OSNOVE STATISTIKE – 1. DEL

OPREDELITEV STATISTIKE

Klasična "učbeniška" definicija statistike opredeljuje statistiko kot vedo, ki se ukvarja s proučevanjem množičnih pojavov. Vedeti pa moramo, da enkratni pojav **ni** množični, prav tako podatke o eni osben **niso** statistični podatki.
Predmet statistike je **reduciranje podatkov**, kar pomeni, da jih združujemo po rangih glede na eno izbrano spremenljivko.
Statistika je veda, ki s kvantitativnim proučevanjem množičnih pojavov, z metodami, ki so njej lastne, odkriva zakonitosti množičnega pojavljanja in podaja kvalitativno analizo pojavov. Statistika niso samo matematični postopki, saj moramo vsako številko tudi interpretirati in ugotoviti kaj pomeni.

Statistika je eno najpomembnejših orodji večine znanstvenih disciplin in jo uporablja večina strok, za sistematiko podatkov, proučevanje kvalitete, tržne raziskave, značilnosti uporabnikov itd., uporabljamo pa jo tudi v vsakdanjem življenju, kot čisto preprosto interpretacijo enostavnih postopkov.
Statistiko lahko v bibliotekarstvu, informatiki in založništvu uporabimo za interpretiranje števila uporabnikov, strukture kupcev knjig po izobrazbi, zadovoljstva uporabnikov s storitvami – razlike glede na spol, kvalitete storitve, uspešnosti izobraževanja, uspešnost promocije, povprečne starosti uporabnikov itd.

Osonovni pojmi:
POPULACIJA – statistična množica, ki jo proučujemo
PARAMETER – lastnost statistične množice (populacije)
STATISTIČNA ENOTA – enota množice, ki jo proučujemo
SPREMENLJIVKA – lastnost statistične enote
VZOREC – del populacije

Naloge statistike so, da OPISUJE POJASNJUJE NAPOVEDUJE

PODROČJI STATISTIKE

1. DESKRIPTIVNA (OPISNA) STATISTIKA

Pojave opisuje. Zanima jo neka frekvenca; število uporabnikov/povprečno število uporabnikov na programih informacijskega izobraževanja.
Podatke deskriptivne statistike lahko obdelujemo z rangiranjem, določanjem mer srednjih vrednosti in mer razpršenosti.

2. INFERENČNA (POJASNJEVALNA) STATISTIKA

Pojave razlaga in napoveduje. Zanima jo npr. Stopnja zadovoljstva s knjižnično zbirko glede na spol (kako se uporabniki razlikujejo).
Podatke inferenčne statistike lahko obdelujemo s korelacijo, x kvadratom, t-testi, ANOVA ali regresijsko analizo.

POPULACIJA
STATISTIČNA MNOŽICA

Opredelitev statistične množice pomeni opredeliti, kdo vanjo spada in kdo ne. Statistično množico opredelimo na tri načine:

1. Stvarno – kdaj ali kdo so enote te množice
2. Krajevno – določimo geografsko razsežnost množice
3. Časovno – določimo čas v katerem zajemamo množico

PRIMER: število obiskov uporabnikov splošne knjižnice v Sloveniji v času poletnih počitnic (junij, julij, avgust).

ENOTA STATISTIČNE MNOŽICE

Vsaka enota statistične množice (populacije) ima vrsto lastnosti. **Spremenljivka** je lastnost enote statistične množice.

SPREMENLJIVKE

Spremenljivka ali statistični znak je lahko spol, zakonski stan, stopnja izobrazbe, gradivo, spletni informacijski vir, višina dohodka, višina sredstev za nakup literature, število kupcev, vrste uporabnikov, dohodki, knjižničar ...
Spremenljivke delimo glede na tip vrednosti, glede na merski nivo – kvaliteto merjenja in glede na vlogo.

SOREMENLJIVKE GLEDE NA TIP VREDNOSTI

1. **Opisne** (atributivne, kategorialne):
   a. Spol
   b. Izobrazna
   c. Bralni interes

2. **Številske**
   a. Zvezne – vse vrednosti v intervalu (starost, finančna sredstva za nakup knjige)
   b. Nezvezne – samo nekatere vrednosti v intervalu (število založb, pogost obiska)

SPREMENLJIVKE GLEDE NA MERSKI NIVO

1. **Nominalne –** imenske

Ureja spremenljivke po značilnostih, kategorije niso logično urejene po velikost, ampak jih lahko samo poimenujemo in štejemo. Opredeljujejo pogostost (frekvenco) pojava: koliko je v raziskavi moških in koliko žensk.

