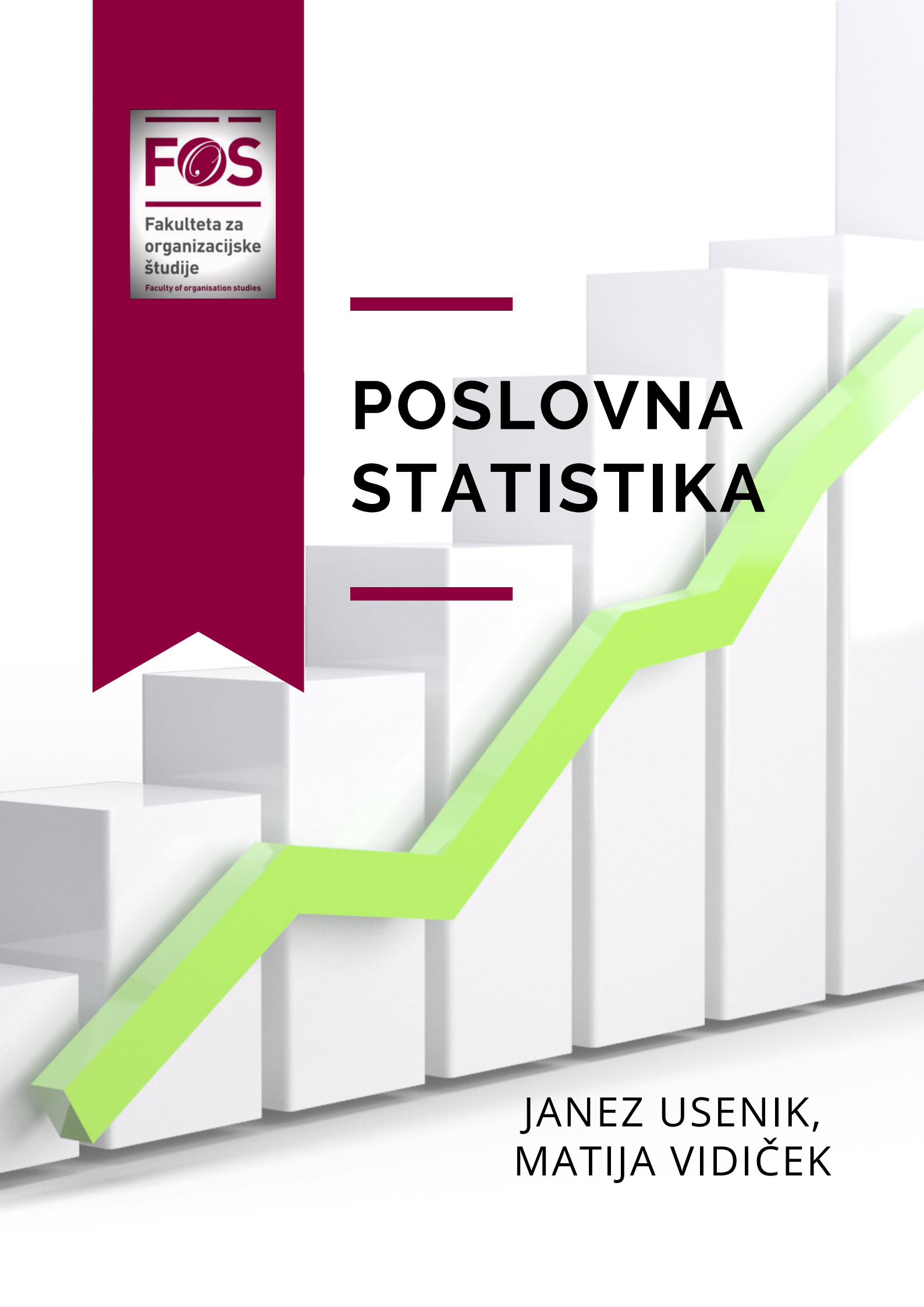


Fakulteta za organizacijske študije
Faculty of organisation studies

POSLOVNA STATISTIKA

JANEZ USENIK,
MATIJA VIDIČEK

Fakulteta za
organizacijske študije
Faculty of organisation studies

Prof. ddr. Janez USENIK

Matija VIDIČEK

POSLOVNA STATISTIKA

Novo mesto, 2020

Poslovna statistika
Janez Usenik, Matija Vidiček

Strokovna recenzija:
Prof. dr. Robert Vodopivec
Doc. dr. Franc Brcar

Založila: Fakulteta za organizacijske študije Novo mesto

Dostop do gradiva: https://www.fos-unm.si/si/dejavnosti/zaloznistvo/

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID=304080384
ISBN 978-961-6974-47-9 (epub)
ISBN 978-961-6974-48-6 (pdf)
ISBN 978-961-6974-49-3 (mobi)

KAZALO VSEBINE

1 UVOD

Beseda in pojem *statistika* pomeni več stvari, običajno pa imamo v mislih obravnavanje množičnih pojavov.

Pod pojmom statistika lahko razumemo tudi državne organe (v naši državi je to Statistični urad Republike Slovenije, http://www.stat.si/), ki zbirajo in analizirajo podatke o vseh množičnih pojavih, pomembnih za delovanje države, npr. število prebivalcev, število stanovanj, število avtomobilov, število študentov, povprečno plačo, rast ali padec industrijske proizvodnje, proizvedena in potrošena električna energija v posameznih obdobjih itd…

Statistika pa je predvsem znanstvena disciplina, ki se ukvarja z zbiranjem podatkov o množičnih pojavih, njihovim razvrščanjem in prikazovanjem, z znanstvenim razvojem metod za obdelavo podatkov, s samo obdelavo in zlasti analizo dobljenih rezultatov.

Statistika se torej ukvarja z obdelavo, analizo in predstavitvijo množičnih pojavov v znanosti, gospodarstvu, industriji, kmetijstvu, prometu, trgovini, šolstvu, varstvu okolja in nasploh na vseh področjih človekovega raziskovanja množičnih pojavov.

1.1 Osnovni pojmi

Osnovni pojmi statistike so naslednji: pojav, populacija, statistična enota, statistična množica, vzorec, statistična spremenljivka (ali tudi statistični znak), statistični parameter.

Pojav je predmet konkretnega statističnega proučevanja in se mora dogajati množično[1] v času in prostoru.

Pojav je lahko:

-oseba (študent, občan, krajan,…)
-žival (konj, krava, ovca,…)
-stvar, izdelek (avto, hiša, hladilnik, …)
-pravna tvorba (podjetje, agencija, zavod, ...)
-administrativna enota (krajevna skupnost, občina, …)
-gospodarska tvorba (trgovina, proizvodno podjetje, svetovalno podjetje, …)
-dogodek (rojstvo, poroka, nesreča, …)
-poskus (cepljenje živali, ljudi,…)

[1] Besedna zveza »množični pojav« pomeni dejstvo, da se mora pojav zgoditi večkrat (množično), ne le npr. dvakrat ali trikrat.

Populacija je množica elementov, na katere konkretni pojav učinkuje. Populacija je npr. lahko množica študentov neke fakultete, lahko so polnoletni občani nekega mesta in podobno.

Statistične enote so posamezni elementi populacije (študenti, občani,…) in morajo biti opredeljene s

- krajevnega,
- časovnega,
- stvarnega vidika.

Statistična množica je množica vseh statističnih enot, ki izpolnjujejo vse vnaprej dogovorjene pogoje, npr. vsi prebivalci v neki občini, ki imajo status študenta.

Vzorec je (neprazna) podmnožica statistične množice.

Statistična spremenljivka ali tudi ***statistični znak*** je neka posebna značilnost statističnih enot.

Statistični parameter pa je konkretna značilnost statistične množice.

Primer:
Vzemimo, da proučujemo odrasle nezaposlene osebe po spolu, starosti in izobrazbi v Mestni občini Novo mesto za leto 2017.
Gornji pojmi bi bili v takšnem primeru naslednji:

- pojav je nezaposlena odrasla oseba v MO NM v letu 2017,
- statistična enota je posamezna odrasla oseba, ki
  - ni zaposlena – stvarna opredelitev,
  - je iz MO NM – krajevna opredelitev,
  - ni zaposlena leta 2017 – časovna opredelitev,
- statistična spremenljivka je proučevana oseba po spolu, izobrazbi, starosti,
- statistična množica je množica vseh odraslih nezaposlenih oseb v MO NM leta 2017,
- statistični parameter je lahko delež žensk, mlajših od 30 let, lahko je delež moških s srednjo izobrazbo in podobno, seveda pa gre za osebe dane statistične množice.

♦♦♦

Statistične enote delimo glede na časovno opredelitev na:

- realne enote s trenutkom, ko jih evidentiramo (npr. študent, podjetje, krajevna skupnost…),
- dogodke, ki se zgodijo tako rekoč v trenutku (npr. prometna nesreča, poroka, rojstvo, smrt, …),
- dogajanja, ki trajajo dalj časa (npr. gradnja ceste, študij, …),.

Statistične spremenljivke delimo glede na način izražanja njihovih vrednosti na

- številske (numerične), izražamo jih *kvantitativno* s števili,
- opisne (atributivne), izražamo jih *kvalitativno* z besedami.

Številske statistične spremenljivke so lahko

- zvezne (zavzamejo lahko vsako vrednosti iz nekega v naprej znanega oz. določenega intervala),
- diskretne (zavzamejo le nekatere vrednosti iz nekega v naprej znanega oz. določenega intervala).

Opisne statistične spremenljivke so

- nominalne (razvrščene po nazivu),
- ordinalne (razvrščene v skupine).

Primer:

Statistična spremenljivka	Vrednost statistične spremenljivke	Vrsta statistične spremenljivke
Mesečna štipendija v EUR	0; 100; 150,50; 165,76; ….	numerična, zvezna
Trajanje študija v letih	3; 3,5; 4,2; 5,58;…..	numerična, zvezna
Število otrok v družini	0, 1, 2, 3, ….	numerična, diskretna
Spol	ženski, moški	opisna, nominalna
Izobrazba	OŠ, SŠ, VSŠ, UNI,…	opisna, ordinalna

♦♦♦

1.2 Statistična raziskava

Metodologija (potek) statistične raziskave je določena z več koraki, ki jih mora izvajalec raziskave opraviti.
Najprej je seveda potrebno *določiti namen in cilje* statistične raziskave ter od tod definirati *vsebino dela.*

Načrtovati moramo oblike in možnosti *pridobivanja podatkov* ter izbrati statistične metode, ki bodo uporabljene v raziskavi.
Podatke je nato potrebno *zbrati*, jih ustrezno *urediti* in *prikazati* (tabele, grafikoni).
Pravilno urejene podatke moramo *analitično obdelati*, pri čemer je pomembno izbrati primerne statistične metode glede na cilj in namen raziskave.
Ob koncu sledi še *analiza* ter *interpretacija* rezultatov.

Nepravilno izvedena statistična obdelava lahko prikaže izkrivljene in napačne rezultate, zato ob tej priliki poudarimo, da mora biti statistična raziskava opravljena strogo nevtralno, kazalniki, ki jih računamo, morajo biti primerno izbrani, pravilno izračunani in jasno, nedvoumno in skrajno objektivno interpretirani.

V nadaljevanju nekoliko podrobneje poglejmo fazo zbiranja podatkov in morebitne napake, ki se v raziskavi lahko pojavijo.

1.2 Zbiranje podatkov

Kot je bilo že rečeno, moramo pri vsaki statistični raziskavi pravočasno in v skladu s postavljenim ciljem in namenom raziskave pridobiti ustrezne podatke.
Podatke delimo na:

- primarne, to so tisti, ki jih moramo pridobiti posebej za to raziskavo,
- sekundarne, to so podatki, ki so že zbrani in jih moremo uporabiti brez dodatnega zbiranja, npr. pri kakšni predhodni raziskavi, morda jih lahko dobimo na statističnem uradu in podobno.

Primarne podatke lahko pridobimo na več različnih načinov:

- z ustnim spraševanjem/anketiranjem (po telefonu, z obiskom, na ulici, …),
- s pisnim spraševanjem/anketiranjem (po pošti, elektronski pošti),
- z opazovanjem in merjenjem (npr. štetje prometa, merjenje temperature v mestih po Sloveniji),
- s poskušanjem (npr. test kvalitete proizvoda),
- s stalnim nadzorom in evidentiranjem (npr. število dnevno narejenih proizvodov v tovarni),
- s tekočo registracijo (rojstvo, prodaja avtomobila,…)

S samimi načini pridobivanja podatkov se ne bomo podrobneje ukvarjali.

1.3 Napake

Pri zbiranju podatkov in kasneje pri njihovi obdelavi lahko pride do različnih napak, ki popačijo dejanske rezultate in s tem izničijo ves vloženi trud, zato moramo skrbno paziti, da se to ne zgodi.

V splošnem govorimo lahko o naslednjih napakah:

- sistematična napaka (napaka merilnega inštrumenta, npr. pri meritvi temperature toplomer ni pravilno umerjen),
- napaka pri odčitavanju (npr. namesto 37,3°C lahko iz različnih vzrokov preberemo 37,4°C),
- napaka pri zapisu podatka (npr. namesto 4,7 napišemo 7,4),
- napake zaradi napačnega razvrščanja podatkov (npr. kakšen podatek pri razvrščanju lahko pomotoma izpade, kakšen podatek morda zapišemo dvakrat, podatek pri vnosu v računalnik vnesemo v napačni stolpec in podobno).

Pri obdelavi podatkov pa lahko pride še do različnih računskih napak, zlasti pri zaokroževanju numeričnih rezultatov. Do zaokroževanja pride, ko približno število vsebuje odvečna (netočna) mesta. Pri zaokroževanju obdržimo le točna (pravilna) mesta oz. številke, odvečna (nezanesljiva) pa izpustimo, pri čemer zadnjo obdržano številko povečamo za ena, če je prva izpuščena številka večja od 4. Pri zaokroževanju se pojavljajo dodatne napake, ki niso večje od polovice enote vrednosti zadnjega pravilnega znaka zaokroženega števila. Da bi bile po zaokroževanju vse številke (znaki) pravilne, pred zaokroževanjem napaka ne sme biti večja od polovice enote tistega mesta, do koder želimo število zaokrožiti.

Pri numeričnem računanju veljajo glede za osnovne računske operacije takšna pravila zaokroževanja:

- pri seštevanju in odštevanju približnih števil v rezultatu obdržimo toliko decimalnih mest, kolikor jih ima podatek z najmanj decimalnimi mesti,
- pri množenju in deljenju v rezultatu obdržimo toliko veljavnih mest, kot jih ima podatek z najmanj veljavnimi mesti.

Primer:
Vzemimo pravokotnik s stranicama: $a=11{,}6\,\mathrm{m}$, $b=13{,}48\ \mathrm{m}$. Izračunajmo njegov obseg in njegovo ploščino in rezultat smiselno zaokrožimo.

Rezultat:

a) Obseg: po formuli za obseg pravokotnika je $o=2(a+b)=2(11{,}6+13{,}48)m=50{,}16m$. Ker smo podatke seštevali in ker ima najmanj natančen podatek eno decimalno mesto, bo rezultat zaokrožen na eno decimalno mesto, torej: $o=50{,}2m$. Pri tem smo upoštevali, da je prva izpuščena številka (na drugi decimalki) 6, zato smo zaokroženemu mestu na prvi decimalki dodali 1.
b) Ploščina: po formuli za ploščino pravokotnika je $S=a\cdot b=11{,}6\cdot 13{,}48m^2=156{,}368m^2$. Dobljeni rezultat smo dobili z množenjem, zato ga bomo zaokrožili na toliko mest, kot jih ima najmanj natančen rezultat. Najmanj natančen rezultat (v tem primeru $a=11{,}6$) ima tri mesta, zato bomo na tri mesta zaokrožili tudi rezultat, torej: $S=156m^2$. Številke na tretjem mestu nismo spreminjali, ker številka na prvem izpuščenem mestu ni večja od 4.

♦♦♦

2 UREJANJE IN PRIKAZOVANJE STATISTIČNIH PODATKOV

Pridobljene podatke je potrebno najprej urediti in jih prikazati.

Uredimo jih v obliki

- statistične vrste,
- absolutne strukture,
- frekvenčne porazdelitve,

prikažemo pa v obliki

- preglednice (tabele),
- grafikona.

2.1 Statistična vrsta

Statistična vrsta je zaporedni zapis vrednosti statistične spremenljivke, torej tako, kot smo jih pridobivali v postopku zbiranja. Takšna vrsta je običajno ***neurejena***. Če jo uredimo po nekem dogovorjenem kriteriju (npr. po velikosti podatkov), je statistična vrsta ***urejena.***

Statistične vrste delimo na

- časovne,
- krajevne,
- stvarne.

Primer za *časovno statistično vrsto*:
Vzemimo, da proučujemo število izposojenih knjig v knjižnici od leta 2011 do vključno 2017.

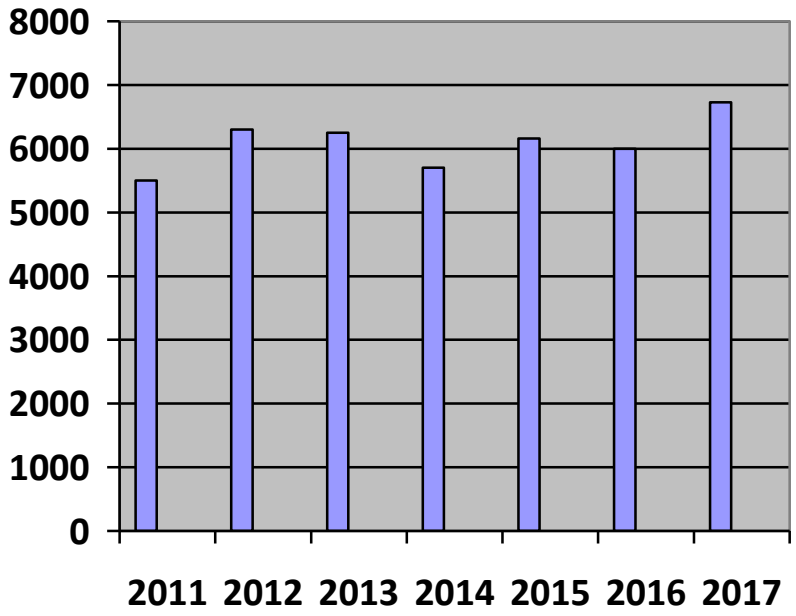
Podatke uredimo v časovno statistično vrsto in jih prikažemo v tabeli (preglednici) in/ali v grafični obliki .

Preglednica:

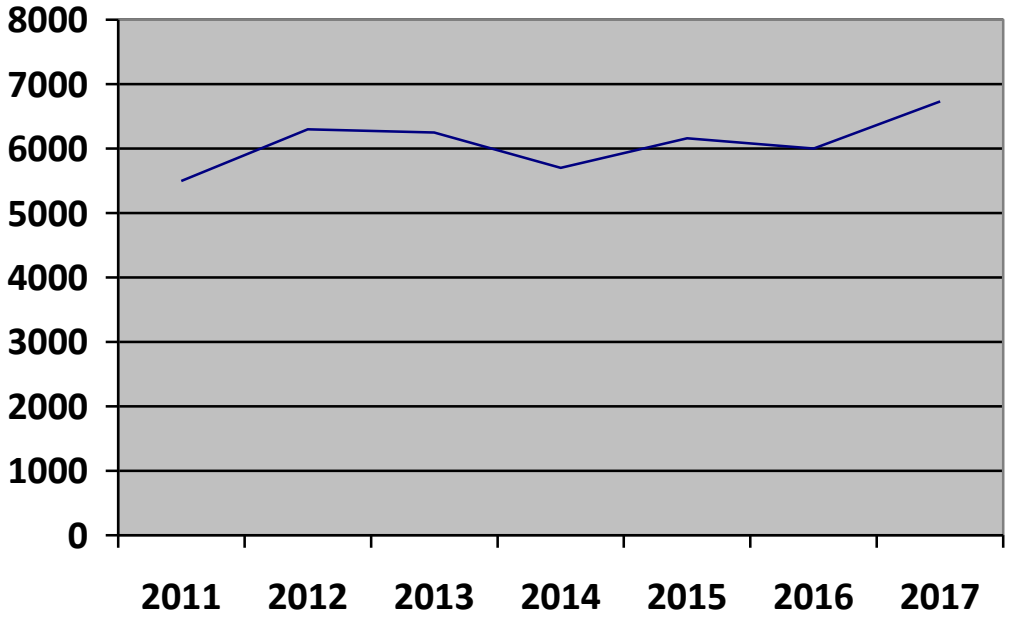
Leto	2011	2012	2013	2014	2015	2016	2017
Št. izposojenih knjig	5500	6300	6250	5700	6160	6000	6730

Grafično (na vodoravno os nanašamo čas, na navpično os pa število izposojenih knjig) podatke časovne statistične vrste prikažemo s stolpnim ali linijskim grafikonom

a) *stolpni grafikon*



b) linijski grafikon



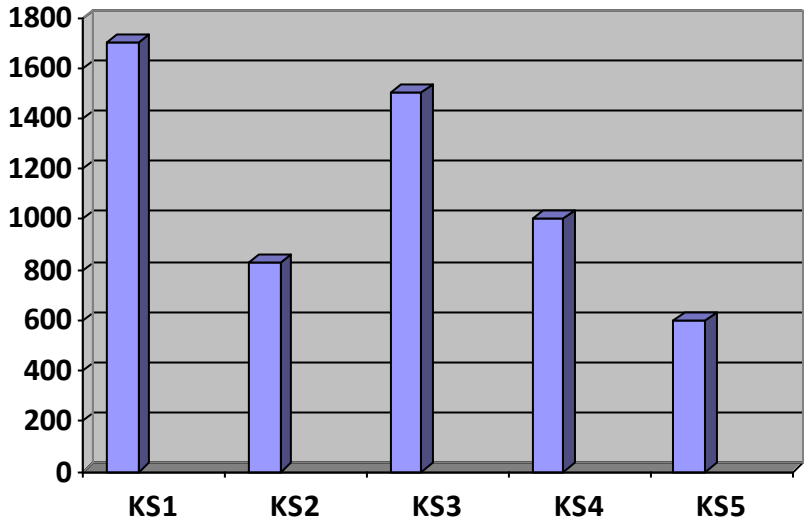
Primer za *krajevno statistično vrsto*:
Proučujemo število registriranih osebnih avtomobilov po krajevnih skupnostih neke občine v letu 2017. Podatke prikažimo v tabeli in grafično.

Tabela:

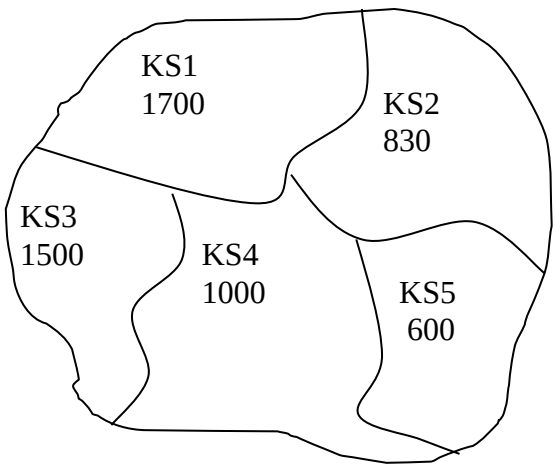
Krajevna skupnost	KS1	KS2	KS3	KS4	KS5
Št. registriranih os. avt.	1700	830	1500	1000	600

Grafično (na vodoravno os nanašamo krajevno skupnost, na navpično os pa število osebnih avtomobilov) podatke krajevne statistične vrste prikažemo s stolpnim diagramom ali s kartogramom. Kartogram pomeni zemljevid (karto) občine s vrisanimi krajevnimi skupnostmi.

a) *stolpni diagram*



b) *kartogram*

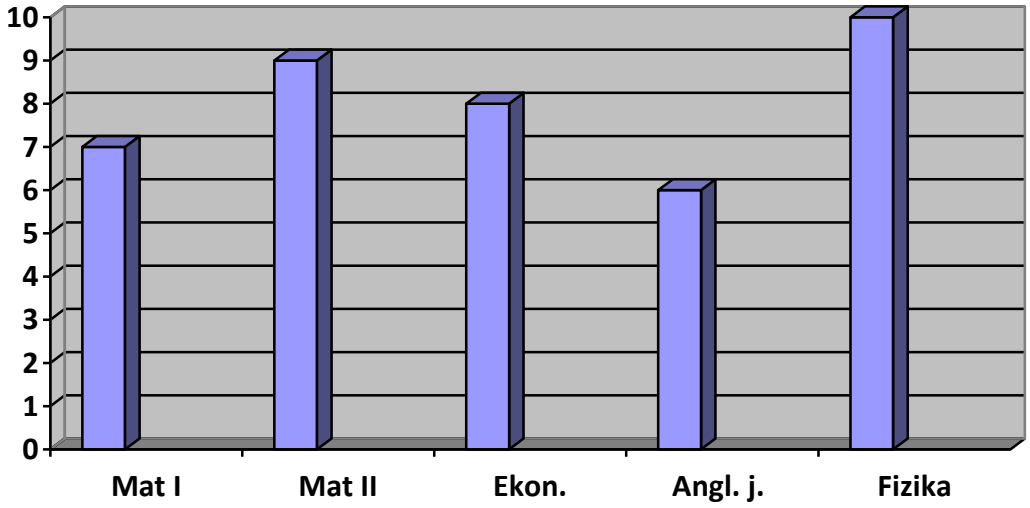


Primer za *stvarno statistično vrsto*
Opazujemo ocene opravljenih izpitov študenta Jožeta po predmetih.

Preglednica

Predmet	Matematika I	Matematika II	Ekonomika	Tuji jezik	Fizika
Ocena	7	9	8	6	10

Grafični prikaz je v primeru stvarne statistične vrste samo stolpni diagram.



2.2 Absolutne strukture

Absolutne strukture so takšne skupine enot neke izbrane statistične množice, ki jim pripada ista vrednost številske ali opisne spremenljivke. Število, ki meri pogostost pojava posamezne spremenljivke, je **frekvenca.**

Strukture prikazujemo v preglednicah, ki so

- **enostavne**: v njih so prikazane skupine enot statistične množice po eni spremenljivki,
- **sestavljene**: v njih so prikazane strukture več množic po isti statistični spremenljivki,
- **kombinirane**: v njih je prikazana struktura množice statističnih enot, ki jih proučujemo po najmanj dveh statističnih spremenljivkah hkrati.

Absolutnih struktur običajno ne prikazujemo grafično.

Primer:
Dve enostavni strukturi

a) struktura študentov iz Novega mesta po spolu (podatek je izmišljen)

Študentke iz NM	Študenti iz NM	Skupaj
245	179	424

b) struktura študentov iz Novega mesta po bivanju med študijem (izmišljen podatek!)

Doma	Drugod (LJ, MB,..)	Skupaj
95	329	424

Obe zgornji enostavni strukturi pa moremo zapisati tudi v eni sami tabeli v obliki kombinirane strukture

Stanovanje med študijem	Študentke	Študenti	Skupaj
Doma	55	40	95
Drugod	190	139	329
Skupaj	245	179	424

♦♦♦

2.3 Frekvenčne porazdelitve

Če imamo zelo veliko vrednosti numeričnih spremenljivk, potem je zapis v vrsti nepregleden. Pregledne postanejo šele, ko jih grupiramo v **razrede.**
Prikaz skupin enot, ki jim pripada ista vrednost (interval) numerične spremenljivke, je frekvenčna porazdelitev.
Število razredov pogojuje širino razredov. Širine vseh razredov naj bodo – če se le da – enake.
Razredov naj bo med 5 in 15.

Ocen za število razredov je več.
Ena je takšna (Sturgesovo pravilo):

$$r = 1 + 3{,}3 \cdot \log N \tag{2.01}$$

Druga ocena je

$$r - 1 \le \frac{y_{\max} - y_{\min}}{i} \le r \tag{2.02}$$

V zgornjih formulah pomenijo:

r – število razredov,
N – število podatkov (numerus),
$y_{\max}$ – največja vrednost spremenljivke (podatek z največjo vrednostjo),
$y_{\min}$ – najmanjša vrednost spremenljivke (podatek z najmanjšo vrednostjo),
i - širina razreda.

Primer:
Za 200 študentov poznamo njihove povprečne ocene. Podatke bomo razvrstili v razrede. Koliko naj bo razredov?

Rešitev:
Po formuli (5.01) dobimo

$$n = 1 + 3,3 \cdot \log N = 1 + 3,3 \cdot \log 200 = 1 + 3,3 \cdot 2,3 = 8,59$$

Razredov bo med 8 in 9.

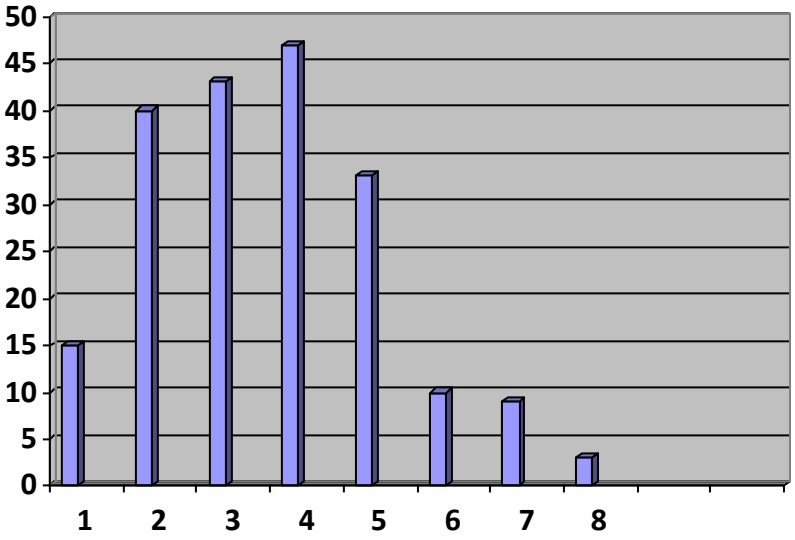
Za boljšo oceno uporabimo še kriterij (2.02). Največja vrednost podatka $y_{\max} = 10$, najmanjša pa $y_{\min} = 6$, zato je razlika $y_{\max} - y_{\min} = 4$. Zaradi pogoja, naj bodo razredi enako široki, je najenostavneje v formuli vzeti $i = 0,5$, tako da je število razredov $r = 8$.

Tabela je potem npr. takšna:

Zap. št.	Povprečna ocena/razred	Frekvenca f_k
1	6,00-6,50	15
2	6,51-7,00	40
3	7,01-7,50	43
4	7,51-8,00	47
5	8,01-8,50	33
6	8,51-9,00	10
7	9,01-9,50	9
8	9,51-10,00	3
	skupaj	200

Opomba: širina razredov v zgornji tabeli pravzaprav ni enaka, prvi razred je za eno stotino večji od ostalih. Ta razlika je tako majhna, da jo brez škode za rezultat lahko zanemarimo.

Grafični prikaz s stolpci se pri frekvenčni porazdelitvi imenuje **histogram.**

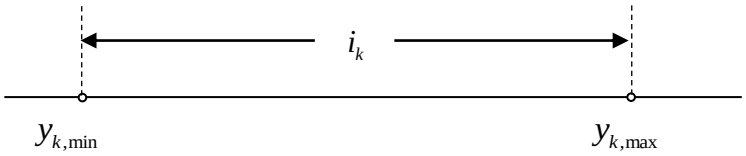


♦ ♦ ♦

Pri frekvenčnih porazdelitvah bomo v nadaljevanju uporabljali naslednje znake:

N – število podatkov/enot,
r – število razredov,
y – spremenljivka,
f_k - frekvenca k-tega razreda, $(k = 1, 2, \ldots, r)$
$y_{\max}$ – največja vrednost spremenljivke (podatek z največjo vrednostjo),
$y_{\min}$ – najmanjša vrednost spremenljivke (podatek z najmanjšo vrednostjo),
$y_{k,\max}$ - zgornja meja k-tega razreda,
$y_{k,\min}$ - spodnja meja k-tega razreda,
$i_k = y_{k,\max} - y_{k,\min}$ - širina k-tega razreda,
$y_k = \frac{y_{k,\max} + y_{k,\min}}{2}$ - sredina k-tega razreda.

Predstavimo grafično k-ti razred:

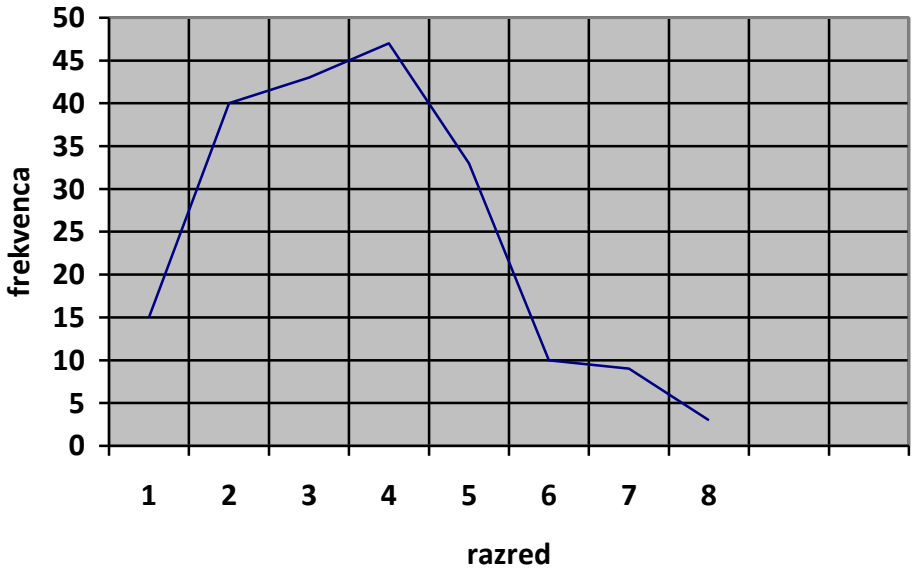


Pri določanju razredov moramo upoštevati nekatere zahteve:

- razredi naj bodo enako široki (če je le mogoče), včasih sta prvi in zadnji razred lahko tudi odprta in s tem različne širine,
- vsak podatek je lahko le v enem razredu,
- meje med razredi morajo biti jasno določene.

Grafični prikaz frekvenčne porazdelitve z linijskim diagramom se imenuje **frekvenčni poligon.** Pri tem poligonu vzamemo na abscisni osi za določanje koordinat *sredino intervala.*

Primer:
Vzemimo 200 študentov iz prejšnjega primera. Frekvenčni poligon je dan na spodnji sliki.



♦ ♦ ♦

Frekvenčne porazdelitve so različnih oblik. Iz histograma ali linijskega poligona je razvidno, kje se podatki zgostijo – tam je grafikon v navpični smeri (naj)višji, doseže torej lokalni maksimum.

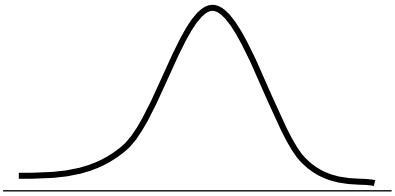
Oblike frekvenčne porazdelitve so lahko:

- unimodalna (ena zgostitev, en maksimum)

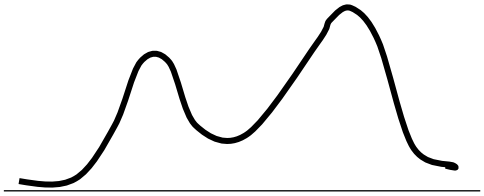
- bimodalna (dve zgostitvi, dva lokalna maksimuma),

- polimodalna (več kot dve zgostitvi, več lokalnih maksimumov)

Grafični prikaz:

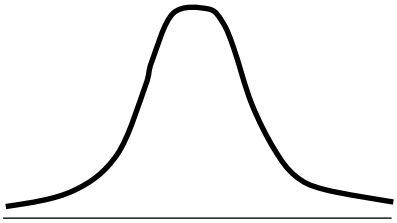


unimodalna

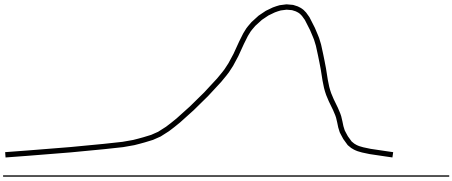


bimodalna

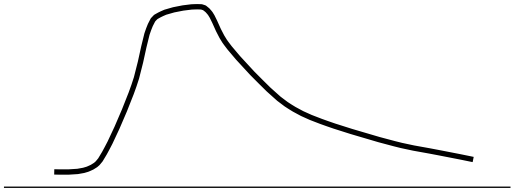
Frekvenčna porazdelitev je lahko še



simetrična



asimetrična v levo



asimetrična v desno

2.4 Kumulativna frekvenčna porazdelitev

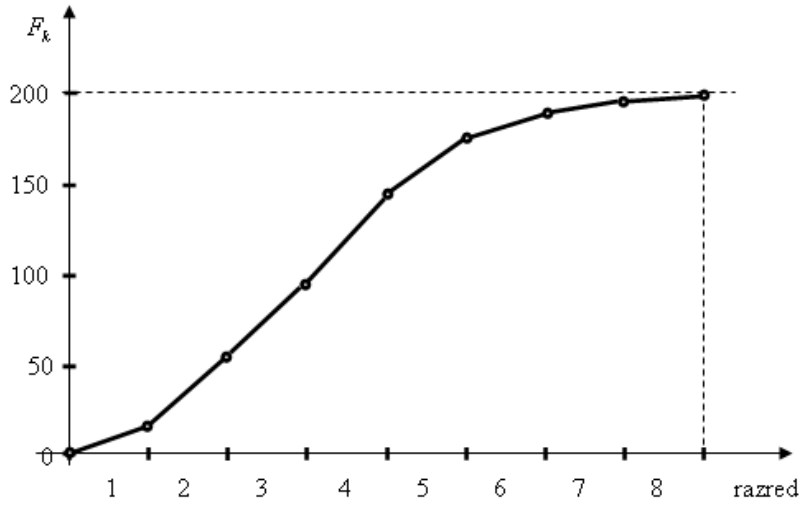
Kumulativno frekvenčno porazdelitev dobimo z vpeljavo kumulativnih frekvenc F_k (označimo jih z veliko črko) po pravilu:

$$F_1 = f_1$$
$$F_2 = f_1 + f_2 = F_1 + f_2$$
$$F_3 = f_1 + f_2 + f_3 = F_2 + f_3$$
$$\cdots\cdots$$
$$F_k = f_1 + f_2 + \cdots + f_k = F_{k-1} + f_k\,, \qquad k = 2, 3, \ldots, r$$

Primer:
Vzemimo spet podatke o povprečnih ocenah študentov, za katere poznamo frekvenčno porazdelitev. Sedaj to tabelo dopolnimo tabelo s stolpcem kumulativnih frekvenc.

Zap. št.	Povprečna ocena/razred	Frekvenca f_k	Kumulativna frekvenca F_k
1	6,00-6,50	15	15
2	6,51-7,00	40	55
3	7,01-7,50	43	98
4	7,51-8,00	47	145
5	8,01-8,50	33	178
6	8,51-9,00	10	188
7	9,01-9,50	9	197
8	9,51-10,00	3	200
-	skupaj	200	-

Grafično:



♦ ♦ ♦

Opomba: Linijski poligon kumulativne frekvence ima značilno obliko. Za unimodalno porazdelitev je oblika krivulje vedno podobna črki S.

3 RELATIVNA ŠTEVILA

Podatki, pridobljeni v procesu statistične raziskave, so za obravnavo običajno smiselni šele tedaj, ko jih primerjamo med seboj ali pa jih primerjamo s kakšno v naprej dogovorjeno vrednostjo. Števila, ki jih dobimo (izračunamo) s takšno primerjavo, imenujemo **relativna števila.**

V tem poglavju bomo spoznali naslednja relativna števila:

- relativne strukture,
- indeksi,
- statistični koeficienti.

3.1 Relativne strukture

Z relativnimi strukturami prikazujemo sestavo statističnih množic. Izražamo jih s:

- strukturnimi deleži,
- strukturnimi odstotki (procenti).

Strukturne deleže izračunamo po formuli

$$f_k^0 = f^0(y_k) = \frac{f_k}{N} \quad , \quad k = 1, 2, \ldots, r \tag{3.01}$$

Strukturne procente pa izračunamo po formuli

$$f_k\% = f\%(y_k) = \frac{f_k}{N} \cdot 100 \quad , \quad k = 1, 2, \ldots, r \tag{3.02}$$

Seveda velja: $\sum_{k=1}^{r} f_k^0 = 1$ in $\sum_{k=1}^{r} f_k\% = 100$.

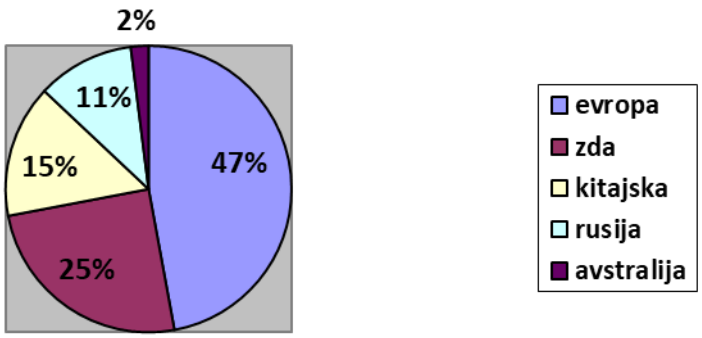
Relativne strukture prikažemo v tabeli, grafično pa imamo več različnih možnosti: prikazovanje s stolpci, s krogi, s kvadrati,…

Primer:
Neko podjetje izvozi za 5.000.000 € izdelkov, od tega v Evropo za 2.355.000 €, v ZDA za 1.245.000 €, na Kitajsko za 750.000 €, v Rusijo za 550.000 € in v Avstralijo za 100.000 €. Prikažimo strukturo izvoza.

Rešitev:
Prikaz v obliki tabele:

Tržišče	Izvoz v €	struktura	
		f_k^0	f_k %
Evropa	2.355.000	0,471	47,1
ZDA	1.245.000	0,249	24,9
Kitajska	750.000	0,150	15,0
Rusija	550.000	0,110	11,0
Avstralija	100.000	0,020	2,0
Skupaj	5.000.000	1,000	100,0

Grafično (prikaz s krogom):



♦ ♦ ♦

Relativna struktura je lahko tudi **sestavljena**. Poglejmo si takšno možnost spet kar na primeru.

Primer:
Podana je sestavljena tabela, ki prikazuje način bivanja študentov in študentk iz neke občine (opomba: podatek je izmišljen) med študijem.

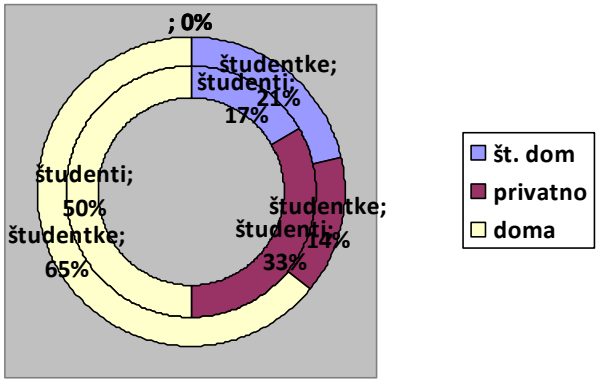
Način bivanja	študenti	študentke	skupaj
Štud. dom	10	15	25
Zasebno	20	10	30

Doma	30	45	75
Skupaj	60	70	130

Tabela strukturnih deležev

Način bivanja	Delež študentov		Delež študentk		Delež skupaj	
	f_k^0	f_k %	f_k^0	f_k %	f_k^0	f_k %
Štud. dom	0,167	16,7	0,214	21,4	0,192	19,2
Zasebno	0,333	33,3	0,143	14,3	0,231	23,1
Doma	0,500	50,0	0,643	64,3	0,577	57,7
Skupaj	1,000	100,0	1,000	100,0	1,000	100,0

Graf:



3.2 Indeksi

Indeksi so posebna oblika relativnih števil.

Ločimo:

- indekse s stalno osnovo,
- indekse s premično osnovo (poseben primer so verižni indeksi)

Zapišimo statistično vrsto takole

$$y_1, y_2, \ldots, y_i, \ldots, y_N$$

Indeksi s stalno osnovo (ta osnova je običajno, ne pa vedno, kar prvi podatek časovne vrste) so števila, določena z izrazom

$$I_{i/1}=\frac{y_i}{y_1}\cdot 100\ ,\ i=1,2,\ldots,N \tag{3.03}$$

V izrazu (3.03) je stalna osnova prvi člen statistične časovne vrste y_1, zato smo indeks označili z znakom $I_{i/1}$. Če bi bila npr. stalna osnova člen y_3, bi indekse zapisali z znakom $I_{i/3}$ in podobno.

Števila

$$I_{i/p}=\frac{I_{i/1}}{I_{p/1}}\cdot 100\ ,i=1,2,\ldots,N \tag{3.04}$$

pa so **indeksi s premično osnovo**.

Poseben primer indeksov s premično osnovo so **verižni indeksi.** Ti so določeni z izrazom

$$I_i=\frac{y_i}{y_{i-1}}\cdot 100\ ,i=2,3,\ldots,N \tag{3.05}$$

Vpeljimo še dva kazalnika, ki sta v neposredni zvezi z indeksi. To sta

a) **koeficient dinamike**

$$K_i=\frac{y_i}{y_{i-1}}\ ,i=2,3,\ldots,N \tag{3.06}$$

b) **stopnja rasti**

$$S_i=I_i-100\ ,i=2,3,\ldots,N \tag{3.07}$$

Izraz (3.07) lahko zapišemo tudi drugače

$$S_i=I_i-100=\frac{y_i}{y_{i-1}}\cdot 100-100=\left(\frac{y_i}{y_{i-1}}-1\right)\cdot 100=(K_i-1)\cdot 100\ ,i=2,3,\ldots,N$$

Torej velja takšna zveza:

$$S_i = (K_i - 1) \cdot 100 \ , i = 2,3,\ldots,N \qquad (3.08)$$

Indekse in oba iz indeksov dobljena kazalnika grafično predstavljamo z linijskimi grafikoni.

Primer:
Vzemimo primer, ki smo ga že omenili, to je knjižnica in število izposojenih knjig v letih od 2011 do 2017.
Izračunajmo indekse s stalno osnovo, verižne indekse, koeficient dinamike in stopnjo rasti.

