2 Elementárne spracovanie štatistických údajov

V tejto kapitole sa sústredíme na opis jednorozmerných štatistických súborov s kvantitatívnym triediacim znakom, pričom v závere kapitoly ukážeme krátko aj elementárne spracovanie štatistických súborov s kvalitatívnym triediacim znakom.

2.1 Elementárne spracovanie súborov s kvantitatívnym triediacim znakom

Výsledkom štatistického zisťovania je spravidla veľké množstvo číselných údajov, ktoré zapisujeme v takom poradí, v akom sme ich získali. Tieto údaje sú väčšinou neprehľadné. Aby vynikli charakteristické rysy a zákonitosti analyzovaného súboru a aby sa údaje stali prehľadnými, musíme ich roztriediť.

Triedenie štatistických údajov spočíva v tom, že sa daný štatistický súbor rozčlení na skupiny (triedy) rovnorodé z hľadiska určitého štatistického znaku. Triedenie je metóda pre usporiadanie údajov do prehľadnej formy a tiež ich zhustenie.

Klasifikácia triedenia (usporiadania) štatistických údajov:

1. *Triedenie prosté*. Hodnoty $x_1, x_2, \dots, x_n$ triediaceho znaku v súbore usporiadame podľa veľkosti ($x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$) do neklesajúcej postupnosti, pričom zapíšeme všetky hodnoty toľkokrát, koľkokrát boli namerané. Výsledkom je tzv. ***variačný rad*** $x_{(1)}, x_{(2)}, \dots, x_{(n)}$. Toto triedenie používame pre súbory s malým rozsahom (za hranicu medzi malými a veľkými súbormi sa obvykle považuje rozsah 30).
2. *Triedenie početností*. Pre každú hodnotu x_i triediaceho znaku určíme jej ***absolútnu početnosť*** n_i, ktorá udáva, koľkokrát sa hodnota x_i v súbore o rozsahu n vyskytuje. Platí, že $\sum_{i=1}^{k} n_i = n$, kde k znamená počet tried (v tomto prípade každý triediaci znak predstavuje samostatnú triedu). Výsledkom je ***tabuľka rozdelenia absolútnych početností*** **(*frekvenčná tabuľka*)**. Toto triedenie používame pre súbory s veľkým rozsahom a malým počtom rôznych hodnôt v súbore.
3. *Triedenie intervalové*. Prvky súboru sa rozdelia do intervalov a pre každý interval sa určí jeho absolútna početnosť n_i, t. j. počet prvkov v tomto intervale. Interval I_i predstavuje samostatnú triedu a je jednoznačne určený, ak poznáme jeho dolnú hranicu t_{i-1} a hornú hranicu t_i. Obvykle sa interval reprezentuje tzv. ***triednym znakom***, ktorý určujeme ako stred daného triedneho intervalu. Triedny znak z_i je stredom i-teho triedneho intervalu, t. j. $z_i = (t_{i-1} + t_i)/2$. Výsledky sa aj v tomto prípade zapíšu prehľadne vo forme tabuľky. Toto triedenie používame pre súbory s veľkým rozsahom a veľkým počtom rôznych hodnôt v súbore.

Pri triedení štatistických údajov do triednych intervalov (tried) je potrebné, aby sme sa riadili týmito zásadami:

- Počet tried nesmie byť príliš malý, aby to neviedlo k veľmi zjednodušenému pohľadu na vlastnosti súboru a nemal by byť ani príliš veľký, pretože by sa mohlo stať, že sa spracovanie stane neprehľadným a zaniknú zákonitosti charakteristické pre daný súbor.
- Jednotlivé pozorované hodnoty štatistického znaku patria do jednej a len jednej triedy. Tento problém je spojený s otázkou, ako určovať hranice tried, aby bolo možné jednotlivé hodnoty začleniť do príslušných tried jednoznačne.
- Pokiaľ je to možné, stanovíme zhodnú šírku, t. j. dĺžku triedneho intervalu, pre všetky triedy.

Pri určovaní počtu tried sa snažíme potlačiť náhodné kolísanie absolútnych početností, ale zároveň nesmieme zastrieť charakteristické črty. Na stanovenie počtu tried neexistuje jednotný názor ani všeobecný predpis. Počet tried k volíme s prihliadnutím na povahu štatistického znaku a rozsah štatistického súboru n.

Často sa používa tzv. *Sturgessovo pravidlo*: $k \approx 1 + 3{,}3 \cdot \log n$. Niekedy je vhodné použiť odhad: $k \approx \sqrt{n}$. V technickej praxi sa používa aj odhad: $0{,}55 \cdot n^{0{,}4} \le k \le 1{,}25 \cdot n^{0{,}4}$.

Pri zaraďovaní štatistických údajov do triednych intervalov (tried) je potrebné pre každú triedu zvoliť vopred jeden z nasledujúcich postupov:

1. Hodnoty, ktoré sú rovné dolnej hranici i-teho triedneho intervalu, sa zaradia do i-tej triedy a hodnoty, ktoré sa rovnajú hornej hranici triedneho intervalu, sa zaradia do (i+1)-vej triedy, t. j. $I_i = \langle t_{i-1}, t_i)$.
2. Hodnoty rovné dolnej hranici i-teho triedneho intervalu sa zaradia do (i–1)-vej triedy a hodnoty rovné hornej hranici triedneho intervalu sa zaradia do i-tej triedy, t. j. $I_i = (t_{i-1}, t_i \rangle$.

Dĺžka triedneho intervalu h sa volí približne podľa vzťahu $h \approx \frac{R_V}{k}$, kde $R_V = x_{\max} - x_{\min}$ je ***variačné rozpätie***, pričom $x_{\max}$ predstavuje najväčšiu a $x_{\min}$ najmenšiu hodnotu štatistického znaku v danom štatistickom súbore.

Okrem *absolútnych početností* n_i sa v opisnej štatistike používajú aj ďalšie druhy početností, a to:

- ***relatívna početnosť*** $f_i = \frac{n_i}{n}$, pre $i = 1, 2, \ldots, k$;
- ***kumulatívna absolútna početnosť*** $N_i = n_1 + n_2 + \cdots + n_i = \sum_{j=1}^{i} n_j$, pre $i = 1, 2, \ldots, k$;
- ***kumulatívna relatívna početnosť*** $F_i = \frac{N_i}{n} = f_1 + f_2 + \cdots + f_i$, pre $i = 1, 2, \ldots, k$.

Platí: $\sum_{i=1}^{k} f_i = F_k = 1$, $\sum_{i=1}^{k} n_i = N_k = n$.

Relatívne početnosti sa často udávajú v percentách.

V tabuľke rozdelenia absolútnych početností n_i môžeme uviesť ešte aj početnosti f_i, N_i a F_i. Dostaneme tak ***tabuľku rozdelenia početností*** v nasledujúcom tvare:

i	x_i	n_i	f_i	N_i	F_i
1	x_1	n_1	f_1	n_1	f_1
2	x_2	n_2	f_2	$n_1 + n_2$	$f_1 + f_2$
⋮	⋮	⋮	⋮	⋮	⋮
k	x_k	n_k	f_k	n	1
$\sum$	—	n	1	—	—

V tejto tabuľke v prípade intervalového triedenia použijeme namiesto hodnôt x_i triediaceho znaku hodnoty triedneho znaku z_i.

Pre intervalové triedenie sa pomocou kumulatívnych relatívnych početností F_i definuje ***empirická distribučná funkcia*** $F : R \to R$ nasledujúcim spôsobom:

$$F(x) = \begin{cases} 0 & \text{pre} \quad -\infty < x < z_1 \\ F_1 & \text{pre} \quad z_1 \le x < z_2 \\ F_2 & \text{pre} \quad z_2 \le x < z_3 \\ \vdots & \qquad \vdots \\ F_{k-1} & \text{pre} \quad z_{k-1} \le x < z_k \\ 1 & \text{pre} \quad z_k \le x < \infty \end{cases}.$$

V prípade obyčajného triedenia početností použijeme namiesto hodnôt triedneho znaku z_i hodnoty triediaceho znaku x_i.

Empirická distribučná funkcia vyjadruje teda závislosť kumulatívnej relatívnej početnosti na hodnotách kvantitatívneho znaku. Má tieto vlastnosti:

1. $F(x)$ je neklesajúca funkcia, t. j. pre každé reálne $x_1 \le x_2$ platí $F(x_1) \le F(x_2)$.
2. $F(-\infty) = 0$ a $F(+\infty) = 1$.
3. $F(x)$ je spojitá sprava na celej množine reálnych čísel R.

Grafom empirickej distribučnej funkcie je schodovitá krivka so skokmi v bodoch, ktoré sú triednymi znakmi, pričom súradnice y bodov grafu sa rovnajú hodnotám kumulatívnej relatívnej početnosti F_i.

Popri štatistických tabuľkách je dôležitou formou zobrazovania štatistických údajov graf. ***Grafické zobrazenie*** štatistického súboru je geometrickým obrazom výsledkov získaných sledovaním nejakého štatistického znaku v štatistickom súbore. Poskytuje rovnakú informáciu o empirickom rozdelení ako štatistická tabuľka, ale iným spôsobom. Grafické zobrazenie dáva rýchlu a prehľadnú predstavu o tendenciách a charakteristických črtách analyzovaných javov. Grafy sú tiež účinným popularizačným prostriedkom štatistických výsledkov.

Bodové grafy sú grafy vytvorené bodmi (z_i, w_i), kde za w_i môžeme dosadiť početnosti n_i, f_i, N_i a F_i, pričom $i = 1, 2, \ldots, k$, umiestnenými v pravouhlej súradnicovej sústave. V prípade obyčajného triedenia početností použijeme namiesto hodnôt triedneho znaku z_i hodnoty triediaceho znaku x_i.

Spojnicové grafy sú v podstate bodové grafy, v ktorých sú jednotlivé body spojené lomenou čiarou.

Polygón je spojnicový graf, v ktorom sú jednotlivé susedné body spojené úsečkami.

Ak cez jednotlivé body bodového grafu preložíme spojitú hladkú krivku, dostaneme ***frekvenčnú krivku***.

Histogram je graf, ktorý zostrojíme tak, že nad úsekom na osi x, rovným dĺžke triedneho intervalu, nakreslíme obdĺžnik, ktorého výška je úmerná w_i.

Stĺpcový graf je podobný histogramu, ale stĺpce sú oddelené. Používa sa hlavne v prípade obyčajného triedenia početností.

Úsečkový graf zostrojíme tak, že nad hodnotou triedneho znaku z_i, resp. triediaceho znaku x_i, nakreslíme úsečku o dĺžke úmernej w_i.

***Krabicový diagram* (*boxplot*)** je schéma, ktorá v jednom obrázku poskytuje informáciu o významných a extrémnych hodnotách štatistického súboru. Zostavenie tohto diagramu ukážeme neskôr.

Príklad 2.1. V strojárskom závode opracoval pracovník za jednu smenu 20 odliatkov. Každý výrobok prešiel technickou kontrolou a výsledky merania v milimetroch boli získane v tomto poradí: 38,5; 38,5; 38,2; 38,3; 38,2; 38,6; 38,8; 38,4; 38,5; 38,5; 38,6; 38,3; 38,5; 38,4; 38,7; 38,4; 38,6; 38,3; 38,6; 38,4.

Pre daný štatistický súbor zostavte: a) variačný rad, b) frekvenčnú tabuľku, c) tabuľku rozdelenia početností, d) polygón absolútnych početností n_i, úsečkový graf relatívnych početností f_i, stĺpcový graf kumulatívnych absolútnych početností N_i, schodovitý graf kumulatívnych relatívnych početností F_i a histogram absolútnych početností n_i.

Riešenie: a) Daný štatistický súbor má rozsah $n = 20$. Namerané hodnoty štatistického znaku usporiadame do *variačného radu*:

38,2 38,2 38,3 38,3 38,3 38,4 38,4 38,4 38,4 38,5 38,5 38,5 38,5 38,5 38,6 38,6 38,6 38,6 38,7 38,8.

Na tomto malom štatistickom súbore budeme ďalej ilustrovať aplikáciu programu MATLAB. Vektor neroztriedených hodnôt uložíme do premennej x a variačný rad, ktorý môžeme získať pomocou štandardnej funkcie MATLABu (M-funkcie) `sort`, uložíme do premennej *vr*:

```
x=[385,385,382,383,382,386,388,384,385,385,386,383,385,384,...
387,384,386,383,386,384]/10,vr=sort(x)
```

b) Pre každú nameranú hodnotu zistíme počet prípadov, v ktorých sa daná hodnota v súbore vyskytuje, t. j. jej *absolútnu početnosť* n_i. Takto získame *frekvenčnú tabuľku*:

x_i	38,2	38,3	38,4	38,5	38,6	38,7	38,8
n_i	2	3	4	5	4	1	1

Jej získanie ilustrujeme opäť použitím MATLABu. Pomocou M-funkcie `unique` najprv získame vektor usporiadaných rôznych hodnôt:

```
xi=unique(x)
```

Výstup z MATLABu je v tvare riadkového vektora:

```
xi = 38.2000   38.3000   38.4000   38.5000   38.6000   38.7000   38.8000
```

Absolútne početnosti n_i získame pomocou M-funkcie `hist` s jedným výstupným argumentom a uložíme ich do premennej *ni*:

```
ni=hist(x,xi)
```

Výstup z MATLABu:

```
ni = 2     3     4     5     4     1     1
```

Frekvenčnú tabuľku zostavíme v tvare blokovej matice:

```
fr_tab=[xi;ni]
```

Výstup z MATLABu:

```
fr_tab =
   38.2000   38.3000   38.4000   38.5000   38.6000   38.7000   38.8000
    2.0000    3.0000    4.0000    5.0000    4.0000    1.0000    1.0000
```

c) Ďalej ešte vypočítame početnosti f_i, N_i a F_i pomocou uvedených vzorcov $f_i = n_i/n$, $N_i = n_1 + n_2 + \cdots + n_i$, $F_i = f_1 + f_2 + \cdots + f_i$, pre $i = 1, 2, \ldots, k$ a zostavíme *tabuľku rozdelenia početností*:

i	x_i	n_i	f_i	N_i	F_i
1	38,2	2	0,1	2	0,1
2	38,3	3	0,15	5	0,25
3	38,4	4	0,2	9	0,45
4	38,5	5	0,25	14	0,7
5	38,6	4	0,2	18	0,9
6	38,7	1	0,05	19	0,95
7	38,8	1	0,05	20	1
∑	—	20	1	—	—

