Statistika

Vanja Wagner

1. Deskriptivna statistika

Mjere srednje vrijednosti

Podsjetimo se mjera srednje vrijednosti (centralne tendencije) kojima želimo jednom procjenom reći nešto o srednjim (očekivanim) kretanjima promatrane variable u danoj populaciji:

- **aritmetička sredina** (srednja vrijednost)
- **medijan**
- **mod**

Srednja vrijednost je najčešće korištena mjera od ove tri (pogodna za simetrične, odnosno neiskrivljene podatke[1]). Srednja vrijednost ima jedan glavni nedostatak: posebno je osjetljiva na utjecaj ekstremnih vrijednosti (*outliera*). To su vrijednosti koje su neobične u usporedbi s ostatkom skupa podataka jer su posebno male ili velike (na numeričkoj skali)[2]. Stoga srednja vrijednost nije pogodna mjera za iskrivljene podatke, za vrijednosti koje su već prosječne ili kada skala na kojoj mjerimo podatke nije linearna (npr. pH jedinice se ne mjere na linearnoj skali).

[1] O simetriji podataka i odnosu srednje vrijednosti i medijana ćemo malo više govoriti danas kod mjera asimetrije.

[2] O ekstremnim vrijednostima ćemo danas reći nešto više kod mjera raspršenja.

Mjere raspršenosti (disperzije)

Usporedimo dva niza podataka iste duljine i srednje vrijednosti

x_i	10	60	50	30	40	20	$(N = 6,\ \overline{x} = 35)$
y_i	35	45	30	35	40	25	$(N = 6,\ \overline{y} = 35)$

Iako su srednje vrijednosti iste, promatrani podaci se ipak razlikuju - drugi skup podataka je više grupiran oko srednje vrijednosti od prvog,

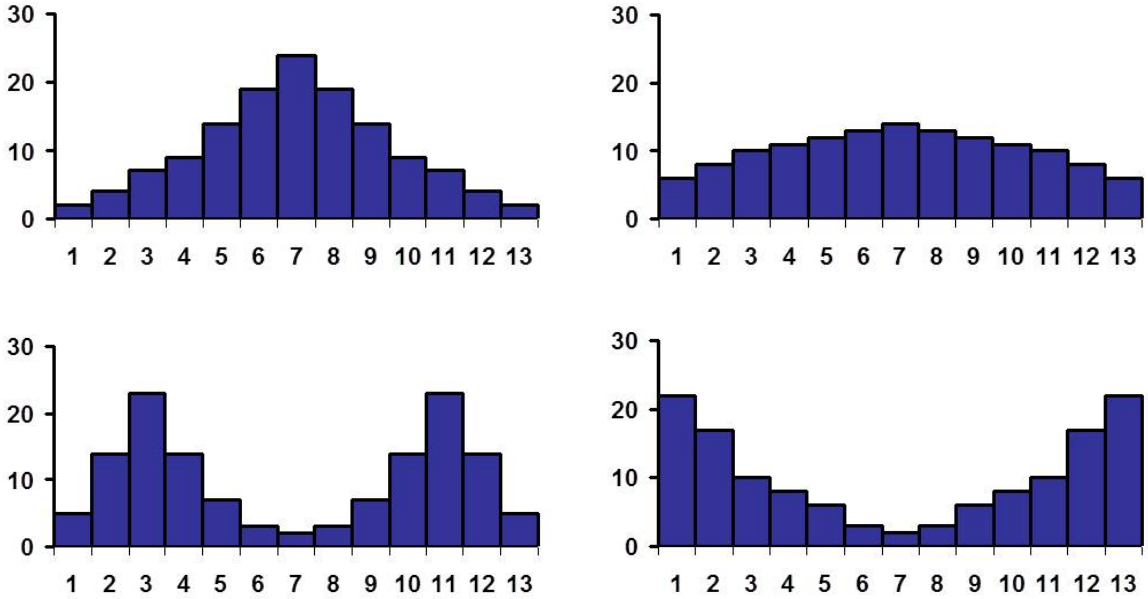
Mjere raspršenosti (disperzije)

Usporedimo dva niza podataka iste duljine i srednje vrijednosti

x_i	10	60	50	30	40	20	$(N = 6,\ \overline{x} = 35)$
y_i	35	45	30	35	40	25	$(N = 6,\ \overline{y} = 35)$

Iako su srednje vrijednosti iste, promatrani podaci se ipak razlikuju - drugi skup podataka je više grupiran oko srednje vrijednosti od prvog,

Promotrimo sljedeće histograme. U kojem od prikazana četiri uzorka su podaci najviše a gdje najmanje raspršeni?



Raspon

Najjednostavnija mjera raspršenja je **raspon**, odnosno razlika najveće i najmanje vrijednosti u uzorku:

$$d = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}.$$

Uočite da uzorci 1, 1, 1, 2, 10 i 1, 4, 7, 8, 10 imaju isti raspon, premda očito nisu jednako disperzirani. Zato ćemo tražiti bolju mjeru za disperziju.

Kvartili

Donji kvartil Q_1 je vrijednost od koje je 25% podataka manje a 75% podataka veće, a **gornji kvartil** Q_3 je vrijednost od koje je 75% podataka manje a 25% podataka veće.

Kvartili

Donji kvartil Q_1 je vrijednost od koje je 25% podataka manje a 75% podataka veće, a **gornji kvartil** Q_3 je vrijednost od koje je 75% podataka manje a 25% podataka veće.

Uočimo, gornji i donji kvartil, zajedno s medijanom, dijele niz podataka na 4 jednaka dijela.

Kvartili

Donji kvartil Q_1 je vrijednost od koje je 25% podataka manje a 75% podataka veće, a **gornji kvartil** Q_3 je vrijednost od koje je 75% podataka manje a 25% podataka veće.

Uočimo, gornji i donji kvartil, zajedno s medijanom, dijele niz podataka na 4 jednaka dijela.

Računanje kvartila

- Niz podataka podijelimo na dva jednaka dijela. Ukoliko je broj podataka neparan, izbacimo medijan.

