KAKO REDUCIRATI MNOŠTVO PODATAKA

Mirjana Kujundžić Tiljak i Davor Ivanković

REDUKCIJA PROSTORA KVANTITATI VNIH PODATAKA

Redukciju mnoštva podataka započeli smo već u prethodnom poglavlju formiranjem empirijskih distribucija.

Nameće se potreba pronalaženja veličina koje distribuciju podataka opisuju dovoljno dobro s obzirom na grupiranje oko neke vrijednosti kao i s obzirom na obratnu tendenciju, rasap.

Okupljanje podataka najradije se ocjenjuje kao *aritmetička sredina* ili *prosjek* (engl. mean) što nipošto ne znači da je taj odabir uvijek najbolji ili uopće odgovarajući. Definira se kao zbroj svih podataka podijeljen s ukupnim brojem podataka.

Ili *aritmetička sredina populacije*:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

odnosno *aritmetička sredina uzorka*:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Aritmetička sredina težište je distribucije podataka.

Sama aritmetička sredina ne opisuje raspodjelu dostatno. Iste aritmetičke sredine ne znače jednake raspodjele. Pokažimo to na ovom hipotetskom primjeru:

niz A: 2 2,5 3 3,5 4 $\quad \bar{x} = 15/5 = 3$

niz B: 1 1 1 2 10 $\quad \bar{x} = 15/5 = 3$

ili

niz A: 2 2,5 3 3,5 4 $\quad \bar{x} = 15/5 = 3$

niz B: 1 1 1 2 10 $\quad \bar{x} = 15/5 = 3$

Očito je da prosjek 5 dobro opisuje niz A a nikako niz B.

Nužno je stoga naći i neku mjeru rasapa podataka, disperzije ili varijabilnosti. Slijedeći logiku aritmetičke sredine, razvijmo novi niz podataka koji predstavljaju razliku pojedinačnih podataka i aritmetičke sredine, zbrojimo ih i uprosječimo:

	niz A:	niz B:
	-1,0	-2
	-5,0	-2
	0,0	-2
	0,5	-1
	1,0	7
Zbroj	0,0	0,0

Taj postupak očito nije bio od koristi zbog svojstva aritmetičke sredine da je zbroj odstupanja oko aritmetičke sredine jednaka nuli.
Simbolički:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Stoga ćemo pojedinačna odstupanja najprije kvadrirati a zatim zbrojiti.

	niz A:	niz B:
	1,00	4,00
	0,25	4,00
	0,00	4,00
	0,25	1,00
	1,00	49,00
Zbroj	2,50	62,00

Zbroj kvadrata odstupanja oko aritmetičke sredine minimalna je:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \rightarrow \min .$$

odnosno

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - a)^2$$

gdje je $a \neq \bar{x}$.

Ovo svojstvo naziva se *„zbroj najmanjih kvadrata odstupanja od aritmetičke sredine“*.

„Zbroj kvadrata odstupanja od aritmetičke sredine“ (engl. sum of squares about the mean) nije neposredno praktična mjera varijabilnosti jer je jako osjetljiva na veličinu skupa. Stoga je uprosječujemo tj. prikazujemo na jedinicu skupa.

Simbolički:

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

Dobiveni izraz zove se *varijanca*. Varijanca je izražena u kvadratima jedinica mjerenja. Stoga ćemo za potrebe opisa varijabilnosti podataka iz varijance izvaditi kvadratni korijen i dobili smo *standardnu devijaciju*..

Simbolički:

$$\sigma = \sqrt{\sigma^2}$$

ili

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

Međutim opisani postupak vrijedi samo za varijancu odnosno standardnu devijaciju populacije ili osnovnog skupa. Ako s populacijom ne raspolažemo, suma kvadrata odstupanja dijeli se sa n-1 umjesto sa n. Obrazlaganje ovog postupka zahtijevalo bi znatnu matematičku argumentaciju. Stoga ćemo se ovdje zadržati na intuitivnom obrazloženju. Slučajni uzorak nikad neće pokazati tako veliku varijabilnost kao cijela populacija stoga dijeljenje sa n-1 umjesto s n kompenzira očekivano podcjenjivanje veličine standardne devijacije populacije.

