ŠTATISTICKÉ SPRACOVANIE DÁT

Štatistická analýza jednorozmerných dát

prof. Ing. Dušan KUDELAS, PhD

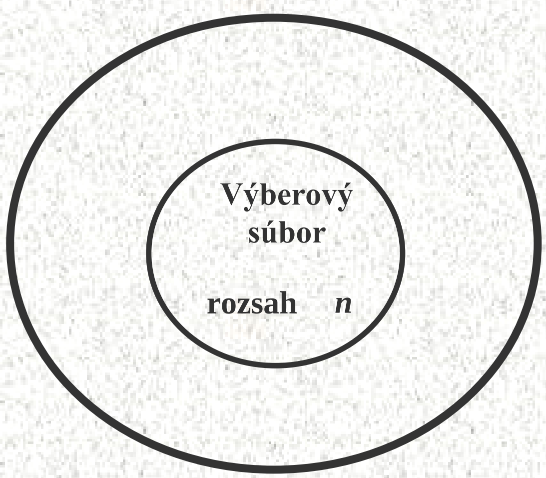
- Prieskumová analýza dát je odvetvie štatistiky, ktoré pomocou rôznych postupov odhaľuje zvláštnosti v dátach, ako sú lokálne koncentrácie dát, tvarové zvláštnosti rozdelenia dát a prítomnosť podozrivých hodnôt.
- Bezprostredné zisťovanie hodnôt náhodných veličín sa vykonáva pozorovaním.

- Štatistickou jednotkou, alebo jednotkou pozorovania sa nazýva ten predmet, osoba, alebo jav, vo vzťahu ku ktorému sa zbierajú fakty (údaje) pozorovaní a z ktorých sa skladá štatistický súbor. Vymedzenie štatistických jednotiek musí byť presné a jednoznačné z hľadiska:

1. vecného (čo je predmetom pozorovania),
2. časového (ako dlho bude skúmaná jednotka predmetom pozorovania),
3. priestorového (kde sa pozorovanie bude realizovať).

- Štatistické znaky (premenné) je možné podľa ich charakteru rozdeliť:
- Kvalitatívne štatistické znaky vyjadrujú vlastnosti štatistických jednotiek, ktoré sa musia opísať slovom, alebo definíciou (národnosť, povolanie, vzdelanie, farba a pod.). Možné vyjadriť aj číselne, ale nie je ju možné použiť k číselným operáciám (**Nominálna a ordinálna premenná (stupnica)**).
- Kvantitatívne štatistické znaky charakterizujú vlastnosti štatistických jednotiek, ktoré sa vyjadrujú číselne a hodnotia mieru prítomnosti kvality (telesná výška, hmotnosť, výška príjmu, zrýchlenie auta, počet prihlásených a pod.). Podľa toho, či kvantitatívne štatistické znaky môžu nadobúdať v rámci nejakého intervalu akékoľvek hodnoty alebo iba niektoré, hovorí sa o *spojitých* a *nespojitých* (diskrétnych) znakoch (**Kardinálna premenná – intervalová, pomerová stupnica**).
  - Spojitý znak môže nadobudnúť akékoľvek reálne hodnoty z určitého intervalu (zrýchlenie, čas, spotreba pohonných látok a pod.).
  - Nespojitý (diskrétny) znak môže nadobúdať len oddelené hodnoty. Typickým príkladom diskrétnych veličín sú počty (počet prihlásených na VŠ, počet predaných výrobkov a pod.).

- Štatistický súbor (základný) - súhrn štatistických jednotiek rovnakého druhu. Rozsah štatistického súboru je určený počtom štatistických jednotiek v súbore
- Náhodný výber (výberový súbor) je súbor náhodne vybraných rovnorodých štatistických jednotiek z jedného základného súboru.
- Základný súbor, je súbor jednotiek, z ktorých sa koná výber.



