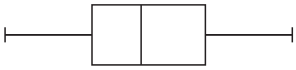
Karakteristična petorka i brkata kutija

Sanja Varošanec, Zagreb

Jedan od ishoda trećeg ciklusa obrazovanja jest prikazati podatke s pomoću tzv. brkate kutije. U engleskom govornom području u upotrebi su nazivi *box-plot*, *box-whiskers*, dok kod nas koristimo nazive *dijagram pravokutnika*, *kutijasti dijagram* ili **brkata kutija**.



Slika 1.

Pogledamo li kako izgleda taj dijagram jasni su nam nazivi koji se javljaju u vezi s njim. Dijagram se sastoji od pravokutnika (kutije) i dvije dužine (brka) koje izviru iz nasuprotnih stranica pravokutnika. Za crtanje dijagrama trebamo neke numeričke pokazatelje skupa podataka i to: najmanji podatak $x_{\min}$, donji kvartil Q_1, medijan M, gornji kvartil Q_3 i najveći podatak $x_{\max}$. Tih pet brojeva

$$(x_{\min}, Q_1, M, Q_3, x_{\max})$$

čine tzv. **karakterističnu petorku** skupa podataka. Medijan smo naučili određivati u članku *Peteljka-list dijagram*. *Medijan i mod* obrađeni u prethodnom broju MIŠ-a, a ovdje ćemo naučiti kako odrediti donji i gornji kvartil.

Primjer 1. Zadani su podatci: $80, 320, 412, 130, 210, 273, 190, 95, 350, 60, 45$. Odredimo medijan, donji i gornji kvartil ovog skupa podataka.

Rješenje. Prvo podatke napišimo u rastućem poretku:

$$45, 60, 80, 95, 130, 190, 210, 273, 320, 350, 412.$$

Ima ih 11, a medijan je onaj podatak koji taj niz dijeli na dva niza jednakih duljina. Dakle, medijan je onaj broj koji ima svojstvo da je broj podataka manjih od njega jednak broju podataka većih od njega. Drugim riječima tražimo "srednji" član. U ovom nizu podataka to je broj 190, tj. $M = 190$. Pet je podataka manje od 190, a pet ih je veće od 190. Kao što već znamo, mogli smo medijan izračunati i prisjetivši se formule

$$M = \begin{cases} x_{k+1} & \text{ako je } n = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2} & \text{ako je } n = 2k \end{cases}$$

pri čemu je n ukupan broj podataka koji su poslagani u rastućem poretku. Dakle, medijan je niz podijelio na dva kraća niza jednakih duljina. Istaknimo samo prvi dio niza:

$$45, 60, 80, 95, 130.$$

To je opet jedan niz. I on posjeduje medijan. To je broj 80 i to je donji kvartil cjelokupnog skupa podataka. Pišemo: $Q_1 = 80$.

Sličnu priču provedimo i za drugu polovinu početnog niza:

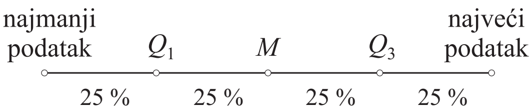
$$210, 273, 320, 350, 412.$$

Njegov srednji element (medijan) je broj 320 i to je gornji kvartil cjelokupnog skupa podataka. Pišemo: $Q_3 = 320$.

$$\begin{array}{ccccccccccc} 45 & 60 & \boxed{80} & 95 & 130 & 190 & 210 & 273 & \boxed{320} & 350 & 412 \\ & & \downarrow & & & \downarrow & & & \downarrow & & \\ & & Q_1 & & & M & & & Q_3 & & \end{array}$$

Slika 2.

Dakle, donji kvartil, medijan i gornji kvartil zamišljamo ovako: to su brojevi koji dijele cijeli niz podataka na četiri jednako velike grupe, odnosno u svakoj od grupa nalazi se 25 % podataka.



Slika 3.

Primjer 2. Zadani su podatci: $21, 17, 32, 45, 59, 73, 29, 63$. Odredimo medijan, donji i gornji kvartil ovog skupa podataka.

Rješenje. Podatci u rastućem poretku izgledaju ovako:

$$17,\ 21,\ 29,\ 32,\ 45,\ 59,\ 63,\ 73.$$

Ukupan broj podataka je 8, to je paran broj, pa za medijan uzimamo aritmetičku sredinu dva srednja podatka, tj.

$$M = \frac{x_4 + x_5}{2} = \frac{32 + 45}{2} = 38.5.$$

Donji kvartil je medijan prve polovine niza, tj. niza $17, 21, 29, 32$.