Kvartili

Donji kvartil Q_1 je vrijednost od koje je 25% podataka manje a 75% podataka veće, a **gornji kvartil** Q_3 je vrijednost od koje je 75% podataka manje a 25% podataka veće.

Uočimo, gornji i donji kvartil, zajedno s medijanom, dijele niz podataka na 4 jednaka dijela.

Računanje kvartila

- Niz podataka podijelimo na dva jednaka dijela. Ukoliko je broj podataka neparan, izbacimo medijan.
- Donji kvartil je medijan prve polovice niza.

Kvartili

Donji kvartil Q_1 je vrijednost od koje je 25% podataka manje a 75% podataka veće, a **gornji kvartil** Q_3 je vrijednost od koje je 75% podataka manje a 25% podataka veće.

Uočimo, gornji i donji kvartil, zajedno s medijanom, dijele niz podataka na 4 jednaka dijela.

Računanje kvartila

- Niz podataka podijelimo na dva jednaka dijela. Ukoliko je broj podataka neparan, izbacimo medijan.
- Donji kvartil je medijan prve polovice niza.
- Gornji kvartil je medijan druge polovice niza.

Kvartili

Donji kvartil Q_1 je vrijednost od koje je 25% podataka manje a 75% podataka veće, a **gornji kvartil** Q_3 je vrijednost od koje je 75% podataka manje a 25% podataka veće.

Uočimo, gornji i donji kvartil, zajedno s medijanom, dijele niz podataka na 4 jednaka dijela.

Računanje kvartila

- Niz podataka podijelimo na dva jednaka dijela. Ukoliko je broj podataka neparan, izbacimo medijan.
- Donji kvartil je medijan prve polovice niza.
- Gornji kvartil je medijan druge polovice niza.

Odnosno, $Q_1 = x_{\left(\frac{1}{4}(n+1)\right)}$ i $Q_3 = x_{\left(\frac{3}{4}(n+1)\right)}$, gdje za $s \notin \mathbf{N}$ definiramo

$$x_{(s)} = (1-r)x_{(k)} + rx_{(k+1)} \tag{1}$$

za $k = \lfloor s \rfloor$ i $r = s - k$.

Interkvartil (IQ) je udaljenost između donjeg i gornjeg kvartila:

$$IQ = Q_3 - Q_1.$$

Interkvartil (IQ) je udaljenost između donjeg i gornjeg kvartila:

$$IQ = Q_3 - Q_1.$$

Uočimo:

- Između donjeg i gornjeg kvartila se nalazi 50% podataka.

Interkvartil (IQ) je udaljenost između donjeg i gornjeg kvartila:

$$IQ = Q_3 - Q_1.$$

Uočimo:

- Između donjeg i gornjeg kvartila se nalazi 50% podataka.
- Interkvartil je raspon u kojem se nalazi 50% središnjih članova niza podataka.

Interkvartil (IQ) je udaljenost između donjeg i gornjeg kvartila:

$$IQ = Q_3 - Q_1.$$

Uočimo:

- Između donjeg i gornjeg kvartila se nalazi 50% podataka.
- Interkvartil je raspon u kojem se nalazi 50% središnjih članova niza podataka.
- Vrijednost prvih i zadnjih 25% podataka ne utječe na vrijednost interkvartila.

Interkvartil (IQ) je udaljenost između donjeg i gornjeg kvartila:

$$IQ = Q_3 - Q_1.$$

Uočimo:

- Između donjeg i gornjeg kvartila se nalazi 50% podataka.
- Interkvartil je raspon u kojem se nalazi 50% središnjih članova niza podataka.
- Vrijednost prvih i zadnjih 25% podataka ne utječe na vrijednost interkvartila.

Uočimo i da je interkvartil izražen u istoj mjernoj jedinici kao i podaci.Ako to želimo izbjeći pogodno je promotriti **koeficijent kvartilne devijacije** VQ definiran s

$$VQ = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Interkvartil (IQ) je udaljenost između donjeg i gornjeg kvartila:

$$IQ = Q_3 - Q_1.$$

Uočimo:

- Između donjeg i gornjeg kvartila se nalazi 50% podataka.
- Interkvartil je raspon u kojem se nalazi 50% središnjih članova niza podataka.
- Vrijednost prvih i zadnjih 25% podataka ne utječe na vrijednost interkvartila.

Uočimo i da je interkvartil izražen u istoj mjernoj jedinici kao i podaci.Ako to želimo izbjeći pogodno je promotriti **koeficijent kvartilne devijacije** VQ definiran s

$$VQ = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Uočimo da smo time dobili jednu relativnu mjeru raspršenosti, s vrijednostima u $[0, 1]$, koja ne ovisi o jedinicama u kojima su izraženi podaci.

Primjer. Odredite donji i gornji kvartil, interkvartil i koeficijent kvartilne devijacije za sljedeći niz podataka:

1.3, 4.1, 4.1, 4.2, 4.4, 4.6, 5.1, 5.2, 5.3, 5.5, 5.5, 5.5, 5.9, 6.1, 7.8

Primjer. Odredite donji i gornji kvartil, interkvartil i koeficijent kvartilne devijacije za sljedeći niz podataka:

1.3, 4.1, 4.1, 4.2, 4.4, 4.6, 5.1, 5.2, 5.3, 5.5, 5.5, 5.5, 5.9, 6.1, 7.8

Rješenje. Prvo određujemo donji i gornji kvartil.

Primjer. Odredite donji i gornji kvartil, interkvartil i koeficijent kvartilne devijacije za sljedeći niz podataka:

1.3, 4.1, 4.1, 4.2, 4.4, 4.6, 5.1, 5.2, 5.3, 5.5, 5.5, 5.5, 5.9, 6.1, 7.8

Rješenje. Prvo određujemo donji i gornji kvartil.

Niz podijelimo na dva jednaka dijela s medijanom u sredini ($n = 15$ je neparan).