Základný

súbor, rozsah *N*, resp ∞

Tri úrovne využívania štatistiky:

- **Opisná štatistika** (sprehľadnenie dát)
- **Analýza dát** (zisťovanie vzťahov v dátach)
- **Induktívna štatistika** (zovšeobecňovanie a extrapolácia získaných výsledkov z analýzy dát)

Opisná (deskriptívna) štatistika

- **Účel**:
  - Sprehľadniť získané údaje (tabuľky, grafické prostriedky)
  - Charakterizovať výbery pomocou kvantitatívnych charakteristík (stredové hodnoty, hodnoty rozptylu)

Triedenie

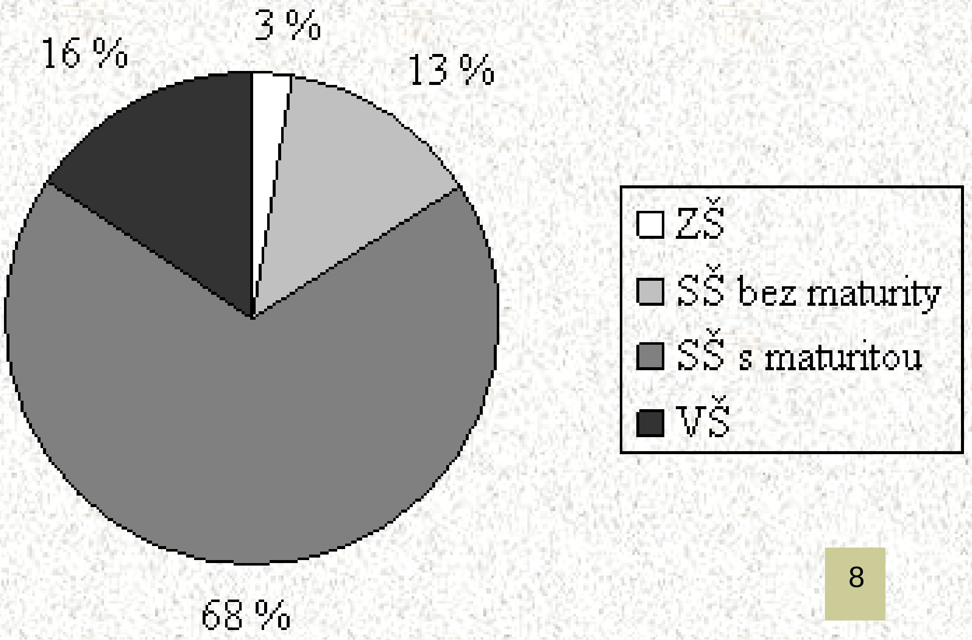
- Prvým krokom v spracovaní štatistických údajov je triedenie
- podľa **kvalitatívnych** znakov sa štatistický súbor rozdeľuje na skupiny (triedy) podľa variantov kvalitatívneho znaku alebo viacerých kvalitatívnych znakov: -
- *podvojné* (dichotomické) triedenie(napr. pohlavie).
- *množné* (multinomické) triedenie (napr. vzdelanie).
- podľa **kvantitatívneho** znaku je proces usporiadania jednotiek štatistického súboru do skupín (tried) podľa veľkosti štatistického znaku.
- Ďalší postup závisí od toho, či sa pracuje s diskrétnou, alebo spojitou náhodnou premennou (veličinou).

Triedenie podľa kvalitatívneho znaku

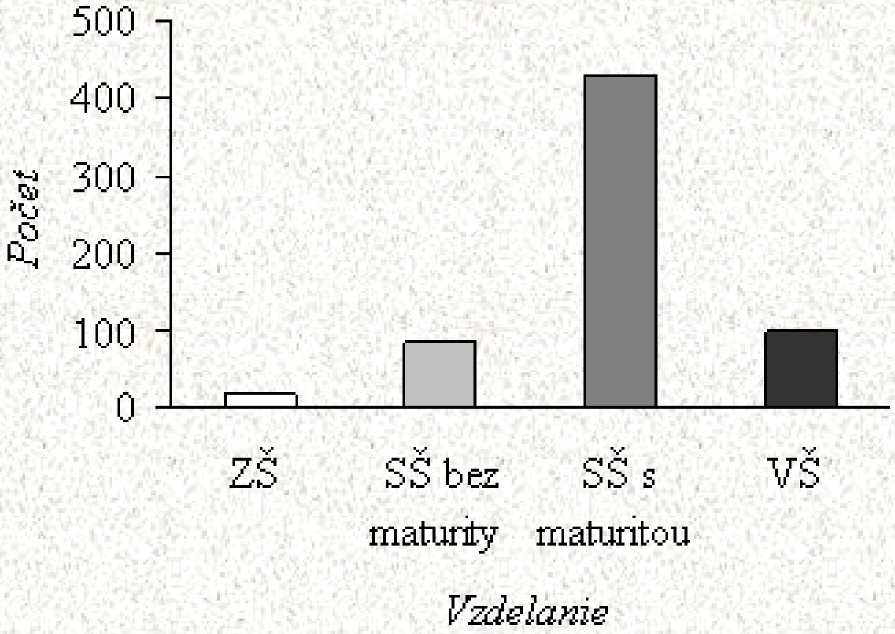
Jednoduchá (frekvenčná) tabuľka zodpovedá jednostupňovému triedeniu. Obsahuje triedne početnosti podľa kategórií (tried) jednej premennej. Zostrojenie frekvenčných tabuliek z údajov sa nazýva tabelácia. Nasledujúci tabuľka obsahuje rozdelenie respondentov podľa vzdelania.

