvedene čokolade bit će nam nezavisna varijabla X, a ukupna zarada dobivena prodajom čokolade zavisna varijabla Y (za korelacijsku je analizu to nebitno). Odgovarajući dijagram rasipanja je:



Izračunima analognima onima iz točke 4.1 dobiva se vrijednost Pearsonova koeficijenta jednostavne linearne korelacije:

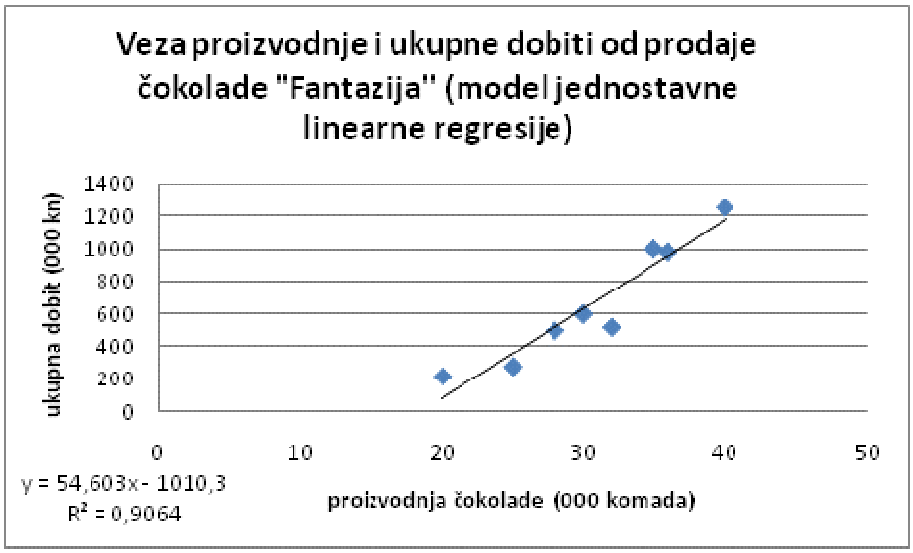
$$r = 0.952039274,$$

a tom koeficijentu pripadna p – vrijednost jednaka je

$$p = 0.000265976 < 0.05.$$

Dakle, dobiveni koeficijent je statistički značajan, tj. *postoji statistički značajna linearna veza između proizvodnje čokolade i ukupne dobiti*. Iz vrijednosti Pearsonova koeficijenta jednostavne linearne korelacije može se zaključiti da je ta veza pozitivna i jaka, tj. da će *povećanje proizvodnje čokolade* vrlo vjerojatno uzrokovati *povećanje ukupne dobiti*.

Regresijski pravac i njegova jednadžba, koja je ujedno i jednadžba odgovarajućega modela jednostavne linearne regresije, prikazani su na donjoj slici.



Osnovni parametri dobivenoga regresijskog modela jednostavne linearne regresije su:

$$a = 54.60276339,$$
$$b = -1010.284974$$

(Definiciju i značenje parametra R^2 navodimo u točki 4.2.3.) Značenje tih parametara je sljedeće:

$a = 54.60276339 \Rightarrow$ Ukoliko se proizvodnja poveća za 1000 čokolada, *očekivano prosječno povećanje* ukupne dobiti iznosi 54.602,76 kn. (Oprez s mjernim jedinicama!)
$b = -1010.284974 \Rightarrow$ Ukoliko tvornica ne proizvede *niti jednu* čokoladu, *procjenjuje* (ili *očekuje*) se *gubitak* u iznosu od 1.010.284,97 kn. (Opet pripazite na mjerne jedinice!)

Teoretski gledano, na temelju dobivenoga regresijskoga koeficijenta a možemo procijeniti prosječnu promjenu ukupne dobiti u slučaju *bilo koje* promjene vrijednosti proizvodnje. No, zbog pouzdanosti regresijskoga modela vrijednost navedene promjene ne bi trebala biti strogo veća od raspona varijacije uzorka nezavisne varijable X jer se procjenjivanjem vrijednosti izvan segmenta određenoga najmanjom i najvećom vrijednosti nezavisne varijable X značajno gubi na pouzdanosti procjene.

Npr. ukoliko se mjesečna proizvodnja poveća za 100 čokolada, očekivano prosječno povećanje ukupne dobiti iznosit će 5.460,27 kn. Smanji li se mjesečna proizvodnja (npr. zbog remonta strojeva) za 200 čokolada, očekivano prosječno smanjenje ukupne dobiti iznosit će 10.920,55 kn. Općenito, poveća (smanji) li se proizvodnja za k čokolada mjesečno, očekivano prosječno povećanje (smanjenje) ukupne dobiti iznosit će $k \cdot a = 54{,}60276 \cdot k$ kn.

Preostaje nam procijeniti ukupnu dobit ukoliko se mjesečno proizvede 30000 komada čokolade. U tu ćemo svrhu u dobivenu jednadžbu modela jednostavne linearne regresije uvrstiti $x = 30$, pa ćemo dobiti:

$$\hat{y} = 54.60276339 \cdot 30 - 1010.284974 = 627.797927$$

Dakle, očekivana ukupna dobit iznosi 627.797,93 kn.

Dobiveni regresijski model, naravno, možemo iskoristiti i u obrnutom slučaju, tj. za procjenjivanje vrijednosti nezavisne varijable x u slučajevima u kojima nam je zadana vrijednost zavisne varijable y. Izrazimo li x iz jednadžbe modela jednostavne linearne regresije, dobivamo:

$$\hat{x} = \frac{1}{a} \cdot y - \frac{b}{a}.$$

Ovdje oznaka $\hat{x}$ sugerira da vrijednost varijable x procjenjujemo, a oznaka y da je vrijednost varijable y unaprijed poznata. U konkretnom slučaju, u navedenu jednakost uvrstimo a = 54.60276339, y = 800 i b = –1010.284974, pa dobivamo da je tražena procijenjena mjesečna proizvodnja

$$\hat{x} = \frac{1}{54.60276339} \cdot 800 + \frac{-1010.284974}{54.60276339} \approx 33.1537245.$$

Dakle, *procijenjena* proizvodnja u mjesecu u kojemu je ostvarena dobit od 800.000,00 kn približno iznosi 33154 komada čokolade. (Oprez s interpretacijom: vrijednosti varijable *proizvodnja* su nužno prirodni brojevi ili nula, pa ne možemo reći npr. da procijenjena proizvodnja iznosi 33153.7 komada čokolade. Slično vrijedi i za vrijednosti varijable y koje ima smisla računati na najviše 5 decimalnih mjesta jer 6. decimala označava tisućiti dio kune koji u praksi ne postoji.)

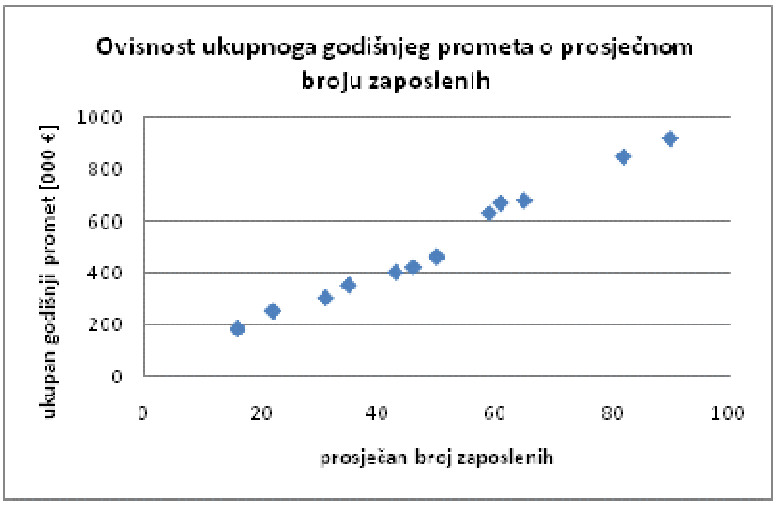
Napomenimo da se u slučaju modela jednostavne linearne regresije p – vrijednost koja odgovara nezavisnoj varijabli X i ukazuje na njezinu statističku značajnost podudara s p – vrijednosti Pearsonova koeficijenta jednostavne linearne korelacije. U ovom slučaju to znači da je nezavisna varijabla X statistički značajna, tj. da *postoji statistički značajan utjecaj proizvodnje čokolade na ukupnu dobit*. Ovaj pokazatelj ujedno je i pokazatelj statističke značajnosti cijeloga dobivenoga modela jednostavne linearne regresije, tj. dobiveni je model također statistički značajan.

Primjer 2. U sljedećoj su tablici navedeni podaci o prosječnom broju zaposlenih tijekom 2007. godine i ukupnom godišnjem prometu za svaku od ukupno 12 prodavaonica tvrtke "*Zumkon* d.o.o." iz Kukače.

prosječan broj zaposlenih	*ukupan godišnji promet* [000 €]
22	250
31	300
90	920
82	850
43	400
65	680
59	630
16	180
61	670
46	420
35	350
50	460

Odnos ukupnoga godišnjega prometa i prosječnoga broja zaposlenih prikažimo grafički odgovarajućim dijagramom (uz kojega navedimo sve potrebne oznake), pa odredimo Pearsonov koeficijent jednostavne linearne korelacije promatranih varijabli i njegovu statističku značajnost na razini pouzdanosti od 95%.
Potom procijenimo jednadžbu i parametre modela jednostavne linearne regresije koji najbolje opisuje ovisnost ukupnoga godišnjega prometa o prosječnom broju zaposlenih, te utvrdimo njihovu statističku značajnost na razini pouzdanosti od 95%.
Procijenimo i ukupan godišnji promet prodavaonice koja je u 2007. godini imala prosječno 40 zaposlenih, prosječan broj zaposlenih u prodavaonici koja je u 2007. godini ostvarila promet u iznosu od pola milijuna eura, te promjenu ukupnoga godišnjega prometa ukoliko se prosječan broj zaposlenih u jednoj prodavaonici smanji za 5.

Zbog drugoga dijela zadatka, prosječan broj zaposlenih u jednoj prodavaonici bit će nam nezavisna varijabla *X*, a ukupan godišnji promet zavisna varijabla *Y*. Odgovarajući grafički prikaz je dijagram raspršenja prikazan na donjoj slici.



Izračunima analognima onima iz točke 4.1. dobiva se vrijednost Pearsonova koeficijenta jednostavne linearne korelacije:

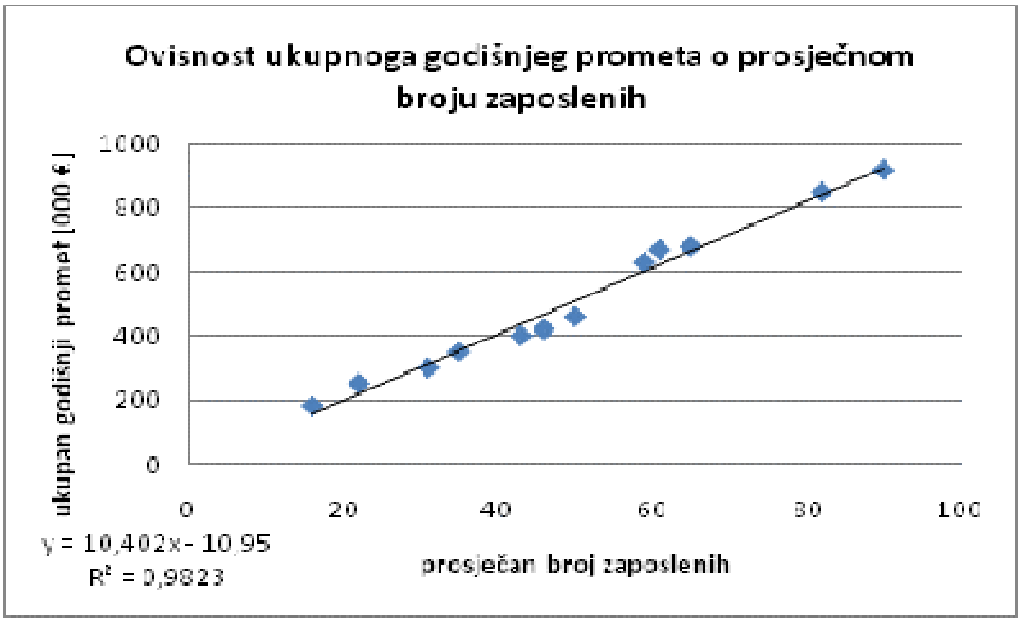
$$r = 0.991128384,$$

a tom koeficijentu odgovarajuća p – vrijednost iznosi

$$p = 4.26414E\text{-}10 = 0.0000000000426414 < 0.05.$$

Prema tome, dobiveni koeficijent statistički je značajan, tj. postoji *statistički značajna linearna veza između prosječnoga broja zaposlenih i ukupnoga godišnjega prometa*. Iz vrijednosti Pearsonova koeficijenta jednostavne linearne korelacije zaključujemo da je ta veza pozitivna i jaka, tj. da će *povećanje prosječnoga broja zaposlenih* vrlo vjerojatno uzrokovati i *povećanje ukupnoga godišnjega prometa*.

Koristeći gornji dijagram, možemo odrediti jednadžbu modela jednostavne linearne regresije koji najbolje opisuje promatranu ovisnost, te osnovne parametre toga modela. Zajedno s odgovarajućim regresijskim pravcem, ti su podaci navedeni na sljedećoj slici.



Iz dobivene jednadžbe modela jednostavne linearne regresije slijedi da su osnovni parametri dobivenoga regresijskog modela:

$$a = 10.4023396,$$
$$b = -10.95031313$$

Njihova značenja su:

a = 10.4023396 ⇒ Ukoliko se prosječan broj zaposlenih poveća za 1, *očekivano prosječno povećanje* ukupnoga godišnjega prometa iznosi 10.402,34 €. Odavde izravno slijedi da ukoliko se prosječan broj zaposlenih smanji za 5, *očekivano prosječno smanjenje* ukupnoga godišnjega prometa iznosi 5 · 10.402,34 = 52.011,70 €. Općenito, ukoliko se prosječan broj zaposlenih poveća (smanji) za k, očekivano prosječno povećanje (smanjenje) ukupnoga godišnjega prometa iznosi $k \cdot a \cdot 1.000{,}00 = k \cdot 10.4023396 \cdot 1.000{,}00 = 10.402{,}3396 \cdot k$ €.

b = –10.95031313 ⇒ U ovom slučaju vrijednost $x = 0$ znači da u prodavaonici tijekom cijele godine nije radio niti jedan zaposlenik. Takva prodavaonica prirodno nije ostvarila nikakav ukupni godišnji promet. Stoga *konstantni član* u ovom slučaju *nema svojega značenja* (ranije je istaknuto da se u praksi nerijetko javljaju ovakvi slučajevi i da je zbog toga regresijski koeficijent najvažniji parametar dobivenoga modela).

Analogno kao i u prethodnom primjeru, očekivani ukupan godišnji promet prodavaonice u kojoj je prosječan godišnji broj zaposlenih jednak 40 dobijemo tako da u jednadžbu regresijskoga modela uvrstimo $x = 40$. Dobiva se:

$$\hat{y} = 10.4023396 \cdot 40 - 10.95031313 = 405.1432707$$

Dakle, traženi očekivani promet iznosi 405.143,27 €. S druge strane, prosječan godišnji broj zaposlenih u prodavaonici čiji je ukupan godišnji promet iznosio pola milijuna eura procijenit ćemo tako da u jednadžbu

$$\hat{x} = \frac{1}{a} \cdot y - \frac{b}{a}$$

uvrstimo a = 10.4023396, y = 500 i b = —10.95031313. Dobivamo:

$$\hat{x} = \frac{1}{10.4023396} \cdot 500 + \frac{10.95031313}{10.4023396} \approx 49.118788\,.$$

Dakle, traženi procijenjeni prosječan broj zaposlenih približno iznosi 49. (Opet oprez s interpretacijom: brojimo ljude, pa dobiveni rezultat nužno mora biti prirodan broj ili nula.)

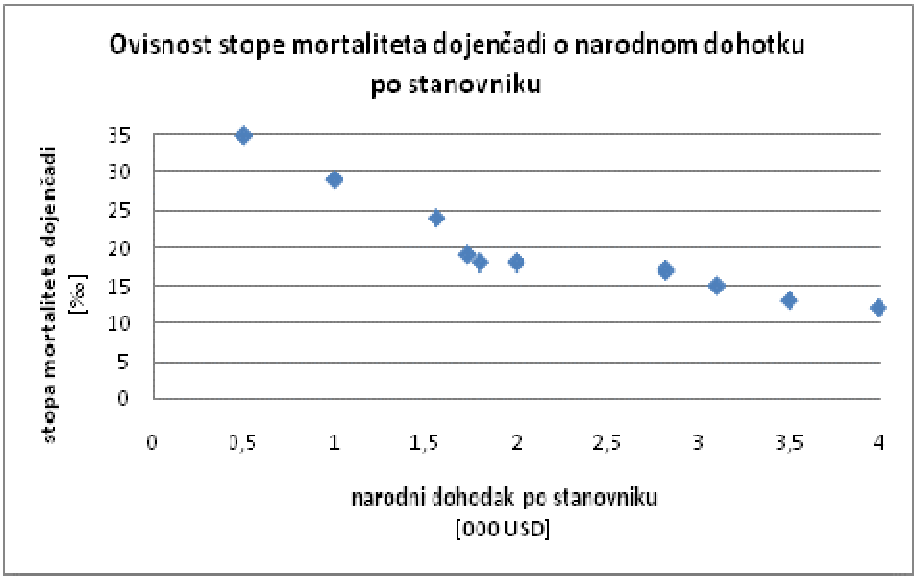
Iz p – vrijednosti izračunate za Pearsonov koeficijent jednostavne linearne korelacije slijedi da su varijabla X i dobiveni model jednostavne linearne regresije statistički značajni. Drugim riječima, *postoji statistički značajan utjecaj prosječnoga broja zaposlenih na ukupan godišnji promet* u promatranim prodavaonicama.

<u>Primjer 3.</u> Za 10 odabranih pokrajina Kraljevine Bušbumbuj utvrđeni su narodni dohodak po stanovniku (*per capita*) i stopa mortalileta dojenčadi u 2007. godini. Dobiveni su podaci prikazani u sljedećoj tablici.

narodni dohodak po stanovniku [000 USD]	*stopa mortaliteta dojenčadi* [‰]
3,99	12
3,5	13
3,1	15
2,82	17
2	18
1,8	18
1,73	19
1,56	24
1	29
0,5	35

Odnos stope mortaliteta dojenčadi i narodnoga dohotka po stanovniku prikažimo grafički odgovarajućim dijagramom (uz kojega navedimo sve potrebne oznake), pa odredimo Pearsonov koeficijent jednostavne linearne korelacije promatranih varijabli i njegovu statističku značajnost uz razinu pouzdanosti od 95%.
Potom procijenimo jednadžbu i parametre modela jednostavne linearne regresije koji najbolje opisuje ovisnost stope mortaliteta dojenčadi o narodnom dohotku po stanovniku, te statističku značajnost regresijskih varijabli i dobivenoga modela jednostavne linearne regresije uz razinu pouzdanosti od 95%.
Procijenimo i stopu mortaliteta dojenčadi pokrajine čiji je narodni dohodak po stanovniku u 2007. godini bio 2.500,00 USD, narodni dohodak po stanovniku u pokrajini koja je u 2007. godini zabilježila stopu mortaliteta dojenčadi 20‰, te promjenu stope mortaliteta dojenčadi ukoliko se narodni dohodak po stanovniku smanji za 800 USD.

Ponovo radi drugoga dijela zadatka, stopa mortaliteta dojenčadi bit će zavisna varijabla *Y*, a narodni dohodak po stanovniku nezavisna varijabla *X*. Odgovarajući grafički prikaz odnosa promatranih varijabli je dijagram rasipanja prikazan na sljedećoj slici.



Izračunima analognima onima iz točke 4.1. dobiva se vrijednost Pearsonova koeficijenta jednostavne linearne korelacije:

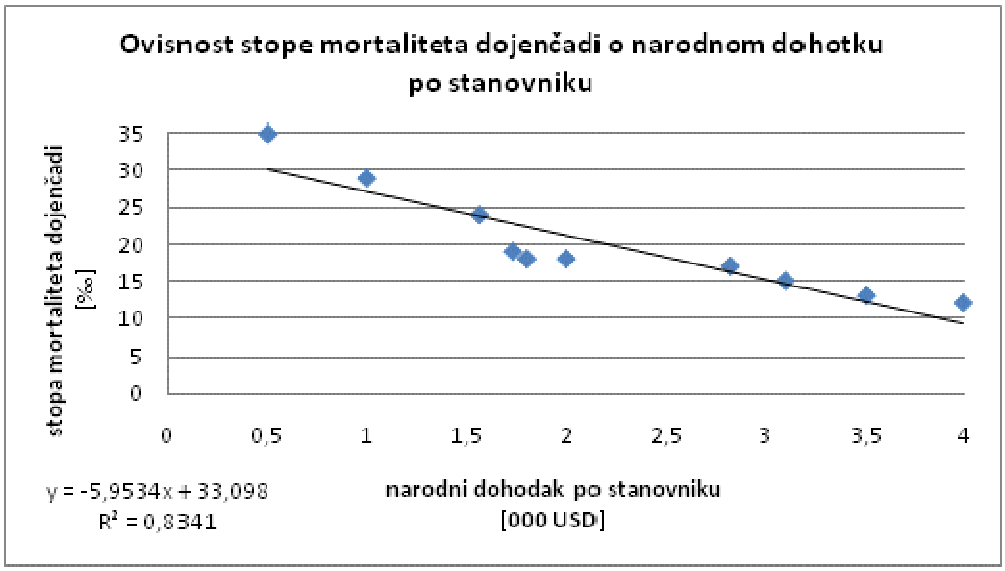
$$r = -0.913291451,$$

a tom koeficijentu pripadna p – vrijednost iznosi

$$p = 0.000222487 < 0.05,$$

pa zaključujemo da je dobiveni koeficijent statistički značajan, tj. da *postoji statistički značajna linearna veza između narodnoga dohotka i stope mortaliteta dojenčadi*. Iz vrijednosti Pearsonova koeficijenta jednostavne linearne korelacije zaključujemo da je ta veza negativna i jaka, tj. da će *povećanje narodnoga dohotka* vrlo vjerojatno uzrokovati *smanjenje stope mortaliteta dojenčadi*.

Parametri modela jednostavne linearne regresije, kao i odgovarajući regresijski pravac prikazani su na donjoj slici.



U ovome je slučaju, dakle,

$$a = -5.953418082;$$
$$b = 33.09751978.$$

Interpretacije tih parametara su:

$a = -5.953418082 \Rightarrow$ Ukoliko se narodni dohodak po stanovniku poveća za 1.000,00 USD, očekivano prosječno smanjenje stope mortaliteta dojenčadi iznosi približno 6‰. Odatle izravno slijedi da ukoliko se narodni dohodak po stanovniku smanji za 800 USD, očekivano

prosječno povećanje stope mortaliteta dojenčadi iznosi približno $0.8 \cdot 5.953418082 \approx 4.762734465‰$, tj. približno 4.76‰. Općenito, ukoliko se narodni dohodak po stanovniku poveća (smanji) za k USD, očekivano prosječno smanjenje (povećanje) stope mortaliteta dojenčadi iznosi $k \cdot |a| \cdot \frac{1}{1000} = k \cdot 5.953418082 \cdot \frac{1}{1000} = 0.005953418082 \cdot k\,‰$.

$b = 33.098 \Rightarrow$ I u ovom slučaju konstantni član nema značenja jer je praktički nemoguće da cjelokupni narodni dohodak po stanovniku (varijabla X) iznosi 0,00 USD.

Preostaje procijeniti stopu mortaliteta dojenčadi pokrajine čiji je narodni dohodak po stanovniku u 2007. godini bio 2.500,00 USD, te narodni dohodak po stanovniku u pokrajini koja je u 2007. godini zabilježila stopu mortaliteta dojenčadi 20‰. U prvom slučaju u dobivenu jednadžbu modela jednostavne linearne regresije uvrstimo $x = 2.5$, pa dobijemo:

$$\hat{y} = -5.953418082 \cdot 2.5 + 33.09751978 = 18.21397458.$$

Dakle, tražena procijenjena stopa mortaliteta iznosi približno 18,214‰. U drugom slučaju, u jednadžbu

$$\hat{x} = \frac{1}{a} \cdot y - \frac{b}{a}$$

uvrstimo $a = -5.953418082$, $y = 20$ i $b = 33.098$, pa dobivamo:

$$\hat{x} = -\frac{1}{5.953418082} \cdot 20 + \frac{33.09751978}{5.953418082} \approx 2.2$$

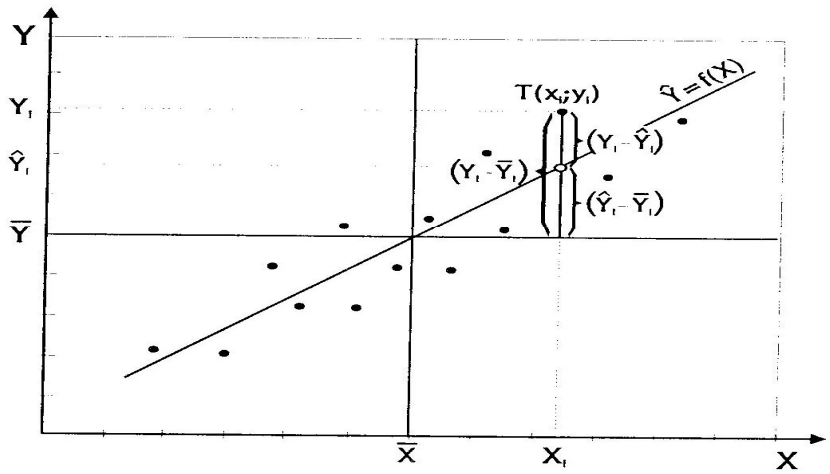
Dakle, traženi procijenjeni narodni dohodak po stanovniku iznosi približno 2.200,00 USD.

Zaključno istaknimo da iz p – vrijednosti izračunate za Pearsonov koeficijent jednostavne linearne korelacije slijedi da su i regresorska varijabla X i dobiveni model jednostavne linearne regresije statistički značajni, tj. da *postoji statistički značajan utjecaj narodnoga dohotka po stanovniku na stopu mortaliteta dojenčadi*.

4.2.2. *Analiza varijance modela jednostavne linearne regresije*

Nakon što smo odredili jednadžbu modela jednostavne linearne regresije nameće se pitanje "kvalitete" dobivenoga modela. Zapravo, promatramo u kojoj mjeri vrijednosti zavisne varijable procijenjene dobivenim modelom odgovaraju empirijskim vrijednostima, tj. s kolikom točnošću možemo *procijeniti* vrijednost zavisne varijable za određenu vrijednost nezavisne varijable. Procjena reprezentativnosti modela temelji se na već spomenutoj analizi

varijance rezidualnih odstupanja empirijskih vrijednosti (y_i) zavisne varijable od procijenjenih vrijednosti ($\hat{y}_i$) te varijable. Tu je analizu grafički zgodno prikazati donjom slikom.



Prigodom ocjenjivanja reprezentativnosti modela jednostavne linearne regresije koristi se tzv. *jednadžba analize varijance* toga modela. Ona predstavlja temelj deskriptivne i inferencijalne statističke analize reprezentativnosti regresijskoga modela, a glasi:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y})^2\,,$$

odnosno, u razvijenom obliku,

$$\sum_{i=1}^{n}y_i^2-n\cdot\overline{y}^2=\left(a\cdot\sum_{i=1}^{n}y_i+b\cdot\sum_{i=1}^{n}x_i\cdot y_i-n\cdot\overline{y}^2\right)+\left(\sum_{i=1}^{n}y_i^2-a\cdot\sum_{i=1}^{n}y_i-b\cdot\sum_{i=1}^{n}x_i\cdot y_i\right)$$

Ova jednadžba je posljedica sljedećega netrivijalnoga svojstva modela jednostavne linearne regresije:

<u>Svojstvo 1.</u> Neka su (x_i, y_i) empirijski dobivene vrijednosti nezavisne varijable X, odnosno zavisne varijable Y, za svaki i = 1, 2, …, n. Nadalje, neka su $\hat{y}_i$ procijenjene vrijednosti dobivene uvrštavanjem vrijednosti x_i u jednadžbu modela jednostavne linearne regresije koji najbolje opisuje ovisnost varijable Y o varijabli X, i to za svaki i = 1, 2, …, n. Tada je *aritmetička sredina svih procijenjenih vrijednosti jednaka aritmetičkoj sredini svih empirijskih vrijednosti*, tj. vrijedi jednakost:

$$\overline{y}=\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}\hat{y}_i\,.$$

Objasnimo pobliže svaki pojedini član u gornjoj jednadžbi:

$ST := \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2$ – ukupan zbroj kvadrata odstupanja svih empirijskih vrijednosti zavisne varijable *Y* od aritmetičke sredine tih vrijednosti[109];

$SP = \sum_{i=1}^{n}(\hat{y} - \overline{y})^2 = a \cdot \sum_{i=1}^{n} y_i + b \cdot \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{y}^2$ – ovaj se član naziva *protumačeni dio ukupnoga zbroja kvadrata odstupanja*, a zapravo je riječ o ukupnom zbroju kvadrata odstupanja svih procijenjenih vrijednosti zavisne varijable *Y* od aritmetičke sredine tih vrijednosti[110] (zbog iskazanoga svojstva 1., ta aritmetička sredina jednaka je onoj koja se pojavljuje u članu *ST*);

$SR = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n} y_i^2 - a \cdot \sum_{i=1}^{n} y_i - b \cdot \sum_{i=1}^{n} x_i \cdot y_i$ – ovaj se član naziva *neprotumačeni dio ukupnoga zbroja kvadrata odstupanja* ili *zbroj kvadrata rezidualnih odstupanja*, a zapravo je riječ o zbroju kvadrata odstupanja procijenjenih vrijednosti od odgovarajućih empirijskih vrijednosti. Slobodno govoreći, ovaj član "mjeri" ukupnu pogrešku svih *n* procjenjivanja.

Budući da je prvi član jednak zbroju preostalih dvaju, praktično se obično računaju samo *protumačeni* i *neprotumačeni* dio ukupnoga zbroja kvadrata odstupanja. Koristeći dobivene rezultate računaju se tzv. *mjere reprezentativnosti* o kojima se govori u sljedećoj točki.

4.2.3. *Mjere reprezentativnosti modela jednostavne linearne regresije*

Za ocjenu reprezentativnosti nekoga regresijskoga modela rabe se pokazatelji analogni pokazateljima reprezentativnosti aritmetičke sredine u deskriptivnoj analizi. Te pokazatelje nazivamo *mjere reprezentativnosti regresijskoga modela*, a dijelimo ih na *apsolutne* i *relativne*.