Simbolički za uzorak:

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

Odnosno

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

Standardna devijacija jedina je mjera varijabilnosti koja zadovoljava zahtjeve koji se postavljaju pred optimalnu mjeru varijabilnosti. Prima vrijednost nula kada varijabilnosti nema, to je veća sto je varijabilnost veća, uzima u obzir svaki podatak, nije osjetljiva na veličinu uzorka, izražena je u istim jedinicama mjerenja kao i aritmetička sredina tj. kao i mjerenje samo i najmanji je broj od drugih mogućih zbog svojstva aritmetičke sredine.

Ovdje želimo još istaknuti da je zbroj kvadrata odstupanja od aritmetičke sredine značajan koncept koji nalazi svoje mjesto u svim parametrijskim analitičkim postupcima, o čemu će više biti riječi kasnije.

Uspoređujemo li varijabilnost dviju ili više skupina podataka to će samo iznimno biti moguće direktnom usporedbom standardnih devijacija. Taj poseban slučaj je kada su aritmetičke sredine uspoređivanih svojstava iste. U svim drugim situacijama uspoređuju se omjeri standardnih devijacija i pripadajućih aritmetičkih sredina sto zovemo *relativnom standardnom devijacijom* odnosno *koeficijentom varijabilnosti* koji se u pravilu izražava u postotku i neovisan je o jedinicama mjerenja..

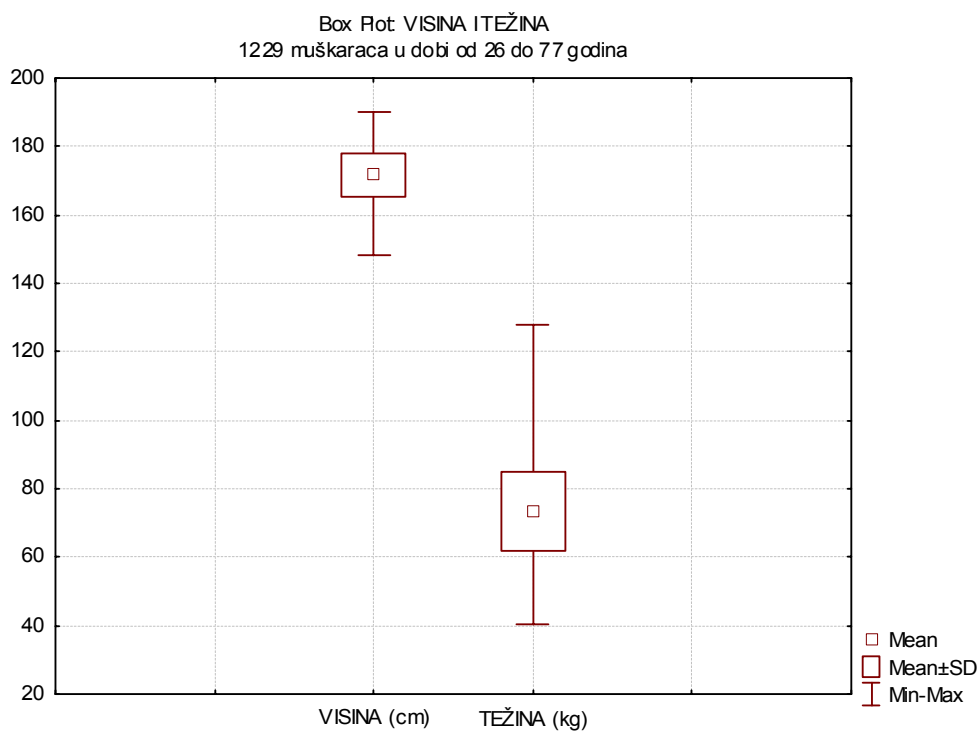
$$KV = \frac{s}{\bar{x}} \times 100$$

Analizom distribucija visine izmjerene u cm i težine izmjerene u kg kod 1229 odraslih muškaraca u dobi od 26 do 77 godina dobit ćemo sljedeće parametre:

	VISINA (cm)	TEŽINA (kg)
$\bar{x}$	171,7	73,3
Minimum	148	40,5
Maksimum	190	128
s	6,3	11,7
KV	3,7%	15,9%

Vidimo da varijabla visina ima veći raspon, tj. distribucija je raspršenija i varijabla je varijabilnija. Isto možemo jasno vidjeti iz grafičkog prikaza („box-plot“) što prikazuje slika13.