Na výpočet ďalších druhov početností použijeme uvedené vzorce, ktoré ľahko prepíšeme do MATLABu. V prípade početností N_i a F_i použijeme M-funkciu `cumsum`, ktorá vytvorí vektor kumulatívnych súčtov prvkov vektora. Tabuľku rozdelenia početností zapíšeme prehľadne vo forme matice a uložíme ju do premennej *tab*:

```
n=length(x);fi=ni/n;Ni=cumsum(ni);Fi=cumsum(fi);
tab=[xi;ni;fi;Ni;Fi]'
```

```
tab =
   38.2000    2.0000    0.1000    2.0000    0.1000
   38.3000    3.0000    0.1500    5.0000    0.2500
   38.4000    4.0000    0.2000    9.0000    0.4500
   38.5000    5.0000    0.2500   14.0000    0.7000
   38.6000    4.0000    0.2000   18.0000    0.9000
   38.7000    1.0000    0.0500   19.0000    0.9500
   38.8000    1.0000    0.0500   20.0000    1.0000
```

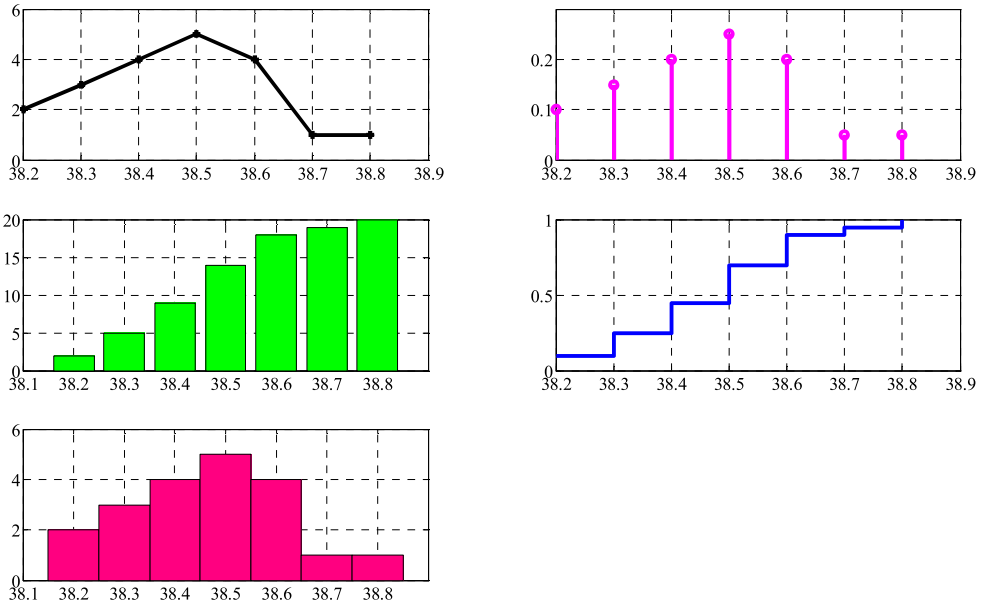
Poznámka: Výsledky budeme uvádzať zaokrúhlené na 4 desatinné miesta, čo je bežný výstup výsledkov MATLABu (`short`). Ak chceme zvýšiť presnosť výstupov, použijeme prepínač `format long`, `format short e`, resp. iné (pozrite `help format`).

d) Daný štatistický súbor znázorníme aj graficky. Pomocou M-funkcie `plot` sa znázorňuje polygón, pomocou `stem` úsečkový graf, pomocou `bar` stĺpcový graf. Pomocou M-funkcie `stairs` s použitím početností F_i dostaneme schodovitý graf, ktorý pripomína graf empirickej distribučnej funkcie. Ak použijeme M-funkciu `hist` bez výstupného argumentu, dostaneme grafické znázornenie rozdelenia absolútnych početností – histogram absolútnych početností.

Ďalej využijeme M-funkciu `subplot`, ktorá nám umožňuje umiestniť viac obrázkov do jedného grafického okna.

Fungovanie príkazov v MATLABe si môžeme pozrieť prostredníctvom „helpu“, kde sú informácie zapísane v angličtine. Napr. do príkazového riadku zapíšeme `help plot`. Pri kreslení grafu máme možnosť zvoliť si farbu, rôzne zvýraznenia bodov, ako aj hrúbku čiary. Dva jednoduché apostrofy ohraničujú reťazec znakov. Pomocou `colormap` môžeme meniť farebnú paletu. Pomocou `grid` pridávame sieťku. Upraviť obrázok môžeme cez *Edit* v hornej časti grafického okna.

```
subplot(3,2,1);plot(xi,ni,'k*-','linewidth',3),grid,
subplot(3,2,2);stem(xi,fi,'m','linewidth',3),grid,
subplot(3,2,3);bar(xi,Ni,'g'),grid,
subplot(3,2,4);stairs(xi,Fi,'linewidth',3),grid,
subplot(3,2,5);hist(x,xi),colormap([1 0 0.5]),grid
```



Príklad 2.2. Máme k dispozícii údaje o príjmoch tridsiatich domácností v Sk v roku 1991 na Slovensku: 9966, 5105, 13911, 7080, 6640, 8003, 6905, 12999, 11110, 3060, 10401, 2880, 6732, 12625, 12012, 11763, 9968, 8720, 9043, 10509, 8826, 7740, 4970, 7260, 10395, 6830, 8390, 14165, 5165, 7260.

Riešte nasledujúce úlohy: a) daný štatistický súbor roztrieďte do triednych intervalov, pričom zvoľte vhodný počet tried k; b) zostavte tabuľku rozdelenia početností, c) znázornite schodovitý graf kumulatívnych relatívnych početností F_i, d) určte empirickú distribučnú funkciu a jej graf.

Riešenie: a) Daný štatistický súbor má rozsah $n = 30$. Ak si pozorne prezrieme namerané hodnoty, tak zistíme, že najmenšou hodnotou daného štatistického súboru je hodnota 2880 a najväčšou je hodnota 14165.

Aplikujeme opäť MATLAB:

```
x=[9966,5105,13911,7080,6640,8003,6905,12999,11110,3060,10401,...
2880,6732,12625,12012,11763,9968,8720,9043,10509,8826,7740,4970,...
7260,10395,6830,8390,14165,5165,7260];
n=length(x),xmin=min(x),xmax=max(x)
```

Vhodný počet tried k určíme pomocou *Sturgessovho pravidla*: $k \approx 1 + 3{,}3 \cdot \log 30 = 5{,}8745$. Zvolíme teda 6 tried. Uvedený vzorec prepíšeme pre zaujímavosť aj do MATLABu pomocou M-funkcie pre dekadický logaritmus `log10` a M-funkcie `round`, ktorá zaokrúhli prvky matice (vektora) na celé čísla podľa pravidiel zaokrúhľovania:

```
k=round(1+3.3*log10(n))
```

Dĺžku triedneho intervalu zvolíme podľa vzťahu $h \approx (x_{\max} - x_{\min})/k = 11285/6 = 1880{,}8\overline{3}$. Pre jednoduchosť a názornosť zvolíme rovnakú dĺžku triedneho intervalu pre všetky triedy $h = 2000$. Minimálnu hodnotu štatistického znaku zaokrúhlime vhodne nadol, napr. na 2500 a maximálnu hodnotu vhodne nahor, napr. na 14500. Takto dostaneme tieto hranice triednych intervalov (tried): 2500, 4500, 6500, 8500, 10500, 12500, 14500. Potom vypočítame triedne znaky z_i, t. j. reprezentantov jednotlivých tried, ako stredy príslušných triednych intervalov. Dostaneme tieto hodnoty triednych znakov: 3500, 5500, 7500, 9500, 11500, 13500.

Aplikujeme teraz znova MATLAB. Vytvoríme vektor *t*, ktorý tvoria hranice triednych intervalov a vektor *zi* triednych znakov:

```
t=2500:2000:14500,zi=(t(1:6)+t(2:7))/2
```

Pri zaraďovaní štatistických údajov do triednych intervalov zvolíme v poradí druhý z dvoch v teoretickej časti spomenutých postupov. Takto získame absolútne početnosti n_i.

Pri tejto prácnej úlohe môžeme efektívne využiť MATLAB. Absolútne početnosti získame pomocou M-funkcie `hist` s jedným výstupným argumentom:

```
ni=hist(x,zi)
```

b) Analogicky ako v predchádzajúcom príklade vypočítane ešte ďalšie druhy početností a nakoniec zostavíme tabuľku rozdelenia početností:

i	$I_i = \langle t_{i-1}, t_i \rangle$	z_i	n_i	f_i	N_i	F_i
1	2500 – 4500	3500	2	1/15	2	1/15
2	4500 – 6500	5500	3	1/10	5	1/6
3	6500 – 8500	7500	10	1/3	15	1/2
4	8500 – 10500	9500	7	7/30	22	11/15
5	10500 – 12500	11500	4	2/15	26	13/15
6	12500 – 14500	13500	4	2/15	30	1
$\sum$	—	—	30	1	—	—

Vytvorenie tabuľky početností použitím programu MATLAB:

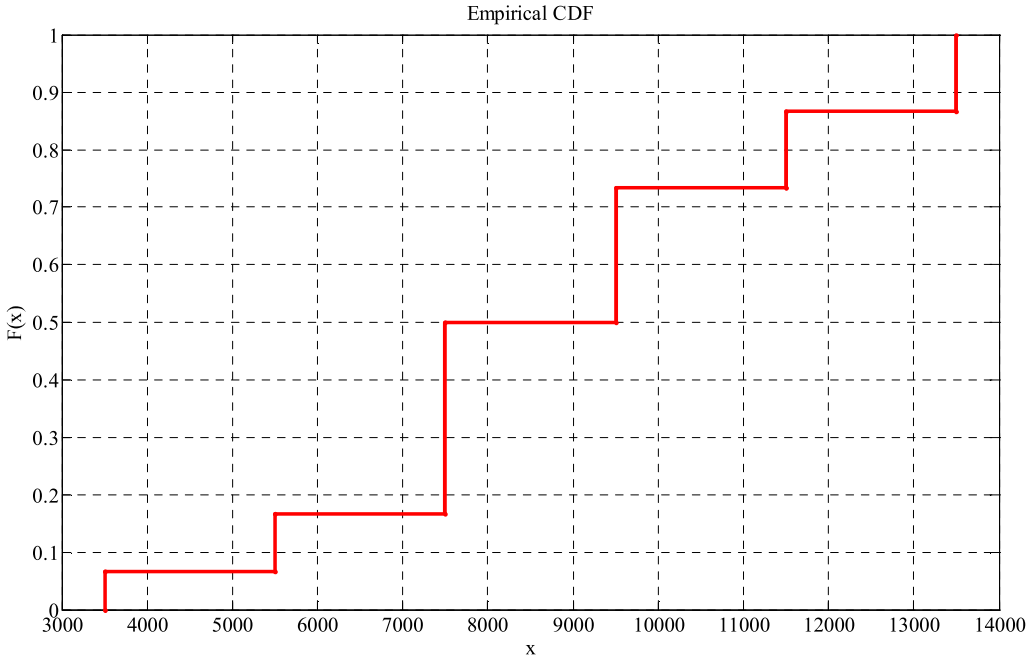
```
format rat;fi=ni/n;Ni=cumsum(ni);Fi=cumsum(fi);tab=[zi;ni;fi;Ni;Fi]'
```

c) Pre daný štatistický súbor máme zostrojiť schodovitý graf kumulatívnych relatívnych početností F_i. Budeme to realizovať pomocou M-funkcie `stairs` s použitím početností F_i:

```
stairs(zi,Fi)
```

Ďalšou možnosťou je aplikácia M-funkcie `cdfplot`:

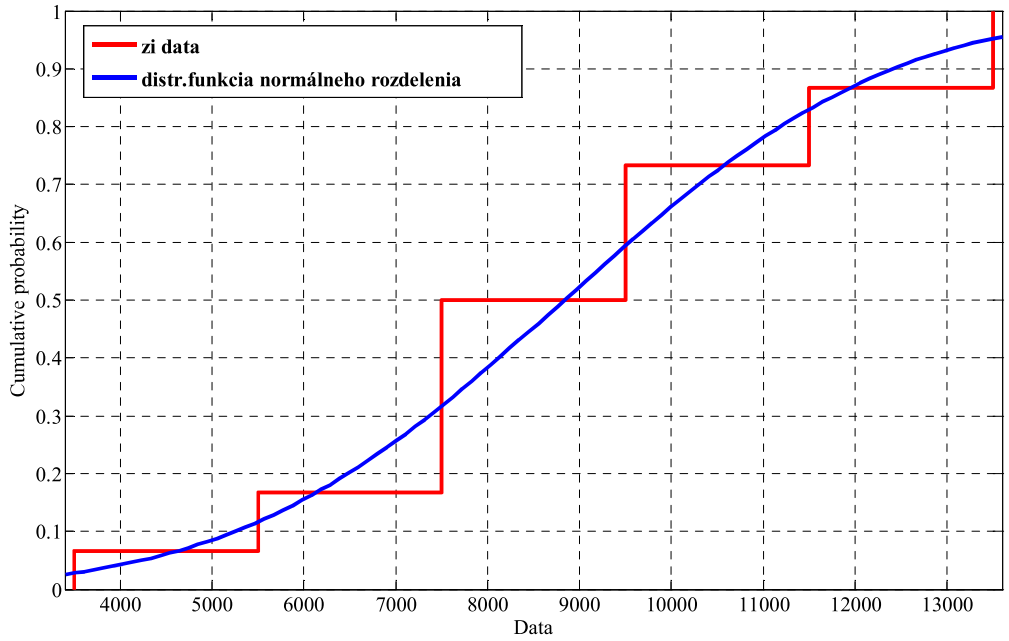
```
z=[3500*ones(1,2),5500*ones(1,3),7500*ones(1,10),9500*ones(1,7),...
11500*ones(1,4),13500*ones(1,4)];cdfplot(z)
```



Môžeme použiť aj prostredie `dfittool`, kde sa dá získať zjednodušený graf empirickej distribučnej funkcie, pričom tam môžeme dokresliť aj graf distribučnej funkcie napr. normálneho rozdelenia.
Odošleme príslušný príkaz:

```
dfittool
```