PRIMER 1: spremenljivka = motivacija za branje (dobra, slaba, odlična ...)

PRIMER 2: nominalne spremenljivke z **dvema** kategorijama; spol = M/Ž ali plačane članarine= DA/NE

PRIMER 3: nominalne spremenljivke z **več** kategorijami; studijski program = bibliotekarstvo /medicina/pravo/... ali literatura = priročniki/monografije/revije/...

2. **Ordinalne –** urejenostne
   Spremenljivke so logično urejene po enostavnem načinu rangiranja po velikosti od večjega do manjšega ali obratno. Stopnje so določene na osnovi količine značilnosti, ki jo posedujejo, vendar razlike niso enako velike (imajo neko vrednost).

   PRIMER 1: stopnja izobrazbe = prva/druga/tretja/četrta/peta/šesta

   PRIMER 2: pogostost branja = redko/včasih/pogosto/zelo pogosto

   Ali so šolske ocene ordinalna lestvica? Da, ker so logično urejene po vrsti.

3. **Intervalne –** razmične
   Spremenljivke so logično urejene z enakimi razmiki med vrednostmi (so bolj natančno določene). Spremenljivkam lahko določimo modus in mediano. Imajo natančno določeno mersko enoto, stopnje pa so določene na osnovi količine značilnosti, ki jo posedujejo. Razlike so velike, vendar ni ničelnega izhodišča (npr. Ne moremo reči da je 20°C dvakrat več kot 10°C)

   PRIMER 1: temperatura = ni absolutne ničle

   PRIMER 2 : višina stopnic = enaka razlika med eno stopnico in drugo

   Ali je število točk na izpitu intervalna spremenljivka? Ne, ker ima ničelno izhodišče (lahko pišeš nič točk)

4. **Racionalne –** razmernostne
   Spremenljivke so logično urejene in imajo ničelno izhodišče. Lahko računamo korelacije in zahtevne statistične analize (tudi aritmetično sredino). Stopnje so logično urejene in določene na osnovi količine značilnosti, ki jo posedujejo. Razlike so enako velike. Ima absolutna ničla in odsotnost pojava, zato lahko zaključimo, da je pojav npr. 2x več ali manj.

PRIMER: starost ali dohodki ali število prebranih knjig
Ali je med intervalno in racionalno lestvico velika razlika? Ne, razlika je le ničelno izhodišče.

Višjega nivoja kot je spremenljivka, več
Statističnih metod lahko uporabimo.

Na višjem nivoju spremenljivke je dovoljeno
Uporabiti vse statistične metode nižjega nivoja.

PRIMER Z DVEMA VRTAMA SOPREMENLJIVK
Koliko študentov (M/Ž) imajo štipendijo med 100-200€ in 200-300€.

ŠTUDENTI (lahko so M/Ž, redni/izredni, prva/druga stopnja)
MESEČNA NAJEMNINA racionalna povprečna najemnina

PREMENLJIVKE GLEDE NA VLOGO

Odvisna spremenljivka se spreminja zaradi vpliva neodvisne spremenljivke. Za obdelavo so potrebne zahtevne statistične metode, za proučevanje vpliva ene spremenljivke na drugo lahko uporabimo zgolj eksperiment.

PRIMERI:

- Koliko časa namenjenega študiju vpliva na rezultat na izpitu
- Količina prejetega zdravila vpliva na potek bolezni
- Kako količina treninga vpliva na končni rezultat
- Kako slabo vreme vpliva na počutje človeka
- Količina porabljenega denarja vpliva na prihranke.

UREJANJE PODATKOV

Urejamo lahko tako podatke opisnih spremenljivk, kot podatke številskih spremenljivk.

Pri urejanju opisnih spremenljivk sestavljamo **frekvenčne (strukturne) tabele.** Za vsako kategorijo spremenljivke določimo **frekvenco** (pogostost), ki pa so lahko **absolutne** (število

enot v kategoriji spremenljivke), **relativne** (delež celotne množice v kategoriji spremenljivke), **kumulativne** (zbirne frekvence, ki pojasnjujejo število enot do določene stopnje) ali **kumulativne relativne** (zbirne frekvence, ki pojasnjujejo delež enot do določene stopnje).

