

Grupisanje podataka neprekidnog obilježja u intervale

Prilikom grupisanja podataka nekog neprekidnog obilježja u intervale javljaju se neki praktični problemi. Na primjer, koliko intervala je potrebno, koja je širina tih intervala, koja je početna tačka. Zato se zadaci ovog tipa rješavaju u nekoliko etapa.

I etapa – broj intervala

Broj intervala se određuje na osnovu formule $k=1+3,3*\log N$, gdje je log oznaka za logaritam, a N za ukupan broj elemenata nekog uzorka.

U situaciji kada se ovi zadaci rade na predavanjima i vježbama, vrijednost logaritma dobijate od nas. U ovoj situaciji, kada budete vježbali, logaritma morate sami da računate. Neophodno je da imate bolji kalkulator (većina ga ima na telefonu). Dok vježbate možete da koristite kalkulator sa telefona ☺. Potrebno je da se unese vrijednost N i samo se pritisne dugme **log**.

Dok računate vrijednost k obratite pažnju na red računskih operacija. Prvo treba vrijednost koju dobijete na kalkulatoru da pomnožite sa 3,3 pa tako dobijeni proizvod saberete sa 1. U najvećem broju slučajeva vrijednost k će biti data u decimalnom zapisu. S obzirom da broj intervala ne može da bude, na primjer 4,6, tada se ova vrijednost mora zaokružiti. Ovdje ćemo zakoruživanje uraditi nešto drugačije, nego ranije. Naime, zakoruživaćemo na najbliži cio broj koji nije manji od k.

Na primjer $k=4,58 \approx 5$

$k=7,18 \approx 8$

Dakle, vidimo da smo vrijednost 7,18 zaokružili da 8, jer je to najbliži cio broj koji nije manji od 7,18.

Ovdje treba naglasiti, da se u zavisnosti od podataka, može desiti da broj intervala i bude i manji od k. Ukoliko neki interval, čiji je redni broj manji od k, obuhvati najveću vrijednost u uzroku tu se zaustavljam.

II etapa – određivanje širine intervala

Kada znamo broj intervala treba da izračunamo i njihovu širinu. Širina se računa po formuli

$$i = \frac{x_{\max} - x_{\min}}{k}$$

gdje je $x_{\max}$ najveća, $x_{\min}$ najmanja vrijednost u uzorku. Vrijednost k je izračunata u prethodnoj etapi i za potrebe izračunavanja širine intervala uzima se nezaokružena. Slično, kao i u prethodnoj etapi, vrijednost i će biti u obliku decimalnog broja. Primjenjuje se isti postupak zaokruživanja, kao u slučaju parametra k.

Primjer: Dat je uzorak 14, 12, 18, 13, 10, 19, 20, 17, 16, 11. Odrediti vrijednost k i i .

Vidi se da je $N=10$. Parametar k dobijamo po formuli $k=1+3,3*\log N=1+3,3*\log 10$. Primjenom kalkulatora lako možete da dobijete da je $\log 10=1$.

Dakle, $k=1+3,3*1=4,3\approx 4$.

Računamo širinu intervala. Lako se vidi da je $x_{\max}=20$ i $x_{\min}=10$. Lako se dobija

$$i = \frac{20 - 10}{4,3} = \frac{2}{4,3} = 0,46 \approx 1$$

III etapa – određivanje početne tačke

Da bi znali od koje tačke počinje prvi interval, potrebno je da nađemo i početnu tačku. Primjenjujemo sljedeću formulu

$$x_0 = x_{\min} - \frac{i}{2}$$

Vrijednost i smo izračunali u II etapi, i za potrebe izračunavanja početne tačke uzimamo nezaokruženu vrijednost. Ako uzmemo vrijednost iz prethodnog primjera, dobijamo

$$x_0 = x_{\min} - \frac{i}{2} = 10 - \frac{0,46}{2} = 9,76 \approx 10.$$

Pošto ćemo uglavnom da radimo sa cjelobrojnim vrijednostima, radi lakšeg računanja, vrijednost početne tačke zaokružićemo po prethodno objašnjenom principu.