Rešitev:
Podatki za ta primer, zapisani v tabeli, so takšni

Leto	2011	2012	2013	2014	2015	2016	2017
Št. izposoj. knjig	5500	6300	6250	5700	6160	6000	6730

Dopolnimo to tabelo z iskanimi kazalniki

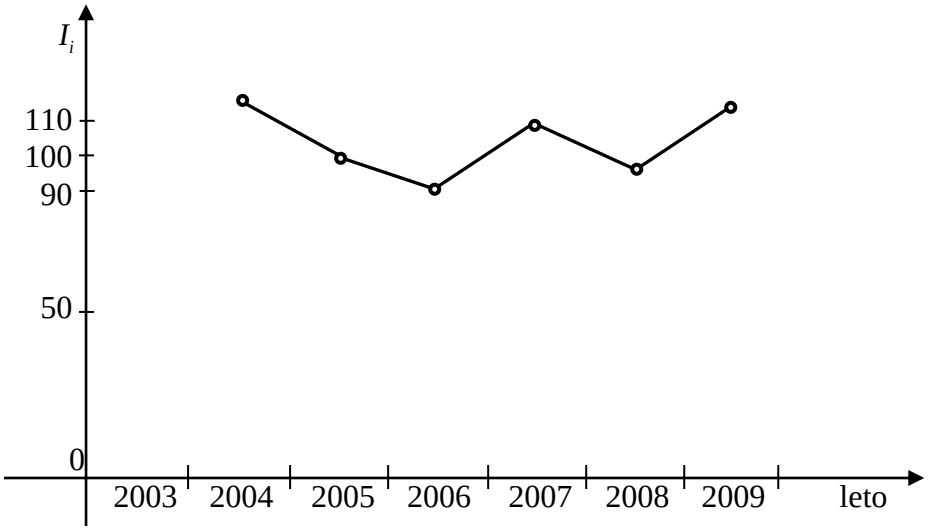
Leto	2011	2012	2013	2014	2015	2016	2017
Št. izposojenih knjig	5500	6300	6250	5700	6160	6000	6730
indeksi s stalno osnovo $I_{i/1}$	100	114,5	113,6	103,6	112,0	109,1	122,4
verižni indeksi I_i	-	114,5	99,2	91,2	108,1	97,4	112,2
koeficient dinamike K_i	-	1,145	0,992	0,912	1,081	0,974	1,122
stopnja rasti S_i	-	14,5	-0,8	-8,8	8,1	-2,6	12,2

Opomba: Tabelo bi lahko tudi obrnili za 90 stopinj in bi jo zapisali v spodnji obliki.

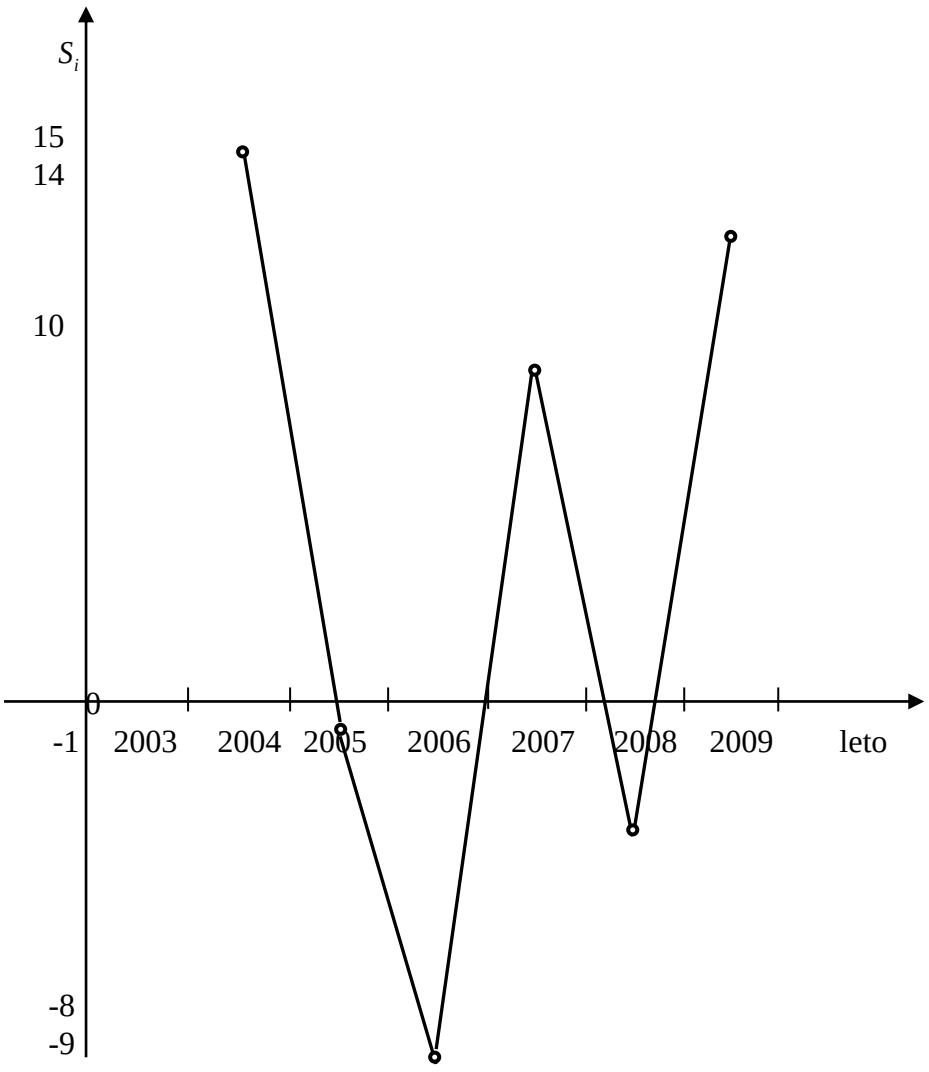
Leto	Št. izposoj. knjig	Indeksi s st. osnovo $I_{i/1}$	Verižni indeksi I_i	Koeficient dinamike K_i	Stopnja rasti S_i
2011	5500	100	-	-	-
2012	6300	114,5	114,5	1,145	14,5
2013	6250	113,6	99,2	0,992	-0,8
2014	5700	103,6	91,2	0,912	-8,8
2015	6160	112,0	108,1	1,081	8,1
2016	6000	109,1	97,4	0,974	-2,6
2017	6730	122,4	112,2	1,122	12,2

Za grafični prikaz indeksov in iz indeksov dobljenih kazalnikov so primerni linijski diagrami.

a) verižni indeks



b) stopnja rasti



♦ ♦ ♦

3.3 Statistični koeficienti

S statističnimi koeficienti primerjamo med seboj razne vrste podatkov, kadar je takšna primerjava seveda vsebinsko smiselna. Opazujemo npr. število študentov na 1000 prebivalcev države (mesta, regije,..), število računalnikov na 100 prebivalcev in podobno. Takšne primerjave opisujemo s statističnimi koeficienti *K*, ki jih izračunamo po formuli

$$K = \frac{X}{Y} \cdot E \tag{3.09}$$

V (3.09) sta *X* in *Y* podatka, ki ju med seboj primerjamo, *E* pa je število, ki pove, na koliko enot izražamo ta koeficient. *E* je potenca števila 10, torej (odvisno od posameznega konkretnega primera) je lahko 1, 10, 100, 1000,…

Primer:
V neki občini je 5 krajevnih skupnosti, za katere poznamo število prebivalcev in število avtomobilov, kar je prikazano v spodnji tabeli.

krajevna skupnost	število prebivalcev	število avtomobilov
KS1	8500	1700
KS2	7315	930
KS3	9450	1500
KS4	3748	1000
KS5	1511	600
skupaj občina	30524	5730

Za vsako posamezno KS in za celotno občino izračunajmo statistični koeficient, ki naj pomeni število avtomobilov na 100 prebivalcev.

Rešitev:
V tem primeru je: *X* – število avtomobilov in *Y* – število prebivalcev.
Ker primerjamo podatka glede na 100 prebivalcev, je *E* = 100.
Od tod dobimo spodnjo tabelo:

krajevna skupnost	število prebivalcev	število avtomobilov	$K\left[\frac{\text{število avtomobilov}}{\text{100 prebivalcev}}\right]$
KS1	8500	1700	20,0
KS2	7315	930	11,3
KS3	9450	1500	15,8
KS4	3748	1000	26,7
KS5	1511	600	39,7
skupaj občina	30524	5730	18,8

♦ ♦ ♦

4 KVANTILI

Položaj posamezne statistične enote ali pa skupine enot v sklopu vseh danih podatkov določamo z več parametri.
Ti statistični parametri so:

- kvantil,
- rang,
- relativni ali kvantilni rang.

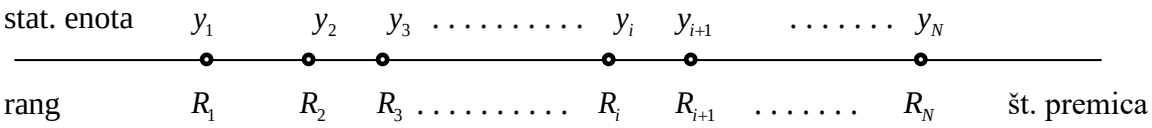
Ko je statističnih enot (podatkov) malo, jih lahko najenostavneje, pregledno in urejeno zapišemo v obliki ranžirne vrste.

Ranžirna vrsta je zapis podatkov, urejenih po velikosti – običajno v naraščajočem zaporedju:

$$y_1, y_2, \ldots, y_N \quad , \quad y_{i+1} \geq y_i \, , \; i = 1,2,3,\ldots,N-1 \tag{4.01}$$

Rang je celoštevilska spremenljivka, ki pove, na katerem mestu v ranžirni vrsti se nahaja statistična enota, ki jo obravnavamo. Rang označimo z znakom R_i, $i = 1,2,\ldots N$.

Grafično si numerične podatke (statistične enote) in njihove range predstavimo na številski premici takole:



Za rang seveda velja:

$$R_i = i \, , \; i = 1,2,\ldots N \tag{4.02}$$

Kvantili so vrednosti numerične spremenljivke y_i, $i = 1,2,\ldots N$ v ranžirni vrsti.

Kvantilu, ki ima vrednost y_i, lahko priredimo rang R_i. Glede na kvantil z vrednostjo y_i, ima v ranžirni vrsti $(i-1)$ statističnih enot manjšo ali enako vrednost od y_i, medtem ko ima $(N-R_i)$ statističnih enot večjo vrednost od y_i.

Relativni ali **kvantilni rang** P_i je število, ki zavzame vrednosti od 0 do 1 in prikazuje ***<u>delež</u>*** statističnih enot, ki so po velikosti manjše od dane statistične enote y_i.

$$0 \le P_i \le 1 \quad i = 1,2,\ldots N$$

Kvantilu vrednosti y_i priredimo relativni (kvantilni) rang $P_i = P(y_i)$. Zanj velja, da je zvezna spremenljivka (lahko zavzame vse vrednosti med 0 in 1) , medtem ko je pripadajoči rang R_i diskretna spremenljivka (zavzame vrednosti 1, 2, 3,…, *N*). Zaradi tega med njima <u>ne velja</u> zveza $P_1 = \frac{R_i}{N}$ oziroma $R_i = P_i N$, pač pa v tem izrazu vedno upoštevamo <u>*numerični popravek*</u> v vrednosti 0,5, kar zapišemo takole

$$R_i = P_i N + 0{,}5 \tag{4.03}$$

Od tod je

$$P_i = \frac{R_i - 0{,}5}{N} \tag{4.04}$$

Nekateri kvantili imajo posebna imena.

Kvartili so kvantili, ki ustrezajo vsaki posamezni četrtini števila podatkov glede na število vseh podatkov. To so vrednosti, ki razdelijo statistično množico na štiri dele, tako da je v vsakem delu ena četrtina (25%) vseh statističnih enot. Kvartili so trije, označimo jih Q_1, Q_2, Q_3. Njihovi zaporedni kvantilni rangi so 0,25, 0,50 in 0,75:

$$P(Q_1) = 0{,}25$$
$$P(Q_2) = 0{,}50$$
$$P(Q_3) = 0{,}75$$

Decili so kvantili, ki ustrezajo vsaki posamezni desetini števila podatkov glede na število vseh podatkov. To so vrednosti, ki razdelijo statistično množico na deset delov, tako da je v vsakem delu ena desetina (10%) vseh statističnih enot. Decilov je devet: $D_1, D_2, \ldots, D_9$. Za njih velja

$$P(D_1) = 0{,}10$$
$$P(D_2) = 0{,}20$$
$$\cdots$$
$$P(D_9) = 0{,}90$$

Centili so kvantili, ki ustrezajo vsaki posamezni stotini števila podatkov glede na število vseh podatkov. To so vrednosti, ki razdelijo statistično množico na sto delov, tako da je v vsakem delu ena stotina (1%) vseh statističnih enot. Centilov je 99. Za njih velja

$$P(C_1)=0{,}01$$
$$P(C_2)=0{,}02$$
$$\dots$$
$$P(C_{99})=0{,}99$$

Seveda velja:

$$P(Q_2)=P(D_5)=P(C_{50})$$

Vrednosti Q_2, D_5 in C_{50} so med seboj enake, imenujemo jih **mediana** ali središčnica. Mediana je tista vrednost statistične enote, od katere ima 50% vseh statističnih enot manjšo ali enako vrednost, 50% pa večjo.
Mediana množico vseh statističnih enot v ranžirni vrsti razdeli na dva dela glede na število statističnih enot in ne na njihovo velikost!

4.1 Računanje kvantilov in kvantilnih rangov

Sedaj bomo določali položaj posamezne statistične enote, tako da bomo

- znanemu kvantilnemu rangu priredili kvantil
  ali pa obratno
- danemu kvantilu določili kvantilni rang.

4.1.1 Iz frekvenčne porazdelitve

Podatki naj bodo razvrščeni v razrede, torej prikazani s frekvenčno porazdelitvijo.
Vzemimo najprej, da *poznamo kvantilni rang P* in želimo določiti ustrezno vrednost spremenljivke *y*, torej *določiti pripadajoči kvantil*.

V tem primeru uporabimo formulo

$$y_k = y_{k,\min} + \frac{R_k - F_{k-1}}{f_k} \cdot i_k \qquad (4.05)$$

V (4.05) je k kvantilni razred, ki ga dobimo iz ocene

$$F_{k-1} < R_k \leq F_k \tag{4.06}$$

Iskani rang dobimo iz formule (4.03), kjer vzamemo $i=k$:

$$R_k = NP_k + 0{,}5 \tag{4.07}$$

Primer:
Za 200 študentov poznamo povprečne ocene opravljenih izpitov. Podatki so že grupirani v razrede in zapisani v spodnji tabeli.

Zap. št.	Povprečna ocena/razred	Frekvenca f_k	Kumulativna frekvenca F_k
1	6,00-6,50	15	15
2	6,51-7,00	40	55
3	7,01-7,50	43	98
4	7,51-8,00	47	145
5	8,01-8,50	33	178
6	8,51-9,00	10	188
7	9,01-9,50	9	197
8	9,51-10,00	3	200
-	skupaj	200	-

Katera povprečna ocena bi bila v ranžirni vrsti natanko v sredini (mediana)?

Rešitev:
Računamo mediano ($R=Me$), njen kvantilni rang je 0,5, torej $P(Me)=0{,}5$. Njen rang dobimo iz relacije $R_k = P_k N + 0{,}5$, torej

$$R_k = NP_k + 0{,}5 = 200 \cdot 0{,}5 + 0{,}5 = 100{,}5$$

Seveda je to podatek, ki leži na 100. mestu, če bi imeli zapis v obliki ranžirne vrste.

Ker imamo podatke podane s frekvenčno porazdelitvijo, moramo dobiti razred, v katerem se nahaja stoti podatek. To dobimo iz relacije (4.06): $F_{k-1} < R \leq F_k$. V tabeli vidimo, da je

$$F_3 < R = 100{,}5 \leq F_4$$

Iskani podatek se nahaja v 4. razredu, kvantilni razred je torej k = 4.

Iz enačbe (4.05) dobimo

$$y = y_{k,\min} + \frac{R - F_{k-1}}{f_k} \cdot i_k = y_{4,\min} + \frac{R - F_3}{f_4} \cdot i_4 = 7{,}51 + \frac{100{,}5 - 98}{47} \cdot 0{,}50 = 7{,}54$$

Povprečna ocena, ki bi bila natanko v sredini vseh podatkov, je 7,54. To pomeni, da ima 50% študentov povprečno oceno manj ali enako 7,54 in 50% več od 7,54.

♦ ♦ ♦

Sedaj pa vzemimo obratno situacijo: *poznamo kvantil y* in želimo *določiti pripadajoči kvantilni rang P.*

V tem primeru ravnamo takole:

- najprej opredelimo kvantilni razred *k* iz ocene

$$y_{k,\min} < y_k \leq y_{k,\max} \tag{4.08}$$

- iz formule (4.05) izračunamo rang

$$R_k = F_{k-1} + \frac{y - y_{k,\min}}{i_k} \cdot f_k \tag{4.09}$$

- iskani kvantilni rang dobimo iz formule $R_k = P_k N + 0{,}5$, torej

$$P_k = \frac{R_k - 0{,}5}{N} \tag{4.10}$$

Primer:
Ponovno vzemimo 200 študentov, za katere poznamo njihove povprečne ocene opravljenih izpitov. Podatki so grupirani v razrede in zapisani v spodnji tabeli.

Zap. št.	Povprečna ocena/razred	Frekvenca f_k	Kumulativna frekvenca F_k
1	6,00-6,50	15	15
2	6,51-7,00	40	55
3	7,01-7,50	43	98
4	7,51-8,00	47	145
5	8,01-8,50	33	178
6	8,51-9,00	10	188
7	9,01-9,50	9	197
8	9,51-10,00	3	200
-	skupaj	200	-

Določimo kvantilni rang študenta, ki ima povprečno oceno 8,90.

Rešitev:
Poznamo kvantil $y_k = 8{,}90$. Kvantilni razred dobimo iz ocene (4.08), to je

$$y_{k,\min} < y \leq y_{k,\max}$$

$$\underset{y_{6,\min}}{8{,}51} < 8{,}90 \leq \underset{y_{6,\max}}{9{,}00} \quad \Rightarrow k = 6$$

Kvantilni razred je 6. razred, ker se ocena 8,90 nahaja v tem razredu.

Uporabimo formulo (5.19)

$$R_k = F_{k-1} + \frac{y - y_{k,\min}}{i_k} \cdot f_k = F_5 + \frac{y - y_{6,\min}}{i_6} \cdot f_6 = 178 + \frac{8{,}90 - 8{,}51}{0{,}50} \cdot 10 = 185{,}8$$

Kvantilni rang, ki pripada rangu $R_6 = 185{,}8$ je po formuli (5.20)

$$P_6 = \frac{R_6 - 0{,}5}{N} = \frac{185{,}8 - 0{,}5}{200} = 0{,}93$$

Komentar: študent s povprečno oceno 8,90 je na 186. mestu, od njega ima slabšo ali enako povprečno oceno 93% študentov, 7% pa ima višjo povprečno oceno.
♦♦♦

5.4.1.2 Iz ranžirne vrste

Kvantile in kvantilne range lahko računamo tudi iz ranžirne vrste. Ranžirno vrsto si lahko predstavljamo kot frekvenčno porazdelitev podatkov, kjer je v vsakem razredu natanko po en element.
Torej glede na zapise pri prejšnji možnosti velja:

$$\begin{aligned} & f_1 = f_2 = \cdots = f_N = 1 \\ & y_{k,\min} = y_{k-1} \\ & i_k = y_k - y_{k-1} \\ & k = 1,2,3,\ldots,N \end{aligned} \tag{4.11}$$

Vzemimo najprej, da *poznamo kvantilni rang P* in želimo določiti ustrezno vrednost spremenljivke *y*, torej *določiti pripadajoči kvantil.*

V formuli (4.05), to je $y_k = y_{k,\min} + \frac{R_k - F_{k-1}}{f_k} \cdot i_k$, ki velja pri frekvenčni porazdelitvi, upoštevamo značilnosti (4.11) in dobimo

$$y_k = y_{k-1} + (R_k - F_{k-1}) \cdot (y_k - y_{k-1}) \tag{4.12}$$

V tej formuli potrebujemo k (kvantilni razred), ki ga dobimo iz ocene

$$F_{k-1} < R_k \leq F_k \tag{4.13}$$

Končno dobimo še iskani rang

$$R_k = NP_k + 0{,}5 \tag{4.14}$$

Primer:
Na zdravniškem pregledu so izmerili višino 20 študentov. Dobljeni rezultati so urejeni v naraščajočem vrstnem redu v ranžirni vrsti in zapisani v spodnji tabeli.

Rang R_i	višina y_i (cm)
1	162
2	163
3	169
4	170
5	170
6	171
7	172
8	173
9	175
10	175
11	177
12	178
13	178
14	180
15	181
16	183
17	183
18	190
19	197
20	199

Izračunajmo 3. kvartil Q_3.

Rešitev:
Tretji kvartil ima kvantilni rang 0,75, torej $P_3 = P(Q_3) = 0{,}75$.

Iz formule (4.14) izračunamo rang: $R_k = NP_k + 0{,}5 = 20 \cdot 0{,}75 + 0{,}5 = 15{,}5$.

Vrednost za *k* dobimo iz ocene (4.13)

$$F_{k-1} < R_k \leq F_k$$

Ker je $15 < 15{,}5 < 16$, sledi od tod: $k = 16$.

Iz formule (4.12) dobimo še iskani kvartil

$$y_k = y_{k-1} + (R_k - F_{k-1}) \cdot (y_k - y_{k-1})$$

$$y_{16} = Q_3 = y_{15} + (R_{16} - F_{15})(y_{16} - y_{15}) = 181 + (15{,}5 - 15)(183 - 181) = 182$$

Tretji kvartil je $Q_3 = 182$ cm. To pomeni, da je 75% statističnih enot (študentov) manjših ali enakih 182 cm in 25% večjih od 182 cm.
♦ ♦ ♦

Sedaj pa vzemimo, da *poznamo kvantil y* in želimo *določiti pripadajoči kvantilni rang P*.

Tudi v tem primeru uporabimo formule za ta postopek iz primera frekvenčne porazdelitve, pri čemer upoštevamo posebnosti, zapisane v (4.11).

Postopek poteka takole:

- najprej opredelimo kvantilni razred *k* iz ocene $y_{k,\min} < y_k \leq y_{k,\max}$, kar sedaj pomeni

$$y_{k-1} < y \leq y_k$$

- iz formule (4.12), to je $y_k = y_{k-1} + (R_k - F_{k-1}) \cdot (y_k - y_{k-1})$, izračunamo kvantilni rang R_k

$$R_k = F_{k-1} + \frac{y - y_{k-1}}{x_k - x_{k-1}} \qquad (4.15)$$

- končno dobimo iskani kvantilni rang iz formule (4.14), torej je

$$P_k = \frac{R_k - 0{,}5}{N} \tag{4.16}$$

Primer:
Vzemimo iste podatke, kot v prejšnjem primeru. Kakšen kvantilni rang ustreza vrednosti $y = 174$ cm?

Rešitev:
Tabela s podatki je zapisana v prejšnjem primeru.

V skladu s postopkom določimo najprej kvantilni razred *k*:

$$y_{k-1} < y \leq y_k$$

$$y_8 = 173 < 174 \leq y_9 = 174 \quad \Rightarrow k = 9$$

Nato iz formule (4.09) izračunamo kvantilni rang

$$R_k = F_{k-1} + \frac{y - y_{k-1}}{x_k - x_{k-1}}$$

$$R_9 = F_8 + \frac{y - y_8}{x_9 - x_8} = 8 + \frac{174 - 173}{185 - 173} = 8{,}5$$

Od tod pa iz formule (4.10) dobimo še iskani kvantilni rang

$$P_k = \frac{R_k - 0{,}5}{N}$$

$$P_9 = \frac{R_9 - 0{,}5}{N} = \frac{8{,}5 - 0{,}5}{20} = 0{,}40$$

Komentar:
Od velikosti 174 cm je manjših ali enakih 40% podatkov, 60% pa je večjih.
♦ ♦ ♦

4.2 Grafično določanje kvantilov in kvantilnih rangov

Kvantile in kvantilne range, ki jim pripadajo, lahko določimo tudi grafično. Dobimo jih tako, da jih enostavno preberemo iz grafa kumulativnih frekvenc.
Poglejmo si takšno možnost kar na primeru.

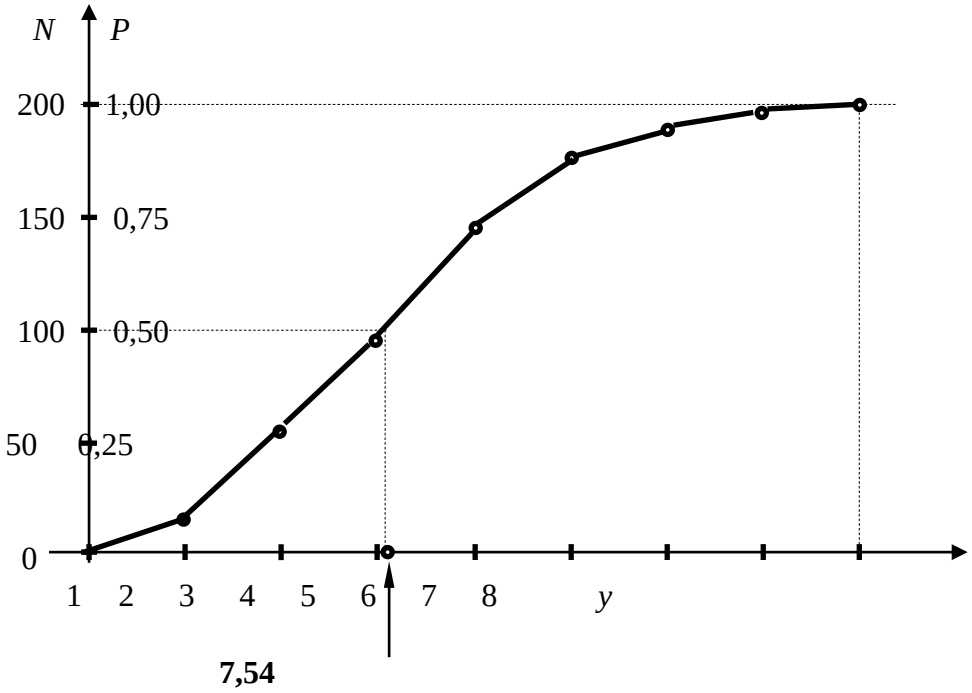
Primer:
Vzemimo že znan primer z 200 študenti in njihovimi povprečnimi ocenami.

Zap. št.	Povprečna ocena/razred	Frekvenca f_k	Kumulativna frekvenca F_k
1	6,00-6,50	15	15
2	6,51-7,00	40	55
3	7,01-7,50	43	98
4	7,51-8,00	47	145
5	8,01-8,50	33	178
6	8,51-9,00	10	188
7	9,01-9,50	9	197
8	9,51-10,00	3	200
-	skupaj	200	-

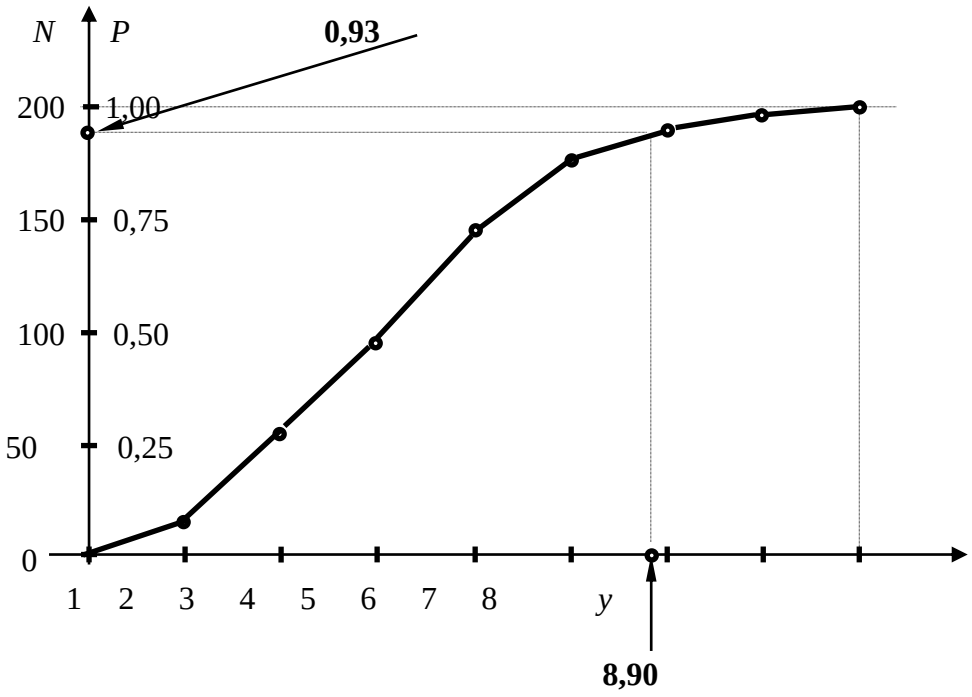
V prejšnjih primerih smo že ugotovili tole:
- mediana, ki ima kvantilni rang $P(Me)=0{,}5$, je podatek $Me=7{,}54$,
- kvantilu $y=8{,}90$ pripada kvantilni rang $P=0{,}93$.

Grafični prikaz kumulativnih frekvenc za ta primer smo tudi že narisali. V tem grafu odmerimo na ordinatni osi še kvantilni rang.

1. če poznamo kvantilni rang in iščemo pripadajoči kvantil, bomo na ordinatni osi (os kumulativnih frekvenc) odmerili kvantilni rang in grafično ugotovili, katera vrednost na abscisni osi (= kvantil) ji ustreza. Mediana, ki ima kvantilni rang $P(Me)=0{,}5$, je podatek $Me=7{,}54$, kar preberemo na abscisni osi.



2. če poznamo kvantil, dobimo kvantilni rang na ordinatni osi, za že znani primer je kvantil 8,90 in na ordinatni osi preberemo pripadajoči kvantilni rang 0,93.



♦ ♦ ♦

5 SREDNJE VREDNOSTI

Srednje vrednosti so statistični parametri, ki jih v statističnih izračunih velikokrat uporabljamo. Srednjih vrednosti je več, omenili bomo naslednje:

- mediana,
- modus,
- aritmetična sredina,
- harmonična sredina,
- geometrična sredina.

5.1 Mediana (Me)

Mediana je vrednost statistične spremenljivke, ki ji pripada kvantilni rang $P=0{,}50$. To pomeni, da je to tista vrednost, od katere je polovica statističnih enot manjših ali enakih, polovica statističnih enot pa je od nje večjih.
Mediano smo natanko spoznali v prejšnjem poglavju. Izračunamo jo iz ranžirne vrste ali iz frekvenčne porazdelitve na način, ki ga že poznamo.
Mediana je lahko razumljiva in predstavljiva ter enostavna za računanje.
Njena pomanjkljivost je zlasti v tem, da je premalo občutljiva za spremembe vrednosti podatkov.

5.2 Modus (Mo)

Modus je naslednja srednja vrednost. To je takšna vrednost, okoli katere se število podatkov najbolj zgosti. Modus je torej najpogostejša vrednost. Iz te definicije že vidimo, da modusa ne moremo ugotavljati (računati) v primeru, ko so podatki dani v ranžirni vrsti, pač pa le pri frekvenčni porazdelitvi, *kjer je širina razredov enaka*. Modus je v tistem razredu, kjer je frekvenca največja, tam torej, kjer je največ podatkov. Razred, v katerem najdemo modus, imenujemo *modusni razred.*

Modus izračunamo po formuli

$$Mo = y_{k,\min} + \frac{f_k - f_{k-1}}{(f_k - f_{k-1}) + (f_k - f_{k+1})} \cdot i_k$$

oziroma

$$\boxed{Mo = y_{k,\min} + \frac{f_k - f_{k-1}}{2f_k - f_{k-1} - f_{k+1}} \cdot i_k} \qquad (5.01)$$

Znaki v formuli (5.01) pomenijo:

k – modusni razred, ki ga dobimo iz pogoja

$$f_k = \max_i (f_i)\ ,\ i = 1,2,\ldots,r \tag{5.02}$$

f_k - frekvenca modusnega razreda,

$y_{k,\min}$ - spodnja meja modusnega razreda,

i_k - širina modusnega razreda.

Opomba: Če je frekvenčna porazdelitev polimodalna, ima več modusov!

Primer:
Vzemimo spet 200 študentov, za katere poznamo povprečno oceno opravljenih izpitov, podatki so dani v tabeli.
Določimo modus.

Rešitev:
Iz tabele frekvenčne porazdelitve vidimo, da je modusni razred četrti, ker zadošča pogoju (5.02).

$$\max_i (f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8) = f_4 = 47 \Rightarrow k = 4$$

Zap. št.	Povprečna ocena/razred	Frekvenca f_k	Kumulativna frekvenca F_k
1	6,00-6,50	15	15
2	6,51-7,00	40	55
3	7,01-7,50	43	98
4	7,51-8,00	47	145
5	8,01-8,50	33	178
6	8,51-9,00	10	188
7	9,01-9,50	9	197
8	9,51-10,00	3	200
-	skupaj	200	-

Modus dobimo po formuli (5.01)

$$Mo = y_{k,\min} + \frac{f_k - f_{k-1}}{2f_k - f_{k-1} - f_{k+1}} \cdot i_k = 7{,}51 + \frac{47-43}{94-43-33} \cdot 0{,}50 = 7{,}62$$

To pomeni, da ima največ študentov srednjo oceno okrog 7,62.
♦ ♦ ♦

5.3 Aritmetična sredina (M, $\bar{y}$)

Aritmetična sredina ali aritmetično povprečje je najpogosteje uporabljana srednja vrednost. Izračunamo jo po formuli

$$M_y = \bar{y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (5.03)$$

Opomba: Za aritmetično sredino uporabljamo znak M_y ali pa $\bar{y}$. Če spremenljivko označimo npr. s črko *x*, potem je znak za aritmetično sredino seveda M_x ali $\bar{x}$ in podobno za kakšno drugačno označevanje.

Primer:
Na zdravniškem pregledu so izmerili višino 20 študentov. Dobljeni rezultati so urejeni v naraščajočem vrstnem redu v ranžirni vrsti in zapisani v spodnji tabeli. Kolika je povprečna višina teh študentov?

zap. št.	višina y_i (cm)
1	162
2	163
3	169
4	170
5	170
6	171
7	172
8	173
9	175
10	175
11	177
12	178
13	178
14	180
15	181
16	183
17	183
18	190
19	197
20	199
skupaj	3547

Po (5.03) dobimo povprečno višino takole:

$$M_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{20}\sum_{i=1}^{20} y_i = \frac{1}{20}\cdot 3547 = 177{,}3$$

Povprečna višina študentov je 177,3 cm.

♦♦♦

5.3.1 Lastnosti aritmetične sredine

1. *Aritmetična sredina konstante je konstanta*

$$\boxed{M_C = C \quad , \; C - \text{konstanta}} \tag{5.04}$$

Relacijo (5.04) dokažemo tako, da upoštevamo definicijo aritmetične sredine. Vsi členi ranžirne vrste so v tem primeru namreč enaki *C*, zato je po definiciji (5.03) res tole:

$$M_y = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\left(\underbrace{C + C + \cdots + C}_{N-krat}\right) = \frac{1}{N}\cdot NC = C$$

2. *Aritmetična sredina produkta spremenljivke s konstanto je enaka produktu konstante z aritmetično sredino*

$$\boxed{y = \lambda x \Rightarrow M_y = \lambda M_x} \tag{5.05}$$

Tudi to lastnost dobimo iz definicije aritmetične sredine. Vzemimo, da imamo spremenljivko *x,* za katero poznamo arirmetično sredino M_x. Spremenljivka *y* naj bo s konstanto λ pomnožena spremenljivka *x*, torej $y_i = \lambda x_i$, $i = 1, 2, \ldots, N$. Sedaj je:

$$M_y = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} \lambda x_i = \frac{1}{N}\cdot\lambda\cdot\sum_{i=1}^{N} x_i = \lambda\cdot\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \lambda M_x$$

3. *Za aritmetično sredino velja lastnost linearnosti.*

Naj bo *z* statistična spremenljivka, dana kot linearna kombinacija dveh drugih statističnih spremenljivk *x* in *y*, torej

$$z = \lambda_1 x + \lambda_2 y$$

Potem je res

$$M_z = \lambda_1 M_x + \lambda_2 M_y \tag{5.06}$$

Relacijo (5.06) spet dobimo iz definicije povprečne vrednosti in lastnosti vsote takole:

$$M_z = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_1 x_i + \lambda_2 y_i\right) = \lambda_1\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \lambda_2\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = \lambda_1 M_x + \lambda_2 M_y$$

4. *Vsota vseh odklonov (odmikov, razlik) posameznih vrednosti od aritmetične sredine je enaka 0:*

$$\sum_{i=1}^{N}\left(y_i - M_y\right) = 0 \tag{5.07}$$

Zgornjo lastnost ugotovimo takole:

$$\sum_{i=1}^{N}\left(y_i - M_y\right) = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} M_y = NM_y - NM_y = 0$$

Pri tem smo upoštevali, da je $\sum_{i=1}^{N} M_y = \underbrace{M_y + M_y + \cdots + M_y}_{N-krat} = NM_y$

Lastnost (5.07) še prav posebej poudarja pomen aritmetične sredine.

5. *Vsota kvadratov odklonov posameznih vrednosti od neke konstante je najmanjša, če je ta konstanta enaka aritmetični sredini.*

$$\sum_{i=1}^{N}\left(y_i - \lambda\right)^2 \text{ min, če je } \lambda = M_x \tag{5.08}$$

Dokažimo še to.

Označimo $Y = \sum_{i=1}^{N}\left(y_i - \lambda\right)^2$. Funkcija $Y = f(\lambda)$ ima ekstrem v stacionarni točki, torej mora biti njen prvi odvod nič.

$$Y' = \frac{dY}{d\lambda} = 2\sum_{i=1}^{N}\left(y_i - \lambda\right)(-1) = 0$$

Od tod je

$$\sum_{i=1}^{N}(y_i-\lambda)=0 \text{ oziroma } \sum_{i=1}^{N}y_i-\sum_{i=1}^{N}\lambda=0$$

Ker je $\sum_{i=1}^{N}\lambda=\underbrace{\lambda+\lambda+\cdots+\lambda}_{N-krat}=N\lambda$

iz enačbe $\sum_{i=1}^{N}y_i-\sum_{i=1}^{N}\lambda=0$ sledi $\sum_{i=1}^{N}y_i=N\lambda$

in od tod res vidimo, da je to število ravno aritmetična sredina:

$$\lambda=\frac{1}{N}\sum_{i=1}^{N}y_i=M_y$$

Z drugim odvodom pa se moramo še prepričati, če gre zares za minimum.
V ta namen zapišimo prvi odvod takole:

$$Y'=\frac{dY}{d\lambda}=-2\sum_{i=1}^{N}(y_i-\lambda)=-2\left[(y_1-\lambda)+(y_2-\lambda)+\cdots+(y_N-\lambda)\right]$$

Drugi odvod je potem

$$Y''=\frac{d^2Y}{d\lambda^2}=-2\left[\underbrace{(-1)+(-1)+\cdots+(-1)}_{N-krat}\right]=2N>0$$

Drugi odvod je pozitiven za vsako vrednost λ, zato je ekstrem res minimum.

6. *Aritmetično sredino statistične množice z N statističnimi enotami je mogoče izračunati iz aritmetičnih sredin r delnih statističnih množic, ki imajo skupaj N enot:*

$$M_y=\overline{y}=\frac{N_1\overline{y}_1+N_2\overline{y}_2+\cdots+N_r\overline{y}_r}{N_1+N_2+\cdots+N_r}=\frac{\sum_{i=1}^{r}N_i\overline{y}_i}{\sum_{i=1}^{r}N_i}$$

Ker je $\sum_{i=1}^{r}N_i=N_1+N_2+\cdots+N_r=N$, dobimo iz zgornjega zapisa

$$\boxed{M_y=\overline{y}=\frac{1}{N}\sum_{i=1}^{r}N_i\overline{y}_i} \qquad (5.09)$$

Aritmetični sredini (5.09) pravimo **tehtana aritmetična sredina** (ali tudi **ponderirana aritmetična sredina**). Števila N_i so **uteži** ali **ponderji**.

Primer:
V podjetju ABC, d.o.o. so imeli v letu 2019 po delovnih urah naslednje število zaposlenih: 5 mesecev 10 delavcev, 3 mesece 8 delavcev in 4 mesece 9 delavcev. Koliko zaposlenih po delovnih urah so imeli v povprečju leta 2019?

Rešitev:
Če vzamemo, da so podatki o zaposlenih kar po vrsti po mesecih, lahko zapišemo tabelo

													skupaj
mesec	jan	feb	mar	apr	maj	jun	jul	avg	sep	okt	nov	dec	12
zaposleni	10	10	10	10	10	8	8	8	9	9	9	9	110

V 12 mesecih so imeli po delovnih urah skupno 110 zaposlenih, v povprečju mesečno torej $\frac{110}{12}=9{,}17$. Gre za tipičen primer tehtane aritmetične sredine. Če uporabimo formulo (5.09), dobimo seveda isti rezultat:

$$M_y=\overline{y}=\frac{1}{N}\sum_{i=1}^{r}N_i\overline{y}_i=\frac{1}{12}\sum_{i=1}^{3}N_i\overline{y}_i=\frac{1}{12}\left(N_1\overline{y}_1+N_2\overline{y}_2+N_3\overline{y}_3\right)=\frac{1}{12}\left(5\cdot10+3\cdot8+4\cdot9\right)=9{,}17$$

♦♦♦

5.3.2 Aritmetična sredina pri frekvenčni porazdelitvi

Ko so podatki dani s frekvenčno porazdelitvijo, aritmetičnega povprečja ne moremo izračunati natanko, dobimo lahko le njegov **približek** z uporabo formule

$$M_y=\overline{y}=\frac{1}{N}\sum_{i=1}^{r}f_iy_i \qquad (5.10)$$

Tu je
r- število razredov
y_i- sredina i-tega razreda, $i=1,2,\ldots,r$
f_i - frekvenca i-tega razreda, $i=1,2,\ldots,r$

Zgornja formula je res le približek (ocena) za aritmetično povprečje, ker smo privzeli, da imajo vsi predstavniki istega razreda enako vrednost, to je sredina tega razreda.

Primer:
Vzemimo 200 študentov in njihove povprečne ocene opravljenih izpitov. Izračunajmo povprečno vrednost statistične spremenljivke (ocene opravljenih izpitov).

Zap. št.	Povprečna ocena/razred	Sredina razreda y_i	Frekvenca f_i	Produkt $f_i y_i$
1	6,00-6,50	6,25	15	93,75
2	6,51-7,00	6,75	40	270,00
3	7,01-7,50	7,25	43	311,75
4	7,51-8,00	7,75	47	364,25
5	8,01-8,50	8,25	33	272,25
6	8,51-9,00	8,75	10	87,50
7	9,01-9,50	9,25	9	83,25
8	9,51-10,00	9,75	3	29,25
-	skupaj	.	200	1512,00

Rešitev:
Povprečno vrednost statistične spremenljivke, ki je podana s frekvenčno porazdelitvijo, bomo dobili z uporabo (5.10).

$$M_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{r} f_i y_i = \frac{1}{200}\sum_{i=1}^{8} f_i y_i = \frac{1}{200}\big(15\cdot 6,25 + 40\cdot 6.75 + 43\cdot 7,25 + 47\cdot 7,75 +$$

$$+ 33\cdot 8,25 + 10\cdot 8,75 + 9\cdot 9,25 + 3\cdot 9,75\big) = \frac{1512,00}{200} = 7,56$$

♦♦♦

5.4 Harmonična sredina

Harmonična sredina za *N* vrednosti statistične spremenljivke *y* je določena s formulo

$$H = \frac{N}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}} = \frac{N}{\sum_{i=1}^{N} \frac{1}{y_i}} \qquad (5.11)$$

Formula (5.11) je uporabna tedaj, ko so podatki dani v ranžirni vrsti. Vrednosti statistične spremenljivke so lahko tudi frekvenčno porazdeljene, torej se pojavljajo v *r* razredih. V

tem primeru imamo opravka s **tehtano** (ponderirano) harmonično sredino, ki je definirana takole

$$H = \frac{f_1 + f_2 + \cdots + f_r}{\frac{f_1}{y_1} + \frac{f_2}{y_2} + \cdots + \frac{f_r}{y_r}} = \frac{\sum_{i=1}^{r} f_i}{\sum_{i=1}^{r} \frac{f_i}{y_i}}$$

Ker je izraz v števcu $\sum_{i=1}^{r} f_i = f_1 + f_2 + \cdots + f_r = N$, zapišemo tehtano harmonično sredino s formulo

$$H = \frac{f_1 + f_2 + \cdots + f_r}{\frac{f_1}{y_1} + \frac{f_2}{y_2} + \cdots + \frac{f_r}{y_r}} = \frac{N}{\sum_{i=1}^{r} \frac{f_i}{y_i}} \qquad (5.12)$$

Harmonično sredino praviloma uporabljamo v primerih, ko se vrednosti statistične spremenljivke izražajo v obliki kvocienta oziroma ko imamo opraviti z *obratno sorazmernimi količinami.*

Primer:
Trije študenti rešujejo vsak zase isto nalogo. Prvi študent jo reši v 5 minutah, drugi študent jo reši v 6 minutah, tretji študent pa jo reši v 10 minutah. Kakšen je povprečni čas reševanja te naloge?

Rešitev:
Če bi skušali dobiti povprečni čas z aritmetično sredino, bi dobili rezultat

$$\bar{y} = \frac{y_1 + y_2 + y_3}{3} = \frac{5+6+10}{3} = 7$$

Ali je to pravilen rezultat?
Poglejmo nalogo z druge plati in prevedimo delo študentov na skupni čas 1 uro. To pomeni naslednje:

Prvi študent reši nalogo v 5 minutah, v eni uri bi rešil 12 takšnih nalog.
Drugi študent reši nalogo v 6 minutah, v eni uri bi rešil 10 takšnih nalog.
Tretji študent reši nalogo v 10 minutah, v eni uri bi rešil 6 takšnih nalog.

Skupno bi ti trije študenti v treh urah (=180 minut) rešili 28 takšnih nalog. Od tod dobimo pravilno povprečje za reševanje 1 naloge: $\bar{y} = \frac{180 \text{ minut}}{28} = 6,43 \text{ minut}$.

Rezultat, ki smo ga dobili z aritmetičnim povprečjem, je torej napačen.

Ker sta pojma »znanje matematike« in »čas reševanja naloge« obratno sorazmerna (več kot zna matematike, manj časa potrebuje za reševanje naloge), moramo uporabiti harmonično sredino in s tem dobimo pravilen rezultat:

$$H = \frac{N}{\frac{1}{y_1}+\frac{1}{y_2}+\frac{1}{y_3}} = \frac{3}{\frac{1}{5}+\frac{1}{6}+\frac{1}{10}} = 6,43$$

♦♦♦

5.5 Geometrična sredina

Geometrično (ali tudi: geometrijsko) sredino uporabljamo pri analizi časovnih vrst, ko proučujemo dinamične kazalnike, kot so verižni indeksi, koeficient dinamike in stopnja rasti.
Geometrijsko sredino podatkov izračunamo po formuli

$$G = \sqrt[N]{y_1 \cdot y_2 \cdots y_N} \tag{5.13}$$

Če so podatki grupirani po razredih, računamo **tehtano** (ponderirano) geometrijsko sredino, ki jo dobimo iz enačbe

$$G = \sqrt[N]{y_1^{N_1} \cdot y_2^{N_2} \cdots y_r^{N_r}} \tag{5.14}$$

Če formulo za geometrijsko sredino, to je (5.13), logaritmiramo, imamo

$$\log G = \log \sqrt[N]{y_1 \cdot y_2 \cdots y_N}$$

$$\log G = \frac{1}{N}\left(\log y_1 + \log y_2 + \cdots + \log y_N\right) = \frac{1}{N}\sum_{i=1}^{N} \log y_i$$

Iz tega zapisa ugotovimo zanimivo lastnost: *logaritem geometrijske sredine je aritmetična sredina logaritmov vrednosti statistične spremenljivke.*

Podobno bi ugotovili, če bi logaritmirali tehtano geometrijsko sredino:

$$\log G = \log \sqrt[N]{y_1^{N_1} \cdot y_2^{N_2} \cdots y_r^{N_r}}$$

$$\log G = \frac{1}{N}\left(N_1 \log y_1 + N_2 \log y_2 + \cdots + N_r \log y_r\right) = \frac{1}{N}\sum_{i=1}^{r} N_i \log y_i$$

Logaritem tehtane geometrijske sredine je tehtana aritmetična sredina logaritmov vrednosti grupiranih podatkov.