PRIMER 1: strukturna tabela za nominalno spremenljivko

Spol	Frekvenca - f	Relativna frekvenca – f%
Moški	22	33,3
Ženski	66	66,7
Skupaj	88 **N - numerus**	100

Numerus (N) – velikost vzorca oz. število statističnih enot v vzorcu

PRIMER 2: strukturna tabela za ordinarno spremenljivko

Obisk knjižnice	f	f%	F%
Nikoli	35	10,5	0
Redko	97	29,2	10,5
Včasih	43	13	39,7
Pogosto	157	47,3	52,7
Skupaj	332	100	100

PRIMER 3: strukturna križna tabela

Obisk knjižnice	Srednja izobrazba	Visoka izobrazba	Skupaj
Nikoli	22	13	35
Redko	57	40	97
Včasih	23	20	43
Pogosto	27	130	157
Skupaj	129	203	332

Križna tabela križa vrednosti več spremenljivk in ne predstavlja vsake celice posebej. Iz križne tabele izluščimo najvišje in najnižje vrednosti.

Pri urejanju številskih spremenljivk sestavljamo **ranžirne vrste** in **frekvenčne porazdelitve.**

RANŽIRNE VRSTE

PRIMER 1: Ranžirna vrsta za število dni izposoje (**absolutni rangi**)

x	3	6	9	10	12	13	14	15	17	21	23	33	45
rang	1	2	3	4	5	6	7	8	9	10	11	12	13

PRIMER 2: Ranžirna vrsta za število dni izposoje (**vezani rangi**)

x	3	6	9	10	12	12	14	15	17	17	23	33	45
rang	1	2	3	4	5,5	5,5	7	8	9,5	9,5	11	12	13

Vezane range dobimo takrat, kadar se pojavi več enakih vrednosti. Rang v tem primeru določimo s povprečno vrednostjo.

FREKVENČNA PORAZDELITEV

Tabela x: Število dni izposoje

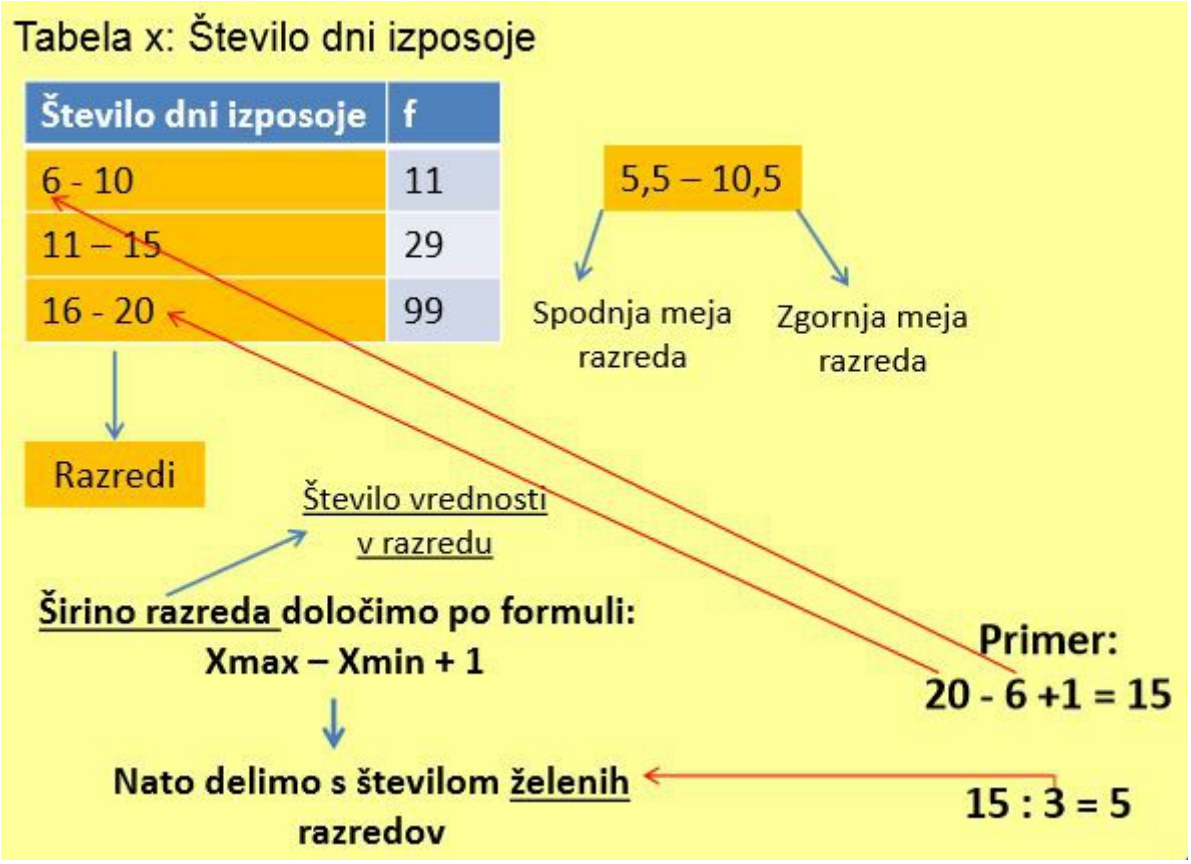
Število dni izposoje	f
7	4
10	7
14	21
15	8
17	4
20	95