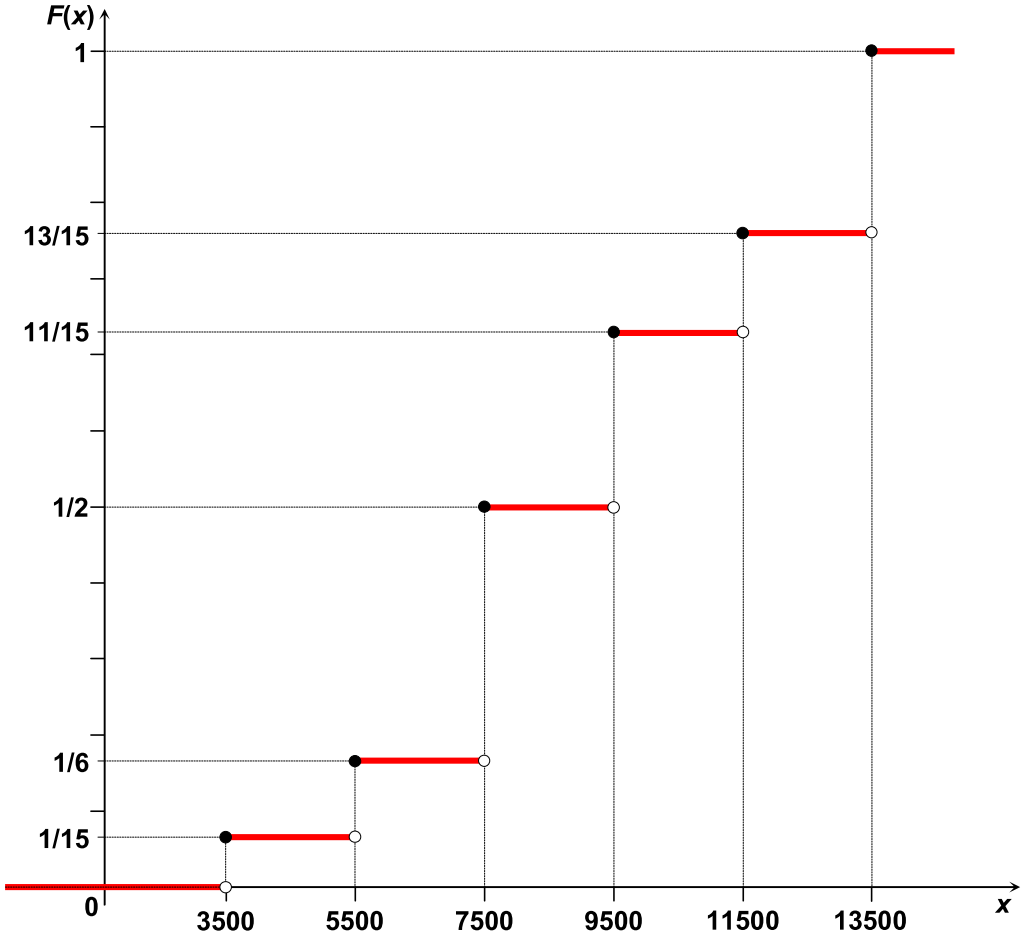
Otvorí sa grafické okno, kde vľavo hore nastavíme *CDF*, potom do *Data* dáme hodnoty z_i, do *Frequency* početnosti n_i, klikneme na *Create Data Set*, ďalej na *New Fit*, do *Distribution* dáme *Normal* a klikneme na *Apply*.



d) Teraz si pomocou početností F_i zostavíme empirickú distribučnú funkciu:

$$F(x)=\begin{cases} 0 & \text{pre} \quad -\infty < x < 3500 \\ \frac{1}{15} & \text{pre} \quad 3500 \le x < 5500 \\ \frac{1}{6} & \text{pre} \quad 5500 \le x < 7500 \\ \frac{1}{2} & \text{pre} \quad 7500 \le x < 9500 \\ \frac{11}{15} & \text{pre} \quad 9500 \le x < 11500 \\ \frac{13}{15} & \text{pre} \quad 11500 \le x < 13500 \\ 1 & \text{pre} \quad 13500 \le x < \infty \end{cases}.$$

Presný graf empirickej distribučnej funkcie bude vyzerať takto:



2.2 Opisné charakteristiky súboru s kvantitatívnym triediacim znakom

V predchádzajúcej časti sme sa dozvedeli, ako usporiadať štatistické údaje do tabuliek a sumarizovať údaje pomocou grafov. Rozdelenie početností poskytuje užitočnú informáciu o štruktúre skúmaného štatistického súboru, ale opisovať a hlavne porovnávať niekoľko súborov len pomocou tabuliek alebo grafov by bolo prácne. Z tohto dôvodu sa snažíme zhrnúť informáciu obsiahnutú v zistených údajoch o štatistickom znaku a vyjadriť ju v koncentrovanej forme pomocou určitých charakteristík.

Pri opise štatistických súborov nás zaujíma predovšetkým poloha (úroveň) rozdelenia početností a jeho variabilita (rozptýlenosť). Menej často sa zameriavame na tvar rozdelenia početností, t. j. na jeho šikmosť a špicatosť. Čísla, ktoré slúžia k opisu súborov dát, sa nazývajú ***opisné charakteristiky*** **(*miery*)**. V tejto časti sa budeme zaoberať niektorými najdôležitejšími opisnými charakteristikami.

2.2.1 Charakteristiky polohy

Charakteristiky polohy (*stredné hodnoty*) charakterizujú polohu znaku na číselnej osi.

Ich hlavnú skupinu tvoria *priemery*, a to aritmetický priemer, geometricky priemer a harmonický priemer, ktorých spoločnou vlastnosťou je, že sú určované zo všetkých nameraných hodnôt znaku.

Druhú skupinu charakteristík polohy tvoria tzv. *pozičné stredné hodnoty*, ktoré sú určené pozíciou niektorých jednotiek súboru. Medzi ne patrí modus a kvantily. Z kvantilov sa najčastejšie používa medián.

Aritmetický priemer $\bar{x}$ je definovaný ako súčet nameraných hodnôt $x_1, x_2, \ldots, x_n$ delený ich počtom n, t. j. platí vzorec $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Aritmetický priemer počítaný podľa tohto vzorca nazývame ***prostý aritmetický priemer***. Používa sa pre neroztriedený štatistický súbor.

Ak máme roztriedený súbor, pre ktorý máme spracovanú frekvenčnú tabuľku, tak na výpočet aritmetického priemeru môžeme použiť vzorec $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$, kde $n = \sum_{i=1}^{k} n_i$. Aritmetický priemer počítaný podľa tohto vzorca nazývame ***vážený aritmetický priemer***. V prípade intervalového triedenia použijeme v tomto vzorci namiesto hodnôt x_i triediaceho znaku hodnoty triedneho znaku z_i. V tomto prípade je ale potrebné si uvedomiť, že dostaneme len približnú hodnotu aritmetického priemeru, presne vypočítaného z pôvodných nameraných hodnôt podľa vzorca pre prostý aritmetický priemer.

Použitie váženého aritmetického priemeru prichádza do úvahy aj tam, kde váhy nie sú odvodené z početností, ale z relatívneho významu (dôležitosti) jednotlivých hodnôt. Napr. pri hodnotení likvidity podniku musíme počítať s tým, že jednotlivé aktíva podniku majú rôznu schopnosť využitia pre splatenie krátkodobých záväzkov. Preto sa v tejto oblasti stretávame s tým, že jednotlivým aktívam sú na základe expertného ohodnotenia priradzované váhy, ktoré určujú dôležitosť danej skupiny aktív z hľadiska likvidity podniku. Celkový (priemerný) ukazovateľ je potom váženým aritmetickým priemerom z objemov peňažných prostriedkov viazaných v jednotlivých skupinách aktív, kde ako váhy vystupujú nejaké koeficienty kvality aktív z hľadiska stupňa likvidity.

Aritmetický priemer má celý rad dôležitých vlastností, z ktorých uvedieme nasledujúce:

- Súčet odchýlok všetkých nameraných hodnôt od ich aritmetického priemeru je nulový, t. j. platí $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.
- Ak ku každej nameranej hodnote pripočítame tú istú konštantu, zväčší sa o túto konštantu aj aritmetický priemer, t. j. platí $\frac{1}{n}\sum_{i=1}^{n} (x_i + c) = \bar{x} + c$, pre $c \in R$.
- Ak všetky namerané hodnoty vynásobíme tou istou konštantou, je touto konštantou vynásobený aj aritmetický priemer, t. j. platí $\frac{1}{n}\sum_{i=1}^{n} c \cdot x_i = c \cdot \bar{x}$, pre $c \in R$.

Matematické vyjadrenie aritmetického priemeru je jednoduché a používa sa v mnohých ďalších odvodeniach dôležitých vzťahov. Jeho výpočet je založený na všetkých nameraných hodnotách. Aritmetický priemer je citlivý voči extrémnym hodnotám súboru. Má zmysel ako veľmi dôležitá charakteristika daného súboru vtedy, ak sú odchýlky nameraných hodnôt náhodné a v súbore sa nevyskytujú extrémne malé alebo extrémne veľké hodnoty.

Príklad 2.3. Vypočítajte aritmetický priemer pre štatistický súbor z príkladu 2.2.

Riešenie: Výpočet aritmetického priemeru pomocou kalkulačky podľa uvedených vzorcov je zrejme jasný. Potrebné údaje máme uvedené v príklade 2.2. Z neroztriedených hodnôt ho

môžeme vypočítať ako prostý aritmetický priemer podľa vzorca $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Dostaneme výsledok $\bar{x} = 8681{,}1$.

Výpočet z roztriedených hodnôt ako vážený aritmetický priemer realizujeme podľa vzorca $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} z_i \cdot n_i$. Tu dostaneme ale len približnú hodnotu $\bar{x} = 8833,\overline{3}$.

Výpočet z neroztriedených hodnôt podľa vzorca pre prostý aritmetický priemer môžeme v MATLABe realizovať takto:

```
xi=[9966,5105,13911,7080,6640,8003,6905,12999,11110,3060,10401,...
2880,6732,12625,12012,11763,9968,8720,9043,10509,8826,7740,4970,...
7260,10395,6830,8390,14165,5165,7260];n=length(xi);ap=sum(xi)/n
```

Výpočet aritmetického priemeru pomocou M-funkcie *mean*:

```
ap=mean(xi)
```

Výpočet z roztriedených hodnôt, t. j. z tabuľky intervalového rozdelenia početností, môžeme v MATLABe realizovať podľa vzorca pre vážený aritmetický priemer nasledujúcim spôsobom:

```
zi=3500:2000:13500;ni=[2,3,10,7,4,4];n=sum(ni);ap=sum(zi.*ni)/n
```

Výpočet aritmetického priemeru pomocou M-funkcie *mean*:

```
z=[3500*ones(1,2),5500*ones(1,3),7500*ones(1,10),9500*ones(1,7),...
11500*ones(1,4),13500*ones(1,4)],format long g,ap=mean(z)
```

K ďalším charakteristikám polohy patrí geometrický a harmonický priemer, ktoré sa však nepoužívajú tak často ako aritmetický priemer.

Geometrický priemer $\bar{x}_g$ z n kladných hodnôt $x_1, x_2, \ldots, x_n$ je definovaný ako n-tá odmocnina z ich súčinu, t. j. platí vzorec $\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$. Geometrický priemer počítaný podľa tohto vzorca nazývame ***prostý geometrický priemer***. Používa sa pre neroztriedený štatistický súbor.

Ak máme roztriedený súbor, pre ktorý máme spracovanú frekvenčnú tabuľku, tak pre výpočet geometrického priemeru môžeme použiť vzorec $\bar{x}_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \cdots \cdot x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$, kde $n = \sum_{i=1}^{k} n_i$. Geometrický priemer počítaný podľa tohto vzorca nazývame ***vážený geometrický priemer***. V prípade intervalového triedenia použijeme v tomto vzorci namiesto hodnôt x_i triediaceho znaku hodnoty triedneho znaku z_i.

Geometrický priemer sa používa napr. na výpočet priemernej inflácie, priemerného úroku, priemerného percentuálneho rastu medzi dvomi časovými obdobiami a pod.

Príklad 2.4. O zmenách objemu produkcie v jednotlivých mesiacoch prvého polroka oproti predchádzajúcemu mesiacu informujú nasledujúce koeficienty rastu: 1,05; 1,1; 1,22; 1,30; 1,17; 0,98. Vypočítajte priemerný koeficient rastu.

Riešenie: Priemerný koeficient rastu vypočítame pomocou vzorca pre prostý geometrický priemer. Je to šiesta odmocnina zo súčinu uvedených čísel. Výsledok je 1,1317.

Aplikácia jednoduchým prepisom uvedeného vzorca do MATLABu:

```
xi=[1.05,1.1,1.22,1.3,1.17,.98];gp=(prod(xi))^(1/6)
```

Aplikácia pomocou M-funkcie *geomean*:

```
gp=geomean(xi)
```

Harmonický priemer $\bar{x}_h$ z *n* nenulových hodnôt $x_1, x_2, \ldots, x_n$ je definovaný ako podiel ich počtu *n* a súčtu ich prevrátených hodnôt. Vzorec pre ***prostý harmonický priemer*** má tvar $\bar{x}_h = \dfrac{n}{\sum_{i=1}^{n} \dfrac{1}{x_i}}$. Analogicky ako v prípade aritmetického a geometrického priemeru môžeme zapísať vzorec pre ***vážený harmonický priemer*** v tvare $\bar{x}_h = \dfrac{n}{\sum_{i=1}^{k} \dfrac{n_i}{x_i}}$.

Zo vzorca $\bar{x}_h = \dfrac{n}{\sum_{i=1}^{n} \dfrac{1}{x_i}}$ je zrejmé, že prevrátená hodnota harmonického priemeru je aritmetickým priemerom prevrátených hodnôt nameraných hodnôt $x_1, x_2, \ldots, x_n$, t. j. platí $\dfrac{1}{\bar{x}_h} = \dfrac{\sum_{i=1}^{n} \dfrac{1}{x_i}}{n}$.

Harmonický priemer sa používa predovšetkým pri výpočte priemernej rýchlosti, priemerného času a pod.

Pre kladné hodnoty $x_1, x_2, \ldots, x_n$ platia medzi uvedenými priemermi týchto hodnôt nerovnosti: $\bar{x}_h \le \bar{x}_g \le \bar{x}$.

Príklad 2.5. Máme 3 stroje, ktoré vyrábajú skrutky. Prvý stroj potrebuje na výrobu jednej skrutky 2,5 minúty, druhý 2 minúty a tretí 3,2 minúty. Vypočítajte priemerný čas potrebný na výrobu jednej skrutky, ak stroje budú pracovať súčasne.

Riešenie: Počítame podľa vzorca pre prostý harmonický priemer. Dostaneme $\bar{x}_h = \dfrac{3}{\dfrac{1}{2,5} + \dfrac{1}{2} + \dfrac{1}{3,2}} = 2,4742$ [min.].

Aplikácia jednoduchým prepisom uvedeného vzorca do MATLABu:

```
xi=[2.5,2,3.2];hp=3/sum(1./xi)
```

Aplikácia pomocou M-funkcie *harmmean*:

```
hp=harmmean(xi)
```

Napr. pri skúmaní miezd pracovníkov určitého priemyselného odvetvia nás budú okrem iného zaujímať minimálne a maximálne mzdy týchto pracovníkov. Je zrejmé, že jednoduché zistenie minimálnej a maximálnej mzdy pracovníkov skúmaného súboru nemá veľký význam. Sú to extrémne hodnoty a nie sú podstatné z hľadiska danej otázky. Môžu byť celkom náhodné. Pri úvahách o minimálnych mzdách nás bude skôr zaujímať napr. horná hranica miezd 1 % pracovníkov daného odvetvia s najnižšími mzdami a pri úvahách o vysokých mzdách nás bude zaujímať napr. dolná hranica miezd 10 % pracovníkov s najvyššími mzdami. Inokedy nás bude zaujímať horná hranica medzi 50 % pracovníkov s nízkymi mzdami, ktorá je súčasne dolnou hranicou 50 % pracovníkov s vysokými mzdami. Charakteristiky, ktoré nám podávajú informáciu tohto druhu, nazývame kvantily.