U *apsolutne mjere reprezentativnosti* ubrajamo *varijancu regresijskoga modela* i iz nje izvedenu *standardnu devijaciju regresijskoga modela*. U računanju ovih vrijednosti koriste se svi empirijski i svi procijenjeni podaci.

U *relativne mjere reprezentativnosti* ubrajamo *koeficijent varijacije regresijskoga modela* i *koeficijent determinacije regresijskoga modela*. Za usporedbu reprezentativnosti dvaju različitih regresijskih modela najčešće se rabe upravo relativne mjere reprezentativnosti.

---

[109] Podijeli li se navedeni član ukupnim brojem parova podataka (*n*), dobiva se varijanca empirijskih vrijednosti varijable *Y*.
[110] Podijeli li se navedeni član ukupnim brojem parova (*n*), dobiva se varijanca procijenjenih vrijednosti varijable *Y*.

4.2.3.1. *Varijanca regresijskoga modela*

Osnova za mjerenje reprezentativnosti regresijskoga modela je raspršenje (disperzija) oko regresijskoga pravca koja se očituje na rezidualnim odstupanjima. Što su odstupanja empirijskih vrijednosti zavisne varijable *Y* od odgovarajućih procijenjenih vrijednosti manja, reprezentativnost regresijskoga modela je bolja. Već smo istaknuli da je aritmetička sredina svih rezidualnih odstupanja uvijek jednaka nuli, pa se pri analizi raspršenja polazi od analize kvadrata rezidualnih odstupanja, odnosno od analize njihove varijance. S tim u vezi definira se pojam varijance regresijskoga modela.

<u>*Varijanca regresijskoga modela*</u> (oznaka: $\sigma^2_{\hat{y}}$) je *apsolutna* mjera reprezentativnosti regresijskoga modela. Ona se računa kao aritmetička sredina kvadrata svih rezidualnih odstupanja, pa je, sukladno oznakama iz točke 4.2.2., jednaka

$$\sigma^2_{\hat{y}} = \frac{SR}{n-2} = \frac{1}{n-2} \cdot \left(\sum_{i=1}^{n} y_i^2 - a \cdot \sum_{i=1}^{n} y_i - b \cdot \sum_{i=1}^{n} x_i \cdot y_i \right)$$

Istaknimo da se u *inferencijalnoj* statistici *procjena* ovoga pokazatelj računa kao omjer veličine *SR* i broja (n – 2) koji, kako smo već vidjeli, označava *ukupan broj stupnjeva slobode*. Takav način izračuna primijenjuju gotovo svi statistički računalni programi, pa je u takvim slučajevima nužno naglasiti da je riječ o *procjeni varijance regresije osnovnoga skupa*.

4.2.3.2. *Standardna devijacija regresijskoga modela*

<u>*Standardna devijacija regresijskoga modela*</u> (oznaka: $\sigma_{\hat{y}}$) je *apsolutna* mjera reprezentativnosti regresijskoga modela. Ona je, prema definiciji, *drugi korijen iz varijance regresijskoga modela*. Nerijetko se, analogno standardnoj devijaciji cjelokupnoga osnovnoga skupa, matematički netočno interpretira kao prosječno odstupanje originalnih od regresijskih vrijednosti. Radi potpunosti, navodimo formulu prema kojoj se računa standardna devijacija regresijskoga modela:

$$\sigma_{\hat{y}} = \sqrt{\frac{SR}{n-2}} = \sqrt{\frac{1}{n-2} \cdot \left(\sum_{i=1}^{n} y_i^2 - a \cdot \sum_{i=1}^{n} y_i - b \cdot \sum_{i=1}^{n} x_i \cdot y_i \right)}.$$

Napomenimo da se u inferencijalnoj statistici standardna devijacija *definira* na potpuno isti način, ali je, zbog različitosti izračuna varijance, pripadna formula za izravni izračun drugačija od gore navedene. U tom se slučaju govori o *procjeni standardne devijacije osnovnoga skupa*.

4.2.3.3. *Koeficijent varijacije regresijskoga modela*

<u>*Koeficijent varijacije regresijskoga modela*</u> (oznaka: $V_{\hat{y}}$) je *relativna* mjera reprezentativnosti modela i predstavlja postotni udio standardne devijacije regresijskoga modela u odnosu na aritmetičku sredinu svih empirijskih (ili, prema Svojstvu 1., procijenjenih) vrijednosti zavisne varijable:

$$V_{\hat{y}} = \frac{\hat{\sigma}_{\hat{y}}}{\overline{y}} \cdot 100$$

Ekvivalentno, možemo reći da koeficijent varijacije regresijskoga modela mjeri relativno prosječno odstupanje empirijskih vrijednosti zavisne varijable Y od procijenjenih vrijednosti te varijable i iskazuje rezultat u postotcima. Što je vrijednost koeficijenta $V_{\hat{y}}$ bliža nuli, reprezentativnost regresijskoga modela je *bolja*.

4.2.3.4 *Koeficijent determinacije regresijskoga modela*

<u>*Koeficijent determinacije*</u> (oznaka: R^2) je *relativna* mjera reprezentativnosti regresijskoga modela. Definira se kao omjer protumačenoga dijela ukupnoga zbroja kvadrata odstupanja (*SP*) i ukupnoga zbroja kvadrata odstupanja (*ST*), a može poprimiti sve vrijednosti iz segmenta [0, 1]. Objašnjava se kao postotni udio modelom protumačenoga dijela zbroja kvadrata svih odstupanja u odnosu na ukupan zbroj kvadrata svih odstupanja, a računa prema formuli:

$$R^2 = \frac{SP}{ST} = \frac{\sum_{i=1}^{n} (\hat{y} - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{a \cdot \sum_{i=1}^{n} y_i + b \cdot \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{y}^2}{\sum_{i=1}^{n} y_i^2 - n \cdot \overline{y}^2}$$

Što je njegova vrijednost bliža 1, reprezentativnost regresijskoga modela je *bolja* (donekle suprotno od koeficijenta varijacije).

Napomenimo da su *u slučaju jednostavne linearne regresije* koeficijent determinacije i Pearsonov koeficijent jednostavne linearne korelacije vezani relacijom:

$$r = \pm\sqrt{R^2} \ ,$$

pri čemu je predznak Pearsonova koeficijenta linearne korelacije jednak predznaku regresijskoga koeficijenta *a*.

Koeficijent determinacije općenito se može objasniti i kao *razmjer varijance zavisne varijable objašnjen nezavisnom varijablom*. Npr., utvrdimo li da je Pearsonov koeficijent linearne korelacije između varijabli *količnik inteligencije* i *prosječna ocjena na studiju* iznosi $r = 0.5$, onda je pripadna vrijednost koeficijenta determinacije modela jednostavne linearne regresije kojim se najbolje može opisati promatrana veza jednaka

$$R^2 = 0.5^2 = 0.25,$$

što, slobodno govoreći, znači da se približno 25% uspjeha na studiju može objasniti inteligencijom studenata.

Istaknimo da u MS EXCEL-u jednadžbu, te vrijednost koeficijenta determinacije *bilo kojega* od ponuđenih regresijskih modela možemo dobiti uključivanjima podopcija *Prikaži jednadžbu na grafikonu* i *Prikaži R-kvadratnu vrijednost na grafikonu* u podopciji *Mogućnosti* opcije *Dodaj crtu trenda*.

<u>Primjer 4.</u> Na temelju podataka iz Primjera 1. izračunajmo koeficijent varijacije i koeficijent determinacije dobivenoga modela jednostavne linearne regresije. Objasnimo značenje dobivenih vrijednosti.

U Primjeru 1. dobivena je sljedeća jednadžba modela jednostavne linearne regresije:

$$\hat{y} = 54.603 \cdot x - 1010.3.$$

Izračunajmo najprije vrijednost koeficijenta varijacije toga modela. Koristeći navedenu jednadžbu najprije za svaku empirijsku vrijednost nezavisne varijable X računamo odgovarajuću procijenjenu vrijednost zavisne varijable Y. Potom izračunamo kvadrat razlike svake pojedine empirijske vrijednosti zavisne varijable Y i njoj odgovarajuće procijenjene vrijednosti te varijable, te zbrojimo dobivene rezultate. Tako dobivamo treći i četvrti stupac sljedeće tablice:

proizvodnja čokolade [000 kom.] (x_i)	*ukupna dobit* [000 kn] (y_i)	*procijenjena ukupna dobit* [000 kn] ($\hat{y}_i$)	$(y_i - \hat{y}_i)^2$
20	220	81,76	19110.3
25	280	354,775	5591.301
32	520	736,996	47087.26
28	500	518,584	345.3651
40	1250	1.173,82	5803.392
35	1000	900,805	9839.648
36	980	955,408	604.7665
30	600	627,79	772.2841
ukupno:	5.350	5.349,938	89154.32

Tako je vrijednost koeficijenta varijacije jednaka:

$$V_{\hat{y}} = \frac{\hat{\sigma}_{\hat{y}}}{\bar{y}} \cdot 100 = \frac{\sqrt{\frac{SR}{n-2}}}{\frac{1}{n} \cdot \sum_{i=1}^{n} y_i} \cdot 100 = \frac{n \cdot \sqrt{\frac{SR}{n-2}}}{\sum_{i=1}^{n} y_i} \cdot 100 = \frac{8 \cdot \sqrt{\frac{89.154,32}{8-2}}}{5350} \cdot 100 \approx 18.23\%,$$

pa možemo reći da je varijabilitet empirijskih vrijednosti oko regresijskoga pravca relativno slab. Vrijednost koeficijenta determinacije već smo dobili u Primjeru 1. i ona iznosi R^2 = 0.9064. To znači da je modelom jednostavne linearne regresije protumačeno 90.64% ukupnih odstupanja, odnosno, da se 90.64% ovisnosti ukupne dobiti o proizvodnji čokolade može objasniti modelom jednostavne linearne regresije. Prema obama pokazateljima, možemo zaključiti da je model reprezentativan.

<u>Primjer 5.</u> Na temelju podataka iz Primjera 2. izračunajmo koeficijent varijacije i koeficijent determinacije dobivenoga modela jednostavne linearne regresije. Objasnimo značenje dobivenih vrijednosti.

U Primjeru 2. dobivena je sljedeća jednadžba modela jednostavne linearne regresije:

$$\hat{y} = 10.402 \cdot x - 10.95.$$

Kao i u Primjeru 4., izračunajmo najprije vrijednost koeficijenta varijacije toga modela. Koristeći navedenu jednadžbu najprije za svaku empirijsku vrijednost nezavisne varijable *X* računamo odgovarajuću procijenjenu vrijednost zavisne varijable *Y*. Potom izračunamo kvadrat razlike svake pojedine empirijske vrijednosti zavisne varijable *Y* i njoj odgovarajuće procijenjene vrijednosti te varijable, te zbrojimo dobivene rezultate. Tako dobivamo treći i četvrti stupac sljedeće tablice:

prosječan broj zaposlenih (x_i)	*ukupan godišnji promet* [000 €] (y_i)	*procijenjeni ukupan godišnji promet* [000 €] ($\hat{y}_i$)	$(y_i - \hat{y}_i)^2$
22	250	217,894	1030.79524
31	300	311,512	132.52614
90	920	925,23	27.3529
82	850	842,014	63.7762
43	400	436,336	1320.3049
65	680	665,18	219.6324
59	630	602,768	741.58182
16	180	155,482	601.13232
61	670	623,572	2155.55918
46	420	467,542	2260.24176
35	350	353,12	9.7344
50	460	509,15	2415.7225
ukupno:	6110	6109,8	10978.35977

Tako je vrijednost koeficijenta varijacije jednaka:

$$V_{\hat{y}} = \frac{\hat{\sigma}_{\hat{y}}}{\bar{y}} \cdot 100 = \frac{\sqrt{\frac{SR}{n-2}}}{\frac{1}{n} \cdot \sum_{i=1}^{n} y_i} \cdot 100 = \frac{n \cdot \sqrt{\frac{SR}{n-2}}}{\sum_{i=1}^{n} y_i} \cdot 100 = \frac{12 \cdot \sqrt{\frac{10978.35977}{12-2}}}{6110} \cdot 100 \approx 6.507631929\%,$$

pa možemo reći da je varijabilitet empirijskih vrijednosti oko regresijskoga pravca vrlo slab, odnosno da je model reprezentativan. Vrijednost koeficijenta determinacije već smo dobili u Primjeru 2. i ona iznosi R^2 = 0.9823. Dakle, 98,23% veze između ukupnoga godišnjega prometa i prosječnoga broja zaposlenika može se objasniti modelom jednostavne linearne regresije.

<u>Zadatak za vježbu</u>: Izračunajte koeficijent varijacije i koeficijent determinacije modela jednostavne linearne regresije iz Primjera 3. Objasnite značenje dobivenih vrijednosti.

4.3. *Primjeri nelinearnih modela jednostavne regresije*

Odnos zavisne i nezavisne varijable često nije linearan (ravnomjerne promjene nezavisne varijable ne uzrokuju ravnomjerne promjene zavisne varijable). U takvim je slučajevima potrebno je pronaći neku drugu (nelinearnu) funkciju koja najbolje pokazuje ovisnost zavisne o nezavisnoj varijabli. U nastavku se razmatraju modeli jednostruke regresije u kojima su te funkcije eksponencijalna funkcija, logaritamska funkcija i opća potencija. Ti se modeli odgovarajućim zamjenama varijabli uvijek mogu svesti na model jednostavne linearne regresije, ali to nećemo učiniti budući da MS Excel omogućuje rad i s takvim vrstama modela jednostavne regresije.

4.3.1. *Model jednostavne eksponencijalne regresije*

<u>*Standardni oblik regresijske jednadžbe*</u> modela jednostavne eksponencijalne regresije je:

$$\hat{Y} = B \cdot A^X$$

pri čemu dodatno vrijede nejednakosti:

$$A, B > 0, A \neq 1.$$

Ove nejednakosti su posljedice činjenice da eksponencijalna funkcija nije definirana za $A \leq 0$, te da za $A = 1$ dobivamo konstantnu funkciju $\hat{Y} = B$ koja je, zapravo, poseban slučaj modela jednostavne linearne regresije.

I za ovaj model navodimo interpretacije obaju njegovih parametara (A i B). Zbog nešto kompliciranije interpretacije parametra A navodimo i odgovarajući matematički izvod te interpretacije.

$B \rightarrow$ *procijenjena* (ili *očekivana*) vrijednost *zavisne* varijable Y kada je vrijednost nezavisne varijable X jednaka 0. (U istinitost ove tvrdnje lako se možete uvjeriti uvrstite li $X = 0$ u gornju jednadžbu modela jednostavne eksponencijalne regresije i primijenite činjenicu da za svaki strogo pozitivan realan broj A vrijedi jednakost $A^0 = 1$.) Kao i kod modela jednostavne linearne regresije, odmah treba napomenuti da interpretacija toga parametra nerijetko *ne odgovara* realnoj situaciji. Zbog toga ovaj parametar nije od osobite važnosti u dodatnim analizama.

Interpretacija parametra A u ovom slučaju nije izravna kao kod modela jednostavne linearne regresije, nego posredna. Naime, vrijedi sljedeće:

***Svojstvo* 1.** Ako se vrijednost nezavisne varijable X *promijeni* za k jedinica mjere, onda će se očekivana vrijednost zavisne varijable Y *promijeniti za prosječno* $100 \cdot (A^k - 1)\%$. Pritom predznak broja k označava povećanje ili smanjenje vrijednosti nezavisne varijable X, a predznak parametra $A^k - 1$ označava prosječno povećanje ili smanjenje očekivane vrijednosti zavisne varijable Y.

Dokaz: Neka je x_1 početna vrijednost nezavisne varijable x. Pripadna regresijska vrijednost zavisne varijable y je:

$$\hat{y}_1 = B \cdot A^{x_1}.$$

Odmah primijetimo da je, zbog uvjeta $A, B > 0$, $A \neq 1$, svaka očekivana vrijednost zavisne varijable y strogo veća od nule (što slijedi i iz činjenice da je područje vrijednosti bilo koje eksponencijalne funkcije skup $\mathbf{R}^+ = \langle 0,+\infty \rangle$), pa posebno i različita od nule.

Ako se vrijednost x_1 promijeni za k jedinica mjere, nova vrijednost te varijable je:

$$x_2 = x_1 + k.$$

(Pritom za $k > 0$ imamo povećanje vrijednosti varijable x_1, a za $k < 0$ smanjenje njezine vrijednosti.) Pripadna regresijska vrijednost zavisne varijable y je

$$\hat{y}_2 = B \cdot A^{x_2} = B \cdot A^{x_1+k} = B \cdot A^{x_1} \cdot A^k = A^k \cdot (B \cdot A^{x_1}) = A^k \cdot \hat{y}_1.$$

Stoga je relativna promjena vrijednosti zavisne varijable y jednaka:

$$\Delta \hat{y} = \frac{\hat{y}_2 - \hat{y}_1}{\hat{y}_1} = \frac{A^k \cdot \hat{y}_1 - \hat{y}_1}{\hat{y}_1} = \frac{\hat{y}_1 \cdot (A^k - 1)}{\hat{y}_1} = A^k - 1,$$

odnosno, u postotcima,

$$\Delta \hat{y} = \left[100 \cdot (A^k - 1)\right]\%,$$

što je i valjalo pokazati. ■

Dakle, kod modela jednostavne eksponencijalne regresije dodatno se računa i interpretira broj $S_k = A^k - 1$. Posebno, za $k = 1$ vrijednost $S_1 = 100 \cdot (A - 1)$ uobičajeno se naziva *prosječna stopa promjene* (očekivanih) vrijednosti zavisne varijable *Y*. Ako model jednostavne eksponencijalne regresije reprezentativno opisuje vezu između zavisne i nezavisne varijable, približno jednaka stopa promjene vrijedi i za empirijske vrijednosti zavisne varijable *Y*.

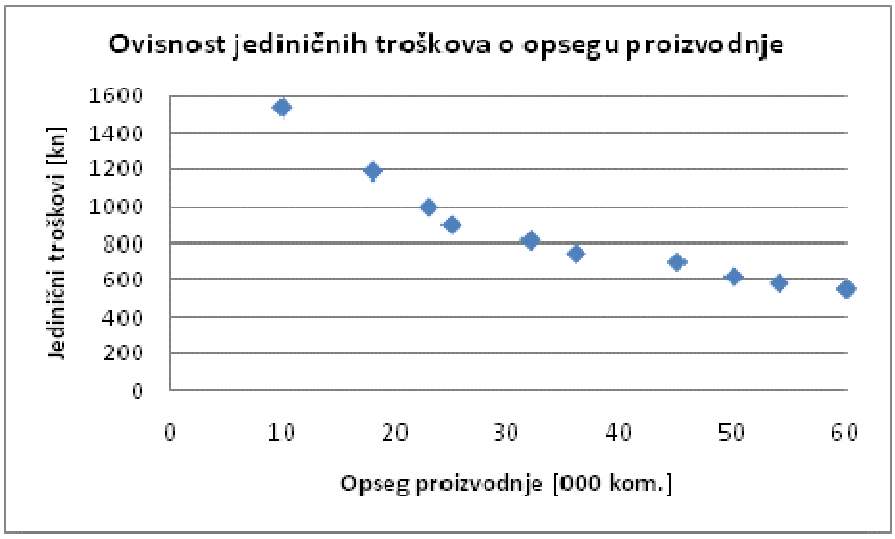
I za model jednostavne eksponencijalne regresije mogu se računati koeficijent varijacije i koeficijent determinacije tako da se u formule navedene u točkama 4.2.3.3. i 4.2.3.4. umjesto originalnih vrijednosti uvrštavaju logaritmi tih vrijednosti. Koeficijent determinacije može se dobiti izravno pomoću MS Excela, dok koeficijent varijacije moramo izračunati "klasično". Ilustrirajmo to na primjerima.

Primjer 1. Uprava tvornice zamjenskih dijelova za željeznička vozila "*Grudelj* d.d." iz Bibinja razmatra ovisnost ukupnih jediničnih troškova o opsegu proizvodnje jednoga od zamjenskih dijelova. Istraživanjem dobiveni podaci navedeni su u sljedećoj tablici:

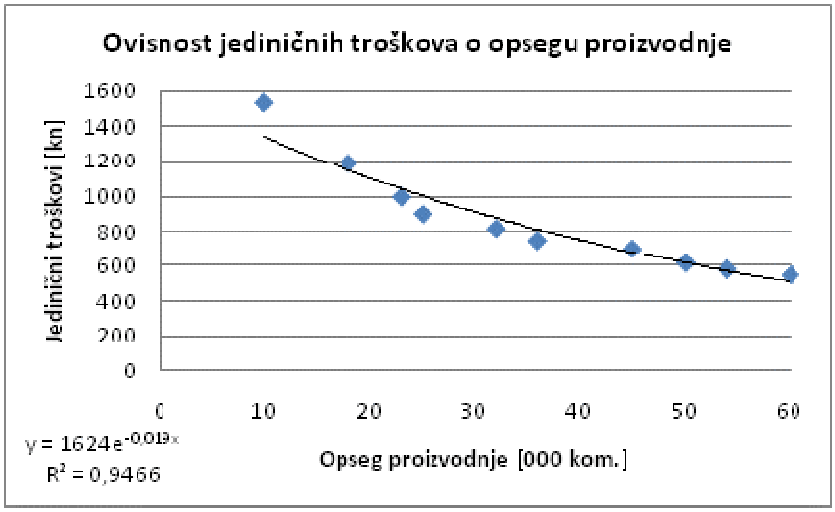
jedinični troškovi [kn]	*opseg proizvodnje* [000 kom.]
549	60
582	54
618	50
700	45
748	36
817	32
900	25
998	23
1.193	18
1.540	10

Prikažimo dobivene podatke odgovarajućim grafikonom (uz kojega navedimo sve potrebne oznake), pa odredimo jednadžbu modela jednostavne eksponencijalne regresije koji najbolje opisuje promatranu ovisnost i interpretirajmo parametre toga modela. Naposljetku, procijenimo reprezentativnost dobivenoga modela pomoću koeficijenta determinacije, ukupne jedinične troškove pri opsegu proizvodnje od 40 000 komada i opseg proizvodnje u slučaju kad ukupni jedinični troškovi iznose 800,00 kn

Odgovarajući grafički prikaz empirijskih podataka je dijagram raspršenja prikazan na sljedećoj slici:



Odgovarajuća jednadžba modela jednostavne eksponencijalne regresije, zajedno s pripadnom krivuljom, dana je na sljedećoj slici:



Zapišimo je najprije u standardnom obliku $\hat{Y} = B \cdot A^X$. U ovome je slučaju $B = 1624$, dok je vrijednost parametra A jednaka

$$A = e^{-0,019} \approx 0.98118.$$

pa jednadžba modela jednostavne eksponencijalne regresije zapisana u standardnom obliku glasi:

$$\hat{Y} = 1624 \cdot 0.98118^X.$$

Odgovarajuće interpretacije parametara su:

$B = 1624 \Leftrightarrow$ očekivana vrijednost jediničnih troškova u slučaju kad nema proizvodnje iznosi 1.624,00 kn. (Ova interpretacija nema smisla jer se jedinični troškovi uvijek računaju u odnosu na jedinicu proizvedenoga proizvoda, a ako nije proizveden niti jedan proizvod, nema smisla govoriti o jediničnim troškovima.)

$A = 0.98118 \Leftrightarrow$ ovu ćemo vrijednost najlakše interpretirati računajući prosječnu stopu promjene:

$$S_1 = 100 \cdot (A - 1) = 100 \cdot (0.98118 - 1) = -1.882.$$

Dakle, *ukoliko se opseg proizvodnje poveća za* 1000 komada, *jedinični troškovi će se očekivano smanjiti za prosječno* 1.882%. Odatle slijedi npr. da ukoliko se opseg proizvodnje poveća za 100 komada, očekivano prosječno smanjenje jediničnih troškova iznosi

$$S_{0.1} = 100 \cdot (A^{0.1} - 1) = 100 \cdot (0.98118^{0.1} - 1) = 0.18982\%,$$

a ukoliko se opseg proizvodnje smanji za 200 komada, očekivano prosječno povećanje jediničnih troškova iznosi

$$S_{-0.2} = 100 \cdot (A^{-0.2} - 1) = 100 \cdot (0.98118^{-0.2} - 1) = 0.380723\% \text{ itd.}$$

Koeficijent determinacije dobivenoga modela je $R^2 = 0.9466$, što znači da se 94.66% *ovisnosti jediničnih troškova o opsegu proizvodnje može objasniti modelom jednostavne eksponencijalne regresije*. Budući da je navedeni koeficijent relativno blizu 1, možemo reći da je dobiveni model reprezentativan.

Preostaje nam procijeniti po jednu vrijednost zavisne, odnosno nezavisne varijable. Procijenjeni iznos ukupnih jediničnih troškova u slučaju kad opseg proizvodnje iznosi 40 000 komada dobit ćemo tako da u jednadžbu regresijskoga modela uvrstimo $x = 40$:

$$\hat{Y} = 1624 \cdot 0.98118^{40} = 759.51.$$

Dakle, procijenjeni ukupni jedinični troškovi u slučaju kad opseg proizvodnje iznosi 40 000 komada iznose 759,51 kn.

Nadalje, iz dobivene jednadžbe regresijskoga modela slijedi

$$\hat{X} = \frac{\ln y - \ln 1624}{\ln 0.98118},$$

pa procijenjenu vrijednost opsega proizvodnje u slučaju kad ukupni jedinični troškovi iznose 800,00 kn dobivamo tako da u navedenu jednadžbu uvrstimo $y = 800$:

$$\hat{X} = \frac{\ln 800 - \ln 1624}{\ln 0.98118} \approx 37.265 .$$

Dakle, procijenjena vrijednost opsega proizvodnje u slučaju kad ukupni jedinični troškovi iznose 800,00 kn jednaka je 37265 komada.

Primjer 2. Promatramo ovisnost temperature netom skuhane kave o vremenu proteklom nakon ulijevanja kave u šalicu[111]. Podaci su dani u donjoj tablici.

vrijeme [min.]	*temperatura* [°C]
0	82
5	76
8	70
11	65
15	61
18	57
22	52
25	51
30	47
34	45
38	43
42	41
45	39
50	38

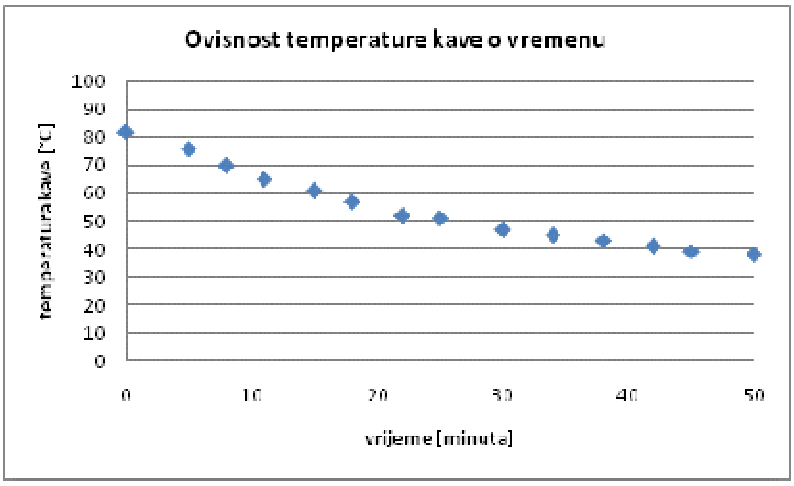
a) Ovisnost temperature kave o vremenu grafički prikažimo odgovarajućim grafikonom (uz kojega navedimo sve potrebne oznake).
b) Odredimo jednadžbu modela jednostavne eksponencijalne regresije koji najbolje opisuje promatranu ovisnost i objasnimo značenje parametara dobivenoga modela.
c) Na temelju koeficijenta determinacije procijenimo reprezentativnost modela.
d) Na temelju rezultata **b)** podzadatka procijenimo početnu temperaturu kave i odredimo relativnu pogrešku u odnosu na empirijski dobivenu vrijednost.
e) Na temelju rezultata **b)** podzadatka procijenimo kolika će biti temperatura kave točno 1 sat nakon ulijevanja u šalicu.
f) 1992. jedna od gošći restorana "*McDonald's*" tužila je restoran jer je slučajnim prolijevanjem kave temperature 82°C po svojim rukama zaradila teške opekotine. Vještak na suđenju tada je posvjedočio da tekućina temperature 82°C koja je u dodiru s ljudskom kožom bila barem pet sekundi izaziva teške opekotine i ustvrdio da bi se takve posljedice izbjegle da se kava poslužuje pri temperaturi ne većoj od 68°C. Žena je dobila spor i odštetu od

[111] Preciznije rečeno, nakon ulijevanja kave iz *džezve* ili *kavokuhala* u šalicu.

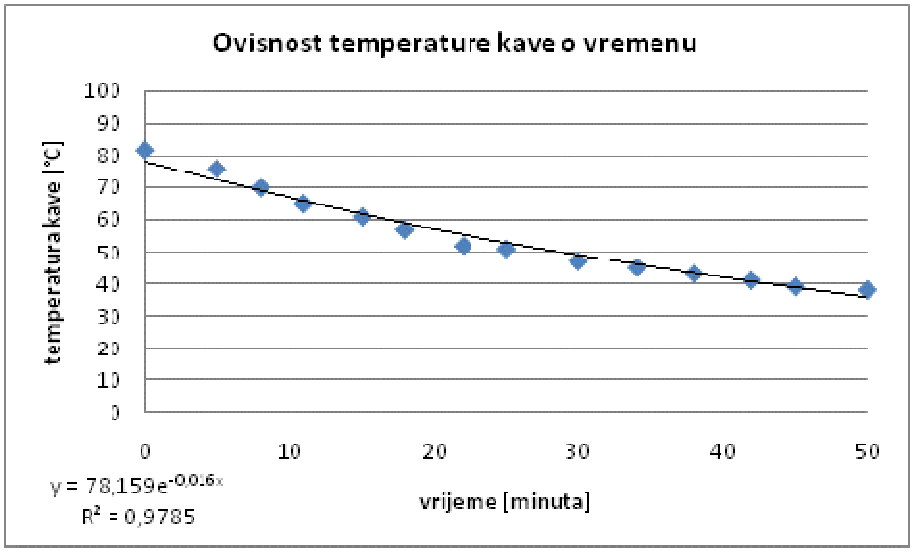
2.700.000,00 USD. Jedna od posljedica toga poznatoga sudskoga procesa bila je da većina restorana u SAD danas poslužuje kavu čija je temperatura približno 68°C. Na temelju rezultata **b)** podzadatka procijenimo koliko dugo treba čekati da se netom skuhana kava ohladi na temperaturu od 68°C.

g) Pretpostavimo da temperatura sobe u kojoj se nalazi poslužena kava iznosi 23°C. Što se može reći o temperaturi kave nakon točno 2 sata na temelju dobivenoga modela jednostavne eksponencijalne regresije? Objasnimo svoj odgovor.