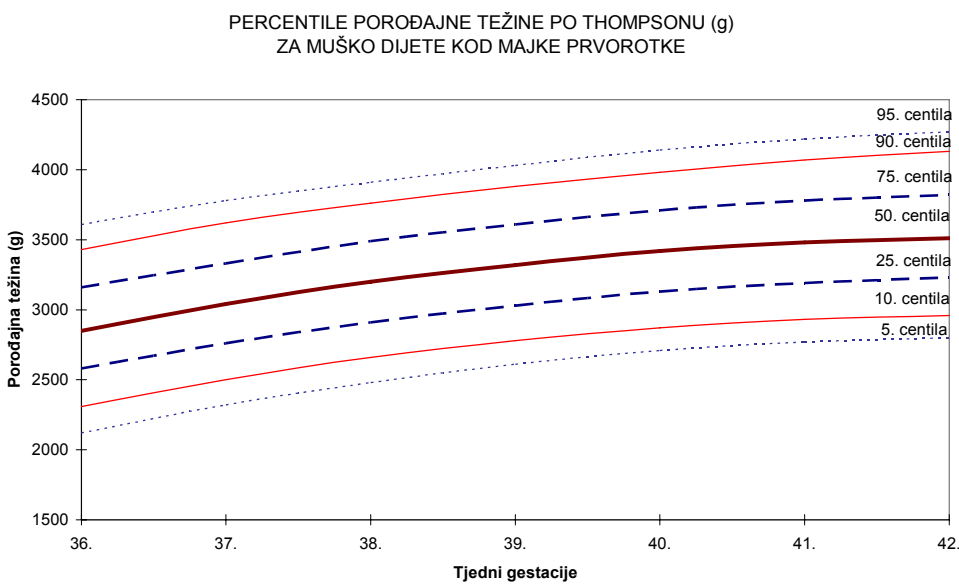
Slika 1. Distribucije visine i težine izmjerene kod 1229 muškaraca u dobi od 26 do 77 godina



Međutim međusobno iste aritmetičke sredine i međusobno iste standardne devijacije još uvijek ne reflektiraju nužno iste distribucije podataka. Moguće je da su jedna skupina pokazatelja opisnice simetrične distribucije a druga neke nesimetrične, krnje distribucije. U takvom je hipotetičkom slučaju očito da aritmetička sredina i standardna devijacija nisu dobro odabrane opisnice tendencije okupljanja odnosno varijabilnosti. Takav slučaj u realnom svijetu nalazimo i u antropometrijskim mjerama predškolske djece. Tada ćemo radije posegnuti za drugim mjerama varijabilnosti i centralne tendencije kao što su *kvantile*..

Kvantile (*medijan, kvartile, decile i centile* odnosno *percentile)* su podaci koji raspodjeljuju raspodjelu u određeni broj jednakih dijelova: medijan na pola, kvartile na četiri jednaka dijela (sic. medijan je jedna od kvartila!), decile na deset a centile na stotinu jednakih dijelova.
Vrijednost obilježja koje prima neka kvantila može se aproksimativno odrediti iz kumulativnog relativnog dijagrama koji prikazuje distribuciju podataka od interesa.

Slika 2. Percentile porođajne težine za muško dijete majke prvorotke po Thompsonu



Aritmetička sredina i medijan *mjere su centralne tendencije*, tj. ukazuju na veličinu u centru distribucije vrijednosti varijable, odnosno oko koje se podaci najviše okupljaju. Osim njih, treća mjera centralne tendencije je i *mod* ili *dominantna vrijednost*, tj. vrijednost varijable koja dominira svojom frekvencijom odnosno vrijednost koju varijabla najčešće poprima.