U tom, kraćem nizu opet imamo paran broj elemenata, pa je donji kvartil

$$Q_1 = \frac{x_2 + x_3}{2} = \frac{21 + 29}{2} = 25.$$

Za određivanje gornjeg kvartila pogledajmo drugu polovinu niza: $45, 59, 63, 73$. Gornji kvartil zadanog niza je ujedno medijan tog kraćeg niza pa je

$$Q_3 = \frac{x_6 + x_7}{2} = \frac{59 + 63}{2} = 61.$$

Konačno, rješenje ovog zadatka je

$$Q_1 = 25,\ M = 38.5,\ Q_3 = 61.$$

$$\begin{array}{ccccccccccc} 17 & 21 & & 29 & 32 & & 45 & 59 & & 63 & 73 \\ & & \uparrow & & & \uparrow & & & \uparrow & & \\ & & Q_1 & & & M & & & Q_3 & & \end{array}$$

Slika 4.

Sad kad znamo izračunati sve članove karakteristične petorke, pristupimo crtanju "brkate kutije", tj. crtanju dijagrama pravokutnika.

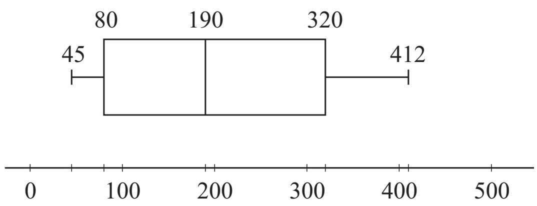
Primjer 3. S pomoću dijagrama pravokutnika grafički prikažimo podatke iz Primjera 1.

Rješenje. Napišimo sve članove karakteristične petorke za podatke iz Primjera 1.

$$x_{\min} = 45,\ x_{\max} = 412,$$

$$Q_1 = 80,\ M = 190,\ Q_3 = 320.$$

Nacrtajmo brojevni pravac i na njemu odredimo tih pet elemenata karakteristične petorke. Pravokutniku stranice paralelne s x-osi počinju na Q_1, a završavaju na Q_3. Još u pravokutniku povučemo dužinu okomitu na x-os na pravcu $x = M$. Najmanji i najveći podatak prikazujemo točkom ili crticom te spojimo s pravokutnikom – tako dobivamo "brkove" pravokutnika.



Slika 5.

Ovaj grafički prikaz pogodan je i za usporedbu dviju i više skupina podataka.

Primjer 4. Na fakultetskom poljoprivrednom dobru uzgajaju se dvije sorte graška: A i B. Nakon berbe, prebrojena su zrna graška na pojedinoj stabljici i dobivena su ova dva skupa podataka.

Broj zrna graška na pojedinoj stabljici sorte A

$$21, 13, 40, 39, 42, 44, 52, 53, 60, 56,$$

$$52, 64, 51, 59, 36, 39, 24, 33, 44, 52.$$

Broj zrna graška na pojedinoj stabljici sorte B

$$20, 47, 64, 61, 55, 27, 15, 59, 44,$$

$$57, 58, 40, 43, 38, 52, 64, 47, 57.$$

Nacrtaj dijagrame pravokutnika za obje sorte i usporedi im distribuciju.

Rješenje. Umjesto da podatke pišemo u obliku niza, napravit ćemo zajednički peteljka-list dijagram za obje sorte.

Sorta A		Sorta B
3	1	5
4 1	2	0 7
9 9 6 3	3	8
4 4 2 0	4	0 3 4 7 7
9 6 3 2 2 2 1	5	2 5 7 7 8 9
4 0	6	1 4 4

Legenda: $1|5$ je 15 zrna graška.

Za sortu A imamo $n = 20$ podataka i vrijedi:

$$x_{\min} = 13,\ x_{\max} = 64,$$

$$M = \frac{x_{10} + x_{11}}{2} = \frac{44 + 44}{2} = 44,$$

$$Q_1 = \frac{x_5 + x_6}{2} = \frac{36 + 39}{2} = 37.5,$$

$$Q_3 = \frac{x_{15} + x_{16}}{2} = \frac{52 + 53}{2} = 52.5.$$

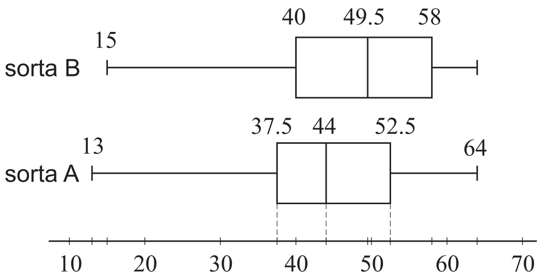
Za sortu B imamo 18 podataka i elementi karakteristične petorke su:

$$x_{\min} = 15,\ x_{\max} = 64,$$

$$M = \frac{x_9 + x_{10}}{2} = \frac{47 + 52}{2} = 49.5,$$

$$Q_1 = x_{4+1} = 40,\ Q_3 = x_{13+1} = 58.$$

Dijagrami pravokutnika za obje sorte u jednoj slici izgledaju ovako.



Slika 6.

