UVOD

Pojam i značaj statistike
Način statističkog istraživanja
Statistički skup
Statističko obilježje
Parametar skupa i statistika uzorka
Nastanak i razvoj statistike
Primjena računara u statistici

CILJEVI POGLAVLJA

Nakon čitanja ovog poglavlja bićete u stanju da:

1. shvatite i cijenite značaj statistike u ekonomiji i biznisu
2. protumačite statističke zakonitosti
3. shvatite razliku između statističkog skupa i uzorka
4. klasifikujete statistička obilježja
5. shvatite osnovni mehanizam statističkog zaključivanja
6. shvatite značaj i ograničenja primjene računara u statističkoj analizi

POJAM I ZNAČAJ STATISTIKE

Statistika ima svoj poseban, statistički – induktivni, način razmišljanja, koji se veoma razlikuje od deduktivnog u matematici.

Varijabilna pojava

Varijabilna pojava je ona na koju djeluje veliki broj faktora i zbog toga uzima različite vrijednosti od jednog do drugog slučaja svoga ispoljavanja. Te pojedinačne vrijednosti nemoguće je sa sigurnošću predvidjeti.

Ekstremni podatak

Ekstremni podatak (opservacija, ili vrijednost) je onaj koji znatno odstupa od vrijednosti ostalih podataka, bilo zato što je znatno veći ili znatno manji.

Apsolutno homogene pojave ne zanimaju statistiku.

Varijabilnost neke pojave nema nikave veze sa brojem slučajeva (masovnošću) na kojima se ta pojava iskazuje.

Zbog čega se javljaju varijacije?

Zbog čega je neophodno istraživati varijacije?

Statističke zakonitosti imaju dvije bitne karakteristike:

- važe samo u masi slučajeva, i
- pojedinačni slučajevi mogu da pokažu odstupanja od opšte tendencije.

Statistika

Statistika je univerzalni kvalitativno-kvantitativni naučni metod analize varijabilnih pojava, zasnovan na teoriji vjerovatnoće.

Neki autori **statistiku definišu kao nauku o podacima.**

Podaci su brojevi sa odgovarajućim kontekstom.

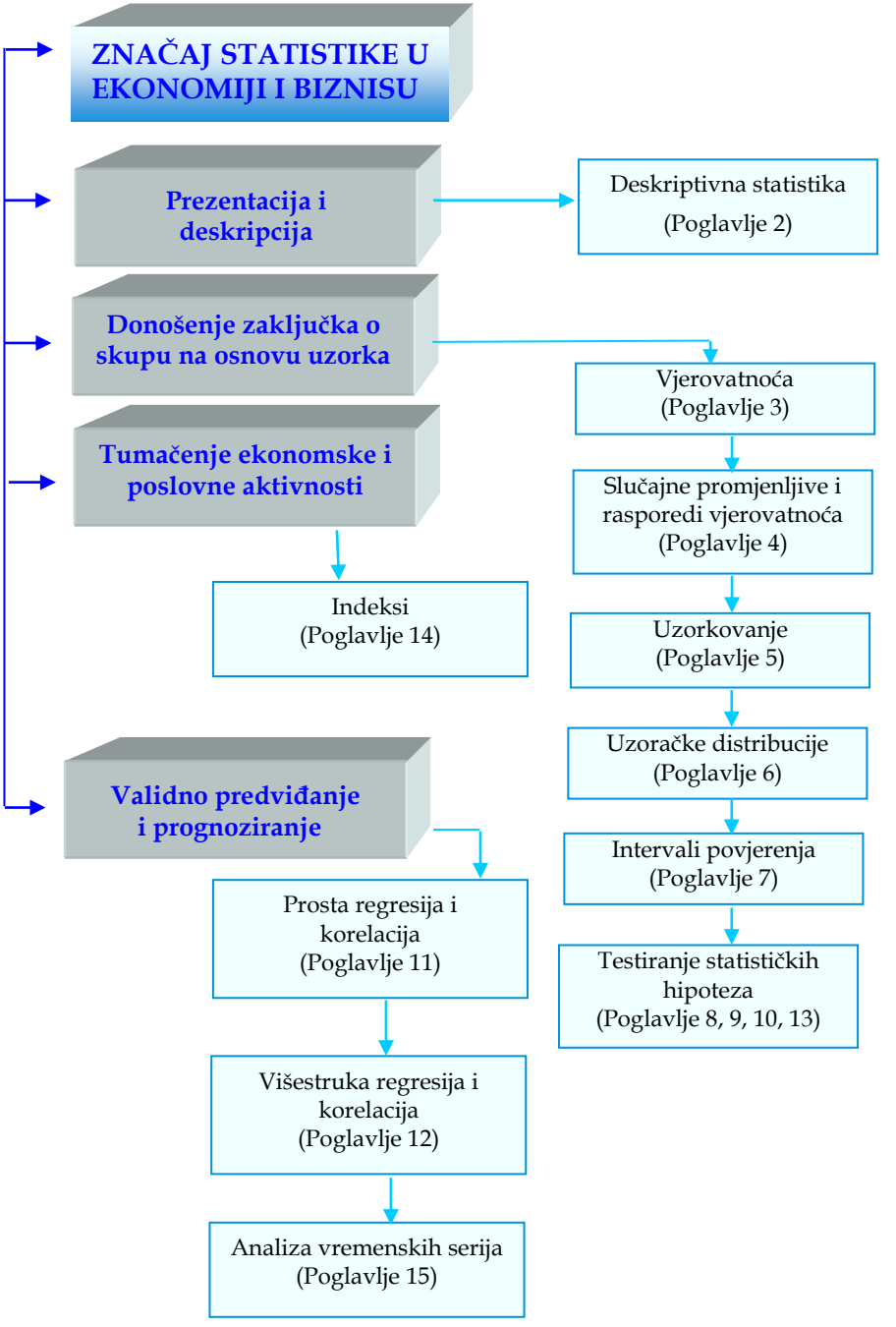
Statističar treba – da na kritički i analitički način izabere onaj statistički metod koji najviše odgovara posmatranim podacima i, shodno tome, interpretira rezultat, tj. formuliše zaključak.

Biznis menadžeri, donosioci odluka u firmi ili vladi, kao i ekonomisti, imaju potrebu za statističkom analizom bar iz sljedeća četiri razloga:

1. da bi znali kako da na ispravan sintetički način prikažu i opišu podatke;
2. da bi znali kako da, na osnovu samo dijela raspoloživih podataka, donesu validan zaključak o cjelini kojoj taj dio pripada;
3. kako da tumače ekonomske indikatore i indikatore poslovne aktivnosti; i
4. da izvrše validno predviđanje.

Kolika je važnost statistike najbolje ilustruje sljedeća, često citirana izjava Marka Tvena: "Statistika će jednog dana biti neophodna običnom građaninu isto onoliko koliko čitanje i pisanje". A taj dan je osvanuo početkom trećeg milenijuma.

NAČIN STATISTIČKOG ISTRAŽIVANJA



STATISTIČKI SKUP

Predmet statističkog istraživanja su varijabilne pojave.

Statistički skup

Skup svih elemenata na kojima se izvjesna varijabilna pojava ispoljava i statistički posmatra naziva se statistički skup ili osnovni skup ili populacija ili skup.

Jedinice skupa

Pojedinačni elementi iz kojih se skup sastoji nazivaju se jedinicama skupa ili jedinicama posmatranja.

Skup mora ispunjavati određene uslove da bi mogao da se nazove statističkim skupom:

1. Statistički skup mora da obuhvati sve elemente koji su predmet posmatranja.

2. Elementi toga skupa moraju imati bar jednu zajedničku osobinu na osnovu koje se i deklarišu kao pripadnici toga skupa.

3. Na elementima takvoga skupa se posmatra neka varijabilna pojava. Iz ovoga slijedi da ti elementi moraju imati bar jednu karakteristiku po kojoj se mogu razlikovati, odnosno koja je varijabilna.

U zavisnosti od cilja istraživanja, osnovni skup se može sastojati od ljudi, bića, predmeta ili događaja.

Statistički skup je potrebno precizno odrediti, odnosno definisati: **sadržinski, prostorno i vremenski.**

Sadržinski odrediti neki statistički skup zahtijeva jasno definisanje osobine koju mora da posjeduje svaka jedinica da bi bila predmet posmatranja.

Prostorno odrediti osnovni skup znači precizirati teritoriju u okviru koje će se posmatrati data varijabilna pojava.

Vremenski odrediti skup znači precizno odrediti jedan momenat ili vremenski interval u kojem ćemo izmjeriti nivo pojave (i/ili snimiti njenu strukturu).

STATISTIČKO OBILJEŽJE

Obilježje

Osobine po kojima se jedinice statističkog skupa među sobom razlikuju, a koje su predmet statističkog istraživanja, nazivamo obilježjima (promjenljivim ili varijablama).

Sva obilježja u statistici možemo **klasifikovati u dvije osnovne grupe:**

- atributivna (kvalitativna, kategorijska), i
- numerička (kvantitativna).

Različiti vidovi u kojima se jedno obilježje može javiti nazivaju se modalitetima ili vrijednostima tog obilježja.

Numerička obilježja:

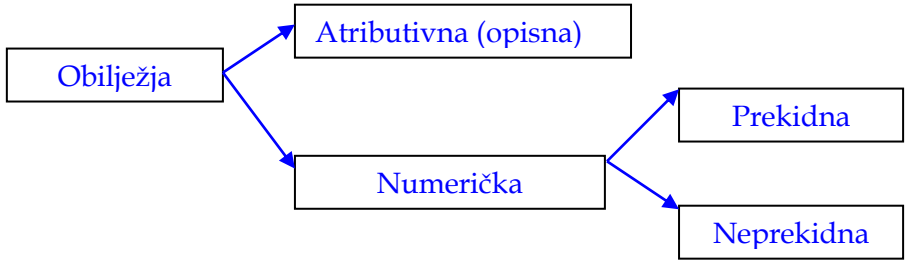
- prekidna (ili diskretna) numerička obilježja, i
- neprekidna (ili kontinuirana) numerička obilježja.

Suštinska razlika između ove dvije grupe je u tome što prekidna obilježja svoje vrijednosti (modalitete) dobijaju na osnovu prebrojavanja, a neprekidna na osnovu mjerenja.

Prekidna obilježja su numeričke karakteristike koje mogu uzimati samo izolovane vrijednosti na mjernoj skali.

Neprekidna obilježja predstavljaju numeričke karakteristike jedinica skupa koje mogu uzeti bilo koju vrijednost unutar nekog intervala.

Obilježje je ono po čemu se jedinice skupa međusobno razlikuju, a ne ono po čemu su slične.



Slika 1 Klasifikacija obilježja (varijabli) u statistici

PARAMETAR SKUPA I STATISTIKA UZORKA

Statistika ispituje varijabilne pojave na osnovu svih podataka statističkog skupa (metod popisa) ili na osnovu dijela toga skupa - uzorka.

Statistički uzorak

Statistički uzorak predstavlja dio statističkog skupa na osnovu čijih osobina donosimo statističke zaključke o odgovarajućim karakteristikama populacije iz koje je izabran.

Uzorak koristimo isključivo da bismo, uopštavanjem dobijenih informacija iz uzorka, došli do informacije o nepoznatim karakteristikama skupa u cjelini.

Reprezentativni uzorak

Uzorak je reprezentativan ako svojim osobinama vjerno odslikava osobine statističkog skupa iz kojeg je izabran.

Uzorak, sam po sebi, nije cilj, već samo sredstvo da se dođe do željene informacije o skupu. Takve karakteristike skupa u statistici se nazivaju **parametrima skupa.**

Parametar skupa

Parametar skupa je neka sumarna numerička karakteristika toga skupa.

Drugačije rečeno, po svojoj statističkoj prirodi parametar je neka konstanta, a ne promjenljiva.

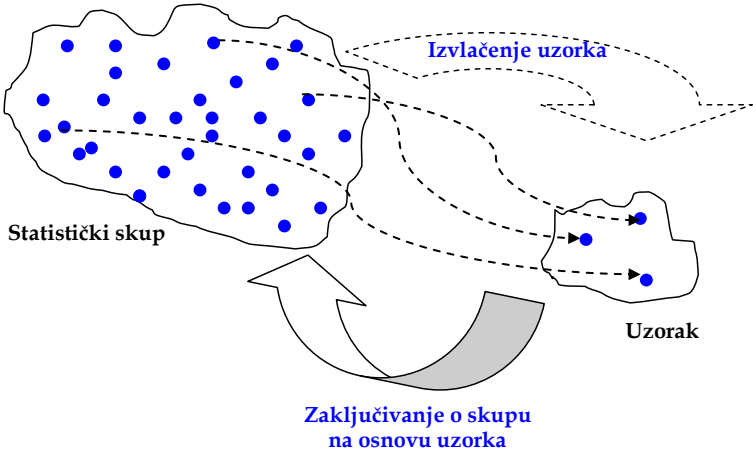
U statistici, o nepoznatoj vrijednosti parametra donijećemo ocjenu koristeći odgovarajući pokazatelj iz uzorka. Takav pokazatelj nazivamo statistika uzorka ili samo statistika.

Statistika uzorka

Statistika uzorka je neka sumarna numerička karakteristika toga uzorka.

Ne treba zaboraviti: parametar je za populaciju ono što je statistika uzorka za uzorak.

Dakle, cilj statističkog zaključivanja jeste da na osnovu statistike uzorka dođemo do informacije o parametru skupa.



Slika 2 Postupak statističkog zaključivanja

NASTANAK I RAZVOJ STATISTIKE

Začetak statistike kao naučnog metoda može se naći u **Njemačkoj i Engleskoj** u XVII i prvoj polovini XVIII vijeka, kada se pojavljuju **dvije statističke koncepcije**.

U to vrijeme se i prvi put spominje riječ statistika.

Danas se smatra da riječ "statistika" potiče od novolatinskog izraza statisticum collegium (predavanja o državnim poslovima).

Ovaj izraz koristio je njemački profesor univerziteta u Jeni, **Martin Schmeizel (1679-1747),** u nizu svojih predavanja o ustavima, resursima i politici raznih država u svijetu.

Međutim, prvi koji je upotrijebio riječ statistika u pisanoj formi je njemački profesor **Gottfried Achenvall 1748.**

Zadatak statistike zasnivao se samo na deskripciji, pa je ovaj koncept nazvan kasnije deskriptivna škola, ili državopis. Kako je statistika u takvom obliku predavana i na univerzitetu, ovakav pravac nazivan je i "univerzitetska statistika".

U Engleskoj se razvila drugačija statistička koncepcija, "Politička aritmetika". Istaknuti predstavnici ovog pravca **J. Graunt, W. Petty i E. Halley** su istakli zahtjev za matematičkom obradom i analizom statističkih podataka i otkrivanja zakonitosti u ponašanju pojava.

Ako izuzmemo statističku teoriju (koja iznalazi nove statističke metode i matematička je prije svega), ***primijenjena statistika*** se dijeli u dvije grupe: **deskriptivnu statistiku i inferencijalnu statistiku.**

Deskriptivna statistika

Deskriptivna statistika se sastoji od metoda za prikupljanje, obradu i prikazivanje podataka korišćenjem tabela, grafikona i sumarnih mjera.

Inferencijalna statistika

Inferencijalna statistika podrazumijeva one statističke metode koji koriste uzorak u cilju donošenja zaključka o osnovnom skupu.

Razvoj statistike možemo grubo podijeliti u **tri etape:**

1. Sakupljanje podataka o brojnom stanju stanovnika, vojnika, poreskih obveznika neophodnih državi, prije svega za vođenje poreske politike.
2. Razvoj teorije vjerovatnoće, koji je statistici dao neophodan mehanizam i omogućio da se na osnovu uzorka donese validan zaključak o osnovnom skupu.
3. Revolucija u razvoju i dostupnosti elektronskih računara posljednjih godina otvorila je neslućene mogućnosti statistici – da postane primjenjiva u skoro svim naučnim oblastima i da koristi nove metode, koji su neupotrebljivi bez kompjuterske podrške, vjerovatno je najveće otkriće u statistici u drugoj polovini XX vijeka.

Tabela 1 Neke najvažnije statističke akcije, otkrića i podaci u evoluciji statistike

Godina	Otkriće ili događaj	Autor
3800. BC[1]	Popisi u Vavilonu sa ciljem oporezivanja	
2323. BC	Popisi stočnog fonda u Egiptu održavaju se svake godine (do tada dvogodišnje)	
1055. BC	Popis stanovništva u Izraelu	Kralj David
550. BC	Prvi popis (cenzus) u Rimu, Rim ima 83,000 građana	Servije Tulije
28. BC	Popis otkriva da u Rimskom carstvu ima 4,063,000 građana	
2.	Najstariji popis čiji su rezultati sačuvani, Kina po ovom popisu ima 47,5 miliona stanovnika	Kineska Hun dinastija
1086.	Najveća statistička akcija srednjeg vijeka – popis u Engleskoj, rezultati objavljeni u *Knjizi strašnog suda*	William I Osvajač
1654.	**Postavljeni temelji teorije vjerovatnoće**	**Blaise Pascal i Pierre de Fermat**

[1] U nedostatku odgovarajuće naše skraćenice, koristili smo englesku BC da označimo prije nove ere.

1662.	Prva publikovana demografska studija zasnovana na tablicama smrtnosti	John Graunt
1676.	**Iz štampe izlazi *Politička aritmetika***	**William Petty**
1710.	Prva upotreba nekog statističkog testa	John Arbuthnott
1713.	Posthumno se štampa najvažniji rad o teoriji vjerovatnoće u XVIII vijeku: *Ars Conjectandi*	Jacob Bernoulli
1733.	Otkriće normalnog rasporeda	Abraham De Moivre
1749.	**Prvi put se koristi riječ statistika u nekom radu**	**Gottfried Achenvall**
1763.	Temelji Bayes-ove statistike, zasnovane na subjektivnom konceptu vjerovatnoće	Thomas Bayes
1801.	Svjetska populacija dostigla **1** milijardu stanovnika	
1805.	Otkriće metoda najmanjih kvadrata	A. M. Legendre
1809.	Gauss ponovo otkriva normalan raspored i proširuje metod najmanjih kvadrata	Carl F. Gauss
1812.	Prvi iscrpni publikovani rad iz teorije vjerovatnoća	Pierre S. Laplace
1853.	U Berlinu, prva međunarodna statistička konferencija	Adolphe Quetelet

1885.	**Uvodi se prvi put ideja regresije**	**Francis Galton**
1896.	Formulisan koeficijent proste linearne korelacije	Karl Pearson
1900.	Formulisan χ^2 test	Karl Pearson
1904.	Formulisan Spearman-ov koeficijent korelacije	Charles Spearman
1908.	Otkriće ocjenjivanja aritmetičke sredine u slučaju kada je standardna devijacija skupa nepoznata	William Gosset ("Student")
1919.	Formulisan koncept analize varijanse	Ronald Fisher
1925.	Svjetska populacija dostiže **2** milijarde stanovnika	
1925.	**Štampana knjiga *Statistički metodi za istraživače***	**Ronald Fisher**
1933.	Formulisani intervali povjerenja, greška II vrste, jačina testa, kritični regioni	Jerzy Neyman i Egon Pearson
1933.	Postavljen aksiomatski koncept vjerovatnoće	Andrei Kolmogorov
1945.	Formulisani najpoznatiji neparametarski testovi: test sume rangova i Wilcoxon-ov test ranga sa znakom	Frank Wilcoxon
1959.	Svjetska populacija dostiže **3** milijarde stanovnika	

1966.	Prva primjena resampling statističkih metoda (metoda ponovljenih uzoraka)	Julian L. Simon
1972.	Formulisan koncept istraživačke analize podataka	John Tukey
1972.	Formulisani generalizovani linearni modeli	J.A. Nelder i R.W.M. Wedderburn
1974.	Populacija u svijetu dostiže **4** milijarde	
1979.	Formulisan b*ootstrap* metod	Bradley Efron
1986.	Populacija u svijetu dostiže **5** milijardi	
2000.	Populacija svijeta broji **6** milijardi	
2002.	Korišćenjem FDR metoda potvrđena teorija Big Benga	

Kada nešto čujem ja to zaboravim,
Kada nešto vidim to zapamtim,
Tek kada nešto uradim sam, to i razumijem.

KINESKA POSLOVICA

PRIMJENA RAČUNARA U STATISTICI

Najpoznatiji statistički paketi:

SPSS, Minitab, SAS, Statistica i Statgraphics. Statističari su formulisali i posebne programerske jezike, od kojih je najpoznatiji jezik koji se naziva *R*.

Koristićemo uglavnom statistički softver 3BStat**,** a po potrebi i neke od gore navedenih.

DESKRIPTIVNA ANALIZA

2.1 Posmatranje, prikupljanje, sređivanje i obrada podataka
2.2 Prikazivanje podataka
2.3 Deskriptivne statističke mjere
2.4 Opis osnovnih karakteristika rasporeda

CILJEVI POGLAVLJA

Nakon čitanja ovog poglavlja bićete u stanju da:

7. shvatite razliku između potpunog i djelimičnog posmatranja statističkog skupa
8. shvatite značaj statističkog popisa, izvještaja i uzorka kao metoda prikupljanja podataka
9. grupišete, sredite i prikažete podatke u vidu statističkih serija, tabela i grafički
10. shvatite značaj deskriptivnih statističkih mjera skupa i uzorka
11. ručno i pomoću statističkog softvera izračunate mjere centralne tendencije, mjere varijacije i mjere oblika rasporeda
12. protumačite rezultate deskriptivne statističke analize

Statističko istraživanje:

Niz postupaka sa određenim ciljem, predmetom, statističkom jedinicom i obilježjima, koji podrazumijeva primjenu metoda i postupaka u različitim aktivnostima ispitivanja određene varijabilne pojave.