Ovisnost temperature kave o vremenu prikazana je donjim dijagramom rasipanja:



Jednadžba modela jednostavne eksponencijalne regresije, zajedno s pripadnom krivuljom, prikazana je na donjoj slici:



Dobivenu jednadžbu najprije zapišimo u standardnom obliku $\hat{Y} = B \cdot A^X$. U ovome je slučaju $B = 78.159$, dok je vrijednost parametra A jednaka

$$A = e^{-0.016} \approx 0.98413.$$

Dakle, jednadžba dobivenoga modela jednostavne eksponencijalne regresije je

$$\hat{Y} = 78.159 \cdot 0.98413^X.$$

Parametre dobivenoga modela interpretiramo na sljedeći način:

$B = 78.159 \Leftrightarrow$ *procijenjena* početna temperatura kave iznosi 78.159°C. Budući da empirijska početna temperatura kave iznosi 82°C, relativna pogreška procjene početne temperature je:

$$p = \frac{82 - 78.159}{82} \cdot 100 \approx 4.68\%.$$

$A = 0.98413 \Leftrightarrow$ izračunat ćemo očekivanu prosječnu stopu promjene temperature kave:

$$S_1 = 100 \cdot (A - 1) = -1.5873.$$

Dakle, *temperatura kave će se očekivano smanjiti za* 1.5873% *u jednoj minuti*.

Temperaturu kave nakon jednoga sata (= 60 minuta) procijenit ćemo tako da u jednadžbu modela jednostavne eksponencijalne regresije uvrstimo $x = 60$:

$$\hat{Y} = 78.159 \cdot 0.98413^{60} \approx 30°\text{C},$$

dok ćemo vrijeme potrebno da se kava ohladi na temperaturu 68°C procijeniti tako da iz jednadžbe

$$\hat{Y} = 78{,}159 \cdot 0.98413^X$$

izrazimo nepoznaznicu X:

$$\hat{X} = \frac{\ln Y - \ln 78.159}{\ln 0.98413}$$

i u dobiveni izraz uvrstimo $Y = 68$:

$$\hat{X} = \frac{\ln 68 - \ln 78.159}{\ln 0.98413} \approx 8{,}7 \text{ minuta} = 8 \text{ minuta i } 42 \text{ sekunde.}$$

Temperatura kave bit će jednaka sobnoj temperaturi, odnosno 23°C, nakon

$$\hat{X} = \frac{\ln 23 - \ln 78.159}{\ln 0.98413} \approx 76{,}5 \text{ minuta} = 76 \text{ minuta i } 30 \text{ sekundi.}$$

Potom se kava prestaje hladiti i zadržava temperaturu okoline (sobe) od 23°C. Stoga će stvarna temperatura kave za 120 minuta biti jednaka 23°C. Prognostička temperatura kave za 120 minuta je dvostruko manja:

$$\hat{Y} = 78.159 \cdot 0.98413^{120} \approx 11.46°\text{C}.$$

Na ovom smo primjeru ukazali na tzv. *problem ekstrapolacije* u regresijskim modelima. Grubo govoreći, regresijski model može davati dobre procjene za vrijednosti nezavisne varijable X koje se nalaze unutar segmenta omeđenoga najmanjom i najvećom empirijskom vrijednošću te varijable. (U promatranom primjeru riječ je o segmentu [0, 50].) Međutim, za vrijednosti izvan toga segmenta kvaliteta procjene opada što je vrijednost udaljenija od donje ili gornje granice segmenta. Npr. u promatranom bismo primjeru za $x = 51$ još mogli očekivati relativno dobru procjenu temperature kave, ali $x = 120$ je vrijednost više nego dvostruko veća od gornje granice segmenta [0, 50], pa se za tu vrijednost dobiva loša procjena odgovarajuće vrijednosti zavisne varijable Y. Stoga, kad god je to moguće, treba izbjegavati procjenu vrijednosti zavisne varijable za vrijednosti nezavisne varijable izvan segmenta omeđenoga najmanjom i najvećom empirijski dobivenom vrijednošću.[112]

4.3.2. *Model jednostavne logaritamske regresije*

<u>*Standardni oblik regresijske jednadžbe*</u> ovoga modela je:

$$\hat{Y} = A \cdot \ln x + B,$$

pri čemu vrijednosti nezavisne varijable X nužno moraju biti strogo pozitivne jer je prirodno područje definicije *bilo koje* logaritamske[113] funkcije skup $\langle 0, +\infty \rangle$. Kao i kod svih prethodnih

---

[112] U promatranom je primjeru opisani problem ekstrapolacije dodatno izražen zbog fizikalne neusklađenosti dobivenoga modela s realnim stanjem: prema fizikalnim zakonima, kava se ne može ohladiti na temperaturu manju od temperature okoline, a prema dobivenom regresijskom modelu, za velike vrijednosti nezavisne varijable x temperatura kave bit će sve bliža 0°C zbog asimptotskoga približavanja odgovarajućega grafa osi Ox.

[113] U ovome je slučaju riječ o tzv. *prirodnom logaritmu*: baza logaritma je broj $e \approx 2{,}718281828459045235360 3$.

modela, značaj parametara A i B u analizi regresijskoga modela nije jednak: parametar A je bitno značajniji od parametra B. U nastavku navodimo njihove interpretacije u obliku dvaju svojstava:

$B \rightarrow$ *procijenjena* (ili *očekivana*) *vrijednost zavisne* varijable Y kada je vrijednost nezavisne varijable X jednaka 1 jedinici mjere. Ova se tvrdnja lako provjeri uvrštavanjem $x = 1$ u jednadžbu regresijskoga modela i korištenjem činjenice da je $\ln 1 = 0$.

Interpretacija varijable A, kao i kod modela jednostavne eksponencijalne regresije, nije izravna, nego "zaobilazna" (neizravna) i temelji se na sljedećem svojstvu:

Svojstvo 1. Neka je $p > -100$ proizvoljan, ali fiksiran realan broj. Ako se vrijednost *nezavisne* varijable X *promijeni* za $p\%$, *procijenjena* (ili *očekivana*) *prosječna promjena vrijednosti zavisne varijable* Y iznosi $A \cdot \ln\left(1+\frac{p}{100}\right)$. Posebno, za $p = 171.82818284590452353603$ procijenjena prosječna promjena vrijednosti zavisne varijable Y iznosi točno A.

Uvjet $p > -100$ postavljen je zbog zahtjeva da vrijednost nezavisne varijable X mora biti strogo pozitivna: smanjenje konkretne vrijednosti nezavisne varijable X za barem 100% povlačilo bi da je nova vrijednost te varijable nepozitivna (strogo negativna ili nula), a takve vrijednosti ne mogu biti obuhvaćene regresijskim modelom.

Dokaz: Analogno kao i kod modela jednostavne eksponencijalne regresije, neka je x_1 početna vrijednost varijable X i neka je p postotak promjene te varijable. Ako je $p > 0$, riječ je o povećanju vrijednosti x_1 za $p\%$, a ako je $p < 0$, riječ je o smanjenju vrijednosti x_1 za $p\%$. Pripadna očekivana vrijednost zavisne varijable Y jednaka je:

$$\hat{y}_1 = A \cdot \ln x_1 + B .$$

Nakon promjene vrijednosti x_1 za $p\%$, nova vrijednost te varijable je:

$$x_2 = x_1 + \frac{p}{100} \cdot x_1 = \left(1+\frac{p}{100}\right) \cdot x_1 .$$

Pripadna očekivana vrijednost zavisne varijable Y jednaka je

$$\hat{y}_2 = A \cdot \ln x_2 + B = A \cdot \ln\left[\left(1+\frac{p}{100}\right) \cdot x_1\right] + B = A \cdot \ln\left(1+\frac{p}{100}\right) + A \cdot \ln x_1 + B = A \cdot \ln\left(1+\frac{p}{100}\right) + \hat{y}_1$$

otkuda slijedi da je ukupna prosječna promjena vrijednosti zavisne varijable Y jednaka

$$\hat{y}_2 - \hat{y}_1 = A \cdot \ln\left(1+\frac{p}{100}\right),$$

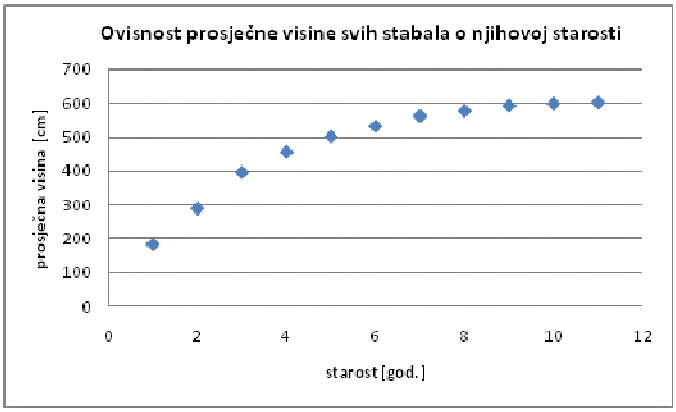
što je i trebalo pokazati. Napomenimo da se pridjev *prosječna* navodi zbog izvoda formule za procijenjenu vrijednost parametra A u kojemu se koriste prosjeci vrijednosti nezavisne i zavisne varijable, pa je i sâm parametar A pokazatelj prosječne promjene. ■

Primjer 1. Prije točno 10 godina istodobno je posađeno ukupno 12 jednogodišnjih sadnica određene vrste višnje. U trenutku sadnje prosječna visina jedne sadnice bila je 182.88 cm. Podaci o prosječnoj visini stabala tijekom svake od minulih 10 godina navedeni su u donjoj tablici. Promatra se ovisnost prosječne visine izraslih stabala o njihovoj starosti.

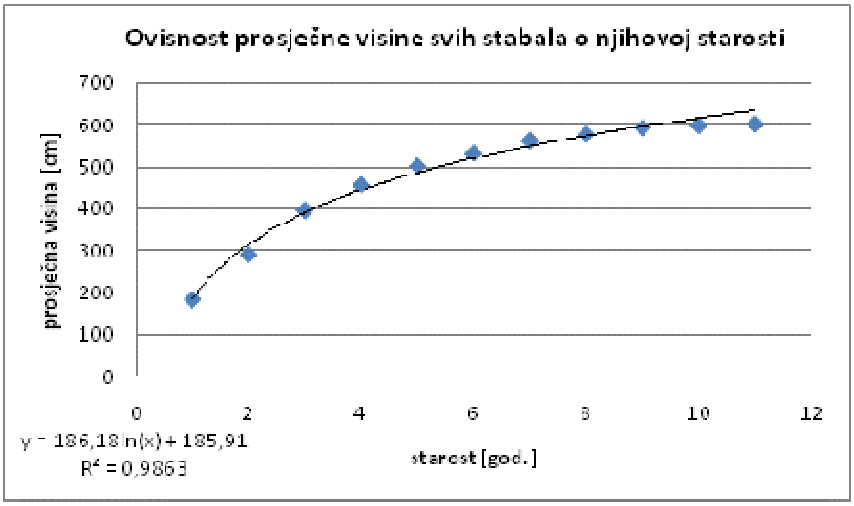
starost [god.]	*prosječna visina* [cm]
1	182.88
2	289.56
3	396.24
4	457.2
5	502.92
6	533.4
7	563.88
8	579.12
9	594.36
10	600.456
11	603.504

a) Prikažimo ovisnost prosječne visine stabala o njihovoj starosti odgovarajućim grafikonom uz kojega navedimo sve potrebne oznake.
b) Odredimo jednadžbu modela jednostavne logaritamske regresije koji najbolje opisuje promatranu ovisnost. Objasnimo značenje svih parametara dobivenoga modela.
c) Procijenimo reprezentativnost modela dobivenoga u rješenju **b)** podzadatka pomoću koeficijenta determinacije.
d) Koristeći rezultat **b)** podzadatka procijenimo prosječnu visinu svih stabala četiri i pol godine nakon sadnje.
e) Koristeći rezultat **b)** podzadatka procijenimo trenutak (računajući od trenutka sadnje) u kojemu će prosječna visina svih stabala biti 3 metra.
f) Koristeći rezultat **b)** podzadatka procijenimo prosječnu visinu svih posađenih stabala točno 20 godina nakon sadnje i utvrdimo realističnost dobivene procjene.

Ovisnost prosječne visine svih stabala o njihovoj starosti grafički prikazujemo donjim dijagramom rasipanja:



Jednadžba modela jednostavne logaritamske regresije, zajedno s pripadnom krivuljom, navedena je na donjoj slici:



Najvažniji parametri dobivenoga modela jednostavne logaritamske regresije su:

A = 186.18 ⇔ Ovaj ćemo parametar interpretirati pomoću p = 10. Dakle, *ako se starost svih sadnica poveća za* 10%, *procijenjeno* (*očekivano*) *povećanje njihove prosječne visine iznosi* $186.18 \cdot \ln\left(1+\frac{10}{100}\right) \approx 17.74485$ cm.

B = 185.91 ⇔ *Procijenjena* (*očekivana*) *prosječna visina svih jednogodišnjih sadnica u trenutku sadnje iznosi* 185.91 cm.

Koeficijent determinacije dobivenoga modela iznosi R^2 = 0.9863, otkuda slijedi da je 98.63% *ovisnosti prosječne visine stabala o njihovoj starosti objašnjeno modelom jednostavne logaritamske regresije*, što znači da je model reprezentativan.

Četiri i pol godine nakon sadnje sva stabla bit će stara ukupno 1 + 4,5 = 5.5 godina, pa će njihova procijenjena prosječna visina u tom trenutku biti

$$\hat{Y} = 186.18 \cdot \ln 5.5 + 185.91 \approx 503.3 \text{ cm.}$$

Trenutak u kojemu će prosječna visina svih stabala biti 3 metra (= 300 cm) odredit ćemo tako da iz dobivene jednadžbe modela jednostavne logaritamske regresije

$$\hat{Y} = 186.18 \cdot \ln x + 185.91$$

izrazimo x:

$$\hat{X} = e^{\frac{y-185.91}{186.18}}$$

i u dobivenu formulu uvrstimo $y = 300$. Dobivamo:

$$\hat{X} = e^{\frac{300-185.91}{186.18}} \approx 1.8455 \text{ godina} = 1 \text{ godina } 10 \text{ mjeseci i } 5 \text{ dana.}$$

Dakle, računajući od trenutka sadnje, procijenjeno vrijeme u kojemu će prosječna visina svih stabala biti točno 3 metra iznosi 10 mjeseci i 5 dana.

Točno 20 godina nakon sadnje procijenjena prosječna visina svih stabala (čija će starost tada biti 1 + 20 = 21 godinu) iznosi

$$\hat{Y} = 186.18 \cdot \ln 21 + 185.91 \approx 752.74 \text{ cm.}$$

Vrijednost $x = 21$ je gotovo dvostruko veća od gornje granice segmenta [1, 11] u kojemu smo radili procjene parametara, pa ova procjena ne mora biti realistična. Ipak, u praksi se pokazalo da stabla višnje prosječno narastu u rasponu od 4.5 do čak 9 metara. Stoga procijenjenu vrijednost ne treba unaprijed odbaciti kao potpuno pogrešnu, tim više što logaritamska funkcija za sve veće i veće vrijednosti nezavisne varijable x raste sve "sporije".

4.3.3. *Model jednostavne potencijske regresije* (*POWER*)[114]

<u>*Standardni oblik jednadžbe modela potencijske regresije*</u> je:

$$\hat{Y} = B \cdot X^A .$$

Podsjetimo da se, za proizvoljan, ali fiksiran realan broj $a \in \mathbf{R}$, funkcija $f(x) = x^a$ naziva *opća potencija*. Njezino je prirodno područje definicije skup $\langle 0, +\infty \rangle$, pa – analogno kao i u modelu jednostavne logaritamske regresije – vrijednosti nezavisne varijable X ne mogu biti nepozitivni realni brojevi (strogo negativni realni brojevi i nula).

Za analizu regresijskoga modela ponovno je značajniji parametar A, ali – kao u svim prethodnim modelima – navodimo interpretacije obaju parametara modela:

$B \rightarrow$ *procijenjena* (*očekivana*) *vrijednost zavisne* varijable Y kada je vrijednost *nezavisne* varijable X jednaka 1 jedinici mjere. Ova je interpretacija jednaka interpretaciji istoga regresijskoga parametra u modelu jednostavne logaritamske regresije, a lako se provjeri tako da se u jednadžbu modela jednostavne potencijske regresije uvrsti $x = 1$ i primijeni činjenica da za svaki realan broj $a \in \mathbf{R}$ vrijedi jednakost: $1^a = 1$.

---

[114] Engleska riječ *power* ima više značenja, a jedno od njih je i *potencija*. Stoga je naziv ovoga modela preveden kao *potencijski*. Uobičajeni hrvatski prijevod naziva modela (*dvostruko logaritamski*) potječe od činjenice da se u tzv. *linearizaciji* modela (svođenju modela odgovarajućim zamjenama na model jednostavne linearne regresije) vrijednosti nezavisne i zavisne varijable pojavljuju kao logaritmandi.

Interpretacija parametra A se temelji na sljedećem svojstvu:

***Svojstvo* 1.** Neka je $p > -100$ proizvoljan, ali fiksiran realan broj. Ako se vrijednost nezavisne varijable X promijeni za $p\%$, *procijenjena* (*očekivana*) *promjena vrijednosti zavisne varijable* Y iznosi $\left\{100 \cdot \left[\left(1+\frac{p}{100}\right)^A - 1\right]\right\}\%$.

Dokaz: Analogno ranijim dokazima, neka je x_1 početna vrijednost nezavisne varijable X. Pripadna očekivana vrijednost zavisne varijable Y jednaka je:

$$\hat{y}_1 = B \cdot x_1^A .$$

Ukoliko se vrijednost x_1 promijeni za $p\%$, nova vrijednost nezavisne varijable bit će:

$$x_2 = x_1 + \frac{p}{100} \cdot x_1 = \left(1+\frac{p}{100}\right) \cdot x_1 .$$

Pripadna vrijednost zavisne varijable Y jednaka je

$$\hat{y}_2 = B \cdot x_2^A = B \cdot \left[\left(1+\frac{p}{100}\right) \cdot x_1\right]^A = B \cdot \left(1+\frac{p}{100}\right)^A \cdot x_1^A = \left(1+\frac{p}{100}\right)^A \cdot \hat{y}_1 ,$$

Već smo utvrdili da je područje vrijednosti svake opće potencije skup svih strogo pozitivnih realnih brojeva, što znači da su obje vrijednosti $\hat{y}_1$ i $\hat{y}_2$ strogo veće od nule, a samim tim i različite od nule. Zbog toga je dobro definirana relativna promjena vrijednosti zavisne varijable Y i ona je jednaka

$$s = \frac{\hat{y}_2 - \hat{y}_1}{\hat{y}_1} = \frac{\left(1+\frac{p}{100}\right)^A \cdot \hat{y}_1 - \hat{y}_1}{\hat{y}_1} = \frac{\hat{y}_1 \cdot \left[\left(1+\frac{p}{100}\right)^A - 1\right]}{\hat{y}_1} = \left(1+\frac{p}{100}\right)^A - 1 ,$$

odnosno, iskazano u postotcima,

$$s = \left\{100 \cdot \left[\left(1+\frac{p}{100}\right)^A - 1\right]\right\}\% ,$$

što je i trebalo pokazati. ■

Koristeći razvoj u binomni red

$$\left(1+\frac{p}{100}\right)^A = 1+\frac{p}{100}\cdot A+\frac{A\cdot(A-1)}{2}\cdot\frac{p^2}{10000}+\frac{A\cdot(A-1)\cdot(A-2)}{6}\cdot\frac{p^3}{1000000}+\ldots,$$

pokazuje se da za relativno male vrijednosti broja p (npr. $p \in [-1, 1]$) i $A > 0$ vrijedi aproksimacija:

$$\left(1+\frac{p}{100}\right)^A \approx 1+\frac{p}{100}\cdot A$$

Odavde slijedi sljedeće

Svojstvo 2. Ako se vrijednost nezavisne varijable X promijeni za p%, pri čemu je p relativno mali realan broj i $A > 0$, *procijenjena* (*očekivana*) *promjena vrijednosti zavisne varijable* Y iznosi $(p \cdot A)$%. Posebno, za $p = \pm 1$ ta promjena iznosi $\pm A$%.

Dakle, pri interpretaciji parametra A obično ćemo uzimati $p = 1$. Istaknimo da se u ekonomskim analizama ovim regresijskim modelom omogućava tzv. *mjerenje elastičnosti* zavisne varijable u odnosu na nezavisnu varijablu.

Ilustrirajmo primjenu modela jednostavne potencijske regresije na dvama primjerima.

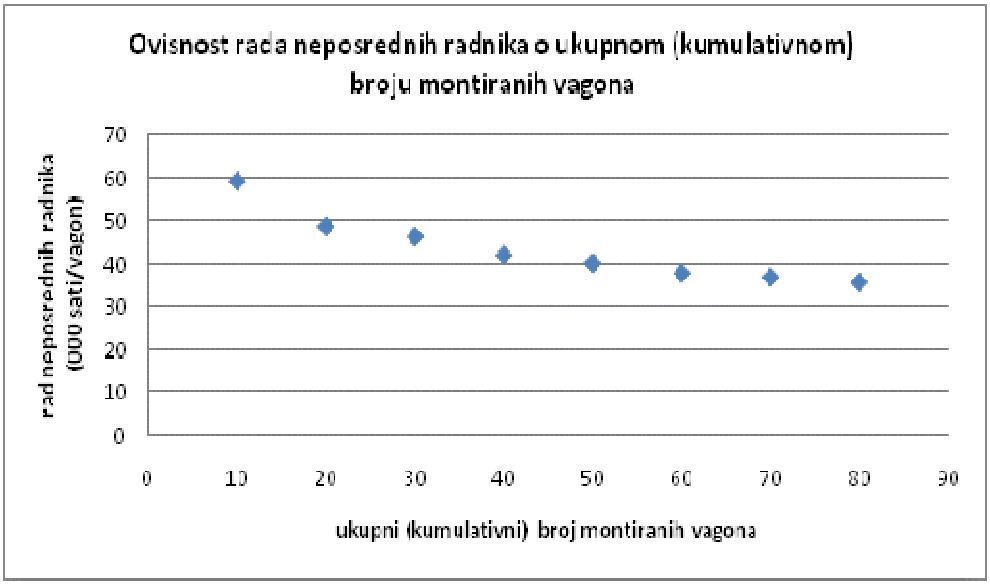
Primjer 1. Tvornica željezničkih vagona "*Šlamperajević*" iz Piškorevaca razmatra ovisnost utrošenih radnih sati neposrednih radnika na montaži vagona o ukupnom (kumulativnom) broju montiranih vagona. Dobiveni podaci navedeni su u sljedećoj tablici.

ukupan (*kumulativni*) *broj montiranih vagona*	*rad neposrednih radnika* [000 *sati/vagon*]
10	59.2
20	48.7
30	46.3
40	42.1
50	40.2
60	37.9
70	36.9
80	35.7

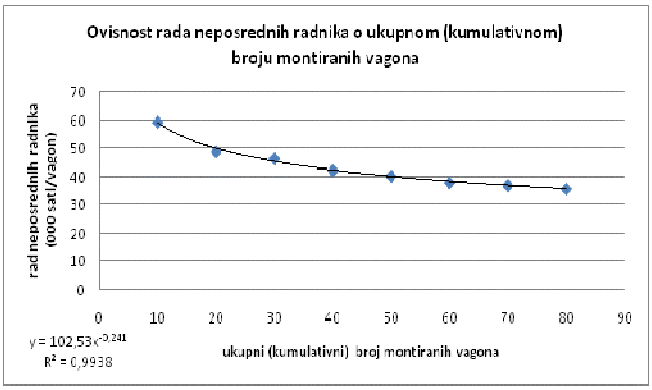
a) Promatranu ovisnost grafički prikažimo odgovarajućim grafikonom uz kojega navedimo sve potrebne oznake.
b) Odredimo jednadžbu modela jednostavne potencijske regresije koji najbolje opisuje promatranu ovisnost i objasnimo značenje parametara dobivenoga modela.

c) Procijenimo reprezentativnost dobivenoga modela na temelju koeficijenta determinacije.
d) Na temelju rezultata **b)** podzadatka procijenimo ukupno vrijeme rada neposrednih radnika po jednom vagonu ako je montirano ukupno 45 vagona.
e) Na temelju rezultata **b)** podzadatka procijenimo ukupan (kumulativni) broj montiranih vagona ako utrošeni radni sati neposrednih radnika iznose 41 000 radnih sati po vagonu.

Ovisnost utrošenih radnih sati neposrednih radnika na montaži o ukupnom (kumulativnom) broju montiranih vagona grafički prikazujemo donjim dijagramom rasipanja:



Jednadžba modela jednostavne potencijske regresije, zajedno s pripadnom krivuljom, navedena je na sljedećoj slici:



Najvažniji parametri dobivenoga modela su:

B = 102.53 ⇔ *Procijenjeni rad neposrednih radnika na montaži prvoga vagona iznosi* 102 530 *radnih sati*.

$A = -0.241 \Leftrightarrow$ (uzimamo $p = 1$) *Ako se ukupan* (*kumulativni*) *broj montiranih vagona poveća za* 1%, *utrošeni radni sati neposrednih radnika po jednom vagonu očekivano će se smanjiti za* 0.241%.

Koeficijent determinacije dobivenoga regresijskoga modela je $R^2 = 0.9938$, što znači da se 99.38% *ovisnosti rada neposrednih radnika o ukupnom* (*kumulativnom*) *broju montiranih vagona može objasniti modelom jednostavne potencijske regresije*, tj. dobiveni je model vrlo reprezentativan.

Ukupno vrijeme rada neposrednih radnika na montaži ukupno 45 vagona procijenit ćemo tako da u jednadžbu dobivenoga regresijskoga modela

$$\hat{Y} = 102.53 \cdot x^{-0.241}$$

uvrstimo $x = 45$:

$$\hat{Y} = 102.53 \cdot 45^{-0.241} = 40.9663$$

Dakle, traženo procijenjeno vrijeme montaže po jednom vagonu iznosi 40 966.3 sati.

Ukupan (kumulativni) broj montiranih vagona u slučaju kad se za montažu jednoga vagona utroši 41 000 radnih sati dobit ćemo tako da iz jednadžbe

$$\hat{Y} = 102.53 \cdot x^{-0.241}$$

izrazimo x:

$$\hat{X} = \left(\frac{102.52}{y}\right)^{0.241}$$

i u dobivenu jednadžbu uvrstimo $y = 41$:

$$\hat{X} = \left(\frac{102.52}{41}\right)^{\frac{1000}{241}} \approx 45.$$

Dakle, procijenjeni ukupan (kumulativni) broj montiranih vagona jednak je 45.

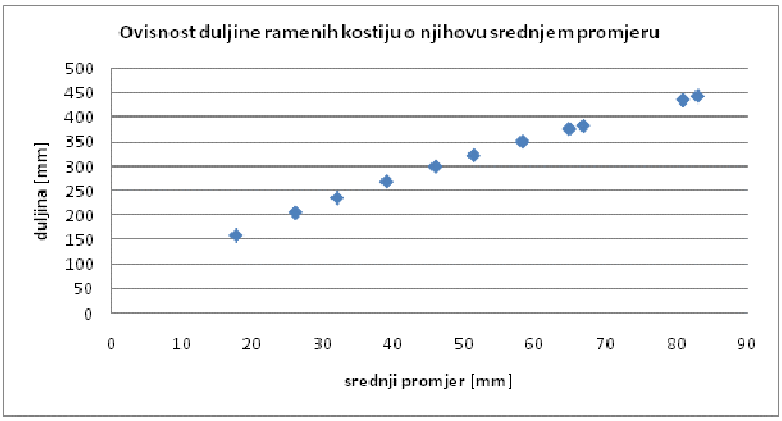
Napomenimo da realna funkcija dobivena kao jednadžba modela jednostavne potencijske regresije pripada tipu tzv. *funkcijâ učenja* (engl. *learning function*), a temelji se na pretpostavci da se utrošak radnih sati smanjuje s ponavljanjem iste radne operacije. Stoga se modeli jednostavne potencijske regresije često primijenjuju pri ugovaranju dinamike izvođenja radova ili isporuke serije određenih proizvoda (npr. željezničkih vagona, aviona, brodova i sl.).

Primjer 2. Na uzorku od ukupno 11 afričkih antilopa (gazela) ispituje se ovisnost između duljine i srednjega promjera tzv. ramenih kostiju. Dobiveni podaci prikazani su u donjoj tablici.

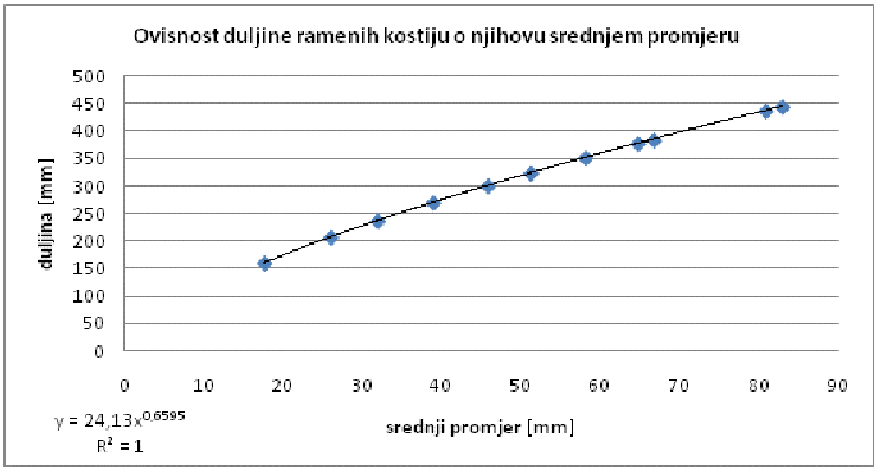
srednji promjer [mm]	*duljina* [mm]
17.6	159.9
26	206.9
31.9	236.8
38.9	269.9
45.8	300.6
51.2	323.6
58.1	351.7
64.7	377.6
66.7	384.1
80.8	437.2
82.9	444.7

a) Ovisnost duljine ramenih kosti o njihovu srednjem promjeru grafički prikažimo odgovarajućim grafikonom uz kojega navedimo sve potrebne oznake.
b) Odredimo jednadžbu modela jednostavne potencijske regresije koji najbolje opisuje promatranu ovisnost. Navedimo sve parametre toga modela i objasnimo njihovo značenje.
c) Procijenimo reprezentativnost dobivenoga regresijskoga modela na temelju koeficijenta determinacije.
d) Na temelju rezultata **b)** podzadatka procijenimo duljinu ramenih kostiju čiji je srednji promjer 84 mm.
e) Na temelju rezultata **b)** podzadatka procijenimo srednji promjer ramenih kostiju čija je duljina 40 cm.