Preglednica

Tabela x: Število dni izposoje

Število dni izposoje	f
6 - 10	11
11 – 15	29
16 - 20	99

Frekvenčna porazdelitev je urejanje podatkov v razrede



RANGI

Rangi določijo položaj posameznega rezultata glede na ostale. Poznamo več različnih vrst rangov. **Absolutni rangi** (razvrstimo jih po vrstnem redu, primerjava pa je možna samo znotraj skupine), **relativni rangi** (kolikšen del skupine je pod ali nad vrednostjo), **kvantil** (je vrednost spremenljivke pri kateri smo določali rang) in **kvantilni rang** (je relativni rang).

Kvantilni rangi (P) razdelijo podatke na več delov, višje vrednosti pomenijo višji položaj rezultata, nižje vrednosti pa nižji položaj. Obstajajo 4 različni kvantilni rangi polovice, kvartili, decili in centili.

	Oznaka	Kvantilni rang
ntil	C1	0,01
centil	C2	0,02
entil	C3	0,03
centil	C4	0,04
ntil	C5	0,05
entil	C6	0,06
	...	...
sti centil	C15	0,15

Primeri kvantilnih rangiov:
Mediana Me = polovica je nižjih vrednosti, polovice višjih
Kvartil Q1 = četrtina je nižjih vrednosti, tri četrtine pa višjih
Decili D6 = šest desetin je nižjih vrednosti, štiri desetine pa višjih
Centil C85 = 85% je nižjih vrednosti, 15% pa višjih

DOLOČANJE KVANTILNIH RANGOV

Rang

$P = R - 0,5/N$ → Numerus

Kvantilni rang = relativni rang

Oseba	Izdatek v €	R	P
		0,5	0
1	23	1	0,08
2	26	2	0,25
3	43	3	0,42
4	55	4	0,58
5	57	5	0,75
6	63	6	0,92
		6,5	1

DOLOČANJE KVANTILOV

Določimo vrednost 50.centila

Oseba	Izdatek v €	R	P
		0,5	0
1	23	1	0,08
2	26	2	0,25
3	43	3	0,42
4	55	4	0,58
5	57	5	0,75
6	63	6	0,92
		6,5	1

Vrednost 50.centila = kvantil

$R_p = N \times P + 0,5 = 6 \times 0,50 + 0,5 = 3,5$

$x_{0,50}=x_0+(x_1-x_0)x(-R_0)= 43 + (55-43)\ (3,5 - 3) = \mathbf{49}$

MERE SREDNJIH VREDNOSTI

Ko želimo primerjati množice med seboj, uporabimo mede rednjih vrednosti (mediana – Me, modus – Mo in aritmetična sredina – M). Mediana razdeli podatke na polovico; z modiano iščemo vrednosr z najvišjo frekvenco; z aritmetično sredino izračunamo povprečno vrednost.

Mediana je srednja vrednost po položaju. Ni občutljiva za skrajnosti, ko imamo velika odstopanja med visokimi in nizkimi vrednosti. Določamo jih pri lihem in sodem številu podatkov. Ker ne upoštevamo vseh vrednosti, primerjava med množicami ni mogoča.