Primer:
Cena nekega izdelka se poveča za 100%, nato se poceni za 50%. Kolikšna je končna sprememba cene?

Rešitev:
Označimo osnovno ceno c_0. Cena, povečana za 100% , naj bo c_1, cena, kasneje znižana za 50%, pa naj bo c_2. Imamo torej podatke, za katere lahko izračunamo verižne indekse.

Cena	c_0	$c_1 = 2c_0$ (za 100% povečana cena c)	$c_2 = \frac{1}{2}c_1$ (za 50% zmanjšana cena c_1)
Verižni indeks	-	$I_1 = \frac{c_1}{c_0} \cdot 100 = 200$	$I_2 = \frac{c_2}{c_1} \cdot 100 = 50$

Jasno je naslednje: če se cena poveča za 100% in se ta cena nato zniža za 50%, je končna cena enaka prvotni ceni. Povprečni verižni indeks bo torej 100.

Izračunajmo povprečje z vsemi tremi znanimi sredinami:

1. aritmetična sredina: $M = \frac{I_1 + I_2}{2} = \frac{200+50}{2} = 125$ NAPAČNO
2. harmonična sredina: $H = \frac{2}{\frac{1}{I_1} + \frac{1}{I_2}} = \frac{2I_1 I_2}{I_1 + I_2} = \frac{2 \cdot 200 \cdot 50}{250} = 80$ NAPAČNO
3. geometrijska sredina: $G = \sqrt[2]{I_1 \cdot I_2} = \sqrt{200 \cdot 50} = \sqrt{10000} = 100$ PRAVILNO

Iz zgornjih izračunov vidimo, da dobimo pravilen rezultat z geometrijsko sredino.
♦ ♦ ♦

Geometrijsko sredino uporabljamo tudi za računanje koeficienta dinamike in stopnje rasti v časovnih vrstah.

Označimo člene v časovni vrsti

$$y_1, y_2, y_3, \ldots, y_N$$

Posamezni koeficienti dinamike so

$$K_2 = \frac{y_2}{y_1},\ K_3 = \frac{y_3}{y_2}, \ldots,\ K_N = \frac{y_N}{y_{N-1}}$$

Pomnožimo zgornje koeficiente med seboj in ta produkt pomnožimo še z y_1, pa dobimo

$$y_1 \cdot K_2 \cdot K_3 \ldots K_N = y_1 \cdot \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \frac{y_4}{y_3} \cdot \ldots \cdot \frac{y_N}{y_{N-1}} = y_N$$

Od tod

$$\frac{y_N}{y_1} = K_2 \cdot K_3 \ldots K_N$$

Koeficienti dinamike so kazalniki, ki povedo, kako pridemo od začetne vrednosti statistične spremenljivke y_1 do končne vrednosti statistične spremenljivke y_N.

Če je *K* povprečni koeficient dinamike, lahko zgornjo ugotovitev napišemo v obliki

$$\frac{y_N}{y_1} = \underbrace{K \cdot K \ldots K}_{(N-1)\text{-krat}} = K^{N-1}$$

Iz primerjave zgornjih izrazov sledi

$$\frac{y_N}{y_1} = K_2 \cdot K_3 \ldots K_N = K^{N-1}$$

Od tod dobimo dve uporabni formuli

$$K = \sqrt[(N-1)]{K_2 \cdot K_3 \ldots K_N} \tag{5.15}$$

in

$$K = \sqrt[(N-1)]{\frac{y_N}{y_1}} \tag{5.16}$$

Povprečni koeficient dinamike je torej geometrijska sredina posameznih koeficientov dinamike.

Primer:
Vzemimo že znani primer izposojenih knjig v knjižnici v letih 2011 do 2017 (tabela). Izračunajmo povprečni koeficient dinamike in povprečno stopnjo rasti.

Leto	2011	2012	2013	2014	2015	2016	2017
Št. izposojenih knjig	5500	6300	6250	5700	6160	6000	6730
indeksi s stalno osnovo $I_{i/1}$	100	114,5	113,6	103,6	112,0	109,1	122,4
verižni indeksi I_i	-	114,5	99,2	91,2	108,1	97,4	112,2
koeficient dinamike K_i	-	1,145	0,992	0,912	1,081	0,974	1,122
stopnja rasti S_i	-	14,5	-0,8	-8,8	8,1	-2,6	12,2

Rešitev:

a) Uporabimo formulo (5.13) in izračunamo povprečni verižni indeks:

$$\bar{I} = \sqrt[6]{I_2 \cdot I_3 \cdot I_4 \cdot I_5 \cdot I_6 \cdot I_7} = \sqrt[6]{114,5 \cdot 99,2 \cdot 91,2 \cdot 108,1 \cdot 97,4 \cdot 112,2} = 103,4$$

b) Uporabimo formulo (5.15) in izračunamo povprečni koeficient dinamike:

$$K = \sqrt[6]{K_2 \cdot K_3 \cdot K_4 \cdot K_5 \cdot K_6 \cdot K_7} = \sqrt[6]{1,145 \cdot 0,992 \cdot 0,912 \cdot 1,081 \cdot 0,974 \cdot 1,122} = \sqrt[6]{1,223741} = 1,034$$

c) Povprečno stopnjo rasti dobimo iz povprečnega koeficienta dinamike iz enačbe $S = (K - 1)100$, kar za ta primer pomeni: $S = (K - 1)100 = (1,0342 - 1)100 = 3,4$.

♦ ♦ ♦

6 MERE VARIABILNOSTI, MERE ASIMETRIJE, MERE SPLOŠČENOSTI

V tem poglavju bomo spoznali tri statistične parametre, s katerimi ugotavljamo posebnosti porazdelitve vrednosti statističnih spremenljivk. To so: mere variabilnosti, mere asimetrije in mere sploščenosti.

6.1 Mere variabilnosti

Množični podatki, s katerimi imamo opravka v statistiki, dosegajo različne vrednosti, rekli bomo, da varirajo. Z merami variabilnosti ugotavljamo, v kakšnih mejah se gibljejo numerične vrednosti pridobljenih podatkov.

6.1.1 Variacijski razmik

Variacijski razmik R_V je najenostavnejša mera variabilnosti in pomeni zgolj razliko med največjo in najmanjšo vrednostjo statistične spremenljivke:

$$R_V = y_{\max} - y_{\min} \tag{6.01}$$

Variacijski razmik je precej groba mera in načeloma ne nudi dovolj informacij o proučevanem pojavu.

6.1.2 Kvartilni razmik

Kvartilni razmik R_Q je razlika med tretjim in prvim kvartilom:

$$R_Q = Q_3 - Q_1 \tag{6.02}$$

5.6.1.3 Kvartilni odklon

Namesto kvartilnega razmika večkrat uporabljamo kvartilni odklon, ki je polovica kvartilnega razmika

$$R_{\frac{Q}{2}} = \frac{1}{2}(Q_3 - Q_1) \tag{6.03}$$

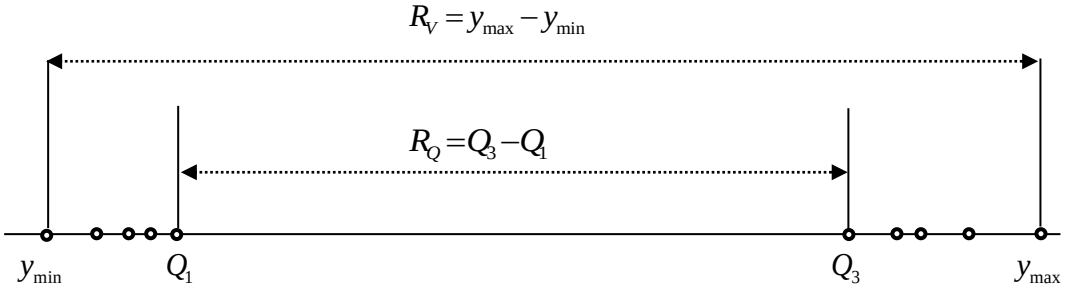
6.1.4 Decilni razmik

Decilni razmik je razlika med devetim in prvim decilom

$$R_D = D_9 - D_1 \tag{6.04}$$

Z decilnim razmikom smo izločili 10% po vrednosti najmanjših in 10% po vrednosti največjih vrednosti statistične spremenljivke.

Prikažimo variacijski in kvartilni razmik grafično.



6.1.5 Povprečni absolutni odklon od aritmetične sredine

Naslednja mera variabilnosti je povprečni absolutni odklon od aritmetične sredine, ki je določen z izrazom

$$d_M = \frac{1}{N}\sum_{i=1}^{N}|y_i - M| \tag{6.05}$$

6.1.6 Povprečni absolutni odklon od mediane

Podobno vpeljemo še povprečni absolutni odklon od mediane

$$d_{Me} = \frac{1}{N}\sum_{i=1}^{N}|y_i - Me| \tag{6.06}$$

Primer:
Vzemimo že večkrat uporabljeni primer meritev velikosti 20 študentov (tabela). Zapišimo omenjene mere variabilnosti.

zap. št.	višina y_i (cm)
1	162
2	163
3	169
4	170
5	170
6	171
7	172
8	173
9	175
10	175
11	177
12	178
13	178
14	180
15	181
16	183
17	183
18	190
19	197
20	199
skupaj	3547

Rešitev:

a) Variacijski razmik: $R_V = y_{\max} - y_{\min} = 199 - 162 = 37$ cm.

b) Za kvartilni razmik $R_Q = Q_3 - Q_1$ potrebujemo prvi in tretji kvartil. Tretji kvartil smo že izračunali in ga poznamo: $Q_3 = 182$ cm. Izračunajmo še prvi kvartil Q_1.

$$P(Q_1) = 0{,}25 \Rightarrow R = NP + 0{,}5 = 20 \cdot 0{,}25 + 0{,}5 = 5{,}5$$

Od tod določimo $k = 6$ in nato

$$Q_1 = y_5 + (R_6 - F_5)(y_6 - y_5) = 170 + (5{,}5 - 5)(171 - 170) = 170{,}5\,\text{cm}$$

Kvartilni razmik je torej: $R_Q = Q_3 - Q_1 = 182 - 170{,}5 = 11{,}5$ cm.

c) Kvartilni odklon: $R_{\frac{Q}{2}} = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2} \cdot 11{,}5\,\text{cm} = 5{,}75$ cm

d) V enem od prejšnjih primerov smo že izračunali aritmetično sredino: $M = 177{,}3$ cm, zato bomo nekoliko kasneje lahko izračunali tudi povprečni absolutni odklon od aritmetične sredine.

e) Če želimo izračunati še povprečni absolutni odklon od mediane, moramo najprej določiti mediano.

Zanjo velja $P(Me)=0{,}5$, zato je $R=NP+0{,}5=20\cdot 0{,}5+0{,}5=10{,}5\Rightarrow k=11$.

Od tod sledi: $Me=y_{10}+(R_{11}-F_{10})(y_{11}-y_{10})=175+(10{,}5-10)(177-175)=176$ cm.

Povprečni absolutni odklon od aritmetične sredine in povprečni absolutni odklon od mediane dobimo iz (6.05) in (6.06).

Računanje si bomo olajšali tako, da bomo dopolnili tabelo podatkov z dvema stolpcema: $|y_i-M_y|$ in $|y_i-Me|$.

zap. št.	višina y_i (cm)	$\lvert y_i-M_y\rvert$	$\lvert y_i-Me\rvert$
1	162	15,3	14,0
2	163	14,3	13,0
3	169	8,3	7,0
4	170	7,3	6,0
5	170	7,3	6,0
6	171	6,3	5,0
7	172	5,3	4,0
8	173	4,3	3,0
9	175	2,3	1,0
10	175	2,3	1,0
11	177	0,3	1,0
12	178	0,7	2,0
13	178	0,7	2,0
14	180	2,7	4,0
15	181	3,7	5,0
16	183	5,7	7,0
17	183	5,7	7,0
18	190	12,7	14,0
19	197	19,7	21,0
20	199	21,7	23,0
skupaj	3547	146,6	146,0

Od tod je

$$d_M=\frac{1}{N}\sum_{i=1}^{N}|y_i-M|=\frac{1}{20}\cdot 146{,}6=7{,}33 \text{ cm}$$

$$d_{Me}=\frac{1}{N}\sum_{i=1}^{N}|y_i-Me|=\frac{1}{20}\cdot 146{,}0=7{,}30 \text{ cm}$$

♦ ♦ ♦

Kadar so podatki dani s **frekvenčno porazdelitvijo**, sta povprečni absolutni odklon od aritmetične sredine in povprečni absolutni odklon od mediane le oceni, ki ju izračunamo po formulah

$$d_M = \frac{1}{N}\sum_{i=1}^{r}\left(f_i \cdot |y_i - M|\right) \tag{6.07}$$

$$d_{Me} = \frac{1}{N}\sum_{i=1}^{r}\left(f_i \cdot |y_i - Me|\right) \tag{6.08}$$

V (6.07) in (6.08) pomenijo:

r – število razredov
f_i - frekvenca i-tega razreda, $i = 1,2,\ldots,r$
y_i - vrednost *sredine i-tega razred*a, $i = 1,2,\ldots,r$.

6.1.7 Varianca (disperzija)

Varianca je naslednja mera variabilnosti podatkov. Označimo jo z znakom *VAR* ali σ^2. Ker statistično spremenljivko lahko označimo z različnimi črkami (npr. *y, x,*…), tudi znaku za varianco običajno dodamo ustrezen indeks: če je spremenljivka označena z znakom *y,* potem bo znak za varianco σ_y^2, če je spremenljivka označena z znakom *x,* potem bo znak za varianco σ_x^2 in podobno.

Definicija variance je naslednja:

$$VAR = \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - M_y\right)^2 \tag{6.09}$$

Iz tega zapisa vidimo, da je varianca aritmetično povprečje kvadriranih odklonov razlik vrednosti posameznih podatkov od aritmetičnega povprečja teh podatkov.

Z varianco merimo *razpršenost* podatkov, torej opazujemo, kako močno se podatki razlikujejo od aritmetične sredine. Ker so nekatere razlike $y_i - M_y$ lahko pozitivne,

nekatere pa negativne, vzamemo v definiciji variance (6.09) kvadrate teh razlik, s čimer se izognemo negativnim vrednostim.

Formulo (6.09) lahko zapišemo tudi v drugačni obliki, ki je v veliko primerih uporabnejša. Izpeljimo jo!

Najprej kvadrirajmo dvočlenik pod sumacijskim znakom

$$\sigma_y^2=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-M_y\right)^2=\frac{1}{N}\sum_{i=1}^{N}\left(y_i^2-2y_iM_y+M_y^2\right)$$

Upoštevamo lastnosti vsote

$$\sigma_y^2=\frac{1}{N}\left(\sum_{i=1}^{N}y_i^2-2\sum_{i=1}^{N}y_iM_y+\sum_{i=1}^{N}M_y^2\right)=\frac{1}{N}\sum_{i=1}^{N}y_i^2-2M_y\cdot\frac{1}{N}\sum_{i=1}^{N}y_i+\frac{1}{N}\sum_{i=1}^{N}M_y^2$$

Zaradi

$$M_y=\frac{1}{N}\sum_{i=1}^{N}y_i \quad \text{in} \quad \sum_{i=1}^{N}M_y^2=\underbrace{M_y^2+M_y^2+\cdots+M_y^2}_{N-krat}=NM_y^2$$

dobimo

$$\sigma_y^2=\frac{1}{N}\sum_{i=1}^{N}y_i^2-2M_y\cdot M_y+\frac{1}{N}\cdot N\cdot M_y^2=\frac{1}{N}\sum_{i=1}^{N}y_i^2-2M_y^2+M_y^2=\frac{1}{N}\sum_{i=1}^{N}y_i^2-M_y^2$$

Druga formula za računanje variance je torej

$$\boxed{\sigma_y^2=\frac{1}{N}\sum_{i=1}^{N}y_i^2-M_y^2} \qquad (6.10)$$

6.1.8 Standardni odklon (standardna deviacija)

Varianca ni prav posebej ugodna mera za razpršenost podatkov. Za podatke, ki so daleč od povprečja, je razlika y_i-M_y velika, kvadrat $\left(y_i-M_y\right)^2$ pa še večji, če je $y_i-M_y>1$. Na ta način imajo ti podatki v vsoti (5.51) prevelik vpliv, s čimer je popačeno objektivno stanje v analizi razpršenosti podatkov. Tudi sicer kvadrat razlike v fizikalnem smislu nima

pomena. Če podatke merimo npr. v kilogramih, bi imel kvadat razlike fizikalno enoto “kg na kvadrat”, ki v realnosti seveda ne obstaja. Takšno anomalijo skušamo deloma popraviti na ta način, da vpeljemo novo enoto variabilnosti, ki jo imenujemo standardni odklon ali standardna deviacija, dobimo pa jo kot kvadratni koren variance.

Torej je

$$\sigma_y = +\sqrt{\sigma_y^2} \tag{6.11}$$

6.1.9 Koeficient variabilnosti

Razpršenost oz. variabilnost podatkov, dobljenih s statistično raziskavo, pa včasih merimo tudi z relativno mero (v procentih). V ta namen poznamo koeficient variabilnosti, ki je določen z naslednjim izrazom:

$$KV\% = \frac{\sigma_y}{M_y} \cdot 100 \tag{6.12}$$

Primer:
Vzemimo ponovno primer, ki ga že poznamo, to je meritev višin 20 študentov (podatki so dani v tabeli). Izračunajmo varianco, standardni odklon in koeficient variabilnosti.
Tabelo podatkov dopolnimo s stolpci $|y_i - M_y|$, $(y_i - M_y)^2$ ter y_i^2. V zadnji vrstici (z oznako »skupaj«) zapišimo še njihove vsote.

zap. št.	višina y_i (cm)	$\lvert y_i - M_y \rvert$	$(y_i - M_y)^2$	y_i^2
1	162	15,3	234,09	26244
2	163	14,3	204,49	26569
3	169	8,3	68,89	28561
4	170	7,3	53,29	28900
5	170	7,3	53,29	28900
6	171	6,3	39,69	29241
7	172	5,3	28,09	29584
8	173	4,3	18,49	29929
9	175	2,3	5,29	30625
10	175	2,3	5,29	30625
11	177	0,3	0,09	31329
12	178	0,7	0,49	31684
13	178	0,7	0,49	31684

14	180	2,7	7,29	32400
15	181	3,7	13,69	32761
16	183	5,7	32,49	33489
17	183	5,7	32,49	33489
18	190	12,7	161,29	36100
19	197	19,7	388,09	38809
20	199	21,7	470,89	39601
skupaj	3547	146,6	1818,20	630524

Varianco izračunajmo po obeh formulah.

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - M_y\right)^2 = \frac{1}{20}\cdot 1818{,}20 = 90{,}91$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - M_y^2 = \frac{1}{20}\cdot 630524 - 177{,}3^2 = 90{,}91$$

Standardni odklon

$$\sigma_y = +\sqrt{\sigma_y^2} = \sqrt{90{,}91} = 9{,}53 \text{ cm}$$

Koeficient variabilnosti pa je

$$KV\% = \frac{\sigma_y}{M_y}\cdot 100 = \frac{9{,}53}{177{,}3}\cdot 100 = 5{,}38\%$$

♦ ♦ ♦

Kadar so podatki dani s **frekvenčno porazdelitvijo**, variance ne moremo izračunati natanko, lahko jo le ocenimo po formuli

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{r} f_i\left(y_i - M_y\right)^2 \qquad (6.13)$$

oziroma

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{r} f_i y_i^2 - M_y^2 \qquad (6.14)$$

V obeh zgornjih formulah je

f_i - frekvenca i-tega razreda, $i = 1, 2, \ldots, r$

y_i - sredina i-tega razreda, $i = 1, 2, \ldots, r$

Varianca, ki jo računamo iz frekvenčne porazdelitve, je le ocena. Izkaže se, da je nekoliko prevelika v primerjavi z natančnim izračunom, ko so podatki dani v ranžirni vrsti. Napaka je tem večja, čim širša je širina razreda.

V primerih, ko so razredi enako široki, lahko to napako v veliki meri popravimo s **Shepardovim popravkom** $\frac{i^2}{12}$, ki ga odštejemo od izračunane ocene (6.13) oziroma (6.14). Pri tem je i širina razreda. Tako korigirana varianca je določena s formulo

$$\sigma_{yp}^2 = \sigma_y^2 - \frac{i^2}{12} \tag{6.15}$$

Primer:
Na zdravniškem pregledu so stehtali 250 študentov. Podatki o masi študentov so grupirani v 6 razredov in zapisani v spodnji tabeli. Izračunajmo varianco!

razred	Masa (kg)	Frekvenca f_i
1	40,0-50,0	11
2	50,1-60,0	51
3	60,1-70,0	86
4	70,1-80,0	64
5	80,1-90,0	33
6	91,1-100,0	5
Skupaj	-	250

Rešitev:
Uporabili bomo obe formuli, (5.55) in (5.56). V ta namen bomo tabelo dopolnili z nekaterimi novimi stolpci. Najprej zapišimo stolpec, v katerega bomo vnašali sredino razredov. Sredina drugega in naslednjih stolpcev je zaokrožena na eno decimalko. Drugi razred npr. ima širino 9,9 kg, zato je sredina 54,95 kg, kar smo zaokrožili na 55,0 kg. Podobno je to narejeno tudi drugod.

razred	Masa (kg)	Sredina razreda y_i (kg)	Frekvenca f_i
1	40,0-50,0	42,5	11
2	50,1-60,0	55,0	51
3	60,1-70,0	65,0	86
4	70,1-80,0	75,0	64
5	80,1-90,0	85,0	33
6	91,1-100,0	95,0	5
Skupaj	-	-	250

Potrebujemo aritmetično sredino, ki jo bomo dobili po formuli (5.36), to je $M_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{r} f_i y_i$, zato tabelo razširimo še z enim stolpcem, v katerega bomo zapisali produkte $f_i y_i$.

razred	Masa (kg)	y_i	f_i	$f_i y_i$
1	40,0-50,0	42,5	11	495
2	50,1-60,0	55,0	51	2805
3	60,1-70,0	65,0	86	5590
4	70,1-80,0	75,0	64	4800
5	80,1-90,0	85,0	33	2850
6	91,1-100,0	95,0	5	475
Skupaj	-	-	250	16970

Z uporabo formule (5.09) izračunamo aritmetično sredino

$$M_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{r} f_i y_i = \frac{1}{250}\cdot 16970 = 67,88 \text{ kg}$$

Ker želimo varianco izračunati po obeh formulah:

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{r} f_i \left(y_i - M_y\right)^2 \quad \text{in}$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{r} f_i y_i^2 - M_y^2$$

bomo tabelo dopolnili še z dodatnimi stolpci: $\left(y_i - M_y\right)$, $\left(y_i - M_y\right)^2$, $f_i\left(y_i - M_y\right)^2$, y_i^2 in $f_i y_i^2$.

Masa (kg)	y_i	f_i	$f_i y_i$	$\left(y_i - M_y\right)$	$\left(y_i - M_y\right)^2$	$f_i\left(y_i - M_y\right)^2$	y_i^2	$f_i y_i^2$
40,0-50,0	45,0	11	495	-22,9	524,41	5768,51	2025	22275
50,1-60,0	55,0	51	2805	-12,9	166,41	8486,91	3025	154275
60,1-70,0	65,0	86	5590	-2,9	8,41	723,26	4225	363350
70,1-80,0	75,0	64	4800	7,1	50,41	3226,24	5625	360000
80,1-90,0	85,0	33	2850	17,1	292,41	9649,53	7225	238425
91,1-100,0	95,0	5	475	27,1	734,41	3672,05	9025	45125
-	-	250	16970	-	1851,73	31526,50	-	1183450

Od tod dobimo:

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{r} f_i \left(y_i - M_y\right)^2 = \frac{1}{250}\cdot 31526,50 = 126,1$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{r} f_i y_i^2 - M_y^2 = \frac{1}{250}\cdot 1183450 - 67,88^2 = 126,1$$

Kot je bilo že rečeno, je dobljeni rezultat le ocena, ki jo moramo korigirati s Shepardovim popravkom $\frac{i^2}{12}$, kjer je *i* širina razreda frekvenčne porazdelitve.

V našem primeru bomo vzeli za širino razreda 10 kg, torej je na eno decimalko zaokrožen Shepardov popravek

$$\frac{i^2}{12} = \frac{10^2}{12} = 8,3$$

Varianca s Shepardovim popravkom je torej

$$\sigma_{yp}^2 = \sigma_y^2 - \frac{i^2}{12} = 126,1 - 8,3 = 117,8$$

♦ ♦ ♦

Računanje variance po zgornjih formulah je dolgotrajno, numeričnega računanja je veliko, zato seveda obstoji veliko možnosti za numerično napako. Da se vsemu temu izognemo, je bila razvita **metoda pomožnega znaka u**, po kateri dobimo varianco v razrede grupiranih podatkov mnogo enostavneje[2].

Metoda pomožnega znaka u poteka v korakih takole:

1. *Izberemo razred približno v sredini frekvenčne porazdelitve in mu dodelimo izhodiščno vrednost* $u=0$. *Od tega razreda navzgor in navzdol dodeljujemo razredom vrednosti* $\pm 1, \pm 2, \ldots$ *Natančneje: za manjše vrednosti grupiranih podatkov so vrednosti pomožnega znaka u negativne, za večje vrednosti grupiranih podatkov pa so vrednosti pomožnega znaka u pozitivne.*

2. *Izračunamo produkte* $f_i u_i$, $i=1,2,\ldots,r$.

3. *Izračunamo produkte* $f_i u_i^2$, $i=1,2,\ldots,r$.

4. *Seštejemo vrednosti* f_i, $f_i u_i$, $f_i u_i^2$, $i=1,2,\ldots,r$.

5. *Izračunamo parameter*

[2] Seveda je pri delu z računalnikom takšna dilema odveč.

$$K_u = \sum_{i=1}^{r} f_i u_i^2 - \frac{\left(\sum_{i=1}^{r} f_i u_i\right)^2}{N} \qquad (6.16)$$

6. *Izračunamo varianco po formuli*

$$\sigma_y^2 = \frac{i^2}{N} \cdot K_u \qquad \textit{(i – širina razreda)} \qquad (6.17)$$

7. *Upoštevamo Shepardov popravek* $\sigma_{yp}^2 = \sigma_y^2 - \frac{i^2}{12}$.

Primer:
Vzemimo kar prejšnji primer, to je masa 250 študentov.

Podatki so dani v tabeli.

razred	Masa (kg)	Frekvenca f_i
1	40,0-50,0	11
2	50,1-60,0	51
3	60,1-70,0	86
4	70,1-80,0	64
5	80,1-90,0	33
6	91,1-100,0	5
Skupaj	-	250

V skladu s koraki opisanega algoritma metode pomožnega znaka u bomo tabeli dodali nekaj novih stolpcev in jo zapisali takole:

razred	Masa (kg)	Frekvenca f_i	u_i	$f_i u_i$	$f_i u_i^2$
1	40,0-50,0	11	-2	-22	44
2	50,1-60,0	51	-1	-51	51
3	60,1-70,0	86	0	0	0
4	70,1-80,0	64	1	64	64
5	80,1-90,0	33	2	66	132
6	91,1-100,0	5	3	15	45
Skupaj	-	250	-	72	336

Seštevki f_i, $f_i u_i$, $f_i u_i^2$, $i = 1,2,\ldots,r$, ki jih potrebujemo, so izračunani v zadnji vrstici (skupaj).

$$K_u = \sum_{i=1}^{r} f_i u_i^2 - \frac{\left(\sum_{i=1}^{r} f_i u_i\right)^2}{N} = 336 - \frac{72^2}{250} = 315{,}26$$

Sedaj po formuli (6.17) izračunamo oceno variance

$$\sigma_y^2 = \frac{i^2}{N} \cdot K_u = \frac{10^2}{250} \cdot 315{,}26 = 126{,}1$$

Še Shepardov popravek:

$$\sigma_{yp}^2 = \sigma_y^2 - \frac{i^2}{12} = 126{,}1 - 8{,}3 = 117{,}8$$

Dobljeni rezultat se ujema z rezultatom, ki smo ga za iste podatke dobili v prejšnjem primeru.

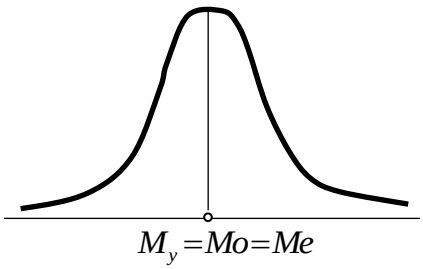
♦♦♦

5.6.2 Mere asimetrije

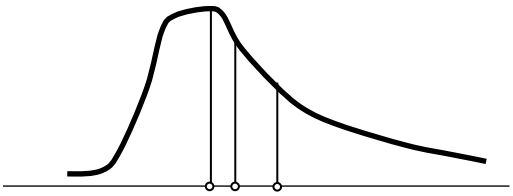
Mere asimetrije ugotavljamo le pri frekvenčnih porazdelitvah.
Od prej že vemo, da so frekvenčne porazdelitve

- simetrične,
- asimetrične v desno,
- asimetrične v levo.

Pri simetrični unimodalni porazdelitvi so srednje vrednosti aritmetična sredina, modus in mediana med seboj enake: $M_y = Mo = Me$.

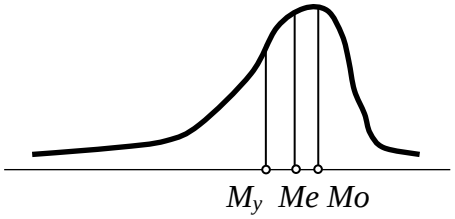


Kadar je porazdelitev asimetrična v desno, velja: $Mo<Me<M_y$



.

Ko je porazdelitev asimetrična v levo, pa je $Mo>Me>M_y$.



Mediana je pri asimetrični(unimodalni) porazdelitvi vedno med modusom in aritmetično sredino.

Stopnjo asimetrije merimo s ***koeficientom asimetrije KA*** na dva načina

a) na osnovi modusa, tedaj ga izračunamo takole

$$KA_{Mo}=\frac{M_y-Mo}{\sigma} \tag{6.18}$$

b) na osnovi mediane, tedaj pa ga izračunamo po formuli

$$KA_{Me}=\frac{3\left(M_y-Me\right)}{\sigma} \tag{6.19}$$

Ko je vrednost koeficientov negativna, je porazdelitev asimetrična v levo. Ko je vrednost koeficientov pozitivna, je porazdelitev asimetrična v desno. Ko je vrednost koeficientov enaka nič, je porazdelitev simetrična.

Porazdelitev je torej

- simetrična, ko je $KA_{Mo} = KA_{Me} = 0$
- asimetrična v levo, ko je $KA_{Mo} < 0,\ KA_{Me} < 0$
- asimetrična v desno, ko je $KA_{Mo} > 0,\ KA_{Me} > 0$

Primer:
Vzemimo ponovno 20 študentov in njihove velikosti.

zap. št.	višina y_k (cm)
1	162
2	163
3	169
4	170
5	170
6	171
7	172
8	173
9	175
10	175
11	177
12	178
13	178
14	180
15	181
16	183
17	183
18	190
19	197
20	199
skupaj	3547

Povprečno velikost že poznamo: $M_y = 177{,}3$ cm.
Izračunajmo še mediano. Zanjo velja

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

$$R = NP + 0{,}5 = 20 \cdot 0{,}5 + 0{,}5 = 10{,}5$$

razred: $10 < 10{,}5 < 11 \Rightarrow k = 11$

in od tod dobimo

$$Me = y_{10} + (R - F_{10})(y_{11} - y_{10}) = 175 + (10{,}5 - 10)(177 - 175) = 176 \text{ cm}$$

Če hočemo izračunati modus, moramo podatke grupirati v razrede. V danem primeru lahko naredimo npr. 8 razredov s širino po 5 cm, kot je prikazano v spodnji tabeli.

razred	f_i	F_i
160-165	2	2
166-170	3	5
171-175	5	10
176-180	4	14
181-185	3	17
186-190	1	18
191-195	0	18
196-200	2	20
Skupaj	20	-

Modusni razred je 3. razred ($f_3=5$) $\Rightarrow k=3$

$$Mo=y_{k,\min}+\frac{f_k-f_{k-1}}{2f_k-f_{k-1}-f_{k+1}}\cdot i_k=171+\frac{5-3}{2\cdot 5-3-4}\cdot 5=174{,}3 \text{ cm}$$

Torej je

$$\left.\begin{array}{l} M_y=177{,}3 \\ Me=176{,}0 \\ Mo=174{,}3 \end{array}\right\} \Rightarrow Mo<Me<M_y$$

Kot je bilo že rečeno, je mediana ***vedno*** *med aritmetično sredino in modusom.*

Izračunajmo še oba koeficienta asimetrija.
Iz enega od prejšnjih primerov vemo, da je za ta primer standardni odklon $\sigma_y=+\sqrt{\sigma_y^2}=\sqrt{90{,}91}=9{,}53$ cm.

Na ta način dobimo

a) koeficient asimetrije na osnovi modusa

$$KA_{Mo}=\frac{M_y-Mo}{\sigma}=\frac{177{,}3-174{,}3}{9{,}53}=0{,}315$$

b) koeficient asimetrije na osnovi mediane

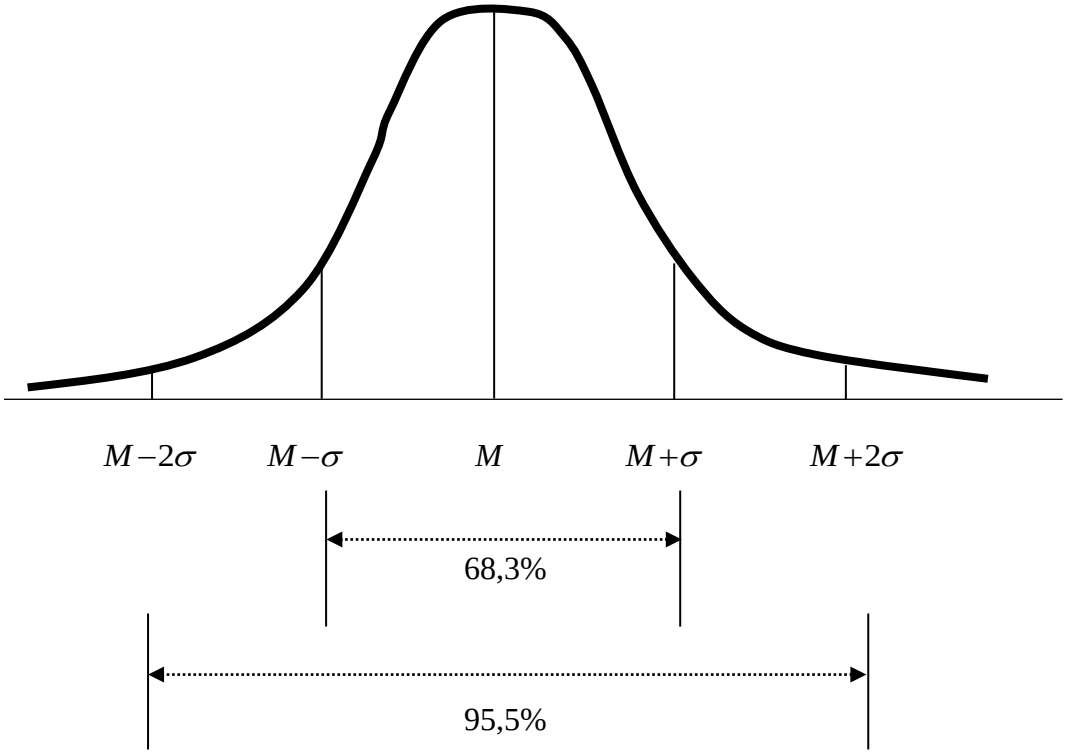
$$KA_{Me}=\frac{3\left(M_y-Me\right)}{\sigma}=\frac{3(177{,}3-176)}{9{,}53}=0{,}409$$

Oba koeficienta sta pozitivna, zato je porazdelitev asimetrična v desno.
♦ ♦ ♦

6.3 Mere sploščenosti

Frekvenčne porazdelitve imajo lahko enako srednjo vrednost, enako mero variabilnosti in enako tudi stopnjo asimetrije, pa kljub temu mordaa niso enake. Medsebojno se lahko razlikujejo po sploščenosti.

Med teoretičnimi porazdelitvami je najpogostejša **normalna (Gaussova) porazdelitev** $\varphi(x)$. Zanjo velja, da je v razmiku $(M-\sigma, M+\sigma)$ približno 68,3% vseh vrednosti, v intervalu $(M-2\sigma, M+2\sigma)$ približno 95,5% vseh vrednosti in v intervalu $(M-3\sigma, M+3\sigma)$ približno 99,7% vseh vrednosti (slika).



Za normalno porazdelitev velja $M_y = Me = Mo$ in $K_{Mo} = K_{Me} = 0$.

Vpeljemo **koeficient sploščenosti *KS*** tako**,** da je pri normalni porazdelitvi enak 1, $KS = 1$.

Koeficient sploščenosti za katerokoli drugo porazdelitev primerjamo s tole po formuli

$$KS = 1{,}9 \cdot \frac{Q_3 - Q_1}{D_9 - D_1} \qquad (6.20)$$

Kadar je

$KS>1$, je porazdelitev sploščena,
$KS<1$, je porazdelitev koničasta.

Primer:
Spet vzemimo 20 študentov, porazdeljenih v razrede po višini. Izračunajmo koeficient sploščenosti!

Rešitev:
Vemo že, da je za ta primer aritmetična sredina $M_y=177{,}3$ cm in standardni odklon $\sigma=9{,}53$ cm. Poznamo tudi že tretji kvartil: $Q_3=182$ cm.
Izračunajmo še Q_1, D_1 in D_9.

Prvi kvartil Q_1: $P(Q_1)=0{,}25 \quad R=NP+0{,}5=20\cdot 0{,}25+0{,}5=5{,}5\Rightarrow k=6$
$Q_1=170+(5{,}5-5)\cdot 1=170{,}5$ cm

Prvi decil D_1: $P(D_1)=0{,}1 \quad R=NP+0{,}5=20\cdot 0{,}10+0{,}5=2{,}5\Rightarrow k=3$
$D_1=163+(2{,}5-2)\cdot 6=166$ cm

Deveti decil D_9: $P(D_9)=0{,}9 \quad R=NP+0{,}5=20\cdot 0{,}90+0{,}5=18{,}5\Rightarrow k=19$
$D_9=190+(18{,}5-18)\cdot 2=191$ cm

Iz (5.62) dobimo iskani koeficient sploščenosti

$$KS=1{,}9\cdot\frac{Q_3-Q_1}{D_9-D_1}=1{,}9\cdot\frac{182-170{,}5}{191-166}=0{,}874<1$$

Ker je $KS<1$, je porazdelitev koničasta.
♦ ♦ ♦

7 MERE KONCENTRACIJE

Posamezne enote populacije statistične spremenljivke *y* dosegajo različne vrednosti $y_1, y_2, \ldots . y_N$. Zanima nas, kako je vsota vseh teh vrednosti porazdeljena po posameznih enotah. Možne so različne možnosti, ta vsota je lahko porazdeljena enakomerno po vseh enotah, lahko pa je porazdeljena neenakomerno ali le po manjšem številu enot populacije.

Želimo proučevali takšne možnosti, zato najprej uvedimo pojem *agregat*, kar pomeni vsoto vseh stvari, ki skupno tvorijo celotno količino nekega pojava.

Primer:
Agregat je lahko skupno kmetijsko zemljišče v ha v Sloveniji.
Agregat je lahko skupna količina denarja, ki je v obtoku v Sloveniji.
Agregat je lahko celotna količina gozdov na Dolenjskem.
Agregat je lahko skupno število zaposlenih v Novem mestu ipd.
♦♦♦

Vsak agregat ima enega ali več nosilcev (lastnikov), skupna količina agregata je običajno med nosilce porazdeljena različno – nekateri lastniki imajo večji delež, nekateri manjši delež agregata.

Takšno različno porazdelitev agregata opisujemo s pojmom **koncentracija**.

Če je pojav enakomerno porazdeljen med nosilce, potem pravimo, da koncentracije ni. Kadar pa je pojav neenakomerno porazdeljen po enotah populacije tako, da ima le nekaj enot razmeroma velik del celotnega agregata, govorimo o večji ali manjši koncentraciji pojava.
Na ta način je pojem koncentracije v tesni zvezi s pojmom *variacija pojava*. Če pojav ne varira, koncentracije ni, če pa varira, koncentracija je. Bolj kot se pojav spreminja, večja je koncentracija.
Koncentracija pojava torej pomeni razdelitev celotne (skupne) vrednosti spremenljivke po statističnih enotah.

Koncentracijo ugotavljamo z *Lorenzovo*[3] *krivuljo* in z *Ginijevim*[4] *koeficientom*.

Oba kazalnika lahko dobimo le iz frekvenčne porazdelitve. Poleg frekvence f_i morajo biti podani še podatki o deležu agregata, ki odpade na posamezni razred Y_i. Poznati pa moramo seveda še podatke o delih agregata (ali vsaj njihove ocene), ki pripadajo posamezni realizaciji pojava.

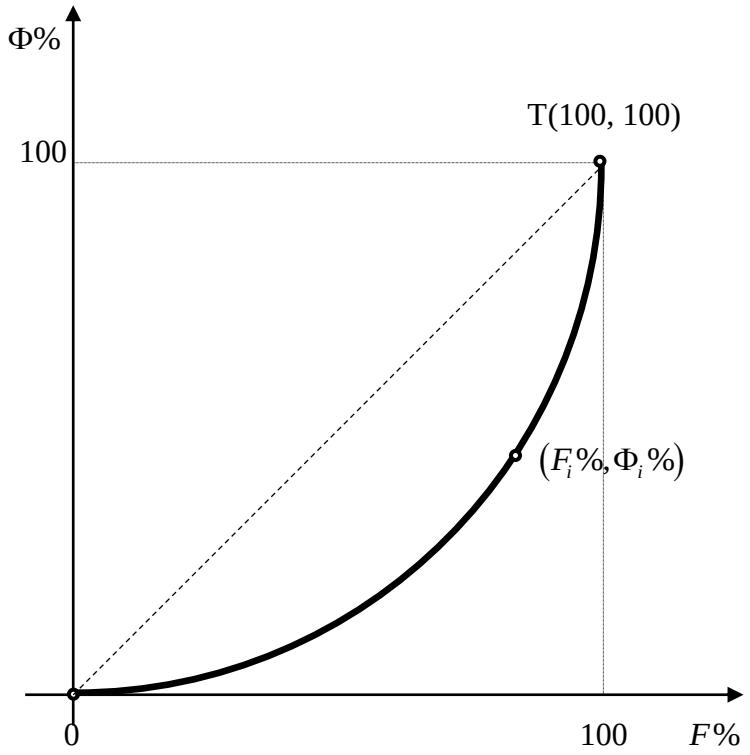
Oglejmo si najprej Lorenzovo krivuljo (ali tudi Lorenzov grafikon).

Lorenzov grafikon narišemo tako, da v kvadrat s stranicama za kumulative relativnih frekvenc $F\%$ na abscisi in kumulative relativnih vsot $\Phi\%$ na ordinati vnesemo točke s koordinatami $\left(F_i\%, \Phi_i\%\right)$.
Zveznica vseh takšnih točk za $k = 1, 2, \ldots, r$ je lomljena črta, ki v loku povezuje točko (0,0) s točko (100,100).
Kadar na vsako enoto odpade enak del vsote pojava (agregata), koncentracije ni in krivulja je daljica, ki povezuje točki (0, 0) in (100, 100), torej diagonala kvadrata na sliki.

Čim večja je koncentracija, tem bolj je dejanska krivulja odmaknjena od takšne diagonale.

[3] Max Otto Lorenz (1880 - 1962), ameriški ekonomist, je krivuljo, ki se po njem imenuje Lorenzova krivulja, prvikrat predstavil leta 1905.
[4] Corrado Gini (1884 – 1965), italijanski statistik, demograf in sociolog.

Zaradi definicije koncentracije je ta krivulja vedno konveksna, torej napeta desno od diagonale kvadrata.
Lorenzov grafikon je skiciran na spodnji sliki.



Primer:
Vzemimo vse zaposlene v gospodarskih podjetjih v neki občini. Podjetja imajo lahko različno število zaposlenih. Grupirajmo podjetja glede na število zaposlenih. Podatki so dani v spodnji tabeli. Narišimo Lorenzovo krivuljo.