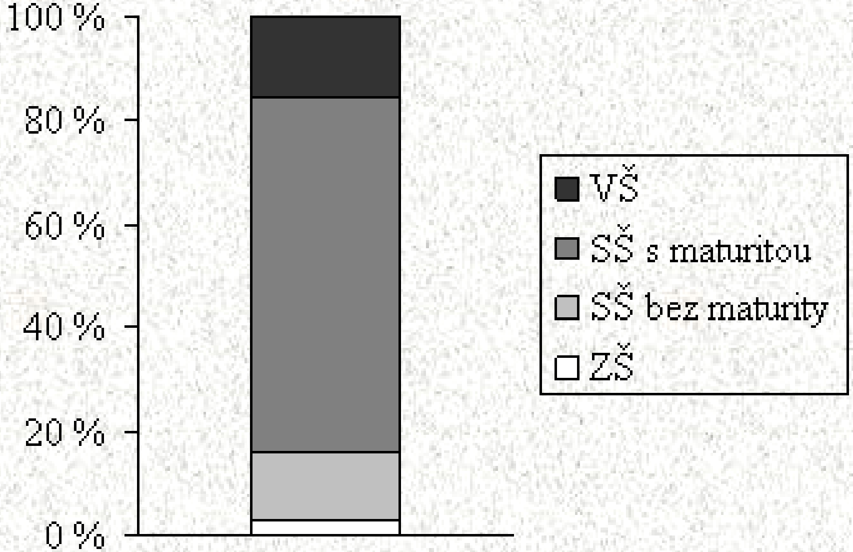
Koláčový graf predstavuje najvhodnejšie grafické znázornenie percent (relatívnych početností) frekvenčnej tabuľky. Nasledujúci obrázok je zostrojený z frekvenčnej tabuľky.

Vzdelanie	nj	fj
Základné	17	0,03
Stredoškolské BM	83	0,13
Stredoškolské M	428	0,68
VŠ	98	0,16



Stĺpcový graf





Triedenie podľa kvantitatívneho znaku -– diskrétna náhodná veličina

- **Pr.**: Istá firma vyrobila počas 10 dní nasledovné množstvá výrobkov: 154; 150; 154; 152; 152; 154; 155; 154; 152; 156.

  Variačný rad: 150; 152; 152; 152; 154; 154; 154; 154; 155; 156.

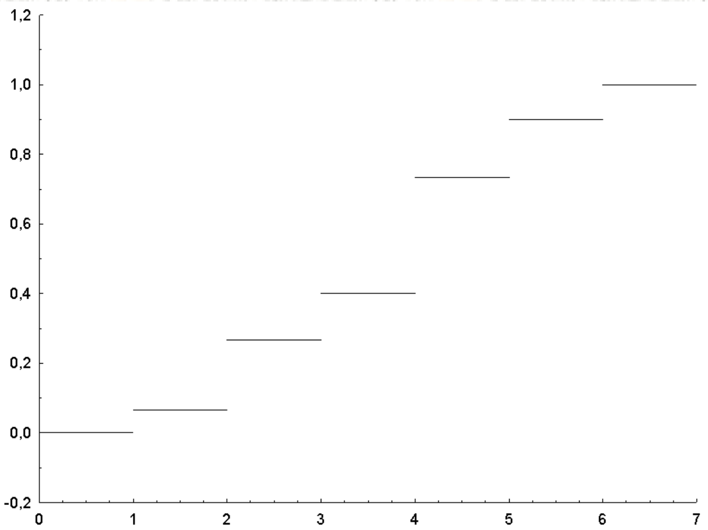
Hodnoty X	150	152	154	155	156
Početnosť n_j	1	3	4	1	1