Varijanca, odnosno standardna devijacija, koeficijent varijabilnosti te kvantile *mjere su varijabilnosti, disperzije ili rasapa* vrijednosti varijable i ukazuju koliko je raspršenje podataka, odnosno kolika je širina promatranog obilježja. Kao mjere varijabilnosti upotrebljavaju se još i *raspon* (razlika između najveće i najmanje vrijednosti varijable) kao i *interkvartilni raspon* (razlika između vrijednosti treće i prve kvartile, interval u kojem se nalazi 50% centralnih podataka)

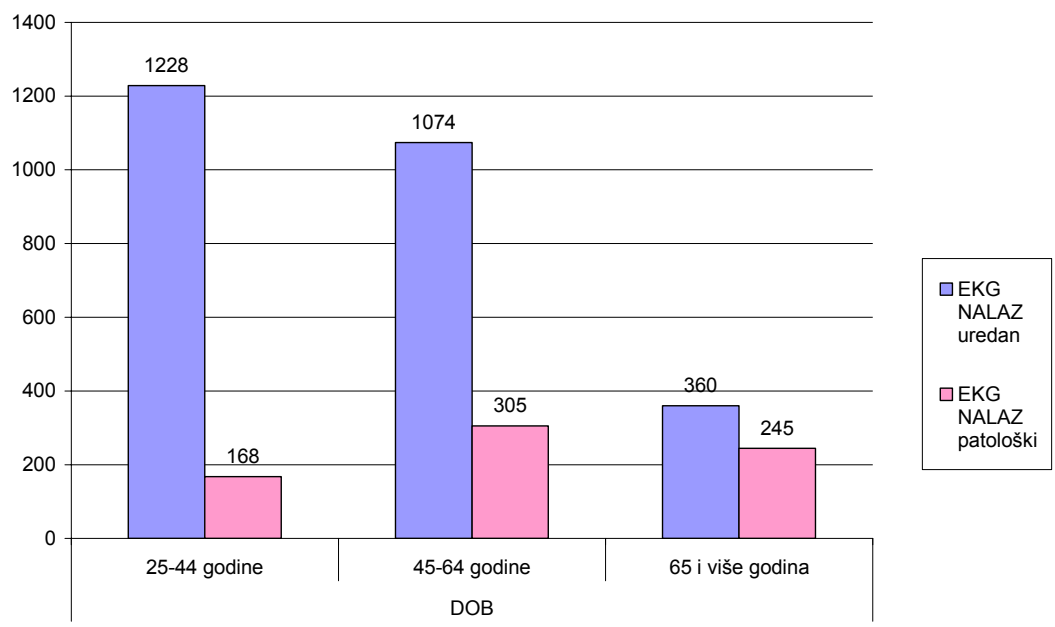
REDUKCIJA PROSTORA KVALITATIVNIH PODATAKA

Formirana *tablica kontingencija* sadrži *apsolutne frekvencije*. Frekvencija u svakom polju tablice dade se prikazati relativno prema zbroju po recima, zbroju po stupcima i prema sveukupnom zbroju. Tako izraženu apsolutnu frekvenciju zovemo *proporcijom*. Pomnožimo li je sa 100 dobijemo postotak: horizontalni postotak (po recima), vertikalni postotak (po stupcima) i dijagonalni postotak (po sveukupnom zbroju).

Tablica 1. Tablica kontingencija nalaza EKG-a prema dobnim skupinama ispianika (apsolutne vrijednsti)

		EKG NALAZ		Ukupno
		uredan	patološki	
DOB	25-44 godine	1228	168	1396
	45-64 godine	1074	305	1379
	65 i više godina	360	245	605
Ukupno		2662	718	3380