Sada možemo da pristupimo grupisanju podataka.

Zadatak 1: Dat je uzorak 17, 12, 18, 21, 17, 10, 15, 16, 21, 22, 10, 14, 13, 15, 9, 19, 12, 16, 20, 19, 21, 10.

Grupisati podatke u intervale.

Rješenje:

Lako se vidi da je $N=22$ i da je $\log 22=1,34$. Sada je $k=1+3,3*1,34=1+ 4,42=5,42 \approx 6$.

Takođe je, $x_{\max}=22$ i $x_{\min}=9$.

$$i = \frac{22 - 9}{5,42} = \frac{13}{5,42} = 2,39 \approx 3$$

Početna tačka je

$$x_0 = 9 - \frac{2,39}{2} = 9 - 1,195 = 7,805 \approx 8.$$

Dakle, intervali su 8 – 11, 12 - 15, 16 - 19, 20 – 23. Iako je $k \approx 6$, mi ćemo se zaustaviti kod četvrtog intervala, jer nema potrebe da idemo dalje. Posljednji, četvrti interval, je obuhvatio najveću vrijednost uzorka.

Formira se tabela

Intervali	f – apsolutna frekvencija
8-11	4 (ove frekvence se dobijaju tako što se prebroji koliko elemenata iz uzorka pripada intervalu)
12-15	6
16-19	7
20-23	5
Ukupno	22

Vrlo je bitno da se suma obojena žutom složi sa ukupnim brojem elemenata u uzorku, dakle sa vrijednošću N.

Zadatak 2: Testirana je nova nastavna metoda u učenju engleskog jezika. Ukupno 35 nastavnika ocjenjivalo je kvalitet primijenjene metode ocjenjujući je jednom ocjenom iz intervala od 0 do 100.

Rezultati su: 36, 51, 40, 57, 63, 44, 32.5, 63, 39, 54, 60, 58, 34, 48, 55, 42, 51, 35, 53, 61, 61, 59, 40, 42, 62, 54, 43, 62, 49, 56, 52, 57, 55, 55, 46.

Grupisati podatke u intervalne serije. Pa zatim izračunati aritmetičku sredinu, IQR, medijanu, standardnu devijaciju, mod.

Rješenje:

Da bi podatke grupisali u intervalne serije, prvo treba da izračunamo broj intervala po formuli

$$k=1+3,3*\log N$$

N= 35.

Kako je $\log 35 = 1,54$, dobijamo

$$k = 1 + 3.3 * 1.54 = 6.082 \approx 7$$

(zaokružujemo na najbliži cio broj koji je veći od 6.082)

Kako je $x_{\max} = 63$, a $x_{\min} = 32.5$ primjenom formule

$$i = \frac{x_{\max} - x_{\min}}{k}$$

Dobijamo

$$i = \frac{63-32.5}{6.082} = 5.014 \approx 6$$

(obratiti pažnju da u formuli ne koristimo približnu vrijednost za k)

Sada pomoću

$$x_0 = x_{\min} - \frac{i}{2}$$

dolazimo do početne tačke.

$$x_0 = 32.5 - \frac{5.014}{2} = 29.996 \approx 30$$

(obratiti pažnju da u formuli ne koristimo približnu vrijednost za i)

Vodeći računa da se kraj jednog i početak narednog intervala ne poklapaju, formiraćemo tabelu raspodjele frekvenci. Počinjemo od broja $x_0 = 30$, širina intervala je $i = 6$. Zaustavljamo se kada obuhvatimo $x_{\max} = 63$. Frekvence f_i određujemo prostim provjeravanjem da li konkretan element pripada intervalu. Formiramo sljedeću tabelu

Intervali	f_i	K	Sredine int. x'	$x' * f_i$	$(x')^2$	$(x')^2 * f_i$
30 -36	4	4	33	33*4 = 132	33*33=1089	1089*4=4356
37-43	6	10	40	40*6 = 240	40*40=1600	1600*6=9600
44-50	4	14	47	47*4=188	47*47=2209	2209*4=8836
51-57	12	26	54	54*12 = 648	54*54=2916	2916*12=34992
58-64	9	35	61	61*9 = 549	61*61=3721	3721*9=33489
Ukupno	35					