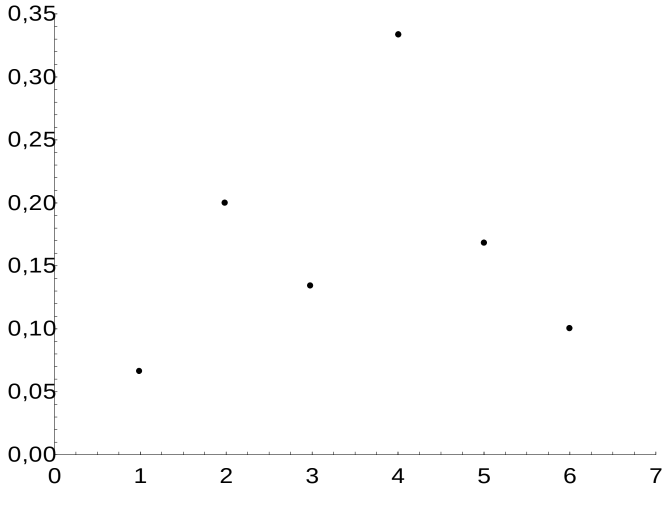
Aby bola tabuľka rozdelenia početnosti úplná pridávajú sa k tabuľke ďalšie riadky (stĺpce) a to riadok s označením:
•Relatívna početnosť, čo je podiel početnosti a celkového počtu hodnôt (je možné ju vyjadriť v %, to znamená vynásobiť číslom 100),
•Kumulatívna početnosť – súčet početností od začiatku, až po danú hodnotu (napr. kumulatívna početnosť hodnoty 154 je 8 = 1 + 3 + 4 (priebežný súčet početností)),
•Kumulatívna relatívna početnosť je podiel kumulatívnej početnosti a celkového počtu hodnôt (je možné ju vyjadriť v %, to znamená vynásobiť číslom 100).

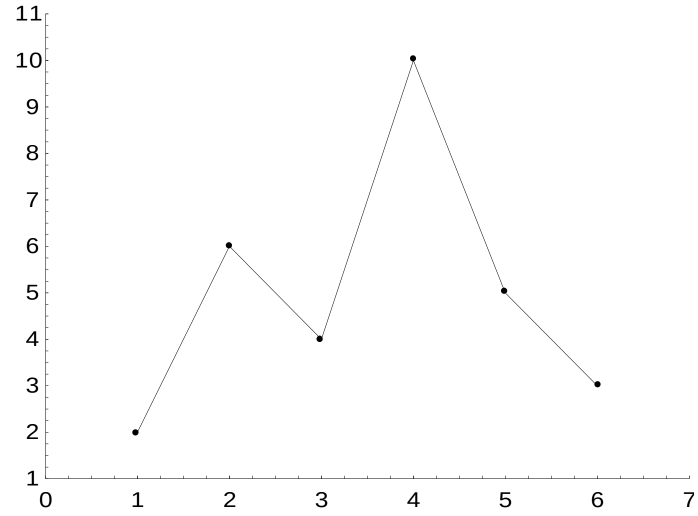
Hodnoty X	150	152	154	155	156
Početnosť n_j	1	3	4	1	1
Relatívna početnosť f_j	1/10	3/10	4/10	1/10	1/10
Kumulatívna početnosť N_j	1	4	8	9	10
Kumulatívna relatívna početnosť F_j	1/10	4/10	8/10	9/10	10/10

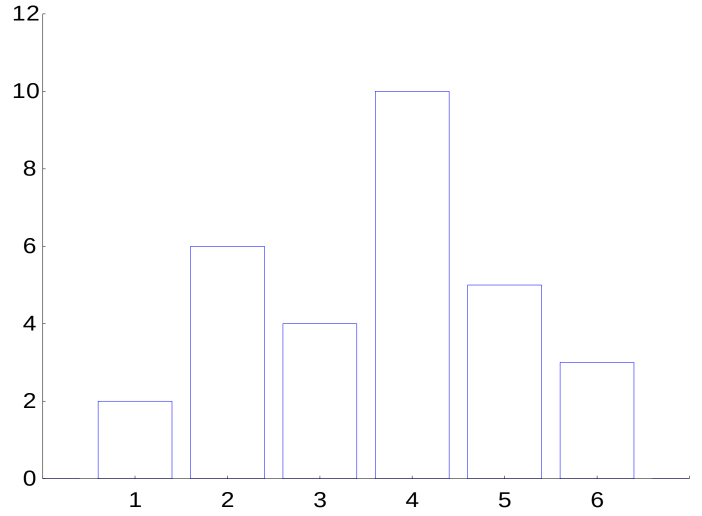
- **Pr.**: V 30 domácnostiach bol zisťovaný počet osôb. Výsledky sú zaznamenané do tabuľky.