Najčešće se govori o tri sljedeće etape statističkog istraživanja:

- statističko posmatranje;
- sređivanje, grupisanje, i obrada podataka;
- statistička analiza.

Posmatranje i prikupljanje, sređivanje, grupisanje, prikazivanje i obrada podataka spadaju u područje deskriptivne statistike.

Statistička analiza obuhvata metode čiji je zadatak objašnjavanje posmatranih varijabilnih pojava i statističko zaključivanje o parametrima na osnovu uzorka.

POSMATRANJE, PRIKUPLJANJE, SREĐIVANJE I OBRADA PODATAKA

Prva faza statističkog istraživanja započinje preciznim postavljanjem cilja i zadatka istraživanja, koji su osnova za rješavanje metodoloških, organizacionih i finansijskih pitanja.

Navedena pitanja utvrđuju se planom statističkog istraživanja.

Planom istraživanja definišu se:

-predmet istraživanja,

-statistički skup i njegovi elementi,

-obilježja jedinica skupa,

-kao i način grupisanja i obrade prikupljenih podataka.

Cilj statističkog posmatranja je da se obezbijede kvalitetni podaci o varijabilnoj pojavi.

Svako prikupljanje podataka podrazumijeva mjerenje.

Mjerenje predstavlja pridruživanje brojeva ili određenih oznaka jedinicama skupa, prema određenom pravilu.

Rezultati statističkog istraživanja mjere se korišćenjem sljedećih mjernih skala: nominalne, ordinalne, intervalne i skale odnosa.

- Nominalna skala data je u vidu liste naziva, kategorija ili

određenih atributa po kojima se jedinice statističkog skupa razlikuju.

Ukoliko je riječ o atributivnim obilježjima koja imaju veliki broj modaliteta, klasifikacija pojedinih modaliteta vrši se u srodne grupe, u okviru posmatranog obilježja. Na taj način formiraju se jednoobrazne grupe i podgrupe, koje se najčešće nazivaju nomenklaturama (npr. nomenklatura zanimanja).

- Ordinalna skala se koristi ukoliko je moguće modalitete obilježja

rangirati prema značaju u odnosu na usvojene kriterijume. Ova skala jedinicama skupa pridružuje brojeve, slovne oznake ili određene simbole, prema stepenu određenog svojstva ili odlike.

Mjesto modaliteta na mjernoj skali predstavlja njegov rang, a ne određenu mjernu veličinu.

- Intervalna skala svakom modalitetu obilježja pridaje određenu

jedinicu mjere. Ovom skalom jedinicama skupa pridružuju se brojevi, pri čemu jednake razlike brojeva predstavljaju jednake razlike mjerene karakteristike. Intervalna skala omogućava utvrđivanje redosljeda modaliteta u skupu, kao i mjeru njihovog razlikovanja.

Tipični primjeri za pojave koje se mogu mjeriti u nivou intervalne skale su temperatura iskazana u Celzijusovim stepenima, i kalendarsko vrijeme. Obje ove pojave nemaju pravu, već arbitrarnu nultu tačku.

- Skala odnosa pokazuje i redosljed modaliteta i mjeru njihovog

razlikovanja.

Ova mjerna skala obezbjeđuje **najviši nivo mjerenja.**

Skalu odnosa ne karakteriše samo upotreba jedinice mjerenja, nego i prava nulta tačka, koja ukazuje na nepostojanje određene karakteristike. Zbog toga ova skala omogućava iskazivanje proporcionalnih odnosa modaliteta obilježja koja se mjere. Skala odnosa je najpreciznija mjerna skala.

Samo intervalna skala i skala odnosa su prave numeričke skale.

Metodi posmatranja i prikupljanja podataka

Posmatranje i prikupljanje podataka vrši se na osnovu prethodno postavljenog plana prikupljanja podataka.

Prema izvodu podataka koji se koriste u statističkom istraživanju, može se govoriti o primarnim i sekundarnim statističkim podacima.

- Primarni statistički podaci prikupljaju se postupkom statističkog

posmatranja i prikupljanja podataka.

- Sekundarni podaci obezbjeđuju se iz sekundarnih izvora (zavodi

za statistiku, ili institucije ovlaštene za prikupljanje primarnih podataka : centralna banka, carinska služba, matične službe opština, izvještaji o poslovanju preduzeća i sl.).

Statističko istraživanje može se zasnivati na potpunom obuhvatu svih jedinica skupa (potpuno posmatranje), ili samo na jednom dijelu njegovih jedinica (djelimično posmatranje).

Potpuno posmatranje, odnosno potpuni obuhvat jedinica skupa može se obezbijediti primjenom statističkog popisa i statističkog izvještaja (tekuće registracije).

Statističkim popisom obuhvataju se sve jedinice skupa u određenom momentu koji se naziva kritični momenat.

Osnovne karakteristike popisa su:

- sveobuhvatnost (posmatranje svih jedinica skupa);
- istovremenost popisa (pri čemu kraći period popisa

obezbjeđuje veću tačnost podataka);

- vrijeme provođenja popisa (kritični momenat, kada je stanje pojave „normalno“);

- ponavljanje popisa (ponovo provođenje popisa u jednakim vremenskim intervalima obezbjeđuje uporedivost podataka);

- normativno regulisanje popisa (zakonski propisi kojima se regulišu prava i obaveze učesnika u popisu i obezbjeđuju normalno odvijanje popisa).

Statistički izvještaj, kao metod potpunog posmatranja, koristi se za prikupljanje podataka o pojavama kod kojih je izražen veći varijabilitet tokom vremena ili prostora.

Izvještaji mogu da budu tipski i specijalni.

S obzirom na vrijeme obuhvata podataka o pojavi, izvještaji mogu da se podnose u sukcesivnim momentima ili u sukcesivnim vremenskim periodima.

Djelimično (nepotpuno) posmatranje zasnovano na statističkom uzorku.

Statističko uzorkovanje predstavlja metod po kome se na osnovu posmatranja jednog dijela jedinica skupa zaključuje o karakteristikama i ponašanju cijelog skupa.

Da bi zaključci na osnovu uzorka bili relevantni za cijeli skup, potrebno je da uzorak bude

reprezentativan.

Uzorak je reprezentativan ako vjerno odslikava strukturu osnovnog skupa iz kojeg je odabran.

Primjena metoda uzorka podrazumijeva anketiranje, kao jednoobrazno prikupljanje podataka, koje se obezbjeđuje adekvatnim upitnicima i pripremom anketara i lica koja će ih popunjavati.

Primjena metoda uzorka neminovno dovodi do mogućnosti greške u statističkom zaključivanju.

Greške mogu da budu slučajne i sistematske. Cilj je smanjenje sistematske greške koja može da utiče na rezultat, dok se slučajna greška smanjuje ili potpuno gubi u velikom broju podataka.

Sređivanje, grupisanje i obrada podataka

Sagledavanje karakteristika jedinica posmatranog skupa zasniva se na prikupljenim podacima, koji se prethodno sređuju prema određenim kriterijumima.

U ovoj fazi statističkog istraživanja prikupljeni statistički materijal pretvara se u brojčane informacije o posmatranom skupu formiranjem statističkih serija i tabela.

Sređivanje statističkih podataka predstavlja uređivanje podataka o jedinicama skupa po svakom obilježju.

Grupisanje podataka predstavlja raščlanjavanje statističkog skupa na određeni broj podskupova, koji se međusobno ne preklapaju.

Pri tome je potrebno uvažavati pravilo da se svaki podatak mora razvrstati u grupe, kao i to da jedan podatak može pripadati samo jednoj grupi.

Sređivanje statističke građe može biti centralizovano, ukoliko se vrši u jednom centru, odnosno u odgovarajućem (centralnom) zavodu za statistiku. Decentralizovano sređivanje vrši se u mjestima posmatranja i prikupljanja podataka o pojavama.

U slučaju većih statističkih akcija koristi se i kombinovano sređivanje podataka, u kome se jedan dio poslova vrši centralizovano, a drugi decentralizovano, radi bržeg obezbjeđenja prethodnih rezultata.

Sređivanje i grupisanje podataka znatno je olakšano i ubrzano korišćenjem računara i gotovih programa za sređivanje i obradu podataka.

PRIKAZIVANJE PODATAKA

Sređeni statistički podaci najčešće se prikazuju u vidu statističkih serija i tabela.

Prema zahtjevima publikovanja ili dalje statističke analize, statistički podaci prikazuju se i grafičkim putem.

Statističke serije

Sređivanjem i grupisanjem podataka dobijaju se nizovi statističkih podataka prema jednom ili više obilježja, ili prema vremenu.

Statističke serije predstavljaju nizove sređenih statističkih podataka koji prikazuju strukturu skupa prema određenom obilježju, raspored statističkog skupa u prostoru, ili promjene skupa u vremenu.

Statističke serije, u zavisnosti od načina formiranja i sadržaja, mogu da budu serije strukture i vremenske serije.

Serije strukture pokazuju raspored statističkog skupa prema modalitetima, odnosno po vrijednostima obilježja.

Serije strukture sadrže modalitete obilježja i frekvencije, odnosno učestalost jedinica skupa po datom

modalitetu obilježja.

Ove serije mogu sadržavati atributivna i numerička obilježja.

Specifična vrsta serija strukture po atributivnim obilježjima su geografske (prostorne, teritorijalne) serije.

Nizovi statističkih podataka formirani na osnovu atributivnih obilježja predstavljaju atributivne serije strukture.

Specifičnu vrstu serija strukture po atributivnim obilježjima predstavljaju geografske (prostorne) serije.

Serije strukture formirane prema numeričkim (brojčanim) obilježjima nazivaju se numeričkim serijama, odnosno rasporedima frekvencija.

Klasifikacija brojčanih vrijednosti obilježja zavisi od toga da li je obilježje **prekidno ili neprekidno**. Prekidne vrijednosti obilježja grupišu se po veličini, od najmanje do najveće vrijednosti modaliteta.

Raspored frekvencija

Raspored frekvencija je numerička serija strukture.

U slučaju prekidnih obilježja sa velikim brojem modaliteta, ili, pak, u slučaju neprekidnih obilježja, postavlja se pitanje broja grupa (intervalnih modaliteta obilježja), kao i pitanje veličine intervala.

Da bi se odredila veličina intervala i broj intervalnih modaliteta neprekidnog obilježja može se koristiti tzv. Sturges-ovo pravilo.

Po tom pravilu, broj grupa (intervala, klasa) određuje se na osnovu obrasca:

$$K = 1 + 3{,}3 \log N,$$

gdje je N ukupan broj podataka.

Veličina intervala i određuje se na osnovu razlike najveće i najmanje vrijednosti obilježja, primjenom sljedećeg obrasca:

$$i = \frac{x_{\max} - x_{\min}}{K}.$$

Da bi intervali bili uporedivi, potrebno je da oni budu iste veličine.

Nekada nije moguće ispoštovati zahtjev da intervali budu iste veličine, a nekada je, u zavisnosti od izabranog obilježja, prihvatljivije da se odrede različite

veličine intervala modaliteta obilježja.

Takođe, nekada se može desiti da prilikom formiranja serija strukture svi intervali ne mogu da budu zatvoreni.

Prilikom formiranja serija mora se voditi računa o razgraničavanju grupnih intervala, pri čemu jedna brojčana vrijednost ne može istovremeno da bude gornja granica jednog i donja granica narednog intervala.

Zbog toga se donja granica svakog intervala obilježava decimalnim, a ne cijelim brojem.

Prilikom obrade ovakvih serija potrebno je intervalne vrijednosti modaliteta obilježja prevesti u prekidne vrijednosti. Utvrđuju se sredine intervala, odnosno razredne sredine, tako što se sabiraju donje i gornje granice intervala i podijele sa dva.

Serije strukture po numeričkim obilježjima nazivaju se serijama distribucije frekvencija (rasporedima ili razdiobama frekvencija).

Rasporedi frekvencija mogu da budu sa prekidnim i neprekidnim obilježjima. Serije distribucija frekvencija, za razliku od serija strukture po atributivnim obilježjima, mogu se iskazivati kao empirijske funkcije.

Osnovni nedostatak serije distribucije frekvencija je u tome da se zamjenom pojedinačnih vrijednosti obilježja grupnim intervalima gubi preciznost informacije o karakteristikama određenih jedinica skupa.

Serije strukture pokazuju koliko jedinica skupa ima određenu vrijednost modaliteta obilježja, odnosno kako su modaliteti obilježja raspoređeni u skupu.

Tu je riječ o apsolutnim frekvencijama, koje pokazuju broj jedinica sa odgovarajućom vrijednošću obilježja.

Nekada se apsolutne frekvencije kumuliraju, pa se umjesto pojedinačnih frekvencija za svaki modalitet obilježja koriste njihove kumulante.

Kumuliranje frekvencija vrši se tako što se, počevši od najniže vrijednosti, frekvencije postupno sabiraju, odnosno sukcesivno dodaju zbiru prethodnih frekvencija.

Na taj način dobija se rastuća kumulanta (odnosno kumulanta "ispod"). Rastuća kumulanta pokazuje broj jedinica u skupu čija je vrijednost ispod (manja od) gornje granice grupnog intervala.

Isto tako, počevši od prvog modaliteta obilježja,

frekvencije se mogu sukcesivno oduzimati od sume svih frekvencija, čime se dobija opadajuća kumulanta (kumulanta „iznad“).

Ukoliko se frekvencija jednog modaliteta stavi u odnos prema ukupnom broju jedinica skupa, dobiće se relativna frekvencija:

$$p_i = \frac{f_i}{\sum f_i}.$$

Relativna frekvencija najčešće se izražava u procentima.

Vremenske serije predstavljaju nizove statističkih podataka koji su složeni hronološkim redosljedom. Zbog toga se nazivaju i hronološkim serijama.

U zavisnosti od prirode podataka o pojavi, vremenske serije mogu da budu momentne i intervalne.

Momentne vremenske serije pokazuju nivo pojave u tačno određenim uzastopnim vremenskim momentima. Momentne serije dobijaju se kao rezultat obrade rezultata više uzastopnih popisa ili statističkih izvještaja o stanju pojave.

Intervalne vremenske serije pokazuju kretanje pojave u uzastopnim vremenskim intervalima. Njima se najčešće

prikazuje kretanje proizvodnje, plata, troškova života i sl. po godinama, kvartalima ili mjesecima.

Statističke tabele

Statistički podaci koji su sređeni i grupisani u statističkim serijama često se prikazuju u statističkim tabelama.

Statističke tabele su jednodimenzionalni i višedimenzionalni prikazi statističkih podataka. Sastavljene su iz većeg broja pravougaonih površina koje nastaju ukrštanjem vertikalnih i horizontalnih linija. Ove površine nazivaju se poljima tabele.

Svaka tabela ima zaglavlje i pretkolonu.

Zaglavlje tabele je prvi red u koji se upišu podaci o modalitetima obilježja, vremenskim intervalima ili geografskim područjima.
U pretkolonu se, takođe, upisuju podaci o obilježjima, vremenu i teritoriji.

Svaka statistička tabela ima svoj broj i naslov. Broj i naslov stavljaju se iznad tabele.

Ukoliko u naslovu nije navedena jedinica mjere, oznaka za jedincu mjere najčešće se stavlja iznad tabele, na desnoj strani. Ako su jedinice mjere različite za pojedina obiljažja ili modalitete obilježja, onda se jedinice mjere stavljaju u posebno pripremljenu kolonu.

Ispod tabele upisuju se izvori podataka i odgovarajuće napomene. Sva polja u tabeli moraju biti popunjena.

S obzirom **na sadržinu**, tabele mogu da budu: proste, složene i kombinovane.

-Proste statističke tabele sadrže podatke o strukturi statističkog skupa prema jednom obilježju, ili, pak, promjene jedne pojave u određenom vremenskom periodu.

-Složene statističke tabele nastaju spajanjem prostih tabela i odnose se na jedno obilježje, ili na jedan vremenski period.

-Kombinovane tabele nastaju ukrštanjem serija statističkih podataka o dva ili više obilježja.

-Tabele kontingencije, u kojima se prikazuje raspored jedinica skupa na osnovu dvije ili više klasifikacija prema modaliteta atributivnih obilježja.

Prema namjeni, tabele mogu da budu obradne i analitičke.

-Analitičke tabele imaju za cilj da iskažu određenu vezu između prikazanih obilježja i da omoguće njihovu analizu.

-Obradne statističke tabele služe za dalju obradu podataka i obimnije su od analitičkih, jer sadrže detaljnije podatke o obilježjima i jedinicama skupa.

Podaci iz obradnih tabela mogu se koristiti i za sastavljanje analitičkih tabela i često služe kao dokumentaciona osnova određenog statističkog istraživanja.

DESKRIPTIVNE STATISTIČKE MJERE

Deskriptivne statističke mjere su odgovarajući pokazatelji, dobijeni na osnovu originalnih numeričkih podataka ili rasporeda frekvencija, koji na sintetizovan način opisuju posmatrane podatke.

Ukoliko se odnose na rasporede frekvencija nazivaju se još pokazatelji rasporeda frekvencija.

Smisao ovih mjera je da:

a) jednim brojem opišu bitne karakteristike posmatranih podataka, i

b) da omoguće poređenje između više statističkih serija.

Deskriptivne mjere klasifikujemo u četiri grupe:

1) mjere centralne tendencije rasporeda (srednje vrijednosti),
2) mjere varijabiliteta (disperzije ili raspršenosti),
3) mjere oblika rasporeda, i
4) relativno učešće (proporciju).

Deskriptivne mjere koje se izračunavaju na osnovu svih podataka skupa nazivaju parametri skupa, a deskriptivne mjere koje se odnose na uzorak nazivaju se statistikama uzorka.

U skoro svim statističkim softverima umjesto izraza "statistike uzorka" koristi izraz deskriptivne statistike.

Mjere centralne tendencije

Osnovni zahtjev statističke obrade rasporeda frekvencija jeste da se sa što manje numeričkih karakteristika dobije što potpunija informacija o karakteristikama jedinica posmatranog skupa ili uzorka.

Mjere centralne tendencije predstavljaju sintezu vrijednosti numeričkih obilježja čijom se upotrebom omogućava statistička analiza sa manjim brojem pokazatelja koji opisuju bitne karakteristike jedinica statističkih skupova.

Mjere centralne tendencije (srednje vrijednosti) mogu grubo da se podijele u dvije gupe: izračunate i pozicione.

Izračunate se dobijaju računskim putem na osnovu određene formule. Najčešće korišćene izračunate srednje vrijednosti su: aritmetička, geometrijska i harmonijska sredina.

Pozicione srednje vrednosti se određuju prema položaju koji data srednja vrijednost ima unutar originalnih podataka. U pozicione srednje vrijednosti ubrajamo modus i medijanu.

2.3.1.1 Izračunate srednje vrijednosti

Aritmetička sredina

Aritmetička sredina je u praksi najčešće korišćena mjera centralne tendencije. Popularno se još naziva prosjek.

Posmatrajmo originalne, negrupisane, podatke nekog statističkog skupa i označimo ih sa $x_1, x_2, ..., x_N$.

Aritmetička sredina (obilježena simbolom μ - čitaj mi) skupa (ili posmatranog obilježja) izračunava se na sljedeći način:

$$\mu = \frac{x_1 + x_2 + x_3 + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

gdje je Σ (čitaj sigma, ili suma) univerzalni znak za sabiranje.

Prostije, može se napisati da je $\mu = \frac{\sum x}{N}$.

Ako je u pitanju uzorak veličine *n*, aritmetička sredina (koju ćemo obilježiti sa $\bar{x}$) iz negrupisanih podataka izračunava se na sljedeći način:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{1=1}^{n} x_i$$

ili, jednostavnije: $\bar{x} = \frac{\sum x}{n}$.

Ukoliko su date vrijednosti obilježja $x_1, x_2, ..., x_k$, sa frekvencijama $f_1, f_2, ..., f_k$, tada će se ponderisana aritmetička sredina skupa izračunati na sljedeći način:

$$\mu = \frac{x_1 f_1 + x_2 f_2 + ... + x_k f_k}{N} = \frac{1}{N} \sum_{i=1}^{k} x_i f_i$$

odnosno, jednostavnije:

$$\mu = \frac{\sum xf}{N},$$

gdje je $N = f_1 + f_2 + ... + f_k = \sum_{i=1}^{k} f_i$, odnosno veličina skupa.

Ukoliko su podaci o uzorku veličine ***n*** grupisani u raspored frekvencija, ponderisana aritmetička sredina uzorka biće:

$$\bar{x} = \frac{1}{n} \sum x_i f_i,$$

gdje je $n = f_1 + f_2 + ... + f_k = \sum_{i=1}^{k} f_i$.

Aritmetička sredina

Prosta aritmetička sredina

Za skup

$$\mu = \frac{1}{N}\sum x_i$$

Za uzorak

$$\bar{x} = \frac{1}{n}\sum x_i$$

Ponderisana aritmetička sredina

Za skup

$$\mu = \frac{1}{N}\sum x_i f_i$$

Za uzorak

$$\bar{x} = \frac{1}{n}\sum x_i f_i$$

Kada se koristi aritmetička sredina?

Aritmetička sredina je pod uticajem svih vrijednosti obilježja (uključujući i ekstremno velike i ekstremno male vrijednosti obilježja) zato što se izračunava na osnovu posmatranja svih jedinica. U slučaju postojanja izrazito ekstremnih vrijednosti, aritmetička sredina predstavlja iskrivljeni prikaz onoga što sadrže podaci o posmatranim jedinicama. Iz tog razloga aritmetička sredina nije najbolja mjera centralne tendencije koju treba u takvom slučaju primijeniti.