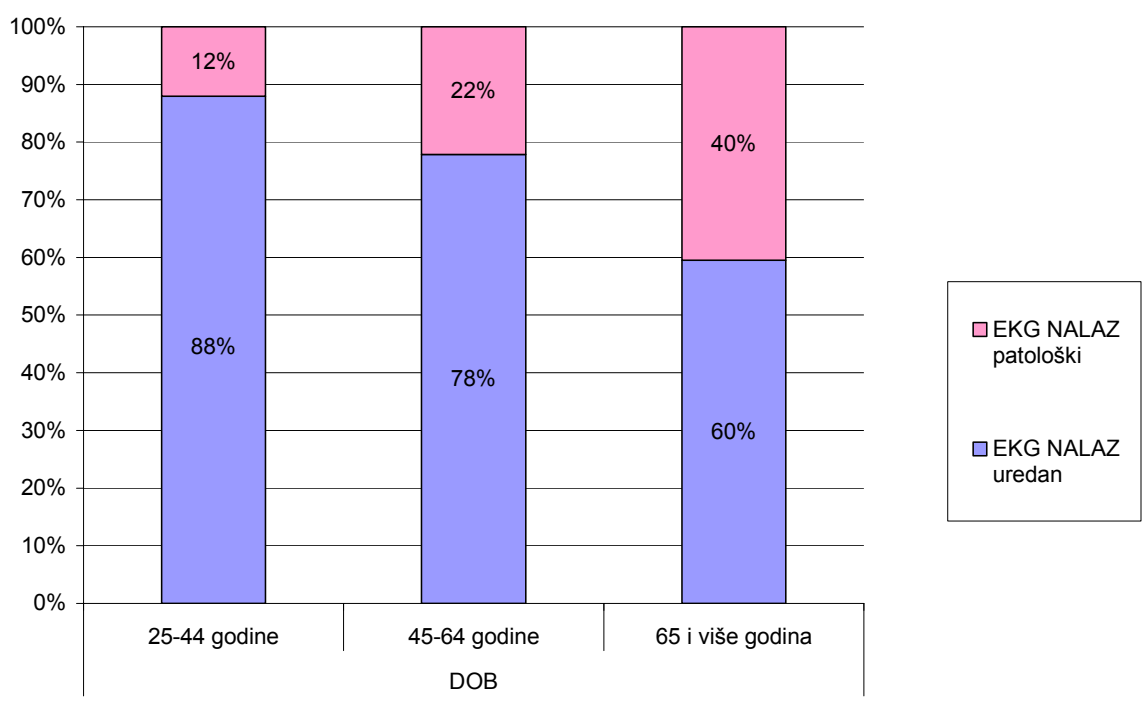
Slika 3. Distribucija EKG nalaza prema dobnim skupinama (apsolutne frekvencije)



Tablica 2. Tablica kontingencija EKG nalaza po dobnim skupinama (apsolutne i relativne frekvencije)

	DOBNA SKUPINA	EKG NALAZ		UKUPNO
		uredan	patološki	
Apsolutna frekvencija	25-44 g	1228	168	1396
Relativna frekvencija u stupcu		46,13%	23,40%	
Relativna frekvencija u retku		87,97%	12,03%	
Ukupna relativna frekvencija		36,33%	4,97%	41,30%
Apsolutna frekvencija	45-64 g	1074	305	1379
Relativna frekvencija u stupcu		40,35%	42,48%	
Relativna frekvencija u retku		77,88%	22,12%	
Ukupna relativna frekvencija		31,78%	9,02%	40,80%
Apsolutna frekvencija	65 i više g	360	245	605
Relativna frekvencija u stupcu		13,52%	34,12%	
Relativna frekvencija u retku		59,50%	40,50%	
Ukupna relativna frekvencija		10,65%	7,25%	17,90%
Apsolutna frekvencija	Ukupno	2662	718	3380
Ukupna relativna frekvencija		78,76%	21,24%	

Slika 4. Distribucija EKG nalaza prema dobnim skupinama (relativne frekvencije)



Literatura:

1. *Ivanković D, i sur. Osnove statističke analize za medicinare. Zagreb: Medicinski fakultet Sveučilišta u Zagrebu, 1989.*
2. *Petrie A, Sabin C. Medical Statistics at a Glance (2nd Ed). Oxford: Blackwell Science Ltd, 2005.*
3. *Glantz. SA. Primer of Biostatistics (4th Ed). New York: McGraww-Hill: 1997.*
4. *Altman DG. Practical Statistics for Medical Research. London. Chapman & Hall, 1991.*
5. *Bland M. An Introduction to Medical Statistics (3rd Ed). Oxford: Oxford University Press, 2005.*
6. *Armitage P, Berry P. Statistical Methods in Medical Research. Oxford: Blackwell Science Ltd, 1994.*