Ovisnost duljine ramenih kostiju o njihovu srednjem promjeru grafički je prikazana donjim dijagramom rasipanja:



Jednadžba modela jednostavne potencijske regresije, zajedno s pripadnom krivuljom, navedena je na donjoj slici:



Najvažniji parametri dobivenoga modela jednostavne potencijske regresije su:

B = 24.13 ⇔ *Procijenjena duljina ramenih kostiju čiji je srednji promjer* 1 mm *iznosi* 24.13 mm. (Interpretacija ovoga parametra ne odgovara realnoj situaciji jer su druga istraživanja pokazala da je najmanja zabilježena duljina ramenih kostiju približno 140 mm = 14 cm.)

A = 0.6595 ⇔ (uzimamo p = 1) *Ako se srednji promjer ramenih kostiju poveća za* 1%, *očekivano relativno prosječno povećanje duljine ramenih kostiju iznosi* 0.6595%.
Koeficijent determinacije dobivenoga modela jednostavne potencijske regresije iznosi $R^2 = 1$, što ukazuje na činjenicu da se dobivenim regresijskim modelom potpuno može objasniti ovisnost duljine ramenih kostiju antilope o njihovu srednjem promjeru. Taj je rezultat posljedica zaokruživanja izračuna koeficijenta determinacije na 4 decimalna mjesta. Izračuna li se R^2 na 10 decimalnih mjesta, dobiva se $R^2 = 0.9999874242$. U svakom slučaju, riječ je o vrlo reprezentativnom modelu jednostavne regresije.

Procijenjenu duljinu ramenih kostiju čiji je srednji promjer 84 mm dobit ćemo tako da u jednadžbu dobivenoga modela jednostavne potencijske regresije

$$\hat{Y} = 24.13 \cdot x^{0.6595}$$

uvrstimo $x = 84$:

$$\hat{Y} = 24.13 \cdot 84^{0.6595} \approx 448.4 \text{ mm.}$$

Stoga tražena procijenjena duljina ramenih kostiju iznosi 448.4 mm.

Srednji promjer ramenih kostiju čija je duljina 40 cm (= 400 mm) procijenit ćemo tako da iz dobivene jednadžbe modela jednostavne potencijske regresije izrazimo x:

$$\hat{X} = \left(\frac{y}{24.13}\right)^{\frac{2000}{1319}}$$

pa u dobivenu jednakost uvrstimo $y = 400$:

$$\hat{X} = \left(\frac{400}{24.13}\right)^{\frac{2000}{1319}} \approx 70.7.$$

Stoga traženi procijenjeni srednji promjer ramenih kostiju iznosi 70.7 mm.

afrička antilopa (gazela)

4.4. *Primjer modela višestruke (multiple) regresije: model višestruke linearne regresije*

U ovoj se točki ukratko opisuje najjednostavniji primjer modela višestruke ili multiple regresije, a to je model višestruke linearne regresije. Podsjetimo da općenito modeli višestruke regresije imaju jednu zavisnu i barem dvije nezavisne varijable. U slučajevima u kojima se pojavljuju najmanje tri nezavisne varijable nemoguće je dati grafički prikaz. No, i u takvim se slučajevima mogu računati odgovarajući analogoni koeficijenata linearne korelacije, koeficijenata determinacije itd., te dati interpretacije parametara dobivenoga modela višestruke regresije. Rečeno ćemo pokazati na primjeru modela višestruke linearne regresije.

Opći oblik modela višestruke linearne regresije je:

$$Y = A_1 \cdot X_1 + A_2 \cdot X_2 + \ldots + A_n \cdot X_n + B + e,$$

a *odgovarajuća jednadžba* toga modela je

$$\hat{Y} = A_1 \cdot x_1 + A_2 \cdot x_2 + \ldots + A_n \cdot x_n + B.$$

U objema jednakostima su, analogno kao i u modelu jednostavne linearne regresije, A_1, A_2, …, A_n *regresijski koeficijenti* modela, B *slobodni član*, a e stohastička (slučajna) varijabla koja "mjeri" pogrešku procjene. Za svaki $i = 1, 2, \ldots, n$ interpretacija regresijskoga koeficijenta A_i glasi:

A_i je procijenjena (očekivana) prosječna promjena vrijednosti varijable X_i u slučaju kada se vrijednost te varijable poveća za 1 jedinicu mjere, a vrijednosti svih ostalih varijabli ostanu nepromijenjene.

Navedena se interpretacija može poopćiti za bilo koji realan broj $k \in \mathbf{R}$:

$k \cdot A_i$ je procijenjena (očekivana) prosječna promjena vrijednosti varijable X_i u slučaju kad se vrijednost te varijable promijeni za k jedinica mjere, a vrijednosti svih ostalih varijabli ostanu nepromijenjene.

Slobodni član B je manje važan parametar modela, ali radi potpunosti navodimo i njegovu interpretaciju:

B je procijenjena (očekivana) vrijednost zavisne varijable Y u slučaju kad su vrijednosti svih nezavisnih varijabli X_1, …, X_n jednake nuli. (Ova interpretacija, naročito u ekonomiji, vrlo često ne odgovara realnoj situaciji.)

Primjer 1. (izbori za kanadski parlament) Ispituje se ovisnost ukupnoga postotka dobivenih glasova (a time i mandata) relativnoga pobjednika na svakim pojedinim parlamentarnim izborima o postotcima glasova dobivenih u četiri kanadske regije: R_1, R_2, R_3 i R_4. Dobiveni podaci su prikazani u sljedećoj tablici.

izborna godina	*postotak na državnoj razini*	*postotak u regiji* R_1	*postotak u regiji* R_2	*postotak u regiji* R_3	*postotak u regiji* R_4
1949.	73	74	78	58	61
1953.	64	34	78	35	38
1957.	42	64	52	29	32
1958.	79	76	73	95	82
1962.	44	55	49	88	27
1963.	49	61	62	6	32
1965.	50	45	69	2	32
1968.	59	22	74	25	70
1972.	41	31	56	7	17
1974.	54	41	71	11	35
1979.	48	56	58	78	68
1980.	52	63	74	4	0
1984.	75	78	74	80	68
1988.	57	34	63	67	38
1993.	60	97	67	39	19
1997.	52	34	72	17	18
2000.	57,5	60	77	17	15
2004.	43,8	69	53	11	22

a) Odredimo jednadžbu modela višestruke linearne regresije i objasnimo značenje parametara dobivenoga modela.
b) Procijenimo reprezentativnost dobivenoga modela na temelju koeficijenta determinacije, te statističku značajnost pojedinih regresorskih varijabli i cijeloga modela (uz razinu pouzdanosti od 95%).
c) Relevantne predizborne ankete (provedene od strane neovisnih agencija) pokazuju da će stranka S u regiji R_1 osvojiti 51% glasova, u regiji R_2 67% glasova, u regiji R_3 12% glasova i u regiji R_4 15% glasova. Na temelju rezultata **a)** podzadatka utvrdimo može li ta stranka računati na relativnu izbornu pobjedu.
d) Na temelju rezultata **a)** podzadatka procijenimo najmanji postotak glasova koji stranka S treba dobiti u njoj tradicionalno sklonoj regiji R_2 tako da bude relativni pobjednik izbora ako se na temelju predizbornih anketa procjenjuje da će ta stranka u regiji R_1 osvojiti 55% glasova, u regiji R_3 38% glasova, a u regiji R_4 41% glasova.

Za dobivanje i analizu modela višestruke linearne regresije u MS Excelu koristimo proceduru *Regression* u okviru alata *Analiza podataka* (*Data Analysis*). Primjenom te procedure dobivamo sljedeće vrijednosti:

koeficijent višestruke linearne korelacije (*Multiple R*): 0.96556752, otkuda slijedi da postoji jaka linearna veza između postotka na državnoj razini, kao zavisne varijable, i postotaka na regionalnoj razini, kao nezavisnih varijabli;
koeficijent determinacije: 0.932320637, otkuda slijedi da je dobiveni model višestruke linearne regresije reprezentativan (objašnjava približno 93.25% ovisnosti zavisne varijable o svim četirima nezavisnim varijablama);
p – vrijednost koeficijenta višestruke linearne korelacije (*p – vrijednost modela višestruke linearne regresije*): 1,76513E-07 = 0.000000176512 ⇔ *dobiveni koeficijent višestruke linearne korelacije* i *dobiveni model višestruke linearne regresije su statistički značajni*;
slobodni član: -17.3211733 (teoretski, to je ukupan postotak mandata na državnoj razini u slučaju da pobjednička stranka niti u jednoj regiji ne osvoji niti jedan glas, ali takva interpretacija ne odgovara realnoj situaciji)
koeficijent uz varijablu Postotak u regiji R_1: 0.129717918 ⇔ *Ako se postotak dobivenih glasova u regiji* R_1 *poveća za* 1, *očekivano prosječno povećanje postotka dobivenih glasova na državnoj razini iznosi* (*približno*) 0.13;
p –vrijednost koja odgovara koeficijentu uz varijablu Postotak u regiji R_1: 0.011263866 ⇔ *uz razinu pozdanosti od* 95%, *postoji statistički značajan utjecaj postotka glasova dobivenih u regiji* R_1 *na ukupan postotak dobivenih glasova na državnoj razini*;
koeficijent uz varijablu Postotak u regiji R_2: 0.857302 ⇔ *Ako se postotak dobivenih glasova u regiji* R_2 *poveća za* 1, *očekivano prosječno povećanje postotka dobivenih glasova na državnoj razini iznosi* (*približno*) 0.86;
p –vrijednost koja odgovara koeficijentu uz varijablu Postotak u regiji R_2: 3.96069E-07 = 0,000000396069 ⇔ *uz razinu pozdanosti od* 95%, *postoji statistički značajan utjecaj postotka glasova dobivenih u regiji* R_2 *na ukupan postotak dobivenih glasova na državnoj razini*;
koeficijent uz varijablu Postotak u regiji R_3: 0.117393926 ⇔ *Ako se postotak dobivenih glasova u regiji* R_3 *poveća za* 1, *očekivano prosječno povećanje postotka dobivenih glasova na državnoj razini iznosi* (*približno*) 0.12;
p –vrijednost koja odgovara koeficijentu uz varijablu Postotak u regiji R_3: 0.00900155 ⇔ *uz razinu pozdanosti od* 95%, *postoji statistički značajan utjecaj postotka glasova dobivenih u regiji* R_3 *na ukupan postotak dobivenih glasova na državnoj razini*;
koeficijent uz varijablu Postotak u regiji R_4: 0.112526063 ⇔ *Ako se postotak dobivenih glasova u regiji* R_4 *poveća za* 1, *očekivano prosječno povećanje postotka dobivenih glasova na državnoj razini iznosi* (*približno*) 0.11;
p –vrijednost koja odgovara koeficijentu uz varijablu Postotak u regiji R_4: 0.049511477 ⇔ *uz razinu pozdanosti od* 95%, *postoji statistički značajan utjecaj postotka glasova dobivenih u regiji* R_4 *na ukupan postotak dobivenih glasova na državnoj razini*;

Primijetimo da prema strožem kriteriju (uz razinu pouzdanosti od 99%) utjecaj postotka glasova dobivenih u regijama R_1 i R_4 ne bi bio statistički značajan

Za svaki i = 1, 2, 3, 4 označimo s x_i postotak glasova koje je pobjednička stranka dobila u regiji R_i. Tada jednadžba modela višestruke linearne regresije koji opisuje promatranu ovisnost glasi:

$$\hat{Y} = 0.129717918 \cdot x_1 + 0.857302 \cdot x_2 + 0.117393926 \cdot x_3 + 0.112526063 \cdot x_4 - 17.3211733.$$

Procjenu može li stranka koja u regiji R_1 osvoji 51% glasova, u regiji R_2 67% glasova, u regiji R_3 12% glasova i u regiji R_4 15% glasova računati na relativnu izbornu pobjedu napravit ćemo tako da u gornju jednadžbu uvrstimo $x_1 = 67$, $x_2 = 52$, $x_3 = 12$ i $x_4 = 15$:

$$\hat{Y} = 0.129717918 \cdot 51 + 0.857302 \cdot 67 + 0.117393926 \cdot 12 + 0.112526063 \cdot 15 - 17.3211733,$$

$$\hat{Y} = 49.83029292.$$

Budući da procijenjeni rezultat nije strogo veći od 50 [%], stranka ne bi trebala računati na relativnu pobjedu na parlamentarnim izborima.
Najmanji postotak glasova koji stranka S treba dobiti u njoj tradicionalno sklonoj regiji R_2 tako da bude relativni pobjednik izbora u drugom podslučaju procijenit ćemo tako da riješimo nejednadžbu

$$\hat{Y} \geq 50,$$

tj. tako da po x_2 riješimo nejednadžbu

$$0.129717918 \cdot 55 + 0.857302 \cdot x_2 + 0.117393926 \cdot 38 + 0.112526063 \cdot 41 - 17.3211733 \geq 50$$

Odavde je

$$x_2 \geq 59.62,$$

pa se može zaključiti da bi stranka S u njoj tradicionalno sklonoj regiji R_2 trebala dobiti najmanje 60% ukupnih glasova birača u toj regiji tako da može računati na relativnu pobjedu na parlamentarnim izborima.

5. INDIVIDUALNI BROJČANI POKAZATELJI RAZVOJA VREMENSKOGA NIZA

Vrijednosti neke pojave u vremenu obično određujemo i uspoređujemo u određenim vremenskim razdobljima. Vremensko razdoblje u kojemu smo posljednji put odredili vrijednost neke pojave naziva se *tekuće* ili *izvještajno razdoblje*, a sama određena vrijednost *tekuća vrijednost* te pojave. Sva razdoblja prije tekućega u kojima smo određivali vrijednost promatrane pojave nazivaju se *prethodna razdoblja*. Od svih njih najznačajnije je *neposredno prethodno razdoblje*: to je, zapravo, vremensko razdoblje u kojemu smo pretposljednji put odredili vrijednost promatrane pojave.

Primjer 1. Promatramo ukupan broj sklopljenih i razvedenih brakova u Republici Hrvatskoj u razdoblju od 2002. do 2006. godine.

	brakovi	
godina	*sklopljeni*	*razvedeni*
2002.	22 806	4 496
2003.	22 337	4 934
2004.	22 700	4 985
2005.	22 138	4 883
2006.	22 092	2 651

izvor: Mjesečno statističko izvješće 3/08, Državni zavod za statistiku, Zagreb, 2008.

U ovome je primjeru tekuće ili izvještajno razdoblje 2006. godina jer je, prema navedenoj tablici, u toj godini posljednji put izvršeno brojanje sklopljenih, odnosno razvedenih brakova. Prethodna razdoblja su 2002., 2003., 2004. i 2005. godina, a neposredno prethodno razdoblje je 2005. godina.

Primjer 2. Promatramo kretanje prosječne mjesečne isplaćene neto-plaće po jednom zaposleniku u svim pravnim osobama u Republici Hrvatskoj koje se bave financijskim posredovanjem za razdoblje od siječnja 2007. godine do siječnja 2008. godine.

godina	*mjesec*	*prosječna neto-plaća* [kn]
2007.	siječanj	6.460,00
	veljača	6.604,00
	rujan	6.567,00
	listopad	6.569,00
	studeni	8.228,00
	prosinac	6.709,00
2008.	siječanj	7.152,00

izvor: Mjesečno statističko izvješće 3/08, Državni zavod za statistiku, Zagreb, 2008.

U ovome je primjeru tekuće ili izvještajno razdoblje siječanj 2008. godine. Prethodna razdoblja su siječanj, veljača, rujan, listopad, studeni i prosinac 2007. godine. Neposredno prethodno razdoblje je prosinac 2007. godine. Uočimo da je razdoblje neposredno prethodno mjesecu rujnu 2007. godine mjesec veljača 2007. godine, a ne mjesec kolovoz 2007. godine. No, to *ne znači* da se u razdoblju od ožujka do kolovoza 2007. godine nije pratilo kretanje promatrane prosječne neto-plaće, nego isključivo da ti podaci nisu navedeni u tablici.

Nerijetko se vrijednosti promatrane pojave iz različitih razloga[115] uspoređuju s jednom te istom vrijednošću pojave određenom u nekomu razdoblju. Ta, jedna te ista vrijednost naziva se *bazna vrijednost*, a razdoblje u kojemu je određena *bazno razdoblje*.

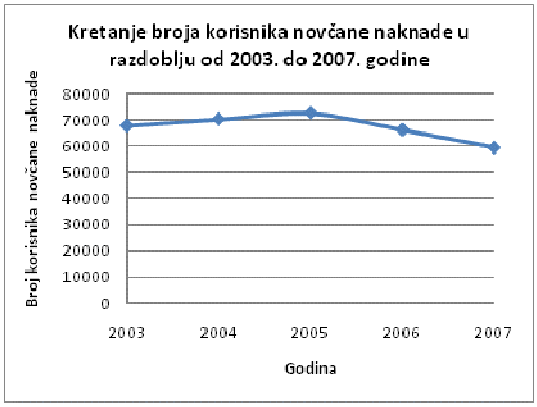
Primjer 3. Promatramo prosječan godišnji broj korisnika novčane naknade na Hrvatskom zavodu za zapošljavanje u razdoblju od 2003. do 2007. godine:

godina	*prosječan godišnji broj korisnika novčane naknade*
2003.	67 977
2004.	70 467
2005.	72 802
2006.	66 407
2007.	59 603

izvor: Mjesečno statističko izvješće 3/08, Državni zavod za statistiku, Zagreb, 2008.

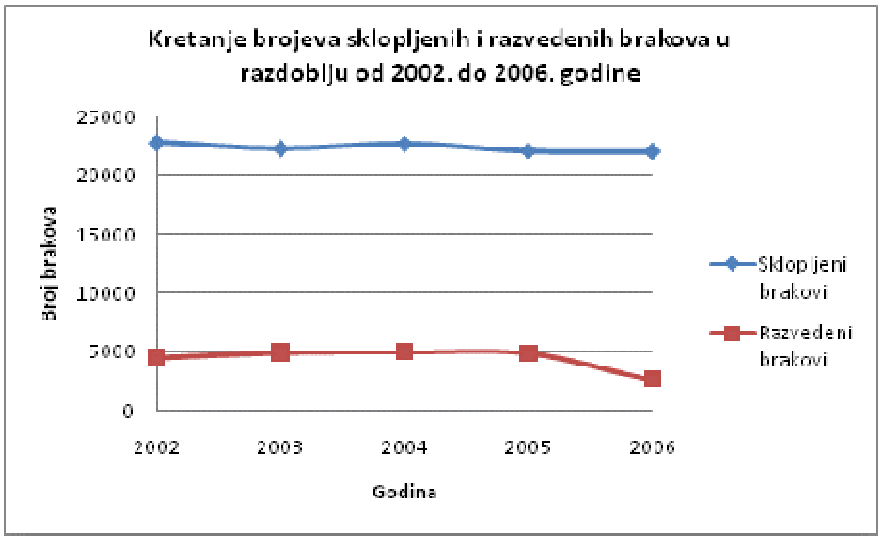
Stanje u svakoj pojedinoj godini možemo usporediti sa stanjem u 2007. godini, pa npr. možemo zaključiti (kako?) da je u 2003. godini bilo za 35.6% više korisnika novčane naknade nego u 2007. godini, u 2004. godini za 14% više korisnika novčane naknade nego u 2007. godini itd. U ovome je primjeru bazno razdoblje 2007. godina, a bazna vrijednost 59 603 (vrijednost promatrane pojave u baznom razdoblju).

Uobičajeni *grafički prikaz* vremenskoga niza jest linijski grafikon. Naslov takvoga grafikona obično je izraz tipa "*Kretanje pojave X u razdoblju od… do…*". Pritom ćemo, radi jednostavnosti, izraz tipa "*razdoblje od* 2000. *do* 2007. *godine*" , interpretirati kao "*vremensko razdoblje od* 01.01.2000. *do* 31.12.2007., tj. u to ćemo razdoblje "uključiti" cijelu 2000. godinu i cijelu 2007. godinu[116]. Grafički prikazi podataka iz Primjera 1. i Primjera 3. su sljedeće dvije slike.



[115] Jedan od najčešćih razloga je političke prirode: npr. u predizbornoj kampanji trenutno vladajuća stranka obično uspoređuje prosječne plaće, cijene, vanjski dug itd. u svakoj godini obnašanja vlasti sa stanjem u godini neposredno prije prošlih parlamentarnih izbora na kojima je preuzela obnašanje vlasti.

[116] Matematički precizno, pod pojmom "razdoblje između trenutka X_1 i trenutka X_2" podrazumijevat ćemo zatvoreni interval (segment) $[X_1, X_2]$.



Budući da kretanje vrijednosti neke pojave pratimo u vremenu, promatranu pojavu možemo shvatiti kao *funkciju vremena*, odnosno, sukladno razmatranjima iz 4. poglavlja, shvatiti vrijeme kao nezavisnu varijablu (X), a promatranu pojavu kao zavisnu varijablu (Y). Uobičajenoj oznaci nezavisne varijable (X) pridodat ćemo indeks t kako bismo naglasili da je riječ o posebnom tipu nezavisne varijable. Dakle, nezavisnu vremensku varijablu označavat ćemo s X_t.

Radi jednostavnosti, promatrat ćemo slučajeve kada se vrijednosti neke pojave određuju u *jednakim* (pravilnim) vremenskim intervalima (npr. točno jednom mjesečno svakoga mjeseca, točno jednom godišnje svake godine itd.). Te vrijednosti poredat ćemo sukladno *kronološkoj skali* modaliteta (vidjeti 1. poglavlje), tj. skup vrijednosti promatrane pojave ćemo *kronološki urediti* radi pojednostavljivanja daljnje statističke analize.

5.1. *Definicija i vrste vremenskih nizova*

<u>(*Konačan*) *Vremenski niz*</u> je niz od ukupno $n \in \mathbf{N}$ određenih, kronološki uređenih vrijednosti neke pojave. Vrijednosti koje tvore vremenski niz obično nazivamo *frekvencije*[117]. Budući da smo istakli kako ćemo kretanje neke pojave promatrati kao funkciju vremena, potrebno je precizno definirati što će biti prirodno područje definicije (domena), a što područje vrijednosti (kodomena) takvoga pridruživanja.

Smatramo intuitivno prihvatljivim pretpostaviti da vrijednost promatrane pojave *ne ovisi* o broju vezanom uz razdoblje u kojemu je izmjerena vrijednost pojave[118]. Tako npr. u Primjeru 1. broj razvedenih brakova u 2003. godini (4 934) ne smatramo na bilo koji način povezanim

[117] Navedeni je naziv svojevrstan vremenski analogon naziva *apsolutna frekvencija* kojega smo susreli u 2. poglavlju.
[118] Javno iznošenje navedene pretpostavke našu malu visoku školu ozbiljno izlaže žestokim prosvjedima numerologa, astrologa i njima sličnih "-loga", ali pojednostavljuje daljnju statističku analizu.

s *brojem* 2002, broj sklopljenih brakova u 2004. godini (22 700) ne smatramo na bilo koji način povezanim s *brojem* 2004 itd. Stoga ćemo za prirodno područje definicije (domenu) nećemo uzimati npr. skupove tipa {2002, 2003, 2004, 2005, 2006}. Kako bismo što više pojednostavnili statističku analizu, *početnom razdoblju* (odnosno, razdoblju u kojemu smo prvi put odredili vrijednost promatrane pojave) dodijelit ćemo *vrijednost* vremenske varijable $X_t = 0$. Osnovni razlog takvom pridruživanju jest korištenje modela jednostavne linearne regresije i jednostavne eksponencijalne regresije u analizi vremenskih nizova: ti modeli pretpostavljaju da se vrijednost zavisne varijable Y može izračunati u slučaju kad je vrijednost nezavisne varijable jednaka nuli[119]. Stoga ćemo, slobodno govoreći, vrijednosti vremenske varijable početi brojati od nule[120]. Preciznije, početnom vremenskom razdoblju dodijelit ćemo vrijednost vremenske varijable $t = 0$, njemu neposredno sljedećem vremenskom razdoblju vrijednost $t = 1$, itd. Imamo li ukupno n vremenskih razdoblja u kojima smo određivali vrijednost neke pojave, posljednjem od njih – a to je, zapravo, tekuće razdoblje – dodijelit ćemo vrijednost vremenske varijable $t = n - 1$. Prema tome, prirodno područje definicije (domena) *svake* funkcije vremena koju budemo razmatrali bit će konačan skup $\{0, 1, \ldots, n - 1\}$.

Područje vrijednosti (kodomena) svake funkcije vremena bit će skup frekvencija. Radi jednostavnosti, s Y_k označavat ćemo frekvenciju pridruženu vrijednosti vremenske varijable $X_t = k$. Tako je npr. Y_0 vrijednost pojave pridružena vrijednosti vremenske varijable $t = 0$, odnosno *vrijednost pojave u početnom razdoblju* (jer je $t = 0$ vrijednost nezavisne varijable dodijeljena početnom vremenskom razdoblju) a Y_{n-1} vrijednost pojave pridružena vrijednosti vremenske varijable $t = n - 1$, odnosno *vrijednost pojave u tekućem razdoblju* (jer je $t = n - 1$ vrijednost vremenke varijable dodijeljena tekućem razdoblju). Ovakav dogovor svakako treba imati na umu prigodom provedbe statističke analize.

U prvi mah nam se možda čini da smo pogriješili kad smo rekli da ćemo kretanje pojave u vremenu opisati kao *funkciju* vremena. Naime, naše dosadašnje shvaćanje pojma funkcije pretpostavlja da, nakon definiranja prirodnoga područja definicije (domene) i područja vrijednosti (kodomene), zadamo "formulu" (preciznije, analitički izraz u zatvorenoj formi) pomoću koje se za svaku pojedinu vrijednost nezavisne varijable *računa* pripadna vrijednost zavisne varijable. Međutim, u matematici se funkcija može potpuno definirati na tri načina: već opisanom "formulom", tablično i grafički. Tablični i grafički prikaz naročito su pogodni načini zadavanja funkcije ukoliko je njezino prirodno područje definicije (domena) relativno "mali" konačan skup, a u analizi vremenskih nizova imamo upravo takav slučaj. Stoga ćemo spomenutu funkciju vremena uvijek najprije definirati tablično, a potom ćemo – potpuno analogno kao u regresijskoj analizi – nastojati pronaći najbolju "formulu", tj. analitički izraz kojim možemo definirati istu tu funkciju. Taj će izraz praktički uvijek biti približno točan jer je korelacija vremena i promatrane pojave vrlo rijetko funkcijska. Iz toga se razloga statistički modeli korišteni u analizi vremenskih nizova mogu ubrojiti u regresijske modele, pa se njihova reprezentativnost može procjenjivati na isti način kao i za sve druge regresijske modele.

---

[119] Štoviše, jedan od parametara svakoga od tih modela interpretira se upravo kao vrijednost odgovarajuće funkcije za $x = 0$.
[120] Za sve vjernike: oprez, $t = 0$ ne znači da je prva vrijednost pojave izmjerena u godini Kristova rođenja.

Zaključno spomenimo da sve vremenske nizove možemo podijeliti u dvije skupine:

1.) <u>*intervalni vremenski niz*</u> – to je niz frekvencija dobivenih mjerenjem u određenomu vremenskom *intervalu* (npr. broj noćenja u svratištu "*Zlatna kokoš*" u Škrabutniku u razdoblju od 2001. godine do 2007. godine, broj vozila na autocesti Zagreb – Lipovac u razdoblju od 08:00 do 20:00 sati itd.).

2. <u>*trenutačni vremenski niz*</u> – to je niz frekvencija dobivenih mjerenjem ili promatranjem u određenomu *trenutku* (npr. broj stanovnika mjesta Sveti Petar u Šumi u razdoblju od 1971. godine do 2001. godine – stanje na dan 30.6., broj vozila na autocesti Zagreb – Lipovac dana 15.svibnja 2008. u 12:00 sati itd.).

5.2. *Pokazatelji pojedinačnih apsolutnih promjena*

Bilo koja razlika vrijednosti dviju frekvencija promatrane pojave naziva se <u>*apsolutna promjena*</u> te pojave. Apsolutna promjena može biti <u>*pozitivna*</u> ili <u>*negativna*</u>: ako je njezin predznak +, govorimo o *povećanju* vrijednosti pojave, a ako je njezin predznak –, govorimo o *smanjenju* vrijednosti pojave.

<u>Primjer 4.</u> Godinama (x) iz Primjera 3. pridružimo vrijednosti vremenske varijable (t) prema formuli: $t = x -$ 2003. Drugim riječima, 2003. godini pridružimo $t = 0$, 2004. godini pridružimo $t = 1$ itd. Izračunajmo razlike $Y_0 - Y_4$ i $Y_3 - Y_1$ i objasnimo njihovo značenje. Imamo redom:

$$Y_0 - Y_4 = 67\,977 - 59\,603 = 8\,374,$$

što znači da je 2003. *godine prosječan godišnji broj korisnika novčane naknade bio za* 8 374 *veći nego* 2007. *godine* (ili, ekvivalentno, da je 2007. godine prosječan godišnji broj korisnika novčane naknade bio za 8 374 manji nego 2003. godine). Nadalje,

$$Y_3 - Y_1 = 66\,407 - 70\,467 = -\,4\,060,$$

što znači da je 2006. *godine prosječan godišnji broj korisnika novčane naknade bio za* 4 060 *manji nego* 2004. *godine* (ili, ekvivalentno, da je 2004. godine prosječan godišnji broj korisnika novčane naknade bio za 4 060 veći nego 2006. godine).