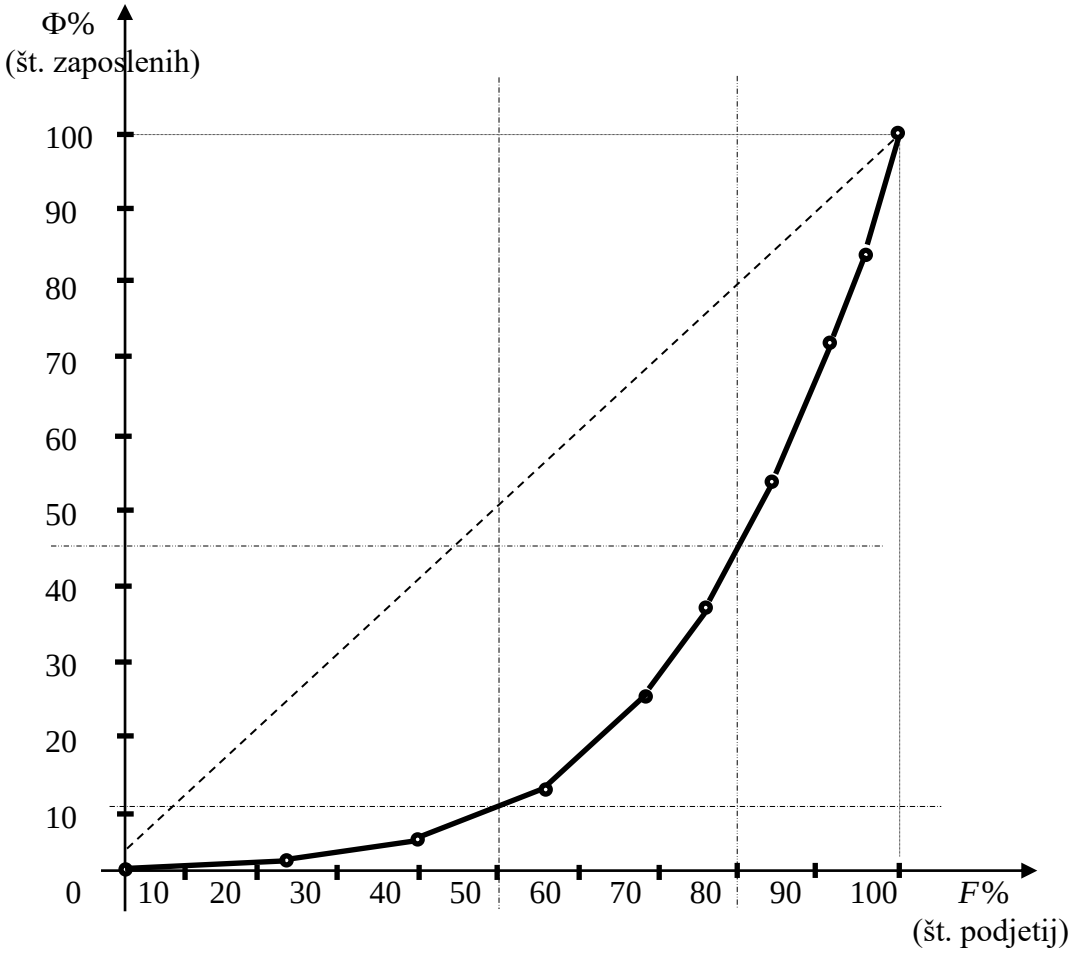
Razred (podjetja po številu zaposlenih)	Število podjetij f_i	Skupno število zaposlenih φ_i
do 10	30	150
11-50	20	600
51-100	20	1300
101-150	15	1700
151-200	10	1750
201-250	10	2300
251-300	10	2700
301-350	5	1700
nad 350	5	2500
skupaj	125	14700

Rešitev:
Imamo torej dva kriterija: število zaposlenih in število podjetij, kjer so ti zaposleni.
Proučujemo porazdelitev zaposlenih v podjetjih različnih velikosti, ugotavljamo torej

kolikšen delež zaposlenih v podjetjih posamezne velikosti. Agregat je v tem primeru skupno število zaposlenih (14.700).
Tabelo s podatki razširimo z dodatnimi stolpci, kjer za frekvenco f_i števila podjetij izračunamo relativno frekvenco $f_i\%$ in kumulativno frekvenco $F_i\%$, za frekvenco skupno zaposlenih φ_i pa relativno frekvenco $\varphi_i\%$ in kumulativno frekvenco $\Phi_i\%$.

Razred (št. zaposl.)	Število podjetij			Skupno število zaposlenih		
	f_i	$f_i\%$	$F_i\%$	φ_i	$\varphi_i\%$	$\Phi_i\%$
do 10	30	24,0	24	150	1,02	1,02
11-50	20	16,0	40	600	4,08	5,10
51-100	20	16,0	56	1300	8,84	13,94
101-150	15	12,0	68	1700	11,56	25,50
151-200	10	8,0	76	1750	11,91	37,41
201-250	10	8,0	84	2300	15,65	53,06
251-300	10	8,0	92	2700	18,37	71,43
301-350	5	4,0	96	1700	11,56	82,99
nad 350	5	4,0	100	2500	17,01	100,00
skupaj	125	100,0	-	14700	100,00	-

Podatki v stolpcu za $F_i\%$ so abscise, podatki v stolpcu za $\Phi_i\%$ pa so ordinate točk, ki določajo Lorenzovo krivuljo.



Komentar:
Iz grafa razberemo, da vsekakor prihaja do koncentracije zaposlenih po podjetjih. Največ zaposlenih je v srednje velikih podjetjih, saj ima 50% manjših podjetij le dobrih 10% zaposlenih.
Med drugim lahko vidimo tudi, da ima 10% največjih podjetij več kot 30% zaposlenih.
♦ ♦ ♦

Opomba:
Če ne poznamo porazdelitve agregata, jo ocenimo po formuli

$$\varphi_i \approx f_i y_i ,\ i=1,2,\ldots,r \qquad (7.01)$$

V (7.01) so y_i sredine razredov.

Če so podatki dani v ranžirni vrsti, dobimo Lorenzovo krivuljo koncentracije tako, da upoštevamo $f_i=1$, $Y_i=y_i$, $i=1,2,\ldots,N$.

Poglejmo sedaj še **Ginijev koeficient koncentracije**. To je mera koncentracije, ki jo dobimo iz Lorenzove krivulje. Kadar koncentracije ni, ima koeficient vrednost nič, kadar je vrednost statistične spremenljivke skoncentrirana le v eni točki, pa je vrednost koeficienta ena. Koncentracija je tem večja, čim bolj je Lorenzova krivulja odmaknjena od diagonale, kar pomeni, da je ***ploščina lika med krivuljo in diagonalo*** čim večja. Če označimo z znakom *S* to ploščino, je Ginijev koeficient določen takole:

$$G_k = 2S$$

Vzemimo sedaj kumulative relativnih frekvenc na abscisi in kumulative relativnih vsot na ordinati *v relativnem deležu (torej brez množenja s 100)* in ju označimo z znakoma F_i in Φ_i , zato je $0 \le F_i \le 1$ in $0 \le \Phi_i \le 1$ za vse $i=1,2,\ldots,r$.

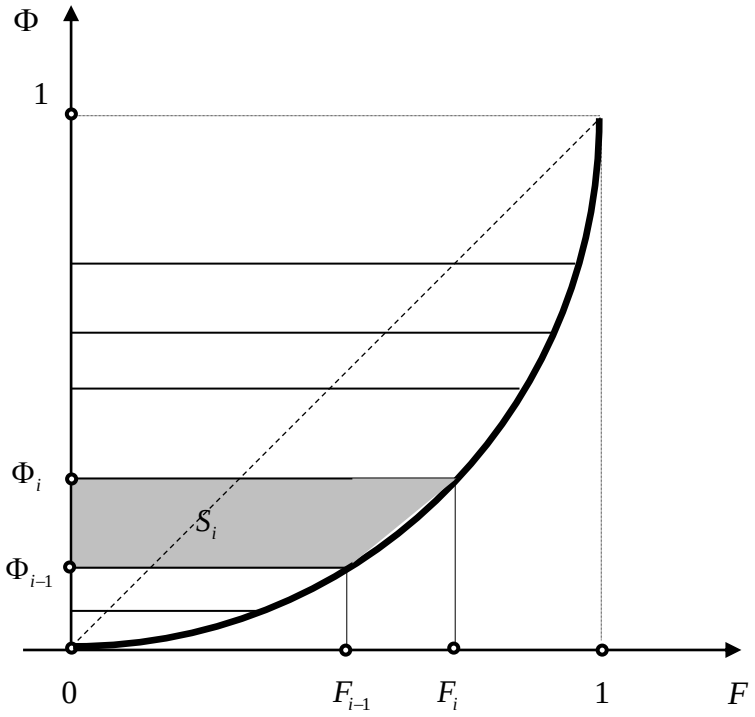
Izračunajmo Ginijev koeficient $G_k = 2S$, pri čemer si pomagamo s sliko. Iz slike vidimo, da z vzporednicami z abscisno osjo (ki jih narišemo v višini ordinat, to je podatkov za Φ_i) lik med ordinatno osjo in Lorenzovo krivuljo razdelimo na *n* likov, ki imajo približno obliko trapezov. Naj bo ploščina posameznega takšnega trapeza $S_i, i=1,2,\ldots,r$ in skupna iskana ploščina *S* je potem enaka seštevku ploščin vseh takšnih trapezov $\sum_{i=1}^{r} S_i$, od katere odštejemo ploščino zgornje polovice kvadrata, to je 0,5.

Velja torej

$$S=\sum_{i=1}^{r}S_i-\frac{1}{2}$$

Ginijev koeficient je potem

$$G_k=2S=2\left(\sum_{i=1}^{r}S_i-\frac{1}{2}\right)=2\sum_{i=1}^{r}S_i-1$$



Iz slike vidimo, da so ploščine trapezov

$$S_i=sred_i\cdot v_i\ ,\ i=1,2,\ldots,r$$

Ker je srednjica *i*-tega trapeza (kar razberemo iz slike) $sred_i=\frac{F_{i-1}+F_i}{2}$, njegova višina pa $v_i=\Phi_i-\Phi_{i-1}$, so ploščine takšne

$$S_i=\frac{F_{i-1}+F_i}{2}\cdot\left(\Phi_i-\Phi_{i-1}\right),\ i=1,2,\ldots,r$$

Vstavimo ta izraz v zapis $G_k=2\sum_{i=1}^{r}S_i-1$ in dobimo Ginijev koeficient v obliki:

$$G_k = \sum_{i=1}^{r}\left[\left(F_{i-1}+F_i\right)\left(\Phi_i-\Phi_{i-1}\right)\right]-1 \qquad (7.02)$$

Kadar ne poznamo porazdelitve agregata in jo moramo oceniti po formuli (7.01), torej $\varphi_i \approx f_i y_i$, $i=1,2,\ldots,r$, kjer je y_i je sredina *i*-tega razreda, lahko zaradi zveze

$$\Phi_i = \frac{\sum_{j=1}^{i} f_j y_j}{\sum_{i=1}^{r} f_i y_i} \quad \text{in} \quad \Phi_{i-1} = \frac{\sum_{j=1}^{i-1} f_j y_j}{\sum_{i=1}^{r} f_i y_i}$$

zgornji zapis zapišemo takole:

$$S_i = \frac{F_{i-1}+F_i}{2}\cdot\left(\Phi_i-\Phi_{i-1}\right) = \frac{F_{i-1}+F_i}{2}\cdot\frac{\left(\sum_{j=1}^{i} f_j y_j - \sum_{j=1}^{i-1} f_j y_j\right)}{\sum_{i=1}^{r} f_i y_i}$$

Ker je izraz v oklepaju v števcu druigega ulomka

$$\sum_{j=1}^{i} f_j y_j - \sum_{j=1}^{i-1} f_j y_j = \left(f_1 y_1 + f_2 y_2 + \cdots + f_{i-1} y_{i-1} + f_i y_i\right) - \left(f_1 y_1 + f_2 y_2 + \cdots + f_{i-1} y_{i-1}\right) = f_i y_i$$

je potem

$$S_i = \frac{F_{i-1}+F_i}{2}\cdot\frac{f_i y_i}{\sum_{i=1}^{r} f_i y_i} \qquad \text{za vsak } i=1,2,\ldots,r$$

in Ginijev koeficient

$$G_k = 2S = 2\left(\sum_{i=1}^{r} S_i - \frac{1}{2}\right) = \sum_{i=1}^{r}\left(F_{i-1}+F_i\right)\cdot\frac{f_i y_i}{\sum_{i=1}^{r} f_i y_i} - 1$$

Torej velja formula

$$G_k = \sum_{i=1}^{r}\left(F_{i-1}+F_i\right)\cdot\frac{f_i y_i}{\sum_{i=1}^{r} f_i y_i} - 1 \qquad (7.03)$$

Za izraz (7.02) oziroma (7.03) velja ocena

$$0 \le G_k \le \frac{r-1}{r} \tag{7.04}$$

To pomeni, da je $G_k = 0$ pri enakomerno razporejenih vrednostih statistične spremenljivke in $G_k = \frac{r-1}{r}$, ko je celotna vrednost skoncentrirana samo v eni enoti.

Kadar so vrednosti statističnega znaka razporejene enakomerno, koncentracije ni in Ginijev koeficient je 0, ko pa je celotna vrednost skoncentrirana samo v eni enoti, je primerno vzeti, da je Ginijev koeficient 1.

Da zadostimo še temu pogoju, uvedemo *normirani Ginijev koeficient* ${}^{(norm)}G_k$. Zanj velja

$$0 \le {}^{(norm)}G_k \le 1$$

Dobimo ga tako, da Ginijev koeficient (7.02) delimo z $\frac{r-1}{r}$.

$${}^{(norm)}G_k = \frac{G_k}{\frac{r-1}{r}} = \frac{r}{r-1} \cdot G_k$$

Iz enačbe (7.02) potem sledi

$${}^{(norm)}G_k = \frac{r}{r-1}\left[\sum_{i=1}^{r}\left[\left(F_{i-1}+F_i\right)\left(\Phi_i - \Phi_{i-1}\right)\right] - 1\right] \tag{7.05}$$

in iz (7.03)

$${}^{(norm)}G_k = \frac{r}{r-1}\left[\sum_{i=1}^{r}\left(F_{i-1}+F_i\right)\cdot\frac{f_i y_i}{\sum_{i=1}^{r} f_i y_i} - 1\right] \tag{7.06}$$

Primer:
Izračunajmo Ginijev koeficient za podatke prejšnjega primera (podjetja in zaposleni).

Rešitev:
Za izračun Ginijevega koeficienta bomo uporabili formuli (7.02) in (7.05).
Najprej preuredimo tabelo tako, da bomo kumulativne frekvence F_i in Φ_i zapisali v relativnem deležu, nato pa vpeljali tri nove stolpce: za $F_i + F_{i-1}$ in $\Phi_i - \Phi_{i-1}$ ter za produkt $\left(F_i - F_{i-1}\right)\left(\Phi_i - \Phi_{i-1}\right)$, $i = 1, 2, \ldots, 9$.
Da ne bo nesporazumov, označimo v tabeli z znakoma f_k in φ_k frekvence razredov za število podjetij in za število zaposlenih, z znaki f_i , φ_i, F_i in Φ_i pa relativne frekvence in kumulativne frekvence razredov, pri čemer so v tem primeru podatki zapisani v devetih vrsticah, torej $k = 1, 2, \ldots, 9$ in $i = 1, 2, \ldots, 9$.

Razred (št. zaposl.)	Število podjetij				Skupno število zaposlenih				
	f_k	f_i	F_i	$F_i + F_{i-1}$	φ_k	φ_i	Φ_i	$\Phi_i - \Phi_{i-1}$	$\left(F_i + F_{i-1}\right)\left(\Phi_i - \Phi_{i-1}\right)$
do 10	30	0,24	0,24	0,24	150	0,0102	0,0102	0,0102	0,002448
11-50	20	0,16	0,40	0,64	600	0,0408	0,0510	0,0408	0,026112
51-100	20	0,16	0,56	0,96	1300	0,0884	0,1394	0,0884	0,084864
101-150	15	0,12	0,68	1,24	1700	0,1160	0,2550	0,1160	0,143344
151-200	10	0,08	0,76	1,44	1750	0,1191	0,3741	0,1191	0,171504
201-250	10	0,08	0,84	1,60	2300	0,1565	0,5306	0,1565	0,250400
251-300	10	0,08	0,92	1,76	2700	0,1837	0,7143	0,1837	0,323312
301-350	5	0,04	0,96	1,88	1700	0,1156	0,8299	0,1156	0,217328
nad 350	5	0,04	1,00	1,96	2500	0,1701	1,0000	0,1701	0,333396
skupaj	125	1,00	-	-	14700	1,0000	-	-	1,552708

Ginijev koeficient je (7.02)

$$G_k = \sum_{i=1}^{9}\left[\left(F_{i-1} + F_i\right)\left(\Phi_i - \Phi_{i-1}\right)\right] - 1 = 1,552708 - 1 = 0,552708$$

To je približek ploščine med Lorenzovo krivuljo in diagonalo kvadrata v tem diagramu.

Normirani Ginijev koeficient pa je (7.05)

$$^{(norm)}G_k = \frac{n}{n-1}\left[\sum_{i=1}^{n}\left[\left(F_{i-1} + F_i\right)\left(\Phi_i - \Phi_{i-1}\right)\right] - 1\right] = \frac{n}{n-1} \cdot G_k = \frac{9}{8} \cdot 0,552708 = 0,6218.$$

♦ ♦ ♦

8. ČASOVNE VRSTE

Časovne vrste prikazujejo zaporedne vrednosti spremenljivk v enakih pretečenih časovnih obdobjih (npr. vsako uro, vsak dan, vsako leto, …). Vrednosti statistične spremenljivke se s časom spreminjajo, gre za dinamični process. Dinamiko tega spreminjanja prikazujemo in opisujemo z raznimi kazalniki. Nekatere že poznamo, nekatere bomo spoznali v nadaljevanju.

Najprej se spomnimo že znanih dinamičnih pokazateljev.

Označimo časovno vrsto

$$y_1, y_2, y_3, \ldots, y_N$$

Znani dinamični kazalniki so naslednji:

a) *indeksi s stalno osnovo*

$$I_{i/1} = \frac{y_i}{y_1} \cdot 100 \qquad i = 1,2,\ldots,N$$

b) *verižni indeksi*

$$I_i = \frac{y_i}{y_{i-1}} \cdot 100 \qquad i = 2,3,\ldots,N$$

c) *koeficient dinamike*

$$K_i = \frac{y_i}{y_{i-1}} \quad , i = 2,3,\ldots,N$$

d) *absolutna razlika* (kaže spremembo pojava) od člena do člena v absolutni meri

$$D_i = y_i - y_{i-1} \quad , i = 2,3,\ldots,N$$

e) *stopnja rasti (ali tudi tempo rasti)*

$$S_i = I_i - 100 \ , i = 2,3,\ldots,N$$

Med zgoraj zapisanimi kazalniki dinamike veljajo nekatere enostavne povezave, ki jih izpeljemo brez posebnih težav:

$$S_i = I_i - 100 = \frac{y_i}{y_{i-1}} \cdot 100 - 100 = \left(\frac{y_i}{y_{i-1}} - 1 \right) \cdot 100 = (K_i - 1) \cdot 100 \ , i = 2,3,\ldots,N \text{ , torej}$$

$$S_i = (K_i - 1) \cdot 100 \ , i = 2,3,\ldots,N$$

ali pa tudi

$$S_i = I_i - 100 = \frac{y_i}{y_{i-1}} \cdot 100 - 100 = \left(\frac{y_i}{y_{i-1}} - 1 \right) \cdot 100 = \frac{y_i - y_{i-1}}{y_{i-1}} \cdot 100 = \frac{D_i}{y_{i-1}} \cdot 100 \ , i = 2,3,\ldots,N$$

torej $S_i = \frac{D_i}{y_{i-1}} \cdot 100 \ , i = 2,3,\ldots,N$

Prikažimo kazalnike v tabeli:

Kazalnik	Pojav		
	raste	zastane (je konstanten)	pada
Člen vrste y_i	$y_{i+1} > y_i$	$y_{i+1} = y_i$	$y_{i+1} < y_i$
Indeks s stalno osnovo $I_{i/1}$	$I_{i+1/1} > I_{i/1}$	$I_{i+1/1} = I_{i/1}$	$I_{i+1/1} < I_{i/1}$
Absolutna razlika D_i	$D_i > 0$	$D_i = 0$	$D_i < 0$
Stopnja (tempo) rasti S_i	$S_i > 0$	$S_i = 0$	$S_i < 0$
Koeficient dinamike K_i	$K_i > 0$	$K_i = 0$	$K_i < 0$
Verižni indeks I_i	$I_i > 100$	$I_i = 100$	$I_i < 100$

Za poglobljeno analizo časovnih vrst pa našteti kazalniki niso dovolj, saj so vrednosti časovnih vrst in predvsem njihove spremembe odvisne od številnih dejavnikov, ki jih v teh kazalnikih nismo zajeli/upoštevali.

Če želimo pojave v časovnem dogajanju zadovoljivo predvidevati tudi "v naprej", moramo časovno vrsto temeljito **analizirati,** zato jo je potrebno najprej razstaviti na njene najpomembnejše komponente.

Dejavnike, ki vplivajo na elemente časovne vrste, delimo na naslednje:

- *trend,*
- *ciklični vplivi,*
- *periodični vplivi,*
- *enkratni slučajni vplivi.*

V posameznih časovnih vrstah ne prihaja vedno hkrati do vseh naštetih komponent, marsikdaj na primer ni zaslediti cikličnih ali periodičnih vplivov.

Ker govorimo o časovnih vrstah, kjer ugotavljamo vrednosti pojava v odvisnosti od časa, je, ko govorimo o funkcijski povezavi, *čas vedno neodvisna spremenljivka, vrednost pojava pa vedno odvisna spremenljivka. Tudi sicer velja: povsod, kjer proučujemo dogajanje v odvisnosti od časa (npr. v fiziki: hitrost, pospešek,…), je čas vedno neodvsina spremenljivka!*

Osnovna komponenta časovne vrste je **trend**.
To je pojem, ki *prikazuje osnovno smer razvoja vrednosti statističnih spremenljivk časovne vrste v daljšem časovnem obdobju.*
Le redki pojavi nimajo trenda.
Ugotavljanje trenda pomeni ugotavljanje funkcijske zveze $y(t)$, če smo z *y* označili statistično spremenljivko (pojav) in s *t* neodvisno spremenljivko (čas).
Definicijsko območje te funkcije je množica vrednosti neodvisne spremenljivke *t*, zaloga vrednosti pa je množica vrednosti spremenljivke *y*. Med njima obstaja funkcijska povezava $y=y(t)$. Časovna enota je od primera do primera različna, npr. dan, mesec, četrtletje in podobno. Ker imamo v splošnem *N* meritev preiskovanega pojava, je definicijsko območje kar zaporedje $D=\{1,2,\ldots,N\}$, neodvisna spremenljivka (čas) pa zavzame vrednosti $t=1, t=2,\ldots, t=N$. Torej je $y=y(t)$ diskretna funkcija neodvisne spremenljivke *t*, to je: $y_1=f(t=1), y_2=f(t=2),\ldots, y_N=f(t=N)$.

Ciklična sestavina časovne vrste se kaže v enakem vzorcu obnašanja/spreminjanja pojava v dajših časovnih intervalih, načeloma enako dolgih, vendar to ni nujno oz. zahtevano v proučevanju tega pojava.

Primer:
Ciklično nihanje gospodarske rasti po vzorcu konjuktura-recesija.
♦♦♦

Periodični vplivi se kažejo v enakem vzorcu spreminjanja pojava v krajših, vendar enako dolgih časovnih intervalih.

Primer:
Zmanjšanje proizvodnje v poletnih mesecih (dopusti), manjši nakupi zimske garderobe v poletnih mesecih, večja poraba kurilnega olja pozimi, znižanje cen zelenjave v poletnih mesecih, zvišanje cen zelenjave pozimi in podobno.
♦♦♦

Enkratni vplivi so večinoma slučajne narave in jih ne moremo predvideti.

Primer: zmanjšanje proizvodnje zaradi požara, zvišanje cen žita zaradi suše, zvišanje cen nafte zaradi terorističnih napadov in podobno.
♦♦♦

8.1 Trend in določanje trenda

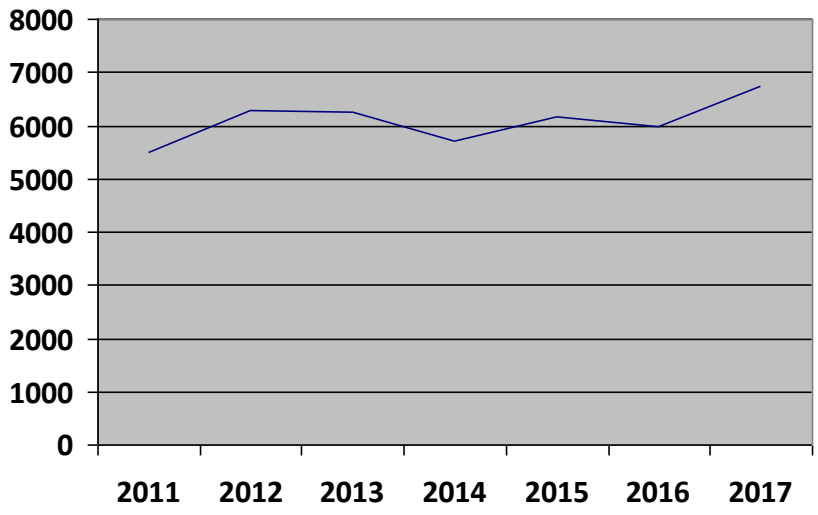
Trend je pomembna sestavina časovne vrste in prikazuje osnovno smer razvoja vrednosti statističnih spremenljivk časovne vrste v daljšem časovnem obdobju
Trend torej dolgoročno opisuje funkcijsko odvisnost vrednosti pojava od časa, kar opišemo v obliki $Y = f(t)$, če z znakom Y označimo trend kot časovno funkcijo. Vrednosti za Y v konkretnih časovnih trenutkih odražajo oceno gibanja pojava.

Časovne vrste običajno prikazujemo z ranžirnimi vrstami in grafično z linijskimi grafikoni.

Primer:
Vzemimo za primer spet izposojo knjig v neki knjižnici v posameznih letih (tabela).

Leto	2011	2012	2013	2014	2015	2016	2017
Št. izpos. knjig	5500	6300	6250	5700	6160	6000	6730

Graf:

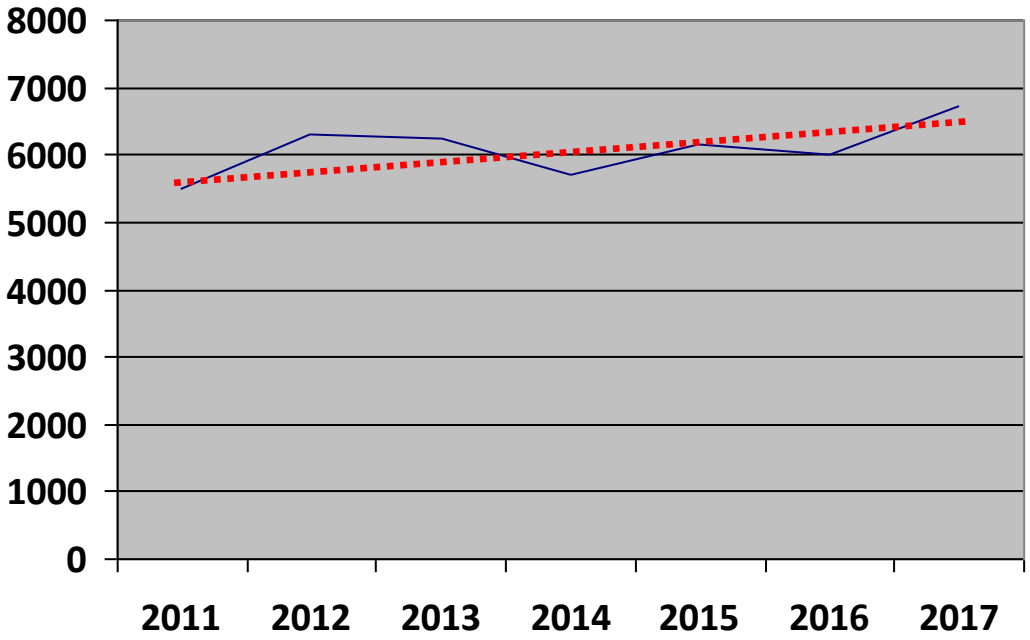


♦♦♦

Funkcijo trenda $Y=f(t)$ lahko določimo grafično ali analitično.

8.1.1 Grafično določanje trenda

Poglejmo si to možnost kar na primeru, ki smo ga spoznali zgoraj. V obstoječi diagram lahko z roko skiciramo krivuljo, ki *se najbolj* ***prilega*** *danim podatkom*. Naredimo to v konkretnem diagramu in dobimo spodnjo sliko. Funkcija trenda je premica, narisana z rdečo barvo.



Krivulja, ki grafično ponazarja funkcijo trenda, je lahko različnih oblik, najenostavnejša je seveda premica, kar smo uporabili tudi v tem primeru.

♦♦♦

8.1.2 Analitično določanje trenda, metoda najmanjših kvadratov

Funkcija trenda v matematični obliki je različna, odvisna je od posameznega pojava, ki ga proučujemo.

Običajno za funkcijo trenda uporabljamo enostavne funkcijske zveze, kot so zlasti:

- linearna funkcija $Y=a+bt$
- kvadratna funkcija $Y=a+bt+ct^2$
- eksponentna funkcija $Y=ab^t$

V vseh zgornjih zapisih je *t* čas kot neodvisna spremenljivka in *Y* (vrednost pojava) kot odvisna spremenljivka, *a, b, c, ...* pa so koeficienti (realna števila), ki jih moramo še določiti. Koeficiente izračunamo iz sistema normalnih enačb, ki jih bomo dobili po metodi najmanjših kvadratov.

Poglejmo, kaj pomeni »*metoda najmanjših kvadratov*«.

Časovna vrsta ima *N* členov, označimo jih $y_1, y_2, y_3, \ldots, y_N$. Funkcijske vrednosti trenda *Y* naj se za vsak čas *t* čim bolj prilegajo dejanskim vrednostim časovne vrste ob istem času. Označimo razlike med dejanskimi podatki in izračunanimi vrednostmi $\left(y_i - Y(t)\right)$. Te razlike kvadriramo, da se izognemo negativnim vrednostim, izračunamo torej $\left(y_i - Y(t)\right)^2$. Iskana funkcija trenda bo optimalna, ko bo vsota vseh teh razlik najmanjša. Metoda najmanjših kvadratov pomeni, da iščemo minimum izraza

$$F = \sum_{i=1}^{N} \left(y_i - Y(t)\right)^2 \quad \min \tag{8.01}$$

8.1.2.1 Trend je linearna funkcija

Za najenostavnejšo, to je linearno funkcijo $Y = a + bt$, pomeni rešitev enačbe (8.01) določitev koeficientov *a* in *b*. V tem primeru smemo vzeti, da je *F* funkcija dveh spremenljivk $F = F(a,b)$. Čas t pri tem zavzame diskretne vrednosti $t = 1, t = 2, \ldots, t = N$, krajše $t_i, i = 1,2,\ldots,N$.

Poiskati moramo torej ekstrem (minimum) funkcije, ki je odvisna od dveh spremenljivk *a* in *b*, to sta koeficienta linearne enačbe $Y = a + bt$. Iščemo minimum funkcije dveh spremenljivk

$$F(a,b) = \sum_{i=1}^{N} \left(y_i - a - bt_i\right)^2$$

Vemo, da ekstremi lahko nastopajo le v stacionarnih točkah, to so tiste, v katerih sta oba prva parcialna odvoda enaka nič.
Izračunamo oba prva parcialna odvoda in ju izenačimo z nič.

$$\frac{\partial F}{\partial a} = 2\sum_{i=1}^{N} \left(y_i - a - bt_i\right) \cdot (-1) = 0$$

$$\frac{\partial F}{\partial b} = 2\sum_{i=1}^{N} \left(y_i - a - bt_i\right) \cdot (-t_i) = 0$$

Uredimo prvo enačbo:

$$\sum_{i=1}^{N}\left(y_i - a - bt_i\right) = 0$$
$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} a - \sum_{i=1}^{N} bt_i = 0$$

Ker je $\sum_{i=1}^{N} a = \underbrace{a + a + \cdots + a}_{N-krat} = Na$, dobimo prvo enačbo v obliki

$$aN + b\sum_{i=1}^{N} t_i = \sum_{i=1}^{N} y_i$$

Uredimo še drugo enačbo:

$$\sum_{i=1}^{N}\left(y_i - a - bt_i\right)\cdot\left(-t_i\right) = 0$$
$$-\sum_{i=1}^{N} t_i y_i + \sum_{i=1}^{N} at_i + \sum_{i=1}^{N} bt_i^2 = 0$$

in od tod

$$a\sum_{i=1}^{N} t_i + b\sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} t_i y_i$$

Na ta način smo dobili sistem dveh enačb, ki ga imenujemo *sistem normalnih enačb*, od koder dobimo koeficienta *a* in *b*.

$$\begin{aligned} aN + b\sum_{i=1}^{N} t_i &= \sum_{i=1}^{N} y_i \\ a\sum_{i=1}^{N} t_i + b\sum_{i=1}^{N} t_i^2 &= \sum_{i=1}^{N} t_i y_i \end{aligned} \qquad (8.02)$$

Seveda moramo še preveriti, če je dobljena rešitev res minimum.

V ta namen izračunamo še vse druge parcialne odvode

$$\frac{\partial F}{\partial a}=-2\sum_{i=1}^{N}\left(y_i-a-bt_i\right)\Rightarrow\frac{\partial^2 F}{\partial a^2}=-2\sum_{i=1}^{N}-1=2N$$

$$\frac{\partial F}{\partial b}=-2\sum_{i=1}^{N}\left(y_i-a-bt_i\right)t_i\Rightarrow\frac{\partial^2 F}{\partial b^2}=-2\sum_{i=1}^{N}-t_i^2=2\sum_{i=1}^{N}t_i^2$$

$$\frac{\partial^2 F}{\partial a\partial b}=-2\sum_{i=1}^{N}-t_i=2\sum_{i=1}^{N}t_i$$

Kvadratna forma, ki jo porebujemo pri določanju ekstrema funkcije dveh spremenljivk, je v tem primeru:

$$K(a,b)=\frac{\partial^2 F}{\partial a^2}.\frac{\partial^2 F}{\partial b^2}-\left(\frac{\partial^2 F}{\partial a\partial b}\right)^2=4N\sum_{i=1}^{N}t_i^2-\left(2\sum_{i=1}^{N}t_i\right)^2=4\left(N\sum_{i=1}^{N}t_i^2-\left(\sum_{i=1}^{N}t_i\right)^2\right)$$

Zaradi

$$\sum_{i=1}^{N}t_i^2>\frac{1}{N}\left(\sum_{i=1}^{N}t_i\right)^2$$

je kvadratna forma pozitivna. Ker je tudi drugi odvod $\frac{\partial^2 F}{\partial a^2}=2N$ pozitiven, je rešitev, ki jo dobimo iz sistema normalnih enačb, res minimum.

Primer:
V podjetju FE, d.o.o so v letih od 2013 do vključno 2019 izdelali naslednje število proizvodov (tabela).

Zap. štev.	leto	Št. izdelkov
1.	2013	9082
2.	2014	10025
3.	2015	9571
4.	2016	11457
5.	2017	11002
6.	2018	12081
7.	2019	11582
-	skupaj	74800

Določimo linearno funkcijo trenda. Koliko izdelkov bodo na osnovi napovedanega trenda izdelali leta 2020, 2021?

Rešitev:
Privzeli bomo, da se trend ravna po linearni funkciji, ima torej enačbo $Y=a+bt$. Določiti moramo koeficienta *a* in *b*.
Uporabili bomo sistem normalnih enačb (8.02).

V ta namen dopolnimo tabelo s tremi stolpci: t_i, t_i^2 in $t_i y_i$, $i=1,2,\ldots,7$ (podatki so znani in zapisani v časovni vrsti za 7 let).

Zap. številka	leto	Št. izdelkov y_i	t_i	t_i^2	$t_i y_i$
1.	2013	9082	1	1	9082
2.	2014	10025	2	4	20050
3.	2015	9571	3	9	28713
4.	2016	11457	4	16	45828
5.	2017	11002	5	25	55010
6.	2018	12081	6	36	72486
7.	2019	11582	7	49	81074
-	skupaj	74800	28	140	312243

V sistem normalnih enačb (8.02) vstavimo podatke tega primera, pri čemer upoštevamo, da je:

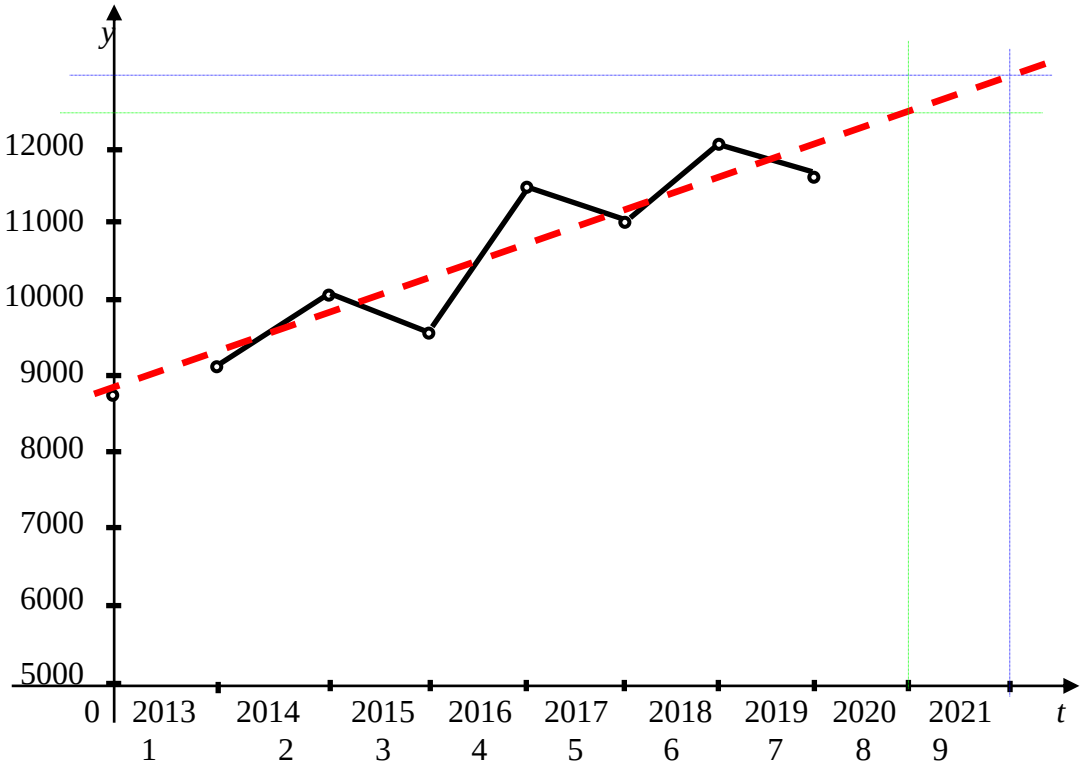
$$N=7,\ \sum_{i=1}^{7} t_i=28,\ \sum_{i=1}^{7} t_i^2=140,\ \sum_{i=1}^{7} t_i y_i=312243$$

dobimo system dveh enačb z dvema naznankama

$$7a+28b=74800$$
$$28a+140b=312243$$

Od tod izračunamo: $a=8822{,}5$ in $b=465{,}8$

Enačba linearne funkcije trenda se glasi: $y=8822{,}5+465{,}8t$.



Leta 2020 (tedaj je v enačbi funkcije trenda $t=8$) bo predvidena proizvodnja

$$y=8822{,}5+465{,}8t=8822{,}5+465{,}8\cdot 8=12548 \text{ proizvodov}$$

Leta 2021 (tedaj je v enačbi funkcije trenda $t=9$) bo predvidena proizvodnja

$$y=8822{,}5+465{,}8t=8822{,}5+465{,}8\cdot 9=13015 \text{ proizvodov}$$

Ta rezultat preberemo tudi na grafu.

♦ ♦ ♦

8.1.2.2 Trend je kvadratna funkcija

Če bi za funkcijo trenda izbrali kvadratno funkcijo $Y=a+bt+ct^2$, bi po metodi najmanjših kvadratov poiskali minimum izraza

$$F(a,b,c)=\sum_{i=1}^{N}\left(y_i-a-bt_i-ct_i^2\right)^2$$

Gre za funkcijo treh spremenljivk *a*, *b* in *c*. To so iskani koeficienti kvadratne funkcije trenda $Y=a+bt+ct^2$.

Vsi prvi parcialni odvodi morajo biti enaki nič:

$$\frac{\partial F}{\partial a}=-2\sum_{i=1}^{N}\left(y_i-a-bt_i-ct_i^2\right)=0$$

$$\frac{\partial F}{\partial b}=-2\sum_{i=1}^{N}\left(y_i-a-bt_i-ct_i^2\right)t_i=0$$

$$\frac{\partial F}{\partial c}=-2\sum_{i=1}^{N}\left(y_i-a-bt_i-ct_i^2\right)t_i^2=0$$

Z nekoliko truda dobimo iz teh treh enačb sistem (treh) normalnih enačb za neznane koeficiente *a*,*b* in *c*.

$$\begin{aligned} aN+b\sum_{i=1}^{N}t_i+c\sum_{i=1}^{N}t_i^2&=\sum_{i=1}^{N}y_i\\ a\sum_{i=1}^{N}t_i+b\sum_{i=1}^{N}t_i^2+c\sum_{i=1}^{N}t_i^3&=\sum_{i=1}^{N}t_iy_i\\ a\sum_{i=1}^{N}t_i^2+b\sum_{i=1}^{N}t_i^3+c\sum_{i=1}^{N}t_i^4&=\sum_{i=1}^{N}t_i^2y_i \end{aligned} \qquad (8.03)$$

Dokaz o tem, da gre res za minimum, bomo izpustili.

9 POVEZANOST IN ODVISNOST MED MNOŽIČNIMI POJAVI (REGRESIJA IN KORELACIJA)

Doslej smo posamezne pojave dane statistične množice proučevali ločeno, vsakega posebej in neodvisno od drugih pojavov iste statistične množice. V realnosti pa so različni pojavi velikokrat med seboj povezani, včasih šibkeje, včasih močneje.
V tem poglavju bomo ugotavljali, kako močno so posamezni pojavi iste statistične množice medsebojno povezani. Proučevali, merili in analizirali bomo njihove različne medsebojne odvisnosti in jih izrazili s posebnimi kriteriji.

Primer:

- proizvodnja električne energije je odvisna od vremena (padavine, sonce, veter), od količine izkopanega premoga, od števila delavcev, od popravil, ….
- cena blaga je odvisna od kvalitete, od povpraševanja, od sezone, …
- plača je odvisna od izobrazbe, od delovnega mesta, od podjetja, …

♦ ♦ ♦

V matematiki zapišemo funkcijsko odvisnost funkcije n spremenljivk v splošnem z izrazom

$$y = f(x_1, x_2, \ldots, x_n)$$

Pri tem je definicijsko območje znano in natanko določeno v n-dimenzionalnem prostoru, tudi pravilo f mora biti natanko definirano.

V statistiki pa (žal) ni tako enostavno. Prav natančne povezave v konkretnih primerih ni, zato takšno povezavo spremenljivk lahko zapišemo v splošnem z izrazom

$$\boxed{y = f(x_1, x_2, \ldots, x_n) + e} \qquad (9.01)$$

Takšnemu zapisu pravimo **korelacijska** ali **regresijska odvisnost**.

Z aditivnim členom e merimo *možnost slučajnih vplivov*.
Torej je e mera za slučajne vplive statistične spremenljivke, to je prav tista komponenta, ki je ne moremo zajeti v matematični obliki funkcijske odvisnosti.

Statističnih spremenljivk je lahko več, recimo da jih je *N*. Zapis oblike (9.01) velja za vsako spremenljivko posebej, torej

$$y_1 = f\left(x_{11}, x_{12}, \ldots, x_{1n}\right) + e_1$$
$$y_2 = f\left(x_{21}, x_{22}, \ldots, x_{2n}\right) + e_2$$
$$\ldots\ldots$$
$$y_N = f\left(x_{N1}, x_{N2}, \ldots, x_{Nn}\right) + e_N$$

Krajše:

$$y_i = f\left(x_{i1}, x_{i2}, \ldots, x_{in}\right) + e_i\,, \quad i = 1,2,\ldots,N\,.$$

Spremenljivki y_i pravimo *odvisna ali endogena spremenljivka*, spremenljivki x_{ij} pa *razlagalna ali eksogena spremenljivka* ($i = 1,2,\ldots,N$, $j = 1,2,\ldots,n$).

Povezavo med vsemi spremenljivkami največkrat pišemo v obliki linearne odvisnosti:

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} + \varepsilon_i \qquad i = 1,2,\ldots,N \tag{9.02}$$

Koeficientom a_i v (9.02) pravimo <u>*regresijski koeficienti*</u>.

Funkcijsko odvisnost v matematiki prikazujemo na tri znane načine:

- tabelarično,
- analitično,
- grafično.

Korelacijsko odvisnost tudi v statistiki zelo podobno prikazujemo na tri načine, to so (vzemimo, da imamo opraviti z <u>dvema statističnima spremenljivkama</u>):

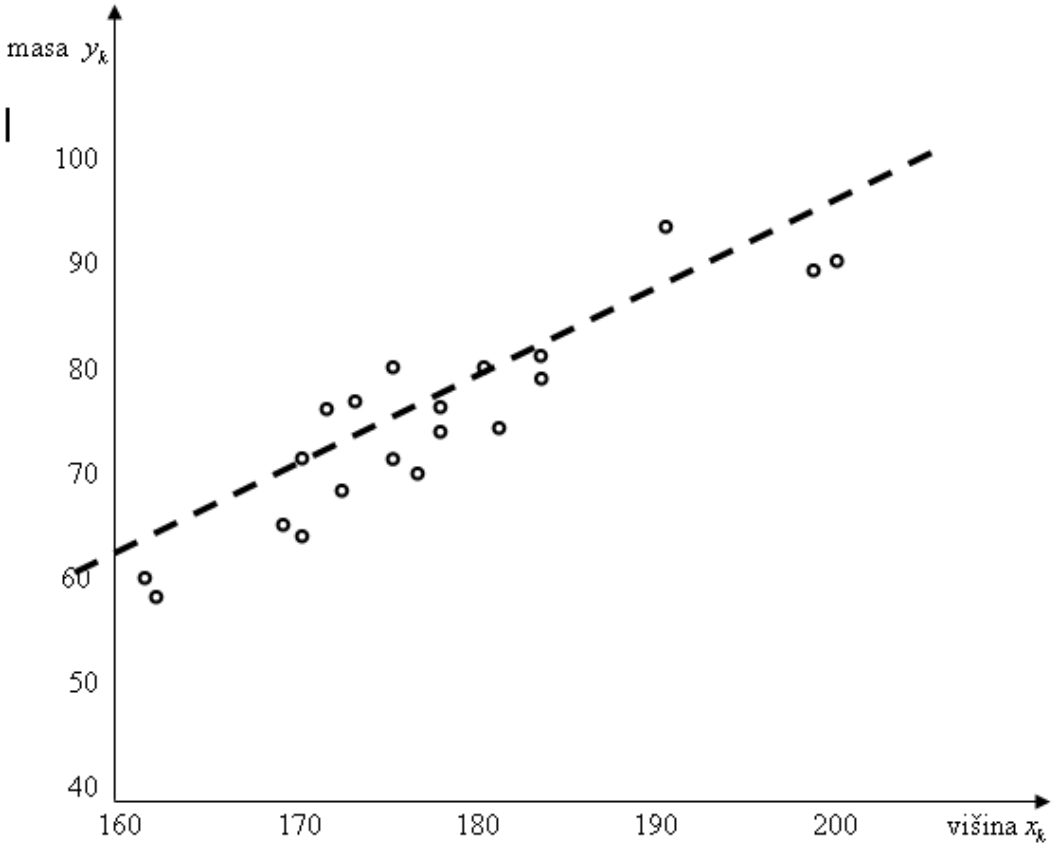
- v *korelacijski tabeli* z nizom dvojic (parov) vrednosti koreliranih podatkov za vsako enoto statistične populacije,
- s točkami v *korelacijskem grafikonu,*
- v funkcijski obliki z *regresijsko krivuljo.*

Primer:
Na zdravniškem pregledu so 20 študentom izmerili velikost (v cm) in maso (v kg). Dobljene podatke zapišemo v **korelacijski tabeli**.

zap. št.	višina x_k (cm)	Masa y_k (kg)
1	162	60
2	163	59
3	169	65
4	170	63
5	170	71
6	171	76
7	172	68
8	173	77
9	175	71
10	175	80
11	177	70
12	178	73
13	178	75
14	180	80
15	181	74
16	183	79
17	183	81
18	190	93
19	197	89
20	199	90

Ugotavljamo korelacijsko odvisnost med maso in višino študentov. Iz izkušenj vemo, da je masa praviloma odvisna od velikosti, obratno, da bi bila velikost študenta odvisna od njegove mase, pa ne drži.