Primjer. Odredite donji i gornji kvartil, interkvartil i koeficijent kvartilne devijacije za sljedeći niz podataka:

1.3, 4.1, 4.1, 4.2, 4.4, 4.6, 5.1, 5.2, 5.3, 5.5, 5.5, 5.5, 5.9, 6.1, 7.8

Rješenje. Prvo određujemo donji i gornji kvartil.

Niz podijelimo na dva jednaka dijela s medijanom u sredini ($n = 15$ je neparan).

$$1.3\ 4.1\ 4.1\ 4.2\ 4.4\ 4.6\ 5.1 \quad | \quad 5.2 \quad | \quad 5.3\ 5.5\ 5.5\ 5.5\ 5.9\ 6.1\ 7.8$$

Primjer. Odredite donji i gornji kvartil, interkvartil i koeficijent kvartilne devijacije za sljedeći niz podataka:

1.3, 4.1, 4.1, 4.2, 4.4, 4.6, 5.1, 5.2, 5.3, 5.5, 5.5, 5.5, 5.9, 6.1, 7.8

Rješenje. Prvo određujemo donji i gornji kvartil.

Niz podijelimo na dva jednaka dijela s medijanom u sredini ($n = 15$ je neparan).

1.3 4.1 4.1 4.2 4.4 4.6 5.1 | 5.2 | 5.3 5.5 5.5 5.5 5.9 6.1 7.8

Donji kvartil (Q1) je medijan prve polovice niza: 1.3 4.1 4.1 **4.2** 4.4 4.6 5.1

Primjer. Odredite donji i gornji kvartil, interkvartil i koeficijent kvartilne devijacije za sljedeći niz podataka:

1.3, 4.1, 4.1, 4.2, 4.4, 4.6, 5.1, 5.2, 5.3, 5.5, 5.5, 5.5, 5.9, 6.1, 7.8

Rješenje. Prvo određujemo donji i gornji kvartil.

Niz podijelimo na dva jednaka dijela s medijanom u sredini ($n = 15$ je neparan).

1.3 4.1 4.1 4.2 4.4 4.6 5.1 | 5.2 | 5.3 5.5 5.5 5.5 5.9 6.1 7.8

Donji kvartil (Q1) je medijan prve polovice niza: 1.3 4.1 4.1 **4.2** 4.4 4.6 5.1

Gornji kvartil (Q3) je medijan druge polovice niza: 5.3 5.5 5.5 **5.5** 5.9 6.1 7.8

Sada jednostavno odredimo interkvartil:

$$IQ = Q_3 - Q_1 = 5.5 - 4.2 = 1.3$$

i koeficijent kvartilne devijacije:

$$VQ = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{1.3}{9.7} = 0.13402.$$

Kvantili (centili)

Centili su vrijednosti koji dijele niz podataka na 100 jednakih dijelova, decili na 10 jednakih dijelova. **p%-ti kvantil** (p-ti centil) je broj koji niz podataka dijeli tako da je p% podataka manje od p%-tog kvantila a $(100 - p)$% veće od njega.

Kvantili (centili)

Centili su vrijednosti koji dijele niz podataka na 100 jednakih dijelova, decili na 10 jednakih dijelova. **p%-ti kvantil** (p-ti centil) je broj koji niz podataka dijeli tako da je p% podataka manje od p%-tog kvantila a $(100 - p)$% veće od njega.

Medijan = 50-ti centil
Donji kvartil = 25-ti centil
Gornji kvartil = 75-ti centil.

Primjer. Dani su sljedeći podaci:

$$6, 3, 3, 6, 3, 5, 6, 1, 4, 6, 3, 5, 5, 2, 2, 2, 2, 3, 2, 3.$$

Odredite donji i gornji kvartil, kao i 5%-tni kvantil uzorka.

Primjer. Dani su sljedeći podaci:

$$6, 3, 3, 6, 3, 5, 6, 1, 4, 6, 3, 5, 5, 2, 2, 2, 2, 3, 2, 3.$$

Odredite donji i gornji kvartil, kao i 5%-tni kvantil uzorka.

Rješenje. Uočimo $n = 20$. Uredimo podatke po veličini:

$$1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6$$

Primjer. Dani su sljedeći podaci:

$$6, 3, 3, 6, 3, 5, 6, 1, 4, 6, 3, 5, 5, 2, 2, 2, 2, 3, 2, 3.$$

Odredite donji i gornji kvartil, kao i 5%-tni kvantil uzorka.

Rješenje. Uočimo $n = 20$. Uredimo podatke po veličini:

$$1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6$$

ili ih zapišemo u obliku frekvencijske tablice:

a_i	1	2	3	4	5	6
f_i	1	5	6	1	3	4.

Možemo koristiti proceduru iz prethodnog primjera ili direktno po formuli za kvartile (1):

$$Q_1 = x_{\left(\frac{21}{4}\right)} = x_{\left(5+\frac{1}{4}\right)} = \tfrac{3}{4}x_{(5)} + \tfrac{1}{4}x_{(6)} = \tfrac{3}{4}\cdot 2 + \tfrac{1}{4}\cdot 2 = 2,$$
$$Q_3 = x_{\left(\frac{63}{4}\right)} = x_{\left(15+\frac{3}{4}\right)} = \tfrac{1}{4}x_{(15))} + \tfrac{3}{4}x_{(16)} = \tfrac{1}{4}\cdot 5 + \tfrac{3}{4}\cdot 5 = 5.$$

Nadalje, 5%-tni kvantil jednak je

$$x_{0.05(n+1)} = x_{(1.05)} = 0.95x_{(1)} + 0.05x_{(2)} = 1.05.$$

Prikaz raspršenja preko kvartila

Kada se ukratko želi okarakterizirati neki skup podataka, često se promatra **karakteristična petorka** koju čine:

$$(x_{(1)}, Q_1, m, Q_3, x_{(n)})$$

minimum, donji kvartil, medijan, gornji kvartil i maksimum uzorka, redom.

Prikaz raspršenja preko kvartila

Kada se ukratko želi okarakterizirati neki skup podataka, često se promatra **karakteristična petorka** koju čine:

$$(x_{(1)}, Q_1, m, Q_3, x_{(n)})$$

minimum, donji kvartil, medijan, gornji kvartil i maksimum uzorka, redom.