U praksi se najčešće računaju vrijednosti niza <u>*uzastopnih apsolutnih promjena*</u> (oznaka: ΔY_t) neke pojave. Te se vrijednosti dobiju tako da se od vrijednosti pojave u *tekućemu* razdoblju oduzmemo vrijednost pojave u *neposredno prethodnomu* razdoblju, tj.

$$\Delta Y_t = Y_t - Y_{t-1}, \quad \text{za svaki } t = 1, 2, 3, \ldots, n-1.$$

<u>Primjer 5.</u> Izračunajmo vrijednosti uzastopnih apsolutnih promjena prosječnoga broja korisnika novčane pomoći (podaci iz Primjera 3.). Dobivamo:

godina	*t*	*prosječan godišnji broj korisnika novčane naknade*	*uzastopne apsolutne promjene* (ΔY_t)
2003.	0	67 977	nepoznata
2004.	1	70 467	2 490
2005.	2	72 802	2 335
2006.	3	66 407	-6 395
2007.	4	59 603	-6 804

U presjeku drugoga retka i trećega stupca gornje tablice upisali smo riječ *nepoznata* jer se vrijednost uzastopne apsolutne promjene prosječnoga godišnjega broja korisnika novčane naknade za 2003. godinu dobije tako da se od prosječnoga godišnjega broja korisnika novčane naknade u 2003. godini oduzme prosječan godišnji broj korisnika novčane naknade u 2002. godini, a potonji podatak nemamo naveden u tablici. Ovo pravilo vrijedi pri svakom računanju apsolutnih promjena:

***Pravilo* 1.** Vrijednost uzastopne apsolutne promjene za početno razdoblje (t = 0) uvijek je nepoznata. **OPREZ! Pogrešno je kao vrijednost uzastopne apsolutne promjene za početno razdoblje (t = 0) navesti broj 0 (nula)** jer broj 0, kao pokazatelj apsolutne promjene, znači *jednakost vrijednosti pojave u odgovarajućim razdobljima*, a ne da taj pokazatelj ne možemo efektivno izračunati.

Iz dobivenoga stupca tablice vidimo npr. da je prosječan broj korisnika novčane naknade u 2004. godini bio za 2 490 veći nego u 2003. godini, prosječan broj korisnika novčane naknade u 2006. godini bio je za 6 395 manji nego u 2005. godini itd.

Istaknimo i da se apsolutne promjene mogu računati u odnosu na *isto* (bazno) vremensko razdoblje. Odgovarajući se pokazatelji dobiju tako da se od frekvencije pojave u *tekućemu* razdoblju oduzme frekvencija pojave u *baznomu* razdoblju (b):

$$\Delta^* Y_t = Y_t - Y_b, \quad \text{za svaki } t = 0, 1, \ldots, n-1.$$

Primjer 6. Izračunajmo pokazatelje apsolutnih promjena u odnosu na 2004. godinu za podatke iz Primjera 3. Budući da je b = 1, tj. $Y_b = Y_1$ = 70 467, dobivamo:

godina	*t*	*prosječan godišnji broj korisnika novčane naknade*	*pojedinačne apsolutne promjene u odnosu na* 2004. *godinu* ($\Delta^* Y_t$)
2003.	0	67 977	-2 490
2004.	1	70 467	0
2005.	2	72 802	2 335
2006.	3	66 407	-4 060
2007.	4	59 603	-10 864

Iz posljednjega stupca dobivene tablice možemo zaključiti npr. da je prosječan broj korisnika novčane pomoći u 2005. godini bio za 2 335 veći, a prosječan broj korisnika novčane pomoći u 2006. godini za 4 060 manji nego u 2004. godini.

5.3. *Modeli trenda*

Ukoliko postoji određena pravilnost u promjenama vrijednosti promatrane pojave u određenom vremenskom razdoblju (tj. ukoliko vremenski niz *ima tendenciju* rasta ili pada), kažemo da vremenski niz *ima trend*. Tendencija rasta ili pada se vrlo dobro može "očitati" izračunavanjem vrijednosti uzastopnih apsolutnih apsolutnih promjena: ako je dobiveni niz *rastući*, tj. ukoliko za svaki $t = 1, 2, \ldots, n - 2$ vrijedi nejednakost

$$\Delta Y_t \leq \Delta Y_{t+1},$$

kažemo da vremenski niz ima *tendenciju rasta*, a ukoliko vrijedi nejednakost

$$\Delta Y_t \geq \Delta Y_{t+1},$$

kažemo da vremenski niz ima *tendenciju pada*. U takvim se slučajevima kretanje promatrane pojave može opisati pomoću odgovarajućega modela trenda. Grubo govoreći, modeli trenda, zapravo, nisu ništa drugo negoli "obični" regresijski modeli u kojima "ulogu" nezavisne varijable X preuzima vremenska varijabla t.

Ukoliko se promatrana pojava mijenja za približno jednake apsolutne iznose u svakoj jedinici vremena, njezino kretanje možemo opisati *modelom linearnoga trenda*. Odgovarajuća jednadžba modela linearnoga trenda i pripadni parametri isti su kao i kod modela jednostavne linearne regresije, samo što se u slučaju modela linearnoga trenda vrijednost parametra A interpretira kao *očekivana prosječna promjena vrijednosti varijable Y u jedinici vremena*.

Ukoliko se, pak, promatrana pojava ne mijenja linearno, nego sve brže (ili sporije) ovisno o vrijednosti vremenske varijable, njezino kretanje možemo opisati *modelom eksponencijalnoga trenda*. Odgovarajuća jednadžba i pripadni parametri toga modela trenda isti su kao i kod modela jednostavne eksponencijalne regresije, samo što se u slučaju modela eksponencijalnoga trenda vrijednost parametra A interpretira kao *očekivana relativna prosječna promjena vrijednosti varijable Y u jedinici vremena*.

Reprezentativnost modela trenda obično se procjenjuje na temelju koeficijenta varijacije toga modela, ali se za procjenu može koristiti i koeficijent determinacije. Formule za izračun navedenih pokazatelja jednake su odgovarajućim formulama iz podtočaka 4.2.3.3. i 4.2.3.4. Kad god je to moguće, treba izračunati oba navedena pokazatelja.

Ilustrirajmo primjenu modela trenda na sljedećem primjeru:

Primjer 1. U donjoj je tablici prikazano kretanje prosječne cijene nafte u SAD u razdoblju od 1. travnja 2007. do 31. travnja 2008. godine.

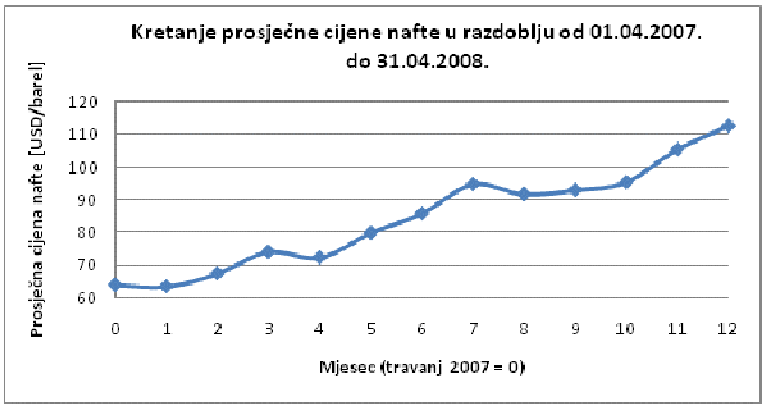
Poslovna statistika

mjesec i godina	*t*	*prosječna cijena nafte* [USD/barel]
travanj 2007.	0	63,98
svibanj 2007.	1	63,45
lipanj 2007.	2	67,49
srpanj 2007.	3	74,12
kolovoz 2007.	4	72,36
rujan 2007.	5	79,91
listopad 2007.	6	85,80
studeni 2007.	7	94,77
prosinac 2008.	8	91,69
siječanj 2008.	9	92,97
veljača 2008.	10	95,39
ožujak 2008.	11	105,46
travanj 2008.	12	112,58

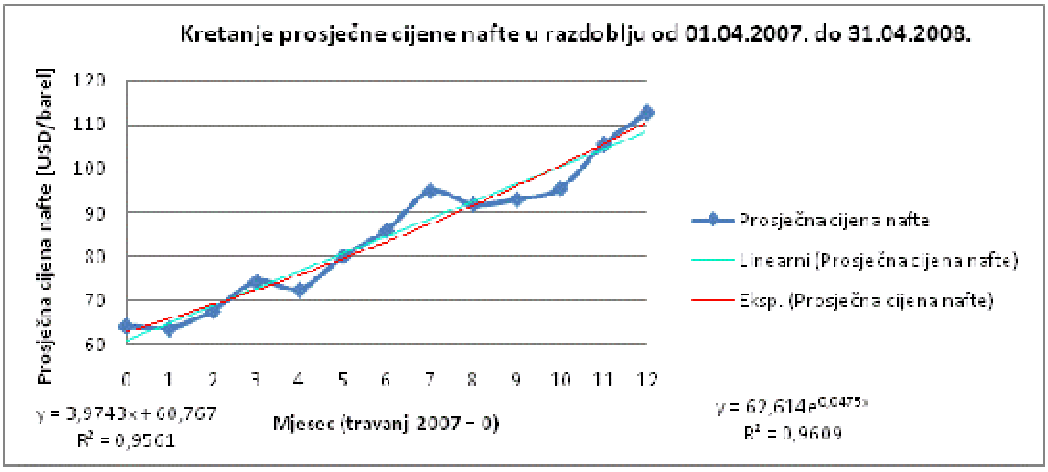
Izvor: Energy Information Administration – službena energetska statistika Vlade SAD

a) Prikažimo kretanje prosječne cijene nafte u promatranom razdoblju linijskim grafikonom s ishodištem u početnoj godini razdoblja promatranja. Uz grafikon navedimo sve potrebne oznake.
b) Odredimo jednadžbu modela linearnoga, odnosno eksponencijalnoga trenda koji najbolje opisuje kretanje promatrane pojave. Objasnimo značenje parametara trenda i prikažimo dobivene modele grafički.
c) Na temelju koeficijenata determinacije procijenimo koji od navedenih modela trenda bolje opisuje kretanje prosječne cijene nafte. Potom na temelju reprezentativnijega modela trenda prognozirajmo kretanje prosječne cijene nafte u svakom pojedinom mjesecu do kraja 2008. godine.
d) Ukoliko se kretanje prosječne cijene nafte nastavi prema modelu odabranom u **c)** podzadatku, odredimo u kojem mjesecu možemo očekivati da cijena nafte prvi put dosegne 200 USD/barel.

Kretanje prosječne cijene nafte na američkom tržištu u promatranom razdoblju grafički je prikazano sljedećim linijskim grafikonom:



Jednadžbe modela linearnoga, odnosno eksponencijalnoga trenda zajedno s pripadajućim krivuljama trenda i koeficijentima determinacije navedene su na sljedećoj slici:



Zasebno ćemo analizirati svaki model trenda:

- *model linearnoga trenda*: njegovi parametri su:
  $A = 3.9743 \Leftrightarrow$ *očekivano mjesečno prosječno povećanje prosječne cijene nafte iznosi približno* 4 USD *po barelu*;
  $B = 60.767 \Leftrightarrow$ *procijenjena prosječna cijena nafte u travnju* 2007. *godine iznosi približno* 60,77 USD
- *model eksponencijalnoga trenda*: njegovi parametri su:
  $A = e^{0.0475} = 1.048646 \Rightarrow 100 \cdot (A - 1) \approx 4.865 \Leftrightarrow$ *očekivani prosječan mjesečni rast prosječne cijene nafte iznosi* 4.865%;
  $B = 62.614 \Leftrightarrow$ *procijenjena prosječna cijena nafte u travnju* 2007. *godine iznosi približno* 62,61 USD

Koeficijent determinacije modela linearnoga trenda iznosi 0.9561, što znači da se 95.61% *periodičnih promjena prosječne cijene nafte može objasniti modelom linearnoga trenda*. S druge strane, koeficijent determinacije modela eksponencijalnoga trenda iznosi 0.9609, što znači da se 96.09% *periodičnih promjena prosječne cijene nafte može objasniti modelom eksponencijalnoga trenda*. Prema kriteriju koeficijenata determinacije, *model eksponencijalnoga trenda je reprezentativniji*.

Prosječnu cijenu nafte u svakom pojedinom mjesecu preostalom do kraja 2008. godine možemo prognozirati na dva načina:

1.) uvrštavanjem $x = 13$, $x = 14$, …, $x = 19$ i $x = 20$ u svaku od dobivenih jednadžbi;
2.) (u MS Excelu) korištenjem funkcije TREND za prognozu pomoću modela linearnoga trenda, odnosno GROWTH za prognozu pomoću modela eksponencijalnoga trenda.

U oba slučaja dobivamo:

mjesec i godina	*t*	*prognoza pomoću modela linearnoga trenda*	*prognoza pomoću modela eksponencijalnoga trenda*
svibanj 2008.	13	112,43 USD	116,05 USD
lipanj 2008.	14	116,41 USD	121,69 USD
srpanj 2008.	15	120,38 USD	127,61 USD
kolovoz 2008.	16	124,36 USD	133,81 USD
rujan 2008.	17	128,33 USD	140,32 USD
listopad 2008.	18	132,30 USD	147,14 USD
studeni 2008.	19	136,28 USD	154,29 USD
prosinac 2008.	20	140,25 USD	161,79 USD

Da bismo procijenili kada će prosječna cijena nafte prvi puta dosegnuti 200 USD po barelu, trebamo riješiti jednadžbu

$$62.614 \cdot 1.048646^t = 200$$

Njezino je rješenje[121]:

$$t = \frac{\ln 200 - \ln 62.614}{\ln 1.048646} \approx 24.$$

Vrijednost vremenske varijable $t = 24$ odgovara travnju 2009. godine. Prema tome, ukoliko se kretanje prosječne cijene nafte nastavi prema modelu eksponencijalnoga trenda, može se očekivati da će prosječna cijena nafte u travnju 2009. godine dosegnuti 200 USD/barel.

5.4. *Osnovni brojčani pokazatelji relativnih promjena*

U praksi se često pojavljuje problem analiziranja i usporedbe dinamike kretanja najmanje dviju različitih pojava. Vrijednosti tih pojava obično su iskazane u različitim mjernim jedinicama ili se, pak, značajno razlikuju po svojoj veličini. Zbog toga se u takvim slučajevima, umjesto apsolutnih promjena, izračunavaju i interpretiraju *relativni pokazatelji* promjena tih pojava.

Jedna od najčešćih mjera relativnih promjena je *koeficijent dinamike* (oznaka: K_t). Koeficijent dinamike jednak je omjeru apsolutne promjene frekvencija dvaju uzastopnih razdoblja i frekvencije pojave u kronološki prvom od tih dvaju razdoblja:

$$K_t = \frac{\Delta Y_t}{Y_{t-1}} = \frac{Y_t}{Y_{t-1}} - 1 \quad \text{za svaki } t = 1, 2, \ldots, n-1.$$

[121] Svima kojima rješavanje eksponencijalno-logaritamskih jednadžbi predstavlja težak i zahtjevan matematički problem preporučuje se da navedenu jednadžbu riješe korištenjem procedure *Traženje rješenja* (*Goal Seek*).

Usko vezana uz koeficijent dinamike je <u>*pojedinačna stopa promjene*</u> (oznaka: S_t) neke pojave. Ona iskazuje relativnu promjenu vrijednosti pojave u dvaju uzastopnim razdobljima, a dobije se množenjem koeficijenta dinamike sa 100:

$$S_t = K_t \cdot 100 = 100 \cdot \left(\frac{Y_t}{Y_{t-1}} - 1 \right), \text{ za svaki } t = 1, 2, \ldots, n-1.$$

Pojedinačna stopa promjene zapravo predstavlja *postotak* promjene frekvencije neke pojave u *tekućemu* razdoblju u odnosu na prethodno razdoblje. Stoga se uz numeričku vrijednost pojedinačne stope promjene uvijek piše znak %.

Osim za dva uzastopna razdoblja, pojedinačna stopa promjene može se računati i prigodom uspoređivanja vrijednosti neke pojave tijekom svakoga od n razdoblja s vrijednošću te pojave u proizvoljno odabranom, ali fiksiranom razdoblju. U takvim slučajevima *pojedinačnu stopu promjene* vrijednosti pojave u *tekućemu* razdoblju u odnosu na vrijednost te pojave u (proizvoljnom, ali fiksiranom) *baznom* razdoblju dobivamo iz izraza:

$$S_t^* = \frac{\Delta^* Y_t}{Y_b} \cdot 100 = \left(\frac{Y_t}{Y_b} - 1 \right) \cdot 100, \text{ za svaki } t = 0, 1, 2, \ldots, n-1.$$

<u>Primjer 1.</u> U sljedećoj su tablici prikazani podaci o kretanju broja osnovanih registriranih pravnih osoba u Republici Hrvatskoj u proteklih godinu dana.

razdoblje	*broj osnovanih registriranih pravnih osoba*
1. kvartal 2007.	3 724
2. kvartal 2007.	3 567
3. kvartal 2007.	3 240
4. kvartal 2007.	3 500
1. kvartal 2008.	3 783

izvor: *Broj i struktura poslovnih subjekata u Republici Hrvatskoj*, priopćenje br. 11.1.1/1, Državni zavod za statistiku, Zagreb, 9.5.2008.

Izračunajmo koeficijente dinamike promjene broja osnovanih registriranih pravnih osoba u uzastopnim vremenskim razdobljima, te odgovarajuće pojedinačne stope promjene, pa objasnimo značenje dobivenih rezultata za 3. kvartal 2007. godine i 1. kvartal 2008. godine. Potom usporedimo brojeve osnovanih registriranih pravnih osoba u svakom pojedinom kvartalu s brojem osnovanih registriranih pravnih osoba u posljednjem kvartalu 2007. godine i objasnimo značenje pokazatelja koji odgovara 1. kvartalu 2007. godine.

Najprije izračunavamo koeficijente dinamike i njima odgovarajuće stope promjene:

razdoblje	*t*	*broj osnovanih registriranih pravnih osoba*	K_t	S_t
1. kvartal 2007.	0	3 724	*nepoznat*	*nepoznata*
2. kvartal 2007.	1	3 567	-0.04216	-4.2159 %
3. kvartal 2007.	2	3 240	-0.09167	-9.16737 %
4. kvartal 2007.	3	3 500	0.080247	8.024691 %
1. kvartal 2008.	4	3 783	0.080857	8.085714 %

U presjecima 2. retka i 3., odnosno 4. stupca upisali smo riječi *nepoznat* i *nepoznata* jer nam za izračunavanje vrijednosti odgovarajućega pokazatelja nedostaju podaci iz 1. kvartalu neposredno prethodnoga razdoblja, tj. iz 4. kvartala 2006. godine.

Iz dobivene tablice vidimo da je *broj osnovanih registriranih pravnih osoba u* 3. *kvartalu* 2007. *godine bio za približno* 9.17% *manji nego u* 2. *kvartalu iste godine*, te da je *broj osnovanih registriranih pravnih osoba u* 1. *kvartalu* 2008. *godine bio za približno* 8.09% *veći nego u* 4. *kvartalu* 2007. *godine*.

Sada izračunavamo pojedinačne stope promjene uzimajući 4. kvartal 2007. godine kao *bazno* razdoblje:

razdoblje	*t*	*broj osnovanih registriranih pravnih osoba*	S_t^*
1. kvartal 2007.	0	3 724	6.4%
2. kvartal 2007.	1	3 567	1.91429%
3. kvartal 2007.	2	3 240	-7.42857%
4. kvartal 2007.	3	3 500	0%
1. kvartal 2008.	4	3 783	8.08571%

Pokazatelj koji odgovara 1. kvartalu 2007. godine jednak je 6.4%, što znači da je *broj osnovanih registriranih pravnih osoba u* 1. *kvartalu* 2007. *godine bio za* 6.4% *veći nego u posljednjem kvartalu* 2007. *godine*.

5.5. *Individualni indeksi*

Jedna od najčešće rabljenih relativnih promjena vrijednosti neke pojave su tzv. *indeksi*. Pomoću njih uspoređujemo odnos stanja jedne ili više pojava u različitim vremenskim razdobljima ("vremenskim točkama") ili u istim vremenskim razdobljima, ali na različitim mjestima (tj. u različitim gradovima, regijama, državama, skupinama država itd.)
Opća podjela razlikuje dvije vrste indeksa: *individualne* i *skupne*. *Individualni indeksi* prate razvoj točno jedne pojave u vremenu ili prostoru, te se uobičajeno dijele na *verižne* (*lančane*) *indekse* i *indekse na stalnoj bazi (bazne indekse)*. Za razliku od jednostavnih, *skupni indeksi* prate razvoj barem dvije različite pojave u vremenu ili prostoru.

5.5.1. *Verižni (lančani) indeksi*

Verižni (lančani) indeks (oznaka: V_t) predstavlja pokazatelj relativne promjene stanja neke pojave u dvama uzastopnim vremenskim razdobljima. On je, dakle, pokazatelj relativne promjene stanja neke pojave u *tekućemu* razdoblju u odnosu na tekućem razdoblju neposredno prethodno razdoblje. Ukoliko kretanje pojave promatramo tijekom n razdoblja, onda ukupno $n - 1$ verižnih indeksa računamo pomoću relacije:

$$V_t = \frac{Y_t}{Y_{t-1}} \cdot 100, \text{ za svaki } t = 1, 2, \ldots, n-1,$$

ili, ekvivalentno, na temelju koeficijenata dinamike:

$$V_t = (K_t + 1) \cdot 100, \text{ za svaki } t = 1, 2, \ldots, n-1.$$

Svaki verižni indeks zapravo predstavlja *postotni udio* vrijednosti promatrane pojave u tekućemu razdoblju u odnosu na vrijednost te pojave u neposredno prethodnomu razdoblju. Ako je $V_t = A$, onda opća interpretacija toga indeksa glasi:

Na svakih 100 *jedinica mjere promatrane pojave u razdoblju* $(t - 1)$ *dolazi A jedinica mjere te pojave u razdoblju t.*

Istaknimo da, kao i koeficijent dinamike, verižni indeks *nije definiran* za početno razdoblje ($t = 0$), pa se u takvim slučajevima obično stavlja riječ *nepoznat*.[122]

Umjesto iz koeficijenta dinamike, pojedinačnu stopu promjene u uzastopnim vremenskim razdobljima možemo izračunati i pomoću odgovarajućega verižnoga indeksa:

$$S_t = V_t - 100, \text{ za svaki } t = 1, 2, \ldots, n-1.$$

Primjer 1. U sljedećoj su tablici navedeni podaci o ukupnom broju zaposlenih u djelatnostima financijskoga posredovanja u svakom od mjeseci koji tvore 1. kvartal 2008. godine.

mjesec	*ukupan broj zaposlenih*
siječanj	1 604
veljača	1 636
ožujak	1 621

izvor: *Zaposleni u obrtu i u djelatnostima slobodnih profesija u prvom tromjesečnju* 2008., priopćenje br. 9.2.2/1, Državni zavod za statistiku, Zagreb, 12.5.2008.

[122] Ova činjenica znači da je *pogrešno* kao verižni indeks za početno razdoblje definirati broj 0. U ovom bi slučaju vrijednost 0 povlačila da znamo vrijednost promatrane pojave u razdoblju koje neposredno prethodi početnom razdoblju i da je ta vrijednost jednaka nuli. Očito, niti jedna niti druga posljedica nisu istinite.

Izračunajmo odgovarajuće koeficijente dinamike, verižne indekse i pojedinačne stope promjene, pa objasnimo značenje vrijednosti dobivenih za ožujak 2008.

Vrijednosti traženih relativnih pokazatelja navedene su u sljedećoj tablici:

mjesec	*t*	*ukupan broj zaposlenih*	K_t	V_t	S_t
siječanj	0	1 604	*nepoznat*	*nepoznat*	*nepoznata*
veljača	1	1 636	0.01995	101.995	1.995012 %
ožujak	2	1 621	-0.00917	99.08313	-0.91687 %

Iz dobivenih pokazatelja zaključujemo da je *na svakih* 100 *zaposlenih u djelatnostima financijskoga posredovanja u ožujku* 2008. *godine dolazilo približno* 102 *zaposlena u veljači iste godine*, odnosno da je *broj zaposlenih u djelatnostima financijskoga posredovanja u ožujku* 2008. *godine bio za približno* 0.92% *manji nego u veljači iste godine.*

Pomoću *svih* $n - 1$ verižnih indeksa kojima mjerimo dinamiku promjene vrijednosti neke pojave u n razdoblja, te zadane vrijednosti te pojave u *bilo kojem* od tih n razdoblja (označimo ga s b) možemo izračunati vrijednosti pojave u *svakom* od preostalih $n - 1$ razdoblja na temelju sljedećih relacija:

$$Y_{t-1} = \frac{Y_t}{V_t} \cdot 100, \quad \text{za} \quad t < b.$$

$$Y_t = \frac{V_t \cdot Y_{t-1}}{100}, \quad \text{za} \quad b < t.$$

Ilustrirajmo taj postupak na primjeru.

<u>Primjer 2.</u> U sljedećoj su tablici navedeni verižni indeksi promjene prosječne isplaćene mjesečne neto-plaće u razdoblju od 2003. do 2007. godine.

godina	*t*	V_t
2003.	0	*nepoznat*
2004.	1	105.9
2005.	2	105.9
2006.	3	105.2
2007.	4	105.2

izvor: *Mjesečno statističko izvješće* 3/2008, Državni zavod za statistiku, ožujak, 2008.

Ako je prosječna mjesečna isplaćena neto-plaća u 2006. godini iznosila 4.603,00 kn, odredimo prosječne mjesečne isplaćene neto-plaće u svakoj od preostalih godina.

2006. godini odgovara vrijednost vremenske varijable $t = 3$, pa je $Y_3 = 4.603{,}00$. Za $t < 3$, odnosno $t = 0, 1, 2$ dobivamo redom:

$$Y_2 = \frac{Y_3}{V_3} \cdot 100 = \frac{4.603,00}{105.2} \cdot 100 \approx 4.375,48$$

$$Y_1 = \frac{Y_2}{V_2} \cdot 100 = \frac{4.375,48}{105.9} \cdot 100 \approx 4.131,70$$

$$Y_0 = \frac{Y_1}{V_1} \cdot 100 = \frac{4.131,70}{105.9} \cdot 100 \approx 3.901,52$$

dok za $t > 3$, odnosno $t = 4$ dobivamo:

$$Y_4 = \frac{V_4 \cdot Y_3}{100} = \frac{105,2 \cdot 4.603,00}{100} \approx 4.842,36\,.$$

Dobivene podatke možemo zapisati u obliku tablice:

godina	*t*	*prosječna mjesečna isplaćena neto–plaća* [kn]	V_t
2003.	0	3.901,52	*nepoznat*
2004.	1	4.131,70	105.9
2005.	2	4.375,46	105.9
2006.	3	4.603,00	105.2
2007.	4	4.842,36	105.2

Pri računanju *prosječnoga* tempa promjene promatrane pojave rabimo <u>*geometrijsku sredinu verižnih indeksa*</u> (oznaka: G). Pojam geometrijske sredine već smo upoznali u točki 3.1.2. pa ćemo još jednom navesti samo osnovnu formulu za njezino računanje (u terminologiji verižnih indeksa):

$$G = \sqrt[n-1]{V_1 \cdot V_2 \cdot \ldots \cdot V_{n-1}}\,.$$

Iz definicije verižnih indeksa slijedi da njihovu geometrijsku sredinu možemo izračunati i prema formuli:

$$G = \sqrt[n-1]{\frac{Y_{n-1}}{Y_0}} \cdot 100$$

Stoga možemo zaključiti da za alternativno računanje vrijednosti geometrijske sredine verižnih indeksa moramo znati samo prvi (Y_0) i posljednji (Y_{n-1}) član vremenskoga niza. To je ujedno i najveći nedostatak geometrijske sredine (kao mjere prosječnoga tempa promjene) jer *ne* uzima u obzir *sve* frekvencije vremenskoga niza (za razliku od, npr., ranije spomenutih modela trenda), nego isključivo prvu i posljednju frekvenciju. Pomoću spomenute geometrijske sredine verižnih indeksa računa se i *prosječna stopa promjene* vrijednosti promatrane pojave (oznaka: $\overline{S}$):

$$\overline{S} = G - 100.$$

Kao i svi drugi relativni pokazatelji, i prosječna stopa promjene iskazuje se u postotcima. Na temelju nje se okvirno mogu *procijeniti* (*prognozirati*) *frekvencije* te pojave i u *budućim vremenskim razdobljima* uz nužnu pretpostavku da se dinamika pojave ne mijenja u odnosu na dinamiku pojave u promatranomu razdoblju. Pri takvim se prognozama koristi relacija:

$$\hat{Y}_t = Y_0 \cdot \left(\frac{G}{100}\right)^t, \text{ za } t = n, n+1, \ldots$$

No, budući da prosječna stopa promjene ovisi isključivo o prvoj i posljednjoj frekvenciji u vremenskom nizu, u praksi se navedene vrijednosti prognoziraju uglavnom na temelju nekoga od modela trenda.