Sedaj narišimo **korelacijski grafikon** (pravimo tudi: *razsevni grafikon*). Na abscisni osi bomo odmerjali višino študentov, na ordinatni osi pa njihovo maso.



♦ ♦ ♦

Krivulja, ki jo potegnemo med vrisanimi točkami tako, da se jim najbolj prilega, je **regresijska krivulja.**

Regresijska krivulja pokaže, kakšna bi bila zveza med koreliranima spremenljivkama, če ne bi bilo slučajnih vplivov. V takšnem primeru bi bile vse točke za vse statistične enote kar na regresijski krivulji in odvisnost med pojavoma (spremenljivkama) bi bila kar funkcijska odvisnost.

Korelacijsko odvisnost med **dvema pojavoma** (spremenljivkama) pišemo v obliki

$$y=f(x)+e$$

Korelacijska odvisnost med pojavoma je lahko različne intenzitete, v nekaterih primerih je velika, včasih je lahko manjša.

Korelacijska odvisnost je največja, ko je $e=0$ (tedaj je to kar funkcijska odvisnost), najmanjša pa je v primeru, ko spremenljivki nista odvisni, ko je $f(x)=C$.

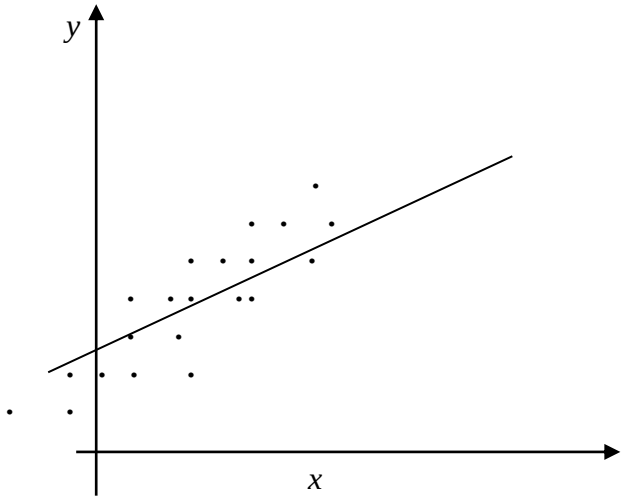
Korelacijska odvisnost je pozitivna, če se povečuje vrednost ene spremenljivke, ko se povečuje tudi vrednost druge, torej ko sta spremenljivki premo sorazmerni.

Korelacijska odvisnost je negativna, če se povečuje vrednost ene spremenljivke, ko se manjša vrednost druge, torej ko sta spremenljivki obratno sorazmerni.

Ko je regresijska krivulja premica, je povezanost med pojavoma linearna, v nasprotnem primeru je nelinearna.

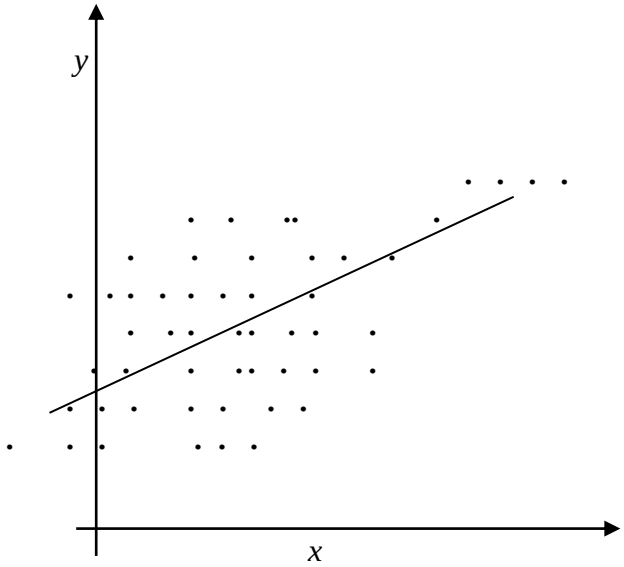
Nekaj primerov:

a)



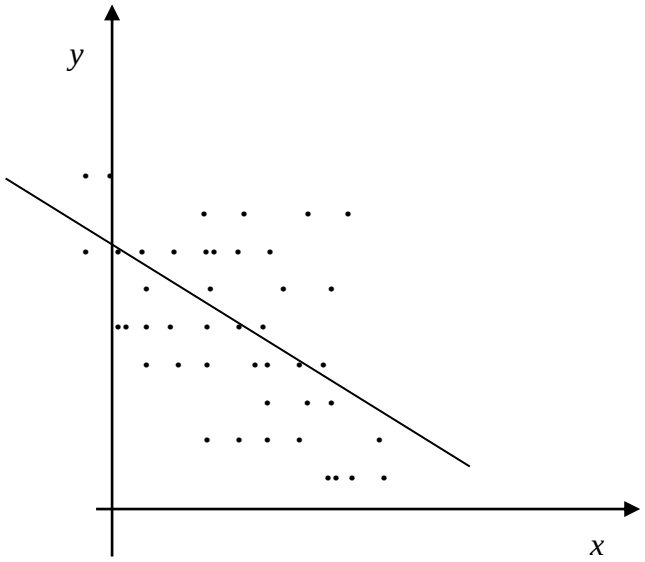
Korelacijska povezava na skici a je
- linearna,
- pozitivna,
- velika (močna).

b)



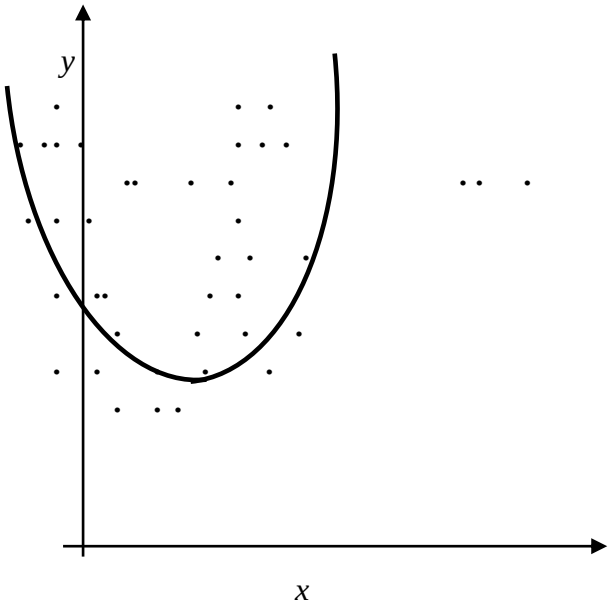
Korelacijska povezava na skici b je
- linearna,
- pozitivna,
- majhna (šibka).

c)



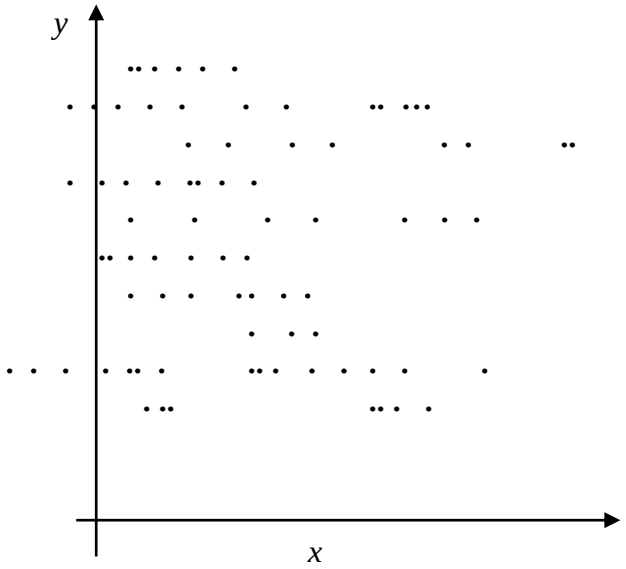
Korelacijska povezava na skici c je
- linearna,
- negativna,
- srednje močna.

d)



Korelacijska povezava na skici d je
- kvadratna (parabolična),
- močna.

e)



Korelacijska povezava na skici e je

- nepovezana.

♦♦♦

9.1 Določanje regresijskih krivulj

Regresijske krivulje lahko določimo grafično prostoročno, lahko pa jo izračunamo analitično. V prejšnjih primerih so bile regresijske krivulje določene prostoročno, kar je sicer enostavno, vendar močno subjektivno.

9.1.1 Analitično določanje regresijske krivulje

Vzemimo, da po analizi podatkov ugotovimo ustrezen tip funkcije

$$y = f(x_i, a_0, a_1, \ldots, a_n)$$

Izračunati moramo koeficiente $a_0, a_1, \ldots, a_n$, ki bodo določali natančno lego in obliko iskane krivulje. Krivuljo seveda določamo tako, da se konkretnim podatkom najbolje prilega. V ta namen bomo spet uporabili metodo najmanjših kvadratov.
Označimo

y_i - dejanske vrednosti podatkov

$\hat{y}_i$ - vrednosti na korelacijski funkciji

x_i - spremenljivka, korelacijsko povezana z y_i

$i = 1, 2, \ldots, N$.

9.1.1.1 Linearna korelacija

Linearna povezanost je analitično podana z enačbo

$$\hat{y}=a+bx \tag{9.03}$$

Po metodi najmanjših kvadratov mora biti

$$F(a,b)=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2=\sum_{i=1}^{N}(y_i-a-bx_i)^2 \quad \min$$

Na podoben način (prva parcialna odvoda morata biti enaka 0), kot smo to storili pri iskanju linearne funkcije trenda, dobimo tudi tu *sistem normalnih enačb*, iz katerih izračunamo koeficienta *a* in *b*.

$$\begin{aligned} aN+b\sum_{i=1}^{N}x_i&=\sum_{i=1}^{N}y_i \\ a\sum_{i=1}^{N}x_i+b\sum_{i=1}^{N}x_i^2&=\sum_{i=1}^{N}x_iy_i \end{aligned} \tag{9.04}$$

Splošna rešitev tega sistema enačb (9.04) je

$$b=\frac{\sum_{i=1}^{N}x_i\cdot\sum_{i=1}^{N}y_i-N\sum_{i=1}^{N}x_iy_i}{\left(\sum_{i=1}^{N}x_i\right)^2-N\sum_{i=1}^{N}x_i^2}=\frac{\frac{1}{N}\sum_{i=1}^{N}x_iy_i-\frac{1}{N}\sum_{i=1}^{N}x_i\cdot\frac{1}{N}\sum_{i=1}^{N}y_i}{\frac{1}{N}\sum_{i=1}^{N}x_i^2-\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2}=$$

$$=\frac{\frac{1}{N}\sum_{i=1}^{N}x_iy_i-\bar{x}\cdot\bar{y}}{\frac{1}{N}\sum_{i=1}^{N}x_i^2-\bar{x}^2}=\frac{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i-\bar{x})\cdot(y_i-\bar{y})\right]}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}=\frac{C_{xy}}{\sigma_x^2}$$

$$a=\frac{1}{N}\sum_{i=1}^{N}y_i-b\cdot\frac{1}{N}\sum_{i=1}^{N}x_i=\bar{y}-b\bar{x}$$

Torej:

$$b=\frac{C_{xy}}{\sigma_x^2} \quad , \quad a=\bar{y}-b\bar{x} \tag{9.05}$$

V (9.05) sta $\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$ in $\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i$ aritmetični povprečji spremenljivk *x* in *y*, izraz

$$C_{xy}=\frac{1}{N}\sum_{i=1}^{N}\left[\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)\right]=\frac{1}{N}\sum_{i=1}^{N}x_iy_i-\bar{x}\,\bar{y} \tag{9.06}$$

pa je **kovarianca.**

Kovarianca C_{xy} (9.06) meri intenzivnost povezanosti obeh spremenljivk.

Iz formule (9.05) vidimo, da kovarianca nastopa v koeficientu *b* linearne regresijske funkcije $\hat{y}=a+bx$, torej določa strmino regresijske premice.

Ko v enačbo (9.03) vstavimo $b=\frac{C_{xy}}{\sigma_x^2}$, dobimo

$$\hat{y}=a+bx=\left(\bar{y}-b\bar{x}\right)+\frac{C_{xy}}{\sigma_x^2}x=\bar{y}-\frac{C_{xy}}{\sigma_x^2}\bar{x}+\frac{C_{xy}}{\sigma_x^2}x=\bar{y}+\frac{C_{xy}}{\sigma_x^2}\left(x-\bar{x}\right)$$

Torej je linearna regresijska funkcija

$$\hat{y}=\bar{y}+\frac{C_{xy}}{\sigma_x^2}\left(x-\bar{x}\right) \tag{9.07}$$

Ko je kovarianca enaka nič, $C_{xy}=0$, iz (9.07) sledi $\hat{y}=\bar{y}=konst$.

Ko je regresijska funkcija konstanta, korelacije med spremenljivkama ni, spremenljivki nista povezani.

Torej velja:

- ko je $C_{xy}=0$, spremenljivki nista povezani,
- ko je $C_{xy}\neq 0$, sta spremenljivki povezani.

Predznak kovariance kaže smer povezanosti spremenljivk.

- ko je $C_{xy} > 0$, je korelacijska povezava spremenljivk pozitivna,
- ko je $C_{xy} < 0$, je korelacijska povezava spremenljivk negativna.

Primer:
Vzemimo že znani primer – 20 študentov, za katere poznamo njihove višine in njihove mase. Ali obstaja med tema dvema statističnima spremenljivkama (višina in masa) korelacija? Izračunajmo regresijsko premico! Podatki so znani že od prej in dani v tabeli.

Rešitev:
Tabelo dopolnimo s tremi stolpci: x_i^2, y_i^2 in $x_i y_i$ in dobimo

zap. št.	višina x_i (cm)	masa y_i (kg)	x_i^2	y_i^2	$x_i y_i$
1	162	60	26244	3600	9720
2	163	59	26569	3481	9617
3	169	65	28561	4225	10985
4	170	63	28900	3969	10710
5	170	71	28900	5041	12070
6	171	76	29241	5776	12996
7	172	68	29584	4624	11696
8	173	77	29929	5929	13321
9	175	71	30625	5041	12425
10	175	80	30625	6400	14000
11	177	70	31329	4900	12390
12	178	73	31684	5329	12994
13	178	75	31684	5625	13350
14	180	80	32400	6400	14400
15	181	74	32761	5476	13394
16	183	79	33489	6241	14457
17	183	81	33489	6561	14823
18	190	93	36100	8649	17670
19	197	89	39809	7921	17533
20	199	90	39601	8100	17920
skupaj	3546	1494	630524	113288	266461

Najprej izračunamo obe aritmetični sredini:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{20} \cdot 3546 = 177{,}3 \text{ cm}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{20} \cdot 1494 = 74{,}7 \text{ kg}$$

Nato izračunamo kovarianco:

$$C_{xy}=\frac{1}{N}\sum_{i=1}^{N}x_i y_i-\bar{x}\,\bar{y}=\frac{1}{20}\cdot 266461-177{,}3\cdot 74{,}7=13323-13244{,}3=78{,}7$$

Varianca spremenljivke *x:*

$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\bar{x}^2=\frac{1}{20}\cdot 630524-177{,}3^2=31526{,}2-31435{,}3=90{,}9$$

Varianca spremenljivke *y:*

$$\sigma_y^2=\frac{1}{N}\sum_{i=1}^{N}y_i^2-\bar{y}^2=\frac{1}{20}\cdot 113288-74{,}7^2=5664{,}4-5580{,}1=84{,}3$$

Regresijska premica, ki kaže korelacijo med spremenljivkama (odvisnost spremenljivke *y* od spremenljivke *x*) ima enačbo $\hat{y}=a+bx$.

Koeficienta *a* in *b* dobimo iz (9.04):

$$b=\frac{C_{xy}}{\sigma_x^2}=\frac{78{,}7}{90{,}9}=0{,}87$$

$$a=\bar{y}-b\bar{x}=74{,}7-0{,}87\cdot 177{,}3=-79{,}6$$

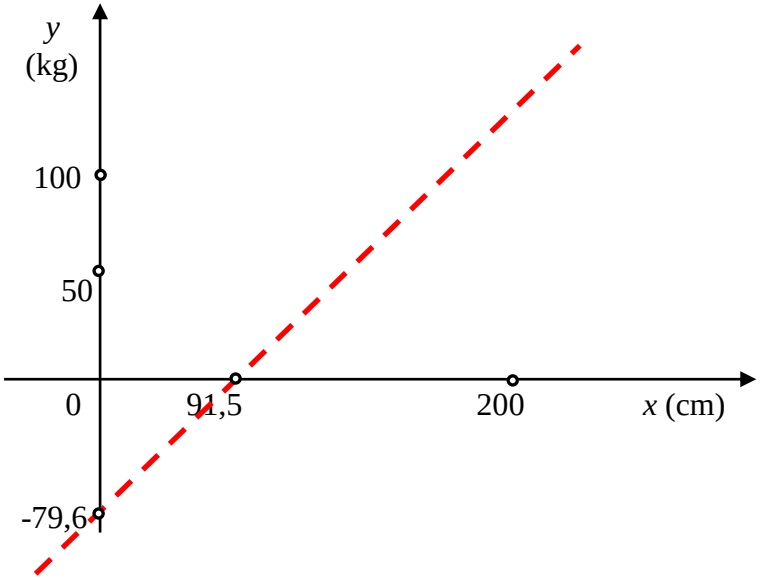
Enačba regresijske premice za maso v odvisnosti od višine študentov je

$$\hat{y}=-79{,}6+0{,}87x$$

Še grafično. Za premico potrebujemo dve točki. Vzemimo (kot običajno)

$$x=0\Rightarrow y=-79{,}6 \quad \text{in} \quad y=0\Rightarrow x=\frac{79{,}6}{0{,}87}=91{,}5$$

Graf:



♦♦♦

9.2 Moč korelacijske povezave

Moč korelacijske povezave med dvema statističnima spremenljivkama ugotavljamo z dvema kazalnikoma:

- determinacijski koeficient,
- korelacijski koeficient.

Skupno varianco σ_y^2 odvisne spremenljivke y lahko pišemo kot vsoto poznane (ali tudi: razložene) variance σ_{xy}^2 in nepojasnjene (ali: nerazložene) variance σ_{ey}^2, to je

$$\sigma_y^2 = \sigma_{xy}^2 + \sigma_{ey}^2$$

Delimo zgornjo enačbo s σ_y^2

$$1 = \frac{\sigma_{xy}^2}{\sigma_y^2} + \frac{\sigma_{ey}^2}{\sigma_y^2} \tag{9.08}$$

Kvocient $\frac{\sigma_{xy}^2}{\sigma_y^2}$ imenujemo **determinacijski koeficient.** Označimo ga

$$\boxed{D = r_{xy}^2 = \frac{\sigma_{xy}^2}{\sigma_y^2}} \tag{9.09}$$

Drugi kvoceint v (9.08) $\frac{\sigma_{ey}^2}{\sigma_y^2}$ pa je **koeficient nedoločenosti**. Označimo ga

$$U=\frac{\sigma_{ey}^2}{\sigma_y^2} \tag{9.10}$$

Zveza med obema koeficientoma je zaradi (9.08) preprosta

$$D+U=1 \tag{9.11}$$

Determinacijski koeficient *D* lahko včasih pišemo tudi takole

$$D=r_{xy}^2=r_{xy}\cdot r_{xy}$$

Z determinacijskim koeficientom merimo stopnjo odvisnosti spremenljivke y od spremenljivke x. Čim večji je vpliv x na y, tem večji je determinacijski koeficient.

Seveda je ta koeficient vedno nenegativen in kvečjemu enak 1:

$$0\le r_{xy}^2\le 1$$

Kadar je korelacijski koeficient

- $r_{xy}^2\to 1$, je korelacijska odvisnost zelo močna,
- $r_{xy}^2\approx 0{,}50$, je korelacijska odvisnost srednje močna,
- $r_{xy}^2\to 0$, je korelacijska odvisnost šibka.

Kadar pa je $r_{xy}^2=1$, sta statistični spremenljivki kar funkcijsko odvisni, saj bi bil tedaj determinacijski koeficient D = 0.

Pomembna značilnost povezave med statističnima spremenljivkama je še *smer povezanosti*, ki jo definiramo s **korelacijskim koeficientom** $r_{xy}=\pm\sqrt{r_{xy}^2}$ po formuli

$$r_{xy}=\frac{C_{xy}}{\sigma_x\sigma_y} \tag{9.12}$$

Ker je $0\le r_{xy}^2\le 1$, je korelacijski koeficient število z intervala $[-1,1]$;

$$-1\le r_{xy}\le 1$$

S korelacijskim koeficientom določimo ***smer*** in in tudi ***jakost*** odvisnosti med statističnima spremenljivkama.

Primer:
Izračunajmo korelacijski in determinacijski koeficient za prejšnji primer (študenti, merjena višina in masa).

Rešitev:
Za te podatke že poznamo oba standardna odklona in kovarianco (vse zaokroženo na eno decimalko)

$$\sigma_x^2 = 90{,}9 \Rightarrow \sigma_x = \sqrt{90{,}9} = 9{,}5$$

$$\sigma_y^2 = 84{,}3 \Rightarrow \sigma_y = \sqrt{84{,}3} = 9{,}2$$

$$C_{xy} = 78{,}7$$

Iz enačbe (9.12) izračunamo korelacijski koeficient

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{78{,}7}{9{,}5 \cdot 9{,}2} = 0{,}9$$

Determinacijski koeficient pa je

$$D = r_{xy}^2 = r_{xy} \cdot r_{xy} = 0{,}9 \cdot 0{,}9 = 0{,}81$$

Komentar:
a) Determinacijski koeficient $D = 0{,}81$ pove, da je 81% variance za maso pojasnjeno z linearno odvisnostjo mase od višine, le 19% pa je odvisno od slučajnih vplivov.
b) Smer korelacije je pozitivna, ker je $r_{xy} > 0$.

♦ ♦ ♦

Varianca je lahko tudi nepojasnjena, kar združimo v kazalniku σ_{ey}. Iz (9.10) in (9.11) dobimo

$$\sigma_{ey} = \sigma_y \sqrt{1 - r_{xy}^2} \qquad (9.13)$$

Kazalnik σ_{ey} imenujemo **standardna napaka ocene odvisne spremenljivke** in ga uporabljamo za *ugotavljanje zaupanja* pri ocenjevanju vrednosti odvisne spremenljivke.

Primer:
Pri normalni porazdelitvi je koeficient zaupanja

a) 0,683 pri razmiku $\hat{y}_x - \sigma_{ey} < y_x < \hat{y}_x + \sigma_{ey}$
b) 0,954 pri razmiku $\hat{y}_x - 2\sigma_{ey} < y_x < \hat{y}_x + 2\sigma_{ey}$
c) 0,997 pri razmiku $\hat{y}_x - 3\sigma_{ey} < y_x < \hat{y}_x + 3\sigma_{ey}$

♦ ♦ ♦

Primer:
Ponovno vzemimo 20 študentov, za katere poznamo višino in maso. Vzemimo, da sta spremenljivki normalno porazdeljeni. S koeficientom zaupanja 0,95 ocenimo razmik za telesno maso študenta, ki je visok 180 cm.

Rešitev:
Regresijsko premico v tem primeru že poznamo: $\hat{y} = -79{,}6 + 0{,}87x$.

Regresijska točkovna ocena za maso pri tej višini je

$$\hat{y}_{x=180} = -79{,}6 + 0{,}87 \cdot 180 = 77{,}1 \text{ kg}$$

Standardna napaka ocene je

$$\sigma_{ey} = \sigma_y \sqrt{1 - r_{xy}^2} = 9{,}2 \cdot \sqrt{1 - 0{,}81} = 4{,}0$$

Ker govorimo o koeficientu zaupanja 0.95, vidimo iz prejšnjega primera, da v tem primeru masa (kot odvisna spremenljivka) leži v intervalu $\hat{y}_x - 2\sigma_{ey} < y_x < \hat{y}_x + 2\sigma_{ey}$.
Razmik za maso je tedaj takšen

$$\hat{y}_{x=180} - 2\sigma_{ey} < y_{x=180} < \hat{y}_{x=180} + 2\sigma_{ey}$$
$$77{,}1 - 2 \cdot 4{,}0 < y_{x=180} < 77{,}1 + 2 \cdot 4{,}0$$
$$69{,}1 < y_{x=180} < 85{,}1$$

Z zaupanjem 0,95 je masa 180 cm visokega študenta med 69,1 kg in 85,1 kg. Ocena je precej slaba, saj vsebuje velik razmik, kar 16 kg.

♦ ♦ ♦

10 SLUČAJNE SPREMENLJIVKE, TEORETIČNE PORAZDELITVE

10.1. Slučajne spremenljivke

Definicija: *Slučajna spremenljivka je merljiva količina, ki ima svojo vrednost odvisno od slučaja in je natanko določena s svojo zalogo vrednosti ter svojim porazdelitvenim zakonom.*

Slučajne spremenljivke običajno označujemo z znaki $X, Y, Z, \ldots$ ali pa tudi $X_1, X_2, X_3, \ldots.$ in podobno. Pripadajoče *vrednosti* slučajnih spremenljivk označujemo $x, y, z, \ldots$ ali $x_1, x_2, x_3, \ldots$

Dogodke, da slučajna spremenljivka zavzame vrednost iz svoje zaloge, označujemo z znaki kot npr. $(X = x)$, $(X < x)$, $(x_1 < X < x_2)$, $(X \geq x)$ ipd.

Primer:
Slučajna spremenljivka je lahko:
- število pik pri metu kocke,
- oddaljenost zadetka od centra pri streljanju v tarčo,
- število avtomobilov v vrsti na bencinski črpalki,
- vrednost dobitka na ruleti itd.

♦ ♦ ♦

Pri slučajnih spremenljivkah moramo poznati dvoje:

- zalogo vrednosti,
- predpis, ki določa verjetnosti za njihove vrednosti. Ta predpis imenujemo *porazdelitveni zakon slučajne spremenljivke.*

Slučajne spremenljivke v splošnem delimo na diskretne in nediskretne slučajne spremenljivke.

Diskretne slučajne spremenljivke so tiste, katerih zaloga vrednosti je neko končno ali neskončno zaporedje.

Slučajne spremenljivke, katerih zalogo vrednosti ne moremo numerirati z naravnimi števili, so *nediskretne slučajne spremenljivke***.** Med nediskretnimi slučajnimi spremenljivkami so zlasti pomembne *zvezno porazdeljene slučajne spremenljivke*.

Splošna oblika porazdelitvenega zakona se imenuje *porazdelitvena funkcija*.

Definicija: *Porazdelitvena funkcija* $F(x)$ *slučajne spremenljivke X je funkcija, ki ima pri vsakem realnem x vrednost enako verjetnosti dogodka* $(X < x)$*, kar zapišemo v obliki*

$$F(x) = P(X < x) \qquad -\infty < x < \infty \tag{10.01}$$

Porazdelitvena funkcija $F(x)$ ima naslednje lastnosti:

- je naraščajoča funkcija med 0 in 1

$$F(-\infty) = 0\ ,\ F(+\infty) = 1 \tag{10.02}$$

- če sta x_1 in x_2 realni števili in je $x_1 < x_2$, velja

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1) \tag{10.03}$$

- za vsak realen x velja še

$$F(x+0) - F(x) = P(X = x) \tag{10.04}$$

Podobno, kot pri dogodkih in poskusih, uvedemo tudi pri slučajnih spremenljivkah pojem medsebojne (ne)odvisnosti.

Definicija: *Če sta slučajni spremenljivki X in Y takšni, da je pri poljubnih realnih x in y* $P(X < x, Y < y) = P(X < x)\,P(Y < y)$*, pravimo, da sta med seboj neodvisni. V nasprotnem primeru sta slučajni spremenljivki med seboj odvisni.*

Za diskretne slučajne spremenljivke pa je splošno obliko porazdelitvenega zakona primerneje podati z *verjetnostno funkcijo*.

Definicija: Verjetnostna funkcija p_k *diskretne slučajne spremenljivke X je funkcija, ki ima pri vsakem mogočem k svojo vrednost enako verjetnosti dogodka* $(X = x_k)$*, torej* $p_k = P(X = x_k)$*. Pri tem k preteče vse tiste cele vrednosti, za katere spada* x_k *v zalogo vrednosti diskretne slučajne spremenljivke X.*

Zveza med verjetnostno funkcijo in porazdelitveno funkcijo je iz obeh definicij jasna:

$$p_k = P(X = x_k) \tag{10.05}$$

$$F(x) = \sum_{x_k < x} p_k \tag{10.06}$$

Porazdelitvena funkcija diskretne slučajne spremenljivke narašča torej samo v *skokih*.

Ker tvorijo dogodki $(X = x_k)$ popolni sistem dogodkov, velja, da je vsota vseh skokov enaka 1, torej

$$\sum_k p_k = 1 \tag{10.07}$$

Podatke o diskretni slučajni spremenljivki običajno zapišemo v obliki *verjetnostne sheme*:

$$X: \begin{pmatrix} x_1 & x_2 & x_3 & . & . \\ p_1 & p_2 & p_3 & . & . \end{pmatrix} \tag{10.08}$$

Primer:
Iz serije 100 izdelkov, kjer je 10% izdelkov slabih, smo na slepo izbrali pet izdelkov in preverjali njihovo kvaliteto. Naj bo slučajna spremenljivka X število slabih izdelkov. Zapišimo njeno verjetnostno shemo.

Rešitev:
Slučajna spremenljivka lahko doseže 6 vrednosti: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$ in $x_6 = 5$. Verjetnosti so

$$p_k = P(X = k) = \frac{K_{10}^k \, K_{90}^{5-k}}{K_{100}^5} \text{ za } k = 0, 1, 2, 3, 4, 5.$$

Rezultate zapišemo v verjetnostno shemo

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0{,}5837524 & 0{,}3393908 & 0{,}0702187 & 0{,}0063834 & 0{,}0002514 & 0{,}0000033 \end{pmatrix}$$

Še preizkus:

$$\sum_{k=0}^{5} p_k = 0{,}5837524 + 0{,}3393908 + 0{,}0702187 + 0{,}0063834 + 0{,}0002514 + 0{,}0000033 = 1$$

♦ ♦ ♦

10.2 Nekatere porazdelitve slučajne spremenljivke

10.2.1 Porazdelitve diskretne slučajne spremenljivke

10.2.1.1 Enakomerna diskretna porazdelitev

Diskretna slučajna spremenljivka je porazdeljena *enakomerno,* če tvorijo njeno zalogo vrednosti poljubna realna števila $x_1, x_2, \ldots, x_n$ in je za vsak k, $k = 1, 2, \ldots, n$, $n \in N$, porazdelitveni zakon podan z verjetnostno funkcijo oblike

$$p_k = P(X = x_k) = \frac{1}{n} \tag{10.09}$$

Primer:
Naj bo pri metanju igralne kocke slučajna spremenljivka število pik pri posameznem metu. To je seveda diskretna spremenljivka, ki doseže vrednosti: 1, 2, 3, 4, 5, in 6. Verjetnosta funkcija prirejena metanju igralne kocke, je porazdeljena enakomerno. Njena verjetnostna shema je:

$$X : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

10.2.1.2 Binomska porazdelitev b(n,p)

Slučajna spremenljivka je porazdeljena po binomskem zakonu, ko je njena zaloga vrednosti množica števil $\{0, 1, 2, \ldots, n\}$, verjetnostna funkcija pa ima obliko

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \qquad n \in N, \quad 0 < p < 1 \tag{10.10}$$

Primer:
Vzemimo poskus, kjer nas zanima le, če se dogodek *A* v posamezni ponovitvi poskusa zgodi ali ne. Verjetnost, da se v posamezni ponovitvi poskusa dogodek *A* zgodi, naj bo *p*, verjetnost, da se dogodek *A* ne zgodi, je potem $1-p=q$.
Posamezni *j*-ti ponovitvi poskusa lahko priredimo slučajno spremenljivko X_j s takšnim dogovorom: če se *A* zgodi, ima ta spremenljivka vrednost 1, če se *A* ne zgodi, pa ima spremenljivka vrednost 0.
Torej smemo vsaki slučajni spremenljivki X_j prirediti verjetnostno shemo oblike

$$X_j : \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix} \tag{10.11}$$

Če smo poskus ponovili *n*-krat in se je dogodek *A* zgodil *k*-krat, ima vsota $X = X_1 + X_2 + \cdots + X_n$ vrednost *k* (tisti členi v vsoti, kjer se je *A* zgodil, imajo vrednost 1, ostali imajo vrednost 0). To pomeni, da je *X* ravno enaka frekvenci dogodka *A* in je porazdeljena po binomskem zakonu $b(n, p)$.

♦ ♦ ♦

10.2.1.3 Poissonova porazdelitev P(a)

Slučajna spremenljivka je porazdeljena po Poissonovem zakonu, ko je njena zaloga vrednosti množica števil $\{0,1,2,\ldots,n\}$, verjetnostna funkcija pa ima obliko

$$\boxed{p_k = \frac{a^k e^{-a}}{k!}, \quad a > 0} \tag{10.12}$$

Matematično upanje (o tem več malo kasneje) in varianca te porazdelitve sta enaki parametru *a*. Poissonova porazdelitev je posebej primerna pri proučevanju npr. števila vozil, ki v času t pridejo do neke točke (cestninska postaja, bencinska črpalka) ali števila ljudi, ki pridejo v vrsto na blagajni, na banki, pošti, bencinski črpalki in podobno. Matematično upanje Poisonove porazdelitve v tem primeru pomeni povprečno število vozil, povprečno število kupcev,.., ki v danem trenutku prispejo do dane točke.

10.2.1.4 Pascalova porazdelitev

Slučajna spremenljivka *X* je porazdeljena po *Pascalovem zakonu reda m*, če je njena zaloga vrednosti množica števil $\{m,\ m+1,\ m+2,\ \ldots\}$, $m \in N$, verjetnostna funkcija pa ima obliko

$$p_k = \binom{k-1}{m-1} p^m (1-p)^{k-m} \tag{10.13}$$

10.2.1.5 Geometrijska porazdelitev

Ko v Pascalovi porazdelitvi vzamemo $m = 1$, je zaloga vrednosti slučajne spremenljivke kar množica naravnih števil $\{1,\ 2,\ 3,\ \ldots\}$, verjetnostna funkcija pa ima obliko

$$p_k = p\,(1-p)^{k-1}\ ,\ k = 1, 2, 3, \ldots \tag{10.14}$$

Zaporedje

$$p_1 = p(1-p)^0,\ p_2 = p(1-p)^1,\ p_3 = p(1-p)^2\ p_4 = p(1-p)^3,\ \ldots$$

je geometrijsko s kvocientom $(1-p)$, zato se porazdelitev s takšno verjetnostno funkcijo imenuje *geometrijska porazdelitev.*

10.2.1.6 Hipergeometrijska porazdelitev

To porazdelitev običajno razložimo s tole nalogo: v posodi je N kroglic, od tega M belih ter $(N - M)$ črnih. Iz posode *n*-krat na slepo izbiramo po eno kroglico. Izbranih kroglic *ne vračamo* v posodo.
Kakšna je verjetnost p_{nk}, da je med temi *n* potegi (kjer je $n \le \min(M, N - M)$) natanko *k* belih kroglic?
Ko *n*-krat izbiramo po eno kroglico brez vračanja, je efekt enak, kot če naenkrat potegnemo *n* kroglic. Ker je v posodi *N* kroglic, je torej vseh možnih potegov v tem primeru $\binom{N}{n}$. Med temi izidi so ugodni tisti, pri katerih izberemo *k* belih kroglic (izmed *M*) ter $(n-k)$ črnih kroglic (izmed danih $N - M$). Takšnih izbir je $\binom{M}{k}\binom{N-M}{n-k}$, zato velja

$$p_{nk} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\ ,\quad k = 0, 1, 2, \ldots \tag{10.15}$$

Porazdelitev (10.15) se imenuje *hipergeometrijska porazdelitev*. Uporablja se v statistični kontroli kvalitete, kjer je treba kontrolirati serijo *N* izdelkov, med katerimi je *M* slabih (nekvalitetnih).
V veliko primerih ne moremo kontrolirati cele serije, zato se kontrola izvaja tako, da pregledujemo kvaliteto proizvodov le v slučajno izbranem *vzorcu* z *n* izdelki. Z zgornjo formulo je potem dana verjetnost, da bo v izbranem vzorcu z *n* izdelki natanko *k* nekvalitetnih proizvodov.

10.2.2 Zvezne porazdelitve

Pravimo, da je slučajna spremenljivka *X zvezno porazdeljena*, če se njena porazdelitvena funkcija izraža v obliki:

$$F(x) = \int_{-\infty}^{x} p(t)dt \tag{10.16}$$

Funkcija $p(t)$, ki jo imenujemo **gostota verjetnosti**, je določena z odvodom porazdelitvene funkcije

$$p(x) = F'(x) \tag{10.17}$$

Za gostoto verjetnosti veljajo naslednje lastnosti:

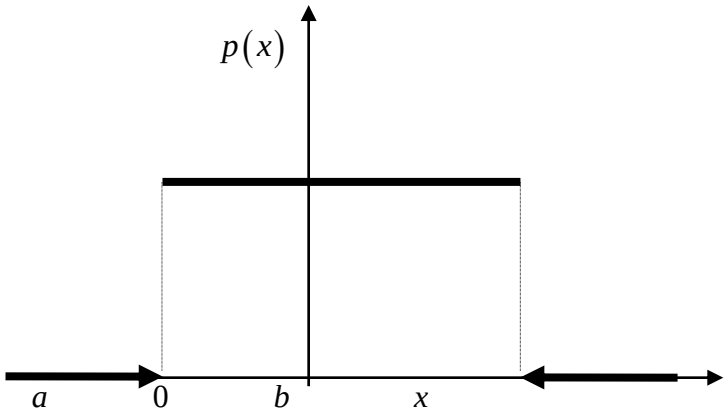
1. ker je $F(x)$ naraščajoča funkcija, je povsod $p(x) \geq 0$,
2. ker je $F(\infty) = 1$, je $\int_{-\infty}^{+\infty} p(x)\,dx = 1$,
3. ker velja $P(x_1 \leq X < x_2) = F(x_2) - F(x_1)$, je $P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} p(x)dx$,
4. zaradi zveze $F(x+0) - F(x) = P(X = x)$ pa velja še $P(X = x) = \int_{x}^{x} p(t)dt = 0$

10.2.2.1 Enakomerna zvezna porazdelitev

Slučajna spremenljivka je enakomerno zvezno porazdeljena, če je njena gostota verjetnosti

$$p(x)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{drugod}\end{cases} \tag{10.18}$$

Grafično:



Ker velja $P(x_1\le X<x_2)=\int_{x_1}^{x_2}p(x)\,dx$, dobimo v posebnem primeru, če je $a\le u<v\le b$, tole povezavo:

$$P(u\le X<v)=\int_{u}^{v}\frac{1}{b-a}\,dx=\frac{v-u}{b-a}$$

10.2.2.2 Normalna (Gaussova) porazdelitev N(a, σ^2)

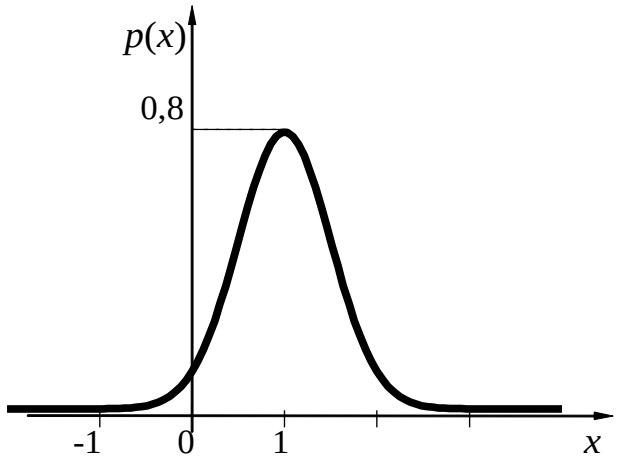
Normalna porazdelitev ima gostoto verjetnosti

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2} \tag{10.19}$$

Porazdelitev (10.19) je odvisna od parametrov *a* in σ. Pri tem je lahko *a* poljubno realno število, σ pa poljubno pozitivno število.

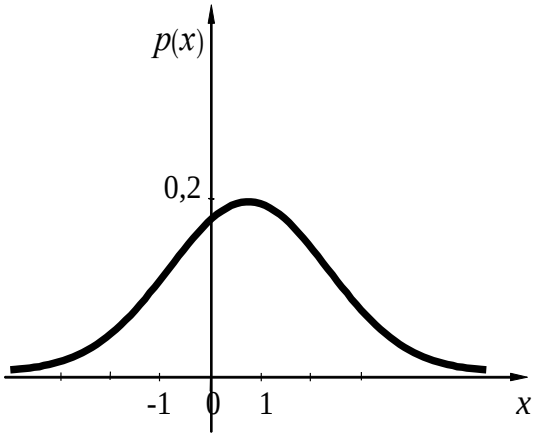
Funkcijo $p(x)$ lahko narišemo. Brez težav z odvajanjem ugotovimo, da ima v točki $E\left(a,\ \frac{1}{\sigma\sqrt{2\pi}}\right)$ maksimum, v točkah $T_1\left(a-\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$ in $T_2\left(a+\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$ pa ima prevoja. Parameter *a* torej določa lego krivulje, parameter σ pa določa njeno obliko (slika 6.2, slika 6.3).

$a = 1,\ \sigma = 0{,}5$



Slika 6.2: Graf funkcije $p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$ za $a = 1$ in $\sigma = 0{,}5$

$a = 1,\ \sigma = 2$



Slika 6.3: Graf funkcije $p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$ za $a = 1$ in $\sigma = 2$

Če vzamemo vrednosti parametrov $a = 0$ in $\sigma = 1$, dobimo tako imenovano ***standardizirano normalno porazdelitev N*(0,1):**

$$p(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{10.20}$$

Standardizirano normalno porazdelitev dobimo iz normalne porazdelitve (10.19), to je $p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$, tako da uvedemo novo slučajno spremenljivko s predpisom

$$z = \frac{x-a}{\sigma} \tag{10.21}$$

Od tod je

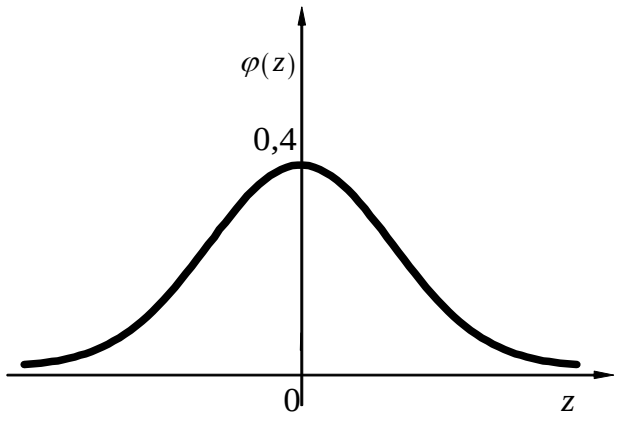
$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \tag{10.22}$$

Opomba: Zakaj je to res, bomo pokazali v primeru pri naslednjem poglavju.

Nekatere lastnosti funkcije $\varphi(z)$:

1. $\varphi(-z) = \varphi(z)$
2. $\varphi(0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}0^2} = \frac{1}{\sqrt{2\pi}} \approx 0,4$
3. $\lim_{z\to\infty} \varphi(z) = 0$
4. $\int_{-\infty}^{\infty} \varphi(x)\, dx = 1$

Graf:

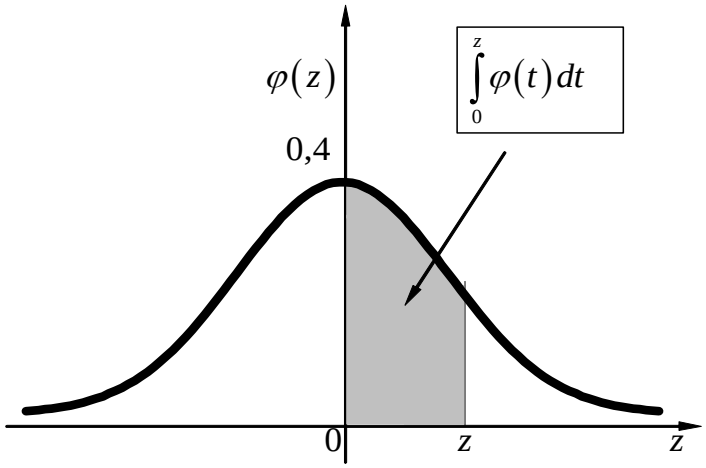


V statistiki je pomembna še funkcija

$$\Phi(z)=\int_0^z \varphi(t)\,dt=\frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}t^2}\,dt$$

ki je – geometrijsko gledano - ploščina lika med krivuljo $\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$ in abscisno osjo na intervalu med 0 in spremenljivim argumentom *z*.

Grafično:



Zaradi sodosti funkcije $\varphi(z)$ je res tole:

$$\int_{-\infty}^{0} \varphi(z)\,dz=\int_0^{\infty} \varphi(z)\,dz=0,5$$

Zato je tudi

$$\int_{-\infty}^{z} \varphi(t)\,dt=0,5+\int_0^z \varphi(t)\,dt=0,5+\Phi(z)$$

Zaradi te lastnosti je dovolj, če poznamo le vrednosti integrala od 0 do $z>0$, to pa je ravno funkcija $\Phi(z)$.

V nadaljevanju sta priloženi tabeli funkcij $\varphi(z)$ in $\Phi(z)$. Tabeli sta dani za vrednosti spremenljivke od 0 do 4,09 na 5 decimalk natančno.

Vidimo, da je vrednost funkcije $\varphi(z)$ za $z=4,09$ zelo blizu nič, zato za $z>4,9$ vzamemo vrednost funkcije kar nič.

Podobno ravnamo pri funkciji $\Phi(z)$. Njena vrednost je za $z = 4,09$ zelo blizu 0,5, zato za $z > 4,9$ vzamemo vrednost funkcije 0,5.