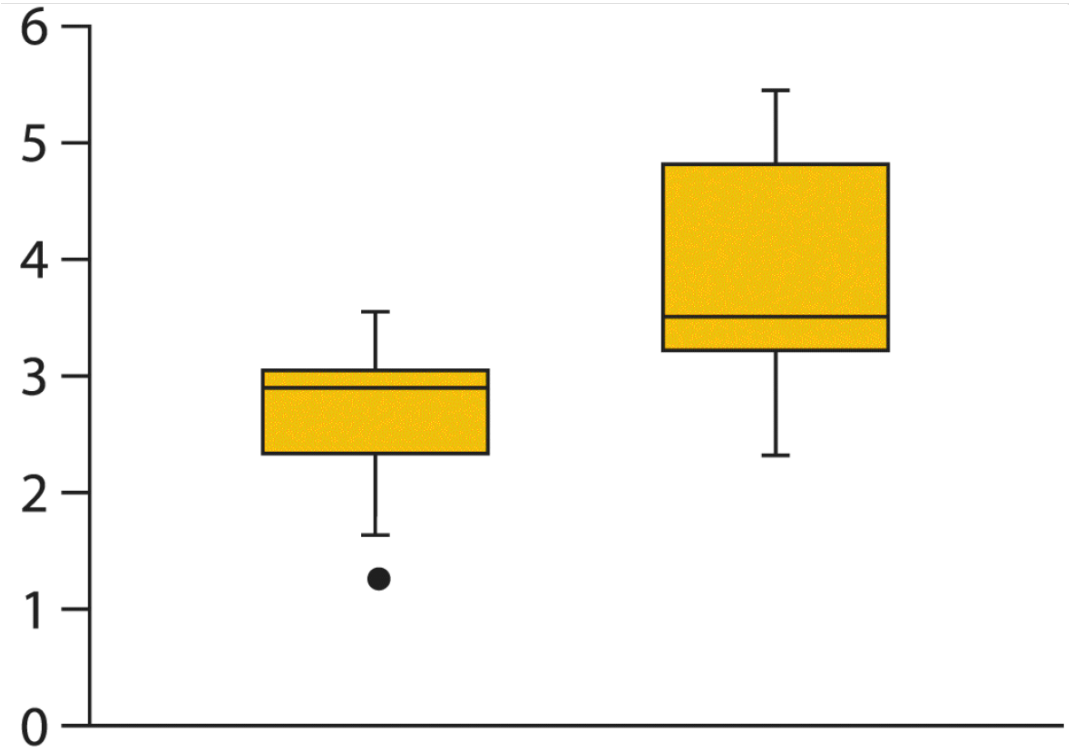
Elegantan grafički prikaz ovih karakteristika uzorka naziva se *box and whiskers* dijagram (doslovno dijagram "kutije i brkova" ili pravokutni dijagram):

- Na brojevnom pravcu označimo medijan m i oko njega izgradimo pravokutnik ("kutiju") od Q_1 do Q_3.
- Nakon toga, pronađemo najmanji (odnosno, najveći) podatak koji je od Q_1 (odnosno Q_3) udaljen za najviše $\frac{3}{2}IQ$.
- Povučemo linije ("brkove") od tih podataka do kutije. Sve podatke u uzorku koji se nalaze izvan tog intervala oko medijana označimo točkama (njih nazivamo ekstremi, odnosno *outliers*).

"Box-whiskers" graf

Promotrimo podatke o brzini u cm/s 16 mužjaka paukova iz roda *Tidarren* prije i nakon dobrovoljne amputacije jednog od palpa (pri sazrijevanju u spolno aktivne jedinke odrasle dobi).

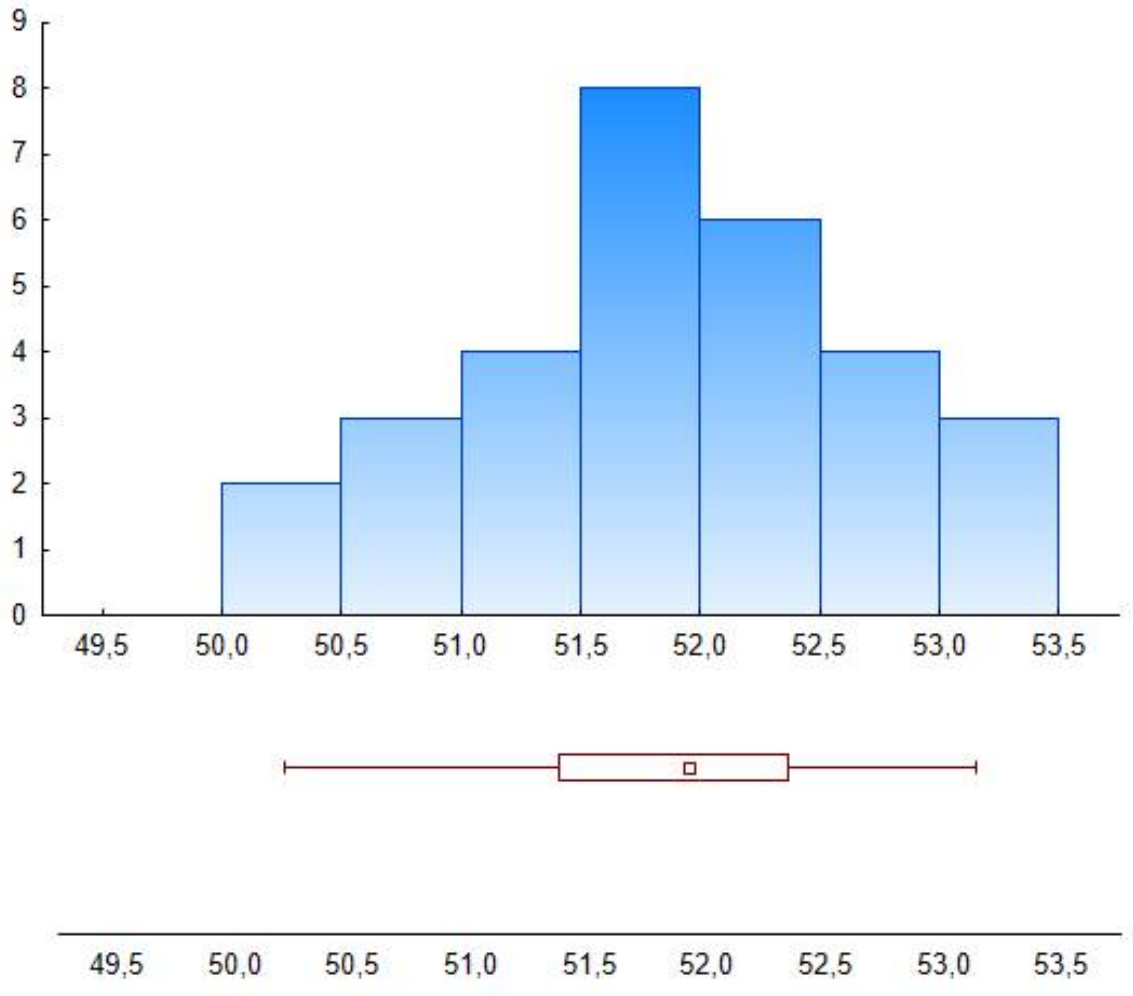
pauk	brzina prije	brzina poslije
1	1.25	2.40
2	2.94	3.50
3	2.38	4.49
4	3.09	3.17
5	3.41	5.26
6	3.00	3.22
7	2.31	2.32
8	2.93	3.31
9	2.98	3.70
10	3.55	4.70
11	2.84	4.94
12	1.64	5.06
13	3.22	3.22
14	2.87	3.52
15	2.37	5.45
16	1.91	3.40