Ono što odmah uočavamo na ovoj zajedničkoj slici, jest da je pravokutnik sorte B pomaknut udesno u odnosu na pravokutnik sorte A, tj. $50\,\%$ podataka sorte B zauzima područje većih vrijednosti nego $50\,\%$ podataka za sortu A. Jedna i druga sorta imaju jednaki maksimalni podatak. Minimalni podatak i medijan su kod sorte B veći nego kod sorte A. Duljina pravokutnika za sortu A je

$$IQR_A = Q_3 - Q_1 = 52.5 - 37.5 = 15,$$

dok je duljina pravokutnika za sortu B jednaka

$$IQR_B = Q_3 - Q_1 = 58 - 40 = 18.$$

To znači da je srednjih $50\,\%$ podataka sorte A manje raspršeno nego kod sorte B. Ta se veličina naziva **interkvartilni raspon**.

Varijante dijagrama pravokutnika

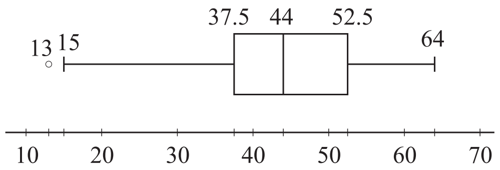
Ponekad se u literaturi susrećemo s malo drukčijim dijagramom pravokutnika. U takvom dijagramu "brkovi" ne počinju, odnosno završavaju, u najmanjem odnosno najvećem podatku, nego u u podatcima koji su najmanji, odnosno najveći u intervalu

$$[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR].$$

Podatci koji se nalaze izvan tog intervala se na grafu označavaju točkama i nazivaju se vanjski podatci (eng. *outliers*). To su, intuitivno rečeno, podatci koji jako odudaraju od većine podataka, ekstremni su, možda im je uzrok pogreška pri mjerenju, zapisu ili neka druga anomalija.

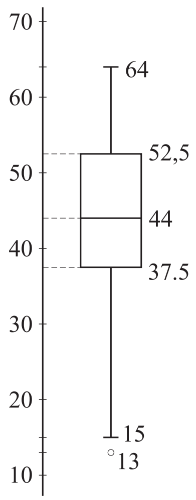
Primjer 5. Nacrtajmo dijagram pravokutnika s istaknutim vanjskim elementima za podatke za sortu A iz prethodnog primjera.

Rješenje. Prvo izračunajmo $1.5 \cdot IQR$ za sortu A. $1.5 \cdot IQR = 1.5 \cdot 15 = 22.5$. Interval $[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$ za sortu A jednak je $[37.5 - 22.5, 52.5 + 22.5] = [15, 75]$. Dakle, vanjski elementi su oni podatci koji leže van intervala $[15, 75]$. To je samo podatak 13. Lijevi brk počinje u 15, a desni završava u $x_{\max} = 64$, jer je u ovom primjeru $x_{\max} < Q_3 + 1.5IQR$. Vanjski podatak se crta kao izolirana točka.



Slika 7.

Osim gornje opisane varijante dijagrama pravokutnika, ponekad ćemo naići i na graf u kojemu je brojevni pravac postavljen vertikalno. Na sljedećoj je slici prikazan dijagram iz prethodnog primjera u takvom položaju.



Slika 8.

Varijante izračunavanja kvartila

Nastavnik matematike treba biti svjestan da je u opisnoj statistici prisutno nekoliko načina izračunavanja kvartila. Metoda opisana u primjerima 3 i 4 (zovimo je metoda I.) koristi se u džepnim računalima i u mnogim *online*-kalkulatorima, te u starim verzijama Excela. Osim nje koriste se još i sljedeće dvije metode.

Metoda II. Neka je n broj podataka danih u rastućem poretku. Označimo s k najveće cijelo broja $\frac{n+1}{4}$, tj. k je najveći cijeli broj koji je manji ili jednak broju $\frac{n+1}{4}$. A s α označimo decimalni dio broja $\frac{n+1}{4}$. Tako primjerice, za $n = 4$ vrijedi da je $k = 1$ i $\alpha = 0.25$, za $n = 9$ je $k = 2$ i $\alpha = 0.5$.

Donji se kvartil Q_1 definira ovako

$$Q_1 = x_k + \alpha(x_{k+1} - x_k).$$

I gornji se kvartil definira istom formulom samo što se za k uzima najveće cijelo broja $\frac{3(n+1)}{4}$.

Metoda III. Ova je metoda prisutna u novoj verziji Excela. Donji se kvartil računa s pomoću formule $Q_1 = x_{k+1} + \alpha(x_{k+2} - x_{k+1})$, pri čemu je k najveće cijelo broja $\frac{n-1}{4}$, α njegov decimalni dio. Gornji se kvartil računa istom formulom samo što je tada k najveće cijelo broja $\frac{3(n-1)}{4}$.

Svim je ovim metodama zajedničko da ako skup podataka ima $4k$ elemenata, dobiveni kvartili zajedno s medijanom dijele skup na četiri dijela u kojima ima k podataka, iako se pri svakoj metodi dobiju različite vrijednosti kvartila.

LITERATURA

1/ A. G. Bluman: *Elementary Statistics*, McGraw Hill, Boston, 2001.

2/ S. Varošanec: *Peteljka-list dijagram. Medijan i mod*, Matematika i škola, 71 (2013), 10–13.