Osobine aritmetičke sredine

Aritmetička sredina, kao prosječna vrijednost obilježja svih jedinica skupa, **izravnava apsolutne razlike** između podataka posmatrane serije.

1. Aritmetička sredina je srednja vrijednost, sa osobinom da je veća od najmanje i manja od najveće vrijednosti obilježja.

Drugačije rečeno, ako su vrijednosti obilježja poređane po veličini, tj: $x_1 < x_2 < ... < x_N$, može se lako pokazati da je:

$$x_1 < \mu < x_N .$$

2. Ako su sve vrijednosti obilježja međusobno jednake, tj:

$$x_1 = x_2 = ... = x_N = a,$$

tada je i aritmetička sredina jednaka vrijednosti *a*, odnosno $\mu = a$.

3. Zbir odstupanja svih vrijednosti obilježja od njihove aritmetičke sredine jednak je nuli, tj.

$$\sum (x_i - \mu) = 0,$$

odnosno u slučaju grupisanih podataka: $\sum_{i=1}^{k} (x_i - \mu) f_i = 0$.

4. Zbir kvadrata odstupanja svih vrijednosti obilježja od aritmetičke sredine je minimalan, tj. manji je od zbira kvadrata odstupanja svih vrijednosti obilježja od bilo koje druge proizvoljno odabrane vrijednosti:

$$\sum (x_i - \mu)^2 = \min,$$

tj. u slučaju grupisanih podataka:

$$\sum_{i=1}^{k} (x_i - \mu)^2 f_i = \min.$$

5. Ako su vrijednosti dva obilježja povezane nekom linearnom funkcijom (vezom), tada su i njihove

aritmetičke sredine povezane istom tom linearnom funkcijom (vezom). Na primjer, ako su vrijednosti obilježja X *i* Y vezane linearnom funkcijom oblika:

$$Y = aX + b,$$

tada su i njihove aritmetičke sredine vezane istom funkcijom, odnosno:

$$\mu_Y = a\mu_X + b.$$

Geometrijska sredina

Geometrijska sredina predstavlja mjeru centralne tendencije koja izravnava proporcionalne promjene između podataka posmatrane serije.

Ako se posmatra obilježje X, sa vrijednostima $x_1, x_2, ..., x_N$, geometrijska sredina će se izračunati na sljedeći način:

$$G = \sqrt[N]{x_1 \cdot x_2 \cdot .... \cdot x_N} = \sqrt[N]{\prod_{i=1}^{N} x_i}.$$

Na ovaj način, iz negrupisanih podataka, izračunava se **prosta geometrijska sredina**.

Izračunavanje ove srednje vrijednosti ima smisla samo ukoliko su vrijednosti obilježja veće od nule.

Polazeći od navedenog uslova, geometrijska sredina se jednostavnije izračunava logaritmovanjem prethodnog izraza, tako da dobijamo:

$$log\ G = \frac{\log x_1 + \log x_2 + ... + \log x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} \log x_i\ .$$

Iz toga je geometrijska sredina jednaka vrijednosti antilogaritma**:**

$$G = \text{antilog.}\ (\frac{1}{N}\sum_{i=1}^{N} \log x_i)$$

Za razliku od računanja proste geometrijske sredine, ako su dati rasporedi frekvencija u kojima se vrijednosti obilježja javljaju više puta, potrebno je izračunati **ponderisanu geometrijska sredina** prema sljedećem obrascu:

$$G = \sqrt[N]{x_1^{\,f_1} \cdot x_2^{\,f_2} \cdot .... \cdot x_k^{\,f_k}}\ .$$

Logaritmovanjem lijeve i desne strane prethodnog izraza dobija se:

$$log\ G = \frac{1}{N}\sum_{i=1}^{k} f_i \log x_i\ ,$$

gdje je $N = f_1 + f_2 + ... + f_k = \sum f_i$.

Geometrijska sredina dobija se antilogaritmovanjem:

$$G = \text{antilog}.\left(\frac{1}{N} \sum_{i=1}^{k} f_i \log x_i \right)$$

Geometrijska sredina je manja od aritmetičke sredine. To je opšte pravilo, izuzimajući slučaj kada su sve vrijednosti jednake.

Međutim, **treba istaći da se geometrijska sredina veoma rijetko koristi kao mjera centralne tendencije obilježja.**

Najvažnija osobina geometrijske sredine je da izravnava proporcionalne promjene podataka u seriji.

Zbog toga je **proizvod odnosa geometrijske sredine prema manjim vrijednostima serije jednak proizvodu odnosa većih vrijednosti prema geometrijskoj sredini**, što je jako korisno **u istraživanju dinamike ekonomskih pojava.**

Harmonijska sredina

Harmonijska sredina se koristi kada su obilježja elemenata iz kojih se izračunava izražena u vidu

recipročnih pokazatelja.

Harmonijska sredina je naročito značajna **u istraživanju pokazatelja produktivnosti rada, pokazatelja brzine obrta poslovnih sredstava, brzine prelaska određenog puta** i sl.

Harmonijska sredina

Harmonijska sredina predstavlja recipročnu vrijednost aritmetičke sredine iz recipročnih vrijednosti obilježja.

Ako su vrijednosti obilježja date kao negrupisane, odnosno ako svaka data vrijednost ima frekvenciju jedan, harmonijska sredina će se izračunati kao prosta sredina, na sljedeći način:

$$H = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}}.$$

Odnosno, kraće, $H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$.

Ukoliko su vrijednosti obilježja grupisane tako da svaka od njih ima frekvenciju veću od jedan, tada se izračunava **ponderisana harmonijska sredina**, prema sljedećem obrascu:

$$H = \frac{f_1 + f_2 + \ldots + f_k}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_k}{x_k}}.$$

Ili, kraće napisano:

$$H = \frac{N}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

gdje je $N = f_1 + f_2 + ... + f_k$.

Upotreba harmonijske sredine ograničena je na pojave čiji podaci su dati kao recipročne vrijednosti i da **računanje ove sredine nema smisla ukoliko je neka vrijednost obilježja jednaka nuli.**

Pozicione srednje vrijednosti

Modus

Pozicione srednje vrijednosti se **određuju na osnovu njihovog mjesta,** odnosno lokacije u seriji.

Kao što ćemo vidjeti, na njihovu veličinu ne utiču ekstremne vrijednosti obilježja.

Modus

Modus je vrijednost obilježja koja se najčešće javlja u seriji, odnosno vrijednost obilježja sa najvećom frekvencijom.

Ukoliko je data negrupisana serija podataka o pojavi u kojoj svaka vrijednost obilježja ima istu frekvenciju

jednaku 1, modus ne postoji.

Međutim, ako u negrupisanoj seriji postoje vrijednosti obilježja sa različitim frekvencijama, onda se modus utvrđuje pronalaženjem vrijednosti obilježja koja se najčešće javlja.

Kod serija sa atributivnim obilježjima modus se određuje na isti način kao i kod rasporeda frekvencija sa prekidnim vrijednostima numeričkih obilježja.

Jedan od nedostataka modusa kao srednje vrijednosti je da postoje serije koje nemaju modus.

Serije mogu imati dva modusa i nazivaju se bimodalne serije.

Ovo je **još jedan nedostatak modusa**, jer ne znamo za koju od navedene dvije vrijednosti da se opredijelimo.

Generalno, ako neka serija ima više modusa naziva se multimodalna serija.

Kod rasporeda frekvencija sa neprekidnim vrijednostima obilježja koje su raspoređene u vidu intervalnih modaliteta obilježja, **koristimo posebnu formulu za izračunavanje modusa.**

Međutim, tako dobijena vrijednost modusa je samo približna.

Promjenom veličine grupnih intervala ili njihovih

granica pri istim intervalima mogu se dobiti različite vrijednosti modusa. Ovo je treći nedostatak modusa.

Za izračunavanje modusa potrebno je najprije pronaći interval sa najvećom frekvencijom (modalni interval).

Tada se modus određuje tzv. interpolacijom, odnosno uzimanjem u obzir i frekvencija dvaju susjednih intervala:

$$Mo = L_1 + \frac{f_2 - f_1}{(f_2 - f_1) + (f_2 - f_3)} \cdot i,$$

gdje su:

L_1 - donja granica modalnog intervala,

f_2 - frekvencija modalnog intervala,

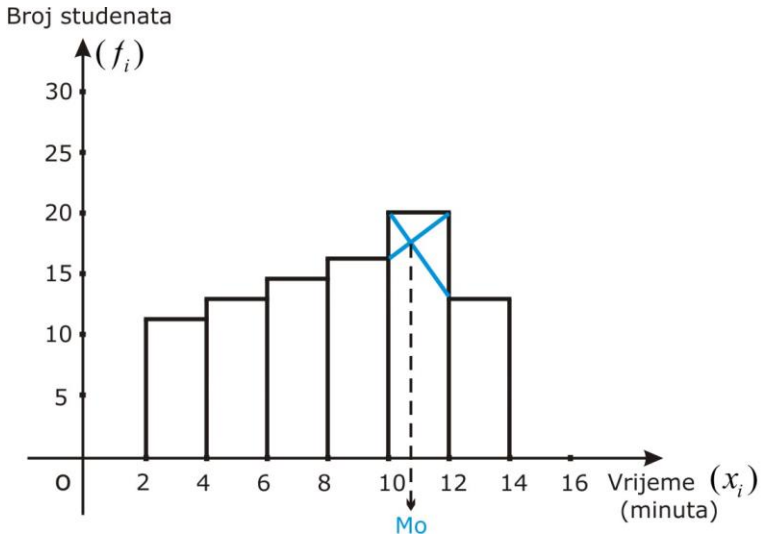
f_1 - frekvencija predmodalnog intervala,

f_3 - frekvencija poslemodalnog intervala,

i - veličina modalnog intervala.

Prilikom tumačenja modusa kažemo "moglo bi se reći" da modus iznosi..., jer kod rasporeda frekvencija sa neprekidnim obilježjem modus gubi svojstvo tipične, odnosno najčešće vrijednosti.

Modus se može približno odrediti i grafički, na osnovu histograma frekvencija.



Slika 2.10 Približno određivanje modusa (grafičkim putem)

Medijana

Medijana čitav skup dijeli na dva jednaka dijela, odnosno polovina jedinica skupa ima vrijednost obilježja manju, a polovina jedinica vrijednost obilježja veću od medijane.

Na medijanu ne utiču ekstremne vrijednosti obilježja.

Zbog toga, **kad su u seriji prisutne jedinice skupa sa ekstremnim malim ili ekstremno velikim vrijednostima obilježja, medijana, kao mjera centralne tendencije, realnije opisuje cijeli skup nego aritmetička sredina.**

Medijana

Medijana je vrijednost obilježja koja se nalazi u sredini serije čiji su podaci sređeni po

veličini.

Da bi se izračunala medijana iz serije podataka, potrebno ih je najprije urediti po veličini od najmanjeg ka najvećem, odnosno kao niz:

$$x_1, x_2, ..., x_N .$$

Tada će medijana biti ona vrijednost obilježja koja se nalazi u sredini formiranog niza.

Prilikom određivanja medijane **bitno je voditi računa o tome da li je broj podataka serije paran ili neparan.**

Mjesto medijane određuje se kao $\frac{N+1}{2}$ podatak u seriji.

Ukoliko je neparan broj podataka serije, medijana će biti vrijednost obilježja koja predstavlja središnji član posmatranog niza.

Međutim, ako se posmatra serija sa parnim brojem podataka, medijana se određuje kao prosjek dviju vrijednosti obilježja koje predstavljaju središnje članove serije.

Međutim, ako su podaci grupisani u distribucije frekvencija sa intervalnim modalitetima obilježja, tada se medijana, kao i modus, izračunava pomoću posebne, aproksimativne formule.

Za grupisane serije podataka medijana će se izračunati na sljedeći način:

$$Me = L_1 + \frac{\frac{N}{2} - \sum f_1}{f_{Me}} \cdot i,$$

gdje je:

L_1 - donja granica medijalnog intervala,

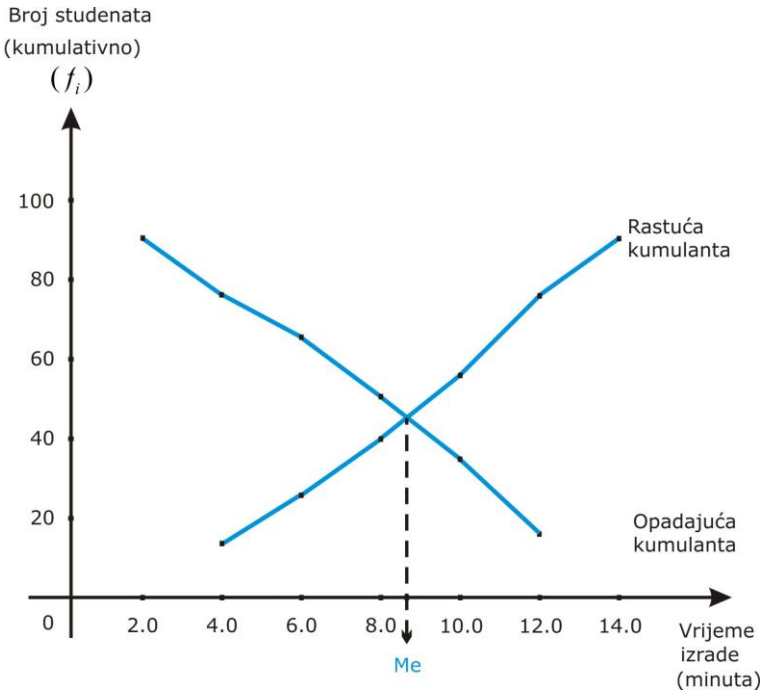
N - broj podataka u seriji,

$\sum f_1$ - suma frekvencija predmedijalnog intervala,

f_{Me} - frekvencija medijalnog intervala.

Potrebno je najprije locirati medijalni interval. Njega ćemo odrediti na osnovu kumuliranih frekvencija. Prvom kumulativu koji u sebi sadrži polovinu frekvencija odgovara medijalni interval.

Medijana se može približno odrediti grafičkim putem, koristeći kumulirane frekvencije. Ukoliko se koristi grafički prikaz rastuće i opadajuće kumulante frekvencija, medijana će se približno odrediti u tački presjeka navedenih kumulanti spuštenoj na *X-osu*.



Slika 2.11 Približno određivanje medijane (grafičkim putem)

Obratimo pažnju da kumulante nisu spojene sa X osom, jer prvi i posljednji interval nisu zatvoreni.

Kvartili

Kvartili su mjere koje dijele seriju podataka na četiri jednaka dijela.

Ako se ukupan broj članova serije podijeli na četiri jednaka dijela, vrijednosti obilježja koje ih dijele nazivaju se kvartilima: prvi kvartil Q_1, drugi kvartil Q_2 i treći kvartil Q_3.

Medijana, u suštini, predstavlja drugi kvartil. **Kvartile ćemo koristiti da opišu svojstva serija numeričkih podataka,** kao i kod izračunavanja interkvartilne razlike.

Prvi kvartil Q_1

Prvi kvartil (Q_1) je vrijednost obilježja za koju je 25% jedinica manje, a 75% jedinica skupa veće od date vrijednosti.

Mjesto Q_1: $\frac{N+1}{4}$, za seriju podataka uređenih po veličini.

Izračunavanje prvog kvartila za raspored frekvencija sa intervalno datim podacima:

$$Q_1 = L_1 + \frac{\frac{N}{4} - \sum f_1}{f_{Q_1}} \cdot i .$$

Treći kvartil Q_3

Treći kvartil (Q_3) je vrijednost obilježja za koju je 75% jedinica manje, a 25% jedinica veće od date vrijednosti.

Mjesto Q_3: $\frac{3(N+1)}{4}$, za podatke uređene po veličini.

Izračunavanje trećeg kvartila za intervalno date rasporede frekvencija:

$$Q_3 = L_1 + \frac{\frac{3N}{4} - \sum f_1}{f_{Q_3}} \cdot i .$$

gdje su:

L_1 - donja granica kvartilnog intervala,

N - broj članova serije,

$\sum f_1$ - zbir frekvencija predkvartilnog intervala,

f_{Q_1}, f_{Q_3} - frekvencije kvartilnih intervala,

i – širina kvartilnog intervala.

Iako je određivanja kvartila bitno, jer se na osnovu njih formuliše jedna važna mjera disperzije (interkvartilna razlika), naglasimo da u statističkoj literaturi ne postoji usaglašenost o tome kako se oni konkretno izračunavaju.

Štaviše, razni statistički softveri daju različite

izračunate vrijednosti za kvartile. Sa povećavanjem broja podataka razlike između izračunatih kvartila postaju zanemarljive.

Metod koji ćemo koristiti je jednostavan, dovoljno precizan, i ne zahtijeva primjenu interpolacije, koja bi bila neophodna da bi se odredile egzaktne vrijednosti.

Određivanje kvartila negrupisanih podataka sprovešćemo kroz četiri etape.

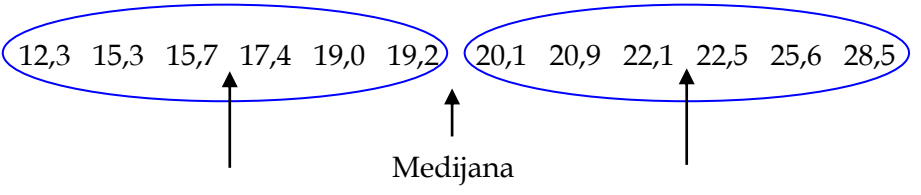
1. U prvom koraku originalni podaci se uređuju po veličini, od najmanjeg ka najvećem.

2. U drugom koraku određuje se medijana.

3. Prvi kvartil određujemo kao medijanu podataka koji su manji od medijane, odnosno nalaze se ulijevo od medijane.

4. Treći kvartil je medijana podataka koji su veći od medijane, tj. nalaze se udesno od medijane.

Pokažimo kako se praktično sprovodi ovaj postupak u našem primjeru.



Medijana = (15,7 + 17,4)/2
= 16,55

Medijana = (22,1 + 22,5)/2
= 22,3

Zaključujemo da prvi kvartil Q_1 iznosi 16,55, a treći Q_3 je 22,3.

Ako se broj članova serije podijeli na 10 ili na 100 jednakih dijelova, **dobiće se decili, odnosno percentili.**

Ovi pokazatelji izračunavaju se na sličan način kao i medijana, odnosno kvartili.

Mjere varijacije

Može se desiti da izračunata srednja vrijednost bude potpuno ista za različite serije podataka, tako da ona ne može biti dovoljna karakteristika svih posmatranih jedinica sa stanovišta njihovog varijabiliteta.

Zbog toga je neophodno utvrditi i odgovarajuću **mjeru varijabiliteta (disperzije ili raspršenosti)** kojom će se dopuniti informacija o karakteru statističkog skupa ili uzorka.

Varijabilitet predstavlja raspršenost podataka u seriji. Utvrđuje se pomoću odgovarajućih mjera varijacije koje, kao i mjere lokacije, mogu da budu pozicione i izračunate u odnosu na neku srednju vrijednost (najčešće aritmetičku sredinu) skupa ili uzorka.

Ukoliko su **mjere varijacije izražene u jedinicama vrijednosti obilježja**, govori se o apsolutnim mjerama, za razliku od **mjera koje se izražavaju u relativnim pokazateljima** (procentualno ili u broju standardizovanih jedinica).

Kao apsolutne mjere varijabiliteta, koje istovremeno predstavljaju pozicione mjere varijacije, koriste se interval (razmak) varijacije i interkvartilna razlika.

Pored njih, u apsolutne mjere spadaju i mjere zasnovane na odstupanjima podataka od aritmetičke

sredine, kao što su: srednje apsolutno odstupanje, varijansa i standardna devijacija.

Od relativnih mjera varijacije najširu upotrebu imaju: koeficijent varijacije i standardizovano odstupanje.

2.3.2.1 Apsolutne mjere varijacije

Interval (razmak) varijacije

Interval ili razmak varijacije se utvrđuje kao razlika između najveće i najmanje vrijednosti obilježja posmatranog rasporeda:

$$I = X_{\max} - X_{\min}$$

Interval varijacije daje samo približnu informaciju o disperziji posmatrane serije, jer na njegovu veličinu utiču samo dvije krajnje vrijednosti obilježja. Ukoliko su te dvije vrijednosti ujedno i ekstremne u seriji, ova mjera će biti nerealno velika i na pogrešan način odslikavati varijacije svih podataka u seriji. Takođe, interval varijacije nije osjetljiv na veličinu serije podataka, što je njegov drugi nedostatak.

Interval varijacije

Interval varijacije je jednak razlici najveće i najmanje vrijednosti obilježja. Može se izračunati samo za konačne skupove podataka.

Interkvartilna razlika

Interkvartilna razlika je raspon između prvog i trećeg kvartila:

$$IQR = Q_3 - Q_1$$

Budući da ova mjera u obzir uzima samo raspon centralnih 50% podataka u sredini serije, jasno je da **ne zavisi od ekstremnih vrijednosti.**

Da sumiramo: kvartili, pa samim tim i interkvartilna razlika, ne mogu biti pod uticajem ekstremnih vrijednosti, jer se jedinice sa vrijednostima obilježja manjim od prvog kvartila ili većim od trećeg kvartila ne uzimaju u obzir. Statističke mjere kao što su medijana, kvartili i interkvartilna razlika, na koje ne utiču ekstremne vrijednosti, nazivaju se rezistentnim ili robustnim mjerama.

Interkvartilna razlika

Interkvartilna razlika predstavlja razliku između trećeg i prvog kvartila.

Srednje apsolutno odstupanje

Postavlja se pitanje: na koji način formulisati takvu mjeru disperzije koja bi bila pod uticajem svih vrijednosti u seriji? U statistici se u tu svrhu najčešće uzima aritmetička sredina.