TABELA I **Vrednosti funkcije** $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$

z	0	1	2	3	4	5	6	7	8	9
0.0	0.39890	39892	39886	39876	39862	39844	39822	39797	39767	39733
0.1	0.39695	39654	39608	39559	39505	39448	39387	39322	39253	39181
0.2	0.39104	39024	38940	38853	38762	38667	38568	38466	38361	38251
0.3	0.38139	38023	37903	37780	37654	37524	37391	37255	37115	36973
0.4	0.36827	36678	36526	36371	36213	36053	35889	35723	35553	35381
0.5	0.35207	35029	34849	34667	34482	34294	34105	33912	33718	33521
0.6	0.33322	33121	32918	32713	32506	32297	32086	31874	31659	31443
0.7	0.31225	31006	30785	30563	30339	30114	29887	29659	29431	29200
0.8	0.28969	28737	28504	28269	28034	27798	27562	27324	27086	26848
0.9	0.26609	26369	26129	25888	25647	25406	25164	24923	24681	24439
1.0	0.24197	23955	23713	23471	23230	22988	22747	22506	22265	22025
1.1	0.21785	21546	21307	21069	20831	20594	20357	20121	19886	19652
1.2	0.19419	19186	18954	18724	18494	18265	18037	17810	17585	17360
1.3	0.17137	16915	16694	16474	16256	16038	15822	15608	15395	15183
1.4	0.14973	14764	14556	14350	14146	13943	13742	13542	13344	13147
1.5	0.12952	12758	12566	12376	12188	12001	11816	11632	11450	11270
1.6	0.11092	10915	10741	10567	10396	10226	10059	09893	09728	09566
1.7	0.09405	09246	09089	08933	08780	08628	08478	08329	08183	08038
1.8	0.07895	07754	07614	07477	07341	07206	07074	06943	06814	06687
1.9	0.06562	06438	06316	06195	06077	05959	05844	05730	05618	05508
2.0	0.05399	05292	05186	05082	04980	04879	04780	04682	04586	04491
2.1	0.04398	04307	04217	04128	04041	03955	03871	03788	03706	03626
2.2	0.03547	03470	03394	03319	03246	03174	03103	03034	02965	02898
2.3	0.02833	02768	02705	02643	02582	02522	02463	02406	02349	02294
2.4	0.02239	02186	02134	02083	02033	01984	01936	01888	01842	01797
2.5	0.01753	01709	01667	01625	01585	01545	01506	01468	01431	01394
2.6	0.01358	01323	01289	01256	01223	01191	01160	01130	01100	01071
2.7	0.01042	01014	00987	00961	00935	00909	00885	00861	00837	00814
2.8	0.00792	00770	00748	00727	00707	00687	00668	00649	00631	00613
2.9	0.00595	00578	00562	00545	00530	00514	00499	00485	00470	00457
3.0	0.00443	00430	00417	00405	00393	00381	00370	00358	00348	00337
3.1	0.00327	00317	00307	00298	00288	00279	00271	00262	00254	00246
3.2	0.00238	00231	00224	00216	00210	00203	00196	00190	00184	00178
3.3	0.00172	00167	00161	00156	00151	00146	00141	00136	00132	00127
3.4	0.00123	00119	00115	00111	00107	00104	00100	00097	00094	00090
3.5	0.00087	00084	00081	00079	00076	00073	00071	00068	00066	00063
3.6	0.00061	00059	00057	00055	00053	00051	00049	00047	00046	00044
3.7	0.00042	00041	00039	00038	00037	00035	00034	00033	00031	00030
3.8	0.00029	00028	00027	00026	00025	00024	00023	00022	00021	00021
3.9	0.00020	00019	00018	00018	00017	00016	00016	00015	00014	00014
4.0	0.00013	00013	00012	00012	00011	00011	00011	00010	00010	00009

TABELA II: **Vrednosti funkcije** $\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}t^2}\,dt$

z	0	1	2	3	4	5	6	7	8	9
0.0	0.00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0.1	0.03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0.2	0.07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0.3	0.11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0.4	0.15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0.5	0.19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0.6	0.22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0.7	0.25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0.8	0.28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0.9	0.31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1.0	0.34134	34375	34614	34849	35083	35314	35543	35769	35993	36214
1.1	0.36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1.2	0.38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1.3	0.40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1.4	0.41924	42073	42220	42364	42507	42647	42785	42922	43056	43189
1.5	0.43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1.6	0.44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1.7	0.45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1.8	0.46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1.9	0.47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2.0	0.47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2.1	0.48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2.2	0.48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2.3	0.48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2.4	0.49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2.5	0.49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2.6	0.49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2.7	0.49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2.8	0.49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2.9	0.49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3.0	0.49865	49869	49874	49878	49882	49886	49889	49893	49896	49900
3.1	0.49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3.2	0.49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3.3	0.49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3.4	0.49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3.5	0.49977	49978	49978	49979	49980	49981	49981	49982	49983	49983
3.6	0.49984	49985	49985	49986	49986	49987	49987	49988	49988	49989
3.7	0.49989	49990	49990	49990	49991	49991	49992	49992	49992	49992
3.8	0.49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3.9	0.49995	49995	49996	49996	49996	49996	49996	49996	49997	49997
4.0	0.49997	49997	49997	49997	49997	49997	49998	49998	49998	49998

10.2.2.3 Porazdelitev $\chi^2(n)$

Gostota verjetnosti porazdelitve "*hi kvadrat*" je določena takole:

$$p(x)=\begin{cases}\dfrac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\dfrac{n}{2}\right)} & x>0\\ 0 & x\le 0\end{cases} \qquad (10.23)$$

V (10.23) je $k \in N$ število prostostnih stopenj, $\Gamma(x)$ pa je *Eulerjeva funkcija gama*, določena z integralom

$$\Gamma(x)=\int_0^{\infty} t^{x-1}e^{-t}dt$$

Pomen te porazdelitve je naslednji:

Če imamo m med seboj neodvisnih slučajnih statističnih spremenljivk, porazdeljenih po standardizirani normalni porazdelitvi, potem je vsota kvadratov teh spremenljivk porazdeljena po porazdelitvi hi kvadrat s številom prostostnih stopenj k = m.

Porazdelitev hi kvadrat se z naraščajočo prostostno stopnjo približuje normalni porazdelitvi z matematičnim upanjem $2k$ in standardnim odklonom $\sqrt{2k}$.

Porazdelitev hi kvadrat uporabljamo tako, da nas zanima verjetnost, da slučajna spremenljivka *y*, ki je porazdeljena po porazdelitvi hi kvadrat, zavzame vrednost, večjo od χ_0^2, torej

$$P\left(\chi^2>\chi_0^2\right).$$

Te verjetnosti so za *k* od 1 do 30 tabelirane v nadaljevanju.

Primer:
Iz tabele za $k=20$ in $\alpha=0,05$ preberemo: χ_0^2=31,4104. To pomeni naslednje: verjetnost, da slučajna spremenljivka, ki je porazdeljena po porazdelitvi hi kvadrat s številom prostostnih stopenj $k=20$ zavzame vrednost, večjo od 31,4104, je 0,05, torej $P\left(\chi^2>31,4104\right)=0,05$. Opomba: verjetnosti α pravimo tudi stopnja značilnosti. O tem več kasneje.

Graf in tabela (VIR: http://www.statsoft.com/textbook/distribution-tables/#chi)

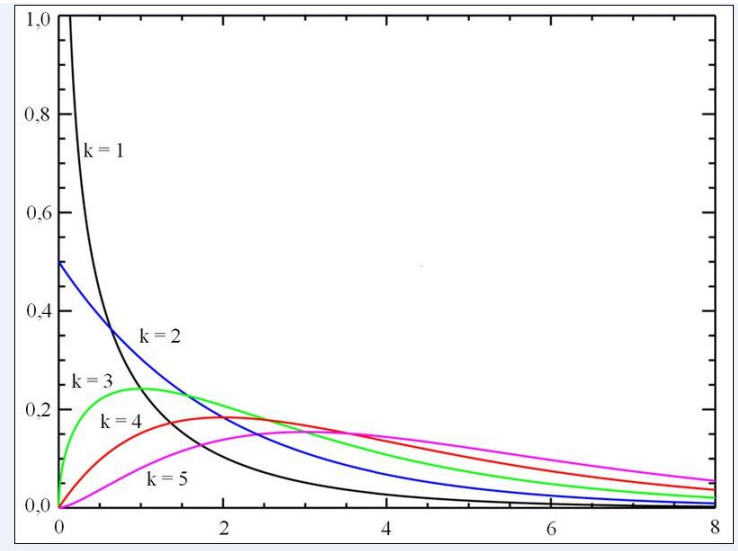


TABELA III: Vrednosti funkcije za porazdelitev hi kvadrat

df\area	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.34840	11.34487
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.14329	13.27670
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.07050	12.83250	15.08627
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.64464	12.59159	14.44938	16.81189
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100	13.70069	17.27501	19.67514	21.92005	24.72497
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032	14.84540	18.54935	21.02607	23.33666	26.21697
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.33976	15.98391	19.81193	22.36203	24.73560	27.68825
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.33850	19.36886	23.54183	26.29623	28.84535	31.99993
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711	30.19101	33.40866
18	6.26480	7.01491	8.23075	9.39046	10.86494	13.67529	17.33790	21.60489	25.98942	28.86930	31.52638	34.80531
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.56200	18.33765	22.71781	27.20357	30.14353	32.85233	36.19087
20	7.43384	8.26040	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043	34.16961	37.56623
21	8.03365	8.89720	10.28290	11.59131	13.23960	16.34438	20.33723	24.93478	29.61509	32.67057	35.47888	38.93217
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444	36.78071	40.28936
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.13730	22.33688	27.14134	32.00690	35.17246	38.07563	41.63840
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503	39.36408	42.97982
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248	40.64647	44.31410
26	11.16024	12.19815	13.84390	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514	41.92317	45.64168
27	11.80759	12.87850	14.57338	16.15140	18.11390	21.74940	26.33634	31.52841	36.74122	40.11327	43.19451	46.96294
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714	44.46079	48.27824
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697	45.72229	49.58788
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297	46.97924	50.89218

10.2.2.4 Studentova t porazdelitev S(n)

Gostota verjetnosti Studentove porazdelitve je določena z izrazom

$$p(x) = \frac{1}{\sqrt{k}\ B\left(\frac{k}{2}, \frac{1}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}} \tag{10.24}$$

V formuli je k število prostostnih stopenj, $B(x,y)$ pa je *Eulerjeva funkcija beta*, določena z integralom

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1}\, dt$$

Studentova *t* porazdelitev in porazdelitev $\chi^2(n)$ imata zelo široko polje uporabe v statistiki.

Studentovo porazdelitev uporabljamo pri vzorčenju tedaj, ko je vzorec majhen, torej ima manj kot 30 statističnih enot.

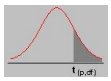
Studentova porazdelitev je (podobno, kot to velja za normalno porazdelitev) simetrična glede na ordinatno os, njena značilnost je, da je bolj sploščena v primerjavi z normalno.

Ko se število enot v vzorcu povečuje, se oblika Studentove porazdelitve približuje obliki standardizirane normalne porazdelitve.

S Studentovo porazdelitvijo izračunamo verjetnost, da je izračunana vrednost Studentove *t* porazdelitve večja od tabelirane vrednosti za $(k-1)$ prostostnih stopenj pri stopnji značilnosti α .

TABELA IV: Vrednosti funkcije za Studentovo t porazdelitev

t table with right tail probabilities



df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460

Vir: http://www.statsoft.com/textbook/distribution-tables/#t

10.3 Številske karakteristike slučajne spremenljivke

Porazdelitve slučajnih spremenljivk so pogosto preširok pojem za konkretno uporabo, zato želimo iz porazdelitvenega zakona najti nekatere osnovne značilnosti/karakteristike, ki jih bomo ocenili s številom. Od tod tudi ime *številske karakteristike*. V nadaljevanju bomo omenili nekatere od njih.

10.3.1 Matematično upanje E(X)

Matematično upanje (angl.: Mathematical expectation, tudi: the expected value) definirajmo posebej za diskretne in posebej za zvezne slučajne spremenljivke.
Opomba: Matematično upanje včasih poimenujemo tudi *povprečna vrednost.*

a) Za *diskretno slučajno spremenljivko*, dano z verjetnostno shemo

$$X:\begin{pmatrix} x_1 & x_2 & . & . & x_n \\ p_1 & p_2 & . & . & p_n \end{pmatrix}$$

je *matematično upanje* določeno z izrazom:

$$E(X)=\sum_{i=1}^{n} x_i\, p_i \tag{10.25}$$

Če ima slučajna spremenljivka neskončno zalogo vrednosti, je

$$E(X)=\sum_{i=1}^{\infty} x_i\, p_i \tag{10.26}$$

vendar le pod pogojem, da je vrsta konvergentna, torej, ko velja $\sum_{i=1}^{\infty} |x_i|\, p_i < \infty$. Če ta pogoj ni izpolnjen, potem za neomejeno diskretno slučajno spremenljivko matematično upanje ne obstaja.

Primer:
Pri streljanju proti tarči zadene strelec v center tarče z verjetnostjo 0,1; dva cm od centra z verjetnostjo 0,3; štiri cm od centra z verjetnostjo 0,3; šest cm od centra z verjetnostjo 0,2 in osem cm od centra z verjetnostjo 0,1. Slučajna spremenljivka *X* naj pomeni oddaljenost

zadetka od središče tarče in lahko zavzame omenjene vrednosti. Podana je s takšno verjetnostno shemo:

$$X:\begin{pmatrix} 0 & 2 & 4 & 6 & 8 \\ 0{,}1 & 0{,}3 & 0{,}3 & 0{,}2 & 0{,}1 \end{pmatrix}$$

Izračunajmo njeno matematično upanje.

Rešitev:
Po definiciji matematičnega upanja je

$$E(X) = \sum_{i=1}^{5} x_i\, p_i = x_1\, p_1 + x_2\, p_2 + x_3\, p_3 + x_4\, p_4 + x_5\, p_5 =$$
$$= 0 \cdot 0{,}1 + 2 \cdot 0{,}3 + 4 \cdot 0{,}3 + 6 \cdot 0{,}2 + 8 \cdot 0{,}1 = 3{,}8$$

Rezultat pomeni, da bo strelec v *povprečju* (zato ime povprečna vrednost) pričakoval zadetek, ki bo 3,8 cm oddaljen od centra.
Opomba: Izraz *matematično upanje* pa izvira od iger na srečo. Če bi bili posamezni rezultati, ki so pri tem poskusu možni, ovrednoteni z nekim denarnim dobitkom, bi povprečna vrednost $E(X)$ dajala njegovo *upanje* na dobitek, ki bi ga osvojil s posameznim strelom.

♦ ♦ ♦

Primer:
Določimo matematično upanje slučajne spremenljivke *X*, porazdeljene po Poissonovem zakonu z verjetnostno funkcijo $p_k = P(X = k) = \dfrac{a^k e^{-a}}{k!}$, $a > 0$, $k = 1, 2, \ldots$.

Rešitev:
Po definiciji je

$$E(X) = \sum_{k=1}^{\infty} k\, p_k = \sum_{k=1}^{\infty} \frac{k a^k}{k!} e^{-a} = \sum_{k=1}^{\infty} \frac{a a^{k-1}}{(k-1)!} e^{-a} = a e^{-a} \sum_{k=1}^{\infty} \frac{a^{k-1}}{(k-1)!}$$

V zgornji vrsti upoštevajmo, da je $\sum_{i=0}^{\infty} \frac{x^i}{i!} = e^x$, zato je $\sum_{k=1}^{\infty} \frac{a^{k-1}}{(k-1)!} = e^a$

in od tod sledi:

$$E(X) = a$$

♦ ♦ ♦

Primer:
Na neki cesti se v povprečju zgodi ena smrtna nesreča na leto. Kakšna je verjetnost, da se v enem letu zgodijo tri nesreče, če se slučajna spremenljivka "na cesti se zgodi nesreča s smrtnim izidom" ravna po Poissonovi porazdelitvi?

Rešitev:
Poissonova porazdelitev, po kateri se ravna slučajna spremenljivka "na cesti se zgodi nesreča s smrtnim izidom", je opisana z verjetnostno funkcijo $p_k = P(X = k) = \frac{a^k e^{-a}}{k!}$.
Ker je za Poissonovo porazdelitev povprečna vrednost (kar smo ugotovili v prejšnjem primeru) $E(X) = a$, preberemo iz podatkov, da je sedaj $a = 1$ in število $k = 3$.

Od tod sledi: $p_3 = P(X = 3) = \frac{1^3 e^{-1}}{3!} \approx 0{,}0613$.

♦ ♦ ♦

Primer:
Slučajna spremenljivka X naj zavzame vrednosti naravnih števil z verjetnostjo $P(X = x_k) = \frac{1}{3} q^k$, $k \in N$. Izračunajmo q in matematično upanje slučajne spremenljivke X!

Rešitev:
$P(X = x_k) = \frac{1}{3}\ q^k\ ,\ k \in N$

Vsota verjetnosti vseh možnih dogodkov je ena: $\sum_{k=1}^{\infty} \frac{1}{3}\ q^k = 1$.

Vsota $\sum_{k=1}^{\infty} q^k$ je za $0 < q < 1$ konvergentna (geometrijska vrsta) in jo izračunamo po znani formuli $\sum_{k=1}^{\infty} q^k = \frac{q}{1-q}$. Od tod dobimo enačbo $\frac{1}{3} \cdot \frac{q}{1-q} = 1 \Rightarrow q = \frac{3}{4}$.

Verjetnostna shema te slučajne spremenljivke je

$$X : \begin{pmatrix} 1 & 2 & \cdots & k & \cdots \\ \frac{1}{4} & \frac{3}{16} & \cdots & \frac{1}{4} \cdot \left(\frac{3}{4}\right)^{k-1} & \cdots \end{pmatrix}$$

Še njeno matematično upanje:

$$E(X) = \sum_{i=1}^{\infty} x_i\ p_i = \frac{1}{4} \sum_{k=1}^{\infty} k \left(\frac{3}{4}\right)^{k-1}$$

♦ ♦ ♦

b) Pri *zvezno porazdeljeni slučajni spremenljivki* je matematično upanje definirano z izrazom

$$E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx \tag{10.27}$$

Matematično upanje obstaja le tedaj, ko je integral v (10.27) absolutno integrabilen, torej ko velja: $E(X) = \int_{-\infty}^{\infty} |x|\, p(x)\, dx < \infty$.

Primer:
Slučajna spremenljivka naj bo porazdeljena po enakomerno zvezni porazdelitvi. Izračunajmo njeno matematično upanje.

Rešitev:
Slučajna spremenljivka je enakomerno zvezno porazdeljena, če je njena gostota verjetnosti dana z izrazom (10.18):

$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{drugod} \end{cases}$$

Matematično upanje takšne slučajne spremenljivke je po formuli (10.27)

$$E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_{a}^{b} x \frac{1}{b-a}\, dx = \frac{1}{b-a} \frac{x^2}{2}\Big|_a^b = \frac{a+b}{2}$$

Opomba: Matematično upanje enakomerno zvezno porazdeljene slučajne spremenljivke je kar aritmetično povprečje obeh meja intervala $[a,b]$.

♦ ♦ ♦

Primer:
Izračunajmo matematično upanje zvezno porazdeljene slučajne spremenljivke, če je njena gostota verjetnosti $p(x) = \dfrac{1}{\pi\sqrt{a^2 - x^2}}$, $-a < x < a$.

Rešitev:
Po definiciji matematičnega upanja za zvezno porazdeljeno slučajno spremenljivko je

$$E(x) = \int_{-a}^{a} x\, p(x)\, dx = \frac{1}{\pi} \int_{-a}^{a} \frac{x}{\sqrt{a^2 - x^2}}\, dx = 0$$

Opomba:

Nedoločeni integral $I = \int \frac{x}{\sqrt{a^2 - x^2}}\, dx$ najenostavneje rešimo s substitucijo $t^2 = a^2 - x^2$.

♦ ♦ ♦

10.3.2 Disperzija ali varianca D(X)

Disperzija ali varianca slučajne spremenljivke *X* je definirana kot matematično upanje kvadriranih odklonov slučajne spremenljivke od njenega matematičnega upanja.

Z disperzijo torej merimo razpršenost slučajne spremenljivke okoli njenega povprečja.

Razlika med nekaterimi vrednostmi slučajne spremenljivke in matematičnim upanjem je lahko tudi negativna, zato te razlike kvadriramo, tako da imamo opravka le s pozitivnimi vrednostmi.
Do disperzije torej pridemo tako, da za slučajno spremenljivko izračunamo njeno matematično upanje $E(X)$ ter nato zapišemo novo slučajno spremenljivko $\left(X - E(X)\right)^2$.

Matematično upanje te slučajne spremenljivke je *disperzija* ali *varianca*:

$$D(X) = E\left(X - E(X)\right)^2 \qquad (10.28)$$

Gornji izraz je splošna definicija disperzije za poljubno slučajno spremenljivko, ne glede na to ali gre za diskretno ali za zvezno slučajno spremenljivko.

Pri ***diskretno porazdeljeni slučajni spremenljivki***, ki je dana z verjetnostno shemo

$$X: \begin{pmatrix} x_1 & x_2 & . & . & x_n \\ p_1 & p_2 & . & . & p_n \end{pmatrix}$$

izračunamo matematično upanje $E(X)$ ter nato zapišemo novo slučajno spremenljivko $\left(X - E(X)\right)^2$:

$$\left(X-E(X)\right)^2:\begin{pmatrix}\left(x_1-E(X)\right)^2 & \left(x_2-E(X)\right)^2 & . & . & \left(x_n-E(X)\right)^2\\ p_1 & p_2 & . & . & p_n\end{pmatrix}$$

in od tod

$$D(X)=\sum_i\left(x_i-E(X)\right)^2 p_i \tag{10.29}$$

Pri ***zvezno porazdeljeni slučajni spremenljivki*** pa je disperzija določena z izrazom

$$D(X)=\int_{-\infty}^{\infty}\left(x-E(X)\right)^2 p(x)\,dx \tag{10.30}$$

Če v formulah (10.29) in (10.30) dvočlenik kvadriramo in upoštevamo, da je $E(X)$ konstanta ter da velja še $\sum_i x_i^2 p_i = E(X^2),\ \sum_i p_i = 1$ za diskretne in $\int_{-\infty}^{\infty} x^2 p(x)dx = E(X^2),\ \int_{-\infty}^{\infty} p(x)dx = 1$ za zvezne slučajne spremenljivke, dobimo:

a) za diskretno slučajno spremenljivko

$$\begin{aligned}D(X) &= E\left(X-E(X)\right)^2 = \sum_i\left(x_i-E(X)\right)^2 p_i = \\ &= \sum_i x_i^2 p_i - 2E(X)\sum_i x_i p_i + E^2(X)\sum_i p_i = \\ &= E(X^2)-2E(X)E(X)+E^2(X) = E(X^2)-E^2(X)\end{aligned}$$

b) za zvezno slučajno spremenljivko

$$\begin{aligned}D(X) &= \int_{-\infty}^{\infty}\left(x-E(X)\right)^2 p(x)\,dx = \\ &= \int_{-\infty}^{\infty} x^2 p(x)\,dx - 2E(X)\int_{-\infty}^{\infty} x\,p(x)\,dx - E^2(X)\int_{-\infty}^{\infty} p(x)\,dx = E(X^2)-E^2(X)\end{aligned}$$

Ugotovili smo, da *vedno* velja

$$D(X)=E(X^2)-E^2(X) \tag{10.31}$$

Formula (10.31) je za računanje disperzije velikokrat preprostejša možnost, kot računanje disperzije po definiciji (10.28).

Primer:
Diskretna slučajna spremenljivka X je podana z verjetnostno shemo

$$X:\begin{pmatrix} -2 & -1 & 0 & 1 & 3 \\ 0,1 & \alpha & 0,4 & 0,2 & 0,1 \end{pmatrix}$$

Določimo:

a) število alfa,
b) matematično upanje slučajne spremenljivke in
c) njeno disperzijo.

Rešitev:

a) Ker je $\sum_{i=1}^{5} p_i = 1 \Rightarrow 0,1 + \alpha + 0,4 + 0,2 + 0,1 = 1 \Rightarrow \alpha = 0,2$.

b) $E(X) = \sum_{i=1}^{5} x_i\, p_i = (-2)\cdot 0,1 + (-1)\cdot 0,2 + 0\cdot 0,4 + 1\cdot 0,2 + 3\cdot 0,1 = 0,1$

c) Disperzijo pa lahko izračunamo na dva načina, bodisi po formuli (10.28) ali po formuli (10.31).

V tem primeru jo izračunajmo na oba načina.

Slučajna spremenljivka $X - E(X)^2$ bo zaradi $E(X) = 0,1$ imela verjetnostno shemo:

$$(X - E(X))^2 : \begin{pmatrix} (-2-0,1)^2 & (-1-0,1)^2 & (0-0,1)^2 & (1-0,1)^2 & (3-0,1)^2 \\ 0,1 & 0,2 & 0,4 & 0,2 & 0,1 \end{pmatrix}$$

oziroma

$$(X - E(X))^2 : \begin{pmatrix} 4,41 & 1,21 & 0,01 & 0,81 & 8,41 \\ 0,1 & 0,2 & 0,4 & 0,2 & 0,1 \end{pmatrix}$$

Od tod dobimo

$$D(X) = E\left[(X - E(X))^2\right] = 4,41\cdot 0,1 + 1,21\cdot 0,2 + 0,01\cdot 0,4 + 0,81\cdot 0,2 + 8,41\cdot 0,1 = 1,69$$

Sedaj izračunajmo disperzijo še po formuli (10.31). Porazdelitvena shema slučajne spremenljivke X^2 je

$$X^2: \begin{pmatrix} 4 & 1 & 0 & 1 & 9 \\ 0{,}1 & 0{,}2 & 0{,}4 & 0{,}2 & 0{,}1 \end{pmatrix}$$

Njeno matematično upanje dobimo kot običajno

$$E\left(X^2\right) = 4 \cdot 0{,}1 \; + \; 1 \cdot 0{,}2 \; + \; 0 \cdot 0{,}4 \; + \; 1 \cdot 0{,}2 \; + \; 9 \cdot 0{,}1 \; = \; 1{,}7$$

Od tod:

$$D(X) = E\left(X^2\right) - \left(E(X)\right)^2 = 1{,}7 - 0{,}1^2 = 1{,}69$$

♦ ♦ ♦

Primer:
Izračunajmo disperzijo slučajne spremenljivke X, porazdeljene po Poissonovem zakonu z verjetnostno funkcijo $p_k = P\left(X = k\right) = \dfrac{a^k e^{-a}}{k!}$, $a > 0$, $k = 1, 2, \ldots$.

Rešitev:
V enem od prejšnjih primerov smo že ugotovili matematično upanje za slučajno spremenljivk, porazdeljeno po Poissonovem zakonu: $E(X) = a$.
Potrebujemo še $E\left(X^2\right)$.

$$E\left(X^2\right) = \sum_{k=1}^{\infty} k^2 \frac{a^k e^{-a}}{k!} = \sum_{k=1}^{\infty} k \frac{a.a^{k-1} e^{-a}}{(k-1)!} = a e^{-a} \sum_{i=0}^{\infty} (i+1) \frac{a^i}{i!} = a e^{-a} \left(a e^a + e^a\right) = a^2 + a$$

Od tod: $D(X) = E\left(X^2\right) - E^2(X) = a^2 + a - a^2 = a$.

Opomba:

$$\sum_{i=0}^{\infty} \frac{i \cdot a^i}{i!} = \sum_{i=1}^{\infty} \frac{a \cdot a^{i-1}}{(i-1)!} = a \sum_{k=0}^{\infty} \frac{a^k}{k!} = a \cdot e^a$$

♦ ♦ ♦

Primer:
Na bencinski postaji je 10 črpalk, od katerih so tri stalno v okvari (to pomeni, da je verjetnost, da je posamezna črpalka pokvarjena, enaka 0,3). Na prazno bencinsko črpalko pripeljejo 3 vozila in se naključno postavijo pred črpalke. Zapišimo verjetnostno funkcijo ter izračunajmo matematično upanje in disperzijo za število vozil, ki se postavijo pred črpalko v okvari!

Rešitev:
Označimo s p verjetnost, da je posamezna črpalka pokvarjena ($p = 0{,}3$) in q verjetnost, da črpalka ni pokvarjena ($q = 0{,}7$). Ker je potrebno zapisati verjetnostno funkcijo ter

izračunati matematično upanje in disperzijo za število vozil, ki se postavijo pred črpalko v okvari, naj bo slučajna spremenljivka X število avtomobilov pred pokvarjenimi črpalkami.

Za 3 avtomobile bo verjetnostna shema:

$$^{(3)}X:\begin{pmatrix} 0 & 1 & 2 & 3 \\ p_0 & p_1 & p_2 & p_3 \end{pmatrix}$$

Verjetnosti izračunamo po Bernoullijevi formuli $P_n(k)=\binom{n}{k}p^k q^{n-k}$, $k=0,1,2,3;\ n=3$, tako da je

$$^{(3)}X:\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0{,}343 & 0{,}441 & 0{,}189 & 0{,}027 \end{pmatrix}$$

Od tod je matematično upanje $E\left(^{(3)}X\right)=0{,}9$. To pomeni, da v povprečju pričakujemo, da se bo od treh vozil eno postavilo pred pokvarjeno črpalko.

Disperzijo bomo dobili po formuli (10.31). Najprej potrebujemo $E\left(^{(3)}X^2\right)$. Ker je

$$^{(3)}X^2:\begin{pmatrix} 0 & 1 & 4 & 9 \\ 0{,}343 & 0{,}441 & 0{,}189 & 0{,}027 \end{pmatrix}$$

je $E\left(^{(3)}X^2\right)=1{,}440$. Od tod dobimo: $D\left(^{(3)}X\right)=E\left(^{(3)}X^2\right)-E^2\left(^{(3)}X\right)=0{,}63$.

♦ ♦ ♦

Primer:
Vzdrževalci električnega omrežja na terenu potrebujejo za svoje delo 6 kamionov. Posamezen kamion je lahko tedaj, ko ga potrebujejo, dober ali pokvarjen. Verjetnost, da je dober, je 0,8, verjetnost, da je pokvarjen, je 0,2. Izračunajmo povprečno število nepokvarjenih kamionov!

Rešitev:
Slučajna spremenljivka X naj bo število dobrih kamionov v nekem trenutku. Ta spremenljivka lahko zavzame vrednosti 0, 1, 2, 3, 4, 5 ali 6. Verjetnosti posameznih dogodkov spet dobimo po Bernoullijevi formuli

$$P_6(k)=\binom{6}{k}\cdot 0{,}8^k\cdot 0{,}2^{6-k} \text{ za } k = 0, 1, 2, 3, 4, 5, 6$$

$$X:\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 0{,}002 & 0{,}015 & 0{,}082 & 0{,}246 & 0{,}393 & 0{,}262 \end{pmatrix}$$

Od tod: $E(X)=4{,}8$ in $D(X)=0{,}96$.

♦ ♦ ♦

Primer:
Za zvezno porazdeljeno slučajno spremenljivko z gostoto verjetnosti $p(x)=\frac{1}{\pi\sqrt{a^2-x^2}}$, $-a<x<a$ smo že izračunali, da je njeno matematično upanje $E(X)=0$. Izračunajmo še njeno disperzijo!

Rešitev:

$$D(x)=\int_{-a}^{a}\left(x-E(x)\right)^2 p(x)\,dx=\int_{-a}^{a}(x-0)^2\frac{1}{\pi\sqrt{a^2-x^2}}\,dx=\frac{1}{\pi}\int_{-a}^{a}\frac{x^2}{\sqrt{a^2-x^2}}\,dx=\frac{a^2}{2}$$

Opomba: $\int\frac{x^2}{\sqrt{a^2-x^2}}\,dx=-\frac{x}{2}\sqrt{a^2-x^2}+\frac{a^2}{2}\arcsin\frac{x}{a}+C$

♦ ♦ ♦

Primer:
Gostota verjetnosti zvezno porazdeljene slučajne spremenljivke X naj bo dana s formulo $p(x)=\frac{2}{\pi}\cos^2 x$ za $x\in\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$. Določimo matematično upanje in disperzijo te slučajne spremenljivke!

Rešitev:

$$E(X)=\int_{-\infty}^{\infty}xp(x)dx=\frac{2}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}x\cos^2 x dx=0$$

$$D(X)=\int_{-\infty}^{\infty}\left(x-E(X)\right)^2 p(x)dx=\frac{2}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}x^2\cos^2 x\,dx=\frac{\pi^2+6}{12}$$

♦ ♦ ♦

Primer:
Porazdelitvena funkcija slučajne spremenljivke X je

$$F(x)=\begin{cases}1-\frac{a^3}{x^3} & \text{za } x\geq a>0\\ 0 & \text{za } x<a\end{cases}$$

Izračunajmo matematično upanje in disperzijo te slučajne spremenljivke.

Rešitev:
Po formuli (10.17) je $p(x) = F'(x)$, zato je sedaj

$$p(x)=\begin{cases}\dfrac{3a^3}{x^4} & x\geq a>0\\ 0 & x<a\end{cases}$$

$$E(X)=\int_{-\infty}^{\infty} xp(x)dx=3a^3\int_{a}^{\infty} x^{-3}dx=\frac{3a}{2}$$

$$E(X^2)=\int_{-\infty}^{\infty} x^2 p(x)dx=3a^3\int_{a}^{\infty} x^{-2}dx=3a^2$$

in končno še

$$D(X)=E(X^2)-E^2(X)=\frac{3a^2}{4}$$

♦ ♦ ♦

Primer:
Za slučajno spremenljivko *X*, porazdeljeno po normalni (Gaussovi) porazdelitvi, ki ima gostoto verjetnosti $p(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$ izračunajmo matematično upanje in disperzijo.

Rešitev:
a) Matematično upanje bomo izračunali po formuli (10.27).

$$E(X)=\int_{-\infty}^{\infty} x\,p(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} xe^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}dx$$

Integral

$$I=\int_{-\infty}^{+\infty} xe^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}dx$$

izračunamo s substitucijo $t=\dfrac{x-a}{\sigma}$ oziroma $x=a+\sigma t\Rightarrow dx=\sigma dt$.

Za spremenljivko *x* sta meji $-\infty$ in ∞, za novo spremenljivko $t=\dfrac{x-a}{\sigma}$ se ti meji ne spremenita, torej ostajata $-\infty$ in ∞.

Integral se tako glasi:

$$I = \int_{-\infty}^{+\infty} xe^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2} dx = a\sigma \int_{-\infty}^{+\infty} te^{-\frac{1}{2}t^2} dt + \sigma^2 \int_{-\infty}^{+\infty} te^{-\frac{1}{2}t^2} dt$$

Iz matematičnega priročnika preberemo

$$\int_{0}^{+\infty} e^{-\frac{1}{2}t^2} dt = \sqrt{\frac{\pi}{2}}$$

Od tod vidimo, da je $\int_{-\infty}^{+\infty} e^{-\frac{1}{2}t^2} dt = 2\sqrt{\frac{\pi}{2}} = \sqrt{2\pi}$.

Ker je še

$$\int_{-\infty}^{+\infty} te^{-\frac{1}{2}t^2} dt = 0$$

dobimo

$$I = a\sigma\sqrt{2\pi} + \sigma^2 \cdot 0 = a\sigma\sqrt{2\pi}$$

in od tod

$$E(X) = a$$

b) Disperzijo bomo izračunali po formuli (10.30):

$$D(X) = \int_{-\infty}^{+\infty} \left[x - \left(E(X)\right)\right]^2 p(x)\, dx$$

Vstavimo pod integralski znak $E(X) = a$ ter $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$ in dobimo integral

$$D(X) = \int_{-\infty}^{+\infty} \left[x - \left(E(X)\right)\right]^2 p(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-a)^2 e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2} dx$$

V integralu $I = \int_{-\infty}^{+\infty} (x-a)^2 e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2} dx$

najprej uvedemo novo spremenljivko

$$t = \frac{x-a}{\sigma} \Rightarrow x = a + \sigma\, t \Rightarrow dx = \sigma\, dt$$

Meji se ohranita, tako da se integral glasi

$$I = \sigma^3 \int_{-\infty}^{+\infty} t^2 e^{-\frac{1}{2}t^2}\, dt$$

Tako zapisan integral bomo rešili z metodo per partes. V ta namen vzamemo

$$u = t \Rightarrow du = dt$$
$$dv = te^{-\frac{1}{2}t^2}\, dt \Rightarrow v = -e^{-\frac{1}{2}t^2}$$

Od tod je

$$I = \sigma^3 \int_{-\infty}^{+\infty} t^2 e^{-\frac{1}{2}t^2}\, dt = \sigma^3 \left(\left[-te^{-\frac{1}{2}t^2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2}\, dt \right)$$

Oba integrala v oklepaju smo spoznali že pri računanju matematičnega upanja, tako da je

$$I = \sigma^3 \int_{-\infty}^{+\infty} t^2 e^{-\frac{1}{2}t^2}\, dt = \sigma^3 \left(\left[-te^{-\frac{1}{2}t^2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2}\, dt \right) = \sigma^3 \sqrt{2\pi}$$

in iskana disperzija

$$D(X) = \frac{1}{\sigma\sqrt{2\pi}} I = \frac{\sigma^3\sqrt{2\pi}}{\sigma\sqrt{2\pi}} = \sigma^2$$

Pri normalno porazdeljeni slučajni spremenljivki je torej: $E(X) = a$ in $D(X) = \sigma^2$.

♦ ♦ ♦

10.3.3 Standardna deviacija

Ker je disperzija številska karakteristika, ki jo računamo iz *kvadriranih* odklonov od povprečja, lahko dobijo veliki odkloni prevelik vpliv. Da ta prevelik vpliv vsaj delno omejimo, vzamemo za novo mero razpršenosti slučajne spremenljivke le (pozitivni) kvadratni koren iz disperzije. Tako dobljeni meri pravimo *standardna deviacija* (*standardni odklon*), ki jo označimo $\sigma(X)$ ali σ_x. Velja torej:

$$\boxed{\sigma(X) = \sigma_x = +\sqrt{D(X)}} \qquad (10.32)$$

Primer:
Za zvezno slučajno spremenljivko *X*, porazdeljeno po normalni porazdelitvi z gostoto verjetnosti $p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$, smo ugotovili, da je njena disperzija $D(X)=\sigma^2$. Njena standardna deviacija je $\sigma(X)=\sigma$.

♦♦♦

10.3.4 Lastnosti matematičnega upanja in disperzije

Matematično upanje in disperzija imata pomembno vlogo v statistiki. Oglejmo si nekatere lastnosti obeh karakteristik.

Izrek 1: *Če je a poljubna konstanta in* $P(X=a)=1$, *velja* $E(X)=a$.

Dokaz:
Slučajna spremenljivka *X* je lahko diskretna in zato lahko napišemo po definiciji matematičnega upanja (6.25):

$$E(X)=\sum_{i=1}^{n}x_i p_i = a.\sum_{i=1}^{n}p_i = a\cdot 1 = a.$$

S tem je izrek dokazan.

Izrek 2: *Če za slučajni spremenljivki X in Y obstajata njuni matematični upanji* $E(X)$ *in* $E(Y)$, *potem obstaja tudi matematično upanje njune vsote in velja relacija*

$$\boxed{E(X+Y)=E(X)+E(Y)} \qquad (10.33)$$

Dokaz:
Naj bosta slučajni spremenljivki diskretni in s končno zalogo vrednosti. Njuni verjetnostni shemi sta

$$X=\begin{pmatrix}x_1 & x_2 & \cdots & x_m\\ p_1 & p_2 & \cdots & p_m\end{pmatrix} \text{ in } Y=\begin{pmatrix}y_1 & y_2 & \cdots & y_n\\ \tilde{p}_1 & \tilde{p}_2 & \cdots & \tilde{p}_n\end{pmatrix}$$

Označimo verjetnost $p_{ij}=P(X=x_i, Y=y_j)$ za vsak $i=1,2,\ldots,m$ in $j=1,2,\ldots,n$.
Slučajna spremenljivka $X+Y$ ima verjetnostno shemo

$$X+Y=\begin{pmatrix} x_1+y_1 & x_1+y_2 & \cdots & x_1+y_j & \cdots & x_m+y_n \\ p_{11} & p_{12} & \cdots & p_{ij} & \cdots & p_{mn} \end{pmatrix}$$

Njeno matematično upanje je

$$E(X+Y)=\sum_{i=1}^{m}\sum_{j=1}^{n}(x_i+y_j)p_{ij}=\sum_{i=1}^{m}x_i\sum_{j=1}^{n}p_{ij}+\sum_{j=1}^{n}y_j\sum_{i=1}^{m}p_{ij}$$

Pri vsakem $i=1,2,\ldots,m$ je

$$\sum_{j=1}^{n}p_{ij}=\sum_{j=1}^{n}P(X=x_i,Y=y_j)=P(X=x_i)\sum_{j=1}^{n}P(Y=y_j\,/\,X=x_i)=P(X=x_i)=p_i$$

Podobno je vsak $j=1,2,\ldots,n$ tudi $\sum_{i=1}^{m}p_{ij}=\tilde{p}_j$.

Zato je

$$E(X+Y)=\sum_{i=1}^{m}x_ip_i+\sum_{j=1}^{n}y_i\tilde{p}_j=E(X)+E(Y)$$

Dokaz je ravno tak tudi za diskretni slučajni spremenljivki z neskončno zalogo vrednosti, podoben pa je tudi za zvezno porazdeljeni slučajni spremenljivki.
S tem je izrek dokazan.

Izrek 2 je mogoče posplošiti. S popolno indukcijo ugotovimo: Če imamo slučajne spremenljivke $X_1,X_2,\ldots,X_n$, velja

$$\boxed{E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)} \qquad (10.34)$$

Primer:
Slučajno spremenljivko *X*, porazdeljeno po binomskem zakonu $b(n,p)$, smemo izraziti kot vsoto končnega števila slučajnih spremenljivk $X_k:\begin{pmatrix}1 & 0\\ p & q\end{pmatrix}$, $k=1,\ 2,\ \ldots,\ n$, torej $X=X_1+X_2+\cdots+X_n$. Izračunajmo njeno matematično upanje.

Rešitev:
Ker je $X_k:\begin{pmatrix}1 & 0\\ p & q\end{pmatrix}$, je $X_k^2:\begin{pmatrix}1 & 0\\ p & q\end{pmatrix}$, $p+q=1$.

Za *posamezno* slučajno spremenljivko X_k, $k=1,\ 2,\ \ldots,\ n$ je potem res:

$$E(X_k)=1\cdot p+0\cdot q=p\ ,\ E(X_k^2)=1\cdot p+0\cdot q=p$$
$$D(X_k)=E(X^2)-E^2(X)=p-p^2=p(1-p)=pq$$

Za slučajno spremenljivko X je potem

$$E(X)=E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)=nE(X_k)=np$$

♦ ♦ ♦

Izrek 3: *Če ima slučajna spremenljivka X matematično upanje* $E(X)$ *in je a poljubno realno število, ima matematično upanje tudi slučajna spremenljivka* aX *in velja relacija*

$$\boxed{E(aX)=a\,E(X)} \qquad (10.35)$$

Dokaz: Za diskretno slučajno spremenljivko je po definiciji $E(X)=\sum_{i=1}^{n} x_i\ p_i$, zato je $E(aX)=\sum_{i=1}^{n}(ax_i)\,p_i=a\sum_{i=1}^{n}x_i\ p_i=a\,E(X)$.

Za zvezno slučajno spremenljivko je po definiciji $E(X)=\int_{-\infty}^{\infty} x\,p(x)\,dx$, zato je zanjo $E(aX)=\int_{-\infty}^{\infty}(a\,x)\,p(x)\,dx=a\int_{-\infty}^{\infty}x\,p(x)\,dx=a\,E(X)$.

S tem je izrek dokazan.

Iz izrekov 2 in 3 sledi še tale izrek.

Izrek 4: *Če za slučajni spremenljivki X in Y obstajata njuni matematični upanji* $E(X)$ *in* $E(Y)$ *in sta* c_1 *in* c_2 *poljubni realni števili, potem velja*

$$\boxed{E(c_1X+c_2Y)=c_1E(X)+c_2E(Y)} \qquad (10.36)$$

Tudi formulo (10.36) lahko posplošimo: če obstajajo matematična upanja $E(X_1), E(X_2), ..., E(X_n)$ slučajnih spremenljivk X_1 , X_2 , ... X_n in so $c_1, c_2, \ldots, c_n$ poljubna realna števila, ima matematično upanje tudi slučajna spremenljivka $c_1X_1+c_2X_2+\cdots+c_nX_n$ in velja relacija

$$E(c_1X_1+c_2X_2+\cdots+c_nX_n)=c_1E(X_1)+c_2E(X_2)+\cdots+c_nE(X_n) \tag{10.37}$$

Izrek 5: *Za vsako slučajno spremenljivko X, ki ima matematično upanje, velja*

$$E\left[E(X)\right]=E(X) \tag{10.38}$$

Dokaz:

Po definiciji matematičnega upanja velja $E(X)=\sum_{i=1}^{n} x_i\, p_i$ za diskretne in

$E(X)=\int_{-\infty}^{\infty} x\, p(x)\,dx$ za zvezno porazdeljene slučajne spremenljivke.

Ko je slučajna spremenljivka kar enaka svojemu matematičnemu upanju, je

$$E\big(E(X)\big)=\sum_{i=1}^{n} E(X)\,p_i = E(X)\sum_{i=1}^{n} p_i = E(X).1 = E(X)$$

in $$E\big(E(X)\big)=\int_{-\infty}^{\infty} E(X)\,p(x)\,dx = E(X)\int_{-\infty}^{\infty} p(x)\,dx = E(X).1 = E(X)$$

S tem je izrek dokazan.

Izrek 6: *Za vsako slučajno spremenljivko X, ki ima matematično upanje, velja*

$$E\ X-E(X)\ =0 \tag{10.39}$$

Dokaz: Po formuli (10.36) je $E\ X-E(X)\ =E(X)-E\big(E(X)\big)$. Po (10.38) je $E\big(E(X)\big)=E(X)$, zato je res tudi: $E\left[X-E(X)\right]=E(X)-E\big(E(X)\big)=E(X)-E(X)=0$.

S tem je izrek dokazan.

Izrek 7: *Če sta slučajni spremenljivki X in Y med seboj neodvisni in imata matematično upanje, ima matematično upanje tudi slučajna spremenljivka XY in velja relacija*

$$E(XY)=E(X)E(Y) \tag{10.40}$$

Dokaz:
Naj bosta slučajni spremenljivki dani z verjetnostnima shemama

$$X=\begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ p_1 & p_2 & \cdots & p_m \end{pmatrix} \text{ in}$$

$$Y=\begin{pmatrix} y_1 & y_2 & \cdots & y_n \\ \tilde{p}_1 & \tilde{p}_2 & \cdots & \tilde{p}_n \end{pmatrix}$$

Označimo verjetnost $p_{ij}=P\left(X=x_i, Y=y_j\right)$ za vsak $i=1,2,\ldots,m$ in $j=1,2,\ldots,n$. Slučajni spremenljivki sta po predpostavki med seboj neodvisni, zato velja $p_{ij}=P\left(X=x_i, Y=y_j\right)=P\left(X=x_i\right)P\left(Y=y_j\right)$. Za vsak $i=1,2,\ldots,m$ in $j=1,2,\ldots,n$ je torej $p_{ij}=p_i\tilde{p}_j$.

Slučajna spremenljivka XY ima verjetnostno shemo

$$XY=\begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_iy_j & \cdots & x_my_n \\ p_1\tilde{p}_1 & p_1\tilde{p}_2 & \cdots & p_i\tilde{p}_j & \cdots & p_m\tilde{p}_n \end{pmatrix}$$

Njeno matematično upanje je

$$E(XY)=\sum_{i=1}^{m}\sum_{j=1}^{n}x_iy_jp_i\tilde{p}_j=\left(\sum_{i=1}^{m}x_ip_i\right)\left(\sum_{j=1}^{n}y_j\tilde{p}_j\right)=E(X)E(Y)$$

Podobno bi dokazali izrek tudi za diskretne spremenljivke z neskončno zalogo vrednosti in za zvezno porazdeljene slučajne spremenljivke.
S tem je izrek dokazan.