Karakteristična petorka za *brzinu prije*: (1.25, 2.34, 2.90, 3.045, 3.55), pa je $\frac{3}{2}IQ = \frac{3}{2}(3.045 - 2.34) = 1.0575$ i $Q_1 - \frac{3}{2}IQ = 1.28$.

Boxplot i histogram

Nekad je zgodno paralelno promotriti histrogram (frekvencija) i boxplot na istom grafu.



Varijanca i standardna devijacija

Za skup podataka $x_1, x_2, \ldots, x_N$ definirajmo **odstupanje** i-tog podatka od srednje vrijednosti:

$$x_i - \overline{x}.$$

Varijanca i standardna devijacija

Za skup podataka $x_1, x_2, \ldots, x_N$ definirajmo **odstupanje** i-tog podatka od srednje vrijednosti:

$$x_i - \overline{x}.$$

Možemo sada pomisliti kako je prosječno odstupanje svih podataka pogodna mjera raspršenja, ali primijetimo da je

$$\frac{1}{N}\sum_i (x_i - \overline{x}) = 0.$$

tj., prosječno odstupanje je uvijek 0 i stoga nam to ništa ne govori o ukupnoj raspršenosti podataka.

Varijanca i standardna devijacija

Za skup podataka $x_1, x_2, \ldots, x_N$ definirajmo **odstupanje** i-tog podatka od srednje vrijednosti:

$$x_i - \overline{x}.$$

Možemo sada pomisliti kako je prosječno odstupanje svih podataka pogodna mjera raspršenja, ali primijetimo da je

$$\frac{1}{N}\sum_i (x_i - \overline{x}) = 0.$$

tj., prosječno odstupanje je uvijek 0 i stoga nam to ništa ne govori o ukupnoj raspršenosti podataka.
Stoga ima smisla u sljedećem koraku promotriti **kvadratno odstupanje** i-tog podatka od srednje vrijednosti:

$$(x_i - \overline{x})^2.$$

Varijanca i standardna devijacija

Za skup podataka $x_1, x_2, \ldots, x_N$ definirajmo **odstupanje** i-tog podatka od srednje vrijednosti:

$$x_i - \overline{x}.$$

Možemo sada pomisliti kako je prosječno odstupanje svih podataka pogodna mjera raspršenja, ali primijetimo da je

$$\frac{1}{N}\sum_i (x_i - \overline{x}) = 0.$$

tj., prosječno odstupanje je uvijek 0 i stoga nam to ništa ne govori o ukupnoj raspršenosti podataka.
Stoga ima smisla u sljedećem koraku promotriti **kvadratno odstupanje** i-tog podatka od srednje vrijednosti:

$$(x_i - \overline{x})^2.$$

- Kvadratno odstupanje je uvijek pozitivno.

Varijanca i standardna devijacija

Za skup podataka $x_1, x_2, \ldots, x_N$ definirajmo **odstupanje** i-tog podatka od srednje vrijednosti:

$$x_i - \overline{x}.$$

Možemo sada pomisliti kako je prosječno odstupanje svih podataka pogodna mjera raspršenja, ali primijetimo da je

$$\frac{1}{N}\sum_i (x_i - \overline{x}) = 0.$$

tj., prosječno odstupanje je uvijek 0 i stoga nam to ništa ne govori o ukupnoj raspršenosti podataka.
Stoga ima smisla u sljedećem koraku promotriti **kvadratno odstupanje** i-tog podatka od srednje vrijednosti:

$$(x_i - \overline{x})^2.$$

- Kvadratno odstupanje je uvijek pozitivno.
- Što je udaljenost podataka od srednje vrijednosti veća to je i kvadratno odstupanje veće.

Varijanca i standardna devijacija

Za skup podataka $x_1, x_2, \ldots, x_N$ definirajmo **odstupanje** i-tog podatka od srednje vrijednosti:

$$x_i - \overline{x}.$$

Možemo sada pomisliti kako je prosječno odstupanje svih podataka pogodna mjera raspršenja, ali primijetimo da je

$$\frac{1}{N}\sum_i (x_i - \overline{x}) = 0.$$

tj., prosječno odstupanje je uvijek 0 i stoga nam to ništa ne govori o ukupnoj raspršenosti podataka.
Stoga ima smisla u sljedećem koraku promotriti **kvadratno odstupanje** i-tog podatka od srednje vrijednosti:

$$(x_i - \overline{x})^2.$$

- Kvadratno odstupanje je uvijek pozitivno.
- Što je udaljenost podataka od srednje vrijednosti veća to je i kvadratno odstupanje veće.
- Srednje kvadratno odstupanje - mjera odstupanja svih podataka.