<u>Primjer 3.</u> Na temelju podataka iz Primjera 2. izračunajmo prosječnu godišnju stopu promjene prosječne mjesečne isplaćene neto-plaće u razdoblju od 2003. do 2007. godine, pa procijenimo prosječnu mjesečnu isplaćenu neto-plaću u Republici Hrvatskoj za svaku od godina u razdoblju od 2008. do 2010. godine.
Geometrijska sredina verižnih indeksa jednaka je

$$G = \sqrt[4]{\frac{Y_4}{Y_0}} \cdot 100 = \sqrt[4]{\frac{4.842,36}{3.901,52}} \cdot 100 = 105.5494197.$$

Stoga je prosječna godišnja stopa promjene

$$\overline{S} = G - 100 = 105.5494197 - 100 \approx 5.55\%,$$

što znači da su *u promatranom razdoblju prosječne mjesečne isplaćene neto–plaće rasle za prosječno* 5.55% *godišnje*. Budući da 2008. godini odgovara vrijednost vremenske varijable t = 5, 2009. godini t = 9, a 2010. godini t = 10, prognozu prosječnih mjesečnih isplaćenih neto-plaća za 2008., 2009. i 2010. godinu dajemo na temelju izraza:

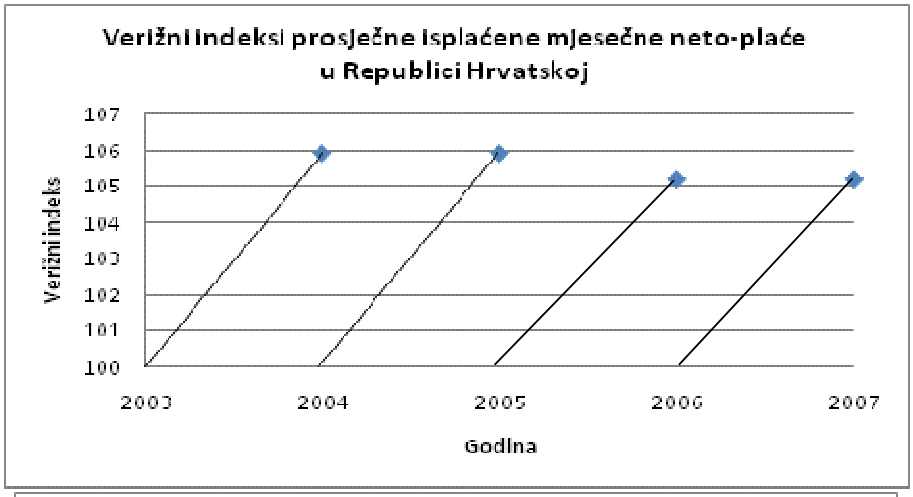
$$\hat{Y}_5 = Y_0 \cdot \left(\frac{G}{100}\right)^5 = 3.901,52 \cdot \left(\frac{105.5494197}{100}\right)^5 \approx 5.111,08$$

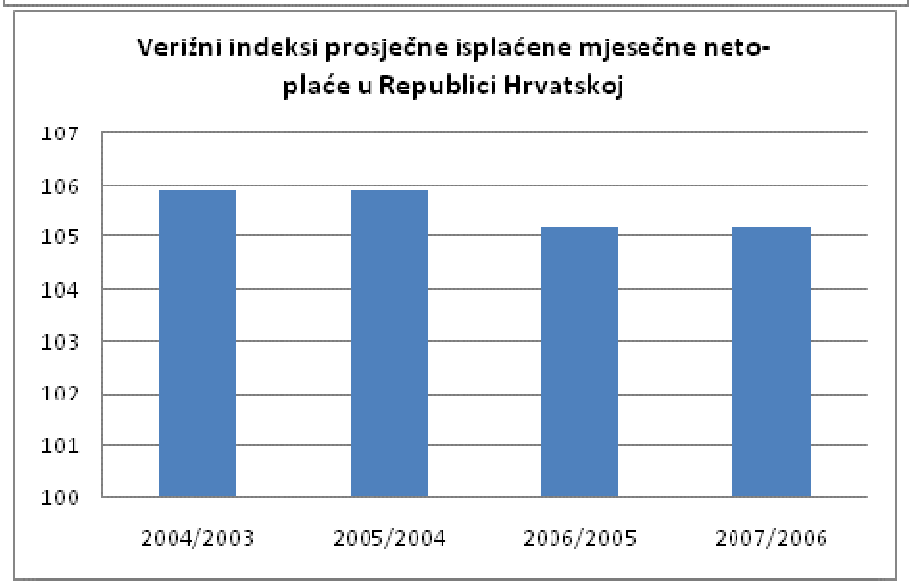
$$\hat{Y}_6 = Y_0 \cdot \left(\frac{G}{100}\right)^6 = 3.901,52 \cdot \left(\frac{105.5494197}{100}\right)^6 \approx 5.394,71$$

$$\hat{Y}_7 = Y_0 \cdot \left(\frac{G}{100}\right)^7 = 3.901,52 \cdot \left(\frac{105.5494197}{100}\right)^7 \approx 5.694,09$$

Dakle, procijenjena prosječna mjesečna isplaćena neto-plaća u Republici Hrvatskoj za 2008. godini iznosi 5.111,08 kn, u 2009. godini 5.394,71 kn, a u 2010. godini 5.694,09 kn (uz smjelu pretpostavku da neće doći do značajnije promjene dinamike kretanja plaća).

Spomenimo zaključno da se verižni indeksi grafički uobičajeno prikazuju specifičnim linijskim grafikonom ili jednostavnim stupcima. Tako su npr. grafički prikazi verižnih indeksa iz Primjera 3. sljedeći grafikoni:





5.5.2. *Indeksi na stalnoj bazi* (*bazni indeksi*)

<u>*Indeksi na stalnoj bazi*</u> (kraće: <u>*bazni indeksi*</u>) pokazatelji su relativnih promjena u *tekućemu* razdoblju u odnosu na neko proizvoljno, ali fiksirano *bazno* razdoblje. Označimo li to razdoblje s b, onda za svako pojedino razdoblje t odgovarajući indeks I_t na stalnoj bazi b računamo pomoću relacije:

$$I_t = \frac{Y_t}{Y_b} \cdot 100 \text{, za svaki } t = 0, 1, 2, \ldots, n-1.$$

Iz navedene relacije proizlaze sljedeća svojstva koja je vrlo korisno primijeniti pri provjeri ispravnosti izračuna indeksa na stalnoj bazi:

Svojstvo 1. Za $t = b$ vrijedi jednakost $I_b = 100$.

Svojstvo 2. Pretpostavimo da je $b \neq n - 1$, tj. da posljednje od ukupno n razdoblja nije ujedno i bazno razdoblje. Tada za $t = b + 1$ vrijedi jednakost:

$$I_t = V_t.$$

Ekvivalentno, *indeks razdoblja t na stalnoj bazi b jednak je verižnomu indeksu razdoblja t ako i samo ako je* $t = b + 1$.

Indeks na stalnoj bazi zapravo predstavlja *postotni udio* frekvencije promatrane pojave u *tekućemu* razdoblju u odnosu na frekvenciju promatrane pojave u *baznomu* razdoblju. Stoga je njegova opća interpretacija sljedeća:

$I_t = A \Leftrightarrow$ *Na svakh* 100 *jedinica mjere pojave A u razdoblju b dolazi A jedinica mjere pojave A u razdoblju B.*

Kao i za verižne indekse, i za indekse na stalnoj bazi može se računati odgovarajuća pojedinačna stopa promjene (oznaka: S_t^*) u odnosu na bazno razdoblje b, i to prema formuli:

$$S_t^* = I_t - 100 \text{, za svaki } t = 0, 1, 2, \ldots, n-1.$$

Primjer 4. U donjoj su tablici navedeni podaci o prosječnom godišnjem broju žena zaposlenih u pravnim osobama, obrtu i djelatnostima slobodnih profesija u Republici Hrvatskoj za razdoblje od 1997. do 2006. godine.

godina	*t*	*prosječan broj zaposlenih žena* [000]
1997.	0	546
1998.	1	574
1999.	2	571
2000.	3	573
2001.	4	578
2002.	5	580
2003.	6	595
2004.	7	606
2005.	8	617
2006.	9	640

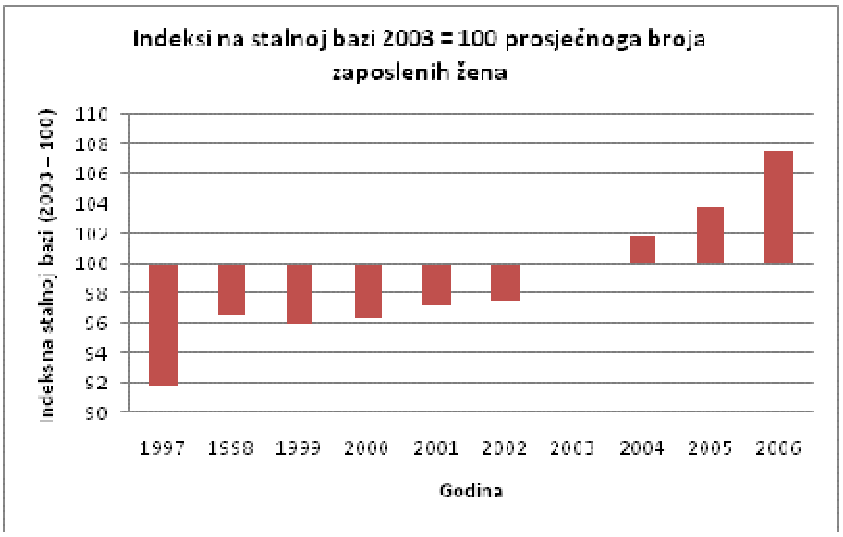
izvor: Statistički ljetopis Republike Hrvatske za 2006. godinu, Državni zavod za statistiku, Zagreb, 2007.

Izračunajmo indekse na stalnoj bazi 2003 = 100 i odgovarajuće pojedinačne stope promjene, pa objasnimo značenje navedenih pokazatelja za 2000. i 2006. godinu. Potom dobivene bazne indekse prikažimo grafički jednostavnim stupcima.

Traženi indeksi na stalnoj bazi 2003 = 100 i njima odgovarajuće stope promjene navedene su u sljedećoj tablici:

godina	*t*	*prosječan broj zaposlenih žena* [000]	I_t	S_t^*
1997.	0	546	91.76471	-8.23529
1998.	1	574	96.47059	-3.52941
1999.	2	571	95.96639	-4.03361
2000.	3	573	96.30252	-3.69748
2001.	4	578	97.14286	-2.85714
2002.	5	580	97.47899	-2.52101
2003.	6	595	100	0
2004.	7	606	101.8487	1.848739
2005.	8	617	103.6975	3.697479
2006.	9	640	107.563	7.563025

Iz dobivenih rezultata vidimo da *na svakih* 100 *zaposlenih žena u* 2003. *godini dolazi približno* 96 *zaposlenih žena u* 2000. *godini*, odnosno *približno* 108 *zaposlenih žena u* 2006. *godini*. Ekvivalentno rečeno, *prosječan broj zaposlenih žena u* 2000. *godini je za približno* 3.69% *manji*, a *u* 2006. *godini za približno* 7.56% *veći nego u* 2003. *godini.*
Indekse na stalnoj bazi obično grafički prikazujemo jednostavnim stupcima. Grafički prikaz indeksa na stalnoj bazi 2003 = 100 je donja slika:



Vidjeli smo da *sve* frekvencije neke pojave možemo odrediti pomoću verižnih indeksa i bilo koje unaprijed zadane frekvencije te pojave. Analogno, možemo ih određivati i pomoću *svih*

unaprijed zadanih indeksa na nekoj stalnoj bazi b i (barem) jedne unaprijed zadane frekvencije te pojave. Ovisno o zadanoj frekvenciji, razlikujemo sljedeća dva podslučaja:

1.) Ako je zadana frekvencija pojave u baznomu razdoblju b (tj. vrijednost Y_b), tada vrijednosti pojave u svakom od n razdoblja[123] dobivamo iz jednakosti:

$$Y_t = \frac{Y_b \cdot I_t}{100}, \text{ za svaki } t = 0, 1, 2, \ldots, n-1.$$

2.) Ako je zadana frekvencija pojave u nekom drugom razdoblju t_1 različitomu od baznoga (tj. ako je zadana frekvencija Y_{t_1}, pri čemu je $t_1 \neq b$), tada najprije izračunamo frekvenciju promatrane pojave u baznomu razdoblju b pomoću relacije:

$$Y_b = \frac{Y_t}{I_t} \cdot 100$$

(I_t je indeks razdoblja t na stalnoj bazi b). Na taj način ovaj podslučaj svodimo na podslučaj 1. pa za računanje vrijednosti pojave u ostalim godinama primijenjujemo formulu navedenu u tomu slučaju.

Napomenimo da i iz indeksa na stalnoj bazi *neovisno o izboru stalne baze b* možemo izračunati prosječnu stopu promjene frekvencija neke pojave u promatranomu razdoblju, i to prema formuli:

$$\overline{S} = \left(\sqrt[n-1]{\frac{I_{n-1}}{I_0}} - 1 \right) \cdot 100 .$$

<u>Primjer 5.</u> Zadani su indeksi na stalnoj bazi 2005 = 100 zaposlenih u pravnim osobama svih oblika vlasništva.

godina	*t*	*indeks na stalnoj bazi* 2005 = 100
1997.	0	89
1998.	1	96
1999.	2	95
2000.	3	95
2001.	4	95
2002.	5	95
2003.	6	98
2004.	7	99
2005.	8	100
2006.	9	104

[123] Navedena formula očito vrijedi i u slučaju $t = b$.

Odredimo prosječan broj zaposlenih u pravnim osobama svih oblika vlasništva u svakoj pojedinoj godini ako je poznato da je prosječan broj zaposlenih u pravnim osobama svih oblika vlasništva u 1997. godini iznosio 995 000. Potom odredimo prosječnu godišnju stopu promjene broja zaposlenih i objasnimo značenje dobivene vrijednosti.

Iz podataka Y_0 = 995 000, b = 8 (jer 2005. godini odgovara t = 8) i I_0 = 89 slijedi:

$$Y_b = Y_8 = \frac{Y_0}{I_0} \cdot 100 = \frac{995000}{89} \cdot 100 \approx 1\ 117\ 978.$$

Dakle, prosječan broj zaposlenih u baznoj godini iznosi 1 117 978. Na temelju toga podatka računamo prosječan broj zaposlenih u svakoj od preostalih godina:

$$Y_1 = \frac{Y_b \cdot I_1}{100} = \frac{Y_8 \cdot I_1}{100} = \frac{1117978 \cdot 96}{100} \approx 1073258$$
$$Y_2 = \frac{Y_8 \cdot I_2}{100} = \frac{1117978 \cdot 95}{100} \approx 1062079$$
$$Y_3 = Y_4 = Y_5 = Y_2 \approx 1062079$$
$$Y_6 = \frac{Y_8 \cdot I_6}{100} = \frac{1117978 \cdot 98}{100} \approx 1095618$$
$$Y_7 = \frac{Y_8 \cdot I_7}{100} = \frac{1117978 \cdot 99}{100} \approx 1106798$$
$$Y_9 = \frac{Y_8 \cdot I_9}{100} = \frac{1117978 \cdot 104}{100} \approx 1162697$$

Dobivene podatke pregledno možemo zapisati u sljedećoj tablici:

godina	*t*	*prosječan broj zaposlenih* [000 000]
1997.	0	0.95
1998.	1	1.073258
1999.	2	1.062079
2000.	3	1.062079
2001.	4	1.062079
2002.	5	1.062079
2003.	6	1.095618
2004.	7	1.106798
2005.	8	1.117978
2006.	9	1.162697

Prosječna godišnja stopa promjene broja zaposlenih iznosi:

$$\overline{S} = \left(\sqrt[n-1]{\frac{I_{n-1}}{I_0}} - 1 \right) \cdot 100 = \left(\sqrt[9]{\frac{I_9}{I_0}} - 1 \right) \cdot 100 = \left(\sqrt[9]{\frac{104}{89}} - 1 \right) \cdot 100 \approx 1.745668 .$$

Prema tome, *prosječan broj zaposlenih u pravnim osobama svih oblika vlasništva u promatranom se razdoblju povećavao za prosječno* 1.75% *godišnje*. Primijetimo da dobiveni pokazatelj relativno loše opisuje kretanje prosječnoga broja zaposlenih u promatranom razdoblju jer se u podrazdoblju od 1999. do 2002. godine navedeni prosječan broj nije mijenjao. Zbog toga je u takvim slučajevima za analizu promjena promatrane pojave primjerenije rabiti modele trenda.

5.5.3. *Pretvorba indeksa*

Jedna od značajnih prednosti individualnih indeksa jest mogućnost pretvorbe (preračunavanja) svake pojedine vrste indeksa u preostalu vrstu indeksa *bez poznavanja frekvencija koje tvore vremenski niz*. Drugim riječima, promjene promatrane pojave možemo analizirati i bez poznavanja originalnih emipirijskih frekvencija te pojave.

Pretvorbe individualnih indeksa možemo podijeliti na tri tipa:

1.) pretvorba indeksa na stalnoj bazi b u indekse na stalnoj bazi $b_1 \neq b$;
2.) pretvorba indeksa na stalnoj bazi b u verižne indekse;
3.) pretvorba verižnih indeksa u indekse na stalnoj bazi b.

Na po jednom primjeru opisat ćemo svaki pojedini tip.

<u>Primjer 6.</u> Indekse na stalnoj bazi 2005 = 100 navedene u Primjeru 5. pretvorimo u indekse na stalnoj bazi 2000 = 100. U tu svrhu koristimo sljedeću formulu:

$$I_t^* = \frac{I_t}{I_{b_1}} \cdot 100 \text{, za svaki } t = 0, 1, 2, \ldots, n-1,$$

pri čemu su:

I_t^* – indeks razdoblja t na "novoj" stalnoj bazi b_1;
I_t – indeks razdoblja t na "staroj" stalnoj bazi b;
I_{b_1} – indeks razdoblja b_1 na "staroj" stalnoj bazi b.

U ovome su primjeru b = 8 (jer 2005. godini odgovara t = 8) i b_1 = 3 (jer 2000. godini odgovara t = 3). Tako dobivamo:

$$I_0^* = \frac{I_0}{I_3} \cdot 100 = \frac{89}{95} \cdot 100 \approx 93.68421$$

$$I_1^* = \frac{I_1}{I_3} \cdot 100 = \frac{96}{95} \cdot 100 \approx 101.0526$$

$$I_2^* = \frac{I_2}{I_3} \cdot 100 = \frac{95}{95} \cdot 100 = 100$$

$$I_4^* = I_5^* = I_2^* = 100$$

$$I_6^* = \frac{I_6}{I_3} \cdot 100 = \frac{98}{95} \cdot 100 = 103.1579$$

$$I_7^* = \frac{I_7}{I_3} \cdot 100 = \frac{99}{95} \cdot 100 = 104.2105$$

$$I_8^* = \frac{I_8}{I_3} \cdot 100 = \frac{100}{95} \cdot 100 = 105.2632$$

$$I_9^* = \frac{I_9}{I_3} \cdot 100 = \frac{104}{95} \cdot 100 = 109.4737$$

Dobivene rezultate pregledno možemo prikazati u sljedećoj tablici:

godina	*t*	*indeks na stalnoj bazi* 2000 = 100
1997.	0	94
1998.	1	101
1999.	2	100
2000.	3	100
2001.	4	100
2002.	5	100
2003.	6	103
2004.	7	104
2005.	8	105
2006.	9	109

<u>Primjer 7.</u> Indekse na stalnoj bazi 2005 = 100 navedene u Primjeru 5. pretvorimo u verižne indekse. U tu svrhu koristimo sljedeću jednadžbu:

$$V_t = \frac{I_t}{I_{t-1}} \cdot 100 \text{, za svaki } t = 1, 2, \ldots, n-1.$$

Tako redom dobivamo:

$$V_1 = \frac{I_1}{I_0} \cdot 100 = \frac{96}{89} \cdot 100 \approx 118.5185$$

$$V_2 = \frac{I_2}{I_1} \cdot 100 = \frac{95}{96} \cdot 100 \approx 98.95833$$

$$V_3 = \frac{I_3}{I_2} \cdot 100 = \frac{95}{95} \cdot 100 = 100$$

$$V_4 = V_5 = V_3 = 100$$

$$V_6 = \frac{I_6}{I_5} \cdot 100 = \frac{98}{95} \cdot 100 \approx 103.1579$$

$$V_7 = \frac{I_7}{I_6} \cdot 100 = \frac{99}{98} \cdot 100 \approx 101.0204$$

$$V_8 = \frac{I_8}{I_7} \cdot 100 = \frac{100}{99} \cdot 100 \approx 101.0101$$

$$V_9 = \frac{I_9}{I_8} \cdot 100 = \frac{104}{100} \cdot 100 = 104$$

Dobivene rezultate pregledno možemo zapisati u sljedećoj tablici:

godina	*t*	*verižni indeks* (V_t)
1997.	0	*nepoznat*
1998.	1	119
1999.	2	99
2000.	3	100
2001.	4	100
2002.	5	100
2003.	6	103
2004.	7	101
2005.	8	101
2006.	9	104

Primjer 8. Zadani su verižni indeksi prosječne isplaćene neto-plaće po zaposlenome u svakom od mjeseci u 2007. godini:

mjesec	*t*	*verižni indeks* V_t
siječanj	0	*nepoznat*
veljača	1	96
ožujak	2	98.9
travanj	3	98.1
svibanj	4	100.7
lipanj	5	100.1
srpanj	6	100.3

mjesec	*t*	*verižni indeks* V_t
kolovoz	7	100.6
rujan	8	98.0
listopad	9	100.6
studeni	10	106.0
prosinac	11	102.4

izvor: Mjesečno statističko izvješće 3/08, Državni zavod za statistiku, Zagreb, travanj 2008.

Odredimo indekse na stalnoj bazi (rujan 2007. = 100), objasnimo značenje indeksa dobivenoga za prosinac 2007. godine i procijenimo prosječnu mjesečnu stopu promjene prosječne isplaćene neto-plaće. U ovom je slučaju $b = 8$ jer rujnu 2007. godine odgovara vrijednost vremenske varijable $t = 8$. Indeks I_t na stalnoj bazi (rujan 2007. = 100) računamo prema sljedećim jednakostima:

$$I_t = 100 \quad \text{za} \quad t = b;$$

$$I_{t-1} = \frac{I_t}{V_t} \cdot 100 \quad \text{za} \quad t < b;$$

$$I_t = \frac{I_{t-1} \cdot V_t}{100} \quad \text{za} \quad b < t.$$

Prigodom pretvorbe, dakle, krećemo od baznoga indeksa I_8. On je jednak $I_8 = 100$. Pomoću njega računamo najprije vrijednosti indeksa I_7, I_6, I_5, I_4, I_2, I_1 i I_0 koristeći drugu od triju gore navedenih jednakosti. Imamo redom:

$$I_7 = \frac{I_8}{V_8} \cdot 100 = \frac{100}{98.0} \cdot 100 \approx 102.04082$$

$$I_6 = \frac{I_7}{V_7} \cdot 100 = \frac{102.04082}{100.6} \cdot 100 \approx 101.43222$$

$$I_5 = \frac{I_6}{V_6} \cdot 100 = \frac{101.43222}{100.3} \cdot 100 \approx 101.12884$$

$$I_4 = \frac{I_5}{V_5} \cdot 100 = \frac{101.12884}{100.1} \cdot 100 \approx 101.02781$$

$$I_3 = \frac{I_4}{V_4} \cdot 100 = \frac{101.02781}{100.7} \cdot 100 \approx 100.32553$$

$$I_2 = \frac{I_3}{V_3} \cdot 100 = \frac{100.32553}{98.1} \cdot 100 \approx 102.26863$$

$$I_1 = \frac{I_2}{V_2} \cdot 100 = \frac{102.26863}{98.9} \cdot 100 \approx 103.4061$$

$$I_0 = \frac{I_1}{V_1} \cdot 100 = \frac{103.4061}{96} \cdot 100 \approx 107.71469$$

Vrijednosti indeksa I_9, I_{10} i I_{11} računamo pomoću posljednje od triju navedenih formula:

$$I_9 = \frac{I_8 \cdot V_9}{100} = \frac{100 \cdot 100.6}{100} \cdot 100 = 100.6$$

$$I_{10} = \frac{I_9 \cdot V_{10}}{100} = \frac{100.6 \cdot 106.0}{100} \cdot 100 = 106.636$$

$$I_{11} = \frac{I_{10} \cdot V_{11}}{100} = \frac{106.636 \cdot 102.4}{100} \approx 109.19526$$

Dobivene rezultate pregledno možemo zapisati u sljedećoj tablici:

mjesec	*t*	*indeks na stalnoj bazi* I_t (rujan = 100)
siječanj	0	107.71
veljača	1	103.41
ožujak	2	102.27
travanj	3	100.33
svibanj	4	101.03
lipanj	5	101.13
srpanj	6	101.43
kolovoz	7	102.04
rujan	8	100
listopad	9	100.6
studeni	10	106.64
prosinac	11	109.2

Vrijednost indeksa za prosinac jednaka je (približno) 109.2, što znači da *na svakih* 100 *kn prosječne isplaćene neto–plaće u rujnu* 2007. *godine dolazi po* 109.2 *kn prosječne isplaćene neto–plaće u prosincu iste godine*. Ekvivalentno možemo reći da je *prosječna isplaćena neto–plaća u prosincu* 2007. *godine bila za* 9.2% *veća nego u rujnu iste godine*.

Prosječna mjesečna promjena isplaćene mjesečne neto–plaće jednaka je:

$$\overline{S} = G - 100 = \sqrt[n-1]{V_1 \cdot V_2 \cdot \ldots \cdot V_{n-1}} - 100 \approx 0.1242\,,$$

što znači da su se u promatranom razdoblju *prosječne mjesečne neto–plaće povećavale za približno* 0.1242% *mjesečno*.

5.6. *Skupni indeksi*

Kako je istaknuto u točki 5.5., individualnim (verižnim ili baznim) indeksima prati se dinamika kretanja neke pojave iskazane *jednim* statističkim nizom. Međutim, često je potrebno pratiti kretanje barem dvije pojave koje, prema nekom kriteriju, pripadaju istoj skupini. Tako se npr. može pratiti kretanje proizvodnje svih vrsta proizvoda neke tvornice, kretanje prometa svih trgovina nekoga trgovačkoga lanca, kretanje cijena na malo osnovnih životnih potrepština itd. Takve se pojave predočavaju barem dvama statističkim nizovima, a prate se tzv. *skupnim indeksima*.

Problem računanja skupnih indeksa ne može se postaviti jednoznačno, tj. nema jedinstvenoga analitičkog izraza koji bi poslužio za računanje *svih* skupnih indeksa. U praksi se najčešće koriste *Laspeyresov*[124] i *Paascheov*[125] oblik skupnih indeksa. Prije nego li objasnimo spomenute indekse, uvedimo neke dogovorne oznake.

Razdoblje *u* kojem iskazujemo dinamiku promjene pojave nazivamo *tekuće* ili *izvještajno razdoblje*, a razdoblje u odnosu *na* koje se iskazuje dinamika pojava naziva se *bazno razdoblje*. Sukladno tome, definiramo sljedeće veličine:

p_0 := cijena u baznomu razdoblju;
p_1 := cijena u tekućemu razdoblju;
q_0 := količina u baznomu razdoblju;
q_1 := količina u tekućemu razdoblju.

Slova p i q su prva slova engleskih riječi *price* = cijena i *quantity* = količina. Brojčani indeks 0 (napisan u subskriptu) *uvijek* označava pojavu u *baznomu* razdoblju, dok brojčani indeks 1 *uvijek* označava pojavu u *tekućemu* razdoblju.

Istaknimo još da se *vrijednost* neke *robe* u određenomu razdoblju definira kao umnožak iznosa količine te robe i njezine cijene *po jedinici mjere količine* u istom razdoblju. Ako je npr. cijena 1 kg kruha jednaka 5 kn, onda je *vrijednost* 2 kg kruha jednaka:

$$2\ \text{kg} \cdot 5\ \frac{\text{kn}}{\text{kg}} = 10\ \text{kn}.$$

(jer je jedinica mjere količine 1 kg). Vrijednost neke robe obično se iskazuje u istoj jedinici mjere kao i odgovarajuća cijena te robe (po jedinici količine).

---

[124] Ernst Louis Étienne Laspeyres (1834. – 1913.), njemački ekonomist i statističar.
[125] Hermann Paasche (1851. – 1925.), njemački ekonomist i statističar.

5.6.1. *Skupni indeksi količina*

5.6.1.1. *Laspeyresov skupni indeks količina*

Laspeyresov skupni indeks količina (oznaka: $q_{01}(p_0)$) je pokazatelj relativne promjene ukupnih količina promatrane skupine od točno n pojava u tekućemu razdoblju u odnosu na ukupne količine te skupine pojava u baznomu razdoblju računajući uz *neizmijenjene cijene iz baznoga razdoblja*. (Dakle, fiksirana veličina pri izračunu ovoga skupnog indeksa jest p_0.)

Praktično najjednostavniji za izračun je tzv. *agregatni Laspeyresov skupni indeks količina* definiran formulom:

$$q_{01}(p_0) = \frac{\sum_{i=1}^{n} q_{i1} \cdot p_{i0}}{\sum_{i=1}^{n} q_{i0} \cdot p_{i0}} \cdot 100$$

U brojniku navedenoga izraza nalaze se *hipotetske* vrijednosti određenih količina robe proizvedenih u tekućem razdoblju, ali računate prema cijenama iz baznoga razdoblja. U nazivniku navedenoga izraza nalaze se stvarne vrijednosti količina iste robe (proizvedenih u baznom razdoblju) računate prema cijenama iz baznoga razdoblja. Budući da i u brojniku i u nazivniku imamo iste cijene, razlika između brojnika i nazivnika može biti isključivo posljedica promjene proizvedenih količina robe, a ni u kojem slučaju posljedica promjene cijena te robe. Stoga navedeni odnos mjeri smjer i jakost relativne promjene cijele skupine proizvedenih količina.

Analogno kao i individualne indekse, i Laspeyresov skupni indeks količina najlakše interpretiramo rabeći odgovarajuću stopu promjene:

$$S_q^{01}(p_0) = q_{01}(p_0) - 100\,[\%]$$

5.6.1.2. *Paascheov skupni indeks količina*

Paascheov skupni indeks količina (oznaka: $q_{01}(p_1)$) je pokazatelj relativne promjene ukupnih količina promatrane skupine od točno n pojava u tekućemu razdoblju u odnosu na ukupne količine te skupine pojava u baznomu razdoblju računajući uz *neizmijenjene cijene iz tekućega razdoblja*. (Dakle, fiksirana veličina pri računanju ovoga skupnog indeksa jest p_1.)

Praktično najjednostavniji za izračun je tzv. *agregatni Paascheov skupni indeks količina* definiran formulom:

$$q_{01}(p_1) = \frac{\sum_{i=1}^{n} q_{i1} \cdot p_{i1}}{\sum_{i=1}^{n} q_{i0} \cdot p_{i1}} \cdot 100 .$$

U brojniku navedenoga izraza nalaze se stvarne vrijednosti određenih količina robe proizvedenih u tekućem razdoblju i računate prema cijenama iz toga razdoblja. U nazivniku navedenoga izraza nalaze se *hipotetske* vrijednosti količina iste robe proizvedenih u baznom razdoblju, ali računate prema cijenama iz tekućega razdoblja. Budući da i u brojniku i u nazivniku imamo iste cijene, razlika između brojnika i nazivnika može biti isključivo posljedica promjene proizvedenih količina robe, a ni u kojem slučaju posljedica promjene cijena te robe. Stoga i navedeni odnos mjeri smjer i jakost relativne promjene cijele skupine proizvedenih količina robe.