Dalje, računamo aritmetičku sredinu po formuli:

$$x = \frac{\sum x' \cdot f_i}{n}$$

$$\bar{x} = \frac{132 + 240 + 188 + 648 + 549}{35} = \frac{1757}{35} = 50.2$$

Računamo medijanu po formuli:

$$M_e = a_i + (b_i - a_i) \cdot \frac{\frac{n+1}{2} - K}{f}$$

Odredimo prvo medijalni interval : $\frac{n+1}{2} = \frac{35+1}{2} = \frac{36}{2} = 18$. U tabeli pronalazimo interval čija je kumulativna frekvencija najbliži broj veći od 18. Uočimo da je to broj 26, pa je medijalni interval 51-57. f je frekvencija intervala 51-57 a to je 12. K je kumulativna frekvencija koja prethodi medijalnom intervalu (51-57) a to je 14. Zamjenom u formulu dobijamo:

$$M_e = 51 + (57 - 51) \cdot \frac{\frac{35+1}{2} - 14}{12} = 51 + 6 \cdot \frac{4}{12} = 51 + 2 = 53$$

Mod računamo po formuli:

$$M_o = a_i + (b_i - a_i) \cdot \frac{f_2}{f_1 + f_2}$$

Interval sa najvećom frekvencijom je 51-57, pa je $a_i = 51$, $b_i = 57$, $f_1 = 4$, $f_2 = 9$.

$$M_o = 51 + (57 - 51) \cdot \frac{9}{4 + 9} = 51 + 6 \cdot \frac{9}{13} = 51 + 4.15 = 55.15$$

Prvi kvartil računamo po formuli:

$$Q_1 = a_i + (b_i - a_i) \cdot \frac{\frac{n+1}{4} - K}{f}$$

Da bi odredili traženi interval prvo izračunajmo $\frac{n+1}{4} = \frac{36}{4} = 9$

U tabeli uzimamo interval čija je kumulativna frekvencija prvi veći broj od 9, uočavamo da je u tabeli to broj 10. Pa je $a_i = 37$, $b_i = 43$, $f = 6$, $K = 4$. Zamijenimo vrijednosti u formuli i dobijamo:

$$Q_1 = 37 + (43 - 37) \cdot \frac{\frac{35 + 1}{4} - 4}{6} = 37 + 6 \frac{5}{6} = 37 + 5 = 42$$

Treći kvartil računamo po formuli:

$$Q_3 = a_i + (b_i - a_i) \cdot \frac{\frac{3(n+1)}{4} - K}{f}.$$

Sprovodimo sličan postupak kao u prethodnom koraku. Prvo računamo poziciju intervala $\frac{3(n+1)}{4} = \frac{3 \cdot 36}{4} = 27$. U tabeli uzimamo interval čija je kumulativna frekvencija prvi veći broj od 27, uočavamo da je u tabeli to broj 35. Pa je $a_i = 58$, $b_i = 64$, $f = 9$, $K = 26$. Zamijenimo vrijednosti u formuli i dobijamo:

$$Q_3 = 58 + (64 - 58) \cdot \frac{\frac{3(35 + 1)}{4} - 26}{9} = 58 + 6 \frac{1}{9} = 58 + 0.67 = 58.67$$

Pa je $IQR = Q_3 - Q_1 = 58.67 - 42 = 16.67$

Još je preostalo da izračunamo standardnu devijaciju. Prvo ćemo izračunati disperziju, koristimo formulu

$$D = \frac{\sum fx^2}{n} - (\bar{x})^2$$

$$D = \frac{4356 + 9600 + 8836 + 34992 + 33489}{35} - (50.2)^2$$

$$D = \frac{91273}{35} - 2520.04 = 2607.8 - 2520.04 = 87.76$$

Standardnu devijaciju računamo po formuli:

$$std = \sqrt{D}$$

Pa je $std = 9.37$, time je zadatak završen.