Varijanca je srednje kvadratno odstupanje podataka od srednje vrijednosti:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Varijanca je srednje kvadratno odstupanje podataka od srednje vrijednosti:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Alternativna formula za računanje varijance:

$$\sigma^2 = \frac{1}{N-1} \sum_{i} x_i^2 - \frac{N}{N-1} \overline{x}^2.$$

Varijanca je srednje kvadratno odstupanje podataka od srednje vrijednosti:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2.$$

Alternativna formula za računanje varijance:

$$\sigma^2 = \frac{1}{N-1}\sum_{i} x_i^2 - \frac{N}{N-1}\overline{x}^2.$$

Često umjesto varijance promatramo **standardnu devijaciju**, odnosno korijen iz varijance:

$$\sigma \quad = \quad \sqrt{\sigma^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

obzirom da je mjerna jedinica za standardnu devijaciju ista kao i za promatrano obilježje.

Primjer. Sandra Perković je na 65. Memorijalu Borisa Hanžekovića (2015) u bacanju diska postigla sljedeće rezultate u pet uspješnih bacanja:

62.30 m, 62.28 m, 69.77 m, 68.95 m, 69.88 m.

Odredite varijancu i standardnu devijaciju ovih bacanja.

Primjer. Sandra Perković je na 65. Memorijalu Borisa Hanžekovića (2015) u bacanju diska postigla sljedeće rezultate u pet uspješnih bacanja:

62.30 m, 62.28 m, 69.77 m, 68.95 m, 69.88 m.

Odredite varijancu i standardnu devijaciju ovih bacanja.

Prvo treba odrediti srednju vrijednost:

$$\overline{x} = \frac{1}{5}\left(62.30 + 62.28 + 69.77 + 68.95 + 69.88\right) = \frac{333.18}{5} = 63.636,$$

a zatim varijancu:

$$\sigma^2 = \frac{1}{4}\left[62.30^2 + 62.28^2 + 69.77^2 + 68.95^2 + 69.88^2\right] - \tfrac{5}{4}63.636^2 = 15.868\mathrm{m}^2.$$

Sada je standardna devijacija jednaka:

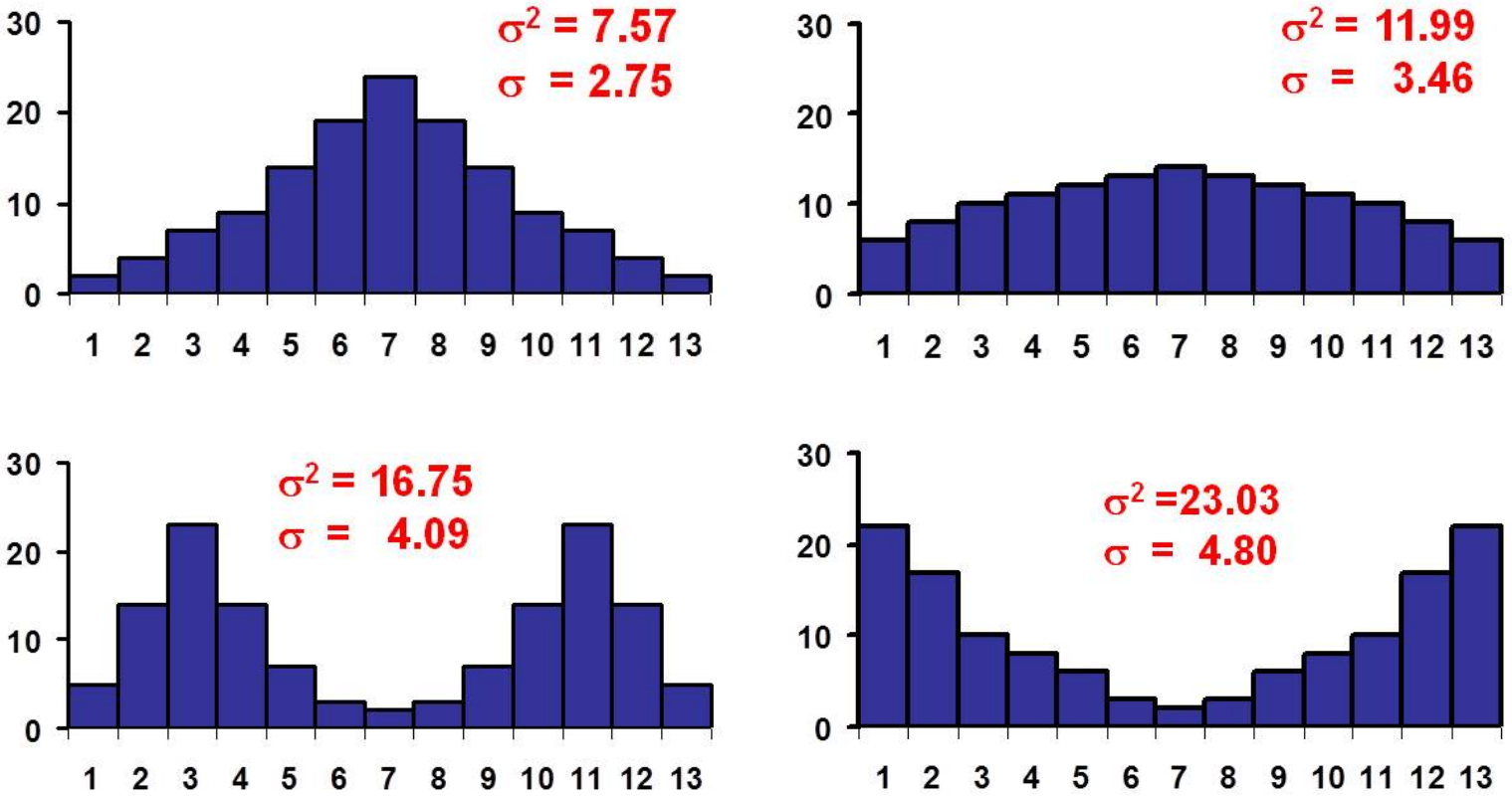
$$\sigma = \sqrt{\sigma^2} = \sqrt{15.868} = 3.98358\mathrm{m}.$$