Modus je najpogostejša vrednost - vrednosti spremenljivke so najbolj zgoščene. Je ne računamo, jo samo določimo. Lahko imamo eno-, dvo-, večmodalne porazdelitev, glede na število modusov. Tudi na modus skrajne vrednosti ne vplivajo. Ker ne upoštevamo vseh vrednosti, primerjava med množicami ni mogoča.

Aritmetična sredina (povprečje) je seštevek vseh rednosti, ki ga delimo s številom vrednosti. Je občutljiv za skrajne vrednosti. Ker upošteva vse vrednosti jo lahko uporabimo za primerjavo med množicami.

PRIMER 1: Izmisli si podatke, s katerimi lahko izračunaš vse 3 mere srednjih vrednosti.

X	8	9	10	10	10	10	11	15	15	17	20	21	22	22	22

Me=15; Mo=10; M=14.8

PRIMER 2: Primerjaj med seboj podatke moške (10 oseb) in ženske (10 oseb) skupine v tem, koliko porabijo v tehnični trgovini.

M	25	30	40	50	50	50	51	55	65	70
Ž	25	25	25	28	30	31	32	40	41	45

M za moške=48.6; M za ženske=32.2
Uporaimo lahko samo aritmetično sredina, saj je edina mera srednjih vrednosti, ki zajame vse podatke in nam s tem omogoča primerjavo dveh skupin.

PRIMER 3: Podatki za 10 dni spremljanja števila kupcev v manjši knjigarni.

X	11	12	16	22	37	43	45	51	55	78

Mo=/; Me=41; M=37

!Mediana je najnižja v hierarhiji. Če imamo multimodalno porazporeditev lahko iščemo samo moduse.

MERE RAZPRŠENOSTI

Mere razpršenosti opredeljujejo različnost vrednosti premenljivke in povejo ali so podatki bolj homogeni ali heterogeni. Več kot imamo enot, večja je možnost za raznolikost.

Z merjenjem razpršenost **nominalnih spremenljivk** ugotavljamo ali so spremenljivke enake ali različne in koliko ta različnost je.
Razlika med **ordinalnimi spremenljivkami** ni merljiva, lahko jih samo primerjamo.
Razlike med intervali **intervalne spremenljivke** lahko merimo (značilnost intervalne lestvice).

1. **ABSOLUTNE MERE RAZPRŠENOSTI**
   a. *Vrijacijski razmik* - je preprosta mera razpršenosti, deloma v povezavi z modusom in mediano. Upoštevamo najvišjo in najnižjo vrednost spremenljivke. Slabost varijacijskega razmika je, da upoštevamo ekstremne/izstopajoče/netipične vrednosti, ker pomeni, da ni najbolj reprezentativen.

      $VR=x_{max}-x_{min}$

Kupec	Poraba v evrih	Variacijski razmik
Prvi	2,21,33,45,47,55	55-2=53
Drugi	12,33,33,34,35,65	65-12=53

   b. *Decilni razmik* – z decilnim razmikom izločimo 10% najvišjih in 10% najnižjih vrednosti (ki so ekstremne vrednosti), ostane pa nam 80% vrednosti, ki so bližje aritmetični sredini. Spodnja meja je 1. decil, zgornja meja pa 9. decil.

      DR=D9-D1

2	5	7	11	14	15	17	18	18	21

      DR=18-5=13

   c. *Kvartilni razmik* – izločimo 25% ekstremnih vrednosti, ostane pa nam 50% vrednosti bližje aritmetični sredini. Nima tolikšne uporabne vrednosti kot decilni razmik, saj tukaj ostane samo središčna vrednost. Spodnja meja je 1. kvartil, zgornja meja je 3. kvartil.

      QR=Q3-Q1

1	3	5	6	6	7	11	12

      QR=7-5=2

   d. *Povprečni absolutni odklon*- pri povprečnem absolutnem odklonu upoštevamo vse vrednosti v množici in s tem izmerimo vse razlike. Uporabljamo ga na najvišjem nivoju podatkov. Izberemo si izhodiščno točko (povprečje) in merimo

individualne odklone. Vse individualne odklone seštejemo, delimo s številom vseh (N) in dobimo povprečni absolutni odklon.