Kvantil súboru dát je hodnota, ktorá rozdeľuje usporiadaný súbor hodnôt určitého kvantitatívneho štatistického znaku na dve časti – jedna obsahuje tie hodnoty, ktoré sú menšie a rovnaké ako kvantil (napr. 1, 15, 50, 90 %), druhá časť naopak obsahuje tie hodnoty, ktoré sú väčšie a rovnaké ako kvantil (t. j. napr. 99, 85, 50, 10 %).

Nech $0 < p < 1$. Kvantilom $\tilde{x}_{100p}$ štatistického znaku X nazývame číslo zvolené tak, aby $100 \cdot p$ % štatistických jednotiek malo hodnotu znaku menšiu (alebo rovnakú) ako toto číslo, t. j. aby $100 \cdot (1-p)$ % štatistických jednotiek malo hodnotu znaku väčšiu (alebo rovnakú) ako toto číslo. Kvantil $\tilde{x}_{100p}$ teda oddeľuje $100 \cdot p$ % malých hodnôt od $100 \cdot (1-p)$ % veľkých hodnôt štatistického znaku a nazývame ho **$100 \cdot p$ %-ný kvantil**.

Medzi najčastejšie používané kvantily patria percentily, decily a kvartily.

Percentily $\tilde{x}_1, \ldots, \tilde{x}_{99}$ rozdeľujú usporiadaný štatistický súbor na 100 rovnako početných častí.

Decily $\tilde{x}_{10}, \ldots, \tilde{x}_{90}$ rozdeľujú usporiadaný štatistický súbor na 10 rovnako početných častí.

Kvartily sú hodnoty, ktoré delia usporiadaný štatistický súbor na štyri časti, pričom každá obsahuje 25 % štatistických jednotiek. Sú dokopy tri.

Prvý* (*dolný*) *kvartil $\tilde{x}_{25}$ oddeľuje zhruba 25 % najmenších hodnôt štatistického znaku od ostatných.

Druhý* (*prostredný*) *kvartil, tzv. ***medián*** $\tilde{x}_{50}$, rozdeľuje usporiadaný súbor hodnôt štatistického znaku na dve čo do počtu rovnaké časti, t. j. každá z nich obsahuje 50 % štatistických jednotiek. Medián označujeme aj $\tilde{x}$.

Tretí* (*horný*) *kvartil $\tilde{x}_{75}$ je taká hodnota štatistického znaku, ktorá oddeľuje zhruba 75 % najmenších hodnôt štatistického znaku od ostatných 25 %.

Medián sa rovná druhému kvartilu, piatemu decilu alebo päťdesiatemu percentilu. Prvý kvartil sa rovná dvadsiatemu piatemu percentilu a tretí kvartil sa rovná sedemdesiatemu piatemu percentilu. Napr. druhý decil sa rovná dvadsiatemu percentilu.

Kvantilom môže byť priamo hodnota znaku určitej štatistickej jednotky sledovaného súboru alebo číslo, ktoré leží medzi hodnotami dvoch susedných štatistických jednotiek vo vzostupne usporiadanej postupnosti podľa hodnôt daného znaku, t. j. vo variačnom rade. V druhom spomenutom prípade sa kvantil stanoví pomocou lineárnej interpolácie týchto susedných hodnôt znaku.

Určiť kvantily v prípade, ak sú k dispozícii údaje o všetkých n jednotkách štatistického súboru, predpokladá usporiadanie daného súboru do *variačného radu* $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Ďalej je nutné určiť *poradové číslo* jednotky, ktorej hodnota je hľadaným kvantilom, resp. poradové čísla jednotiek, z hodnôt ktorých sa tento kvantil určí.

100·p %-ný kvantil určíme pre $0 < p < 1$ podľa nasledujúceho vzorca

$$\tilde{x}_{100p} = \begin{cases} x_{([np]+1)}, & \text{pokiaľ } np \text{ nie je celé číslo} \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & \text{pre } np \text{ celé} \end{cases},$$

kde $[np]$ je celá časť čísla, t. j. najbližšie menšie celé číslo.

Špeciálne pre medián platí vzorec $\tilde{x}_{50} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{pre } n \text{ nepárne} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{pre } n \text{ párne} \end{cases}.$

Ak máme pre stanovenie kvantilov k dispozícii len intervalové rozdelenie početností znaku X, potom môžeme získať len približné hodnoty daného kvantilu.

Medián môžeme v tomto prípade počítať podľa vzorca, ktorý vznikne lineárnou interpoláciou v histograme kumulatívnych absolútnych početností. Platí $\tilde{x}_{50} = a + h \cdot \frac{\frac{n}{2} - N_{i-1}}{n_i}$, kde a je začiatok triedneho intervalu obsahujúceho medián (mediánového intervalu), n_i je absolútna početnosť intervalu obsahujúceho medián, N_{i-1} je kumulatívna absolútna početnosť predchádzajúceho intervalu pred intervalom obsahujúcim medián, h je dĺžka triedneho intervalu, n je rozsah súboru.

Pre prvý kvartil použijeme v prípade intervalového rozdelenia početností vzorec $\tilde{x}_{25} = a + h \cdot \frac{\frac{n}{4} - N_{i-1}}{n_i}$, kde a je začiatok triedneho intervalu obsahujúceho prvý kvartil, n_i je absolútna početnosť intervalu obsahujúceho prvý kvartil, N_{i-1} je kumulatívna absolútna početnosť predchádzajúceho intervalu pred intervalom obsahujúcim prvý kvartil, h je dĺžka triedneho intervalu, n je rozsah súboru.

Pre tretí kvartil použijeme v prípade intervalového rozdelenia početností vzorec $\tilde{x}_{75} = a + h \cdot \frac{\frac{3n}{4} - N_{i-1}}{n_i}$, kde a je začiatok triedneho intervalu obsahujúceho tretí kvartil, n_i je absolútna početnosť intervalu obsahujúceho tretí kvartil, N_{i-1} je kumulatívna absolútna početnosť predchádzajúceho intervalu pred intervalom obsahujúcim tretí kvartil, h je dĺžka triedneho intervalu, n je rozsah súboru.

Jednou z foriem prezentácie štatistického súboru je ***krabicový diagram***, tzv. ***boxplot***. Je založený na pätici čísel, a to $x_{\min}$, $\tilde{x}_{25}$, $\tilde{x}_{50}$, $\tilde{x}_{75}$, $x_{\max}$, ktorým hovoríme aj *päťčíselná charakteristika*. V uvedenom diagrame dolná a horná strana základného obdĺžnika (krabice) zodpovedajú prvému a tretiemu kvartilu, vodorovná čiara vnútri tohto obdĺžnika zodpovedá mediánu, t. j. druhému kvartilu. Uvedené tri vodorovné úsečky teda delia súbor nameraných a podľa veľkosti usporiadaných hodnôt na štyri zhruba rovnako početné časti. Výška obdĺžnika sa nazýva *kvartilové rozpätie* $R_Q = \tilde{x}_{75} - \tilde{x}_{25}$. Dolná zvislá úsečka zodpovedá hodnotám, ktoré ležia pod obdĺžnikom vo vzdialenosti najviac rovnom 1,5-násobku výšky obdĺžnika. Koniec tejto úsečky zodpovedá najmenšej takej hodnote zo súboru. Podobne je to u hornej úsečky. Teda horná a dolná úsečka zodpovedajú tým hodnotám, ktoré nie sú medzi

kvartilmi a sú od nich vzdialené najviac o 1,5-násobok kvartilového rozpätia. Mimo týchto úsečiek (pod nimi a nad nimi) sa znázorňujú body zodpovedajúce prípadným tzv. *extrémnym hodnotám*. Boxplot slúži aj ako *grafický test extrémnych hodnôt*.

Príklad 2.6. 20 vybraných televíznych divákov bolo požiadaných, aby si celý týždeň zaznamenávali čas [hod.], ktorý venovali sledovaniu televíznych programov. Nasleduje získaný variačný rad: 5, 15, 16, 20, 21, 25, 26, 27, 30, 30, 31, 32, 32, 34, 35, 38, 38, 41, 43, 66. Vypočítajte hodnoty kvartilov daného štatistického súboru.

Riešenie: Namerané hodnoty sú už usporiadané do variačného radu. Rozsah je $n = 20$, čo je párne číslo. Platí $\frac{n}{2} = 10$ a $\frac{n}{2} + 1 = 11$, teda medián (druhý kvartil) určíme ako aritmetický priemer hodnoty s poradovým číslom 10 a hodnoty s poradovým číslom 11 v uvedenom variačnom rade. Pre medián teda platí $\tilde{x}_{50} = \frac{30+31}{2} = 30,5$.

To isté dostaneme aj podľa všeobecného vzorca pre $100p$ %-ný kvantil, kde v prípade mediánu dosadíme $p = 0,5$.

Teraz stanovíme prvý kvartil. Platí $np = 20 \cdot 0,25 = 5$, čo je celé číslo. Teda prvý kvartil určíme ako aritmetický priemer hodnoty s poradovým číslom 5 a hodnoty s poradovým číslom 6. Potom prvý kvartil bude $\tilde{x}_{25} = \frac{21+25}{2} = 23$.

Ešte treba určiť tretí kvartil. Platí $np = 20 \cdot 0,75 = 15$, čo je celé číslo. Teda tretí kvartil určíme ako aritmetický priemer hodnoty s poradovým číslom 15 a hodnoty s poradovým číslom 16. Potom tretí kvartil bude $\tilde{x}_{75} = \frac{35+38}{2} = 36,5$.

Výpočet v programe MATLAB pomocou M-funkcie `prctile`:

```
vr=[5,15,16,20,21,25,26,27,30,30,31,32,32,34,35,38,38,41,43,66];
med=prctile(vr,50),q1=prctile(vr,25),q3=prctile(vr,75)
```

Medián môžeme určiť aj pomocou M-funkcie `median`:

```
med=median(vr)
```

Príklad 2.7. Je daná frekvenčná tabuľka veku študentov (x_i) obývajúcich internát:

x_i	18	19	20	21	22	23	24	25	26
n_i	1	23	25	13	10	8	4	2	4

Vypočítajte hodnoty kvartilov daného štatistického súboru a znázornite boxplot.

Riešenie: Daný súbor je usporiadaný v tabuľke rozdelenia absolútnych početností. Platí $n = 90$, $\frac{n}{2} = 45$ a $\frac{n}{2} + 1 = 46$, a teda medián určíme ako aritmetický priemer hodnôt s poradovými číslami 45 a 46. Z tabuľky absolútnych početností je vidieť, že obe tieto hodnoty sú rovné 20 (všetky hodnoty s poradovými číslami 25 až 49 sú rovné 20). Medián (druhý kvartil) je teda $\tilde{x}_{50} = 20$.

Prvý a tretí kvartil veku študentov budú $\tilde{x}_{25} = 19$ a $\tilde{x}_{75} = 22$, pretože všetky hodnoty s poradovými číslami 2 až 24 sú rovné 19, rovnako ako aj všetky hodnoty s poradovými číslami 63 až 72 sú rovné 22.

Môžeme aplikovať aj MATLAB. Napr. si zostavíme variačný rad a pre zaujímavosť si skontrolujeme aj poradové čísla predtým spomenutých hodnôt:

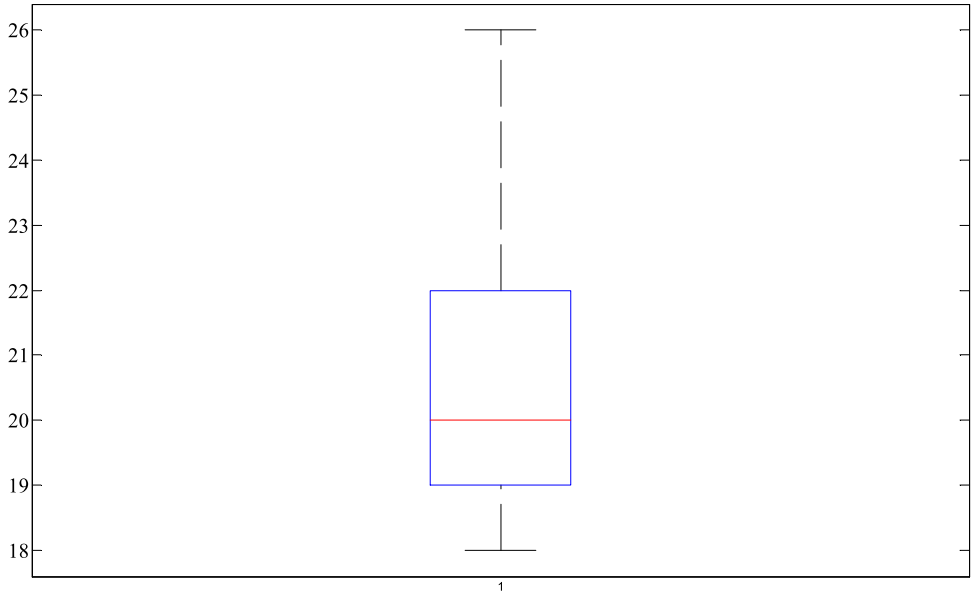
```
vr=[18,19*ones(1,23),20*ones(1,25),21*ones(1,13),22*ones(1,10),...
23*ones(1,8),24*ones(1,4),25,25,26*ones(1,4)],n=length(vr)
vr(25:49),vr(2:24),vr(63:72)
```

Na výpočet kvartilov použijeme už v predchádzajúcom príklade spomenutú M-funkciu *prctile*. Môžeme počítať aj viacero percentilov naraz:

```
prctile(vr,[25,50,75])
```

Pre znázornenie krabicového diagramu použijeme M-funkciu *boxplot*:

```
boxplot(vr)
```



Vidíme, že daný štatistický súbor neobsahuje extrémne hodnoty.

Príklad 2.8. Je daná tabuľka intervalového rozdelenia početností:

I_i	9 – 11	11 – 13	13 – 15	15 – 17	17 – 19	19 – 21
n_i	5	15	25	40	10	5

Pre uvedený štatistický súbor znázornite stĺpcový graf kumulatívnych absolútnych početností a vypočítajte hodnoty jeho kvartilov.