Analogno kao i Laspeyresov skupni indeks količina, i Paascheov skupni indeks količina najlakše interpretiramo rabeći odgovarajuću stopu promjene:

$$S_q^{01}(p_1) = q_{01}(p_1) - 100\ [\%].$$

<u>Primjer 9.</u> Promatraju se prosječne prodajne cijene (PDV uračunat) i prodane količine četiri različite vrste proizvoda u trgovini "*Sve po malo kuna*" u Frkljevcima tijekom listopada i studenoga 2008. g. Podaci su navedeni u donjoj tablici.

vrsta proizvoda	*prosječna prodajna cijena u listopadu* 2008. [kn/kg]	*prodana količina u listopadu* 2008. [kg]	*prosječna prodajna cijena u studenom* 2008. [kn/kg]	*prodana količina u studenom* 2008. [kg]
brašno T-850	2,19	190	2,39	200
smeđa espresso kava	89	50	88	54
sitna morska sol	3,05	170	2,99	150
bijeli šećer	12,10	190	12,20	180

Izračunajmo Laspeyresov i Paascheov skupni indeks količina, pa objasnimo njihovo značenje. U ovome je primjeru *bazno* razdoblje listopad 2008., a *tekuće* (izvještajno) razdoblje studeni 2008. Stoga ćemo prosječnu prodajnu cijenu u listopadu označiti s p_0, prodanu količinu u listopadu s q_0, prosječnu prodajnu cijenu u studenom s p_1, a prodanu količinu u studenom s q_1. Laspeyresov skupni indeks količina jednak je:

$$q_{01}(p_0)=\frac{\sum_{i=1}^{4} q_{i1}\cdot p_{i0}}{\sum_{i=1}^{4} q_{i0}\cdot p_{i0}}\cdot 100=\left(\frac{200\cdot 2.19+54\cdot 89+150\cdot 3.05+180\cdot 12.10}{190\cdot 2.19+50\cdot 89+170\cdot 3.05+190\cdot 12.10}\right)\cdot 100=\frac{7879.50}{7683.60}\cdot 100$$

$$q_{01}(p_0)=102.54959$$

Pripadna stopa promjene iznosi

$$S_q^{01}(p_0)=q_{01}(p_0)-100 = 102.54959 - 100 = 2.54959.$$

Zaključujemo da su *prodane količine* svih promatranih vrstâ proizvoda *u studenom bile za* 2.54959% *veće od nego u listopadu računajući uz neizmijenjene cijene iz listopada*.

Paascheov skupni indeks količina jednak je

$$q_{01}(p_1)=\frac{\sum_{i=1}^{4} q_{i1}\cdot p_{i1}}{\sum_{i=1}^{4} q_{i0}\cdot p_{i1}}\cdot 100=\left(\frac{200\cdot 2.39+54\cdot 88+150\cdot 2.99+180\cdot 12.20}{190\cdot 2.39+50\cdot 88+170\cdot 2.99+190\cdot 12.20}\right)\cdot 100=\frac{7874.50}{7680.60}\cdot 100$$

$$q_{01}(p_1)=102.52721$$

Pripadna stopa promjene iznosi

$$S_q^{01}(p_1)=q_{01}(p_1)-100 = 102.52721 - 100 = 2.52721.$$

Zaključujemo da su *prodane količine* svih promatranih vrstâ proizvoda *u studenom bile za* 2.54959% *veće od nego u listopadu računajući uz neizmijenjene cijene iz studenoga*.

Dakle, za iste prodane količine dobivaju se različite vrijednosti skupnoga indeksa ako se u izračun uzmu cijene iz različitih razdoblja. Stoga skupni indeks *nije jednoznačna* precizna mjera relativnih promjena neke skupine pojava, nego isključivo mjera uz unaprijed zadane pretpostavke. U ovom su primjeru izračunati skupni indeksi prodanih količina promatranih vrsta proizvoda izračunati *uz pretpostavku* da su *odnosi vrijednosti* pojedinih vrsta proizvoda onakvi kakve izražavaju cijene u listopadu 2008.g., odnosno cijene u studenom 2008.g. Time se ne umanjuje značaj skupnih indeksa, nego se daje naglasak na obvezu njihova tumačenja uz uvažavanje pretpostavki pod kojima su izračunati.

5.6.1.3. *Fisherov skupni indeks količina*

U praksi se često koristi i *Fisherov*[126] *skupni indeks količina* ili *idealni skupni indeks količina* (oznaka: Q_{01}) Taj je indeks, po definiciji, jednak geometrijskoj sredini Laspeyresova i Paascheova skupnoga indeksa količina, tj.

[126] Irving Fisher (1867. – 1947.), američki ekonomist, jedan od tvoraca neoklasične ekonomije.,

$$Q_{01} = \sqrt{q_{01}(p_0) \cdot q_{01}(p_1)}\,.$$

Fisherov skupni indeks količina se najjednostavnije interpretira pomoću pripadne stope promjene. Ona se računa pomoću izraza:

$$S_Q^{01} = Q_{01} - 100\ [\%]$$

i predstavlja *prosječnu relativnu promjenu količina u tekućemu razdoblju u odnosu na količine u baznomu razdoblju*.

Primjer 10. Izračunajmo Fisherov skupni indeks količina na temelju podataka iz prethodnoga primjera i objasnimo njegovo značenje. Dobivamo:

$$Q_{01} = \sqrt{q_{01}(p_0) \cdot q_{01}(p_1)} = \sqrt{\frac{7879.50}{7683.60} \cdot \frac{7874.50}{7680.60}} \cdot 100 = 102.5384\,,$$

pa je pripadna stopa promjene

$$S_Q^{01} = Q_{01} - 100 = 102.5384 - 100 = 2.5384.$$

Zaključujemo da su *količine* svih promatranih vrstâ proizvoda *u studenom* 2008. *bile za prosječno* 2.5384% *veće nego u listopadu* 2008.

5.6.2. *Skupni indeksi cijena*

Skupni indeksi cijena mjere relativnu promjenu razine cijena za određeni asortiman robe, usluga itd. U praksi postoji više različitih vrsta skupnih indeksa cijena: skupni indeks cijena na malo, skupni indeks cijena na veliko, skupni indeks izvoznih cijena, skupni indeks cijena u turizmu itd. Odmah napomenimo da se niti jednim od tih indeksa ne mogu obuhvatiti cijene svih roba i usluga, pa se time nameću nimalo trivijalni problemi izbora proizvoda i(li) usluga koji će biti reprezentanti skupine za koju želimo izračunati skupne indekse cijena, problemi izbora cijene (kada i gdje treba prikupiti podatke o cijenama odabranih proizvoda i(li) usluga) itd. Sve to ukazuje na činjenicu da, iako skupne indekse nije teško izračunati, treba biti vrlo oprezan u njihovoj interpretaciji, a napose u poopćavanju zaključaka izvedenih pomoću odabranih reprezentanata na cijelu skupinu.

Dva najčešća skupna indeksa cijena su *Laspeyresov skupni indeks cijena* i *Paascheov skupni indeks cijena* koje ćemo upoznati u nastavku.

5.6.2.1. *Laspeyresov skupni indeks cijena*

<u>*Laspeyresov skupni indeks cijena*</u> (oznaka: $p_{01}(q_0)$) je pokazatelj relativne promjene ukupnih cijena promatrane skupine od točno n pojava u tekućemu razdoblju u odnosu na ukupne cijene te skupine pojava u baznomu razdoblju računajući uz *neizmijenjene količine iz baznoga razdoblja*. (Dakle, fiksirana veličina pri računanju ovoga skupnog indeksa jest q_0.)
Praktično najjednostavniji za izračun je tzv. *agregatni Laspeyresov skupni indeks cijena* definiran formulom

$$p_{01}(q_0) = \frac{\sum_{i=1}^{n} p_{i1} \cdot q_{i0}}{\sum_{i=1}^{n} p_{i0} \cdot q_{i0}} \cdot 100$$

Uočimo da se u brojniku navedenoga izraza nalaze *hipotetske* vrijednosti određenih količina robe proizvedenih u baznom razdoblju, ali računate prema cijenama iz tekućega razdoblja. U nazivniku navedenoga izraza nalaze se stvarne vrijednosti količina iste robe (proizvedenih u baznom razdoblju) računate prema cijenama iz baznoga razdoblja. Budući da i u brojniku i u nazivniku imamo iste količine, razlika između brojnika i nazivnika može biti isključivo posljedica promjene cijena promatrane robe, a ni u kojem slučaju posljedica promjene količina te robe. Stoga navedeni odnos mjeri smjer i jakost relativne promjene cijele skupine promatranih cijena.
Analogno kao i skupne indekse količina, i skupne indeks cijena najlakše interpretiramo rabeći odgovarajuće stope promjene. Za Laspeyresov skupni indeks cijena pripadajuća stopa promjene računa se prema formuli:

$$S_p^{01}(q_0) = p_{01}(q_0) - 100\ [\%]$$

5.6.2.2. *Paascheov skupni indeks cijena*

<u>*Paascheov skupni indeks cijena*</u> (oznaka: $p_{01}(q_1)$) je pokazatelj relativne promjene ukupnih cijena promatrane skupine od točno n pojava u tekućemu razdoblju u odnosu na ukupne cijene te skupine pojava u baznomu razdoblju računajući uz *neizmijenjene količine iz tekućega razdoblja*. (Dakle, fiksirana veličina pri računanju ovoga skupnoga indeksa jest q_1.)
Praktično najjednostavniji za izračun je tzv. *agregatni Paascheov skupni indeks cijena* definiran formulom:

$$p_{01}(q_1) = \frac{\sum_{i=1}^{n} p_{i1} \cdot q_{i1}}{\sum_{i=1}^{n} p_{i0} \cdot q_{i1}} \cdot 100$$

Uočimo da se u brojniku navedenoga izraza nalaze stvarne vrijednosti određenih količina robe proizvedenih u tekućem razdoblju računate prema cijenama iz tekućega razdoblja. U nazivniku navedenoga izraza nalaze se *hipotetske* vrijednosti količina iste robe (proizvedenih u tekućem razdoblju), ali računate prema cijenama iz baznoga razdoblja. Budući da i u brojniku i u nazivniku imamo iste količine, razlika između brojnika i nazivnika može biti isključivo posljedica promjene cijena promatrane robe, a ni u kojem slučaju posljedica promjene količina te robe. Stoga navedeni odnos mjeri smjer i jakost relativne promjene cijele skupine promatranih cijena.

I Paascheov skupni indeks cijena najlakše interpretiramo rabeći odgovarajuću stopu promjene. Ona se računa prema formuli:

$$S_p^{01}(q_1) = p_{01}(q_1) - 100\,[\%]$$

<u>Primjer 11.</u> Za prosječne prodajne cijene (PDV uračunat) i prodane količine četiri različite vrste proizvoda iz Primjera 9. izračunajmo Lasyperesov i Paascheov skupni indeks cijena, pa objasnimo njihovo značenje.

Podsjetimo da je *bazno* razdoblje listopad 2008., a *tekuće* (izvještajno) razdoblje studeni 2008. Stoga prosječnu prodajnu cijenu u listopadu označavamo s p_0, prodanu količinu u listopadu s q_0, prosječnu prodajnu cijenu u studenom s p_1, a prodanu količinu u studenom s q_1. Laspeyresov skupni indeks cijena jednak je:

$$p_{01}(q_0) = \frac{\sum_{i=1}^{4} p_{i1} \cdot q_{i0}}{\sum_{i=1}^{4} p_{i0} \cdot q_{i0}} \cdot 100 = \left(\frac{2.39 \cdot 190 + 88 \cdot 50 + 2.99 \cdot 170 + 12.2 \cdot 190}{2.19 \cdot 190 + 89 \cdot 50 + 3.05 \cdot 170 + 12.1 \cdot 190} \right) \cdot 100 = \frac{7680.40}{7683.60} \cdot 100$$

$$p_{01}(q_0) = 99.958353$$

Pripadna stopa promjene iznosi

$$S_p^{01}(q_0) = p_{01}(q_0) - 100 = 99.958353 - 100 = -0.041647$$

Zaključujemo da su *prosječne prodajne cijene* svih promatranih vrstâ proizvoda *u studenom bile za* približno 0.042% *manje nego u listopadu računajući uz neizmijenjene prodane količine iz listopada.*

Paascheov skupni indeks cijena jednak je

$$p_{01}(q_1) = \frac{\sum_{i=1}^{4} p_{i1} \cdot q_{i1}}{\sum_{i=1}^{4} p_{i0} \cdot q_{i1}} \cdot 100 = \left(\frac{2.39 \cdot 200 + 88 \cdot 54 + 2.99 \cdot 150 + 12.2 \cdot 180}{2.19 \cdot 200 + 89 \cdot 54 + 3.05 \cdot 150 + 12.1 \cdot 180} \right) \cdot 100 = \frac{7874.50}{7879.50} \cdot 100$$

$$p_{01}(q_1) = 99.936544$$

Pripadna stopa promjene iznosi

$$S_p^{01}(q_1) = p_{01}(q_1) - 100 = 99.936544 - 100 = -0.063456$$

Dakle, *prosječne prodajne cijene* svih promatranih vrstâ proizvoda *u studenome bile su za* približno 0,063% *manje nego u listopadu računajući uz neizmijenjene prodane količine iz studenoga.*

5.6.2.3. *Fisherov skupni indeks cijena*

Analogno Fisherovu skupnu indeksu količina, definira se *Fisherov skupni indeks cijena* ili *idealni skupni indeks cijena* (oznaka: P_{01}) Taj je indeks, po definiciji, jednak geometrijskoj sredini Laspeyresova i Paascheova skupnoga indeksa cijena, tj.

$$P_{01} = \sqrt{p_{01}(q_0) \cdot p_{01}(q_1)}\,.$$

Fisherov skupni indeks količina se najjednostavnije interpretira pomoću pripadne stope promjene. Ona se računa pomoću izraza:

$$S_P^{01} = P_{01} - 100\,[\%]$$

i predstavlja *prosječnu relativnu promjenu cijena u tekućemu razdoblju u odnosu na cijene u baznomu razdoblju.*

<u>Primjer 12.</u> Izračunajmo Fisherov skupni indeks cijena na temelju podataka iz Primjera 9. i objasnimo njegovo značenje. Dobivamo:

$$P_{01} = \sqrt{p_{01}(q_0) \cdot p_{01}(q_1)} = \sqrt{\frac{7680.40}{7683.60} \cdot \frac{7874.50}{7879.50}} \cdot 100 = 99.94745\,,$$

pa je pripadna stopa promjene

$$S_P^{01} = P_{01} - 100 = 99.94745 - 100 = -0.52555.$$

Zaključujemo da su *cijene* svih promatranih vrstâ proizvoda *u studenom* 2008. *bile za prosječno* 0.52555% *manje nego u listopadu* 2008.

5.6.3. *Skupni indeksi vrijednosti*

Skupni indeks vrijednosti (oznaka: V_{01}) je pokazatelj relativne promjene ukupne vrijednosti promatrane skupine od točno n pojava u tekućemu razdoblju u odnosu na ukupnu vrijednost te skupine pojava u baznomu razdoblju. Računa se prema formuli:

$$V_{01} = \frac{\sum_{i=1}^{n} p_{i1} \cdot q_{i1}}{\sum_{i=1}^{n} p_{i0} \cdot q_{i0}} \cdot 100$$

Kao i svi ostali indeksi, najlakše ga je interpretirati pomoću pripadajuće stope promjene. Ona se računa prema formuli:

$$S_V^{01} = V_{01} - 100 \ [\%].$$

Napomenimo da se skupni indeks vrijednosti može izračunati i kao umnožak Laspeyresova indeksa cijena i Paascheova indeksa količina podijeljenoga sa 100, odnosno Paascheova indeksa cijena i Laspeyresova indeksa količina podijeljenoga sa 100. Točnije, vrijede sljedeće jednakosti:

$$V_{01} = \frac{p_{01}(q_0) \cdot q_{01}(p_1)}{100} = \frac{p_{01}(q_1) \cdot q_{01}(p_0)}{100}$$

Također, skupni indeks vrijednosti jednak je umnošku Fisherova skupnoga indeksa cijena i Fisherova skupnoga indeksa količina podijeljenoga sa 100, tj. vrijedi jednakost:

$$V_{01} = \frac{P_{01} \cdot Q_{01}}{100}.$$

Primjer 13. Za promatrane vrste proizvoda iz Primjera 9. izračunajmo skupni indeks vrijednosti i objasnimo njegovo značenje. Radi točnosti izračuna koristit ćemo originalnu formulu za skupni indeks vrijednosti, a potom ćemo provjeriti dobiveni rezultat koristeći netom navedene jednakosti. Imamo redom:

$$V_{01} = \frac{\sum_{i=1}^{4} p_{i1} \cdot q_{i1}}{\sum_{i=1}^{4} p_{i0} \cdot q_{i0}} \cdot 100 = \left(\frac{2.39 \cdot 200 + 88 \cdot 54 + 2.99 \cdot 150 + 12.2 \cdot 180}{2.19 \cdot 190 + 89 \cdot 50 + 3.05 \cdot 170 + 12.1 \cdot 190} \right) \cdot 100 = \frac{7874.50}{7683.60} \cdot 100$$

$$V_{01} = 102.484512$$

Pripadna stopa promjene iznosi

$$S_V^{01} = V_{01} - 100 = 102.484512 - 100 = 2.484512.$$

Zaključujemo da je *ukupna vrijednost promatrane skupine robâ u studenom* 2008. *bila za* približno 2.48% *veća nego u listopadu* 2008.

Izračun skupnoga indeksa vrijednosti pomoću Laspeyresovih i Paascheovih, odnosno Fisherovih indeksa daje sljedeće rezultate:

$$V_{01} = \frac{p_{01}(q_0) \cdot q_{01}(p_1)}{100} = \frac{99.958353 \cdot 102.52721}{100} = 102.48451$$

$$V_{01} = \frac{p_{01}(q_1) \cdot q_{01}(p_0)}{100} = \frac{99.936544 \cdot 102.54959}{100} = 102.48442$$

$$V_{01} = \frac{P_{01} \cdot Q_{01}}{100} = \frac{99.94745 \cdot 102.5384}{100} = 102.48452$$

Za postizanje veće točnosti u izračunima odgovarajuće skupne indekse trebali bismo izračunati na više decimalnih mjesta. Najtočnija je svakako vrijednost dobivena iz originalne formule za izračun skupnoga indeksa vrijednosti jer sadrži najmanje podataka dobivenih približnim izračunima pa tu formulu treba primjenjivati kad god je to moguće.

Primjer 14. U sljedećoj su tablici navedeni podaci o broju ostvarenih noćenja i prosječnim cijenama jednoga noćenja u gradu Ždralovgradu u tri različita sektora za 2007. i 2008. godinu.

vrsta usluge	*prosječna cijena jednoga noćenja u* 2007. *g.* [€]	*ostvareni broj noćenja u* 2007. g. [000]	*prosječna cijena jednoga noćenja u* 2008. *g.* [€]	*ostvareni broj noćenja u* 2008. g. [000]
hoteli	80	32	90	30
kampovi	29	12	25	13
privatni smještaj	60	44	59	51

Odredimo:

a) Za koliko su se ukupno promijenile prosječne cijene jednoga noćenja u 2008. godini u odnosu na 2007. godinu računajući uz neizmijenjen ostvaren broj noćenja u 2007. godini?
b) Za koliko se ukupno promijenio ostvaren broj noćenja u 2008. godini u odnosu na 2007. godinu računajući uz neizmijenjene prosječne cijene jednoga noćenja u 2008. godini?
c) Za koliko su se *prosječno* promijenile prosječne cijene jednoga noćenja u 2008. godini u odnosu na 2007. godinu?
d) Je li ostvarena prognoza gradonačelnika Ždralovgrada Mutimira Lupiglupostića da će ukupna vrijednost svih noćenja ostvarenih u 2008. godini biti za 10% veća nego u prethodnoj godini? Objasnimo svoj odgovor.

Najprije primijetimo da je u ovome primjeru 2007. godina bazna godina, a 2008. godina tekuća ili izvještajna godina. Stoga su veličine u drugomu stupcu cijene u baznoj godini, pa ćemo ih označiti s p_0. Na isti način označavamo i ostale stupce: u trećemu su stupcu vrijednosti količina u baznoj godini koje označavamo s q_0, u četvrtomu stupcu vrijednosti veličine p_1, a u posljednjemu stupcu vrijednosti veličine q_1. Stoga početnu tablicu možemo zamijeniti sljedećom:

vrsta usluge	p_0	q_0	p_1	q_1
hoteli	80	32	90	30
kampovi	29	12	25	13
privatni smještaj	60	44	59	51

a) U ovome je slučaju neizmijenjena veličina *broj noćenja u* 2007. *godini* odnosno, sukladno gornjim oznakama, veličina q_0. Promatramo relativnu promjenu cijena uz pretpostavku o neizmijenjenoj veličini q_0, pa računamo *Laspeyresov skupni indeks cijena* ($p_{01}(q_0)$):

$$p_{01}(q_0) = \frac{\sum_{i=1}^{3} p_{i1} \cdot q_{i0}}{\sum_{i=1}^{3} p_{i0} \cdot q_{i0}} \cdot 100 = \left(\frac{90 \cdot 32 + 25 \cdot 12 + 59 \cdot 44}{80 \cdot 32 + 29 \cdot 12 + 60 \cdot 44} \right) \cdot 100 = \frac{5776}{5548} \cdot 100 = 104.10959$$

Pripadna stopa promjene iznosi

$$S_p^{01}(q_0) = p_{01}(q_0) - 100 = 104.10959 - 100 = 4.10959.$$

Stoga zaključujemo: *Računajući uz neizmijenjen ostvareni broj noćenja u* 2007. *godini*, *prosječne cijene jednoga noćenja u* 2008. *godini bile su za* približno 4.11% *nego u* 2007. *godini.*

b) U ovome je slučaju neizmijenjena veličina *prosječna cijena jednoga noćenja u* 2008. *godini*, tj., sukladno ranijim oznakama, veličina p_1. Promatramo relativnu promjenu ostvarenih noćenja uz pretpostavku o fiksiranoj veličini p_1, pa računamo *Paascheov skupni indeks količina* ($q_{01}(p_1)$):

$$q_{01}(p_1) = \frac{\sum_{i=1}^{3} q_{i1} \cdot p_{i1}}{\sum_{i=1}^{3} q_{i0} \cdot p_{i1}} \cdot 100 = \left(\frac{30 \cdot 90 + 13 \cdot 25 + 51 \cdot 59}{32 \cdot 90 + 12 \cdot 25 + 44 \cdot 59} \right) \cdot 100 = \frac{6034}{5776} \cdot 100 = 104.46676$$

Pripadna stopa promjene iznosi

$$S_q^{01}(p_1) = q_{01}(p_1) - 100 = 4.46676.$$

Stoga zaključujemo: *Računajući uz neizmijenjene prosječne cijene jednoga noćenja u* 2008. *godini*, *ostvareni broj noćenja u* 2008. *godini bio je za* 4.46676% *veći nego u* 2007. *godini.*

c) U ovome slučaju računamo prosječnu promjenu cijena u tekućem razdoblju u odnosu na cijene u baznom razdoblju. U tu ćemo svrhu izračunati *Fisherov skupni indeks cijena*, pa će rješenje biti tom indeksu pripadajuća stopa promjene. Fisherov skupni indeks cijena jednak je:

$$P_{01} = \sqrt{p_{01}(q_0) \cdot p_{01}(q_1)} = \sqrt{\frac{90 \cdot 32 + 25 \cdot 12 + 59 \cdot 44}{80 \cdot 32 + 29 \cdot 12 + 60 \cdot 44} \cdot \frac{90 \cdot 30 + 25 \cdot 13 + 59 \cdot 51}{80 \cdot 30 + 29 \cdot 13 + 60 \cdot 51}} \cdot 100$$

$$P_{01} = 103.7417$$

a njemu pripadna stopa promjene

$$S_P^{01} = P_{01} - 100 = 103.7417 - 100 = 3.7417.$$

Dakle, *prosječne cijene jednoga noćenja u* 2008. *godini bile su za prosječno* 3.7417% *veće nego u* 2007. *godini.*

d) U ovome slučaju uspoređujemo stopu promjene koja odgovara skupnom indeksu vrijednosti s očekivanom stopom promjene. Najprije računamo skupni indeks vrijednosti:

$$V_{01} = \frac{\sum_{i=1}^{3} p_{i1} q_{i1}}{\sum_{i=1}^{3} p_{i0} q_{i0}} \cdot 100 = \left(\frac{90 \cdot 30 + 25 \cdot 13 + 59 \cdot 51}{80 \cdot 32 + 29 \cdot 12 + 60 \cdot 44} \right) \cdot 100 = \frac{6034}{5548} \cdot 100 = 108.76,$$

pa je pripadna stopa promjene

$$S_V^{01} = V_{01} - 100 = 108.76 - 100 = 8.76$$

Stoga možemo zaključiti: *Ukupna vrijednost noćenja u* 2008. *godini je za* 8.76% *veća nego u* 2007. *godini*. Dakle, gradonačelnikova prognoza se *nije* ostvarila[127].

[127] Gradonačelnikova prognoza ne bi bila istinita niti da smo skupni indeks V_{01} računali kao količnik umnoška indeksa $p_{01}(q_0) \cdot q_{01}(p_1)$ i broja 100. Provjerite to.

5.6.4. *Neki posebni oblici skupnih indeksa i njihova primjena*

U određenim ekonomskim analizama često se koriste različiti oblici individualnih i skupnih indeksa. Tako se npr. skupni indeks cijena može računati za cijene na malo, cijene na veliko, cijene poljoprivrednih proizvoda, cijene tzv. "sindikalne košarice" itd. Pogledajmo to na sljedećim primjerima.

5.6.4.1. *Izračunavanje realnih plaća na osnovu skupnih indeksa troškova života*

Jedan od najvažnijih oblika *skupnih indeksa cijena* je *indeks potrošačkih cijena*[128]. Indeks potrošačkih cijena odražava promjene na razini cijena dobara i usluga koje u tijeku vremena nabavlja, koristi se njima ili plaća referentno stanovništvo radi osobne potrošnje, te je jedinstvena i opća mjera inflacije u Republici Hrvatskoj. On je samo mjera kretanja cijena dobara i usluga osobne potrošnje, a ne stvarnih promjena razlike i strukture osobne potrošnje, za koje se podaci prikupljaju posebnim anketama. Izračunava se i objavljuje na osnovi oko 590 reprezentativnih dobara i usluga. Svakog mjeseca prikupi se više od 27 000 cijena na unaprijed definiranu uzorku prodajnih mjesta na unaprijed definiranu uzorku prodajnih mjesta na devet lokacija (gradova) u Republici Hrvatskoj (Zagreb, Slavonski Brod, Osijek, Sisak, Rijeka, Pula, Split, Dubrovnik i Varaždin), odabranih prema kriteriju broja stanovnika i reprezentativnosti za pojedinu statističku regiju. Za klasificiranje izdataka koristi se Klasifikacija osobne potrošnje prema namjeni (COICOP) koja izdatke dijeli u sljedećih 12 osnovnih kategorija dobara i usluga:

- prehrana i bezalkoholna pića;
- alkoholna pića i duhan;
- odjeća i obuća;
- stanovanje, voda, energija, plin i druga goriva;
- pokućstvo, oprema za kuću i redovito održavanje;
- zdravstvo;
- promet;
- komunikacije;
- rekreacija i kultura;
- obrazovanje;
- ugostiteljske usluge;
- ostala dobra i usluge[129].

Indeks potrošačkih cijena služi ponajprije za *mjerenje inflacije* (porasta cijena) *u privredi*, a može se koristiti i u druge svrhe (npr. za očuvanje vrijednosti kod ugovora s indeksnim klauzulama, deflacioniranje (uklanjanje utjecaja inflacije) određenih vrijednosnih pokazatelja itd.).

---

[128] Sve do 2004. godine umjesto indeksa potrošačkih cijena računao se tzv. indeks troškova života koji je imao istu namjenu kao i indeks potrošačkih cijena. Od 2004. godine u statističkim se izvješćima primjenjuje indeks potrošačkih cijena, a u svrhu usklađivanja hrvatskih statističkih izvješća sa svjetskim izvješćima i pokazateljima.
[129] Definicije navedene u prvom dijelu ove točke preuzete su iz *Statističkoga ljetopisa Republike Hrvatske za* 2008. *godinu*, DZS, Zagreb, 2009.

Pomoću indeksa potrošačkih cijena mjeri se utjecaj potrošačkih cijena na *nominalne plaće*. Tako se dobiju *realne plaće* čije kretanje služi za opis kretanja životnoga standarda stanovništva. Realne plaće i pripane indekse realnih plaća kao pokazatelje njihove dinamike kretanja izračunavamo prema formulama:

$$RP = \frac{NP}{IPC} \cdot 100$$

$$IRP = \frac{INP}{IPC} \cdot 100$$

pri čemu su:

RP – iznos realne plaće;
NP – iznos nominalne plaće;
IPC – indeks potrošačkih cijena na stalnoj bazi;
IRP – indeks realne plaće na stalnoj bazi;
INP – indeks nominalne plaće na stalnoj bazi.

Istaknimo da se *svi* ovdje navedeni *indeksi na stalnoj bazi obavezno moraju odnositi na isto bazno razdoblje* (tj. moraju se odnositi na razdoblje u odnosu na čije potrošačke cijene računamo iznose realnih plaća).

Primjer 15. Prosječne mjesečne neto-plaće u Republici Hrvatskoj i verižni indeksi potrošačkih cijena za razdoblje od rujna 2008. do siječnja 2009. godine navedeni su u donjoj tablici.

Razdoblje	*Prosječna mjesečna neto-plaća* [kn]	*Verižni indeks potrošačkih cijena*
rujan 2008.	5.203	———
listopad 2008.	5.263	100.0879507
studeni 2008.	5.397	99.73637961
prosinac 2008.	5.410	99.38325991
siječanj 2009.	5.307	101.2411348

Izračunajmo realne plaće i pripadne indekse realnih plaća na stalnoj bazi prosinac 2008. = 100. Objasnimo vrijednosti izračunate za listopad 2008. godine i odredimo ukupnu relativnu promjenu realnih plaća u promatranom razdoblju.