Pošto je zbir odstupanja svih podataka od njihove aritmetičke sredine jednak nuli, odnosno $\Sigma(x_i - \mu)=0$, jasno je da sâm zbir odstupanja ne možemo uzeti kao mjeru disperzije.

Kao prva ideja javlja se uzimanje apsolutnih vrijednosti svih odstupanja, odnosno $\Sigma|x_i - \mu|$. Takva mjera, međutim, imala bi veliki nedostatak, jer bi se automatski povećavala sa povećavanjem broja podataka. Da bismo to prevazišli, jednostavno ćemo podijeliti navedenu sumu sa brojem podataka.

Budući da dijelimo sa brojem podataka, dobijena mjera će pokazivati prosjek sume apsolutnih odstupanja svih podataka od njihove aritmetičke sredine. Takva mjera varijabiliteta naziva se srednje apsolutno odstupanje.

Konkretno, srednje apsolutno odstupanje računa se za negrupisane serije na sljedeći način:

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$$

a za grupisane serije, odnosno rasporede frekvencija, primjenom sljedećeg obrasca:

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{k} f_i |x_i - \mu|$$

gdje je $N = f_1 + f_2 + ... + f_k = \sum f_i$.

Varijansa

Srednje apsolutno odstupanje se rijetko koristi u statistici, jer u sebi sadrži apsolutne vrijednosti koje su nepogodne za dalju matematičku obradu. Ponovo se nalazimo pred istim pitanjem: **kako eliminisati negativna odstupanja, a da ne koristimo apsolutne vrijednosti?**

Odgovor je jednostavan – uzećemo kvadrate odstupanja.

Tako dobijena mjera disperzije pokazuje, stoga, prosjek

sume kvadrata odstupanja svih podataka od njihove aritmetičke sredine i naziva se varijansa ili srednje kvadratno odstupanje.

Ukoliko podaci nisu grupisani, odnosno ako su date vrijednosti obilježja X: x_1, x_2, x_3,..., x_N, varijansa skupa se izračunava na sljedeći način:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

a za podatke grupisane u vidu distribucije frekvencija:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \mu)^2$$

jer odstupanja moramo ponderisati njihovim frekvencijama. Varijansa se može izračunati i na jednostavniji način, koristeći **tzv. radni obrazac:**

$$\sigma^2 = \frac{\sum_{i=1}^{k} x_i^2 f_i}{N} - \mu^2 .$$

Na ovaj način varijansa se izračunava neposredno iz samih vrijednosti posmatranog obilježja, a ne preko odstupanja podataka od aritmetičke sredine.

Varijansa

Varijansa pokazuje prosjek kvadrata odstupanja svih podataka od njihove aritmetičke sredine.

Varijansa uzorka predstavlja prosjek zbira kvadrata odstupanja svih vrijednosti obilježja jedinica u uzorku od aritmetičke sredine uzorka. Kod negrupisanih podataka izračunava se na sljedeći način:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Primjetimo da se zbir kvadrata odstupanja kod varijanse uzorka dijeli sa $n - 1$, a ne sa N, kao kod varijanse skupa.

Kao što ćemo kasnije vidjeti, takav pokazatelj se naziva ocjena. U teorijskoj statistici je pokazano da se preciznija ocjena varijanse skupa dobija ako se suma kvadrata odstupanja podijeli sa $n - 1$, a ne sa n.

Standardna devijacija

Iako varijansa, kao apsolutna mjera varijacije, ima široku primjenu u statističkim istraživanjima, ona ima i

jedan značajan nedostatak. Naime, radi se o kvadriranju odstupanja podataka od aritmetičke sredine, pa je ona iskazana u kvadratima mjernih jedinica (kao, npr KM^2, godine starosti na kvadrat i sl.). Također, time se znatno povećava veličina izračunate mjere varijabiliteta.

Da bi se taj nedostatak otklonio, izračunava se kvadratni korijen iz varijanse i dobija se najčešće korišćena mjera apsolutnog varijabiliteta poznata kao standardna devijacija.

Standardna devijacija se može izračunati direktno iz varijanse, odnosno kao pozitivna vrijednost kvadratnog korijena varijanse, tj:

$\sigma = +\sqrt{\sigma^2}$ za skup, ili $s = +\sqrt{s^2}$ za uzorak.

Ukoliko nije prethodno izračunata varijansa u skupu, standardna devijacija se može izračunati na osnovu podataka o odstupanjima od aritmetičke sredine, na sljedeći način:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

ili direktno iz podataka:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^{\,2} - \mu^2}\ .$$

Standardna devijacija skupa **na osnovu grupisanih podataka** u vidu rasporeda frekvencija izračunava se na sljedeći način:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \mu)^2}$$

Kao i varijansa, standardna devijacija podataka skupa datih u vidu rasporeda frekvencija se može odrediti i direktno iz podataka:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} x_i^{\,2} f_i - \mu^2}.$$

Kada raspolažemo rasporedom frekvencija uzorka, standardna devijacija uzorka se može izračunati jednostavnije, pomoću "radne formule":

$$s = \sqrt{\frac{\sum f_i x_i^{\,2} - n\bar{x}^{\,2}}{n-1}}$$

Standardna devijacija pokazuje da prosječno odstupanje vremena izrade zadatka svih studenata od prosječnog vremena iznosi 3,24 minuta.

Standardna devijacija

Standardna devijacija predstavlja prosječno odstupanje svih pojedinačnih podataka od njihove aritmetičke sredine**.**

Da sumiramo: iako **varijansa posjeduje korisna matematička svojstva,** njena vrijednost je uvijek iskazana u mjernim jedinicama na kvadrat. To je razlog što je **standardna devijacija najčešće korišćena mjera disperzije, budući da je njena vrijednost izražena u originalnim jedinicama mjere obilježja.**

Interesantno je da je standardna devijacija kao statistička mjera uvedena mnogo ranije od varijanse. Nju je formulisao Karl Pearson 1893. godine. Tek kada se ukazala potreba za složenijom statističkom analizom, 1918. godine, Ronald Fisher je uveo pojam varijanse.

Razumijevanje varijabliteta podataka u seriji

1. Što su podaci više raspršeni u posmatranoj seriji, veći će biti interval varijacije, interkvartilna razlika, varijansa i standardna devijacija.
2. Što su podaci više skoncentrisani, ili homogeni, biće manji interval varijacije, interkvartilna razlika, varijansa i standardna devijacija.
3. Ako sve jedinice skupa (ili uzorka) imaju istu vrijednost obilježja (istu vrijednost imaće i aritmetička sredina, tako da neće biti ni varijacije među podacima), tada će sve mjere varijacije biti jednake nuli.
4. Nijedna od apsolutnih mjera varijacije (interval varijacije, interkvartilna razlika, varijansa i standardna devijacija) ne mogu imati negativnu vrijednost.

Relativne mjere varijacije

Za razliku od mjera varijacije koje su izražene u apsolutnim jedinicama vrijednosti obilježja, **relativne mjere izražene su u procentima ili u jedinicama standardne devijacije.**

Ove mjere omogućavaju da se upoređuje varijabilitet numeričkih serija podataka koji su izraženi u različitim jedinicama mjere, ili serija čija su obilježja izražena istim jedinicama, ali sa različitim aritmetičkim sredinama.

Najčešće korišćene relativne mjere varijacije su: koeficijent varijacije i standardizovano odstupanje.

Koeficijent varijacije predstavlja relativni odnos standardne devijacije i aritmetičke sredine. Izračunava se na sljedeći način:

$$Kv = \frac{\sigma}{\mu}$$

Koeficijent varijacije se često izražava u procentima[2] i pokazuje koliko procentualno iznosi standardna devijacija od aritmetičke sredine. Njegove **velike vrijednosti ukazuju na relativno veliki stepen varijabilnosti podataka u seriji, i suprotno, male na relativno malu disperziju u skupu ili uzorku.**

PREPORUKA: Prilikom poređenja disperzije

[2] Iako se koeficijent varijacije iskazuje u procentima, njegova maksimalna vrijednost može biti veća od 100%, kao, na primjer, kod tzv. hipereksponencijalnog rasporeda.

dvije ili više serija podataka koristite isključivo koeficijent varijacije.

Standardizovano odstupanje predstavlja mjeru odstupanja nekog pojedinačnog podatka od aritmetičke sredine izražena u jedinicama standardne devijacije. Izračunava se na sljedeći način:

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

Standardizovano odstupanje, kao i koeficijent varijacije, omogućava upoređivanje varijabiliteta individualnih podataka više pojava, zbog toga što se odstupanja podataka od aritmetičke sredine ne izražavaju u apsolutnim jedinicama mjere obilježja.

Značaj standardizovanog odstupanja kao mjere disperzije proizilazi iz činjenice da se varijabilitet ne ocjenjuje samo sa gledišta rasporeda frekvencija kao cjeline, nego i sa gledišta pojedinačnih podataka u skupu ili uzorku.

Mjere oblika rasporeda frekvencija

Rasporedi frekvencija, odnosno serije distribucije frekvencija, imaju različite oblike u pogledu načina rasporeda članova serije, s obzirom na vrijednosti obilježja. Navedene razlike se uglavnom odnose na simetričnost i spljoštenost, odnosno zaobljenost (ili, suprotno, izduženost) rasporeda frekvencija.

Za mjerenje oblika rasporeda frekvencija prema osi simetrije ili u pogledu zaobljenosti najčešće se koriste tzv. centralni momenti. Centralni momenti sukcesivno mjere prosječna odstupanja podataka, nultog, prvog, drugog ili n-tog stepena, u odnosu na aritmetičku sredinu.

Centralni statistički moment r-tog reda definisan je na sljedeći način:

$$M_r = \frac{1}{N}\sum_{i=1}^{k} f_i \left(x_i - \mu\right)^r$$

Centralni moment nultog reda uvijek je jednak 1. Centralni moment prvog reda jednak je nuli (on pokazuje da je prosjek odstupanja svih podataka od aritmetičke sredine jednak nuli), pa se, kao i nulti

moment, ne može upotrijebiti za mjerenje oblika rasporeda. Lako je shvatiti da je centralni moment drugog reda, u stvari, varijansa.

Za mjerenje oblika rasporeda ostaju nam, dakle, centralni momenti viših redova od drugog. Za mjerenje simetričnosti koristi se treći centralni momenat.

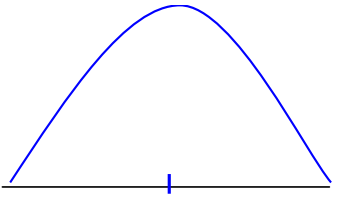
Za mjerenje oblika rasporeda u pogledu spljoštenosti (zaobljenosti) koristi se četvrti centralni momenat.

2.3.3.1 Mjere asimetrije

Za simetričnu distribuciju karakteristično je da svakom odstupanju vrijednosti obilježja od aritmetičke sredine negativnog predznaka odgovara isto toliko odstupanje pozitivnog predznaka.

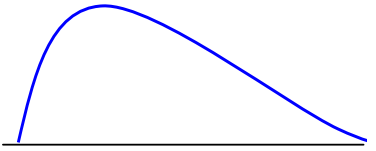
Ako je raspored pozitivno asimetričan, pozitivna i negativna odstupanja neće se izravnati nego će preovladati odstupanja sa pozitivnim predznakom. Suprotno, kod negativno asimetričnih rasporeda preovladaće odstupanja sa negativnim predznakom.

Na Slici 2.12 grafički su prikazane ove tri mogućnosti.

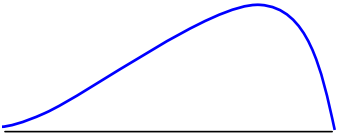


μ = Modus = Medijana

a) Simetričan raspored



b) Raspored asimetričan udesno
(pozitivno asimetričan)



c) Raspored asimetričan ulijevo
(negativno asimetričan)

Slika 2.12 Grafički prikaz oblika rasporeda prema simetričnosti

Treći centralni momenat za negrupisane podatke dat je sljedećim izrazom:

$$M_3 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^3,$$

dok se, za distribucije frekvencija, treći centralni momenat izračunava na sljedeći način:

$$M_3 = \frac{1}{N}\sum_{i=1}^{k} f_i\left(x_i - \mu\right)^3.$$

Ukoliko je raspored frekvencija simetričan, prethodni izraz biće jednak nuli, jer neparan eksponent ne mijenja

predznak odstupanja.

Za pozitivno asimetrične rasporede treći centralni momenat je veći od nule, dok je za negativno asimetrične rasporede manji od nule.

Zbog navedenih osobina, treći centralni momenat se može koristiti za utvrđivanje mjere stepena i smjera asimetrije rasporeda. Međutim, pošto vrijednost trećeg centralnog momenta zavisi od jedinice mjere obilježja X, zaključivanje o asimetriji rasporeda nije moguće samo na osnovu njegove apsolutne veličine.

Zbog toga se za mjeru asimetrije koristi količnik trećeg centralnog momenta i standardne devijacije dignute na treći stepen, koji se obilježava sa α_3.

Pokazatelj asimetrije, dakle, dat je sljedećim izrazom:

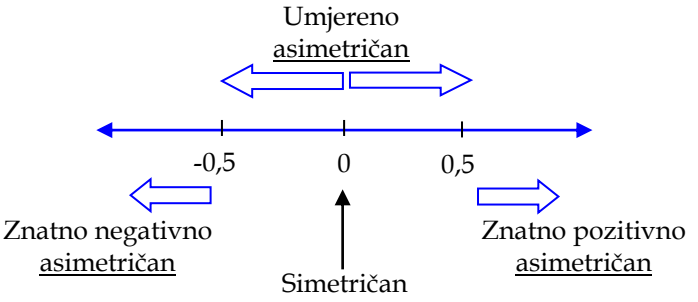
$$\alpha_3 = \frac{M_3}{\sigma^3}$$

Iz navedenog obrasca zaključuje se da je pokazatelj asimetrije relativna mjera stepena i smjera asimetrije, koja je za egzaktno simetrične rasporede jednaka nuli.

Što je veća apsolutna vrijednost ovog koeficijenta, veća je i asimetrija posmatranog rasporeda. Smatraćemo da

je raspored umjereno asimetričan ako se vrijednost α_3 nađe u intervalu -0,5 do +0,5. U suprotnom, kazaćemo da je raspored znatno asimetričan.

Različita tumačenja asimetričnosti s obzirom na vrijednost α_3 prikazana su na Slici 2.13.



Slika 2.13 Interpretacija stepena asimetrije s obzirom na vrijednost α_3

Iako se u najvećem broju knjiga iz statistike navodi da se smjer asimetrije može približno ocijeniti i na osnovu odnosa između medijane i modusa, u odnosu na aritmetičku sredinu, mi to ne bismo preporučili.[3]

[3] Paul Hippel sa Ohio State univerziteta, u članku "Mean, Median and Skew: Correcting a Textbook Rule", *Journal of Statistics Education*, Volume 13, Number 2, je 2005. godine pokazao da "pravilo" po kojem je raspored pozitivno asimetričan, ako je aritmetička sredina veća od medijane (i obrnuto, da je raspored negativno asimetričan kada je medijana veća od aritmetičke sredine) ima isuviše veliki broj izuzetaka. Preciznije, ono je netačno kod velikog broja prekidnih rasporeda. Kod neprekidnih rasporeda bi moglo grubo da se primijeni, pod uslovom da raspored ima samo jedan modus. Članak se može naći i na Internetu: (www.amstat.org/publications/jse/v13n2/vonhippel.html).

Pokazatelji spljoštenosti rasporeda

Zaobljenost u okolini modalnog maksimuma krive distribucije frekvencija mjeri se koeficijentom spljoštenosti (izduženosti), koji se bazira na četvrtom centralnom momentu.

Četvrti centralni momenat za podatke date u vidu distribucije frekvencija dat je sljedećim izrazom:

$$M_4 = \frac{1}{N}\sum_{i=1}^{k} f_i \left(x_i - \mu\right)^4$$

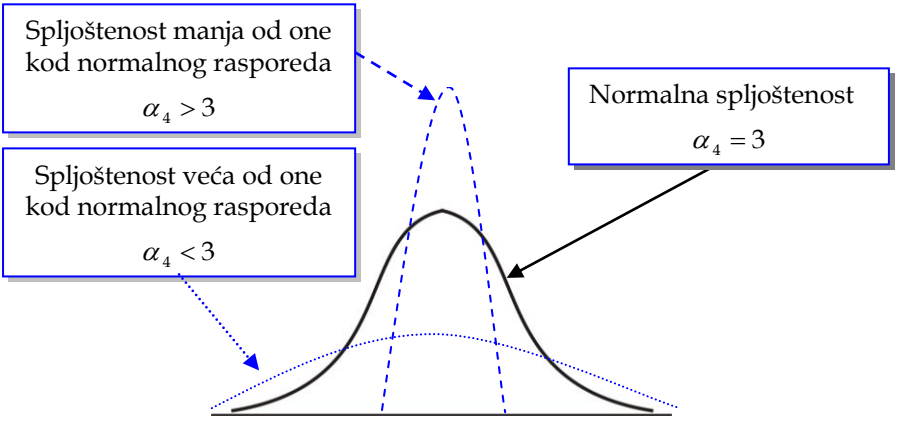
Stavljanjem u odnos navedenog centralnog momenta i standardne devijacije na četvrti stepen, dobija se relativna mjera spljoštenosti rasporeda.

Označava se sa α_4:

$$\alpha_4 = \frac{M_4}{\sigma^4}$$

Prilikom tumačenja relativne mjere spljoštenosti, **dobijenu vrijednost uspoređujemo sa teorijskom spljoštenošću tzv. normalnog rasporeda frekvencija, koja iznosi 3.**

Ako je **koeficijent spljoštenosti veći od 3,** tada je raspored više izdužen, a **ako je vrijednost koeficijenta manja od 3**, tada je serija više spljoštena u odnosu na normalnu distribuciju. Grafički, to se može prikazati kao na Slici 2.14.



Slika 2.13 Oblici rasporeda u pogledu spljoštenosti

Napomenimo da se u **najvećem broju statističkih softvera mjera spljoštenosti modifikuje tako što se od α_4 u izrazu (2.24) oduzima 3.** Jasno je da tako formulisana mjera spljoštenosti kod normalnog rasporeda iznosi 0. Prilikom tumačenja kompjuterskog izlaza treba voditi računa da se koeficijent spljoštenosti poredi sa 0, a ne sa 3.

OPIS OSNOVNIH KARAKTERISTIKA RASPOREDA

Vidjeli smo da **kod opisivanja centralne tendencije i disperzije postoji veći broj različitih mjera.** Postavlja se pitanje, koje od njih odabrati i opisati konkretan raspored, dobijen na osnovu uzorka ili popisa.

Predložićemo da se raspored, uz pomoć odgovarajućeg statističkog softvera, opiše na osnovu sljedećih etapa.

1. Najprije je potrebno **odrediti relativnu mjeru asimetrije α_3.**

2. U drugom koraku ispituje se **da li serija sadrži, i koliko,**

ekstremnih vrijednosti.

U tu svrhu najbolje je koristiti poseban dijagram koji se naziva boxplot. Ako u seriji ima ekstremnih vrijednosti, na boxplotu su prikazane zvjezdicama. Najčešće se

neki podatak označava kao ekstremni (outlier) ako je on $1{,}5 \times IQR$ veći (ili manji) od trećeg (prvog) kvartila.

3. U trećoj etapi **određujemo onu mjeru centralne tendencije**

koja najviše odgovara našim podacima. Poznato nam je da aritmetička sredina nije dobar pokazatelj ukoliko serija sadrži ekstremne vrijednosti. Također, vidjeli smo da je medijana rezistentna na ekstremne vrijednosti.

Međutim, **ako podaci ne sadrže takve opservacije, aritmetička sredina je bolja mjera, jer zavisi od svih podataka.**

Na osnovu navedenog formulisaćemo sljedeću preporuku:

Ako serija ne sadrži ekstremne vrijednosti i ako je umjereno asimetrična, najbolja mjera centralne tendencije je aritmetička sredina. U suprotnom, ako serija sadrži ekstremne vrijednosti, ili ako je znatno asimetrična, bolje je koristiti medijanu.

4. U posljednjoj etapi određujemo **koja mjera disperzije najviše**

odgovara konkretnim podacima. Pri tome, rukovodimo se sličnom logikom kao u prethodnoj.

Naime, budući da se standardna devijacija izračunava na osnovu aritmetičke sredine, ona može nerealno prikazivati varijacije podataka u prisustvu ekstremnih

vrijednosti. Sa druge strane, interkvartilna razlika je robustna na takve vrijednosti.

Stoga preporučujemo:

A) Ako serija ne sadrži ekstremne vrijednosti i ako je umjereno asimetrična, najbolja mjera disperzije je standardna devijacija.

B) U suprotnom, ako serija sadrži ekstremne vrijednosti, ili ako je znatno asimetrična, bolje je koristiti interkvartilnu razliku.

Dakle, aritmetičku sredinu uvijek "uparujemo" sa standardnom devijacijom, a medijanu sa interkvartilnom razlikom.