Vrednosti nakupa v evrih	|x - M|
2	6
6	2
8	0
9	1
15	7
N = 5 nakupov M = 8 evrov	Σ|x - M| = 16

$$PO_M = \Sigma|x - M|/N = 16: 5 = 3,2$$

e. *Standardni odklon* – standardni odklon upošteva vse vrednosti v množici in je vedno vezan na aritmetično sredino. Izberemo izhodiščno točko (povprečje) in merimo individualne odklone (tako kot pri povprečnem absolutnem odklonu). Individualne odklone kvadriramo in določimo povprečni kvadrat varianca
Varianca se uporablja pri povezanosti spremenljivk, zato varianco korenimo in dobimo standardni odklon (SD). Standardni odklon nam pomaga samo v primeru, da ga lahko primerjamo s standardnim odklonom neke druge skupine (torej, potrebujemo dve skupini, ki ju med sabo primerjamo)

2. **RELATIVIVNE MERE RAZPRŠENOSTI**
   a. *Koeficient variacije* – omogoča primerjanje razpršenosti dveh ali več spremenljivk/skupin med seboj.

   KV = SD/M x 100

   PRIMER 1:
   M = 37,5 dni dopusta SD = 9,5 **KV = 25,3**
   M = 38,0 dni dopusta DS = 10,1 **KV = 26,6**
   Delavci se med seboj ne razlikujejo preveč.

   PRIMER 2:
   M = 37,5 dni dopusta SD = 9,5 **KV = 25,3**
   M = 16,4 dni dopusta SD = 13,7 **KV = 83,5**
   Delavci se med seboj zelo razlikujejo.
   b. *Z vrednost* – Z = X – M / SD
   Z njo določimo položaj v normalni porazdelitvi.

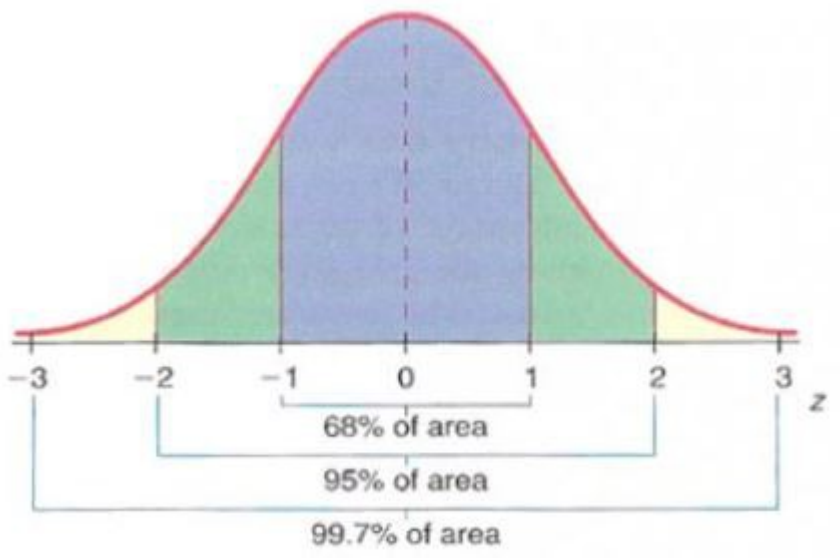
M = 37,5 dni dopusta, SD = 9,5 ***z = 10–37,5/9,5= -2,9***

M = 16,4 dni dopusta, SD = 13,7 ***z = 10–16,4/13,7= -0,5***

Gl. drugi primer KV

NORMALNA PORAZDELITEV

Standardna normalna porazdelitev je porazdelitev vrednosti s povprečjem 0 in standardnim odklonom 1. Ima stalen odnos med standardnim odklonom in verjetnostjo. Razpon od +1SD do–1SD vključuje 68% podatkov oz. je 68%verjetnost, da bo podatek v tem razponu. Razpon od +2SD do–2SD vključuje 95% podatkov oz. je 95%verjetnost, da bo podatek v tem razponu.
Odkloni v levo predstavljajo vrednosti, ki so nižje od aritmetične sredine, odkloni v desno pa predstavljajo vrednosti, ki so višje od aritmetične sredine.



INTERVAL	% VERJETNOSTI
Od M-1 do M+1	68,27%
Od M-2 do M+2	95,45%
Od M-3 do M+3	99,73%

INTERVAL ZAUPANJA/STOPNJA ZAUPANJA

Interval zaupanja vključuje vrednosti/podatke, po navadi znotraj 2SD. Stopnja zaupanja vključuje verjetnost, s katero lahko določeno domnevo sprejmemo; po navadi je 95%. Uporabljamo pri preverjanju hipotez. Uporabljamo pri posploševanju z vzorca na populacijo.