Riešenie: Použijeme opäť MATLAB:

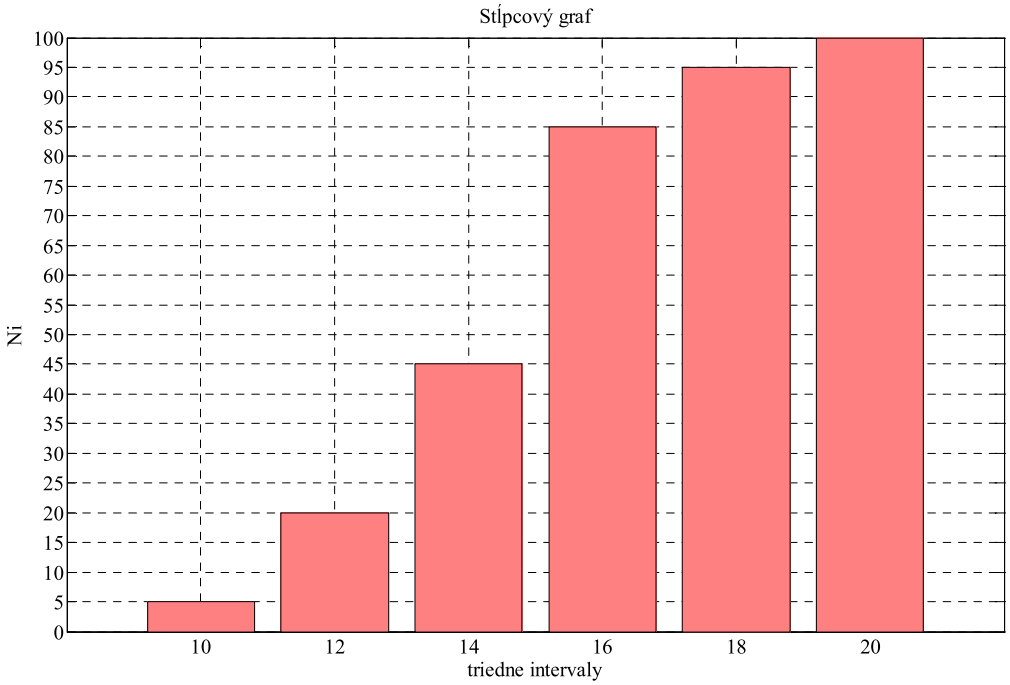
```
zi=10:2:20;ni=[5,15,25,40,10,5];n=sum(ni),Ni=cumsum(ni)
```

Pre rozsah štatistického súboru dostaneme výsledok $n = 100$. Pre kumulatívne absolútne početnosti bude výstupom z MATLABu riadkový vektor:

Ni = 5 20 45 85 95 100

Použitie MATLABu na grafické znázornenie:

```
bar(zi,Ni),colormap([1,0.5,0.5]),grid,
xlabel('triedne intervaly'),ylabel('Ni'),
title('Stĺpcový graf')
```



Aplikáciou MATLABu pomocou M-funkcie `prctile` dostaneme v prípade intervalového triedenia len triedny znak prislúchajúci intervalu obsahujúceho príslušný kvartil:

```
z=[10*ones(1,5),12*ones(1,15),14*ones(1,25),16*ones(1,40),...
18*ones(1,10),20*ones(1,5)];kvartily=prctile(z,[25,50,75])
```

Výstup z MATLABu:
kvartily = 14 16 16

Tento výsledok nám v spojení so vstupnou tabuľkou hovorí, že prvý kvartil sa nachádza v treťom triednom intervale, druhý kvartil (medián) a tretí kvartil sa nachádzajú v štvrtom triednom intervale.
To isté však môžeme zistiť aj zo znázorneného stĺpcového grafu kumulatívnych absolútnych početností. Keďže je $n = 100$, číslo $n/4 = 25$ (viď na osi y príslušnú kumulatívnu početnosť N_i) prislúcha tretiemu triednemu intervalu, číslo $n/2 = 50$ prislúcha štvrtému triednemu intervalu a číslo $3n/4 = 75$ prislúcha tiež štvrtému triednemu intervalu.

Teraz si spresníme hodnoty kvartilov výpočtom podľa vzorcov uvedených v teoretickej časti.

Pre prvý kvartil platí $\tilde{x}_{25} = a + h \cdot \frac{\frac{n}{4} - N_{i-1}}{n_i}$. Po dosadení do vzorca a vyčíslení dostaneme:

$$\tilde{x}_{25} = 13 + 2 \cdot \frac{25 - 20}{25} = 13{,}4 .$$

Pre medián platí: $\tilde{x}_{50} = a + h \cdot \frac{\frac{n}{2} - N_{i-1}}{n_i} = 15 + 2 \cdot \frac{50 - 45}{40} = 15{,}25 .$

Pre tretí kvartil dostaneme: $\tilde{x}_{75} = a + h \cdot \frac{\frac{3n}{4} - N_{i-1}}{n_i} = 15 + 2 \cdot \frac{75 - 45}{40} = 16{,}5 .$

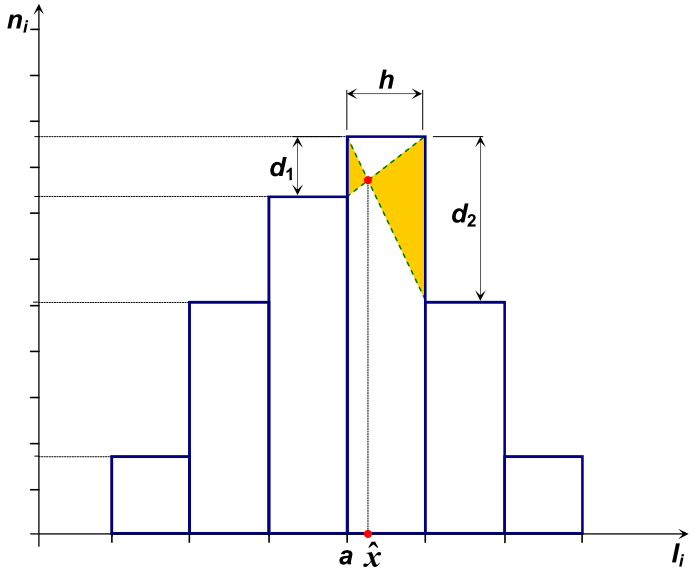
Poslednou charakteristikou polohy, ktorou sa budeme zaoberať, je modus. Predstavuje hodnotu, ktorá je v rámci skúmaného štatistického súboru najtypickejšou. Považujeme ho za dôležitú doplnkovú charakteristiku k najčastejšie používaným charakteristikám polohy, a to k aritmetickému priemeru a mediánu. Podobne ako medián, ani modus nie je ovplyvnený extrémnymi hodnotami.

Modus $\hat{x}$ súboru dát je každá hodnota, ktorej početnosť výskytu je väčšia ako 1 a je rovnaká alebo väčšia ako početnosť výskytu akejkoľvek inej hodnoty.

Ak početnosť žiadnej hodnoty súboru dát nie je väčšia ako 1, potom hovoríme, že daný súbor nemá modus. Inak povedané, každá hodnota, ktorá má najväčšiu početnosť sa nazýva modus. Súbor hodnôt môže mať teda viac ako jeden modus.

V prípade intervalového rozdelenia početností sa pri stanovení modusu uspokojujeme buď s určením modálneho (najpočetnejšieho) intervalu, alebo v rámci tohto intervalu modus odhadujeme, napr. stredom intervalu. Existujú však aj presnejšie postupy, ktoré vychádzajú z rekonštrukcie vrcholu súboru podľa rozdelenia početností v okolí modálneho intervalu.

Ukážeme postup, pri ktorom sa vychádza z histogramu absolútnych početností a z predpokladu, že namerané hodnoty sa v modálnom intervale viac koncentrujú k tej hranici, ktorej susedný interval má väčšiu početnosť, ako je to znázornené na nasledujúcom obrázku:



Z podobnosti trojuholníkov vyplýva, že platí $\frac{\hat{x}-a}{d_1}=\frac{(a+h)-\hat{x}}{d_2}$. Odtiaľ dostaneme vzorec pre výpočet modusu v tvare $\hat{x}=a+h\cdot\frac{d_1}{d_1+d_2}$, kde a je začiatok modálneho intervalu, d_1 je rozdiel medzi absolútnou početnosťou modálneho a predchádzajúceho intervalu, d_2 je rozdiel medzi absolútnou početnosťou modálneho a nasledujúceho intervalu, h je dĺžka triedneho intervalu.

Príklad 2.9. Vo výrobni bolo za jednu hodinu vyrobených 15 oceľových tyčí. Pri meraní ich dĺžky boli zistené tieto výsledky [cm]: 20,48; 20,62; 21,21; 20,00; 18,22; 18,65; 18,89; 19,50; 21,54; 18,22; 18,19; 20,47; 19,19; 20,62; 18,48. Určte modus daného štatistického súboru.

Riešenie: Aplikáciou MATLABu si vytvoríme frekvenčnú tabuľku:

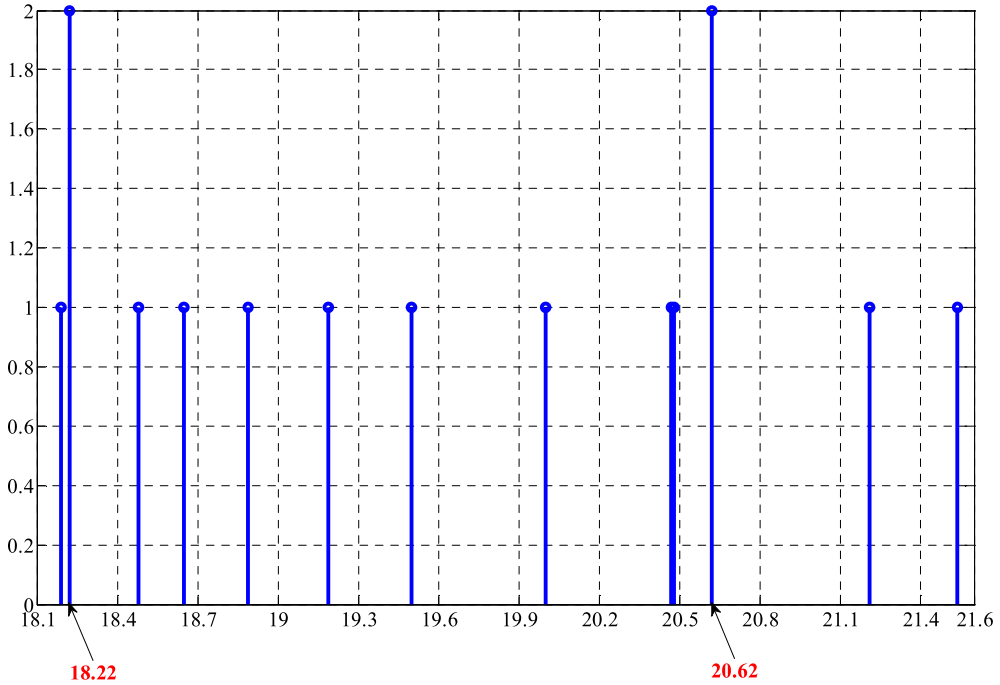
```
x=[2048,2062,2121,2000,1822,1865,1889,1950,2154,1822,1819,2047,1919,
2062,1848]/100; xi=unique(x); ni=hist(x,xi);fr_tab=[xi;ni]
```

x_i	18,19	18,22	18,48	18,65	18,89	19,19	19,50	20,00	20,47	20,48	20,62	21,21	21,54
n_i	1	2	1	1	1	1	1	1	1	1	2	1	1

Z uvedenej frekvenčnej tabuľky je vidieť, že daný štatistický súbor má 2 modusy. Sú nimi hodnoty 18,22 a 20,62.

Pomocou MATLABu si môžeme aj graficky zobraziť túto frekvenčnú tabuľku. Použijeme napr. úsečkový graf, kde je opäť dobre vidieť hodnoty s rovnakou najväčšou početnosťou:

```
stem(xi,ni),grid
```



Poznámka: V grafickom okne si môžeme robiť ďalšie úpravy, napr. v hornej časti lišty cez *Edit* a *Axes Properties*. Pridať šípky s hodnotou môžeme cez *Insert* a *Text Arrow*.

Aplikácia MATLABu pomocou M-funkcie `mode` s 1 výstupným argumentom:

```
modus=mode(x)
```

Výstup z MATLABu: modus = 18.2200

Poznámka: M-funkcia `mode` s 1 výstupným argumentom vyberá len jednu hodnotu modusu. V tomto prípade je preto pri neroztriedených štatistických súboroch nutné sledovať pozorne hodnoty s rovnakou najväčšou početnosťou buď prostredníctvom variačného radu, resp. frekvenčnej tabuľky. Ďalšou možnosťou je aplikácia príkazu `[m,f,c]=mode(x)` s 3 výstupnými argumentmi. Pozrite si to podrobne prostredníctvom `help mode`. V premennej *c* (vo *Workspace*) nájdeme všetky hodnoty modusu.

Príklad 2.10. Určte modus štatistického súboru z príkladu 2.8.

Riešenie: Z danej tabuľky intervalového rozdelenia početností je zrejmé, že modus sa nachádza v štvrtom triednom intervale s triednym znakom 16. Hodnotu modusu spresníme aplikáciou vzorca $\hat{x} = a + h \cdot \frac{d_1}{d_1 + d_2}$. Dosadíme hodnoty: $a = 15$, $h = 2$, $d_1 = 40 - 25 = 15$, $d_2 = 40 - 10 = 30$. Teda modusom je hodnota $\hat{x} = 15 + 2 \cdot \frac{15}{15 + 30} = 15,\overline{6}$.

2.2.2 Charakteristiky variability

Variabilitou (*rozptylom*) kvantitatívneho štatistického znaku rozumieme kolísanie hodnôt tejto veličiny. Pokiaľ sú všetky hodnoty súboru rovnaké (x_i = konštanta), hovoríme o nulovej variabilite. Kolísanie hodnôt v súbore môžeme posudzovať buď ako vzájomnú rozdielnosť jednotlivých hodnôt sledovanej veličiny alebo ako rozdielnosť jednotlivých hodnôt od aritmetického priemeru (prípadne mediánu). Tento druhý princíp merania variability prevláda.

Meranie variability je možné využiť na hodnotenie rovnorodosti (homogenity) súboru a tiež na posudzovanie kvality informácie, ktorú o úrovni hodnôt v súbore poskytla niektorá zo stredných hodnôt. Vychádzame pritom z úvahy, že čím je súbor homogénnejší, s menšou variabilitou, tým je napr. aritmetický priemer výstižnejší z hľadiska hodnotenia úrovne hodnôt súboru.

K základným charakteristikám variability patria: rozptyl (disperzia), smerodajná odchýlka, priemerná odchýlka, variačné rozpätie, kvartilové rozpätie a variačný koeficient.

Najznámejšou a najpoužívanejšou mierou variability je ***rozptyl*** S^2. Je definovaný ako aritmetický priemer zo štvorcov odchýlok jednotlivých nameraných hodnôt $x_1, x_2, \ldots, x_n$ od ich aritmetického priemeru $\bar{x}$, t. j. platí vzorec $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Tento tzv. *prostý* tvar vzorca sa používa pre neroztriedený štatistický súbor.