PRIMJER:

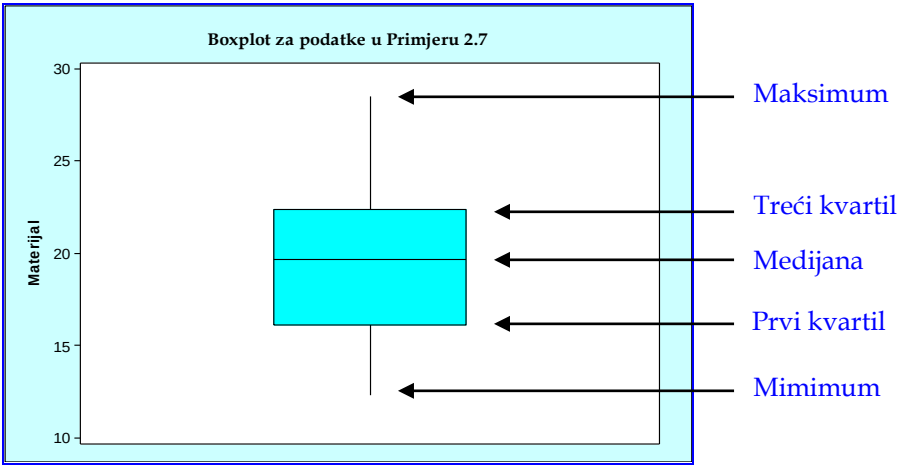
Opišimo podatke serije iz Primjera 2.7 o količini download-ovanog materijala sa Interneta.

1. Najprije ćemo da odredimo relativnu mjeru asimetrije korišćenjem izlaza iz 3BStata, datog sljedećom tabelom:

Deskriptivne Statistike							
Varijabla:					Materijal		
n	*Min.*	*Maks.*	*Arit. Sred.*	*Modus*	Q_1	*Medijana*	Q_3
12	12,3	28,5	19,883	Nema modus	16,125	19,65	22,4
Varijansa	*St. dev.*	*IQR*	*Interval varijac.*	α_3	α_4	*Stand. greška*	

20,291	4,505	6,275	16,2	0,28	0,08	1,3

Vidimo da koeficijent asimetrije iznosi 0,28, što znači da je raspored umjereno asimetričan. (Da li je spljoštenost približno normalna?)



2. U drugoj etapi izvršićemo "detekciju" ekstremnih vrijednosti pomoću boxplota dobijenog u statističkom paketu Minitab.

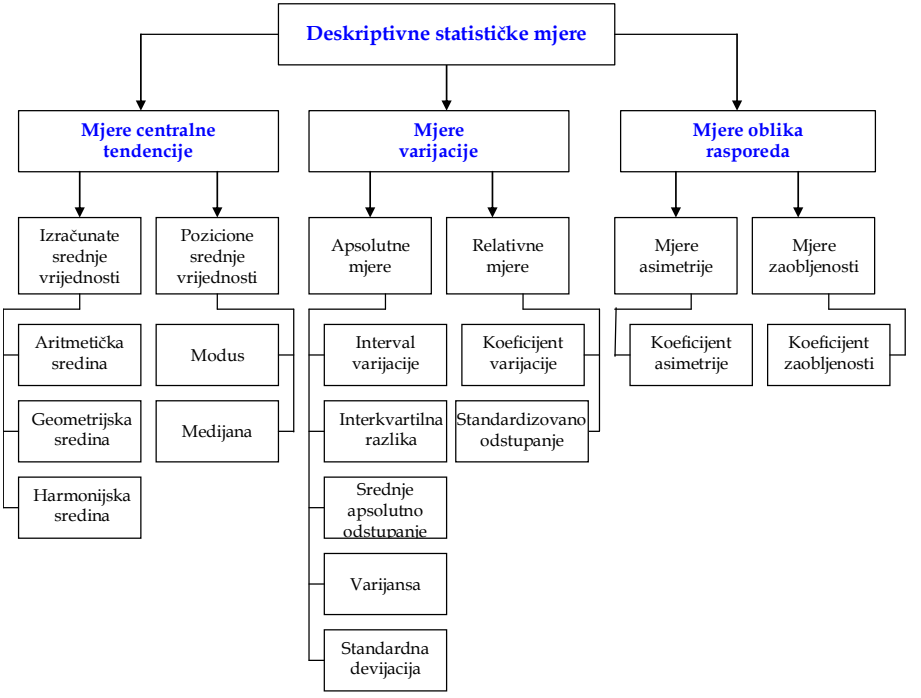
Boxplot nam sugeriše da u seriji nema ekstremnih vrijednosti. Boxplot je, sâm po sebi, veoma pogodno grafičko sredstvo za prikaz rasporeda. Kao što se može vidjeti, on se zasniva na opisu rasporeda pomoću pet brojeva (five number summary):

Minimum	Prvi kvartil	Medijana	Treći kvartil	Maksimum

3. Budući da je raspored umjereno asimetričan i da nema ekstremnih vrijednosti, kao odgovarajuću mjeru centralne tendencije izabraćemo aritmetičku sredinu. Dakle, centar datog rasporeda opisujemo aritmetičkom sredinom. To znači da je prosječna količina download-ovanog materijala u posmatranoj firmi iznosila 19,883 megabajta.

4. Na osnovu identičnih argumenata zaključujemo da je varijacije date serije najbolje opisati pomoću standardne devijacije, koja iznosi 4,505 megabajta.

Da sumiramo: **Posmatrani raspored je umjereno asimetričan, nema ekstremnih vrijednosti, aritmetička sredina mu iznosi 19,883 megabajta i standardna devijacija 4,505 megabajta.**



Slika 2.14 Klasifikacija deskriptivnih mjera u statistici

VJEROVATNOĆA

Osnovni pojmovi
Neke operacije sa slučajnim događajima
Vjerovatnoća događaja
Klasična definicija vjerovatnoće
Statistička definicija vjerovatnoće
Subjektivna vjerovatnoća
Pravilo sabiranja vjerovatnoća (aditivno pravilo)
Uslovna vjerovatnoća i nezavisnost događaja
Vjerovatnoća uzroka (Bayes-ova teorema)

CILJEVI POGLAVLJA

Nakon čitanja ovog poglavlja bićete u stanju da:

13. shvatite značaj teorije vjerovatnoće za statističko zaključivanje
14. objasnite razliku između klasične, statističke i subjektivne vjerovatnoće
15. shvatite termine eksperiment, slučajni događaj i ishod
16. razumijete i primijenite pravilo sabiranja vjerovatnoća
17. izračunavate vjerovatnoće primjenom pravila množenja
18. shvatite značenje uslovne vjerovatnoće i nezavisnih događaja
19. izračunavate vjerovatnoće korišćenjem Bayes-ove teoreme

Vjerovatnoća, **kao izraz mjere očekivanja da se neki događaj desi,** važno je uporište statističkog

zaključivanja, koje se izvodi u uslovima manje ili veće neizvjesnosti.

Teorija vjerovatnoće potpomaže zaključivanje u uslovima imperfektne informacije i neizvjesnosti. To se posebno odnosi na formulisanje metoda za zaključivanje o vrijednostima parametara populacije na osnovu informacije iz uzorka.

OSNOVNI POJMOVI

Osnovni pojmovi u teoriji vjerovatnoće su: **eksperiment, ishod i prostor uzorka,** odnosno **elementarnih događaja**.

Eksperimenti, **su potpuno precizirane operacije posmatranja ili prikupljanja podataka, koje se u nepromijenjenim uslovima mogu ponavljati proizvoljno mnogo puta i čiji se ishod ne može sa sigurnošću predvidjeti.**

Realizacije eksperimenata su ishodi (elementarni događaji), kao skup unaprijed poznatih mogućnosti realizacije.

U svakom pojedinom izvođenju eksperimenta realizuje se samo jedan ishod - elementarni događaj, što znači da su elementarni događaji međusobno isključivi.

Prostor uzorka (prostor elementarnih događaja) je skup svih elementarnih događaja, a može se označiti kao:

$$S = \{e_1, e_2, ..., e_i, ..., e_n\}, \qquad e_i \in S, 1 \leq i \leq n$$

Slučajni događaj predstavlja podskup skupa elementarnih događaja, koji imaju neku zajedničku osobinu.

Posmatrajmo bacanje kocke, čije su strane numerisane brojevima od 1 do 6. Elementarni događaji su brojevi od 1 do 6, koji čine prostor uzorka $S = \{1, 2, 3, 4, 5, 6\}$.

Slučajni događaj čine jedan, dva ili više elementarnih događaja. Na primjer, slučajni događaj je A < dobijeni broj je neparan > , što se može zapisati kao $A = \{1, 3, 5\}$.

Siguran događaj (S) jednak je skupu elementarnih događaja i realizuje se svaki put kada se izvodi određeni eksperiment.

U ovom slučaju može se odrediti kao:

S < dobijeni broj je manji od 7>.

Nemoguć događaj (N) **je onaj koji se ne može realizovati prilikom izvođenja nekog eksperimenta.**

Za eksperiment bacanja kocke nemoguć događaj može se odrediti kao:

N < dobijeni broj veći je od 6 >.

NEKE OPERACIJE SA SLUČAJNIM DOGAĐAJIMA

Neka su dati događaji A i B.

Kažemo da A implicira B ako se svaki put kada se ostvari događaj A ostvari i događaj B, što se označava sa:

$$A \subseteq B$$

Događaji A i B su jednaki ako istovremeno A implicira B i B implicira A, što se označava sa:

$$A \subseteq B \ \wedge B \subseteq A$$

Unija događaja A i B je događaj koji se ostvaruje onda i samo onda ako se ostvari bar jedan od događaja A i B. Označava se sa:

$$A \cup B = \{e_i \mid e_i \in A \vee e_i \in B\} \ \ \forall i$$

Presjek događaja A i B je događaj koji se ostvaruje istovremenim ostvarenjem i događaja A i događaja B.

Označava se sa:

$$A \cap B = \{e_i \mid e_i \in A \wedge e_i \in B\} \ \ \forall i$$

Za dva događaja A i B kažemo da se međusobno isključuju (da su disjunktni) ako je njihov presjek nemoguć događaj, odnosno prazan skup:

$$A \cap B = \phi$$

gdje je sa ϕ označen prazan skup, odnosno ϕ je oznaka da nema zajedničkih elemenata događaja A i B.

Unija dva događaja A i B, koji su međusobno isključivi (disjunktni), predstavlja zbir tih događaja:

$$A \cup B = A + B$$

U slučaju kada događaji A i B pokrivaju cjelokupan prostor elementarnih događaja S ($A \cup B = S$), a međusobno su isključivi, tj. $A \cap B = \phi$, tada za ove događaje kažemo da su komplementarni.

VJEROVATNOĆA DOGAĐAJA

Neka je slučajni događaj E definisan u prostoru elementarnih događaja S. Vjerovatnoća događaja E je neki realan broj, koji se obilježava sa $P(E)$.

Vjerovatnoća je funkcija P definisana na slučajnim događajima, koja ih preslikava u realne brojeve.

Vjerovatnoća ima sljedeće osobine:

1. svakom slučajnom događaju E odgovara nenegativni broj $P(E)$ koji predstavlja njegovu vjerovatnoću, tako da je:

$$P(E) \geq 0$$

2. vjerovatnoća sigurnog događaja S je:

$$P(S) = 1$$

3. za međusobno disjunktne događaje E_1, E_2, ..., E_n vjerovatnoća njihove unije jednaka je zbiru njihovih vjerovatnoća:

$$P(E_1 \cup E_2 \cup \cdots \cup E_n \cup \cdots) = P(E_1) + P(E_2) + \cdots + P(E_n) + \cdots$$

KLASIČNA DEFINICIJA VJEROVATNOĆE

Ako skup elementarnih događaja nekog eksperimenta ima n elemenata sa jednakim mogućnostima realizacije i ako realizacija ukupno k elementarnih događaja, pri čemu je $0 < k < n$, povlači realizaciju događaja E, tada je vjerovatnoća događaja E data izrazom:

$$P(E) = \frac{k}{n}$$

Ova vjerovatnoća označava se kao vjerovatnoća a priori i veže se za slučaj kada je prostor elementarnih događaja konačan, a elementarni događaji imaju jednaku vjerovatnoću ostvarenja.

STATISTIČKA DEFINICIJA VJEROVATNOĆE

Označimo sa N broj ponavljanja nekog eksperimenta, a sa N_E broj javljanja događaja E u N tih ponavljanja.

Statistička ili vjerovatnoća a posteriori određuje se nakon izvršenog eksperimenta, na osnovu dobijenih empirijskih vrijednosti, kao granična vrijednost relativnog učešća javljanja događaja E prilikom N ponavljanja datog eksperimenta:

$$P(E) = \lim_{N \to \infty} \frac{N_E}{N}$$

S obzirom na to da se u praksi eksperiment izvodi samo konačan broj puta, ova vjerovatnoća se ne može precizno odrediti kao granična vrijednost, nego se samo ocjenjuje kao relativna frekvencija događaja E u N ponavljanja eksperimenta i izračunava se kao:

$$\hat{P}(E) = \frac{N_E}{N}$$

Pri tome, empirijske vrijednosti pokazuju ispravnost ovog izraza, jer što se više eksperimenata izvodi, to se ova vjerovatnoća više približava vjerovatnoći a priori.

Ukoliko je broj ponavljanja nekog eksperimenta veći, utoliko se vjerovatnoća a posteriori više približava vjerovatnoći a priori, što je, također, jedna od osnovnih postavki u statističkom zaključivanju, a podrazumijeva sljedeće: prilikom posmatranja nekog osnovnog skupa, što više elemenata tog skupa bude opservisano, to će i zaključci o tom skupu (prije svega o njegovim parametrima) biti tačniji, precizniji i pouzdaniji.

SUBJEKTIVNA VJEROVATNOĆA

Kada se događaji javljaju samo jedanput ili kada su uslovi njihove realizacije značajno različiti, tada se njihova vjerovatnoća može odrediti i na osnovu mišljenja određenih lica koja su na neki način upućena u realizaciju tih događaja.

Polaznu osnovu za davanje ocjene o vjerovatnoći dešavanja nekog događaja čini znanje, iskustvo i raspoložive informacije tih lica (eksperata), koja prema svom ubjeđenju daju ocjenu tražene vjerovatnoće.

U nekim praktičnim istraživačkim procedurama ovo je neophodna zamjena u slučajevima kada se ne raspolaže podacima potrebnim za određivanje egzaktne vjerovatnoće.

PRAVILO SABIRANJA VJEROVATNOĆA (ADITIVNO PRAVILO)

U slučajevima kada se određuje vjerovatnoća da se desi događaj E_1 ili E_2 ili događaj ... E_n, primjenjuje se pravilo sabiranja vjerovatnoća.

U čisto praktičnom smislu, ovakvu kombinaciju očekivanja realizacije događaja prepoznajemo po upravo ovakvoj formulaciji ... "ili" ... "ili".

Neka su događaji E_1 i E_2 slučajni događaji iz prostora elementarnih događaja S. Tada je **vjerovatnoća njihove unije jednaka zbiru pojedinačnih vjerovatnoća tih događaja umanjenom za vjerovatnoću njihovog istovremenog ostvarenja**:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

što predstavlja vjerovatnoću da će se ostvariti ili događaj E_1 ili događaj E_2.

Ako su događaji E_1 i E_2 međusobno disjunktni (isključivi), tada je vjerovatnoća njihove unije jednaka zbiru njihovih pojedinačnih vjerovatnoća:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2), \qquad P(E_1 \cap E_2) = \phi$$

USLOVNA VJEROVATNOĆA I NEZAVISNOST DOGAĐAJA

Za dva događaja E_1 i E_2 uslovna vjerovatnoća označava se sa $P(E_2/E_1)$ i predstavlja vjerovatnoću ostvarenja događaja E_2 pod uslovom da se ostvario događaj E_1.

Ova vjerovatnoća određuje se pomoću izraza:

$$P(E_2 / E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}$$

uz uslov da je $P(E_1) > 0$.

Nezavisni događaji su oni kod kojih ostvarenje ili neostvarenje jednog događaja nema uticaja na vjerovatnoću ostvarenja drugog događaja, tako da je:

$$P(E_2 / E_1) = P(E_2) \quad \wedge \quad P(E_1 / E_2) = P(E_1).$$

Ovo se može iskazati i na sljedeći način: dva događaja E_1 i E_2 su nezavisni ako je:

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2),$$

Odnosno, dva događaja su nezavisna ako je vjerovatnoća njihovog istovremenog ostvarenja jednaka proizvodu njihovih pojedinačnih vjerovatnoća.

PRAVILO MNOŽENJA VJEROVATNOĆA (MULTIPLIKATIVNO PRAVILO)

Pravilo množenja vjerovatnoća primjenjuje se za određivanje vjerovatnoće istovremenog (zajedničkog) javljanja dva ili više događaja prilikom izvođenja nekog eksperimenta.

U čisto praktičnom smislu, ovakvu kombinaciju očekivanja realizacije događaja prepoznajemo po formulaciji ... "i" ... "i", odnosno kada se određuje vjerovatnoća da se desi "i" jedan "i" drugi "i" ... "i" *n*-ti događaj.

Za dva događaja E_1 i E_2 vjerovatnoća njihovog istovremenog ostvarenja dobije se na osnovu izraza:

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2 \,/\, E_1)$$
$$P(E_1 \cap E_2) = P(E_2) \cdot P(E_1 \,/\, E_2)$$

Za tri događaja E_1, E_2 i E_3 vjerovatnoća njihovog istovremenog ostvarenja dobije se na osnovu izraza:

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) \cdot P(E_2 \,/\, E_1) \cdot P(E_3 \,/\, E_1 \cap E_2)$$

Za slučajne događaje koji su međusobno nezavisni vjerovatnoća njihovog istovremenog ostvarenja dobije se kao proizvod njihovih pojedinačnih vjerovatnoća:

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

ako je $P(E_2 \,/\, E_1) = P(E_2)$, odnosno $P(E_1 \,/\, E_2) = P(E_1)$.

VJEROVATNOĆA UZROKA (BAYES-OVA TEOREMA)

Vrlo često nas zanima **vjerovatnoća uzroka nekog događaja.**

U ovakvim slučajevima određivanje vjerovatnoće uzroka jedan je od načina njihovog otkrivanja.

Neka su dati događaji D_1, D_2, ... , D_k, koji su međusobno disjunktni:

$$D_i \cap D_j = \phi \ , \ \forall \ i, j \ ,$$

a čija unija predstavlja siguran događaj, odnosno neki potpun prostor elementarnih događaja:

$$\cup \ D_i = S \ , \ i = 1, ..., k.$$

Tada je vjerovatnoća događaja X koji se realizuje pod uslovom da se realizovao jedan i samo jedan od događaja D_1, D_2, ... , D_k :

$$P(X) = \sum_{i=1}^{k} P(D_i) \cdot P(X / D_i) = P(D_1) \cdot P(X/D_1) + P(D_2) \cdot P(X/D_2) + \cdots + P(D_k) \cdot P(X/D_k).$$

Prema pravilu množenja vjerovatnoća važi:

$$P(D_1) \cdot P\,(D_2 / D_1) = P(D_2) \cdot P\,(D_1 / D_2) \Rightarrow P\,(D_2 / D_1) = \frac{P(D_2)}{P(D_1)} \cdot P\,(D_1 / D_2),$$

Odavde se uopštavanjem dobija opšti oblik Bayes-ove teoreme:

$$\boldsymbol{P\,(D_i/X)} = \frac{P(D_i)}{P(X)} \cdot \boldsymbol{P\,(X/D_i)} = \frac{P(D_i)\cdot P(X/D_i)}{\sum P(D_i)\cdot P(X/D_i)}$$

Pri tome se događaji D_1, D_2, ... , D_k smatraju uzrocima događaja X koji je, dakle, njihova posljedica. Zato se ova vjerovatnoća označava kao vjerovatnoća uzroka.

Vjerovatnoće $P(D_i)$ su a priori, a vjerovatnoće $P(X/D_i)$ a posteriori, što u konačnom rezultatu daje a posteriorne vjerovatnoće, koje pokazuju da je neki od događaja D_1, D_2, ... , D_k bio uzrok nastanka događaja X.

Ova teorema posebnu primjenu nalazi prilikom određivanja vjerovatnoće nekih događaja za koje je poznata vjerovatnoća uzroka njihovog nastanka.

SLUČAJNE PROMJENLJIVE I MODELI RASPOREDA VJEROVATNOĆA

Slučajne promjenljive
Raspored vjerovatnoće prekidne slučajne promjenljive
Funkcija rasporeda prekidne slučajne promjenljive
Očekivana vrijednost i varijansa prekidne slučajne promjenljive
Modeli prekidnih rasporeda vjerovatnoća
Raspored vjerovatnoća neprekidne slučajne promjenljive
Normalan raspored

CILJEVI POGLAVLJA

Nakon čitanja ovog poglavlja bićete u stanju da:

20. razumijete slučajnu promjenljivu, raspored vjerovatnoća i očekivanu vrijednost, koji predstavljaju osnov za pravilno razumijevanje statističkog zaključivanja
21. shvatite značaj i primjenu osnovnih teorijskih rasporeda u ekonomiji
22. koristite tablice standardizovanog normalnog rasporeda
23. shvatite postupak standardizacije i primijenite ga u rješavanju različitih problema

SLUČAJNE PROMJENLJIVE

Promjenljiva koja svoje vrijednosti uzima "na slučaj" naziva se slučajna promjenljiva (a ne numeričko obilježje).

PRIMJER Posmatrajmo statistički eksperiment koji se sastoji u bacanju dva novčića. Definisali smo prostor uzorka kao skup svih elementarnih ishoda eksperimenta.

U našem eksperimentu prostor uzorka S sastoji se od 4 elementarna ishoda:

$$S = \{(P,P),(P,G),(G,P),(G,G)\},$$

gdje je sa P označen pad pisma, a sa G pad grba.

Ako nas interesuje pojavljivanje grba, **možemo uvesti promjenljivu veličinu X tako što ćemo svakom ishodu eksperimenta pridružiti jedan broj** koji će pokazivati koliko se puta pojavljuje grb u tom ishodu.