Rješenje možemo provesti i tablično:

i	x_i	x_i^2
1	62.30	3881.290
2	62.28	3878.798
3	69.77	4867.853
4	68.95	4754.103
5	69.88	4883.214
$\sum$	333.18	22265.258

$$\begin{aligned} \overline{x} &= \frac{1}{N}\sum_i x_i = \frac{1}{5}\cdot 333.18 = 66.636 \\ \sigma^2 &= \frac{1}{N-1}\sum_i x_i^2 - \frac{N}{N-1}\overline{x}^2 = \frac{1}{4}\cdot 22265.258 - \frac{5}{4}66.636^2 = 15.868 \\ \sigma &= \sqrt{\sigma^2} = \sqrt{15.868} = 3.98358 \end{aligned}$$

Varijanca i standardna devijacija za podatke iz uvodnog primjera:



Računanje varijance iz tablice frekvencija

Neka su $x_1, x_2, x_3, \ldots, x_k$ vrijednosti obilježja i neka su $f_1, f_2, f_3, \ldots, f_k$ pripadne frekvencije. Tada je

$$\sigma^2 = \frac{1}{N-1}\sum_i f_i(x_i - \overline{x})^2 = \frac{1}{N-1}\sum_i f_i x_i^2 - \frac{N}{N-1}\overline{x}^2.$$

Računanje varijance iz tablice frekvencija

Neka su $x_1, x_2, x_3, \ldots, x_k$ vrijednosti obilježja i neka su $f_1, f_2, f_3, \ldots, f_k$ pripadne frekvencije. Tada je

$$\sigma^2 = \frac{1}{N-1}\sum_i f_i(x_i - \overline{x})^2 = \frac{1}{N-1}\sum_i f_i x_i^2 - \frac{N}{N-1}\overline{x}^2.$$

Primjer. Izračunajte varijancu iz distribucije frekvencija obilježja dob:

x_i (Dob)	f_i (Frekvencija)	$f_i \cdot x_i$	$f_i \cdot x_i^2$
19	1	19	361
20	0	0	0
21	5	105	2205
22	2	44	968
23	2	46	1058
24	1	24	576
25	4	100	2500
26	1	26	676
27	0	0	0
28	1	28	784
29	0	0	0
30	1	30	900
31	1	31	961
32	1	32	1024
$\sum$	20	485	12013

Računanje varijance iz tablice frekvencija

Neka su $x_1, x_2, x_3, \ldots, x_k$ vrijednosti obilježja i neka su $f_1, f_2, f_3, \ldots, f_k$ pripadne frekvencije. Tada je

$$\sigma^2 = \frac{1}{N-1}\sum_i f_i(x_i - \overline{x})^2 = \frac{1}{N-1}\sum_i f_i x_i^2 - \frac{N}{N-1}\overline{x}^2.$$

Primjer. Izračunajte varijancu iz distribucije frekvencija obilježja dob:

x_i (Dob)	f_i (Frekvencija)	$f_i \cdot x_i$	$f_i \cdot x_i^2$
19	1	19	361
20	0	0	0
21	5	105	2205
22	2	44	968
23	2	46	1058
24	1	24	576
25	4	100	2500
26	1	26	676
27	0	0	0
28	1	28	784
29	0	0	0
30	1	30	900
31	1	31	961
32	1	32	1024
$\sum$	20	485	12013

$$\begin{aligned}\sigma^2 &= \frac{1}{N-1}\sum_i f_i x_i^2 - \frac{N}{N-1}\overline{x}^2 = \\ &= \frac{1}{19}\cdot 12031 - \frac{20}{19}\cdot 25.52^2 = \\ &= 14.1974.\end{aligned}$$

Koeficijent varijacije

Varijanca i standardna devijacija ovise o mjernim jedinicama.

Koeficijent varijacije

Varijanca i standardna devijacija ovise o mjernim jedinicama.

- Mjerna jedinica za varijancu je kvadrirana mjerna jedinica varijable.

Koeficijent varijacije

Varijanca i standardna devijacija ovise o mjernim jedinicama.

- Mjerna jedinica za varijancu je kvadrirana mjerna jedinica varijable.
- Mjerna jedinica za standardnu devijaciju je jednaka mjernoj jedinici varijable.

Koeficijent varijacije

Varijanca i standardna devijacija ovise o mjernim jedinicama.

- Mjerna jedinica za varijancu je kvadrirana mjerna jedinica varijable.
- Mjerna jedinica za standardnu devijaciju je jednaka mjernoj jedinici varijable.

Koeficijent varijacije:

$$CV = \frac{\sigma}{\overline{x}}.$$

Koeficijent varijacije nema mjernu jedinicu i govori nam kakva je standardna devijacija uzorka u odnosu na srednju vrijednost (npr. smatramo da je uzorak kojem je $\overline{x} = 1$ i $\sigma = 1$ više raspršen od uzorka s $\overline{x} = 100$ i $\sigma = 1$.)

Ostale mjere raspršenosti - srednje apsolutno odstupanje

Umjesto kvadratnog odstupanja podataka od srednje vrijednosti:

$$(x_i - \overline{x})^2$$

mogli smo promatrati apsolutno odstupanje:

$$|x_i - \overline{x}|.$$

Time dobvamoi mjeru raspršenosti poznatu kao **srednje apsolutno odstupanje** (*(engl. MAD - mean absolute deviation)*):

$$\text{s.a.o.} = \frac{1}{N} \sum_i |x_i - \overline{x}|.$$

Može se gledati i srednje apsolutno odstupanje od drugih mjera centralne tendencije (medijan, mod...).