Pre roztriedený súbor daný tabuľkou absolútnych početností používame vzorec v tzv. *váženom* tvare $S^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i$, kde $n = \sum_{i=1}^{k} n_i$. V prípade intervalového triedenia použijeme v tomto vzorci namiesto hodnôt x_i triediaceho znaku hodnoty triedneho znaku z_i.

Čím väčšia je variabilita súboru, tým väčší je aj jeho rozptyl. Rozptyl nie je rezistentný, jeho hodnota môže byť silne ovplyvnená niekoľkými extrémnymi hodnotami.

Rozptyl sám o sebe nie je interpretovateľnou veličinou, pretože výsledok je daný v štvorcoch merných jednotiek. Napr. ak sú hodnoty súboru merané v cm, ich rozptyl je udávaný v cm^2. Preto sa pri hodnotení variability dáva prednosť druhej odmocnine rozptylu, tzv. ***smerodajnej odchýlke*** S (uvažovanej s kladným znamienkom), t. j. platí $S = \sqrt{S^2}$. Smerodajná odchýlka je už udaná v rovnakých jednotkách ako merané hodnoty a je to (zhruba povedané) priemerná odchýlka nameraných hodnôt od aritmetického priemeru.

Ďalšou mierou rozptylu je ***priemerná odchýlka***, ktorá je definovaná ako aritmetický priemer z absolútnych hodnôt odchýlok všetkých nameraných hodnôt od aritmetického priemeru, t. j. platí vzorec $\bar{d} = \frac{1}{n}\sum_{i=1}^{n}\left| x_i - \bar{x} \right|$. Tento tzv. *prostý* tvar vzorca sa používa pre neroztriedený štatistický súbor.

Pre roztriedený súbor daný tabuľkou absolútnych početností použijeme vzorec v tzv. *váženom* tvare $\bar{d} = \frac{1}{n}\sum_{i=1}^{k}\left| x_i - \bar{x} \right| \cdot n_i$, kde $n = \sum_{i=1}^{k} n_i$. V prípade intervalového triedenia použijeme v tomto vzorci namiesto hodnôt x_i triediaceho znaku hodnoty triedneho znaku z_i.

Priemerná odchýlka nedáva veľmi spoľahlivé výsledky rozptylu. V modernej štatistike sa používa len zriedka a nahradzuje ju smerodajná odchýlka.

Variačné rozpätie R_V je definované ako rozdiel medzi najväčšou a najmenšou hodnotou v danom súbore nameraných hodnôt, t. j. platí vzorec $R_V = x_{\max} - x_{\min}$. Predstavuje rýchlu, ale len orientačnú charakteristiku variability. Jeho nevýhodou je jeho závislosť od extrémnych

hodnôt a tiež aj to, že neposkytuje informáciu o variabilite hodnôt medzi najväčšou a najmenšou hodnotou súboru. Z tohto dôvodu sa častejšie používa ***kvartilové rozpätie*** R_Q, ktoré je definované ako rozdiel medzi horným a dolným kvartilom, t. j. platí vzorec $R_Q = \tilde{x}_{75} - \tilde{x}_{25}$. Zhruba povedané, R_Q udáva rozpätie stredných 50 % hodnôt. Jeho nevýhodou (podobne ako u variačného rozpätia) je to, že nevyužíva všetky číselné hodnoty, ale len ich poradie.

Pri porovnávaní variability viacerých súborov narážame na problém rozdielnych merných jednotiek a rozdielnej úrovne hodnôt v súboroch. V týchto prípadoch je z hľadiska potreby porovnania najvhodnejšou charakteristikou variability ***variačný koeficient*** V_k, ktorý je definovaný ako pomer smerodajnej odchýlky a aritmetického priemeru, t. j. platí vzorec $V_k = \frac{S}{\bar{x}}$. Ľahká interpretácia hodnôt variačného koeficientu ho radí medzi najpoužívanejšie charakteristiky variability. Variačný koeficient určuje akou časťou sa smerodajná odchýlka podieľa na aritmetickom priemere dát. Patrí medzi relatívne miery variability, pretože nevyjadruje variabilitu v pôvodných merných jednotkách. Je to bezrozmerné číslo. Obvykle sa variačný koeficient vyjadruje v percentách. Podľa veľmi hrubého pravidla, ak je variačný koeficient väčší ako 0,5, tak je to prejavom značnej nerovnorodosti štatistického súboru.

Príklad 2.11. Bolo vykonaných 32 chemických analýz na overenie koncentrácie x_i [%] chemickej látky v roztoku. Výsledky sú dané frekvenčnou tabuľkou:

x_i	9	11	12	14	15	16	17	18	20	21
n_i	1	2	4	4	6	5	4	3	2	1

Pre daný štatistický súbor vypočítajte: a) rozptyl, b) smerodajnú odchýlku, c) priemernú odchýlku, d) variačné rozpätie, e) kvartilové rozpätie, f) variačný koeficient.

Riešenie: a) Najprv vypočítame aritmetický priemer pomocou vzorca vo váženom tvare $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$. Výpočet aritmetického priemeru sme už ukázali v príklade 2.3.

Dostaneme $\bar{x} = \left(9 \cdot 1 + 11 \cdot 2 + \cdots + 20 \cdot 2 + 21 \cdot 1\right) / 32 = 15,25$.

Kvôli aplikácii MATLABu pre všetky úlohy si vytvoríme nasledujúce vektory:

```
xi=[9,11,12,14:18,20,21];ni=[1,2,4,4,6,5,4,3,2,1];
x=[9,11,11,12*ones(1,4),14*ones(1,4),15*ones(1,6),16*ones(1,5),...
17*ones(1,4),18*ones(1,3),20,20,21]
```

Výpočet aritmetického priemeru pomocou M-funkcie *mean*:

```
ap=mean(x)
```

Rozptyl budeme počítať pomocou vzorca vo váženom tvare $S^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i$.

Dostaneme výsledok:

$$S^2 = \frac{1}{32}\left((9-15,25)^2 \cdot 1 + (11-15,25)^2 \cdot 2 + \cdots + (20-15,25)^2 \cdot 2 + (21-15,25)^2 \cdot 1\right) = 7,5.$$

MATLAB tu môžeme aplikovať jednoduchým prepisom uvedeného vzorca alebo pomocou M-funkcie *var(x,1)*:

```
n=length(x);disp=sum(((xi-ap).^2).*ni)/n,disp=var(x,1)
```

b) Smerodajnú odchýlku vypočítame podľa vzorca $S = \sqrt{S^2}$. Teda $S = \sqrt{7,5} = 2,7386$.

Výpočet pomocou MATLABu dvojakým spôsobom, a to prepisom uvedeného vzorca a pomocou M-funkcie `std(x,1)`:

```
sm_odch=sqrt(disp),sm_odch=std(x,1)
```

c) Priemernú odchýlku vypočítame pomocou vzorca vo váženom tvare $\overline{d} = \frac{1}{n}\sum_{i=1}^{k} |x_i - \overline{x}| \cdot n_i$.

Teda $\overline{d} = \frac{1}{32}\left(|9-15,25| \cdot 1 + |11-15,25| \cdot 2 + \cdots + |20-15,25| \cdot 2 + |21-15,25| \cdot 1\right) = 2,1406$.

Aplikácia MATLABu priamym prepisom uvedeného vzorca a tiež pomocou M-funkcie `mad`:

```
priem_odch=sum(abs(xi-ap).*ni)/n,priem_odch=mad(x)
```

d) Variačné rozpätie vypočítame podľa vzorca $R_V = x_{\max} - x_{\min}$. Teda bude $R_V = 21-9 = 12$.

Ukážeme aplikáciu MATLABu dvojakým spôsobom, a to prepisom uvedeného vzorca a pomocou M-funkcie `range`:

```
var_rozp=max(x)-min(x),var_rozp=range(x)
```

e) Pre kvartilové rozpätie platí vzorec $R_Q = \tilde{x}_{75} - \tilde{x}_{25}$. Výpočet prvého a tretieho kvartilu bez MATLABu a aj s ním sme si ukázali v príklade 2.7.

Teraz si vypočítame kvartilové rozpätie len aplikáciou programu MATLAB, a to ako rozdiel medzi horným a dolným kvartilom alebo použitím M-funkcie `iqr` :

```
kv_rozp=prctile(x,75)-prctile(x,25),kv_rozp=iqr(x)
```

Dostaneme výsledok $R_Q = 3$.

f) Variačný koeficient vypočítame podľa vzorca $V_k = \frac{S}{\overline{x}}$.

Použijeme výsledky predchádzajúcich úloh a dostaneme $V_k = \frac{2,7386}{15,25} = 0,1796$.

Aplikácia MATLABu:

```
var_koef=std(x,1)/mean(x)
```

2.2.3 Charakteristiky tvaru rozdelenia

Charakteristiky tvaru rozdelenia delíme na charakteristiky šikmosti a špicatosti. *Charakteristiky šikmosti* sú založené na porovnávaní stupňa koncentrácie malých hodnôt sledovaného štatistického znaku so stupňom koncentrácie veľkých hodnôt tohto znaku. *Charakteristiky špicatosti* sú založené na porovnávaní stupňa koncentrácie hodnôt strednej veľkosti so stupňom koncentrácie ostatných hodnôt.

Nech $x_1, x_2, \ldots, x_n$ sú namerané hodnoty sledovaného štatistického znaku *X*, ktorých aritmetický priemer je $\overline{x}$ a S je ich smerodajná odchýlka.

Nech $r \in N \cup \{0\}$. ***Centrálny moment r-tého rádu*** μ_r štatistického súboru definujeme vzťahom $\mu_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^r$. Tento *prostý* tvar vzorca sa používa pre neroztriedený štatistický

súbor. Pre roztriedený štatistický súbor sa používa *vážený* tvar vzorca $\mu_r = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^r \cdot n_i$.

Platí napr.: $\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = S^2$.

Pomocou centrálnych momentov tretieho a štvrtého rádu definujeme koeficient šikmosti a koeficient špicatosti.

***Koeficient šikmosti* (*asymetrie*)** γ_3 je definovaný vzťahom $\gamma_3 = \frac{1}{S^3 \cdot n}\sum_{i=1}^{n}(x_i - \bar{x})^3$. Tento *prostý* tvar vzorca sa používa pre neroztriedený štatistický súbor. Pre roztriedený štatistický súbor sa používa *vážený* tvar vzorca $\gamma_3 = \frac{1}{S^3 \cdot n}\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i$. Platí, že $\gamma_3 = \frac{\mu_3}{S^3}$.

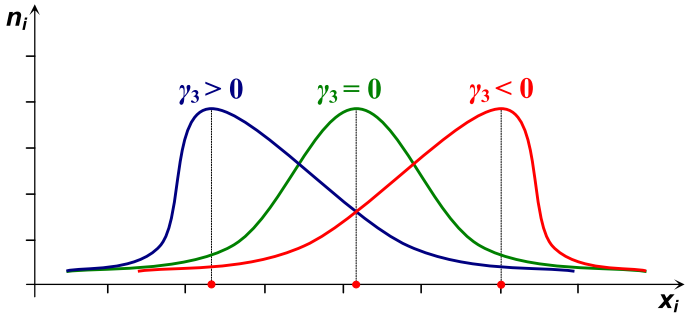
Koeficient šikmosti vyjadruje stupeň asymetrie rozdelenia početností hodnôt štatistického znaku. Rovnaký stupeň koncentrácie malých a veľkých hodnôt sa spravidla prejavuje v symetrii tvaru rozdelenia početností.

Koeficient šikmosti je charakteristika, ktorá nám pomáha rozhodnúť o zhode našich dát s modelom normálneho rozdelenia z hľadiska symetrie okolo aritmetického priemeru.

Pre úplne *symetrické rozdelenie* platí $\gamma_3 = 0$ a je pre neho charakteristické, že všetky hlavné charakteristiky polohy sú totožné, t. j. platí $\bar{x} = \tilde{x} = \hat{x}$. Hodnoty γ_3 blízke nule zodpovedajú rozdeleniu, ktoré sa blíži symetrickému.

Pre $\gamma_3 < 0$ je rozdelenie početností *zošikmené záporne* (*negatívne*), čo sa prejavuje väčšou koncentráciou väčších hodnôt štatistického znaku. Je to rozdelenie s predĺženým ľavým koncom.

Pre $\gamma_3 > 0$ je rozdelenie početností *zošikmené kladne* (*pozitívne*), čo znamená väčšiu koncentráciu menších hodnôt štatistického znaku. Je to teda rozdelenie s predĺženým pravým koncom.

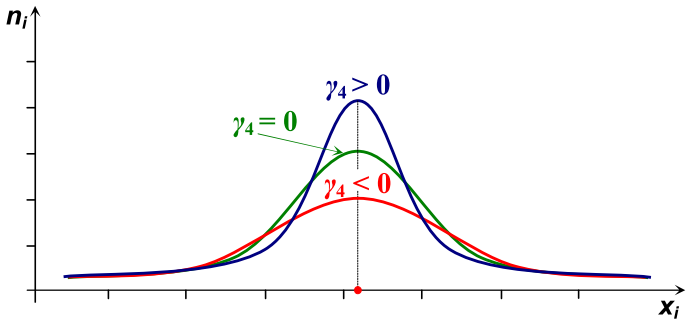


***Koeficient špicatosti* (*excesu*)** γ_4 sa definuje vzťahom $\gamma_4 = \frac{1}{S^4 \cdot n}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3$. Tento *prostý* tvar vzorca sa používa pre neroztriedený štatistický súbor. Pre roztriedený štatistický súbor sa používa *vážený* tvar vzorca $\gamma_4 = \frac{1}{S^4 \cdot n}\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i - 3$. Platí, že $\gamma_4 = \frac{\mu_4}{S^4} - 3$.

Symetrické rozdelenia môžu mať rovnaký rozptyl, ale odlišnú špicatosť. Koeficient špicatosti vyjadruje strmosť rozdelenia početností hodnôt štatistického znaku. Je založený na meraní odchylky špicatosti skúmaného empirického rozdelenia od normálneho rozdelenia, pre ktoré je $\gamma_4 = 0$. Hodnoty γ_4 blízke nule zodpovedajú rozdeleniu, ktoré sa blíži

normálnemu. Hodnoty $\gamma_4 < 0$ reprezentujú plochejšie rozdelenie ako je normálne rozdelenie, kým hodnoty $\gamma_4 > 0$ reprezentujú špicatejšie rozdelenie ako je normálne rozdelenie.

Vyššia špicatosť rozdelenia tiež znamená nižšiu variabilitu hodnôt štatistického znaku.



Príklad 2.12. V rybníku bolo vylovených 425 kaprov, pričom bola zisťovaná ich hmotnosť v kg. Potom sa vhodne zvolili hmotnostné intervaly a bola zostavená nasledujúca tabuľka rozdelenia absolútnych početností, kde z_i je triedny znak:

z_i	0,75	1,25	1,75	2,25	2,75	3,25	3,75
n_i	75	90	97	63	48	42	10

Pre uvedený štatistický súbor vypočítajte koeficienty šikmosti a špicatosti. Znázornite polygón absolútnych početností.