Tabela 4.1 Promjenljiva X = broj grbova pri bacanju dva novčića

Prostor uzorka eksperimenta	Broj grbova
P, P	0
P, G	1
G, P	1
G, G	2

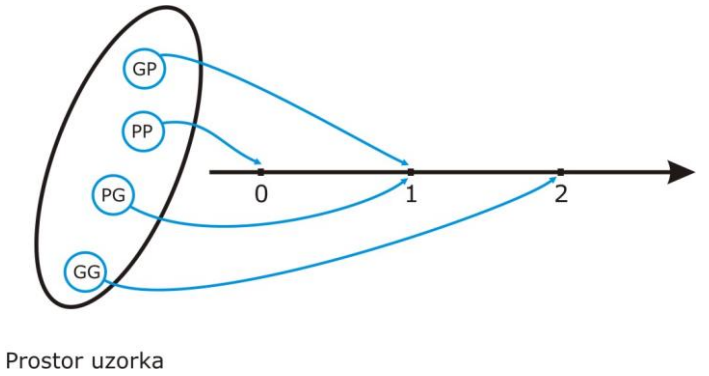
Nikada ne možemo *unaprijed* predvidjeti ni vrijednosti koje će uzeti promjenljiva X. Ona te vrijednosti uzima na slučaj, pa se naziva slučajna promjenljiva.

Neki autori je zovu i aleatorna promjenljiva, slučajna varijabla ili stohastična promjenljiva. Možemo je definisati na sljedeći način:

Slučajna promjenljiva

Slučajna promjenljiva je numerička funkcija koja svakom ishodu statističkog eksperimenta pridružuje jedan realan broj.

Uočimo da elementarni ishodi koji sačinjavaju prostor uzorka ne moraju imati numeričke vrijednosti.



Slika 4.1 Slučajna promjenljiva kao funkcija definisana na prostoru uzorka

U opštem slučaju možemo zaključiti da svaka transformacija, tj. funkcija slučajne promjenljive i sama predstavlja slučajnu promjenljivu.

Slučajne promjenljive možemo podijeliti na osnovu toga da li uzimaju samo izolovane vrijednosti ili sve moguće vrijednosti u nekom intervalu, dakle isto kao i numerička obilježja.

Za slučajnu promjenljivu kažemo da je prekidna (ili diskretna) ako može uzeti konačan broj izolovanih vrijednosti ili prebrojivo mnogo vrijednosti (tj. vrijednosti koje se mogu prebrojati skupom cijelih nenegativnih brojeva: 0, 1, 2, 3, ... itd.).

Tabela 4.2. Primjeri prekidnih slučajnih promjenljivih

Slučajne promjenljive Definicija	Vrijednosti slučajne promjenljive	Broj vrijednosti
1. X = broj korisnika Interneta u zgradi sa 80 stanara	0, 1, 2, ..., 80	Konačan
2. Y = broj neispravnih kompakt diskova u uzorku od 7 diskova	0, 1, 2, ..., 7	Konačan
3. Z = broj zastoja u proizvodnji artikla A	0, 1, 2, ...	Prebrojivo mnogo

Slučajna promjenljiva je neprekidna (ili kontinuirana) ako može da uzme bilo koju vrijednost u nekom intervalu.

Naime, **između bilo koje dvije vrijednosti x_1 i x_2 slučajne promjenljive postoji sljedeća moguća vrijednost x_3, koja je različita od x_1 i x_2. Broj vrijednosti koje može uzeti neprekidna slučajna**

promjenljiva je beskonačan. Primjeri neprekidnih slučajnih promjenljivih su: visina i težina studenata, vrijeme potrebno da se taksijem stigne od stana do fakulteta, prečnik kugličnog ležaja proizvedenog na mašini određenog tipa, itd.

RASPORED VJEROVATNOĆA PREKIDNE SLUČAJNE PROMJENLJIVE

Prekidna slučajna promjenljiva može se opisati kao veličina koja uzima određene izolovane vrijednosti sa odgovarajućim vjerovatnoćama.

Vjerovatnoću da slučajna promjenljiva X uzme neku od navedenih vrijednosti označimo sa $P\,(X = x_i) = p_i$.

Dakle:

$$P(X = x_1) = P(X = 0) = p_1 = \frac{1}{4}$$
$$P(X = x_2) = P(X = 1) = p_2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$
$$P(X = x_3) = P(X = 2) = p_3 = \frac{1}{4}.$$

Raspored vjerovatnoća prekidne slučajne promjenljive

Raspored vjerovatnoće (raspodjela vjerovatnoće ili funkcija vjerovatnoće ili zakon vjerovatnoće) prekidne slučajne promjenljive

X je skup parova svih vrijednosti koje može da uzme slučajna promjenljiva X i odgovarajućih vjerovatnoća.

Tabela 4.3 Raspored vjerovatnoće za broj grbova u eksperimentu sa dva novčića

Broj grbova	Vjerovatnoća
(Različite vrijednosti x)	P
0	¼
1	¼
2	½
$\sum$	1

Opšte karakteristike svih prekidnih slučajnih promjenljivih su:

1. Nijedna vjerovatnoća u rasporedu vjerovatnoće ne može biti negativna, tj:

$$P(X = x_i) \geq 0 \text{ za svako } i.$$

2. Suma vjerovatnoća koje odgovaraju svim vrijednostima slučajne promjenljive X mora biti jednaka 1, tj:

$$\sum_i p_i = 1.$$

U opštem slučaju, raspored vjerovatnoće prekidne slučajne promjenljive možemo prikazati pomoću Tabele 4.4, koja sadrži dva niza informacija: vrijednosti slučajne promjenljive i njihove vjerovatnoće, uz uslov da zbir vjerovatnoća bude jednak 1.

Tabela 4.4 Raspored vjerovatnoće slučajne promjenljive X

Različite vrijednosti x	$x_1\ x_2\ \ldots\ x_i\ \ldots\ x_n$
Vjerovatnoća p	$p_1\ p_2\ \ldots\ p_i\ \ldots\ p_n$

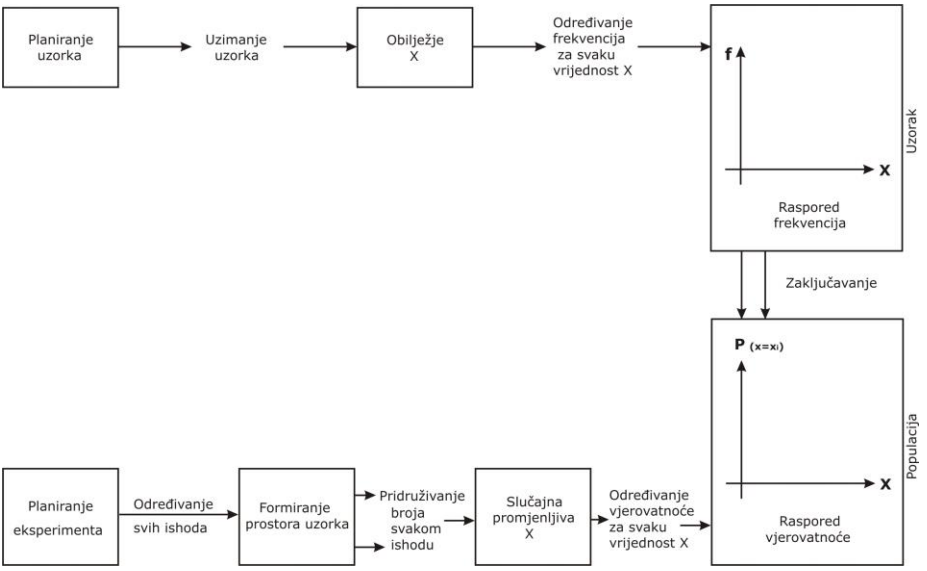
Raspored vjerovatnoće ne smijemo poistovjećivati sa rasporedom frekvencija.

Raspored vjerovatnoće je teorijski model koji pridružuje vjerovatnoće pojedinim vrijednostima slučajne promjenljive; raspored frekvencija može se formirati tek nakon prikupljanja podataka na osnovu statističkog eksperimenta.

Tabela 4.5 Raspored vjerovatnoće za pravilnu kocku

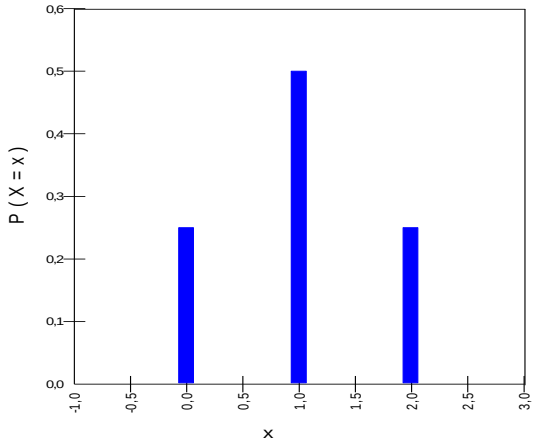
x	1	2	3	4	5	6
$p_i = P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Tipičan postupak statističkog zaključivanja i uspostavljanja veze između rasporeda frekvencija i rasporeda vjerovatnoće ilustrovaćemo na Slici 4.2.



Slika 4.2 Veza između rasporeda frekvencija, rasporeda vjerovatnoće i statističkog zaključivanja

Kao i prekidne rasporede frekvencija, tako i rasporede vjerovatnoće prekidnih slučajnih promjenljivih možemo grafički predstaviti u vidu dijagrama vjerovatnoće, kao na Slici 4.3.



Slika 4.3 Grafički prikaz rasporeda vjerovatnoće iz Tabele 4.3.

Mehanički ovo možemo interpretirati kao da je izvjesna masa jednaka jedinici raspoređena na takav način da se u tačkama x_1, x_2, ..., x_n nalaze odgovarajući dijelovi mase p_1, p_2, ..., p_n.

Usljed toga se često raspored vjerovatnoće prekidne slučajne promjenljive naziva funkcija mase vjerovatnoće.

FUNKCIJA RASPOREDA PREKIDNE SLUČAJNE PROMJENLJIVE

Svaka slučajna promjenljiva ima svoju funkciju rasporeda.

Funkcija rasporeda

Funkcija rasporeda (naziva se još i kumulativna funkcija rasporeda) prekidne slučajne promjenljive pokazuje vjerovatnoću da slučajna promjenljiva X uzme vrijednost koja je manja ili jednaka bilo kojoj proizvoljnoj vrijednosti x.

Funkcija rasporeda slučajne promjenljive X se označava sa $F(x)$ i data je vjerovatnoćom:

$$F(x) = P(X \leq x)$$

gdje x može biti bilo koji realan broj.

Neka je X prekidna slučajna promjenljiva koja može uzeti vrijednosti x_1, x_2, ..., x_r, ..., x_n , gdje su $x_1 < x_2 < ... < x_n$. Neka P $(X = x_i)$ označava vjerovatnoću da X uzme vrijednost x_i.

Tada $F(x_r)$ predstavlja vjerovatnoću da slučajna promjenljiva X uzme vrijednost koja će biti manja ili jednaka x_r i dobija se sabiranjem (kumuliranjem) vjerovatnoća za sve vrijednosti slučajne promjenljive koje su manje ili jednake x_r tj:

$$F(x_r) = P(X \leq x_r) = \sum_{i=1}^{r} P(x = x_i) = P(x = x_1) + P(x = x_2) + ... + P(x = x_r).$$

Svaka funkcija rasporeda mora da zadovolji sljedeće matematičke karakteristike:

a) Za bilo koju vrijednost *a*:

$$0 \le F(a) \le 1,$$

što je i razumljivo, jer je funkcija rasporeda vjerovatnoća;

b) $$F(-\infty) = 0 \text{ i } F(+\infty) = 1,$$

jer se $F(-\infty) = P(X \le -\infty)$ odnosi na nemoguć događaj, a $F(+\infty) = P(X \le +\infty)$ na siguran događaj;

c) ako je $a < b$, tada je $F(a) \le F(b)$,

tj. funkcija rasporeda bilo koje slučajne promjenljive je neopadajuća funkcija.

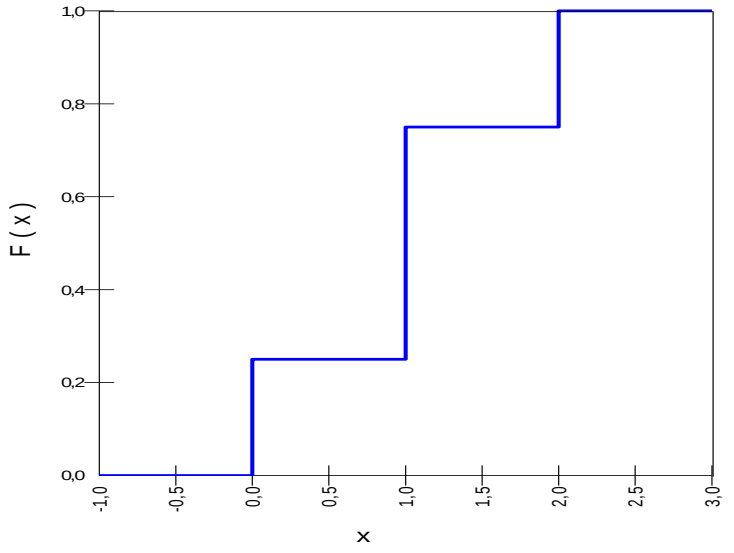
Funkcija rasporeda za prekidnu slučajnu promjenljivu dobija se kumuliranjem vjerovatnoća.

Dobijene vrijednosti nam omogućuju da funkciju rasporeda za eksperiment sa bacanjem dva novčića, zajedno sa rasporedom vjerovatnoće, prikažemo u vidu tabele.

Tabela 4.7 Raspored vjerovatnoće i funkcija rasporeda u primjeru dva novčića

x_i	$p_i = P(X = x_i)$	$F(x) = P(X \leq x)$
0	1/4	1/4
1	2/4	3/4
2	1/4	1

Funkciju rasporeda možemo grafički prikazati u koordinatnom sistemu pomoću tzv. stepenaste funkcije, kao na Slici 4.4.



Slika 4.4 Funkcija rasporeda za broj grbova u eksperimentu sa bacanjem dva novčića

Grafik funkcije rasporeda sastoji se samo od horizontalnih linija. Vertikalne linije nisu dio funkcije rasporeda, ali se obično uključuju da bi grafik bio pregledniji.

OČEKIVANA VRIJEDNOST I VARIJANSA PREKIDNE SLUČAJNE PROMJENLJIVE

Od svih pokazatelja rasporeda vjerovatnoće slučajne promjenljive zadržaćemo se samo na dva:

1. očekivanoj vrijednosti slučajne promjenljive, kao mjeri centralne tendencije i

2. varijansi slučajne promjenljive, kao mjeri disperzije.

Očekivana vrijednost

Očekivana vrijednost za neku slučajnu promjenljivu ima isto značenje kao i aritmetička sredina za neko obilježje.

Posmatrajmo sada prekidnu slučajnu promjenljivu X sa sljedećim rasporedom vjerovatnoća:

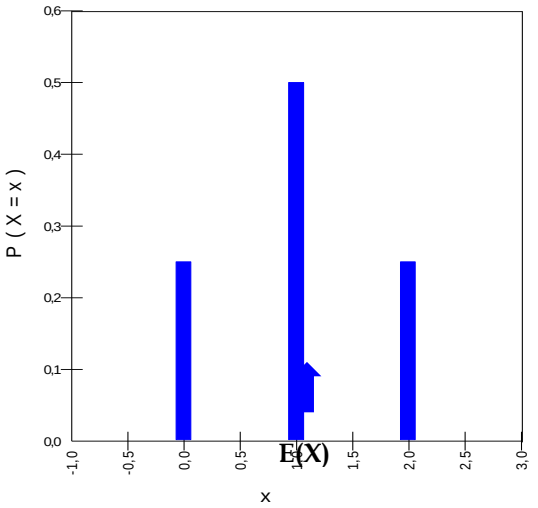
x_i	$p_i = P(X = x_i)$
x_1	$p_1 = P(X = x_1)$
x_2	$p_2 = P(X = x_2)$
$\vdots$	$\vdots$
x_n	$p_n = P(X = x_n)$

Očekivana vrijednost slučajne promjenljive X označava se sa $E(X)$ i jednaka je zbiru proizvoda vrijednosti koje uzima slučajna promjenljiva i odgovarajućih vjerovatnoća.

Očekivana vrijednost slučajne promjenljive X	$E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i$

Očekivana vrijednost naziva se još matematičko očekivanje ili, jednostavno, očekivanje od X.

Očekivana vrijednost može se intuitivno shvatiti kao centar gravitacije rasporeda vjerovatnoće, odnosno svih vrijednosti koje uzima slučajna promjenljiva.



Slika 4.5 $E(X)$ kao centar gravitacije rasporeda vjerovatnoće

U praktičnim istraživanjima, pojam očekivana vrijednost slučajne promjenljive X najčešće se poistovjećuje sa aritmetičkom sredinom osnovnog skupa. Zbog toga možemo i napisati sljedeću jednakost:

$$E(X) = \mu_X = \mu,$$

odnosno očekivana vrijednost jednaka je aritmetičkoj sredini populacije.

Dok se pojam očekivana vrijednost u početku vezivao za hazardne igre, danas predstavlja osnovno sredstvo u analizi slučajnih promjenljivih.

Na osnovu navedenih primjera može se sagledati da se:

- očekivana vrijednost $E(X)$ nalazi između minimalne i maksimalne vrijednosti slučajne promjenljive, i
- očekivana vrijednost se ne mora poklapati sa nekom vrijednošću slučajne promjenljive.

Znači, nikako ne možemo tumačiti pojam očekivana vrijednost kao vrijednost koja se očekuje u statističkom eksperimentu, već kao prosječan očekivani ishod svih mogućih opita u eksperimentu.

Drugačije rečeno, **očekivana vrijednost je prosjek slučajne promjenljive**.

Varijansa prekidne slučajne promjenljive

Slijedeći istu logiku kao kod formulisanja varijanse rasporeda frekvencija, potrebno je kvadrirati odstupanja pojedinih vrijednosti slučajne promjenljive od njene očekivane vrijednosti.

Tako dolazimo do mjere disperzije rasporeda vjerovatnoće slučajne promjenljive, koja se naziva varijansa slučajne promjenljive.

Varijansa slučajne promjenljive označava se sa Var X, ili σ_X^2 ili σ^2 i dobija se na osnovu obrasca:

Varijansa prekidne slučajne promjenljive

$$Var(X) = \sigma_X^2 = E\left[X - E(X)\right]^2 = \sum_i \left[x_i - E(X)\right]^2 p_i$$

Ako očekivanu vrijednost slučajne promjenljive $E(X)$ označimo sa μ_X i shvatimo kao aritmetičku sredinu osnovnog skupa (odnosno populacije), zaključujemo da se formula za varijansu može napisati u obliku:

$$\sigma_X^2 = \sum_i (x_i - \mu_X)^2 p_i \,.$$

Varijansa slučajne promjenljive može se jednostavnije izračunati pomoću radne, alternativne formule, na analogni način kao i varijansa rasporeda frekvencija:

$$Var(X)=\sigma_X^2=E\left[X-E(X)\right]^2=E(X^2)-\left[E(X)\right]^2=\sum x_i^2 p_i-\mu_X^2$$

tj. kao očekivanje kvadrata slučajne promjenljive minus kvadrat njenog očekivanja.

Da bismo izračunali varijansu na osnovu navedene formule potrebno je izračunati proizvode $x_i p_i$ i $x_i^2 p_i$, što je i učinjeno u trećoj i četvrtoj koloni Tabele 4.8.

Tabela 4.8 Izračunavanje varijanse slučajne promjenljive u primjeru sa dva novčića

x	p	$x\cdot p$	x^2p
0	$\frac{1}{4}$	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{1}{2}$	1
$\sum$		$1=E(X)=\mu$	$\frac{3}{2}=E(X^2)$

$$E(X)=\sum x_i p_i=1$$

$$E(X^2)=\sum x_i^2 p_i=\frac{3}{2}$$

$$\sigma_X^2=E(X^2)-\mu_X^2=\frac{3}{2}-(1)^2=\frac{1}{2}$$

Naravno, do istog rezultata bismo došli direktnom primjenom definicione formule 4.2:

$$\sigma_X^2=\sum_{i=1}^{3}\left[x_i-E(X)\right]^2 p_i=(0-1)^2\cdot\frac{1}{4}+(2-1)^2\cdot\frac{1}{2}+(2-1)^2\cdot\frac{1}{4}=\frac{1}{2}.$$

Budući da je $[x_i - E(X)]^2$ uvijek nenegativno, ni varijansa slučajne promjenljive ne može biti negativna, tj. Var $X \geq 0$. Ukoliko je Var $X = 0$, odnosno ne postoje nikakva odstupanja, X nije slučajna promjenljiva već neka konstanta. Važi i obrnuto, tj. varijansa konstante jednaka je nuli.

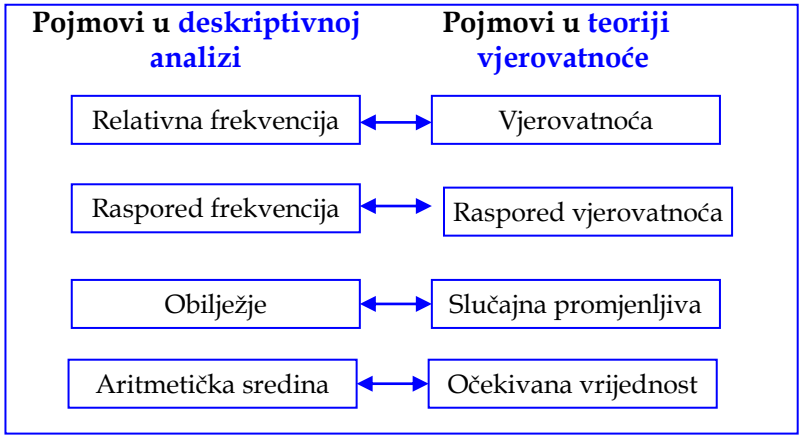
Da bi se disperzija mjerila u istim mjernim jedinicama kao i X, potrebno je uzeti kvadratni korijen varijanse. Na taj način dolazimo do standardne devijacije slučajne promjenljive X, koja se označava sa σ ili σ_X.