$x_{[j]}$	n_j	f_j	N_j	F_j
1	2	2/30	2	2/30
2	6	6/30	8	8/30
3	4	4/30	12	12/30
4	10	10/30	22	22/30
5	5	5/30	27	27/30
6	3	3/30	30	1









Spojitá náhodná veličina

- Variačný rad
- Variačné rozpätie
- Určenie počtu tried
  (5 – 20)
- Určenie šírky triedy
- Zostrojenie tabuľky početnosti (frekvečnej tabuľky)
- Výpočet opisných charakteristík
- Grafická prezentácia - histogram

$$R = x_{\max} - x_{\min}$$

$$k = \sqrt{n}$$

$$k = 5\log n$$

$$k = 1 + 3{,}3\log n$$

$$c = \frac{R}{k}$$

Spojitá náhodná veličina

- **Pr.**: Zostrojme tabuľku početnosti z dát spojitej náhodnej veličiny. Dáta sú výsledky analýz % obsahu Fe v železnej rude.

30,4	30,6	30,7	30,8	30,8	30,8	30,9	30,9	31	31
31	31,1	31,1	31,2	31,2	31,2	31,2	31,2	31,2	31,2
31,2	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3
31,3	31,3	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4
31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,6	31,6	31,6	31,6
31,6	31,6	31,6	31,6	31,6	31,6	31,6	31,7	31,7	31,7
31,7	31,7	31,7	31,7	31,7	31,8	31,8	31,8	31,8	31,8
31,9	31,9	31,9	32	32,1	32,1	32,1	32,2	32,2	32,5

$$R = x_{\max} - x_{\min} = 32{,}5 - 30{,}4 = 2{,}1$$

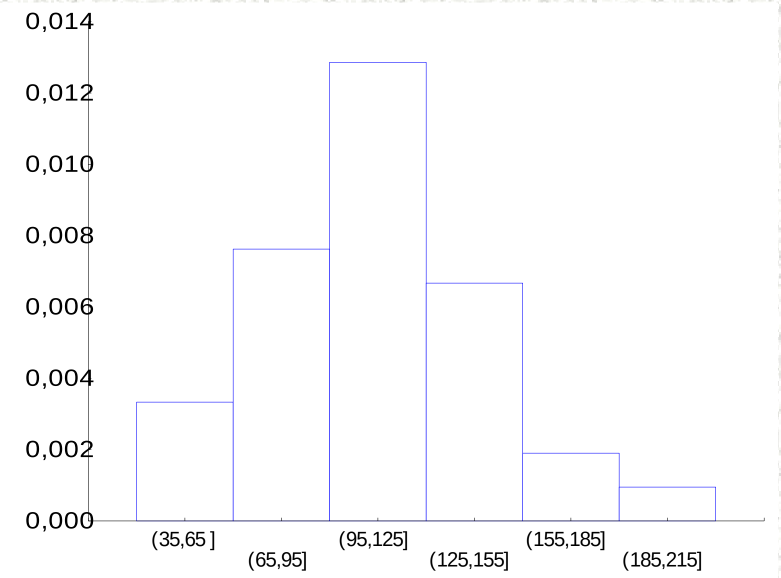
$$k = \sqrt{n} = \sqrt{100} = 10$$

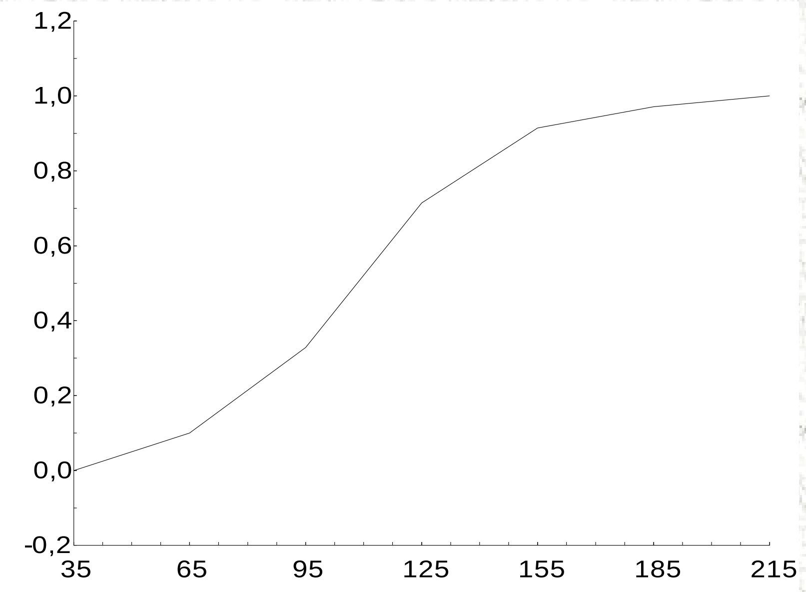
$$c = \frac{R}{k} = \frac{2{,}1}{10} = 0{,}21$$