Riešenie: Koeficient šikmosti budeme počítať podľa vzorca $\gamma_3 = \frac{1}{S^3 \cdot n} \sum_{i=1}^{k} (z_i - \bar{x})^3 \cdot n_i = \frac{\mu_3}{S^3}$

a koeficient špicatosti podľa vzorca $\gamma_4 = \frac{1}{S^4 \cdot n} \sum_{i=1}^{k} (z_i - \bar{x})^4 \cdot n_i - 3 = \frac{\mu_4}{S^4} - 3$.

Výpočet pomocou kalkulačky nebudeme uvádzať, nakoľko presne kopíruje uvedené vzorce a podobné výpočty sme už podrobne uviedli pri výpočte aritmetického priemeru, rozptylu a smerodajnej odchýlky. Uvedieme len aplikáciu v programe MATLAB, a to dvoma spôsobmi.

Najprv si do vektora z uložíme vstupné hodnoty:

```
z=[.75*ones(1,75),1.25*ones(1,90),1.75*ones(1,97),...
2.25*ones(1,63),2.75*ones(1,48),3.25*ones(1,42),3.75*ones(1,10)];
```

Výpočet γ_3 a γ_4 použitím M-funkcie *moment*:

```
gama3=moment(z,3)/std(z,1)^3,gama4=(moment(z,4)/std(z,1)^4)-3
```

Výpočet γ_3 a γ_4 použitím M-funkcií *skewness* a *kurtosis*:

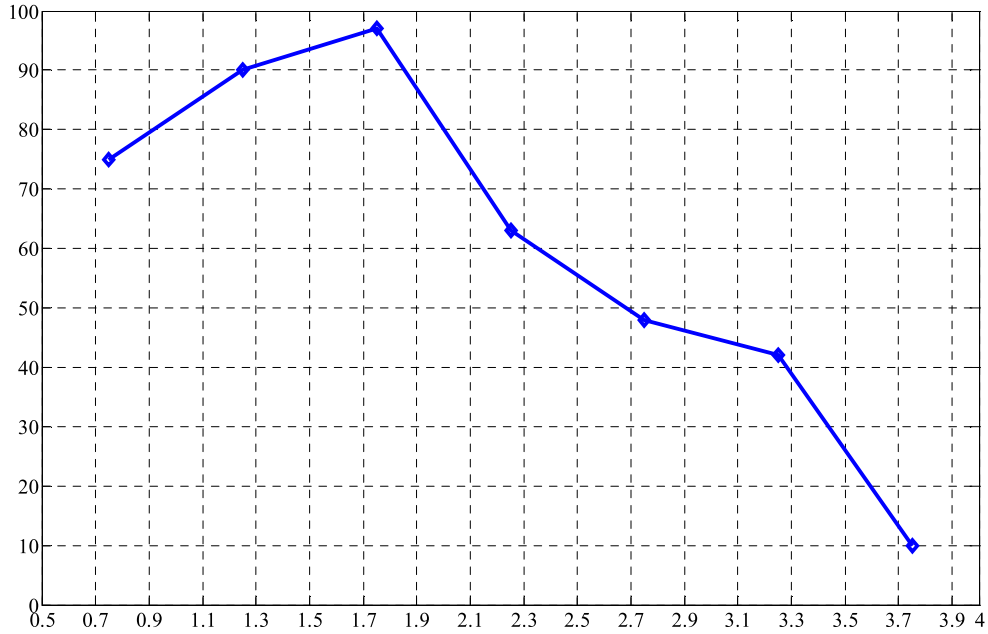
```
gama3=skewness(z),gama4=kurtosis(z)-3
```

Dostaneme tieto výsledky: $\gamma_3 = 0,4287$ a $\gamma_4 = -0,7680$.

Keďže $\gamma_3 > 0$, tak rozdelenie má pozitívnu asymetriu. Pretože $\gamma_4 < 0$, tak dané rozdelenie je plochejšie ako normálne rozdelenie.

Ešte znázorníme polygón absolútnych početností n_i:

```
zi=0.75:0.5:3.75;ni=[75,90,97,63,48,42,10];plot(zi,ni,'d-'),grid
```



2.3 Elementárne spracovanie súborov s kvalitatívnym triediacim znakom

Často nás zaujímajú také vlastnosti štatistických jednotiek, ktoré nemôžeme vyjadriť číselne, ale charakterizujeme ich pomocou kvalitatívnych štatistických znakov. Jednotlivé varianty znaku vyjadrujeme v tomto prípade slovne.

Ak bol sledovaný kvalitatívny štatistický znak, je potrebné získané neusporiadané údaje o tomto znaku nejakým spôsobom sprehľadniť. V prípade alternatívnych znakov používame tzv. *asociačné tabuľky* a v prípade množných znakov tzv. *kontingenčné tabuľky*.

Alternatívny kvalitatívny štatistický znak je možné ľahko kvantifikovať. Jednému variantu tohto znaku sa priradí hodnota 1 a druhému variantu hodnota 0. O tomto znaku sa potom hovorí ako o *nulovo-jedničkovej veličine*.

V prípade množného kvalitatívneho štatistického znaku nie je možné jednoznačne usporiadať štatistické jednotky podľa variantov kvalitatívneho znaku ako to bolo v prípade kvantitatívnych znakov, kde sa varianty hodnôt, resp. skupiny hodnôt znaku usporadúvali od najnižšej hodnoty po najvyššiu. Niekedy je usporiadanie variantov množného kvalitatívneho štatistického znaku dané logicky, niekedy ich usporadúvame napr. od variantu s najväčšou početnosťou k variantu s najnižšou početnosťou alebo naopak.

Ak máme usporiadať štatistické jednotky podľa množného kvalitatívneho štatistického znaku, ktorý nadobúda veľký počet variantov (napr. obyvateľstvo SR podľa zamestnania), potom je nutné stanoviť aj skupiny, do ktorých sa zahrnie vždy niekoľko kvalitatívnych znakov zhodných z určitého hľadiska.

Skupiny musia byť presne definované, aby nedochádzalo k nejasnostiam pri zaraďovaní jednotlivých znakov. Ak sú dané jednotlivé varianty kvalitatívneho znaku alebo ak sú definované skupiny, potom je sprehľadnenie neusporiadaných údajov o kvalitatívnom štatistickom znaku založené na spočítaní jednotiek s jednotlivými variantmi, prípadne spočítaní jednotiek patriacich do jednotlivých skupín.

Aj v prípade kvalitatívnych dát môžeme určovať absolútne a relatívne početnosti a taktiež je vhodné zobrazovať usporiadané údaje graficky. Dve najčastejšie používané metódy znázornenia kvalitatívnych dát sú ***stĺpcové grafy*** a ***kruhové diagramy***, ktoré sa nazývajú aj ***koláčové grafy***.

Stĺpcový graf sme spomínali už pri grafickom zobrazovaní kvantitatívnych dát. Je to graf podobný histogramu, pričom jeho stĺpce sa navzájom nedotýkajú.

Kruhový diagram je kruh rozdelený na časti v tvare „kúskov koláča“, ktoré získame rozdelením stredového uhla kružnice úmerne k podielu jednotlivých častí zobrazovaného javu vyjadrených v percentách.

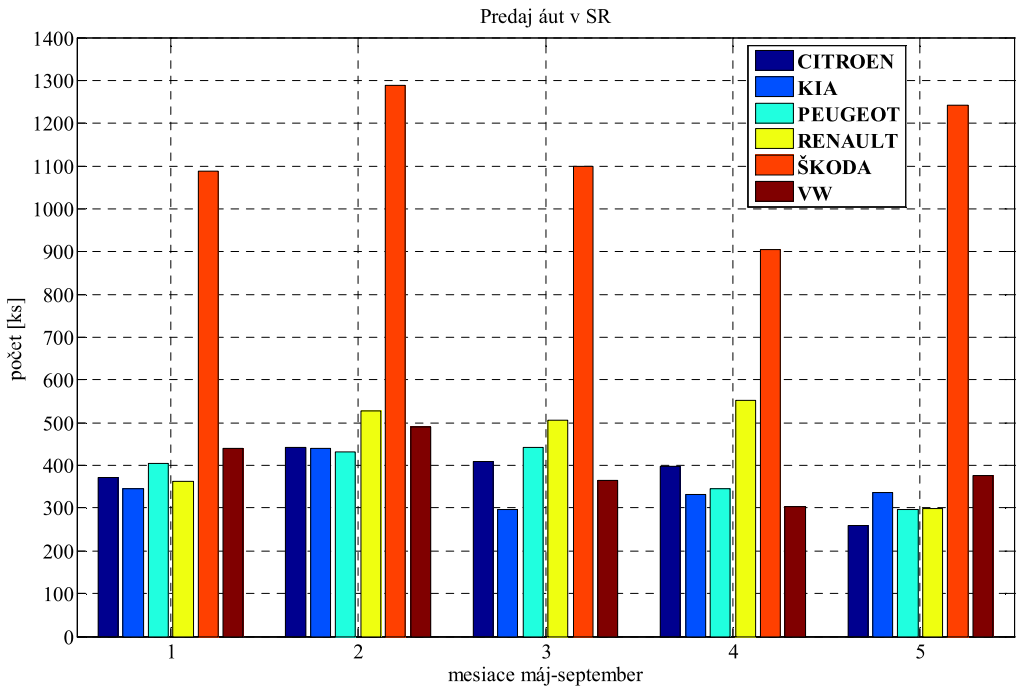
Príklad 2.13. Použili sme údaje zo stránky http://www.zapsr.sk/kategoria/Statistiky/Predaj-automobilov-v-SR/. V tabuľke sú uvedené dáta o počte predaných kusov šiestich najpredávanejších značiek osobných áut v SR za obdobie máj – september v roku 2010:

Značka	máj	jún	júl	august	september
CITROEN	372	441	408	398	260
KIA	346	439	297	332	336
PEUGEOT	404	431	442	345	297
RENAULT	363	529	505	552	298
ŠKODA	1089	1289	1099	905	1243
VW	440	491	365	304	377

Uvedené dáta znázornite graficky pomocou stĺpcového grafu a koláčového 3D-grafu.

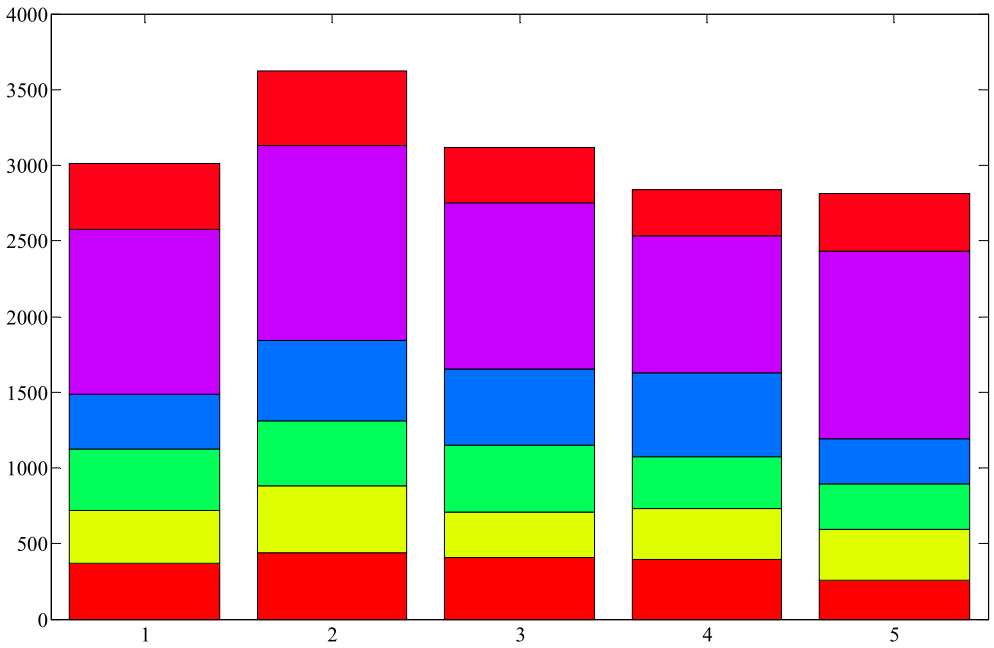
Riešenie: Dáta uložíme do matice *poc* a nakreslíme stĺpcový graf pomocou M-funkcie `bar`:

```
poc=[372,346,404,363,1089,440;441,439,431,529,1289,491;...
408,297,442,505,1099,365;398,332,345,552,905,304;...
260,336,297,298,1243,377];bar(poc),grid,
xlabel('mesiace máj-september'),ylabel('počet [ks]'),
legend('CITROEN','KIA','PEUGEOT','RENAULT','ŠKODA','VW'),
title('Predaj áut v SR')
```



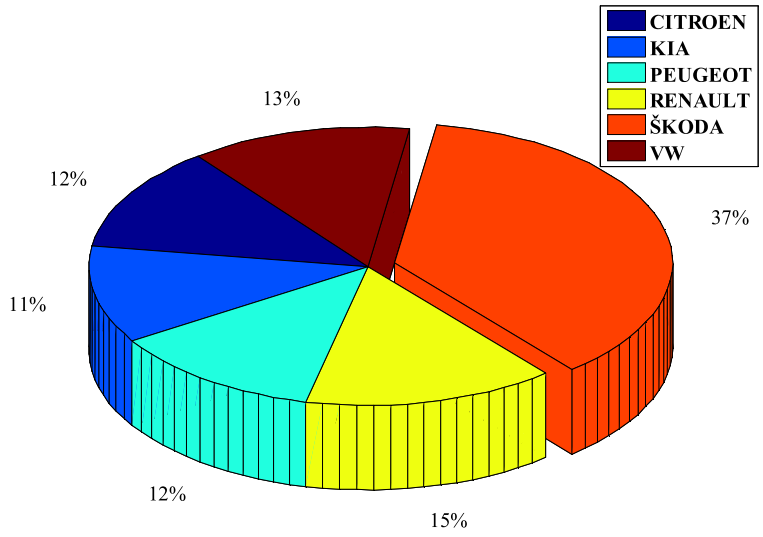
Prepínačom *stack* docielime, že sa jednotlivé hodnoty zobrazia na seba, a teda výsledný stĺpec bude súčtom všetkých čiastkových hodnôt. V našom príklade teda pôjde o súčet počtu predaných kusov uvedených značiek áut v jednotlivých mesiacoch máj – september.

```
bar(poc,'stack'),colormap(hsv)
```



Na grafické znázornenie koláčového 3D-grafu použijeme M-funkciu *pie3*. Koláčový diagram môžeme mierne modifikovať tak, že napr. niektorú časť zvýrazníme jej vysunutím. Slúži na to druhý parameter v príkaze *pie*, ktorý reprezentuje vektor rovnakej dĺžky ako je počet zobrazovaných hodnôt. Ak na príslušné miesto dáme nenulovú hodnotu, tak daná položka grafu bude vysunutá.

```
vyber=[0 0 0 0 1 0];pie3(sum(poc),vyber),
legend('CITROEN','KIA','PEUGEOT','RENAULT','ŠKODA','VW')
```



Úlohy:

2.1. Je daný súbor nameraných hodnôt:

105, 95, 100, 98, 105, 110, 95, 98, 112, 120, 105, 105, 110, 100, 98, 100, 112, 110, 110, 105, 105, 98, 100, 100, 110, 105, 110, 112, 98, 100, 105, 105.