Tudi izrek 7 je mogoče posplošiti. S popolno indukcijo bi ugotovili: če so slučajne spremenljivke $X_1, X_2, \ldots, X_n$ med seboj neodvisne in imajo matematično upanje, velja

$$\boxed{E(X_1X_2\ldots X_n)=E(X_1)E(X_2)\ldots E(X_n)} \qquad (10.41)$$

Izreka 7 pa ne moremo obrniti. *Če za slučajni spremenljivki X in Y velja relacija (10.40), to ne pomeni nujno, da sta spremenljivki med seboj neodvisni.* Takšna relacija lahko včasih velja tudi za odvisni spremenljivki.

Definicija: *Slučajni spremenljivki X in Y imenujemo korelirani, če velja* $E(XY)\neq E(X)E(Y)$. *Slučajni spremenljivki X in Y imenujemo nekorelirani, če velja* $E(XY)=E(X)E(Y)$.

Definicija: Izraz

$$K(X,Y) = E\left(X - E(X)\right)\left(Y - E(Y)\right) \tag{10.42}$$

imenujemo *kovarianca* med slučajnima spremenljivkama *X* in *Y*.

Kovarianco lahko zapišemo tudi drugače. Pomnožimo dvočlenika v oglatem oklepaju in imamo

$$K(X,Y) = E\left(X - E(X)\right)\left(Y - E(Y)\right) = E\ XY - E(X)Y - E(Y)X + E(X)E(Y)$$

Upoštevajmo, da za matematično upanje velja lastnost linearnosti

$$K(X,Y)=E(XY)-E\big(E(X)Y\big)-E\big(E(Y)X\big)+E\big(E(X)E(Y)\big)$$

Zaradi (10.38) in (10.40) dobimo za kovarianco pripravnejšo formulo

$$K(X,Y) = E(XY) - E(X)E(Y) \tag{10.43}$$

Iz formule (10.43) vidimo, da sta slučajni spremenljivki *X* in *Y* nekorelirani natanko tedaj, ko je $K(X,Y) = 0$.

Izrek 8: *Če je a poljubna konstanta in* $P(X = a) = 1$, *velja* $D(X) = 0$

Dokaz:
Zaradi izreka 1 je v tem primeru $E(X) = a$ in disperzija je res enaka nič.

$D(X) = E\left(X - E(X)\right)^2 = E\left((a-a)^2\right) = E(0) = 0$.

S tem je izrek dokazan.

Izrek 9: *Če za slučajno spremenljivko X obstaja njena disperzija in je a poljubna realna konstanta, velja*

$$D(aX) = a^2\ D(X) \tag{10.44}$$

Dokaz:
Za disperzijo lahko uporabimo formulo $D(X) = E(X^2) - E^2(X)$, torej je v tem primeru $D(aX) = E(a^2X^2) - E^2(aX) = a^2E(X^2) - a^2E^2(X)$.
Izpostavimo a^2 in dobimo $D(aX) = a^2\ E(X^2) - E^2(X)\ = a^2D(X)$.
S tem je izrek dokazan.

Izrek 10: *Če imata slučajni spremenljivki X in Y disperzijo, jo ima tudi njuna vsota in velja*

$$D(X+Y) = D(X) + D(Y) + 2K(X,Y) \tag{10.45}$$

Dokaz:
Neposredno iz definicije disperzije sledi

$$\begin{aligned} D(X+Y) &= E\left((X+Y) - E(X+Y)\right)^2 = E\left(\left(X - E(Y)\right)^2 + \left(Y - E(Y)\right)^2\right) = \\ &= E\left(X - E(X)\right)^2 + E\left(Y - E(Y)\right)^2 + 2E\left(X - E(X)\right)\left(Y - E(Y)\right) = \\ &= D(X) + D(Y) + 2K(X,Y) \end{aligned}$$

S tem je izrek dokazan.

Za *nekorelirani* slučajni spremenljivki *X* in *Y* velja

$$D(X+Y) = D(X) + D(Y) \tag{10.46}$$

Formulo (10.46) lahko posplošimo: Za slučajne spremenljivke $X_1, X_2, \ldots, X_n$, ki imajo disperzijo, velja

$$D\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} D(X_k) + 2\sum_{1 \le i < j \le n} K(X_i, Y_j) \tag{10.47}$$

Posledica zgornje lastnosti je še tale: če za slučajne spremenljivke $X_1, X_2, \ldots, X_n$ obstojajo disperzije in če so te slučajne spremenljivke *paroma nekorelirane* (to pomeni, da so vse kovariance $K(X_i, Y_j) = 0$), velja

$$D\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} D(X_k) \tag{10.48}$$

Primer:
Neodvisne in enako porazdeljene slučajne spremenljivke X_1, X_2,..., X_n naj imajo (vsaka zase) znano matematično upanje *a* in znano disperzijo *b*. Izračunajmo matematično upanje in disperzijo slučajne spremenljivke *Y*, ki jo dobimo kot aritmetično povprečje danih slučajnih spremenljivk.

Rešitev:
Iščemo matematično upanje in disperzijo slučajne spremenljivke *Y*, ki je dana kot aritmetično povprečje danih slučajnih spremenljivk, torej velja

$$Y = \frac{X_1 + X_2 + ... + X_n}{n}$$

Na osnovi zgoraj zapisanih lastnosti je matematično upanje

$$E(Y) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}\ E(X_1) + E(X_2) + \cdots + E(X_n)\ =$$
$$= \frac{1}{n}(a + a + \cdots + a) = \frac{1}{n}\, n\, a = a$$

Podobno dobimo še disperzijo

$$D(Y) = D\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\ D(X_1) + D(X_2) + \cdots + D(X_n)\ =$$
$$= \frac{1}{n^2}(b + b + \cdots + b) = \frac{1}{n^2}\, n\, b = \frac{b}{n}$$

Ugotovili smo zanimivo lastnost: a*ritmetično povprečje n neodvisnih spremenljivk z enakim matematičnim upanjem in enako disperzijo ima isto matematično upanje* in *n-krat manjšo disperzijo kot vsaka posamezna spremenljivka.*

♦ ♦ ♦

Primer:
Slučajno spremenljivko *X*, porazdeljeno po binomskem zakonu $b(n, p)$, smemo izraziti kot vsoto končnega števila slučajnih spremenljivk $X_k : \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$, $k = 1,\ 2,\ \ldots,\ n$, torej $X = X_1 + X_2 + \cdots + X_n$. Izračunajmo njeno disperzijo in varianco.

Rešitev:
Ker je $X_k : \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$, je $X_k^2 : \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$, $p+q=1$.

Za <u>posamezno</u> slučajno spremenljivko X_k, $k = 1, 2, \ldots, n$ smo v enem od prejšnjih primerov že izračunali matematično upanje in disperzijo: $E(X_k) = p$ in $D(X_k) = pq$.
Za slučajno spremenljivko $X = X_1 + X_2 + \cdots + X_n$ je potem po formuli (10.48)

$$D(X) = n\,D(X_k) = npq$$
$$\sigma_x = \sqrt{npq}$$

♦♦♦

<u>Primer:</u>
Izračunajmo matematično upanje, disperzijo in standardno deviacijo za slučajno spremenljivko $Z = \dfrac{X - E(X)}{\sigma(X)}$.

<u>Rešitev:</u>

Matematično upanje slučajne spremenljivke $Z = \dfrac{X - E(X)}{\sigma(X)}$ je

$$E(Z) = E\left(\frac{X - E(X)}{\sigma(X)}\right) = \frac{1}{\sigma(X)} \cdot E\big(X - E(X)\big) = 0$$

Disperzija je

$$D(Z) = D\left(\frac{X - E(X)}{\sigma(X)}\right) = \frac{1}{\sigma^2(X)} \cdot D\big(X - E(X)\big) = \frac{1}{\sigma^2(X)} \cdot \sigma^2(X) = 1$$

Standardna deviacija pa je:

$$\sigma(Z) = +\sqrt{D(Z)} = 1$$

Slučajna spremenljivka $Z = \dfrac{X - E(X)}{\sigma(X)}$ ima torej vedno matematično upanje 0 in standardni odklon 1.

To lastnost uporabimo pri normalni porazdelitvi, ko jo želimo standardizirati.
Normalno porazdelitev $p(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$ z matematičnim upanjem $E(X) = a$ in standardno deviacijo σ lahko s substitucijo $z = \dfrac{x-a}{\sigma}$ prevedemo v standardizirano normalno porazdelitev $\varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.

♦♦♦

11 OSNOVE VZORČENJA IN STATISTIČNO OCENJEVANJE

Vzorčenje pomeni posebno metodo zbiranja, obdelave in razlage statističnih parametrov. To je metoda statističnega sklepanja o statističnih parametrih statistične množice na podlagi opazovanja samo enega dela celotne statistične množice.

Velikokrat ne moremo vedno opazovati in proučevati celotne populacije, zato se odločimo za opazovanje samo enega njenega dela. Temu delu populacije pravimo vzorec. Seveda mora biti vzorec takšen, da v čim večji meri izraža vse lastnosti cele populacije, torej mora biti reprezentativen. Enote za vzorec izbiramo z različnimi načini vzorčenja, med katerimi je najbolj običajno slučajno vzorčenje. To pomeni, da dobimo naključno izbrane reprezentante celotne populacije, kar zahteva, da mora biti verjetnost za izbor statističnega elementa v vzorec za vse enote enaka oziroma, da mora biti enaka verjetnost izbora različnih, vendar po številu enot enako močnih vzorcev iz osnovne celotne statistične množice.

Vzorec izberemo običajno po metodi slučajnih števil. Na spodnji tabeli je primer 200 slučajnih števil, izbranih med 0 in 999999. Uporabljen je bil GraphPad Software (http://graphpad.com/quickcalcs/randomn2.cfm).

Vsak vzorec je seveda drugačen od prejšnjega, zato so tudi njegove karakteristike drugačne. Statistični parametri, izračunani kazalniki, ugotovljene lastnosti in podobno so torej natančni in pravilni le za konkretni vzorec, za ostale vzorce, oziroma, kar je pomembno, za celotno populacijo, pa so to le boljše ali slabše *ocene*.

Vzorce delimo na velike in majhne. Pri majhnih vzorcih je število enot $n < 30$, pri velikih vzorcih pa je $n > 30$.
Opomba: ponekod je v rabi tudi ocena, da je velik vzorec tisti, ki ima več kot 50 enot.

Proučevanje vzorca je seveda enostavnejše, kot proučevanje celotne populacije, mnogokrat pa je sploh edina možnost, ker ima populacija v mnogo primerih enostavno preveč elementov in je nemogoče obdelati vse.

Seveda pa ima proučevanje vzorca bistveno pomanjkljivost. Podatki, ki jih dobimo za vzorec in statistični kazalniki, ki jih iz meritev vzorca ugotavljamo, namreč natanko in pravilno veljajo samo za natanko določen vzorec, ne pa natanko tako tudi za ostale možne vzorce ali celo za celotno populacijo.

Tabela: Slučajna števila

Row	A	B	C	D	E	F	G	H	I	J
1	714636	625278	559056	470114	802096	256354	411520	749164	918576	633292
2	504740	417187	116632	317640	649043	603843	387236	986286	508400	90623
3	509916	623304	444937	382222	178538	580910	522692	993668	251408	290403
4	265649	477110	844290	528833	175131	932138	1917	144411	619671	131765
5	277129	110699	648309	67040	357033	622696	161463	867976	446153	127411
6	178548	931439	18183	587702	917678	824687	236763	34155	893896	34940
7	389626	92356	50639	19132	152849	184898	512705	998902	930754	964
8	267070	145496	854224	629087	699715	652570	899977	732479	889858	75397
9	978752	986402	266236	589983	920912	81337	909965	981412	550975	259875
10	174970	175469	740508	240001	772314	732409	850686	371742	578406	933783
11	161309	561207	660776	630938	659059	232250	836877	545480	756306	925779
12	819503	54150	450840	422051	190059	470719	339569	86686	730976	760101
13	674028	982813	153451	289827	762053	668258	190934	975795	908177	190328
14	423693	27084	404815	811500	188639	534845	255224	503859	975721	577969
15	710045	181282	547080	871572	890593	856381	826714	565990	175951	305217
16	105872	950477	186780	658604	214155	526996	328976	471913	505310	759155
17	840741	275363	748463	147586	645085	758200	307012	985151	161448	715303
18	704143	837008	262104	546041	409781	522568	139904	725511	45808	105374
19	505258	635146	279854	817776	540456	656140	692737	195260	719223	9197
20	652695	827408	761631	12738	544188	868284	640104	11002	726809	663888

Vir: http://graphpad.com/quickcalcs/randomn2.cfm

Iz vsake statistične populacije namreč lahko izberemo več vzorcev. Koliko je možnih različnih vzorcev?

Vzemimo, da ima populacija N, vzorec pa n elementov (seveda je $n<N$). Število možnih vzorcev ugotovimo z uporabo osnov kombinatorike. Iščemo podmnožice n elementov množice z N elementi. Enote v vzorcu se ne smejo ponavljati, njihovo mesto v podmnožici pa je nepomembno, v jeziku kombinatorike gre torej za kombinacije. Vseh možnih različnih vzorcev je potem toliko, kot je kombinacij N elementov n-tega reda, torej $\text{št. vzorcev} = \binom{N}{n} = \frac{N!}{n!(N-n)!}$. Iz teorije kombinatorike in verjetnostnega računa vemo, da je število kombinacij za velike N veliko.

Primer:
Koliko različnih vzorcev po 10 elementov lahko tvorimo iz 100 elementov?

Rešitev:

$N=100$ in $n=10 \Rightarrow$ št. vzorcev je $\binom{100}{10}=\frac{100!}{10!\cdot 90!}\approx 17{,}3\cdot 10^{12}$.

Različnih vzorcev je v tem primeru kar 17 tisoč milijard!

♦♦♦

Rezultate, ki jih dobimo pri vzorčenju, grupiramo v razrede in ugotavljamo obliko njihove porazdelitve. Te porazdelitve primerjamo s teoretičnimi porazdelitvami slučajne spremenljivke, kot so binomska porazdelitev, Poissonova porazdelitev, Studentova porazdelitev, porazdelitev χ^2,…

Najbolj znana in tudi najbolj uporabljana je ***normalna porazdelitev,*** ki smo jo že omenili. Normalna porazdelitev, ki jo označimo *N(a, σ^2),* ima gostoto verjetnosti

$$p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2} \qquad (11.01)$$

Ta porazdelitev je odvisna od parametrov *a* in σ. Pri tem je lahko *a* poljubno realno število, σ pa poljubno pozitivno število. V primerih prejšnjega poglavja smo že ugotovili, da sta to ravno matematično upanje in disperzija, torej $E(X)=a$ in $D(X)=\sigma^2$.

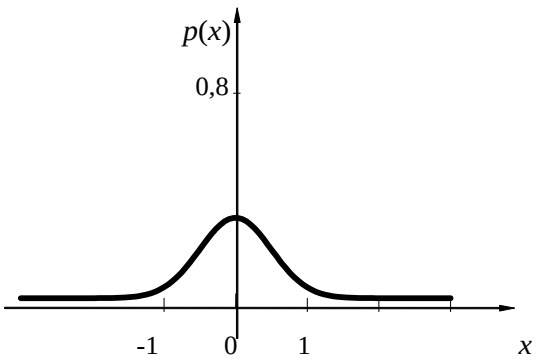
Drugače rečeno: $a=M_x=\overline{x}$ je aritmetična sredina, σ pa je standardni odklon porazdelitve.

Z uporabo prvih treh odvodov (dela je kar precej) izračunamo, da ima funkcija $p(x)$ en ekstrem in dva prevoja:

- v točki $E\left(a,\ \frac{1}{\sigma\sqrt{2\pi}}\right)$ je maksimum,
- v točkah $T_1\left(a-\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$ in $T_2\left(a+\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$ sta prevoja.

Parameter *a* torej določa lego krivulje, parameter σ pa določa njeno obliko. Na spodnji sliki je narisana funkcija $p(x)$ za dana parametra $a=0$ in $\sigma=0{,}5$.

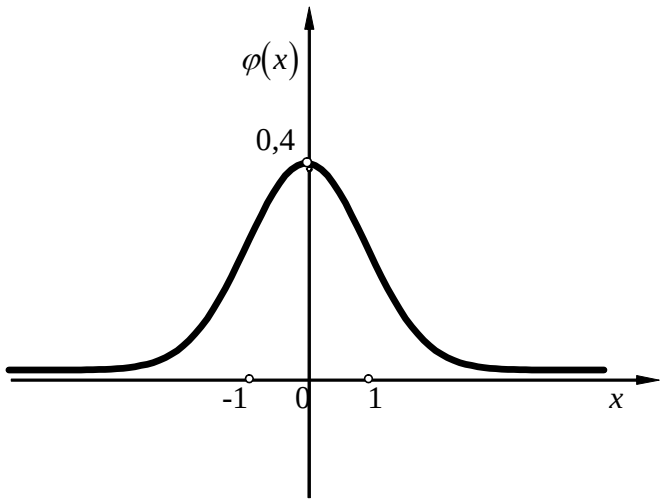
Primer za $a = 0,\ \sigma = 0{,}5$



Če vzamemo vrednosti parametrov $a = 0$ in $\sigma = 1$, dobimo <u>*standardizirano normalno porazdelitev N*(0,1),</u> pri kateri je gostota verjetnosti dana z izrazom:

$$p(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{11.02}$$

Graf funkcije $\varphi(x)$ je prikazan na spodnji sliki.



Za funkcijo (11.02) veljajo naslednje lastnosti:

$$\int_{-\infty}^{x} \varphi(x)dx = P(X < x) \tag{11.03}$$

$$\int_0^x \varphi(x)dx = P(0 \le X < x) \tag{11.04}$$

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1 \tag{11.05}$$

$$\varphi(-x) = \varphi(x) \tag{11.06}$$

Vrednosti funkcije (11.02) so tabelirane. V tabeli so navedene vrednosti funkcije le za pozitivne *x*, saj je funkcija soda, kot vidimo iz (11.03).

Zaradi lastnosti (11.03) - (11.06) je standardizirana normalna porazdelitev (11.02) zelo koristna in predvsem zelo enostavna za praktično uporabo. Njene vrednosti so dane v tabeli pri poglavju o porazdelitvah.

Iz tega razloga je smiselno poljubno normalno porazdelitev

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$$

spremeniti v standardizirano normalno porazdelitev.

Ugotovili smo že, da je to moč storiti s tole substitucijo

$$\boxed{z = \frac{x-a}{\sigma}} \tag{11.07}$$

Na ta način res dobimo standardizirano normalno porazdelitev za novo spremenljivko *z*:

$$\boxed{\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}} \tag{11.08}$$

S posameznim izbranim vzorcem dobimo le **vzorčno oceno statističnega parametra**, ki je seveda odvisna od karakteristik slučajno izbranega vzorca. To pomeni, da vzorčne ocene statističnega parametra niso enake za vse vzorce, pač pa se spreminjajo, torej so spremenljivke. Ker so odvisne od posameznega vzorca, ki je odvisen od slučajne izbire, so torej vrednosti statističnega parametra slučajne spremenljivke. Spremenljivko, za katero so vzorčne vrednosti določenega statističnega parametra vzorčne ocene vseh vzorcev, imenujemo **cenilka statističnega parametra**.

Ocene parametrov populacije, ki jih dobimo na podlagi vzorca, imenujemo ***statistike*.**

Posebej znani statistiki vzorca sta aritmetična sredina vzorca

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{11.09}$$

in varianca vzorca

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \tag{11.10}$$

Vse statistike so slučajne spremenljivke.

Ko realiziramo slučajne spremenljivke $y_1,y_2,\ldots,y_n$, torej šele tedaj, ko vzorec dejansko izberemo, lahko izračunamo konkretne *vrednosti* statistik, kot recimo

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

in

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$$

Pri statističnem sklepanju so zelo pomembni porazdelitveni zakoni teh slučajnih spremenljivk (statistik), imenovali jih bomo vzorčne porazdelitve.

Za *vzorčno porazdelitev aritmetične sredine* velja naslednji izrek, ki ga povejmo brez dokaza. Dokaz je možno prebrati npr. v knjigi Jesenko, Jože.: *Statistika v organizaciji in management*, Moderna organizacija, Kranj, 2001.

Izrek:
Če je $y_1,y_2,\ldots,y_n$ *slučajni vzorec iz neskončne populacije z aritmetično sredino M in varianco* σ^2*, potem velja*

$$E\left(\bar{y}\right)=M=M_{\bar{y}}$$
$$V\left(\bar{y}\right)=\frac{\sigma^2}{n}=\sigma_{\bar{y}}^2 \tag{11.11}$$

Standardni odklon $\sigma_{\bar{y}}=\frac{\sigma}{\sqrt{n}}$ imenujemo *standardna ocena napake* aritmetične sredine $M_{\bar{y}}$.

To pomeni, da se aritmetična sredina vzorcev porazdeli normalno, torej je statistika z imenom »aritmetična sredina vzorca« normalna slučajna spremenljivka.

Iz formule (11.11) vidimo, da je matematično upanje aritmetične sredine vzorca enako aritmetični sredini celotne populacije, iz katere so bili vzorci vzeti, varianca vzorca pa je enaka varianci celotne populacije, deljeno z *n*, to je s številom elementov vzorca .

V statistiki je izjemnega pomena še ***centralni limitni izrek***, ki ga tudi navedimo brez dokaza.

Centralni limitni izrek: *Če je* $y_1,y_2,\ldots,y_n$ *slučajni vzorec iz neskončne populacije z aritmetično sredino* M *in standardno deviacijo* σ, *potem slučajna spremenljivka* $Z=\frac{\bar{y}-M}{\sigma}\cdot\sqrt{n}$ *teži k standardizirani normalni slučajni spremenljivki, ko gre* $n\to\infty$.

Centralni limitni izrek torej pove, da je aritmetična sredina vzorca $\bar{y}$ porazdeljena normalno z matematičnim upanjem M in standardnim odklonom $\frac{\sigma}{\sqrt{n}}$.

Pomembno dejstvo je, da centralni limitni izrek velja za ***vsak vzorec*** ne glede na populacijo, iz katere je izbran.

Aritmetična sredina vzorca je vedno normalna slučajna spremenljivka porazdeljena po $N\left(M,\frac{\sigma}{\sqrt{n}}\right)$, *če je le vzorec dovolj velik.* Izkaže se, da je dovolj že 30 enot.

Iz tega lahko preberemo naslednjo lastnost:

Če je $\bar{y}$ *aritmetična sredina slučajnega vzorca velikosti n iz normalne populacije z aritmetično sredino M in standardnim odklonom* σ, *potem je* $\bar{y}$ *normalna slučajna spremenljivka z matematičnim upanjem M in standardnim odklonom* $\frac{\sigma}{\sqrt{n}}$, *torej* $E(\bar{y})=M$, $E(s^2)=\sigma^2$.

Sedaj se posvetimo še statističnemu sklepanju.

Statistično sklepanje pomeni dvoje:

- ocenjevanje
  - točkasto,
  - intervalno,

- testiranje hipotez.

Pri statističnem ocenjevanju moramo določiti vrednosti nekega parametra na zvezni množici vzorcev.
Pri testiranju hipotez pa moramo odločati o zavrnitvi (ali pa ne) neke posebne vrednosti parametra.

Ko uporabimo vrednost statistike, da z njo ocenimo kakšen parameter populacije, pravimo postopku **točkasto ocenjevanje**, vrednost statistike pa imenujemo *točkasta ocena parametra*.

Postopek, po katerem iz izbranega vzorca izračunamo točkovno oceno parametra, imenujemo **cenilka parametra.** Tako je npr. S^2 cenilka variance σ^2, medtem ko je s^2 točkasta ocena parametra σ^2.

Cenilke so slučajne spremenljivke, zato je potrebno določati in ugotavljati njihove verjetnostne porazdelitve.

Ko izbiramo najprimernejše cenilke, upoštevamo različne lastnosti posameznih cenilk, saj želimo čim bolj natančno oceniti dejansko vrednost parametra na celotni populaciji.

Cenilke morajo imeti naslednje lastnosti:

- *nepristranost,*
- *učinkovitost,*
- *konsistentnost,*
- *zadostnost,*
- *krepkost.*

Cenilka je nepristranska, ko je njeno matematično upanje enako parametru cele populacije, ki jo s to cenilko ocenjujemo.
Če je npr. S^2 cenilka variance σ^2 in populacija neskončna, potem je ta cenilka nepristranska, če velja $E\left(S^2\right)=\sigma^2$.

Intervalsko ocenjevanje parametra, označimo ga npr. z znakom λ, pomeni, da določimo nek interval $\left[\lambda_1, \lambda_2\right]$, kjer sta obe krajišči in vse vrednosti vmes realizacije ustreznih statistik.

Pri tem naj velja

$$P(\lambda_1 \le \lambda \le \lambda_2) = 1-\alpha \qquad (11.12)$$

Pri vnaprej določeni verjetnosti $(1-\alpha)$ pravimo, da je $[\lambda_1, \lambda_2]$ interval zaupanja za parameter λ.

Verjetnosti $(1-\alpha)$ pravimo stopnja zaupanja, krajišči intervala $[\lambda_1, \lambda_2]$ pa sta meji zaupanja.

Zgornje lahko povemo tudi takole*: dogodek* $(\lambda \in [\lambda_1, \lambda_2])$ *je slučajni dogodek. Verjetnost, da se bo ta dogodek zgodil, je* $(1-\alpha)$*, verjetnost, da se ta dogodek ne bo zgodil, pa je* α.

Verjetnosti α pravimo stopnja tveganja ali tudi napaka 1. vrste in pomeni verjetnost, da ocenjeni interval $[\lambda_1, \lambda_2]$ ne bo vseboval vrednosti parametra λ.

11.1 Ocenjevanje aritmetične sredine populacije

Vzemimo populacijo, normalno porazdeljeno po $N(M,\sigma)$, za katero *ne poznamo* niti matematičnega upanja niti standardnega odklona.
Izberimo iz te populacije vzorec z *n* enotami, označimo ga, kot običajno: $y_1, y_2, \ldots, y_n$. Aritmetična sredina vzorca $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ je, glede na vse predhodno povedano, spet slučajna spremenljivka. Aritmetična sredina zanjo naj bo enaka *M*, varianca pa $\frac{\sigma}{\sqrt{n}}$, torej

$$M_{\bar{y}} = M$$

$$\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$$

Za izbrani vzorec je vrednost aritmetične sredina vzorca $\bar{y}$ točkovna ocena za matematično upanje *M* dane populacije.

Intervalsko oceno parametra pa lahko določimo s pomočjo standardizirane slučajne spremenljivke *Z*:

$$Z=\frac{\bar{y}-M}{\sigma_{\bar{y}}}=\frac{\bar{y}-M}{\sigma}\cdot\sqrt{n} \tag{11.13}$$

Sedaj z vnaprej določeno verjetnostjo trdimo, da se parameter nahaja na intervalu. Ob predpisani verjetnosti α obstaja natanko določeno število *Z* (11.13), tako da velja

$$P\left(-z\le\frac{\bar{y}-M}{\sigma_{\bar{y}}}\le z\right)=1-\alpha$$

Iz zahteve $-z\le\frac{\bar{y}-M}{\frac{\sigma}{\sqrt{n}}}\le z$, ki jo rešujemo kot sistem dveh neenačb, dobimo

interval zaupanja

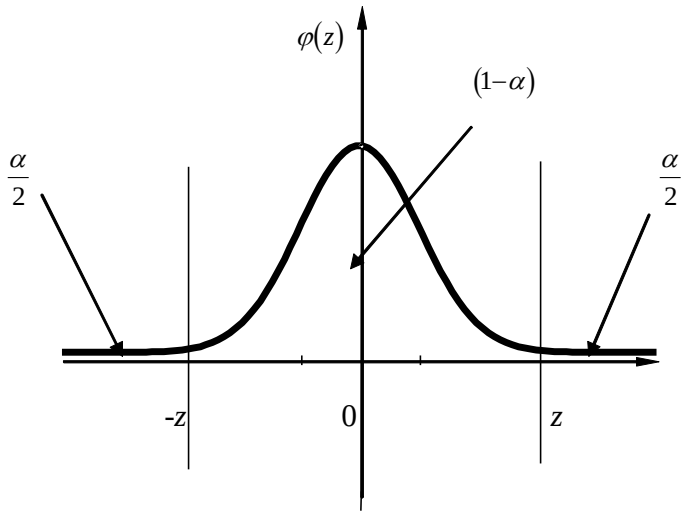
$$\bar{y}-\frac{z\sigma}{\sqrt{n}}\le M\le\bar{y}+\frac{z\sigma}{\sqrt{n}} \tag{11.14}$$

in verjetnost

$$P\left(\bar{y}-\frac{z\sigma}{\sqrt{n}}\le M\le\bar{y}+\frac{z\sigma}{\sqrt{n}}\right)=1-\alpha \tag{11.15}$$

Uporabili smo standardizirano normalno spremenljivko. To pomeni, da je število (-z) tista standardizirana vrednost, od katere ima $\frac{\alpha}{2}\cdot 100\%$ cenilk manjšo vrednost, število (+z) pa je tista standardizirana vrednost, od katere ima $\frac{\alpha}{2}\cdot 100\%$ cenilk večjo vrednost (slika).

Vrednosti števil *z* preberemo iz tabele za vrednosti porazdelitvene funkcije $F(z)$ za standardizirano normalno porazdelitev. Pri tem velja $|-z|=|z|=z$.



Za verjetnost α običajno vzamemo vrednosti: $\alpha=0{,}1$ *ali* $\alpha=0{,}05$ ali $\alpha=0{,}01$.

Velja (glej tabelo za funkcijo $\Phi(z)$):

$$\alpha=0{,}10\Rightarrow\Phi(z)=\frac{1-\alpha}{2}=\frac{1-0{,}1}{2}=0{,}4500\Rightarrow z=1{,}64$$

$$\alpha=0{,}05\Rightarrow\Phi(z)=\frac{1-\alpha}{2}=\frac{1-0{,}05}{2}=0{,}4750\Rightarrow z=1{,}96$$

$$\alpha=0{,}01\Rightarrow\Phi(z)=\frac{1-\alpha}{2}=\frac{1-0{,}01}{2}=0{,}4950\Rightarrow z=2{,}58$$

11.1.1 Velikost vzorca

Z verjetnostjo $(1-\alpha)$ predpišemo neko dovoljeno *napako d*, ki naj pomeni največjo dopustno absolutno razliko med aritmetično sredino vzorca in aritmetično sredino populacije, to je

$$|\bar{y}-M|\leq d$$

Iz formule (11.14) vemo, da je

$$|\bar{y}-M|\leq\frac{z\sigma}{\sqrt{n}}$$

Od tod dobimo $\frac{z\sigma}{\sqrt{n}} \le d$ in nato oceno za velikost vzorca:

$$\boxed{n \ge \left(\frac{z\sigma}{d}\right)^2} \qquad (11.16)$$

V formuli (11.16) je vrednost *z* določena z verjetnostjo α.

Primer:
Na vzorcu 50 študentov ugotovimo, da je njihova povprečna masa 75,5 kg in standardni odklon 8,5 kg. Z zaupanjem 0,99 določimo interval zaupanja aritmetične sredine tega vzorca.

Rešitev:
$n = 50$ (velikost vzorca)
$\bar{y} = 75{,}5$
$\sigma = 8{,}5$
$1 - \alpha = 0{,}99 \Rightarrow \alpha = 0{,}01$

Za interval zaupanja velja formula (11.14), to je : $\bar{y} - \frac{z\sigma}{\sqrt{n}} \le M \le \bar{y} + \frac{z\sigma}{\sqrt{n}}$.
Vemo tudi že, da je pri $\alpha = 0{,}01$ vrednost $z = 2{,}58$. Torej je

$$\frac{z\sigma}{\sqrt{n}} = \frac{2{,}58 \cdot 8{,}5}{\sqrt{50}} = 3{,}1$$

Interval zaupanja (11.14) je v tem primeru:

$$\bar{y} - \frac{z\sigma}{\sqrt{n}} \le M \le \bar{y} + \frac{z\sigma}{\sqrt{n}}$$

$$75{,}5 - 3{,}1 \le M \le 75{,}5 + 3{,}1$$

$$72{,}4 \le M \le 78{,}6$$

Interval zaupanja z zaupanjem 0,99 za aritmetično povprečje *M* je $[72{,}4; 78{,}6]$.
Komentar: Zgornji zapis pomeni, da z verjetnostjo 0,99 lahko sklepamo, da je povprečna masa študentov med 72,4 in 78,6 kg.
♦ ♦ ♦

Primer:
Na slučajno izbranem vzorcu 100 srednješolcev so bili zbrani rezultati pri maturi ter uspeh v 4. in 3. letniku, skupno ocenjeni z največ 100 točkami.
Rezultati so dani s frekvenčno porazdelitvijo in prikazani v tabeli.

Št. točk	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
frekvenca	1	3	5	8	12	17	25	18	9	2

Izračunajmo:

a) mediano, modus, aritmetično sredino,
b) tretji kvartil,
c) interval zaupanja povprečno doseženega števila točk za celotno populacijo, če je stopnja zaupanja 0,99

in preverimo

d) ali dana velikost vzorca zadošča, da z verjetnostjo 0,95 trdimo, da se vzorčna aritmetična sredina ne razlikuje od aritmetične sredine populacije za več kot 3% aritmetične sredine vzorca,
e) kolikšna je natančnost vzorca pri 100 enotah?

Rešitev:
Tabelo dopolnimo s stolpci F_k, y_k, $f_k y_k$, y_k^2 in $f_k y_k^2$.

	razred	f_k	F_k	y_k (sredina razreda)	$f_k y_k$	y_k^2	$f_k y_k^2$
1	0-10	1	1	5	5	25	25
2	11-20	3	4	15	45	225	675
3	21-30	5	9	25	125	625	3125
4	31-40	8	17	35	280	1225	9800
5	41-50	12	29	45	540	2025	24300
6	51-60	17	46	55	935	3025	51425
7	61-70	25	71	65	1625	4224	105625
8	71-80	18	89	75	1350	5625	101250
9	81-90	9	98	85	765	7225	65025
10	91-100	2	100	95	190	9025	18050
skupaj	-	100	-	-	5860	33250	379300

Sedaj lahko izračunamo:

- aritmetično sredino

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} f_k y_k = \frac{1}{100} \cdot 5860 = 58{,}6 \text{ točk}$$

- <u>modus</u>

  $k=7$

  $$Mo=y_{k,\min}+\frac{f_k-f_{k-1}}{2f_k-f_{k-1}-f_{k+1}}\cdot i=66{,}3\ točk$$

- <u>mediano</u>

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

  $R=NP+0{,}5=50{,}5$

  $\Rightarrow k=7$

  $$Me=y_{k,\min}+\frac{R-F_{k-1}}{f_k}\cdot i=62{,}8\ točk$$

- <u>tretji kvartil</u>

  $P(Q_3)=0{,}75$

  $R=NP+0{,}5=75{,}5$

  $\Rightarrow k=8$

  $$Q_3=y_{k,\min}+\frac{R-F_{k-1}}{f_k}\cdot i=73{,}5\ točk$$

Za iskani interval zaupanja povprečno doseženega števila točk velja

$$1-\alpha=0{,}99\Rightarrow\alpha=0{,}01\Rightarrow z=2{,}58$$

Za standardni odklon

$$S^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(f_k y_k-\bar{y}\right)^2=\frac{1}{n-1}\left(\sum_{k=}^{n}f_k y_k^2-n\bar{y}^2\right)$$

je vrednost cenilke

$$s^2=\frac{1}{99}\left(379300-100\cdot 58{,}6^2\right)=362{,}67$$

$$s=\sqrt{s^2}=19{,}04\quad \text{točk}$$

Za interval zaupanja velja formula: $\bar{y}-\frac{z\sigma}{\sqrt{n}}\le M\le\bar{y}+\frac{z\sigma}{\sqrt{n}}$.

Ker je v tem primeru $\alpha=0{,}01\Rightarrow z=2{,}58$, je

$$\frac{z\sigma}{\sqrt{n}}=\frac{2{,}58\cdot 19{,}04}{\sqrt{100}}=4{,}9$$

in od tod sledi iskani <u>interval zaupanja</u>

$$\bar{y}-\frac{z\sigma}{\sqrt{n}} \le M \le \bar{y}+\frac{z\sigma}{\sqrt{n}}$$

$$58,6-4,9 \le M \le 58,6+4,9$$

$$53,7 \le M \le 63,4$$

Z verjetnostjo 0,99 bo aritmetična sredina celotne populacije srednješolcev v intervalu $[53,7 \text{ točk},\ 63,4 \text{ točk}]$.

Še velikost vzorca:

$$\alpha = 0,05 \Rightarrow z = 1,96$$

3% od aritmetične sredine vzorca pomeni

$$d = 0,03 \cdot \bar{y} = 0,03 \cdot 58,6 = 1,76 \text{ točk}$$

Po formuli (11.16) dobimo

$$n \ge \left(\frac{z\sigma}{d}\right)^2 = \left(\frac{1,96 \cdot 19,04}{1,76}\right)^2 = 449,6$$

Vzorec bi moral imeti vsaj 450 enot.

Ugotovimo še, kolikšna je natančnost vzorca pri 100 enotah (*n=100*).

Iz formule (11.56) sledi

$$n = \left(\frac{z\sigma}{d}\right)^2 \Rightarrow d = \frac{z\sigma}{\sqrt{n}}$$

Ker že od prej vemo, da je v našem primeru $\sigma = 19,04$, dobimo

$$d = \frac{z\sigma}{\sqrt{n}} = \frac{1,96 \cdot 19,04}{\sqrt{100}} = 3,73 \text{ točk}$$

Iz zahteve $d = x\% \cdot \bar{y}$ dobimo

$$x\% = \frac{d}{\bar{y}} = \frac{3,73}{58,6} = 0,064 = 6,4\%$$

♦ ♦ ♦

11.2 Testiranje hipotez

Pod pojmom statistična hipoteza (domneva) razumemo neko predpostavko o porazdelitvi ene ali več slučajnih spremenljivk. Takšno predpostavko moramo preveriti in jo z neko vnaprej dano verjetnostjo potrditi ali pa ne, kar storimo z obdelavo in preverjanjem vzorčnih podatkov.

Če statistična hipoteza v celoti določa porazdelitev, jo imenujemo *enostavna,* v nasprotnem primeru je hipoteza *sestavljena*.

Enostavna hipoteza torej ne določa le obliko verjetnostne porazdelitve slučajne spremenljivke, pač pa tudi vrednosti njenih parametrov.

Test je *parametričen*, ko se nanaša na kvantitativne parametre populacije, ko pa se test nanaša na kvalitativne parametre populacije, je *neparametričen*.

Poleg postavljene hipoteze vedno definiramo tudi *nasprotno (alternativno) hipotezo.*

Testiranje hipotez torej zahteva najprej dvoje:

- določitev hipoteze, s katero preverjamo predpostavko o porazdelitvi ali samo o nekem parametru in jo imenujemo *ničelna hipoteza* H_0,
- določitev hipoteze, ki predpostavlja pogoje, pri katerih ničelna hipoteza H_0 ne drži (ničelno hipotezo tedaj zavrnemo); to je *nasprotna ali alternativna hipoteza* H_1.

Testiranje statistične hipoteze zahteva natanko določeno proceduro (postopek) za odločitev o tem, ali ničelno hipotezo zavrnemo ali ne.
Opomba: postavljene hipoteze, bodisi ničelne bodisi nasprotne, ne sprejmemo, čeprav je test zanjo ugoden, pač pa jo v takšnem primeru ne zavrnemo. Če ne zavrnemo ničelne hipoteze, s tem avtomatsko zavrnemo nasprotno hipotezo in obratno: če ne sprejmemo ničelne hipoteze, s tem ne zavrnemo nasprotne hipoteze.

Če označimo pojem, ki ga proučujemo, z znakom λ, je formalni zapis pri testiranju hipotez takšen:

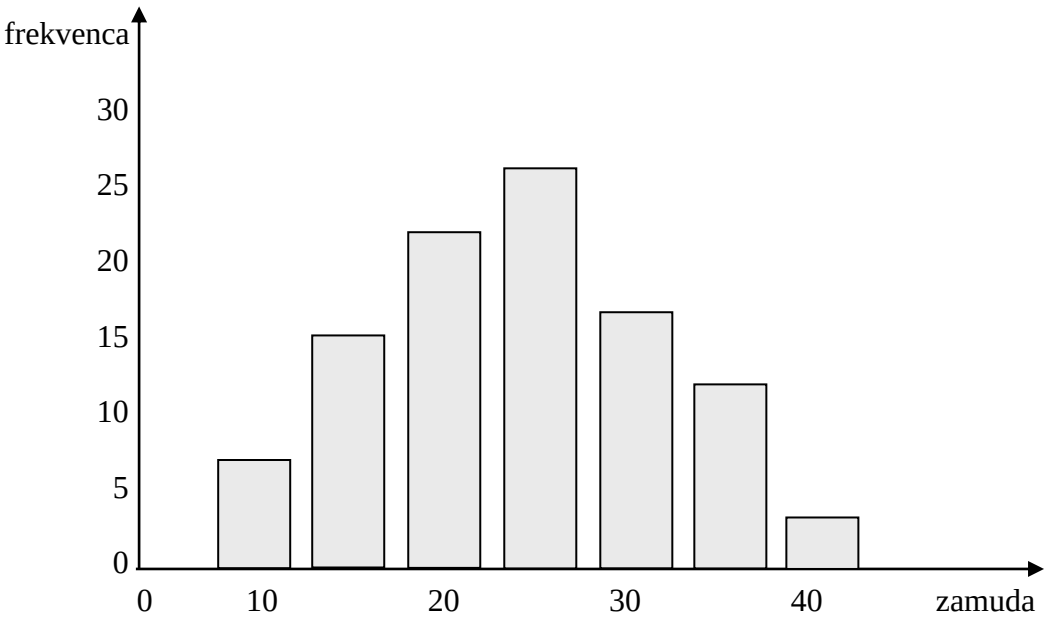
Želimo testirati ničelno hipotezo $H_0: \ \lambda = \lambda_0$
pri nasprotni hipotezi $H_1: \ \lambda = \lambda_1$

Primer:
Vzemimo, da na nekem letališču merimo zamude vzletov letal v nekem konkretenem dnevu. Čase merimo v minutah. Meritve prikažemo v obliki frekvenčne porazdelitve:

Razred (zamuda v minutah)	Frekvenca f_k	Kumulativna frekvenca F_k	Sredina razreda y_k	$f_k y_k$
5,01- 10,0	7	7	7,5	52,5
10,01-15,0	15	22	12,5	187,5
15,01-20,0	22	44	17,5	385,0
20,01-25,0	26	70	22,5	585,0
25,01-30,0	16	86	27,5	440,0
30,01-35,0	11	97	32,5	357,5
35,01-40,0	3	100	37,5	112,5
skupaj	100	-	-	2120,0

Povprečna zamuda v natanko določenem dnevu, ko smo izvajali meritev, je $\bar{y} = \frac{1}{N} \cdot \sum_{k=1}^{8} f_k y_k = 21,2$ minut.

Graf:



Ničelna hipoteza je sedaj lahko naslednja: povprečna zamuda vzletov na tem letališču v enem dnevu je vedno 21 minut, nasprotna hipoteza pa je seveda: povprečna zamuda vzletov letal na tem letališču v enem dnevu ni vedno 21 minut. Če z znakom $\bar{y}$ zaznamujemo povprečno zamudo vzletov v enem dnevu, potem zapišemo

$H_0: \quad \bar{y} = 21$
$H_1: \quad \bar{y} \neq 21$

Ničelna in nasprotna hipoteza pa sta npr. lahko tudi

H_0 : frekvenčna porazdelitev se pokorava normalni porazdelitvi

H_1 : frekvenčna porazdelitev se ne pokorava normalni porazdelitvi

♦ ♦ ♦

Pri postopku testiranja hipotez moramo seveda najprej imeti podatke, ki jih pridobimo iz meritev konkretnega parametra na slučajnem vzorcu. Iz tej podatkov pridobimo oz. izračunamo posebej določene statistike. Vsaki taki statistiki pravimo *statistika testa hipoteze* Γ. Glede na njeno vrednost se odločamo o tem ali hipotezo H_0 zavrnemo ali ne.

Celoten postopek testiranja torej pomeni ugotavljanje možne vrednosti ali območja *statistike testa hipoteze* Γ, torej moramo

- ugotoviti območje, kjer ne zavrnemo ničelne hipoteze H_0 (in s tem zavrnemo naprotno hipotezo H_1)
- ugotoviti območje, kjer zavrnemo ničelno hipotezo H_0 (in s tem ne zavrnemo nasprotne hipoteze H_1).

Pri tem postopku lahko pride do napak, ki so načeloma dveh vrst:

1. Kadar je prava vrednost domneve $\lambda = \lambda_0$, mi pa smo na osnovi podatkov slučajnega vzorca in opravljenega testa zaključili, da je $\lambda = \lambda_1$, potem smo naredili *napako 1. vrste*, ki jo označimo α.
2. Kadar pa je prava vrednost parametra $\lambda = \lambda_1$, mi pa smo na osnovi podatkov vzorca sklenili, da je $\lambda = \lambda_0$, potem smo naredili *napako druge vrste,* označimo jo β.

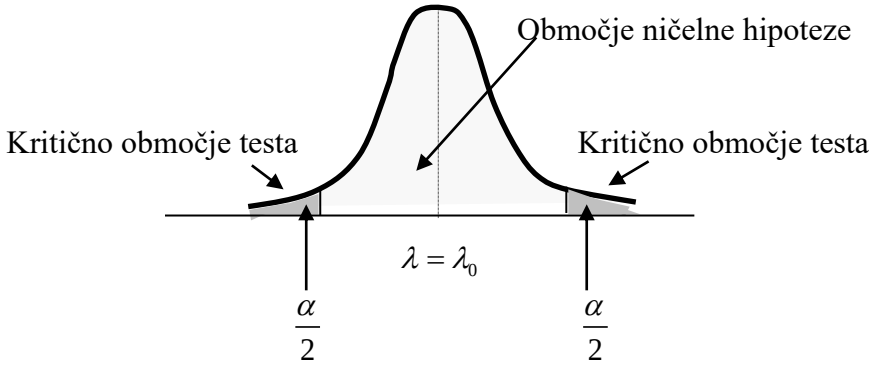
Napako 1. vrste torej naredimo, če zavrnemo ničelno hipotezo H_0, pa to ni pravilno. Napako 2. vrste pa naredimo, če ne zavrnemo ničelne hipotezo, vendar bi jo morali, ker dejansko ne drži.

Povedano drugače:

- zavrnitev ničelne hipoteze, če je pravilna, je napaka prve vrste α,
- nezavrnitev ničelne hipoteze, če je napačna, pa je napaka 2. vrste β.

Območje zavračanja ničelne hipoteze imenujemo *kritično območje testa*. Verjetnost, da vrednost statistike testa pade v kritično območje, je α in jo imenujemo *stopnja pomembnosti testa.*

Če bi bila porazdelitev slučajne spremenljivke normalna, si zgornje pojme grafično predstavimo takole, kot je skicirano na spodnji sliki.