Trieda	x_d	x_h	x_j	n_j	f_j	N_j	F_j
1	30,4	30,6	30,5	2	2/100	2	2/100
2	30,6	30,8	30,7	4	4/100	6	6/100
3	30,8	31,0	30,9	5	5/100	11	11/100
4	31,0	31,2	31,1	10	10/100	21	21/100
5	31,2	31,4	31,3	27	27/100	48	48/100
6	31,4	31,6	31,5	29	29/100	77	77/100
7	31,6	31,8	31,7	13	13/100	90	90/100
8	31,8	32,0	31,9	4	4/100	94	94/100
9	32,0	32,2	32,1	5	4/100	99	99/100
10	32,2	32,4	32,3	0	0/100	99	99/100
11	32,4	32,6	32,5	1	1/100	100	1

30,4	30,6	30,7	30,8	30,8	30,8	30,9	30,9	31	31
31	31,1	31,1	31,2	31,2	31,2	31,2	31,2	31,2	31,2
31,2	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3
31,3	31,3	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4
31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,6	31,6	31,6	31,6
31,6	31,6	31,6	31,6	31,6	31,6	31,6	31,7	31,7	31,7
31,7	31,7	31,7	31,7	31,7	31,8	31,8	31,8	31,8	31,8
31,9	31,9	31,9	32	32,1	32,1	32,1	32,2	32,2	32,5

(x_d, x_h)	n_j	f_j	N_j	F_j
$\langle 35, 65 \rangle$	7	7/70	7	7/70
$\langle 65, 95 \rangle$	16	16/70	23	23/70
$\langle 95, 125 \rangle$	27	27/70	50	50/70
$\langle 125, 155 \rangle$	14	14/70	64	64/70
$\langle 155, 185 \rangle$	4	4/70	68	68/70
$\langle 185, 215 \rangle$	2	2/70	70	1





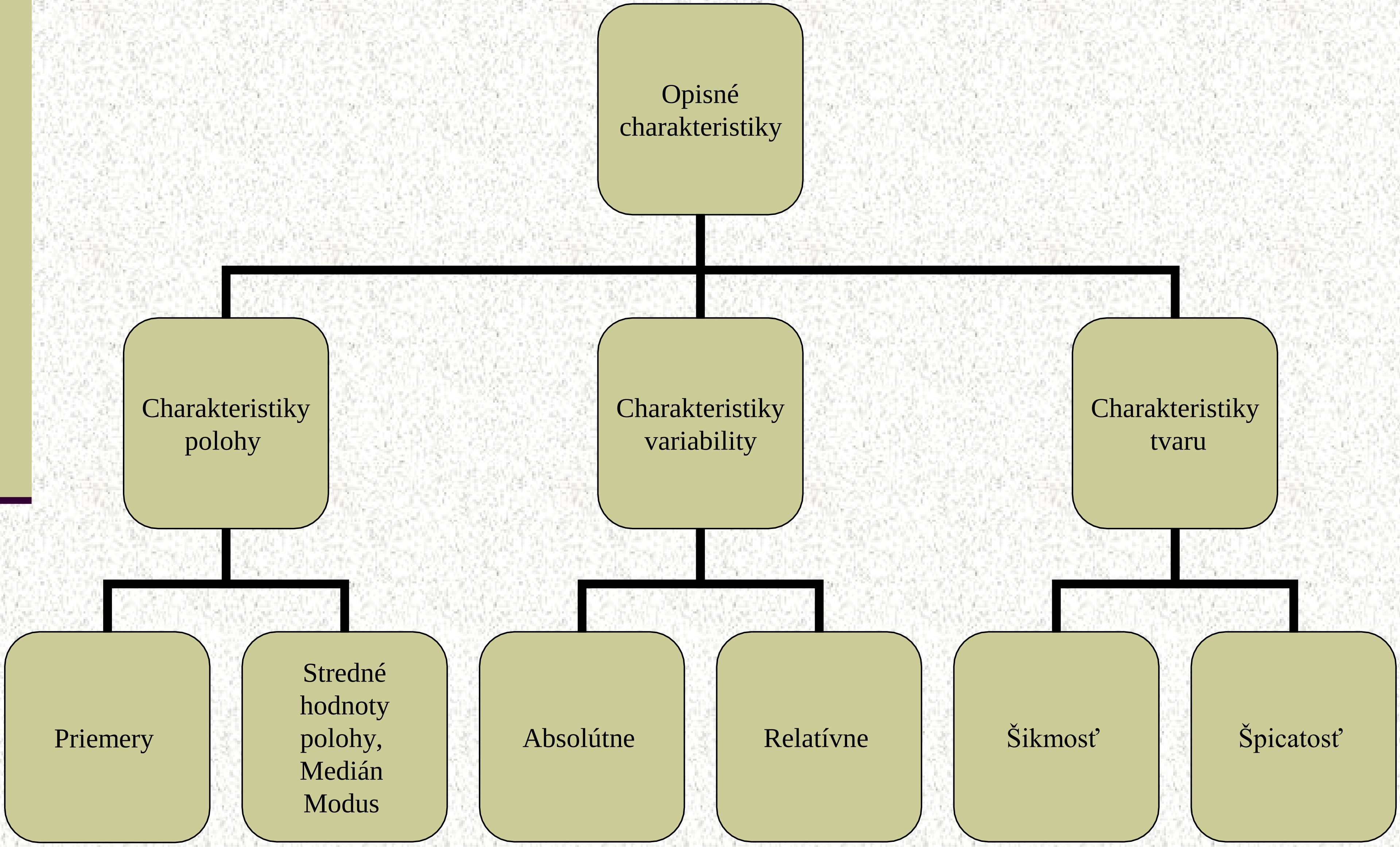
Opisné charakteristiky – Základné číselné charakteristiky dátového súboru