Pre daný štatistický súbor riešte nasledujúce úlohy: a) zostavte variačný rad, b) zostavte frekvenčnú tabuľku, c) zostavte tabuľku rozdelenia početností, d) zobrazte polygón absolútnych početností n_i, úsečkový graf relatívnych početností f_i, stĺpcový graf kumulatívnych absolútnych početností N_i, schodovitý graf kumulatívnych relatívnych početností F_i ; e) vypočítajte aritmetický priemer, modus, medián, prvý kvartil, tretí kvartil, rozptyl, smerodajnú odchýlku, priemernú odchýlku, variačné rozpätie, kvartilové rozpätie, variačný koeficient, koeficient šikmosti a koeficient špicatosti.

Výsledok:

a) 95, 95, 98, 98, 98, 98, 98, 100, 100, 100, 100, 100, 100, 105, 105, 105, 105, 105, 105, 105, 105, 105, 110, 110, 110, 110, 110, 110, 112, 112, 112, 120.

b)

x_i	95	98	100	105	110	112	120
n_i	2	5	6	9	6	3	1

c)

i	x_i	n_i	f_i	N_i	F_i
1	95	2	1/16	2	1/16
2	98	5	5/32	7	7/32
3	100	6	3/16	13	13/32
4	105	9	9/32	22	11/16
5	110	6	3/16	28	7/8
6	112	3	3/32	31	31/32
7	120	1	1/32	32	1
$\sum$	—	32	1	—	—

e) $\bar{x} = 104,4063$; $\hat{x} = 105$, $\tilde{x}_{50} = 105$, $\tilde{x}_{25} = 100$, $\tilde{x}_{75} = 110$, $S^2 = 34,5537$; $S = 5,8782$; $\bar{d} = 4,8301$; $R_V = 25$, $R_Q = 10$, $V_k = 0,0563$; $\gamma_3 = 0,4101$; $\gamma_4 = -0,3481$.

2.2. V priebehu päťdesiatich týždňov bola sledovaná kvalita vyrábaných betónových panelov. Počet nekvalitných betónových panelov bol nasledovný:

14, 16, 11, 10, 8, 13, 12, 14, 16, 12, 15, 13, 12, 10, 16, 12, 17, 9, 12, 12, 14, 18, 15, 13, 17, 9, 13, 11, 11, 12, 15, 13, 14, 13, 13, 8, 10, 15, 11, 11, 14, 14, 11, 9, 13, 10, 16, 15, 13, 12.

Pre daný štatistický súbor riešte úlohy: a) zostavte tabuľku rozdelenia absolútnych početností a zobrazte ju ako histogram, b) vypočítajte aritmetický priemer, modus, medián, prvý kvartil, tretí kvartil, rozptyl, smerodajnú odchýlku, priemernú odchýlku, variačné rozpätie, kvartilové rozpätie, variačný koeficient, koeficient šikmosti a koeficient špicatosti.

Výsledok: a)

x_i	8	9	10	11	12	13	14	15	16	17	18
n_i	2	3	4	6	8	9	6	5	4	2	1

b) $\bar{x} = 12,74$; $\hat{x} = 13$, $\tilde{x}_{50} = 13$, $\tilde{x}_{25} = 11$, $\tilde{x}_{75} = 14$, $S^2 = 5,6324$; $S = 2,3733$; $\bar{d} = 1,9208$; $R_V = 10$, $R_Q = 3$, $V_k = 0,1863$; $\gamma_3 = 0,0322$; $\gamma_4 = -0,5407$.

2.3. Pri skúškach pevnosti [MPa] lán boli namerané tieto výsledky:
302, 310, 312, 310, 313, 318, 305, 309, 301, 309, 310, 311, 307, 308, 311, 300, 310, 308, 310, 307, 313, 299, 315, 312, 310, 308, 314, 333, 305, 310, 309, 314.

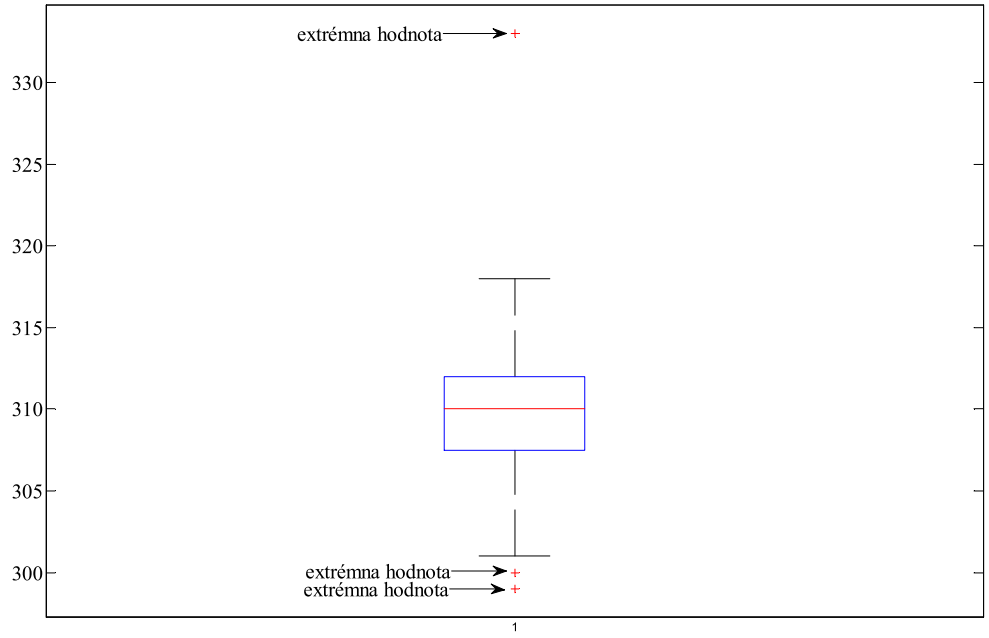
Pre daný štatistický súbor riešte úlohy: a) zostavte variačný rad, b) vypočítajte aritmetický priemer, modus, medián, prvý kvartil, tretí kvartil, rozptyl, smerodajnú odchýlku, priemernú odchýlku, variačné rozpätie, kvartilové rozpätie, variačný koeficient, koeficient šikmosti a špicatosti; c) znázornite boxplot.

Výsledok:

a) 299, 300, 301, 302, 305, 305, 307, 307, 308, 308, 308, 309, 309, 309, 310, 310, 310, 310, 310, 310, 310, 311, 311, 312, 312, 313, 313, 314, 314, 315, 318, 333.

b) $\bar{x} = 309,7813$; $\hat{x} = 310$, $\tilde{x}_{50} = 310$, $\tilde{x}_{25} = 307,5$, $\tilde{x}_{75} = 312$, $S^2 = 35,1709$; $S = 5,9305$; $\bar{d} = 3,7461$; $R_V = 34$, $R_Q = 4,5$, $V_k = 0,0191$; $\gamma_3 = 1,4742$; $\gamma_4 = 5,3448$.

c)



2.4. Zisťovali sme hmotnosť [g] 40 vzoriek jedľového semena s týmito výsledkami:
4,20; 4,54; 4,37; 4,47; 4,29; 4,52; 4,38; 4,21; 4,53; 4,60; 4,56; 4,70; 4,59; 4,73; 4,61; 4,31; 4,42; 4,73; 4,37; 4,64; 4,46; 4,66; 4,50; 4,71; 4,32; 4,69; 4,39; 4,59; 4,34; 4,70; 4,70; 4,28; 4,41; 4,63; 4,80; 4,61; 4,44; 4,85; 4,76; 4,79.

Riešte nasledujúce úlohy: a) daný štatistický súbor roztrieďte do triednych intervalov, pričom zvoľte počet tried $k = 7$; b) zostavte tabuľku rozdelenia početností, c) znázornite stĺpcový graf kumulatívnych absolútnych početností N_i a schodovitý graf kumulatívnych relatívnych početností F_i.

Výsledok: a) $R_V = 4,85 - 4,2 = 0,65$; $h \approx R_V / k = 0,0929$. Volíme $h = 0,1$.

$I_i = \langle t_{i-1}, t_i \rangle$	4,15–4,25	4,25–4,35	4,35–4,45	4,45–4,55	4,55–4,65	4,65–4,75	4,75–4,85
z_i	4,2	4,3	4,4	4,5	4,6	4,7	4,8
n_i	2	5	7	6	8	8	4

b)

i	z_i	n_i	f_i	N_i	F_i
1	4,2	2	0,05	2	0,05
2	4,3	5	0,125	7	0,175
3	4,4	7	0,175	14	0,35
4	4,5	6	0,15	20	0,5
5	4,6	8	0,2	28	0,7
6	4,7	8	0,2	36	0,9
7	4,8	4	0,1	40	1
$\sum$	—	40	1	—	—

2.5. V mechanickej dielni bolo vyrobených za jednu smenu 65 čapov. Hodnoty priemeru čapu v stotinách milimetra, ktoré boli zistené meraním týchto čapov, sú nasledovné:

1002, 1004, 1006, 1011, 1003, 1018, 996, 993, 994, 1014, 1002, 1004, 1017,
1006, 1015, 1005, 1011, 1002, 997, 1006, 1012, 998, 994, 993, 1012, 1014,
1012, 1004, 981, 1018, 1014, 1012, 997, 1014, 996, 1005, 982, 1018, 987,
997, 1020, 1017, 988, 1012, 1003, 1014, 987, 993, 995, 1011, 1012, 1004,
994, 996, 1002, 995, 1003, 1004, 1003, 1006, 1003, 1011, 1002, 996, 994.

Riešte nasledujúce úlohy: a) daný štatistický súbor roztrieďte do triednych intervalov, pričom zvoľte počet tried $k = 5$; b) zostavte tabuľku rozdelenia početností, c) znázornite schodovitý graf kumulatívnych relatívnych početností F_i, d) určte empirickú distribučnú funkciu.

Výsledok: a) – b)

i	$I_i = \langle t_{i-1}, t_i \rangle$	z_i	n_i	f_i	N_i	F_i
1	980 – 988	984	5	0,0769	5	0,0769
2	988 – 996	992	13	0,2	18	0,2769
3	996 – 1004	1000	19	0,2923	37	0,5692
4	1004 – 1012	1008	16	0,2462	53	0,8154
5	1012 – 1020	1016	12	0,1846	65	1
$\sum$	—	—	65	1	—	—

d) $$F(x) = \begin{cases} 0 & \text{pre} \quad -\infty < x < 984 \\ 0,0769 & \text{pre} \quad 984 \le x < 992 \\ 0,2769 & \text{pre} \quad 992 \le x < 1000 \\ 0,5692 & \text{pre} \quad 1000 \le x < 1008 \\ 0,8154 & \text{pre} \quad 1008 \le x < 1016 \\ 1 & \text{pre} \quad 1016 \le x < \infty \end{cases}.$$

2.6. V nasledujúcej tabuľke je uvedené skupinové triedenie dealerov počítačovej firmy podľa počtu predaných kusov počítačových zostáv v poslednom štvrťroku bežného roku:

$I_i = \langle t_{i-1}, t_i \rangle$	50–150	150–250	250–350	350–450	450–550	550–650	650–750	750–850
n_i	8	9	7	13	11	75	12	8

Pre daný štatistický súbor riešte úlohy: a) zostavte tabuľku rozdelenia početností, b) vypočítajte aritmetický priemer, modus, medián, prvý kvartil, tretí kvartil, rozptyl, smerodajnú odchýlku, priemernú odchýlku, variačný koeficient, koeficient šikmosti a koeficient špicatosti.

Výsledok: a)

i	$I_i = (t_{i-1}, t_i\rangle$	z_i	n_i	f_i	N_i	F_i
1	50 – 150	100	8	0,0559	8	0,0559
2	150 – 250	200	9	0,0629	17	0,1189
3	250 – 350	300	7	0,0490	24	0,1678
4	350 – 450	400	13	0,0909	37	0,2587
5	450 – 550	500	11	0,0769	48	0,3357
6	550 – 650	600	75	0,5245	123	0,8601
7	650 – 750	700	12	0,0839	135	0,9441
8	750 – 850	800	8	0,0559	143	1
$\sum$	—	—	143	1	—	—

b) $\bar{x} = 525,8741$; $\hat{x} = 600,3937$; $\tilde{x}_{50} = 581,\overline{3}$; $\tilde{x}_{25} = 440,3846$; $\tilde{x}_{75} = 629$, $S^2 = 3,0449 \cdot 10^4$; $S = 174,4976$; $\bar{d} = 137,6478$; $V_k = 0,3318$; $\gamma_3 = -1,0232$; $\gamma_4 = 0,3534$.

2.7. Je daná tabuľka intervalového rozdelenia početností počtu škôl podľa priemerného počtu žiakov na jednu triedu:

I_i	16 – 18	18 – 20	20 – 22	22 – 24	24 – 26	26 – 28	28 – 30
n_i	6	10	22	16	10	4	2

Pre daný štatistický súbor riešte nasledujúce úlohy: a) zostavte tabuľku rozdelenia početností, b) znázornite stĺpcový graf početností N_i, c) vypočítajte aritmetický priemer, modus, medián, prvý kvartil, tretí kvartil, rozptyl, smerodajnú odchýlku, priemernú odchýlku, variačný koeficient, koeficient šikmosti a koeficient špicatosti.

Výsledok: a)

i	$I_i = (t_{i-1}, t_i\rangle$	z_i	n_i	f_i	N_i	F_i
1	16 – 18	17	6	0,0857	6	0,0857
2	18 – 20	19	10	0,1429	16	0,2286
3	20 – 22	21	22	0,3143	38	0,5429
4	22 – 24	23	16	0,2286	54	0,7714
5	24 – 26	25	10	0,1429	64	0,9143
6	26 – 28	27	4	0,0571	68	0,9714
7	28 – 30	29	2	0,0286	70	1
$\sum$	—	—	70	1	—	—

c) $\bar{x} = 21,9714$; $\hat{x} = 21,\overline{3}$; $\tilde{x}_{50} = 21,7273$; $\tilde{x}_{25} = 20,1364$; $\tilde{x}_{75} = 23,8125$; $S^2 = 8,0849$; $S = 2,8434$; $\bar{d} = 2,3118$; $V_k = 0,1294$; $\gamma_3 = 0,2974$; $\gamma_4 = -0,2109$.