Jedna od važnih primjena navedenih karakteristika je u transformaciji slučajne promjenljive X sa očekivanom vrijednošću $E(X)$ i standardnom devijacijom σ_X u slučajnu promjenljivu Z sa očekivanom vrijednošću 0 i standardnom devijacijom jednakom 1.

Takva slučajna promjenljiva Z naziva se standardizovana slučajna promjenljiva.

Bilo koja slučajna promjenljiva X može se standardizovati kada se od nje oduzme njena očekivana vrijednost i zatim se podijeli sa standardnom devijacijom (čime se eliminiše jedinica mjere).

Standardizovana slučajna promjenljiva	$Z = \dfrac{X - E(X)}{\sigma_X}$	$E(Z) = 0$ Var $Z = 1$


Pojmovi u deskriptivnoj analizi
Pojmovi u teoriji vjerovatnoće
Relativna frekvencija
Vjerovatnoća
Raspored frekvencija
Raspored vjerovatnoća
Obilježje
Slučajna promjenljiva
Aritmetička sredina
Očekivana vrijednost

RASPORED VJEROVATNOĆE NEPREKIDNE SLUČAJNE PROMJENLJIVE

Postavlja se pitanje, **kako primijeniti binomne ili Poisson-ove tablice vjerovatnoće u slučaju kada se parametri rasporeda nalaze van opsega tablica**?

Tada nam modeli neprekidnih rasporeda vjerovatnoće mogu pružiti odgovarajuću aproksimaciju traženih vjerovatnoća. Osim toga, **veliki broj ekonomskih i drugih društvenih pojava može uzeti ma koju vrijednost iz nekog konačnog ili beskonačnog intervala,** odnosno po svojoj prirodi moraju se tretirati kao neprekidne promjenljive.

Navedene promjenljive teorijski mogu uzeti bilo koju vrijednost u nekom intervalu, iako je u praksi broj tih vrijednosti konačan zbog nesavršenih mjernih instrumenata.

Kod prekidne slučajne promjenljive, do rasporeda vjerovatnoće se dolazi jednostavno, tako što se formira lista pojedinih vrijednosti slučajne promjenljive i odgovarajućih vjerovatnoća.

Međutim, kod neprekidne slučajne promjenljive nemoguće je sastaviti takvu listu, jer je broj njenih vrijednosti beskonačan. **Zbog toga nema ni smisla govoriti o vjerovatnoći da slučajna promjenljiva X uzme jednu određenu vrijednost $P(X = x)$,** budući da takvih vrijednosti ima beskonačno mnogo, pa je ta vjerovatnoća jednaka nuli za svako x.

Dakle, kod neprekidne slučajne promjenljive ima

smisla određivati samo vjerovatnoću da se X nalazi u nekom intervalu.

Druga ključna razlika između neprekidnih i prekidnih slučajnih promjenljivih je u tome da prekidne mogu uzimati samo izolovane vrijednosti, a neprekidne sve vrijednosti u nekom intervalu.

Stoga je i razumljivo da se grafički prikaz neprekidne promjenljive neće sastojati od ordinata, već od neprekidne, glatke krive. Takva kriva naziva se kriva gustine vjerovatnoće.

Matematička funkcija označena sa $f(x)$, čiji je grafik predstavljen tom krivom, naziva se funkcija gustine vjerovatnoće (ili raspored vjerovatnoće) neprekidne slučajne promjenljive X.

Osnovne karakteristike funkcije gustine vjerovatnoće su analogne onima kod prekidnih.

1. Funkcija gustine nikada nije negativna, tj. $f(x) \geq 0$.

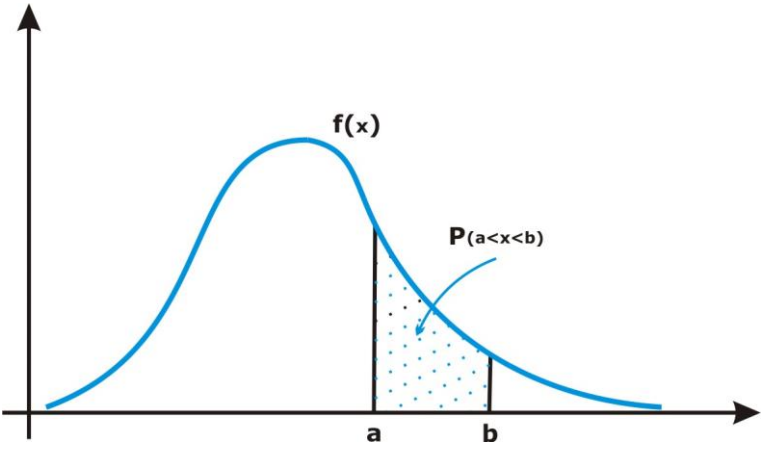
2. Ukupna površina ispod krive gustine vjerovatnoće uvijek je jednaka 1.

Budući da se radi o neprekidnoj krivi, umjesto znaka za sabiranje $\sum$ moramo koristiti integral, tj:

$$\int_D f(x)dx = 1,$$

gdje je D oblast definisanosti X (npr. $-\infty < X < +\infty$).

Slika 4.10 prikazuje hipotetički raspored neprekidne slučajne promjenljive X.



Slika 4.10 Grafik neprekidnog rasporeda vjerovatnoće

Vjerovatnoća da X uzme vrijednost u nekom intervalu, npr. (a, b), jednaka je površini između krive $f(x)$ i x ose duž intervala (a, b). Ova površina je na grafiku šrafirana.

Ako je funkcija $f(x)$ integrabilna, ta se površina može izraziti određenim integralom:

$$P(a < X < b) = \int_a^b f(x)dx$$

Pošto je slučajna promjenljiva neprekidna i može uzeti beskonačno mnogo vrijednosti, vjerovatnoća da uzme jednu određenu vrijednost jednaka je $1 / \infty = 0$.

Stoga je $P(X = a) = P(X = b) = 0$, pa je:

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b).$$

Znači, kod neprekidne slučajne promjenljive uključivanje graničnih vrijednosti intervala neće mijenjati vjerovatnoću da slučajna promjenljiva X pada u taj interval.

Na osnovu navedenog zaključujemo da vrijednost funkcije gustine $f(x)$ ne predstavlja vjerovatnoću, kao kod prekidne slučajne promjenljive $P\ (X = x)$, već nam samo daje informaciju o veličini ordinate.

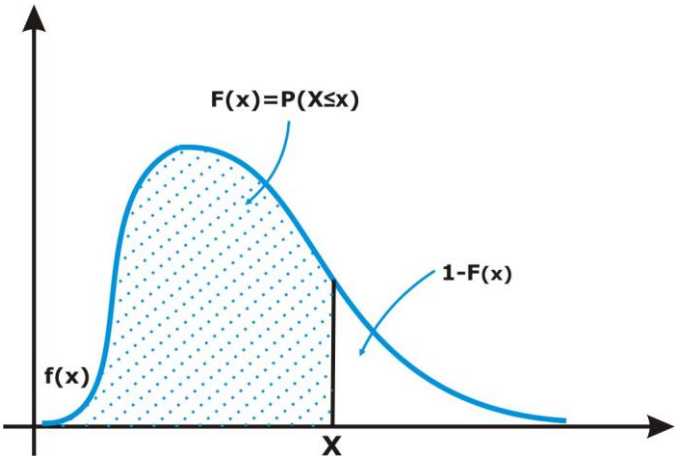
Međutim, nije potrebno u svakom konkretnom slučaju izračunavati određeni integral, jer su za veliki broj vrijednosti najvažnijih neprekidnih rasporeda slučajnih promjenljivih konstruisane odgovarajuće tablice vjerovatnoće.

Funkcija rasporeda neprekidne slučajne promjenljive ima veliku ulogu pri određivanju vjerovatnoće da se X nađe u nekom intervalu. Naime, tablice vjerovatnoće pojedinih modela rasporeda najčešće su date na osnovu funkcije rasporeda.

Uočimo najprije da $F(x)$ predstavlja površinu ispod krive funkcije gustine od njenog početka do tačke x, odnosno vjerovatnoću da će slučajna promjenljiva X uzeti neku vrijednost u intervalu čija je gornja granica x.

Podsjetimo se da je, slično, u prekidnom slučaju $F(x)$

predstavljala kumulativ vjerovatnoća do tačke x i da se dobijala sumiranjem pojedinih vjerovatnoća.



Slika 4.11 Grafički prikaz proizvoljne funkcije gustine i funkcija rasporeda

Pošto površina ispod krive predstavlja vjerovatnoću, funkcija rasporeda (osjenčeni dio na slici) i ostatak površine u zbiru moraju biti jednaki 1. Usljed toga je taj ostatak površine jednak 1 - $F(x)$ = $P(X > x)$.

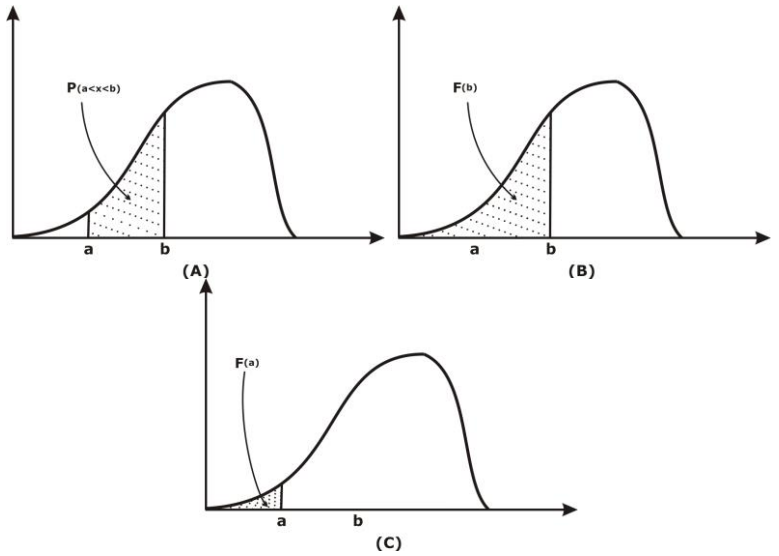
Vjerovatnoću da neprekidna slučajna promjenljiva uzme vrijednost u nekom intervalu (a, b) možemo sada odrediti kao:

$$P(a < X < b) = F(b) - F(a)$$

tj. kao razliku funkcije rasporeda gornje i donje granice intervala.

Ovo se lako može shvatiti ako se podsjetimo da je $F(b)$

vjerovatnoća da X uzme vrijednost manju ili jednaku b, $F(a)$ vjerovatnoća da X uzme vrijednost manju ili jednaku a; njihova razlika mora biti jednaka vjerovatnoći da se X nađe između a i b. Slika 4.12 nam ilustruje nalaženje vjerovatnoće da se X nađe u intervalu (a, b) korišćenjem relacije 4.8. Površina šrafirana na grafiku (A) dobija se kada se od šrafirane površine na grafiku (B) oduzme šrafirana površina na grafiku (C).



Slika 4.12 Određivanje vjerovatnoće da se X nađe u intervalu (a, b) korišćenjem funkcije rasporeda

NORMALAN RASPORED

Normalan raspored prvi je otkrio 1733.

francuski matematičar Abraham de Moivre, **kao granični oblik binomnog rasporeda,** tj. posmatrajući šta se događa sa binomnim rasporedom kada se broj opita neograničeno povećava (podsjetimo se Slike 4.6 c, d i e).

Ovaj raspored bio je poznat i P. Laplace-u u drugoj polovini XVIII vijeka, ali njegovo otkriće, kao ni Moivre-ovo, nije privuklo nikakvu pažnju. Tek kada ga je C. Gauss opisao 1809. godine, normalan raspored je potpuno bio prihvaćen od strane matematičke i statističke javnosti. Gauss je izveo normalan raspored kao matematičku funkciju namijenjenu opisu rasporeda grešaka u mjerenjima astronomskih opservacija. Usljed toga se ovaj raspored dugo nazivao Gausov raspored grešaka.

Za slučajnu promjenljivu X kažemo da ima normalan raspored ako je karakterišu neprekidne vrijednosti, a njena funkcija gustine vjerovatnoće ima izraz kao u (4.9).

Normalan raspored

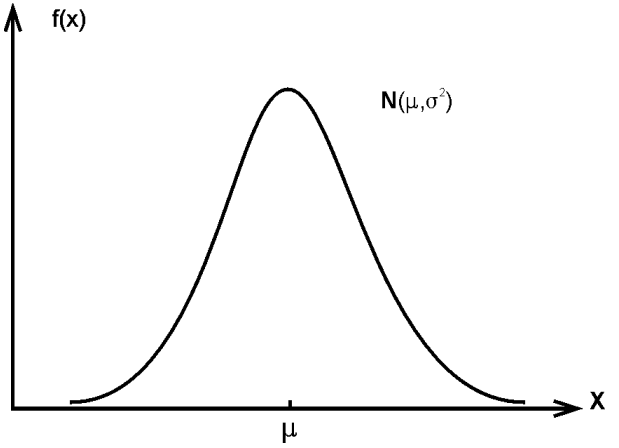
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < +\infty$$

gdje su:

π = matematička konstanta, približno jednaka 3,14159
e = matematička konstanta, približno jednaka 2,71828
μ = aritmetička sredina normalne slučajne promjenljive
σ = standardna devijacija normalne slučajne promjenljive

Takvu slučajnu promjenljivu označavaćemo sa $X : N(\mu, \sigma^2)$, što se čita X ima normalan raspored sa parametrima μ i σ^2.

Karakteristike normalnog rasporeda, s obzirom na složenost njegove funkcije, lakše je uočiti uz pomoć grafičkog prikaza, koji se naziva normalna kriva.



Slika 4.13 Normalna kriva

Osobine normalnog rasporeda

Navedimo najznačajnije karakteristike normalnog rasporeda.

1. Normalna kriva ima oblik zvona, unimodalna je i simetrična u odnosu na vrijednost $x = \mu$.

2. Budući da je normalan raspored simetričan, njegova aritmetička sredina, modus i medijana su međusobno jednaki.

3. Normalna kriva se proteže od $-\infty$ do $+\infty$, tj. asimptotski se približava x osi, pa je njen interval varijacije beskonačan ($i = \infty$).

4. Relativna mjera asimetrije α_3 jednaka je 0, a relativna mjera spljoštenosti α_4 ima vrijednost 3.

5. Ukupna površina ispod krive, kao kod svake krive gustine vjerovatnoće, jednaka je 1.

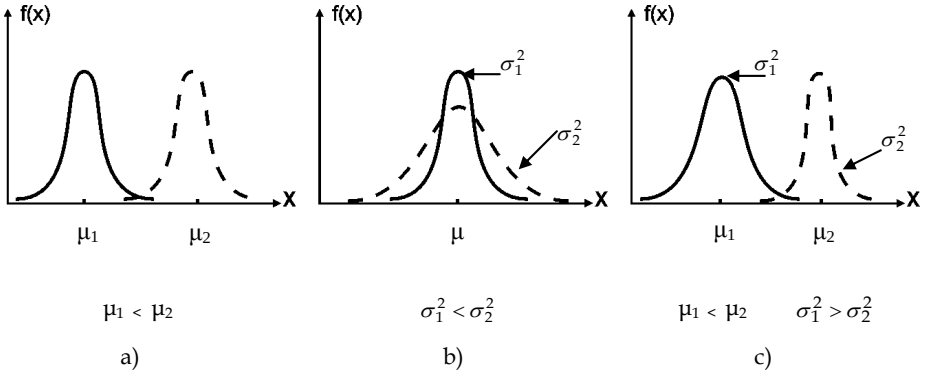
Kako je kriva simetrična, 50% njene površine nalazi se lijevo od normale, podignute nad aritmetičkom sredinom, a 50% desno. Pošto je površina ispod krive, u stvari, vjerovatnoća, slijedi da vjerovatnoća da slučajna promjenljiva X uzme neku vrijednost manju (ili veću) od aritmetičke sredine iznosi 0,5.

6. Normalan raspored je u potpunosti definisan sa

dva parametra, μ i σ^2. Dakle, kao i kod svih do sada obrađivanih rasporeda, postoji čitava familija normalnih rasporeda, u zavisnosti od mogućih vrijednosti μ i σ^2.

Bilo koji normalan raspored možemo, u opštem obliku, označiti sa $N(\mu, \sigma^2)$.

Slika 4.l4 prikazuje različite kombinacije dva normalna rasporeda.



Slika 4.14 Različiti oblici normalnih krivih, u zavisnosti od vrijednosti parametra μ i σ^2

7. Pretpostavimo da smo povukli normale na

udaljenosti od jedne standardne devijacije u oba smjera u odnosu na aritmetičku sredinu. Površina ograničena ovim linijama, X osom i krivom $f(x)$ iznosiće približno 68% od čitave površine (koja, naravno, iznosi 1). Znači, u razmaku ± 1 standardne devijacije od aritmetičke sredine nalazi se približno 68% površine svake normalne krive.

Time smo upravo i odredili vjerovatnoću da normalna slučajna promjenljiva X uzme neku vrijednost u navedenom intervalu:

$$P(\mu-\sigma<X<\mu+\sigma)\approx 0,68\,.$$

Ako povučemo normale na razdaljini od dvije standardne devijacije od aritmetičke sredine u oba smjera do presjeka sa krivom, dobijena površina iznosiće približno 95% čitave površine, odnosno u terminima vjerovatnoće:

$$P(\mu-2\sigma<X<\mu+2\sigma)\approx 0,95\,.$$

Konačno, u intervalu ± 3 standardne devijacije od aritmetičke sredine obuhvaćeno je približno 99,7% površine čitave krive, tj:

$$P(\mu-3\sigma<X<\mu+3\sigma)\approx 0,997\,.$$

Vidimo da je površina koja se nalazi na krajevima normalnog rasporeda izvan intervala ± 3 standardne devijacije od aritmetičke sredine zanemarljivo mala.

Ova osobina ima izuzetno veliku primjenu u statističkom zaključivanju.

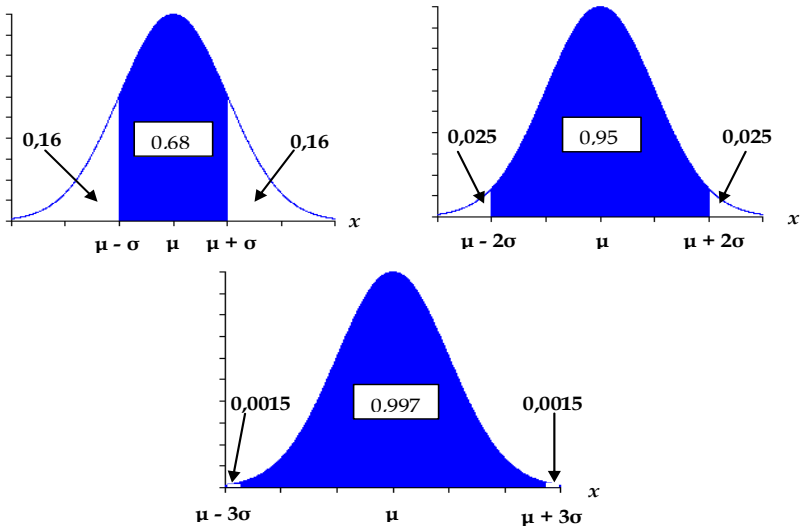
Naziva se često "pravilo 68 − 95 − 99,7".

Može se dokazati da normalna slučajna promjenljiva ispunjava još dvije značajne osobine:

- ako slučajna promjenljiva X ima normalan raspored, tada će i njena linearna transformacija $Y = a + bX$ takođe imati normalan raspored, i
- suma n nezavisnih promjenljivih takođe ima normalan raspored. Ako je: $X_1 : N(\mu_1, \sigma_1^2)$ i $X_2 : N(\mu_2, \sigma_2^2)$,

tada je: $(X_1 + X_2) : N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$;

takođe je: $(X_1 - X_2) : N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.



Slika 4.15 Odnos između površine formirane na udaljenosti $\pm(1, 2 \text{ i } 3\sigma)$ od aritmetičke sredine i površine čitave normalne krive

Značaj normalnog rasporeda

Normalan raspored predstavlja najznačajniji teorijski raspored vjerovatnoće iz sljedećih razloga:

1. Veliki broj pojava u prirodi i društvu ima približno

normalan raspored. Tipični primjeri su visina, težina, krvni pritisak, rezultati na testovima inteligencije, greške u mjerenju itd. Generalno, ako veliki broj faktora utiče na neku pojavu na aditivan način, i uticaj svakog od njih je veoma mali, može se očekivati da ta pojava slijedi normalan raspored.

Normalan raspored može poslužiti kao odlična aproksimacija raznih prekidnih rasporeda (najčešće binomnog i Poisson-ovog), a veliki broj prekidnih rasporeda, pod posebnim uslovima, teži normalnom rasporedu.

2. Iz normalnog rasporeda je izveden veliki broj drugih

neprekidnih rasporeda, koji također imaju značajno mjesto u statističkoj analizi, kao što su Student-ov (ili t) raspored, χ^2 raspored i F raspored, o kojima će kasnije biti riječi.

3. Normalan raspored predstavlja osnovu za

parametarsko statističko zaključivanje zbog: a) njegove veze sa tzv. Centralnom Graničnom Teoremom,

i b) zbog toga što parametarski metodi imaju zajedničku pretpostavku da osnovni skup iz koga se uzima uzorak ima normalan raspored.