Opisná charakteristika predstavuje číslo vypočítané podľa príslušného vzorca zo štatistického súboru. Cieľom opisných charakteristík je charakterizovať súbor.

Opisné charakteristiky sa delia na tri skupiny:

- a) Miery polohy (stredná hodnota, priemer, kvantily)
- b) Miery variability (rozptyl, sm. odchýlka)
- c) Miery tvaru (šikmosť, špicatosť)

Opisné charakteristiky



Charakteristiky polohy – stredné hodnoty

1. **priemery** - aritmetický
   - geometrický
   - harmonický

každý môže byť jednoduchý alebo vážený

2. **ostatné stredné hodnoty** - modus
   - medián

Xmin <= Xstr<= Xmax

Stredná hodnota

Miery polohy - priemer

- **Aritmetický priemer** je najjednoduchším a súčasne najdôležitejším typom strednej hodnoty. Zohľadňuje rovnocennosť všetkých spracovávaných dát, ale je veľmi citlivý, nerezistentný voči omylom a môže prudko reagovať na jediný vychýlený údaj vo výbere, pretože je jeho ťažiskom. Pre stručnosť sa v ďalších kapitolách bude nazývať strednou hodnotou..

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{x} = \frac{x_1.n_1 + x_2.n_2 + \ldots + x_k.n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} x_j.n_j}{\sum_{i=1}^{k} n_j} = \frac{\sum_{i=1}^{k} x_j.n_j}{n}$$