Verižne indekse potrošačkih cijena najprije moramo pretvoriti u indekse potrošačkih cijena na stalnoj bazi (uzimajući prosinac 2008. godine a bazno razdoblje, tj. *b* = prosinac 2008.). Koristimo formule:

$$I_t = 100 \quad \text{za } t = \text{ prosinac 2008.};$$

$$I_{t-1} = \frac{I_t}{V_t} \cdot 100 \quad \text{za } t < \text{prosinac 2008.};$$

$$I_t = \frac{I_{t-1} \cdot V_t}{100} \quad \text{za prosinac 2008.} < t.$$

Tako najprije imamo:

$$I_{prosinac} = 100,$$

pa

$$I_{studeni} = \frac{I_{\text{prosinac}}}{V_{\text{prosinac}}} \cdot 100 = \frac{100}{99.38326} \cdot 100 \approx 100.6206$$

$$I_{listopad} = \frac{I_{studeni}}{V_{studeni}} \cdot 100 = \frac{100,6206}{99.73638} \cdot 100 \approx 100.8865\,,$$

$$I_{rujan} = \frac{I_{listopad}}{V_{listopad}} \cdot 100 = \frac{100.8865}{100.088} \cdot 100 \approx 100.7979$$

dok za siječanj 2009. dobivamo:

$$I_{siječanj} = \frac{I_{\text{prosinac}} \cdot V_{siječanj}}{100} = \frac{100 \cdot 101.2411}{100} = 101.2411\,.$$

(To smo mogli zaključiti i bez računanja. Zašto?) Dobivene podatke pregledno zapišimo u posljednji stupac sljedeće tablice:

Razdoblje	*Prosječna mjesečna neto-plaća* [kn]	*Verižni indeks potrošačkih cijena*	*Indeks potrošačkih cijena* (prosinac 2008. = 100)
rujan 2008.	5.203	———	100.7978724
listopad 2008.	5.263	100.0879507	100.8865248
studeni 2008.	5.397	99.73637961	100.6205674
prosinac 2008.	5.410	99.38325991	100
siječanj 2009.	5.307	101.2411348	101.2411348

Realne plaće (uz prosinac 2008. kao bazno razdoblje) računamo prema formuli:

$$RP = \frac{\text{prosječna mjesečna neto-plaća}}{\text{indeks potrošačkih cijena na stalnoj bazi (prosinac 2008. = 100)}} \cdot 100\,.$$

Tako redom imamo:

$$RP_{rujan} = \frac{5.203}{100.7979} \cdot 100 \approx 5.161,82$$

$$RP_{listopad} = \frac{5.263}{100.8865} \cdot 100 \approx 5.216,75$$

$$RP_{studeni} = \frac{5.397}{100.6206} \cdot 100 \approx 5.363,72$$

$$RP_{\text{prosinac}} = \text{prosječna mjesečna netoplaća za prosinac} = 5.410$$

$$RP_{siječanj} = \frac{5.307}{101.2411} \cdot 100 \approx 5.241,94$$

Potom računamo indeks realnih plaća (*IRP)* ponovno uzimajući prosinac 2008. godine kao bazno razdoblje. Taj indeks možemo izračunati izravno iz gore navedenih realnih plaća kao:

$$IRP = \frac{\text{realna plaća u tekućem razdoblju}}{\text{realna plaća u prosincu 2008.}} \cdot 100$$

pa dobivamo:

$$IRP_{rujan} = \frac{5.161,82}{5.410} \cdot 100 \approx 95.41248$$

$$IRP_{listopad} = \frac{5.216,75}{5.410} \cdot 100 \approx 96.42795$$

$$IRP_{studeni} = \frac{5.363,72}{5.410} \cdot 100 \approx 99.14445$$

$$IRP_{\text{prosinac}} = 100$$

$$IRP_{siječanj} = \frac{5.241,94}{5.410} \cdot 100 \approx 96.89354$$

Sve dobivene podatke možemo pregledno zapisati tablično:

Razdoblje	*Prosječna mjesečna neto-plaća* [kn]	*Verižni indeks potrošačkih cijena*	*Iindeks potrošačkih cijena* (prosinac 2008. = 100)	*Realne plaće* (bazno razdoblje: prosinac 2008.)	*Indeks realnih plaća* (prosinac 2008. = 100)
rujan 2008.	5.203	———	100.7978724	5.161,82	95.41248245
listopad 2008.	5.263	100.0879507	100.8865248	5.216,75	96.42795188
studeni 2008.	5.397	99.73637961	100.6205674	5.363,71	99.14444616
prosinac 2008.	5.410	99.38325991	100	5.410	100
siječanj 2009.	5.307	101.2411348	101.2411348	5.241,94	96.89353887

Iz tablice vidimo da su *realne plaće u listopadu* 2008. *godine*, *uz potrošačke cijene iz prosinca* 2008. *godine*, *iznosile* 5.216,75 kn. Pripadni indeks realnih plaća interpretiramo pomoću odgovarajuće stope promjene:

$$S_t = IRP - 100.$$

Stoga imamo:

$$S_{\text{listopad}} = IRP_{\text{listopad}} - 100 = 96.42795188 - 100 = -3.57204812.$$

Zaključujemo da su, *uz potrošačke cijene iz prosinca* 2008. *godine*, *u listopadu* 2008. *godine realne plaće bile za približno* 3.57% *manje nego u prosincu iste godine*.

Ukupnu relativnu promjenu, tj. ukupan postotak promjene realnih plaća u promatranom razdoblju možemo izračunati na dva različita načina:

1.) određivanjem omjera posljednje i prve vrijednosti realne plaće uz potrošačke cijene iz prosinca 2008. godine:

$$P = \frac{Y_{siječanj}}{Y_{rujan}} \cdot 100 - 100$$

$$P = \frac{5.241,94}{5.161,82} \cdot 100 - 100$$

$$P \approx 1.552267$$

2.) određivanjem omjera posljednjeg i prvog indeksa realnih plaća (uz prosinac 2008. godine kao bazno razdoblje):

$$P = \frac{IRP_{siječanj}}{IRP_{rujan}} \cdot 100 - 100$$

$$P = \frac{96.89354}{95.41248} \cdot 100 - 100$$

$$P \approx 1.552267$$

Zaključujemo da su se, *uz potrošačke cijene iz prosinca* 2008. *godine*, u promatranom razdoblju *realne plaće ukupno povećale za približno* 1.55%.

<u>*Napomena*</u>: Apsolutna vrijednost realnih plaća ovisi o razdoblju za kojega su vezane potrošačke cijene (tj. za različita bazna razdoblja dobivamo različite vrijednosti realnih plaća), pa se prigodom njihova računanja <u>obvezatno</u> mora navesti uz koje su bazno razdoblje vezane potrošačke cijene (upravo kako smo to učinili u prethodnomu primjeru). No, prigodom računanja indeksa ili odgovarajućih stopa promjene nije bitno uz koje su razdoblje vezane potrošačke cijene jer vrijednost indeksa (odnosno, stope promjene) <u>ne</u> ovisi o izboru razdoblja za kojega vežemo potrošačke cijene. U tome je prednost uporabe pokazatelja *relativnih* promjena.

5.6.4.2. *Vrijednost u stalnim cijenama i indeks fizičkoga obujma*

Jedna od vrlo čestih pojava u ekonomskim analizama jest tzv. *deflacija*. Tim se postupkom uklanja utjecaj promjena cijena na vrijednosno izražene pojave, a provodi se tako da se nominalne vrijednosti (vrijednosti izražene u tekućim cijenama) podijele s odgovarajućim skupnim indeksom cijena (kojega nazivamo *deflator*). Njegova obilježja općenito ovise o strukturi pojave čije vrijednosti deflacioniramo. Potreba za deflacijom osobito se javlja pri inflacijskim kretanjima u kojima promjene cijena onemogućuju uvid u stvarni razvoj promatranih pojava (npr. količina proizvoda, količina usluga itd.). Deflacionirane vrijednosti nazivamo i *vrijednosti u stalnim cijenama*, a općenito ih određujemo pomoću formule

$$V_{SC} = \frac{V_{TC}}{IC} \cdot 100$$

gdje su:

V_{SC} – vrijednost iskazana u stalnim cijenama;
V_{TC} – vrijednosti iskazana u tekućim cijenama;
IC – indeks cijena na stalnoj bazi.

Tipični primjeri deflatora su ranije spomenuti indeks potrošačkih cijena i tzv. <u>*indeks fizičkoga obujma*</u>. Taj je indeks poseban oblik *skupnoga indeksa količina*, a računamo ga pomoću izraza:

$$IFO_t = \frac{(V_{SC})_t}{(V_{SC})_b} \cdot 100 \text{, za } t = 0, 1, \ldots, n-1$$

gdje su:

$(IFO)_t$ – indeks fizičkoga obujma za razdoblje t;
$(V_{SC})_t$ – vrijednost u stalnim cijenama za razdoblje t;
$(V_{SC})_b$ – vrijednost u stalnim cijenama za bazno razdoblje b.

Indeks fizičkoga obujma rabimo, primjerice, kada želimo uspoređivati bruto domaći proizvod u određenoj godini (iskazan u stalnim cijenama) s bruto domaćim proizvodom u nekoj odabranoj baznoj godini.

Pokažimo primjenu ovih vrsta indeksa na primjeru.

<u>Primjer 16.</u> U donjoj su tablici navedeni podaci o proizvodnji u tekućim cijenama, indeksu cijena na stalnoj bazi 2005 = 100 i proizvodnji u stalnim cijenama iz 2005. godine.

Godina	*Verižni indeks cijena*	*Proizvodnja u tekućim cijenama* [000 000 €]	*Indeks cijena na stalnoj bazi* (2005 = 100)	*Proizvodnja u stalnim cijenama iz* 2005. *godine* [000 000 €]	*Indeks fizičkoga obujma* (2005 = 100)
1997.			69.76990634	0,90296810	
1998.		0.75	74.65379979		
1999.			76.89341378	1,14444132	
2000.		0.96	77.66234792		
2001.			82.32208879	1,29977266	
2002.		1.23	85.61497235		
2003.			88.18342152	1,54224000	
2004.		1.52	92.59259259		
2005.			100	1,85000000	
2006.		1,98	110		
2007.			118,8	2,11279461	
2008.		2,75	125.928		

Dopunimo navedenu tablicu podacima koji nedostaju, pa potom pomoću reprezentativnijega od dvaju modela trenda (jednostavni linearni i jednostavni eksponencijalni model trenda) prognozirajmo kretanje proizvodnje (iskazane u stalnim cijenama iz 2005. godine) u razdoblju od 2009. do 2013. godine. Zaključno, odredimo ukupnu relativnu promjenu i prosječnu godišnju relativnu promjenu cijena u promatranom razdoblju, te očekivanu ukupnu relativnu promjenu i očekivanu prosječnu godišnju relativnu promjenu proizvodnje (iskazane u stalnim cijenama iz 2005. godine) u razdoblju od 2009. do 2013. godine.

Izračunajmo najprije verižne indekse cijena koristeći zadane indekse na stalnoj bazi 2005 = 100. Podsjetimo da se indeksi na stalnoj bazi pretvaraju u verižne indekse pomoću relacije:

$$V_t = \frac{I_t}{I_{t-1}} \cdot 100 \text{, za svaki } t = 1, 2, \ldots, n-1.$$

Pritom 1997. godini odgovara vrijednost vremenske varijable $t = 0$, 1998. godini vrijednost vremenske varijable $t = 1$ itd. Verižni indeks cijena za 1997. godinu nije moguće izračunati, pa ćemo u odgovarajući stupac tablice upisati riječ *nepoznat*. Za ostale godine dobivamo:

$$V_1 = \frac{IC_1}{IC_0} \cdot 100 = \frac{74.65379979}{69.76990634} \cdot 100 = 107;$$

$$V_2 = \frac{IC_2}{IC_1} \cdot 100 = \frac{76.89341378}{74.65379979} \cdot 100 = 103;$$

$$V_3 = \frac{IC_3}{IC_2} \cdot 100 = \frac{77.66234792}{76.89341378} \cdot 100 = 101;$$

itd. Tako drugi stupac zadane tablice izgleda ovako:

Verižni indeks cijena
nepoznat
107
103
101
106
104
103
105
108
110
108
106

Nepoznate proizvodnje u tekućim cijenama izračunat ćemo koristeći relaciju

$$V_{SC} = \frac{V_{TC}}{IC} \cdot 100$$

iz koje je

$$V_{TC} = \frac{V_{SC} \cdot IC}{100}.$$

Tako je vrijednost proizvodnje u tekućim cijenama za 1997. godinu jednaka

$$(V_{TC})_0 = \frac{(V_{SC})_0 \cdot IC_0}{100} = \frac{0.90296810 \cdot 69.76990634}{100} = 0.63,$$

vrijednost proizvodnje u tekućim cijenama za 1999. godinu

$$(V_{TC})_2 = \frac{(V_{SC})_2 \cdot IC_2}{100} = \frac{1.14444132 \cdot 76.89341378}{100} = 0.88$$

itd. Stoga treći stupac zadane tablice izgleda ovako:

Proizvodnja u tekućim cijenama [000 000 €]
0.63
0.75
0.88
0.96
1.07

Proizvodnja u tekućim cijenama [000 000 €]
1.23
1.36
1.52
1.85
1.98
2.51
2.75

Peti stupac tablice popunjavamo koristeći definicijsku relaciju za vrijednost u stalnim cijenama:

$$V_{SC} = \frac{V_{TC}}{IC} \cdot 100 .$$

Tako se dobiva:

$$(V_{SC})_{1998} = \frac{(V_{TC})_{1998}}{IC_{1998}} \cdot 100 = \frac{0.75}{74.65379979} \cdot 100 = 1.00463741;$$

$$(V_{SC})_{2000} = \frac{(V_{TC})_{2000}}{IC_{2000}} \cdot 100 = \frac{0.96}{77.66234792} \cdot 100 = 1.23612024;$$

$$(V_{SC})_{2002} = \frac{(V_{TC})_{2002}}{IC_{2002}} \cdot 100 = \frac{1.23}{85.61497235} \cdot 100 = 1.43666460;$$

itd. Stoga peti stupac zadane tablice izgleda ovako:

Proizvodnja u stalnim cijenama iz 2005. *godine* [000 000 €]
0.90296810
1.00463741
1.14444132
1.23612024
1.29977266
1.43666460
1.54224000
1.64160000
1.85000000
1.80000000
2.11279461
2.18378756

Indeks fizičkoga obujma izračunat ćemo koristeći relaciju

$$IFO = \frac{(V_{SC})_t}{(V_{SC})_b} \cdot 100 \text{, za } t = 0, 1, 2, \ldots, 10, 11 \text{ i } b = 8,$$

jer 2005. godini odgovara vrijednost vremenske varijable $t = 8$. Zbog toga je:

$$IFO_0 = \frac{(V_{SC})_0}{(V_{SC})_8} \cdot 100 = \frac{0.90296810}{1.85} \cdot 100 = 48.80908667;$$

$$IFO_1 = \frac{(V_{SC})_1}{(V_{SC})_8} \cdot 100 = \frac{1.00463741}{1.85} \cdot 100 = 54.30472482;$$

$$IFO_2 = \frac{(V_{SC})_2}{(V_{SC})_8} \cdot 100 = \frac{1.14444132}{1.85} \cdot 100 = 61.86169300;$$

itd. Prema tome, posljednji stupac zadane tablice izgleda ovako:

Indeks fizičkoga obujma (2005 = 100)
48.80908667
54.30472482
61.86169300
66.81731017
70.25798141
77.65754595
83.36432432
88.73513514
100
97.29729730
114.20511421
118.04257087

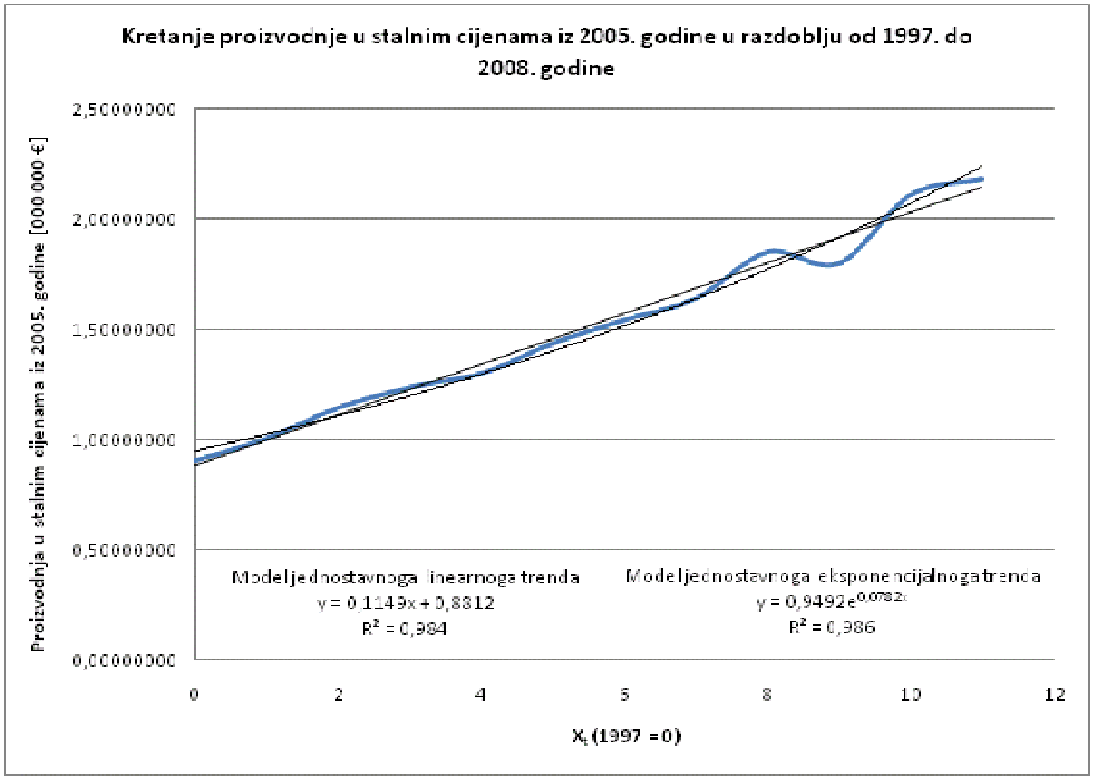
a cijela tablica ovako:

Godina	*Verižni indeks cijena*	*Proizvodnja u tekućim cijenama* [000 000 €]	*Indeks cijena na stalnoj bazi* (2005 = 100)	*Proizvodnja u stalnim cijenama iz* 2005. *godine* [000 000 €]	*Indeks fizičkoga obujma* (2005 = 100)
1997.	nepoznat	0.63	69.76990634	0.90296810	48.80908667
1998.	107	0.75	74.65379979	1.00463741	54.30472482
1999.	103	0.88	76.89341378	1.14444132	61.861693
2000.	101	0.96	77.66234792	1.23612024	66.81731017
2001.	106	1.07	82.32208879	1.29977266	70.25798141
2002.	104	1.23	85.61497235	1.43666460	77.65754595
2003.	103	1.36	88.18342152	1.54224000	83.36432432
2004.	105	1.52	92.59259259	1.64160000	88.73513514
2005.	108	1.85	100	1.85000000	100

Godina	*Verižni indeks cijena*	*Proizvodnja u tekućim cijenama* [000 000 €]	*Indeks cijena na stalnoj bazi* (2005 = 100)	*Proizvodnja u stalnim cijenama iz* 2005. *godine* [000 000 €]	*Indeks fizičkoga obujma* (2005 = 100)
2006.	110	1.98	110	1.80000000	97.2972973
2007.	108	2.51	118.8	2.11279461	114.2051142
2008.	106	2.75	125.928	2.18378756	118.0425709

Iz navedenih podataka vidimo npr. da je 2001. godine *proizvodnja iskazana u stalnim cijenama iz* 2005. *godine bila za* 100 – 70.25798141 = 29.74201859% *manja nego u* 2005. *godini*, a 2007. godine *za* 114.20511421 – 100 = 14.20511421% *veća nego u* 2005. *godini.*

Kretanje proizvodnje u stalnim cijenama iz 2005. godine, kao i jednadžbe modela jednostavnoga linearnoga trenda i modela jednostavnoga eksponencijalnoga trenda s odgovarajućim koeficijentima determinacije navedene su na donjoj slici:



Koeficijent determinacije modela jednostavnoga eksponencijalnoga trenda (R^2 = 0.986) je veći od koeficijenta determinacije modela jednostavnoga linearnoga trenda (R^2 = 0.984), pa je model jednostavnoga eksponencijalnoga trenda reprezentativniji. Koristeći funkciju GROWTH u MS Excelu prognoziramo vrijednosti proizvodnje u stalnim cijenama iz 2005. godine u razdoblju od 2009. do 2013. godine:

godina	*t*	*prognozirana proizvodnja* [000 000 €]
2009	12	2.42576156
2010	13	2.62304438
2011	14	2.83637186
2012	15	3.06704888
2013	16	3.31648645

(Ista se prognoza može dobiti tako da se u jednadžbu modela jednostavnoga eksponencijalnoga trenda

$$y = 0.9492 \cdot e^{0.0782 \cdot x} = 0.9492 \cdot 1.0813389048^{x}$$

uvrštavaju redom vrijednosti $x = 12$, $x = 13$, …, $x = 16$.)

Za izračun relativnih pokazatelja promjene cijenâ, odnosno proizvodnje bit će korisna sljedeća tri svojstva:

***Svojstvo* 1.** Prosječna stopa promjene vrijednosti neke pojave (iskazana u postotcima) jednaka je razlici geometrijske sredine verižnih indeksa te pojave i broja 100.

Dokaz: Neka je Y_i vrijednost promatrane pojave u razdoblju i, $i = 0, 1, 2, \ldots, n-1$. Označimo sa $\overline{S}$ prosječnu stopu promjene vrijednosti te pojave (iskazanu u postotcima). Tada mora vrijediti jednakost:

$$Y_{n-1} = Y_0 \cdot \left(1 + \overline{S}\right)^{n-1},$$

a odavde je

$$1 + \overline{S} = \sqrt[n-1]{\frac{Y_{n-1}}{Y_0}}.$$

Geometrijska sredina G svih $n-1$ verižnih indeksa te pojave jednaka je:

$$G = \sqrt[n-1]{\prod_{i=1}^{n-1} V_i} = \sqrt[n-1]{\prod_{i=1}^{n-1}\left(\frac{Y_i}{Y_{i-1}} \cdot 100\right)} = \sqrt[n-1]{100^{n-1} \cdot \frac{Y_1}{Y_0} \cdot \frac{Y_2}{Y_1} \cdot \frac{Y_3}{Y_2} \cdot \ldots \cdot \frac{Y_{n-2}}{Y_{n-3}} \cdot \frac{Y_{n-1}}{Y_{n-2}}} = 100 \cdot \sqrt[n-1]{\frac{Y_{n-1}}{Y_0}},$$

odnosno, uvažavajući prethodnu jednakost,

$$G = 100 \cdot (1 + \overline{S}).$$

Odavde slijedi

$$\overline{S} = \frac{G}{100} - 1 = \frac{G - 100}{100} = G - 100\,[\%],$$

što je i trebalo pokazati. ■

Iz navedenoga dokaza još se jednom može uočiti nedostatak geometrijske sredine (i iz nje izvedene prosječne relativne stope promjene) kao pokazatelja za prognozu vrijednosti neke pojave. U dokazu se ništa ne govori o međuvrijednostima $Y_1, Y_2, \ldots, Y_{n-2}$, već se isključivo zahtijeva da se podudaraju prva (nulta) i posljednja ($(n-1)$–va) vrijednost pojave. Zbog toga je za prognoziranje budućih vrijednosti te pojave bolje primijeniti ranije spomenute modele trenda.

Svojstvo 2. Za svaki $i = 0, 1, \ldots, n-1$ neka su Y_i vrijednosti neke pojave u razdoblju i, a I_i pripadni indeksi na stalnoj proizvoljnoj, ali fiksiranoj bazi b, za bilo koji $b \in \{0, 1, \ldots, n-1\}$. Tada je prosječna stopa $\overline{S}$ promjene vrijednosti te pojave (iskazana u postotcima) jednaka

$$\overline{S} = 100 \cdot \left(\sqrt[n-1]{\frac{I_{n-1}}{I_0}} - 1 \right) [\%].$$

Dokaz: Prema prethodnom dokazu je

$$1 + \overline{S} = \sqrt[n-1]{\frac{Y_{n-1}}{Y_0}},$$

odnosno

$$\overline{S} = \sqrt[n-1]{\frac{Y_{n-1}}{Y_0}} - 1 = 100 \cdot \left(\sqrt[n-1]{\frac{Y_{n-1}}{Y_0}} - 1 \right) [\%].$$

Stoga je dovoljno dokazati da vrijedi jednakost:

$$\frac{I_{n-1}}{I_0} = \frac{Y_{n-1}}{Y_0},$$

a ona izravno slijedi iz definicije indeksa na stalnoj bazi:

$$\frac{I_{n-1}}{I_0} = \frac{\frac{Y_{n-1}}{Y_b}}{\frac{Y_0}{Y_b}} = \frac{Y_{n-1}}{Y_0}.$$ ■

Svojstvo 3. Neka je $\overline{S}$ prosječna stopa promjene neke pojave (iskazana u postotcima) u nekom razdoblju. Tada je ukupna relativna promjena vrijednosti te pojave (iskazana u postotcima) u promatranom razdoblju:

$$p = \left(1+\overline{S}\right)^{n-1} - 1\,[\%].$$

Dokaz: Prema dokazu Svojstva 1. vrijedi jednakost:

$$Y_{n-1} = Y_0 \cdot \left(1+\overline{S}\right)^{n-1}.$$

Ukupna relativna stopa promjene iskazana u postotcima jednaka je:

$$p = \frac{Y_{n-1} - Y_0}{Y_0} \cdot 100.$$

Iz tih dviju jednakosti slijedi:

$$p = \frac{Y_{n-1} - Y_0}{Y_0} \cdot 100 = \frac{Y_0 \cdot \left(1+\overline{S}\right)^{n-1} - Y_0}{Y_0} \cdot 100 = \frac{Y_0 \cdot \left[\left(1+\overline{S}\right)^{n-1} - 1\right]}{Y_0} \cdot 100,$$

odnosno

$$p = \left[\left(1+\overline{S}\right)^{n-1} - 1\right] \cdot 100,$$

što je i trebalo pokazati. ■

Na temelju prethodnih svojstava rješavamo preostali dio zadataka iz Primjera 16. Ukupnu relativnu promjenu i prosječnu godišnju promjenu cijena u promatranom razdoblju izračunat ćemo primjenom Svojstava 2. i 3. na indekse cijena na stalnoj bazi 2005 = 100. Tako je najprije prosječna godišnja promjena cijena u promatranom razdoblju jednaka

$$\overline{S}_c = 100 \cdot \left(\sqrt[12-1]{\frac{I_{12-1}}{I_0}} - 1\right) = 100 \cdot \left(\sqrt[11]{\frac{I_{11}}{I_0}} - 1\right) = 100 \cdot \left(\sqrt[11]{\frac{125.928}{69.76990634}} - 1\right) \approx 5.044\,[\%],$$

tj. cijene su *rasle prema prosječnoj godišnjoj stopi od približno* 5,044%. Ukupna relativna promjena cijena je

$$p_c = \left[\left(1+\overline{S}\right)^{n-1} - 1\right] \cdot 100 = \left[\left(1 + \frac{5.0439829422782}{100}\right)^{12-1} - 1\right] \cdot 100 \approx 71.8237\,[\%],$$

tj. u promatranom se su razdoblju cijene *povećale za ukupno približno* 71.8237%.

Očekivanu ukupnu relativnu promjenu i očekivanu prosječnu godišnju promjenu proizvodnje u stalnim cijenama iz 2005. godine za razdoblje od 2009. do 2013. godine izračunat ćemo koristeći Svojstvo 3. i formulu

$$\overline{S} = \sqrt[n-1]{\frac{Y_{n-1}}{Y_0}} - 1 = 100 \cdot \left(\sqrt[n-1]{\frac{Y_{n-1}}{Y_0}} - 1 \right) [\%].$$

Izračunajmo najprije očekivanu prosječnu godišnju promjenu proizvodnje. Budući da imamo ukupno 5 prognoziranih vrijednosti proizvodnje u stalnim cijenama iz 2005. godine, to je n = = 5, pa je Y_{n-1} = krajnja prognozirana vrijednost proizvodnje = 3.31648645 i Y_0 = početna prognozirana vrijednost proizvodnje = 2.42576156. Stoga je

$$\overline{S}_p = 100 \cdot \left(\sqrt[5-1]{\frac{3.31648645}{2.42576156}} - 1 \right) = 100 \cdot \left(\sqrt[4]{\frac{3.31648645}{2.42576156}} - 1 \right) \approx 8.13282\ [\%],$$

tj. proizvodnja u stalnim cijenama iz 2005. godine u razdoblju od 2009. do 2013. godine očekivano će *rasti po prosječnoj godišnjoj stopi od približno* 8.13282%. Očekivano ukupno povećanje proizvodnje u stalnim cijenama iz 2005. godine u razdoblju od 2009. do 2013. godine dobit ćemo iz jednakosti:

$$p_p = \left[\left(1 + \overline{S}_p\right)^{n-1} - 1 \right] \cdot 100 = \left[\left(1 + \frac{8.13282}{100}\right)^{5-1} - 1 \right] \cdot 100 \approx 36.7194 [\%],$$

tj u promatranom će se razdoblju proizvodnja u stalnim cijenama iz 2005. godine očekivano ukupno *povećati za približno* 36.7194%.

Zaključno napomenimo da bi sve gore izračunate *relativne* vrijednosti promjena cijena, odnosno proizvodnje ostale *nepromijenjene* ukoliko bismo za bazno razdoblje uzeli bilo koju drugu godinu promatranoga razdoblja. To, dakako, ne vrijedi za prognoziranu proizvodnju u stalnim cijenama jer njezine vrijednosti izravno ovise o tome koju smo godinu odabrali za baznu godinu (za drugu baznu godinu dobit ćemo druge vrijednosti proizvodnje u stalnim cijenama iz te godine). Zbog toga se pri ekonomskim analizama u ovakvim slučajevima češće primjenjuju relativni pokazatelji promjena vrijednosti neke pojave.

6. LITERATURA

1. M. Papić: "*Primijenjena statistika u* MS *Excelu*" (4. izdanje), VPŠ "Libertas" – Naklada "Zoro", Zagreb, 2012.
2. M. Vukičević, M. Papić: *Matematičko-statistički priručnik za poduzetnike*, Golden-marketing – Tehnička knjiga, Zagreb, 2003.
3. I. Šošić: *Primijenjena statistika*, Školska knjiga, Zagreb, 2004.
4. V.Serdar, I. Šošić: *Uvod u statistiku*, Školska knjiga, Zagreb, 2000.
5. A. Rozga, B. Grčić: *Poslovna statistika*, Veleučilište u Splitu[130], Split, 1999.
6. Statistički ljetopisi Republike Hrvatske i Mjesečna statistička izvješća [objavljeni na www.dzs.hr], Državni zavod za statistiku, Zagreb
7. „Žene i muškarci u Hrvatskoj“ (publikacija objavljena na www.dzs.hr), Državni zavod za statistiku, Zagreb, 2007.
8. Sveučilišna anketa o kvaliteti nastave objavljena na www.unizg.hr, Sveučilište u Zagrebu, Zagreb, 2008.
9. nastavni materijali iz kolegija *Statistika* objavljeni na www.efos-statistika.com, Ekonomski fakultet Sveučilišta u Osijeku, Osijek, 2009.
10. nastavni materijal doc.dr.sc. Snježane Pivac iz kolegija *Poslovna statistika* objavljen na www.efst.hr/nastava/44/12.**Skupni_indeksi**._Doc._dr._S._Pivac.pdf, Ekonomski fakultet Sveučilišta u Splitu, Split, 2009.
11. članci objavljeni na *Wikipedii*:
    http://en.wikipedia.org/wiki/Correlation,
    http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient,
    http://en.wikipedia.org/wiki/Regression_analysis,
    http://en.wikipedia.org/wiki/Dollar_cost_averaging,
    http://en.wikipedia.org/wiki/Trend_estimation

[130] Današnji Sveučilišni studijski centar za stručne studije Sveučilišta u Splitu.