Mjere centralne tendencije i raspršenosti za grupirane podatke

Promotrimo kako računamo mjere centralne tendencije i raspršenosti za grupirane podatke. tj. ako pretpostavimo da su nam zadani razredi i njihove frekvencije (f_i). Mjere centralne tendencije i raspršenosti ne možemo egzaktno izračunati iz grupiranih podataka, ali ih možemo aproksimirati na sljedeći način:

Mjere centralne tendencije i raspršenosti za grupirane podatke

Promotrimo kako računamo mjere centralne tendencije i raspršenosti za grupirane podatke. tj. ako pretpostavimo da su nam zadani razredi i njihove frekvencije (f_i). Mjere centralne tendencije i raspršenosti ne možemo egzaktno izračunati iz grupiranih podataka, ali ih možemo aproksimirati na sljedeći način:

- Za svaki razred izračunamo sredinu razreda x_i.

Mjere centralne tendencije i raspršenosti za grupirane podatke

Promotrimo kako računamo mjere centralne tendencije i raspršenosti za grupirane podatke. tj. ako pretpostavimo da su nam zadani razredi i njihove frekvencije (f_i). Mjere centralne tendencije i raspršenosti ne možemo egzaktno izračunati iz grupiranih podataka, ali ih možemo aproksimirati na sljedeći način:

- Za svaki razred izračunamo sredinu razreda x_i.
- Pretpostavimo da svi podaci u i-tom razreda imaju istu vrijednost x_i, odnosno formiramo novi uzorak u kojem se srednja vrijednost x_i pojavljuje kao podatak f_i puta.

Mjere centralne tendencije i raspršenosti za grupirane podatke

Promotrimo kako računamo mjere centralne tendencije i raspršenosti za grupirane podatke. tj. ako pretpostavimo da su nam zadani razredi i njihove frekvencije (f_i). Mjere centralne tendencije i raspršenosti ne možemo egzaktno izračunati iz grupiranih podataka, ali ih možemo aproksimirati na sljedeći način:

- Za svaki razred izračunamo sredinu razreda x_i.
- Pretpostavimo da svi podaci u i-tom razreda imaju istu vrijednost x_i, odnosno formiramo novi uzorak u kojem se srednja vrijednost x_i pojavljuje kao podatak f_i puta.
- Izračunamo promatranu mjeru.

Mjere centralne tendencije i raspršenosti za grupirane podatke

Promotrimo kako računamo mjere centralne tendencije i raspršenosti za grupirane podatke. tj. ako pretpostavimo da su nam zadani razredi i njihove frekvencije (f_i). Mjere centralne tendencije i raspršenosti ne možemo egzaktno izračunati iz grupiranih podataka, ali ih možemo aproksimirati na sljedeći način:

- Za svaki razred izračunamo sredinu razreda x_i.
- Pretpostavimo da svi podaci u i-tom razreda imaju istu vrijednost x_i, odnosno formiramo novi uzorak u kojem se srednja vrijednost x_i pojavljuje kao podatak f_i puta.
- Izračunamo promatranu mjeru.

Točnije, srednju vrijednost odredimo kao:

$$\overline{x} = \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{\sum f_i X_i}{N},$$

Mjere centralne tendencije i raspršenosti za grupirane podatke

Promotrimo kako računamo mjere centralne tendencije i raspršenosti za grupirane podatke. tj. ako pretpostavimo da su nam zadani razredi i njihove frekvencije (f_i). Mjere centralne tendencije i raspršenosti ne možemo egzaktno izračunati iz grupiranih podataka, ali ih možemo aproksimirati na sljedeći način:

- Za svaki razred izračunamo sredinu razreda x_i.
- Pretpostavimo da svi podaci u i-tom razreda imaju istu vrijednost x_i, odnosno formiramo novi uzorak u kojem se srednja vrijednost x_i pojavljuje kao podatak f_i puta.
- Izračunamo promatranu mjeru.

Točnije, srednju vrijednost odredimo kao:

$$\overline{x} = \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{\sum f_i X_i}{N},$$

a varijancu kao:

$$\sigma^2 = \frac{1}{N-1} \sum_i f_i (x_i - \overline{x})^2.$$

Mjere centralne tendencije i raspršenosti za grupirane podatke

Primjer: Odredite srednju vrijednost, medijan, varijancu i donji i gornji kvartil sljedećeg grupiranog uzorka:

interval I_i	f_i	r_i
$1.0 - 1.2$	1	$1/14$
$1.3 - 1.4$	3	$3/14$
$1.5 - 1.6$	6	$6/14$
$1.7 - 1.8$	1	$1/14$
$1.9 - 2.0$	3	$3/14$
$\sum$	14	1

Uočimo da se pretpostavlja da su negrupirani podaci bili zaokruženi do na jednu decimalu (što opravdava razmak od 0.1 između razreda).

Mjere centralne tendencije i raspršenosti za grupirane podatke

Primjer: Odredite srednju vrijednost, medijan, varijancu i donji i gornji kvartil sljedećeg grupiranog uzorka:

interval I_i	f_i	r_i
$1.0 - 1.2$	1	$1/14$
$1.3 - 1.4$	3	$3/14$
$1.5 - 1.6$	6	$6/14$
$1.7 - 1.8$	1	$1/14$
$1.9 - 2.0$	3	$3/14$
$\sum$	14	1

Uočimo da se pretpostavlja da su negrupirani podaci bili zaokruženi do na jednu decimalu (što opravdava razmak od 0.1 između razreda).

Rješenje: Ideja je da kreiramo *pomoćni* uzorak, tako da svaki element razreda reprezentiramo srednjom vrijednosti tog razreda. Time dobivamo uzorak:

$$1.10,\ 1.35,\ 1.35,\ 1.35,\ 1.55,\ 1.55,\ 1.55,\ 1.55,\ 1.55,\ 1.55,\ 1.75,\ 1.95,\ 1.95,\ 1.95.$$

Sada na isti način kao prije odredimo aritmetičku sredinu (1.60), medijan (1.55), varijancu (0.0641) i kvartile uzorka (1.40 i 1.70).