Čak i neparametarski statistički metodi koriste normalan raspored kao neophodno sredstvo za donošenje odluke u slučaju velikih uzoraka.

Standardizovan normalan raspored

Postoji čitava familija različitih normalnih rasporeda, u zavisnosti od vrijednosti parametara μ i σ^2. Kako izračunati vjerovatnoću da normalna slučajna promjenljiva X uzme vrijednost u nekom intervalu?

Ideja je u tome da se za ovaj raspored konstruišu tablice vjerovatnoće i, zatim, da se bilo koji normalan

raspored, postupkom standardizacije, svede na standardizovani raspored, i rješenje problema, tj. vjerovatnoće pronađu u tablicama.

Takav normalan raspored se naziva standardizovani normalan raspored, a postupak koji je neophodno primijeniti, standardizacija.

Za normalan raspored kažemo da je u standardizovanom obliku ako je njegova aritmetička sredina jednaka nuli, a varijansa, odnosno standardna devijacija jednaka jedinici.

Formulu za funkciju gustine standardizovane normalne promjenljive možemo dobiti ako u formuli za normalan raspored u opštem obliku zamijenimo $\mu = 0$ i $\sigma^2 = 1$.

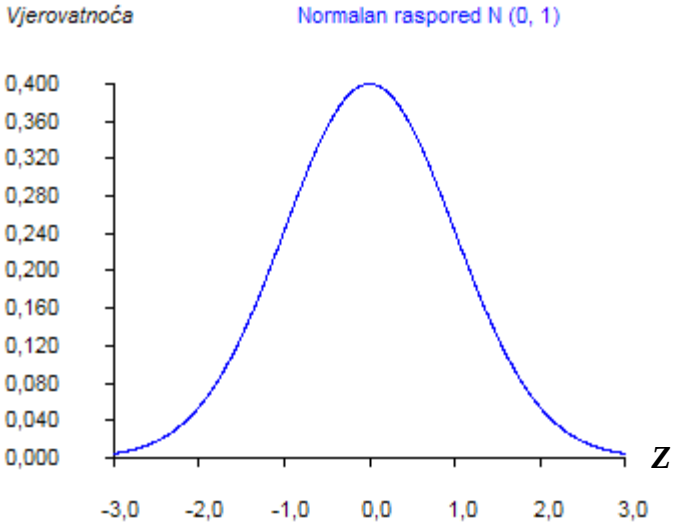
Uobičajeno je da se standardizovana normalna promjenljiva označava sa Z, pa se formula za njen raspored vjerovatnoće može napisati u vidu:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{z^2/2} \qquad -\infty < z < +\infty$$

Kraće se piše:

$$Z : N(0,1)$$

Slika 4.16 prikazuje grafikon standardizovanog normalnog rasporeda.



Slika 4.16 Standardizovan normalan raspored

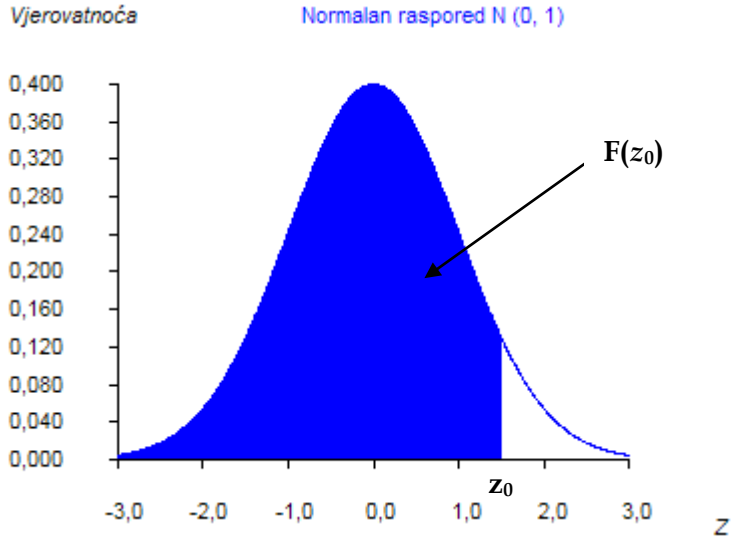
Tablica 3 daje vrijednost funkcije rasporeda standardizovane normalne promjenljive.

Kao kod svake slučajne promjenljive, i ovdje funkcija rasporeda, $F(z_0)$, pokazuje vjerovatnoću da slučajna promjenljiva Z uzme vrijednost manju ili jednaku određenoj vrijednosti z_0:

$$F(z_0) = P(Z \leq z_0).$$

Grafički, ova vjerovatnoća je jednaka osjenčenoj površini ispod standardizovane normalne krive od $-\infty$

do tačke z_0 na Slici 4.17.



Slika 4.17 Funkcija rasporeda standardizovane normalne promjenljive

Budući da je ukupna površina ispod krive jednaka 1, jasno je da nešrafirani dio površine ispod normalne krive iznosi 1 - $F(z_0)$.

Ako bi se direktno određivala vrijednost funkcije rasporeda. morao bi se izračunavati određeni integral:

$$F(z_0)=\int_{-\infty}^{z_0}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\,.$$

Međutim, za ovim nema potrebe − rezultati integracije su već sadržani u Tablici 3.

Objasnimo upotrebu ovih tablica.

Na primjer, vjerovatnoća da slučajna promjenljiva Z uzme vrijednost manju ili jednaku 2,25 iznosi:

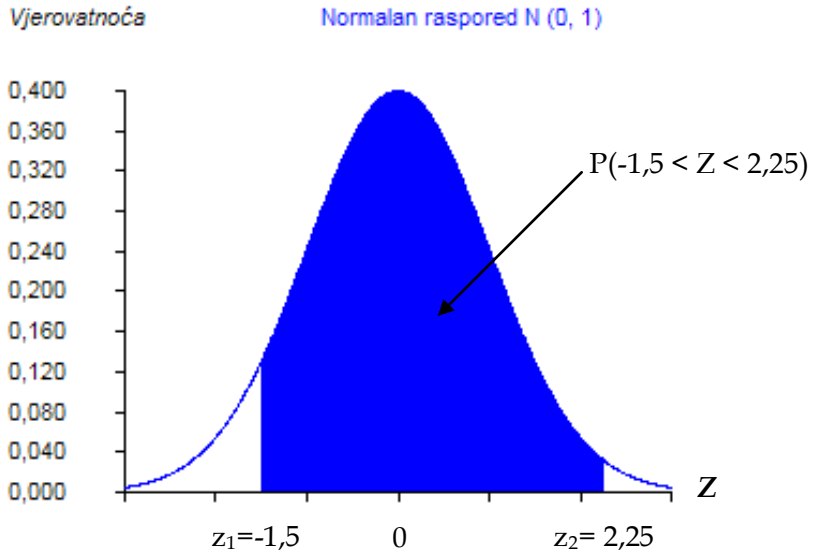
$$P(Z \leq 2,25) = F(2,25) = .9878 = 0,9878 .$$

Očigledno je da Tablicu 3 možemo koristiti i u suprotnom smjeru, odnosno za proizvoljnu vjerovatnoću možemo odrediti onu tačku z_0 na Z osi za koju je funkcija rasporeda jednaka toj vjerovatnoći.

Na primjer, ako je poznato da vrijednost funkcije rasporeda iznosi 0,9878, na osnovu Tablice možemo vidjeti da tačka z_0 iznosi 2,25.

Upotreba Tablice standardizovanog rasporeda u određivanju vjerovatnoće da Z uzme vrijednost u nekom intervalu:

Potražimo najprije vjerovatnoću da Z pada u interval (-1,5, 2,25).



Slika 4.18 Vjerovatnoća da Z uzme vrijednost u intervalu (-1,5, 2,25)

Tražena vjerovatnoća se u opštem slučaju izračunava korišćenjem relacije (4.8), kao razlika funkcije rasporeda gornje i donje granice intervala:

$$P(z_1 < Z < z_2) = F(z_2) - F(z_1),$$

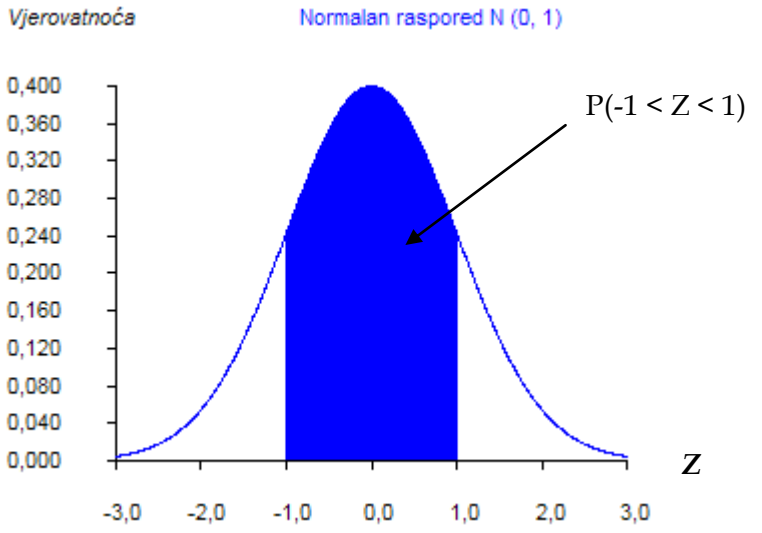
odnosno, u našem primjeru:

$$P(-1,5 < Z < 2,25) = F(2,25) - F(-1,5) = 0,9878 - 0,0668 = 0,921.$$

Odredimo sada vjerovatnoću da se Z nađe u intervalu čije su granice simetrične u odnosu na Y osu, tj. aritmetičku sredinu.

Ovo će nam posebno biti potrebno kod formiranja intervala pouzdanosti.

Ako posmatramo interval ograničen $\pm 1\sigma$ od aritmetičke sredine, tražena vjerovatnoća se može sagledati kao šrafirani dio na Slici 4.19, budući da je $\sigma = 1$.



Slika 4.19 Vjerovatnoća da se Z nađe u intervalu $\pm\sigma$ od aritmetičke sredine, tj. (-1, 1)

Ovu vjerovatnoću možemo izračunati na isti način kao u prethodnom primjeru, odnosno kao:

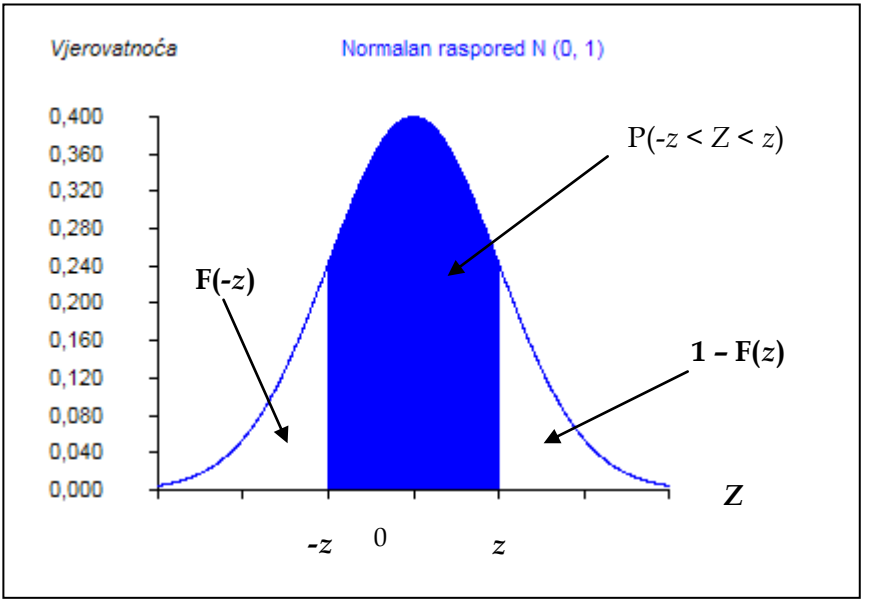
$$P(-1 < Z < 1) = F(1) - F(-1) = 0,8413 - 0,1587 = 0,6826.$$

Uopštimo sada ovaj rezultat i potražimo vjerovatnoću da Z uzme vrijednost između bilo koje dvije simetrične

tačke (-z, z).

Tražena vjerovatnoća se, na osnovu do sada iznijetog, može odrediti kao:

$$P(-z < Z < z) = F(z) - F(-z)$$



Slika 4.20 Vjerovatnoća da Z pada u interval između dvije simetrične tačke

Primjetimo da su zbog simetrije normalne krive obje nešrafirane površine na Slici 4.21 jednake, odnosno $F(-z) = 1 - F(z)$.

$$P(-z < Z < z) = F(z) - [1 - F(z)] = 2F(z) - 1$$

Dakle, da bi se odredila vjerovatnoća da Z uzme

vrijednost između dvije simetrične tačke, dovoljno je koristiti funkciju rasporeda od gornje granice intervala.

U našem primjeru vjerovatnoća iznosi:

$$P(-1 < Z < 1) = 2F(1) - 1 = 2 \cdot 0,8413 - 1 = 0,6825.$$

Ako je, na primjer, poznato da vjerovatnoća iznosi 0,95, granice intervala možemo, na osnovu 4.12, odrediti postavljajući jednakost:

$$0,95 = 2F(z) - 1,$$

pa će $F(z)$ iznositi 0,975. Odgovarajuća vrijednost u Tablici 3 je z = +1,96, a na osnovu simetrije normalne krive zaključujemo da donja granica iznosi -1,96.

Znači, traženi interval iznosi (-1,96, 1,96).

Do sada smo govorili samo o primjeni Tablice 3 u izračunavanju pojedinih vjerovatnoća kod standardizovanog normalnog rasporeda.

Kako, dakle, odrediti vjerovatnoće za normalan raspored u opštem obliku, odnosno sa proizvoljnom aritmetičkom sredinom i standardnom devijacijom? Uzmimo Primjer 4.2 da objasnimo ovaj postupak, koji se, kako smo vidjeli, naziva standardizacija.

PRIMJER Posmatrajmo proizvodnju sijalica i pretpostavimo da njihov vijek trajanja ima približno

normalan raspored sa aritmetičkom sredinom μ = 100 časova i standardnom devijacijom σ = 20 časova. Kolika iznosi proporcija sijalica sa vijekom trajanja između 60 i 90 časova?

Prvi korak u rješavanju je u tome da se dati problem iskaže "jezikom vjerovatnoće", odnosno simbolički. Problem se može prevesti u sljedeće pitanje: koliko iznosi vjerovatnoća da će na slučaj uzeta sijalica imati vijek trajanja između 60 i 90 časova, tj:

$$P\ (60 < X < 90).$$

U drugoj etapi primijenićemo postupak standardizacije. Da bismo izračunali navedenu vjerovatnoću potrebno je najprije prevesti dati normalan raspored u standardizovan oblik.

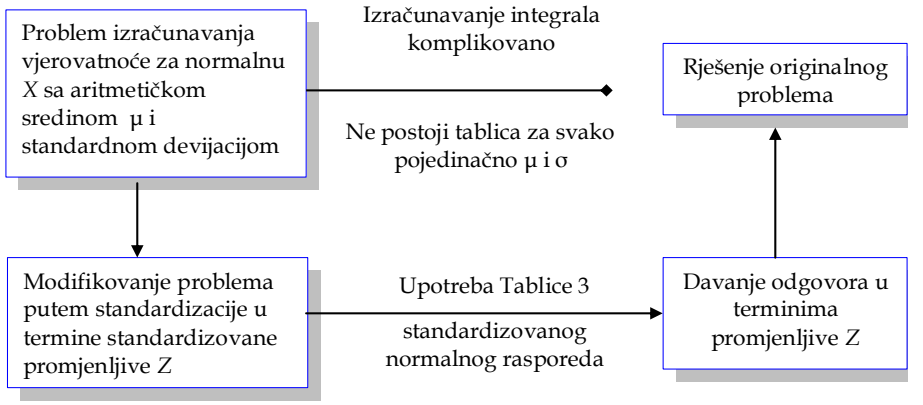
Bilo koja normalna slučajna promjenljiva X, sa aritmetičkom sredinom μ i standardnom devijacijom σ može se transformisati u standardizovanu normalnu promjenljivu Z preko formule:

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

Dok je originalna promjenljiva X bila izražena u određenim mjernim jedinicama (u našem primjeru, u časovima), putem transformacione formule nova slučajna promjenljiva Z ima standardizovanu skalu, nezavisnu od mjernih jedinica.

Na taj način, bez obzira na mjerne jedinice i kombinaciju parametra μ i σ^2, uvijek ćemo biti u stanju da podatke prevedemo u standardizovane, iskazane u standardnim devijacijama, i da potrebne vjerovatnoće

odredimo korišćenjem tablice standardizovanog normalnog rasporeda.



Slika 4.21 Postupak analize korišćenjem standardizovanog normalnog rasporeda

Vjerovatnoća da se slučajna promjenljiva X nalazi u nekom intervalu (x_1, x_2) može se sada izračunati pomoću sljedeće relacije:

$$P(x_1 < X < x_2) = P(z_1 < Z < z_2)$$

gdje su:

$$z_1 = \frac{x_1 - \mu}{\sigma} \quad \text{i} \quad z_2 = \frac{x_2 - \mu}{\sigma}.$$

Vidimo da je vjerovatnoća da slučajna promjenljiva $X\left[X:N(\mu,\sigma^2)\right]$ uzme vrijednost u intervalu (x_1,x_2) jednaka vjerovatnoći da standardizovana promjenljiva $Z[Z:N(0,1)]$ uzme vrijednost u razmaku $(x_1\text{-}\mu)/\sigma$ i $(x_2\text{-}\mu)/\sigma$. Ovo ćemo ilustrovati grafički Slikom 4.22.

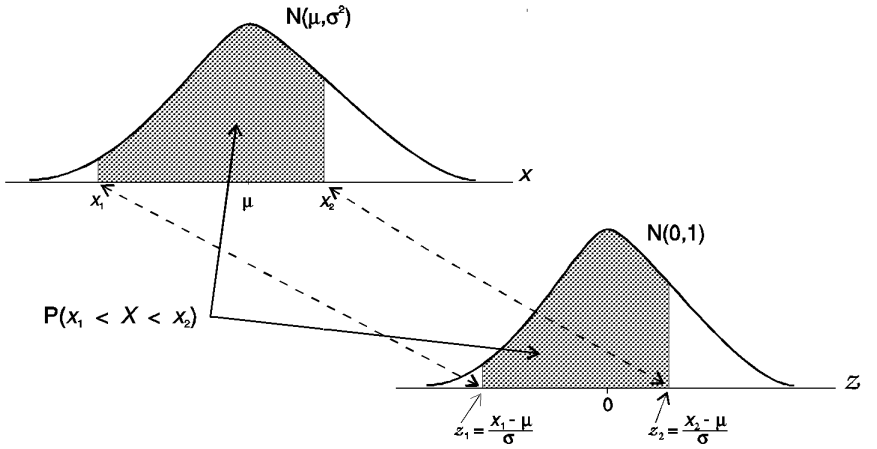
Na osnovu relacije u mogućnosti smo da izračunamo vjerovatnoću da će na slučaj uzeta sijalica imati vijek trajanja između 60 i 90 časova:

$$P(60<X<90)=P(\frac{60-\mu}{\sigma}<\frac{X-\mu}{\sigma}<\frac{90-\mu}{\sigma})=$$

$$=P(\frac{60-\mu}{\sigma}<Z<\frac{90-\mu}{\sigma})=$$

$$=P(\frac{60-100}{20}<Z<\frac{90-100}{20})=P(-2<Z<-0,5)=$$

$$=F(-0,5)-F(-2)=0,3085-0,0227=0,2758.$$



Slika 4.22 Povezanost između normalnog rasporeda sa aritmetičkom sredinom μ i varijansom σ^2, i standardizovanog normalnog rasporeda sa aritmetičkom sredinom μ=0 i varijansom $\sigma^2 = 1$

U trećoj etapi potrebno je dati interpretaciju rezultata, odnosno odgovor na postavljeno pitanje: u skupu svih sijalica, tačno učešće sijalica sa vijekom trajanja između 60 i 90 časova iznosi 27,58%.

PRIMJER Koliko u Primjeru 4.2 iznosi vrijeme do kojeg će trajati ukupno 95% sijalica u skupu?

Analiziranjem ovog problema vidimo da se traži postupak suprotan u odnosu na prethodni primjer, možemo ga nazvati "antistandardizacija".

U prvoj etapi potrebno je da postavimo problem u vidu probabilističkog iskaza. Pošto se podaci odnose na čitav skup, riječ "proporcija" ćemo zamijeniti sa vjerovatnoća.

Slijedi da se problem može postaviti na sljedeći način:

Koliko iznosi vrijednost x za koju je $P(X \le x) = 0,95$?

U drugoj etapi koristimo tablice standardizovanog normalnog rasporeda da potražimo analognu tačku (vrijednost) z, za koju važi prethodna relacija. Budući da je:

$$P(X \le x) = P(Z \le z) = F(z) = 0,95,$$

iz Tablice 3 nalazimo vrijednost z =1,64.

U trećoj etapi (suprotnoj od standardizacije) koristimo izraz (4.13) da na osnovu poznatog z i σ izračunamo nepoznato x:

$$x = \mu + z\sigma = 100 + 1,64 \cdot 20 = 132,8.$$

Konačno, četvrta i posljednja etapa je davanje odgovora na postavljeno pitanje, odnosno interpretacija rezultata: vjerovatnoća da na slučaj izabrana sijalica ima vijek trajanja do 132,8 časova je 0,95, odnosno, tačno učešće sijalica sa vijekom trajanja do 132,8 časova iznosi 95%.

Na kraju razmatranja standardizovanog normalnog rasporeda napomenimo da je uz pomoć statističkog softvera neuporedivo lakše izračunavati vjerovatnoće.