Vzemimo, da želimo testirati hipotezo

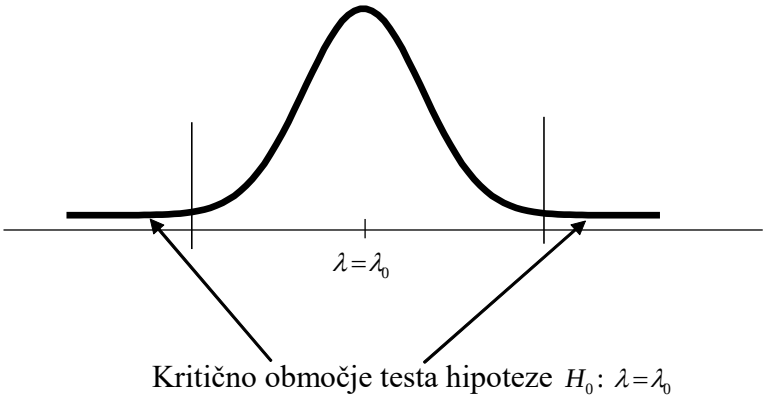
$$H_0: \lambda = \lambda_0$$

pri nasprotni hipotezi

$$H_1: \lambda \neq \lambda_0$$

Ničelne hipoteze ne zavrnemo, če je točkasta ocena parametra λ blizu λ_0 in jo zavrnemo, če je ta ocena veliko večja ali veliko manjša od λ_0. V tem primeru je kritično območje testa sestavljeno iz obeh delov gostote verjetnosti statistike testa hipoteze Γ. Takšnemu testu pravimo *dvostranski test.*

Grafično:



Vzemimo sedaj možnost, da želimo testirati ničelno hipotezo

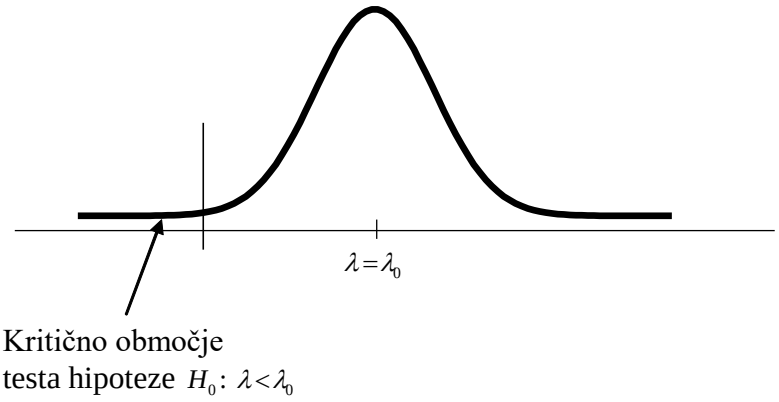
$$H_0: \lambda = \lambda_0$$

pri enostranski nasprotni hipotezi

$$H_1: \lambda < \lambda_0$$

V tem primeru zavrnemo ničelno hipotezo le, če je ocenjena vrednost veliko manjša od λ_0. Kritično območje testa je v tem primeru le na spodnjem (levem) delu porazdelitvene statistike testa hipoteze.

Grafično:

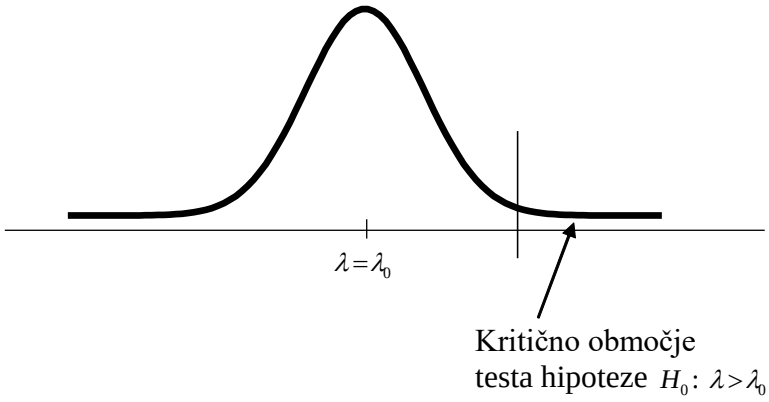


Podobno za primer testiranja hipoteze $H_0: \lambda=\lambda_0$

pri enostranski nasprotni hipotezi $H_1: \lambda>\lambda_0$

Sedaj ničelno hipotezo zavrnemo le, če je ocena veliko večja od λ_0.

Grafično:



Drugi in tretji primer testa imenujemo enostranski test hipoteze.

Postopek testiranja hipotez je natanko določen in ga izvedemo v nekaj korakih.

1. Postavimo hipotezo H_0 proti nasprotni hipotezi H_1 in določimo stopnjo pomembnosti testa α.
2. Izberemo primerno statistiko testa hipoteze in določimo kritično območje glede na stopnjo α.
3. Izračunamo vrednost statistike testa hipoteze iz podatkov vzorca.
4. Preverimo, če vrednost statistike testa hipoteze pade v kritično območje ali ne in glede na to hipotezo bodisi zavrnemo bodisi jo ne zavrnemo. Opomba: hipoteze ne sprejmemo, jo lahko le zavrnemo ali ne zavrnemo!

Določanje kritičnega območja je povezano z verjetnostjo α in z obliko porazdelitve slučajne spremenljivke *»statistika testa hipoteze* Γ*«.* Običajne vrednosti za α so 0,01 ali 0,05 ali 0,1.

11.2.1 Testiranje aritmetične sredine

Pri testiranju aritmetične sredine preverjamo, če je aritmetična sredina enaka neki vnaprej predpostavljeni vrednosti.

Torej testiramo ničelno hipotezo

$$H_0: \quad M = M_0$$

pri eni od nasprotnih hipotez:

a) $H_1: \quad M \neq M_0$ dvostranski test
b) $H_1: \quad M > M_0$ enostranski test
c) $H_1: \quad M < M_0$ enostranski test

Vzemimo naslednje predpostavke:

- vzorec je velikosti *n*,
- vzorec je izbran iz normalno porazdeljene populacije z aritmetično sredino M in standardnim odklonom σ, torej ***populacija*** je porazdeljena po $N(M,\sigma)$,
- poznamo napako prve vrste α.

V skladu s centralnim limitnim izrekom je statistika testa hipoteze H_0 v tem primeru standardizirana slučajna spremenljivka

$$Z = \frac{\overline{y} - M_0}{\sigma} \cdot \sqrt{n}$$

V tem izrazu je $\bar{y}$ aritmetična sredina ***vzorca***, M_0 je vnaprej predpostavljena vrednost aritmetične sredine ***populacije,*** σ je standardni odklon ***populacije*** , *n* pa je število enot v ***vzorcu.***

Opomba 1: če ne poznamo standardnega odklona σ, dobimo zanj oceno *s* iz statistike S^2. ($S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$).

Opomba 2: Statistika Z je uporabna le za velike vzorce ($n > 30$).

Opomba 3: Če je vzorec majhen, torej $n < 30$, potem uporabimo statistiko *t*, ki je dana z izrazom

$$t = \frac{\bar{y} - M_0}{s} \cdot \sqrt{n}$$

Pri dvostranskem testu hipoteze velja za kritično območje

$$|z| \geq z_{\frac{\alpha}{2}}$$

Pri enostranskem testu hipoteze pa

$$z \geq z_\alpha \quad \text{ali} \quad z \leq z_\alpha$$

Pri tem je z vrednost statistike testa hipoteze pri izbranem vzorcu z aritmetično sredino $\bar{y}$, to je

$$z = \frac{\bar{y} - M_0}{\sigma} \cdot \sqrt{n}$$

Primer:
Na vzorcu 200 študentov smo ugotavljali njihovo povprečno oceno. Podatki so dani v tabeli.

Z. št.	Povprečna ocena	Frekvenca f_k
1.	6,00 – 6,50	15
2.	6,51 – 7,00	40
3.	7,01 – 7,50	43
4.	7,51 – 8,00	47
5.	8,01 – 8,50	33
6.	8,51 – 9,00	10
7.	9,01 – 9,50	9
8.	9,51 – 10,00	3
Skupaj		200

Skušajmo iz tega vzorca dobiti aritmetično povprečje in ocenitev variance celotne populacije študentov.
Ali lahko ob napaki prve vrste $\alpha = 0,05$ trdimo, da je povprečna ocena celotne populacije študentov 7,50?

Rešitev:
Testiramo ničelno hipotezo

$$H_0: \quad M = 7,50$$

ob nasprotni hipotezi

$$H_1: \quad M \neq 7,50$$

Vidimo, da gre za dvostranski test hipoteze.

Vzeli bomo, da je vzorec standardizirano normalno porazdeljen s statistiko

$$Z = \frac{\bar{y} - M}{\sigma} \cdot \sqrt{n}$$

Ker so podatki grupirani v razrede, bomo aritmetično sredino vzorca $\bar{y}$ in standardni odklon vzorca določili s statistikama

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{r} f_i y_i$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^{r} (f_i y_i - \bar{y})^2$$

oziroma $S^2 = \frac{1}{n-1} \left(\sum_{i=1}^{r} f_i x_i^2 - n\bar{y}^2 \right)$

Dopolnimo tabelo podatkov z novimi stolpci, tako da bomo čim enostavneje dobili vrednosti statistik $\bar{y}$ in s^2.

Z. št.	Povprečna ocena	f_i	Sredina razr. y_i	y_i^2	$f_i y_i$	$f_i y_i^2$
1.	6,00 – 6,50	15	6,25	39,0625	93,75	585,9375
2.	6,51 – 7,00	40	6,75	45,5625	270,00	1822,5000
3.	7,01 – 7,50	43	7,25	52,5625	311,75	2260,1880
4.	7,51 – 8,00	47	7,75	60,0625	364,25	2822,9380
5.	8,01 – 8,50	33	8,25	68,0625	272,25	2246,0630
6.	8,51 – 9,00	10	8,75	76,5625	87,50	765,0625
7.	9,01 – 9,50	9	9,25	85,5625	83,25	770,0625
8.	9,51 – 10,00	3	9,75	95,0625	29,25	285,1875
Skup.		200	-	-	1512,00	11557,939

Ocena za aritmetično sredino vzorca je

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{8} f_i y_i = \frac{1}{200}\cdot 1512,00 = 7,56$$

Ocena za varianco pa je

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{r} f_i x_i^2 - n\bar{y}^2\right) = \frac{1}{199}\left(11557,939 - 200\cdot 7,56^2\right) = 0,639$$

Od tod dobimo oceno za standardno deviacijo

$$\sigma = \sqrt{s^2} = \sqrt{0,693} = 0,799$$

Sedaj ocenimo parametre, postavimo ničelno hipotezo

$$H_0: \quad M = 7,50$$

pri nasprotni hipotezi

$$H_1: \quad M \neq 7,50$$

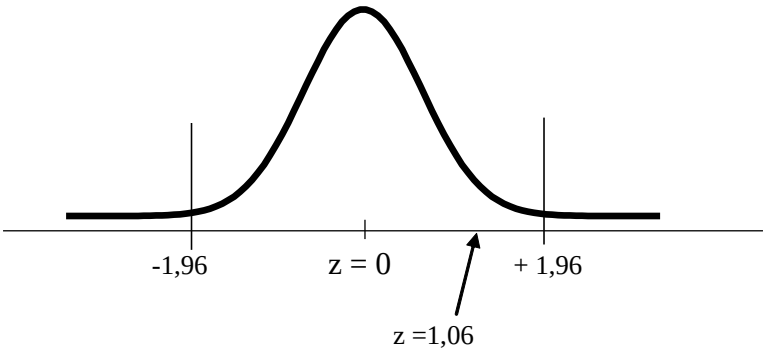
Uporabimo Z statistiko $Z = \frac{\bar{y} - M}{\sigma}\cdot\sqrt{n}$

Njena vrednost pri danih podatkih je

$$z = \frac{\bar{y} - M}{\sigma}\cdot\sqrt{n} = \frac{7,56 - 7,50}{0,799}\cdot\sqrt{200} = 1,062$$

Za predvideno napako prve vrste (ali stopnjo pomembnosti testa) $\alpha = 0,05$ pa je $z_{0,05} = 1,96$. To pomeni, da je izračunana vrednost standardizirane slučajne spremenljivke znotraj intervala $[-1,96;1,96]$, zato hipoteze H_0 ne zavrnemo.

Še graf:



♦ ♦ ♦

Primer:
V bolnišnici so uvedli nov postopek neke operacije. Predvidevajo, da bo čas okrevanja pacientov, operiranih po novi metodi, v povprečju 30 dni po operaciji. Novo metodo so preizkusili na vzorcu 30 pacientov. Povprečni čas okrevanja je bil 32,3 dni, standardna deviacija 4,2 dni.
Ali pri stopnji zaupanja 0,05 lahko sprejmemo domnevo, da so dobro predvideli povprečni čas okrevanja?

Rešitev:
$M = 30$
$\bar{y} = 32,3$
$n = 30$
$\sigma = 4,2$
$\alpha = 0,05$

Testiramo

$$H_0: \quad M = 30$$
$$H_1: \quad M \neq 30$$

Izračunamo

$$z = \frac{\bar{y} - M}{\sigma} \cdot \sqrt{n} = \frac{32,3 - 30}{4,2} \cdot \sqrt{30} = 2,999$$

Od prej vemo, da je za $\alpha = 0,05$ pripadajoči $z_{0,05} = 1,96$. Vidimo, da v tem primeru $z_{0,05} \notin [-1,96; 1,96]$, zato ničelno hipotezo H_0 zavrnemo.

♦ ♦ ♦

Primer:
Letalska družbe je izmerila zamude (v minutah) svojih letal pri 10 vzletih. Podatki so dani v tabeli.

Vzlet	1	2	3	4	5	6	7	8	9	10
zamuda	15	10	12	5	40	30	17	5	12	8

Družba zahteva, da je povprečna zamuda manj kot 15 minut. Ali je na osnovi danih podatkov pri stopnji zaupanja 0,1 mogoče sklepati, da bi to bilo možno?

Rešitev:
Ničelna in nasprotna hipoteza bosta

H_0: $M < 15$ minut
H_1: $M \geq 15$ minut

Zapišimo in dopolnimo tabelo s podatki takole:

vzlet	Zamuda y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	15	-0,4	0,16
2	10	-5,4	29,16
3	12	-3,4	11,56
4	5	-10,4	108,16
5	40	24,6	605,16
6	30	14,6	213,16
7	17	1,6	2,56
8	5	-10,4	108,16
9	12	-3,4	11,56
10	8	-7,4	54,76
skupaj	154	-	1144,4

Povprečna vrednost vzorca je

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10} \cdot 154 = 15,4 \text{ minut}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{9} \cdot 1144,4 = 127,16$$

$$\sigma = \sqrt{s^2} = \sqrt{127,16} = 11,3 \text{ minut}$$

Ker imamo majhen vzorec ($n<30$), ne moremo uporabiti normalne porazdelitve, pač pa Studentovo *t* porazdelitev, ki smo jo spoznali v prejšnjem poglavju. Če izhajamo iz formule (11.15), velja

$$P\left(\bar{y}-\frac{t_{n-1,\frac{\alpha}{2}}\cdot\sigma}{\sqrt{n}}\le M\le\bar{y}+\frac{t_{n-1,\frac{\alpha}{2}}\cdot\sigma}{\sqrt{n}}\right)=1-\alpha$$

oziroma za enostranski test v temle primeru

$$P\left(M\le\bar{y}+\frac{t_{n-1,\alpha}\cdot\sigma}{\sqrt{n}}\right)=1-\alpha$$

Pri tem je *n* število elementov v vzorcu, za ta primer je $n=10$.

Iz tabele za Studentovo porazdelitev odčitamo vrednost *t* za $k=n-1=9$ prostostnih stopenj pri stopnji zaupanja $\alpha=0{,}10$: $t_{9;0.1}=1{,}383029$.

Sedaj je

$$\bar{y}+\frac{t_{n-1,\alpha}\cdot\sigma}{\sqrt{n}}=15{,}4+\frac{1{,}383029\cdot 11{,}3}{\sqrt{10}}=20{,}3$$

To pomeni, da moremo z verjetnostjo 0,90 pričakovati, da bo vzorčna ocena aritmetične sredine manjša od 20,3 minut.

Hipotezo, da je povprečna vrednost zamud v vsakem vzorcu manjša od 15 minut, torej zavrnemo.

♦ ♦ ♦

11.2.2 Testiranje hipotez o porazdelitvi

Sedaj bomo ugotavljali, kako konkretni (stvarni) frekvenčni porazdelitvi priredimo eno od teoretičnih porazdelitev, ki smo jih spoznali v prejšnjem poglavju. Dano (stvarno) frekvenčno porazdelitev primerjamo z eno od teoretičnih porazdelitev in ugotavljamo, v kakšni zvezi sta.

Primerjajmo **grafični obliki** stvarne frekvenčne porazdelitve in teoretične porazdelitve. Takšna primerjava je možna na dva načina: primerjamo histogram stvarne frekvenčne porazdelitve z verjetnostno funkcijo teoretične porazdelitev ali pa primerjamo graf kumulativne frekvenčne porazdelitve s porazdelitveno funkcijo teoretične porazdelitve. Iz

takšnih primerjav je mogoče ugotoviti, ča sta si stvarna in teoretična porazdelitev podobni, torej če ju lahko primerjamo glede na vrednosti kazalnikov.
Ko želimo grafično primerjati stvarno frekvenčno porazdelitev z neko teoretično porazdelitvijo, moramo najprej izračunati teoretične frekvence (teoretične) porazdelitve ob domnevi, da se stvarna frekvenčna porazdelitev dejansko ujema z izbrano teoretično porazdelitvijo.
V nadaljevanju bomo pogledali več možnosti.

11.2.2.1 Prilagoditev normalne porazdelitve stvarni frekvenčni porazdelitvi

Osnovni princip primerjave (prilagoditve) teoretične porazdelitve stvarni frekvenčni porazdelitvi je ta, da izenačimo aritmetično sredino in varianco stvarne porazdelitve z matematičnim upanjem in varianco teoretične porazdelitve. Drugače povedano: stvarni frekvenčni porazdelitvi s parametroma $\bar{y}$ in σ prilagodimo normalno porazdelitev $N\left(\bar{y},\sigma^2\right)$. Če je pravilna domneva, da je stvarni frekvenčni porazdelitvi mogoče prilagoditi normalno porazdelitev, se stvarne frekvence ne smejo preveč razlikovati od teoretičnih frekvenc, ki jih izračunamo iz podatkov, pri čemer upoštevamo parametra normalne porazdelitve. To pomeni, da za vsak razred frekvenčne porazdelitve izračunamo teoretično frekvenco, pri čemer upoštevamo verjetnostno funkcijo normalne porazdelitve, zapisane z že od prej poznano enačbo:

$$p\left(y\right)=\varphi\left(y\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-\bar{y}}{\sigma}\right)^2} \tag{11.17}$$

V (11.17) je *y* slučajna spremenljivka, ki je porazdeljena po normalni porazdelitvi. Verjetnost, da *y* zavzame vrednost v mejah posameznega razreda, je enaka ploščini med verjetnostno funkcijo in abscisno osjo v mejah tega razreda. Če vzamemo za predstavnika k-tega razreda ($k=1,2,\ldots,r$) kar sredino tega razreda, to je y_k ter z i_k označimo njegovo širino, lahko zapišemo

$$P\left(y_{k,\min}<y<y_{k,\max}\right)=i_k\cdot\varphi\left(y_k\right)$$

Ko uporabimo standardizirano normalno porazdelitev, torej substitucijo

$$z_k=\frac{y_k-\bar{y}}{\sigma} \tag{11.18}$$

in upoštevamo znano zvezo $\varphi\left(y_k\right)=\frac{1}{\sigma}\varphi\left(z_k\right)$, dobimo enačbo

$$P\left(y_{k,\min} < y < y_{k,\max}\right) = i_k \cdot \frac{1}{\sigma}\varphi\left(z_k\right) \tag{11.19}$$

Teoretična frekvenca za število statističnih enot v mejah k-tega razreda je

$$f_{t,k} = N \cdot P\left(y_{k,\min} < y < y_{k,\max}\right) = \frac{N \cdot i_k}{\sigma} \cdot \varphi\left(z_k\right) \tag{11.20}$$

Primer:
Vzemimo, da na nekem letališču merimo zamude vzletov letal v nekem konkretenem dnevu. Čase merimo v minutah. Meritve prikažemo v obliki frekvenčne porazdelitve:

Razred (zamuda v minutah)	Frekvenca f_k
5,00- 10,00	7
10,01-15,00	15
15,01-20,00	22
20,01-25,00	26
25,01-30,00	16
30,01-35,00	11
35,01-40,00	3
skupaj	100

Izračunajmo povprečno vrednost in varianco te porazdelitve.

Rešitev:
Tabelo dopolnimo z nekaterimi stolpci, kot vidimo v nadaljevanju.

Razred (zamuda v minutah)	f_i	F_i	y_i (sredina razreda)	$f_i y_i$	y_i^2	$f_i y_i^2$
5,01- 10,00	7	7	7,5	52,5	56,25	393,75
10,01-15,00	15	22	12,5	187,5	156,25	2343,75
15,01-20,00	22	44	17,5	385,0	306,25	6737,50
20,01-25,00	26	70	22,5	585,0	506,25	13162,50
25,01-30,00	16	86	27,5	440,0	756,25	12100,00
30,01-35,00	11	97	32,5	357,5	1056,25	11618,75
35,01-40,00	3	100	37,5	112,5	1406,25	4218,75
skupaj	100	-	-	2120,0	4243,75	50575,00

Uporabimo formule (5.10), (6.13) in (6.15):

$$M_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{r} f_i y_i = \frac{2120}{100} = 21,2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{r} f_i y_i^2 - M_y^2 = \frac{50575}{100} - 21,2^2 = 56,31$$

Upoštevamo še Shepardov popravek

$$\sigma_{yp}^2 = \sigma_y^2 - \frac{i^2}{12} = 56,31 - \frac{25}{12} = 54,23$$

Povprečna zamuda v natanko določenem dnevu, ko smo izvajali meritev, je $\bar{y} = 21,2$ minut, varianca pa 54,23 kvadratnih minut (opomba: fizikalna enota je seveda nesmiselna) oz. standardni odklon $\sigma = 7,36$ minut.

Razmik $(\bar{y} - \sigma, \bar{y} + \sigma)$ je v tem primeru: $(13,84; 28,56)$. Za spodnjo in zgornjo mejo tega intervala izračunajmo kvantilna ranga:

a) $y_i = 13,84 \Rightarrow R_i = 18,45 \Rightarrow P_i = 0,18$

b) $y_i = 28,56 \Rightarrow R_i = 81,3 \Rightarrow P_i = 0,81$

To pomeni, da je v intervalu $(\bar{y} - \sigma, \bar{y} + \sigma)$ skupen delež $0,81 - 0,18 = 0,63$, torej 63% vseh podatkov.

Normalna porazdelitev predvideva (glej poglavje 6.3), da je v intervalu $(\bar{y} - \sigma, \bar{y} + \sigma)$ skupno 68,3% podatkov.

Vidimo, da je razlika glede števila podatkov v intervalu $(\bar{y} - \sigma, \bar{y} + \sigma)$ med porazdelitvijo v danem primeru in normalno porazdelitvijo majhna, zato lahko sklepamo, da se zamude iz našega primera pokoravajo normalni porazdelitvi z aritmetično sredino 21,2 in varianco 56,31, torej $y \sim N(21,2; 56,31)$.

Če torej privzamemo, da se meritve v danem primeru prilagajo tej normalni porazdelitvi, izračunamo pripadajoče teoretične frekvence, dane s formulo (11.20).

V ta namen najprej tabelo podatkov dopolnimo s stolpcem izračunanih standardiziranih vrednosti (11.18), za katere iz tabele dobimo vrednosti $\varphi_k(z)$ in nato po formuli (11.20) izračunamo teoretične frekvence, ki jih zaokrožimo na cele vrednosti.

Takole gre:

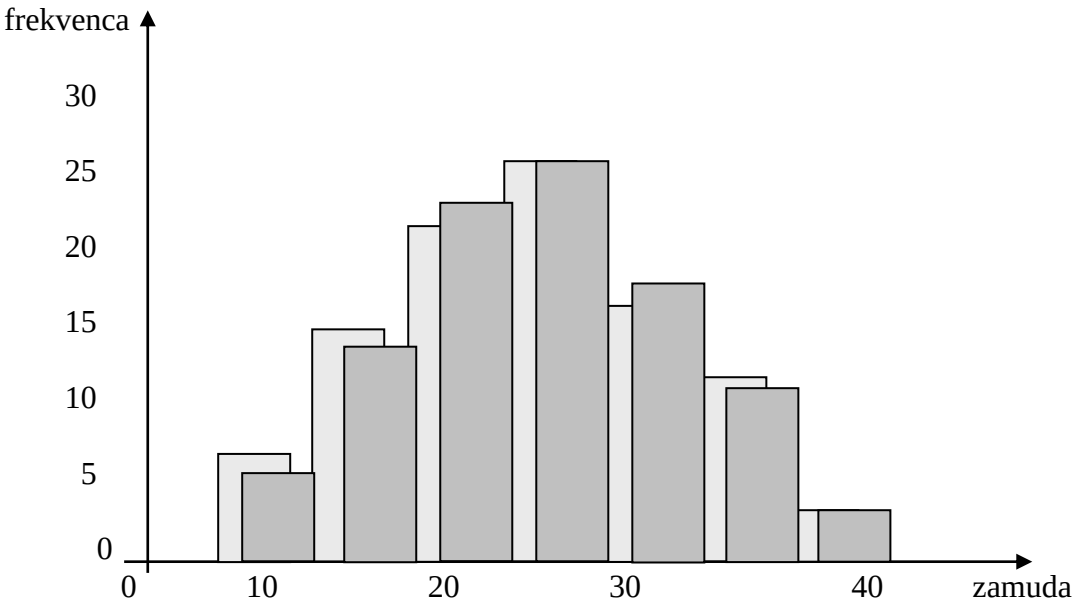
$$z_1 = \frac{y_1 - \bar{y}}{\sigma} = \frac{7,5 - 21,2}{7,36} = -1,86 \Rightarrow \varphi(-1,86) = \varphi(1,86) = 0,07074$$

$$f_{t,1} = \frac{N \cdot i_1}{\sigma} \cdot \varphi(z_1) = \frac{100 \cdot 4,99}{7,36} \cdot 0,07074 = 4,80$$

Na takšen način izpolnimo tabelo do konca.

Razred (zamuda v minutah)	f_i	y_i (sredina razreda)	z_i	$\varphi(z_i)$	$f_{t,i}$ (zaokroženo na cela mesta)
5,01- 10,00	7	7,5	-1,86	0,07074	5
10,01-15,00	15	12,5	-1,18	0,19886	14
15,01-20,00	22	17,5	-0,50	0,35207	24
20,01-25,00	26	22,5	0,18	0,39253	26
25,01-30,00	16	27,5	0,86	0,27562	18
30,01-35,00	11	32,5	1,53	0,12376	10
35,01-40,00	3	37,5	2,21	0,03470	3
skupaj	100	-			100

Še grafični prikaz, kjer močneje osenčeni stolpci predstavljajo teoretične frekvence normalne porazdelitve.



♦ ♦ ♦

Sedaj smo porazdelitve primerjali grafično. Iz grafa vidimo, da so stvarne frekvence in teoretične frekvence »na oko« precej podobne. Seveda gre za subjektivno primerjavo, ki je lahko tudi varljiva, zato potrebujemo neodvisno in objektivno testiranje, kar bomo opravili z uporabo testa hi kvadrat.

11.2.2.2 Test χ^2 (hi kvadrat)

Imejmo vrednosti statistične spremenljivke urejene v frekvenčni porazdelitvi z *r* razredi. Za vsaj razred naj bo dana stvarna frekvenca f_i, za vsak razred tudi izračunajmo teoretično frekvenco $f_{t,i}$ teoretične porazdelitve, za katero domnevamo, da jo moremo prilagoditi stvarni porazdelitvi. Razlike med stvarnimi in teoretičnimi frekvencami so lahko večje ali manjše, lahko so tudi nič. Naša naloga je ugotoviti, če so te razlike dovolj majhne, da jih statistično lahko dopustimo ali ne. Gre torej za statistični preizkus ničelne hipoteze H_0, ki jo lahko definiramo takole:

H_0 : porazdelitev spremenljivke *y*, na katero se nanašajo stvarni podatki, se pokorava prilagojeni teoretični porazdelitvi z določeno vrednostjo parametrov.

Nasprotna hipoteza je lahko marsikaj, zato jo običajno v takšnih situacijah ne postavljamo.

Statistični preizkus hipotez takšne oblike temelji na Pearsonovem[5] izreku, ki ga povejmo brez dokaza:

Spremenljivka χ^2 (hi kvadrat)

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(f_i - f_{t,i} \right)^2}{f_{t,i}} \tag{11.21}$$

je porazdeljena približno po porazdelitvi "hi kvadrat" s številom prostostnih stopenj k, ki je odvisno od števila razredov r in od teoretične porazdelitev, ki smo jo prilagodili stvarnim podatkom.

Če označimo z *m* število statističnih parametrov, ki določajo teoretično porazdelitev, je število prostostnih stopenj enako

$$k = r - m - 1 \tag{11.22}$$

Pri prilagoditvi **normalne porazdelitve** imamo dva statistična parametra (aritmetično sredino in varianco), torej $m = 2$, zato je tedaj število prostostnih stopenj porazdelitve hi kvadrat $k = r - 3$.

Kadar so razlike med stvarnimi in teoretičnimi frekvencami majhne, potem bo tudi vrednost χ_0^2, ki jo izračunamo po Pearsonovem izreku (5.101), majhna. Kadar pa so te razlike (pre)velike, ne moremo trditi, da so nastale slučajno, zato je ničelno hipotezo smiselno zavrniti.

[5] Karl Paerson (1857-1936), angleški matematik, ki se je prvenstveno ukvarjal z matematično statistiko

Izračunana vrednost χ_0^2 je prevelika tedaj, ko pade izven območja, ko domneve ne zavračamo, določenega s stopnjo značilnosti α.
Kadar torej velja $P\left(\chi^2 > \chi_0^2\right) < \alpha$, bomo ničelno hipotezo H_0 zavrnili. Pravimo, da je v takem primeru χ_0^2 statistično značilen in razlike med stvarnimi in teoretičnim frekvencami so tudi statistično značilne. Pri tem je α verjetnost, da zavrnemo ničelno hipotezo, ki v resnici velja.

Test hi kvadrat se za statistično preizkušanje hipotez o dopustnosti prilagoditve določene teoretične porazdelitve stvarni porazdelitvi uporablja v primerih, ko so teoretične frekvence v vseh razredih večje od 5. Če temu ni tako, moramo razrede, kjer je teoretična frekvenca premajhna, združiti s sosednim razredom. Pri določanju števila prostostnih stopenj seveda potem upoštevamo število razredov, ki jih dobimo z morebitnim združevanjem.

Primer:
Preizkusimo hipotezo, da je pri stopnji tveganja $\alpha = 0,05$ stvarni frekvenčni porazdelitvi za zamude letal dopustno prilagoditi normalno porazdelitev.
Podatki so zapisani v tabeli prejšnjega primera.

Rešitev:
Preizkušamo ničelno hipotezo

H_0 : stvarna frekvenčna porazdelitev se pokorava prilagojeni normalni porazdelitvi $N\left(21,2;56,31\right)$

Po enačbi (11.21) izračunamo vrednost χ_0^2. Pred tem pa še upoštevamo, da prvi in zadnji razred nimata frekvenco večje od 5.

Razred (zamuda v minutah)	f_i	$f_{t,i}$	$\left(f_i - f_{t,i}\right)^2$	$\frac{\left(f_i - f_{t,i}\right)^2}{f_{t,i}}$
5,91- 15,90	22	19	9	0,474
15,91-20,09	22	24	4	0,167
20,91-25,09	26	26	0	0
25,91-30,09	16	18	4	0,222
30,91-40,09	14	13	1	0,077
skupaj	100	100	18	0,940

$$\chi_0^2 = \sum_{i=1}^{r} \frac{\left(f_i - f_{t,i}\right)^2}{f_{t,i}} = 0,940$$

Število prostostnih stopenj je $k = r - m - 1 = 5 - 2 - 1 = 2$

Iz tabele za vrednosti hi kvadrat dobimo pri $k = 2$

$$0,50 < P\left(\chi^2 > 0,940\right) < 0,70$$

Verjetnost $P\left(\chi^2 > \chi_0^2\right)$ je torej večja od 0,05, zato ničelne hipoteze pri stopnji značilnosti $\alpha = 0,05$ ne zavrnemo. Izračunana vrednost χ_0^2 ni statistično značilna, to pomeni, da razlike med stvarnimi in teoretičnim frekvencami niso prevelike in jih moremo pripisati slučajnim vplivom.

♦ ♦ ♦

TABELE

GRŠKA ABECEDA

velike črke	male črke	slovensko ime	angleško ime
Α	α	alfa	Alpha
Β	β	beta	Beta
Γ	γ	gama	Gamma
Δ	δ	delta	Delta
Ε	ε	epsilon	Epsilon
Ζ	ζ	ceta	Zeta
Η	η	eta	Eta
Θ	θ	theta	Theta
Ι	ι	jota	Iota
Κ	κ	kapa	Kappa
Λ	λ	lambda	Lambda
Μ	μ	mi	Mu
Ν	ν	ni	Nu
Ξ	ξ	ksi	Xi
Ο	ο	omikron	Omicron
Π	π	pi	Pi
Ρ	ρ	ro	Rho
Σ	σ, ς	sigma	Sigma
Τ	τ	tau	Tau
Υ	υ	ipsilon	Upsilon
Φ	ϕ, φ	fi	Phi
Χ	χ	hi	Chi
Ψ	ψ	psi	Psi
Ω	ω	omega	Omega

Vrednosti funkcije $\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$

z	0	1	2	3	4	5	6	7	8	9
0.0	0.39890	39892	39886	39876	39862	39844	39822	39797	39767	39733
0.1	0.39695	39654	39608	39559	39505	39448	39387	39322	39253	39181
0.2	0.39104	39024	38940	38853	38762	38667	38568	38466	38361	38251
0.3	0.38139	38023	37903	37780	37654	37524	37391	37255	37115	36973
0.4	0.36827	36678	36526	36371	36213	36053	35889	35723	35553	35381
0.5	0.35207	35029	34849	34667	34482	34294	34105	33912	33718	33521
0.6	0.33322	33121	32918	32713	32506	32297	32086	31874	31659	31443
0.7	0.31225	31006	30785	30563	30339	30114	29887	29659	29431	29200
0.8	0.28969	28737	28504	28269	28034	27798	27562	27324	27086	26848
0.9	0.26609	26369	26129	25888	25647	25406	25164	24923	24681	24439
1.0	0.24197	23955	23713	23471	23230	22988	22747	22506	22265	22025
1.1	0.21785	21546	21307	21069	20831	20594	20357	20121	19886	19652
1.2	0.19419	19186	18954	18724	18494	18265	18037	17810	17585	17360
1.3	0.17137	16915	16694	16474	16256	16038	15822	15608	15395	15183
1.4	0.14973	14764	14556	14350	14146	13943	13742	13542	13344	13147
1.5	0.12952	12758	12566	12376	12188	12001	11816	11632	11450	11270
1.6	0.11092	10915	10741	10567	10396	10226	10059	09893	09728	09566
1.7	0.09405	09246	09089	08933	08780	08628	08478	08329	08183	08038
1.8	0.07895	07754	07614	07477	07341	07206	07074	06943	06814	06687
1.9	0.06562	06438	06316	06195	06077	05959	05844	05730	05618	05508
2.0	0.05399	05292	05186	05082	04980	04879	04780	04682	04586	04491
2.1	0.04398	04307	04217	04128	04041	03955	03871	03788	03706	03626
2.2	0.03547	03470	03394	03319	03246	03174	03103	03034	02965	02898
2.3	0.02833	02768	02705	02643	02582	02522	02463	02406	02349	02294
2.4	0.02239	02186	02134	02083	02033	01984	01936	01888	01842	01797
2.5	0.01753	01709	01667	01625	01585	01545	01506	01468	01431	01394
2.6	0.01358	01323	01289	01256	01223	01191	01160	01130	01100	01071
2.7	0.01042	01014	00987	00961	00935	00909	00885	00861	00837	00814
2.8	0.00792	00770	00748	00727	00707	00687	00668	00649	00631	00613
2.9	0.00595	00578	00562	00545	00530	00514	00499	00485	00470	00457
3.0	0.00443	00430	00417	00405	00393	00381	00370	00358	00348	00337
3.1	0.00327	00317	00307	00298	00288	00279	00271	00262	00254	00246
3.2	0.00238	00231	00224	00216	00210	00203	00196	00190	00184	00178
3.3	0.00172	00167	00161	00156	00151	00146	00141	00136	00132	00127
3.4	0.00123	00119	00115	00111	00107	00104	00100	00097	00094	00090
3.5	0.00087	00084	00081	00079	00076	00073	00071	00068	00066	00063
3.6	0.00061	00059	00057	00055	00053	00051	00049	00047	00046	00044
3.7	0.00042	00041	00039	00038	00037	00035	00034	00033	00031	00030
3.8	0.00029	00028	00027	00026	00025	00024	00023	00022	00021	00021
3.9	0.00020	00019	00018	00018	00017	00016	00016	00015	00014	00014
4.0	0.00013	00013	00012	00012	00011	00011	00011	00010	00010	00009

Vrednosti funkcije $\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}t^2}\,dt$

x	0	1	2	3	4	5	6	7	8	9
0.0	0.00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0.1	0.03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0.2	0.07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0.3	0.11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0.4	0.15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0.5	0.19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0.6	0.22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0.7	0.25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0.8	0.28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0.9	0.31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1.0	0.34134	34375	34614	34849	35083	35314	35543	35769	35993	36214
1.1	0.36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1.2	0.38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1.3	0.40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1.4	0.41924	42073	42220	42364	42507	42647	42785	42922	43056	43189
1.5	0.43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1.6	0.44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1.7	0.45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1.8	0.46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1.9	0.47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2.0	0.47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2.1	0.48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2.2	0.48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2.3	0.48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2.4	0.49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2.5	0.49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2.6	0.49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2.7	0.49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2.8	0.49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2.9	0.49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3.0	0.49865	49869	49874	49878	49882	49886	49889	49893	49896	49900
3.1	0.49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3.2	0.49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3.3	0.49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3.4	0.49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3.5	0.49977	49978	49978	49979	49980	49981	49981	49982	49983	49983
3.6	0.49984	49985	49985	49986	49986	49987	49987	49988	49988	49989
3.7	0.49989	49990	49990	49990	49991	49991	49992	49992	49992	49992
3.8	0.49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3.9	0.49995	49995	49996	49996	49996	49996	49996	49996	49997	49997
4.0	0.49997	49997	49997	49997	49997	49997	49998	49998	49998	49998

Vrednosti funkcije za porazdelitev hi kvadrat

df\area	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.34840	11.34487
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.14329	13.27670
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.07050	12.83250	15.08627
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.64464	12.59159	14.44938	16.81189
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100	13.70069	17.27501	19.67514	21.92005	24.72497
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032	14.84540	18.54935	21.02607	23.33666	26.21697
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.33976	15.98391	19.81193	22.36203	24.73560	27.68825
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.33850	19.36886	23.54183	26.29623	28.84535	31.99993
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711	30.19101	33.40866
18	6.26480	7.01491	8.23075	9.39046	10.86494	13.67529	17.33790	21.60489	25.98942	28.86930	31.52638	34.80531
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.56200	18.33765	22.71781	27.20357	30.14353	32.85233	36.19087
20	7.43384	8.26040	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043	34.16961	37.56623
21	8.03365	8.89720	10.28290	11.59131	13.23960	16.34438	20.33723	24.93478	29.61509	32.67057	35.47888	38.93217
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444	36.78071	40.28936
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.13730	22.33688	27.14134	32.00690	35.17246	38.07563	41.63840
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503	39.36408	42.97982
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248	40.64647	44.31410
26	11.16024	12.19815	13.84390	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514	41.92317	45.64168
27	11.80759	12.87850	14.57338	16.15140	18.11390	21.74940	26.33634	31.52841	36.74122	40.11327	43.19451	46.96294
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714	44.46079	48.27824
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697	45.72229	49.58788
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297	46.97924	50.89218

Vrednosti funkcije za Studentovo t porazdelitev

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460

Vir: http://www.statsoft.com/textbook/distribution-tables/#t

REFERENCE

1. Jesenko, J. (2001). *Statistika v organizaciji in management*. Kranj: Moderna organizacija.
2. Tominc, P. (2000). *Statistične metode: uporaba v prometu*. V Mariboru: Fakulteta za gradbeništvo.
3. Artenjak, J. (2000). *Poslovna statistika*. Maribor: Ekonomsko-poslovna fakulteta.
4. Usenik, J. (2017). *Kvantitativne metode*. Fakulteta za energetiko UM; zapiski predavanj.
5. Usenik, J., Vidiček, M. (2018). *Poslovni račun in statistika*. Elektronska izd. Novo mesto: Visoka šola za upravljanje podeželja Grm.

Založniška dejavnost
Fakultete za organizacijske študije v Novem mestu
I Z D A N E M O N O G R A F I J E

1. **Izbrana poglavja iz matematike** (e-knjiga)
   Usenik, Janez 2020
2. **Vpliv uporabe orodij managerjev na ekonomsko donosnost**
   Markič, Mirko; Kreslin, Damijan 2019
3. **Delovni terapevt v inkluzivni šoli: Trenutno stanje in smernice**
   Šuc, Lea 2019
4. **Zrna odličnosti Fakultete za organizacijske študije v Novem mestu: nove paradigme organizacijskih teorij 2018**
   Bukovec, Boris (ur.) 2019
5. **Zrna odličnosti Fakultete za organizacijske študije v Novem mestu: nove paradigme organizacijskih teorij 2017**
   Bukovec, Boris (ur.) 2018
6. **Značilnosti sistemov vodenja kakovosti v slovenskih organizacijah in njihov vpliv na poslovno uspešnost organizacij**
   Vinko Bogataj; Gordana Žurga; Adolf Šostar 2018
7. **Menedžment kakovosti in odličnost zdravnikov v javnem zdravstvu**
   Rumpf, Dean; Voga, Gorazd; Meško Štok, Zlatka 2018
8. **Sistemi vodenja kakovosti in modeli odličnosti: ključni dejavniki (ne)uspešnega delovanja** (e-knjiga)
   Škafar, Branko 2018
9. **Temelji avtopoieze v orgaizaciji: Avtopoietska 4.0 (r)evolucija človeka**
   Balažič Peček, Tanja; Bukovec, Boris 2018
10. **Management varnosti pri delu, delovne razmere in gospodarska učinkovitost** (e-knjiga)
    Pavlič, Miran, Markič, Mirko 2018
11. **Kvalitativno raziskovanje koncepta avtopoieze v organizaciji** (e-knjiga)
    Balažic Peček, Tanja 2018
12. **Podlage in metode za raziskovanje in projektiranje organizacije**
    Ivanko, Štefan 2017
13. **Celostna obravnava dolgotrajne oskrbe v Sloveniji**
    Kavšek, Marta; Bogataj, David 2017
14. **Sodelovalno mreženje in izraba inovacijskega potenciala v turističnem prostoru**
    Colarič-Jakše, Lea-Marija 2017

15. **Model McKinsey 7-S kot kazalnik odličnosti organizacije**
Kalan, Mateja; Meško, Maja 2017
16. **Nova doktrina organizacije – 2.del: Preusmeritev pozornosti**
Ovsenik, Jožef; Ovsenik, Marija 2017
17. **Avtopoietska organizacija**
Bukovec, Boris (ur.) 2017
18. **Kakovost v slovenski javni upravi: Delovanje Odbora za kakovost 1999-2012**
Žurga, Gordana 2017
19. **Poslovne vrednote mladih v Sloveniji**
Pinterič, Uroš 2016
20. **Selected topics in modern society** (e-knjiga)
Kaplánová, Patrícia 2016
21. **Glocalisation of the crisis: could Slovenia survive economic crisis better?** (e-knjiga)
Pinterič, Uroš 2016
22. **Zrna odličnosti Fakultete za organizacijske študije v Novem mestu: nove paradigme organizacijskih teorij 2016**
Bukovec, Boris (ur.) 2016
23. **Pisanje strokovnih in znanstvenih del**
Brcar, Franc 2016
24. **Education policy as the factor of development** (e-knjiga)
Pinterič, Uroš 2016
25. **Spregledane pasti informacijske družbe**
Pinterič, Uroš 2015
26. **Psihosocialni dejavniki tveganja za bolečino v križu pri slovenskih poklicnih voznikih in absentizem**
Kresal, Friderika; Meško, Maja 2015
27. **Zgodovina organizacijske misli**
Ivanko, Štefan 2015
28. **Karierno načrtovanje: kako najti v sebi skriti zaklad?**
Turnšek Mikačić, Marija; Ovsenik, Marija 2015
29. **Sodobni trendi v turizmu**
Ovsenik, Rok 2015
30. **Selected topics in change management** (e-knjiga)
Kaplánová, Patrícia (ur.); Pinterič, Uroš (ur.) 2015
31. **Political legacy and youth civic engagement in Slovakia**
Mihálik, Jaroslav 2015
32. **Turistični prostori različnosti: turizem, turisti in fotografska podoba**
Ambrož, Milan; Bukovec, Boris 2015
33. **Izobraževanje za turizem v Sloveniji**
Ovsenik, Rok; Bukovec, Boris; Ovsenik, Marija 2015

34. **Zrna odličnosti Fakultete za organizacijske študije v Novem mestu: nove paradigme organizacijskih teorij 2015**
Bukovec, Boris (ur.) 2015
35. **Kontrolna teorija sistemov: model za sistemsko razmišljanje v sistemu zdravstvenega varstva**
Mlakar, Tatjana 2014
36. **Featuring Norden in ten episodes**
Czarny, Ryszard M. 2014
37. **Sociológia mládeže**
Macháček, Ladislav 2014
38. **Inter-municipal cooperation in Slovakia : the case of regions with highly fragmented municipal structure**
Klimovsky, Daniel 2014
39. **Rethinking public policies** (e-izdaja)
Pinterič, Uroš 2014
40. **Local Governance between democracy and efficiency**
Jüptner, Petr. (et al.) 2014
41. **Pasti razumevanja politične realnosti : pregled konceptov sodobnega političnega sistema**
Pinterič, Uroš 2014
42. **Selected issues of administrative reality**
Pinterič, Uroš; Prijon, Lea 2013
43. **Organizacijske paradigme: podlage za nastanek in razvoj organizacijskih teorij**
Ivanko, Štefan 2012

RUO

Revija za univerzalno odličnost

Journal of Universal Excellence

Letnik x, številka x, mesec 20xx

Volume x, Issue x, Month 20xx

interdisciplinarna revija, ki združuje organizacijske vede oz. menedžment in univerzalno odličnost, tj. poslovno, organizacijsko in osebno odličnost

znanstvena revija, ki poskuša odgovarjati na ključna vprašanja družbene teme pri čemer akademsko rigoroznost nadgrajuje z inovativnostjo v tematikah in pristopu.

ISSN 24

Izzivi prihodnosti

Challenges of the Future

Letnik x, številka x, mesec

Volume x, Issue x, Month 20xx