$$\bar{x} = \sum_{j=1}^{k} x_j . f_j$$

Príklad – vážený aritmetický priemer

Vypočítajme priemerný obrat predajne

$$\bar{x} = \frac{95.1 + 125.3 + \ldots + 365.1}{1 + 3 + \ldots + 1} = 223$$

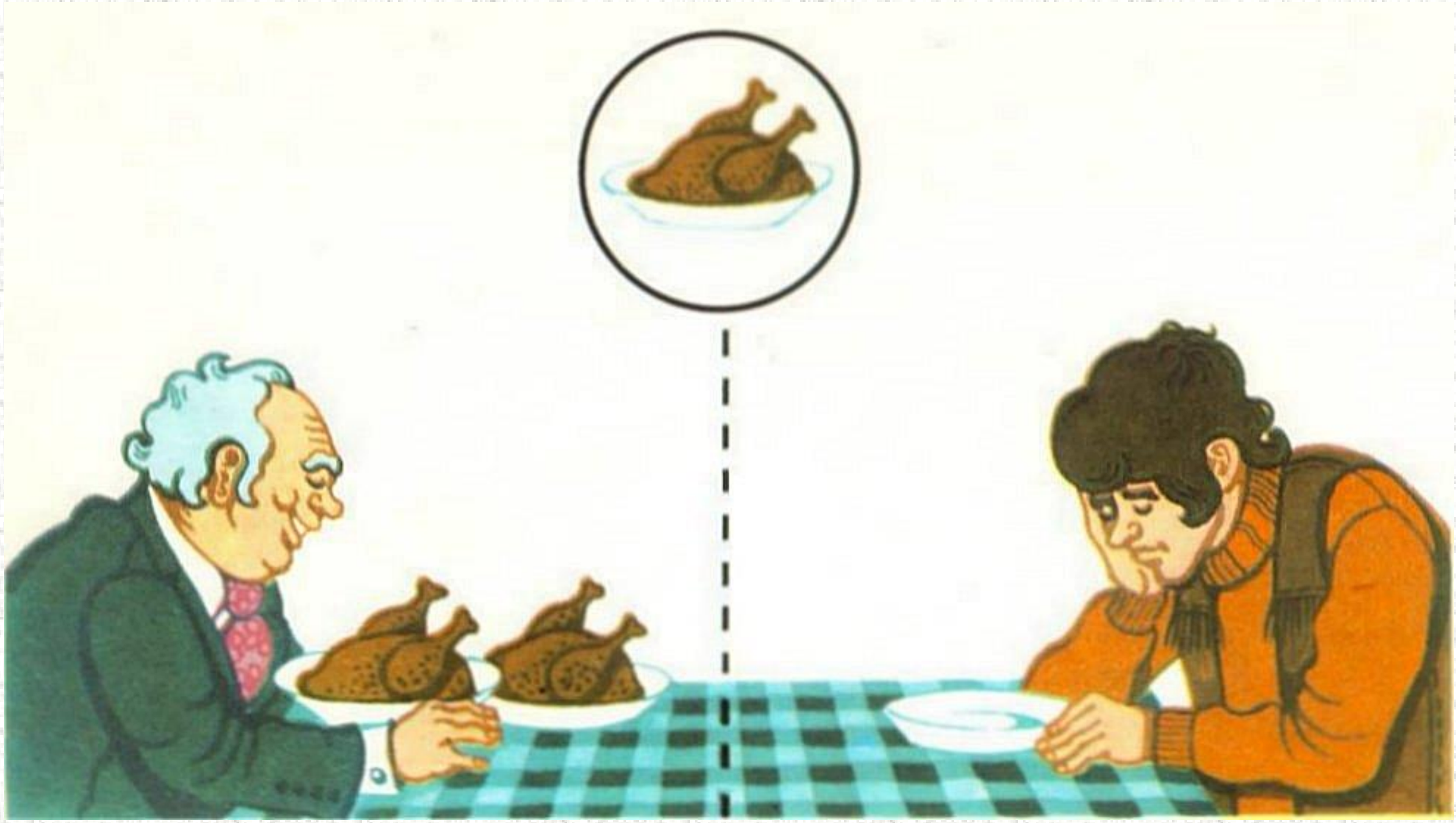
$$\bar{x} = 95.0{,}0217 + 125.0{,}0652 + \ldots + 365.0.0217 = 223$$

Trieda	x_d	x_h	x_j	n_j	f_j	N_j	F_j
1	-	110	95	1	0,0217	1	0,0217
2	110	140	125	3	0,0652	4	0,0869
3	140	170	155	5	0,1087	9	0,1956
4	170	200	185	8	0,1739	17	0,3695
5	200	230	215	9	0,1957	26	0,5652
6	230	260	245	8	0,1739	34	0,7391
7	260	290	275	5	0,1087	39	0,8478
8	290	320	305	4	0,087	43	0,9348
9	320	350	335	2	0,0435	45	0,9783
	350	-	365	1	0,0217	46	1

Príklad: Aká je priemerná známka zo Štatistiky

znám.	p.št.	xj .nj	fj	xj.fj
1	12	12	0,324	0,3243
2	16	32	0,432	0,8649
3	9	27	0,243	0,7297
spolu	37	71		1,92

Priemerna známka bude: $\bar{x} = 1.92$



Priemer predstavuje často rovnomernosť alebo normu, ktorá neexistuje. Keď v priemere každý zje hus, je možné, že niektorí zjedia dve, resp. viac, iní žiadnu.

Aritmetický priemer nemá väčšinou žiadny odraz v skutočnosti. Každá priemerná rodina má 2,2 dieťaťa, našťastie to neznamená to, čo vidíme na obrázku.

Harmonický priemer (jednoduchý)- podiel počtu pozorovaní a súčtu prevrátených (inverzných) hodnôt x_i znaku.

Jednoduchý harmonický priemer

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

Vážený harmonický priemer

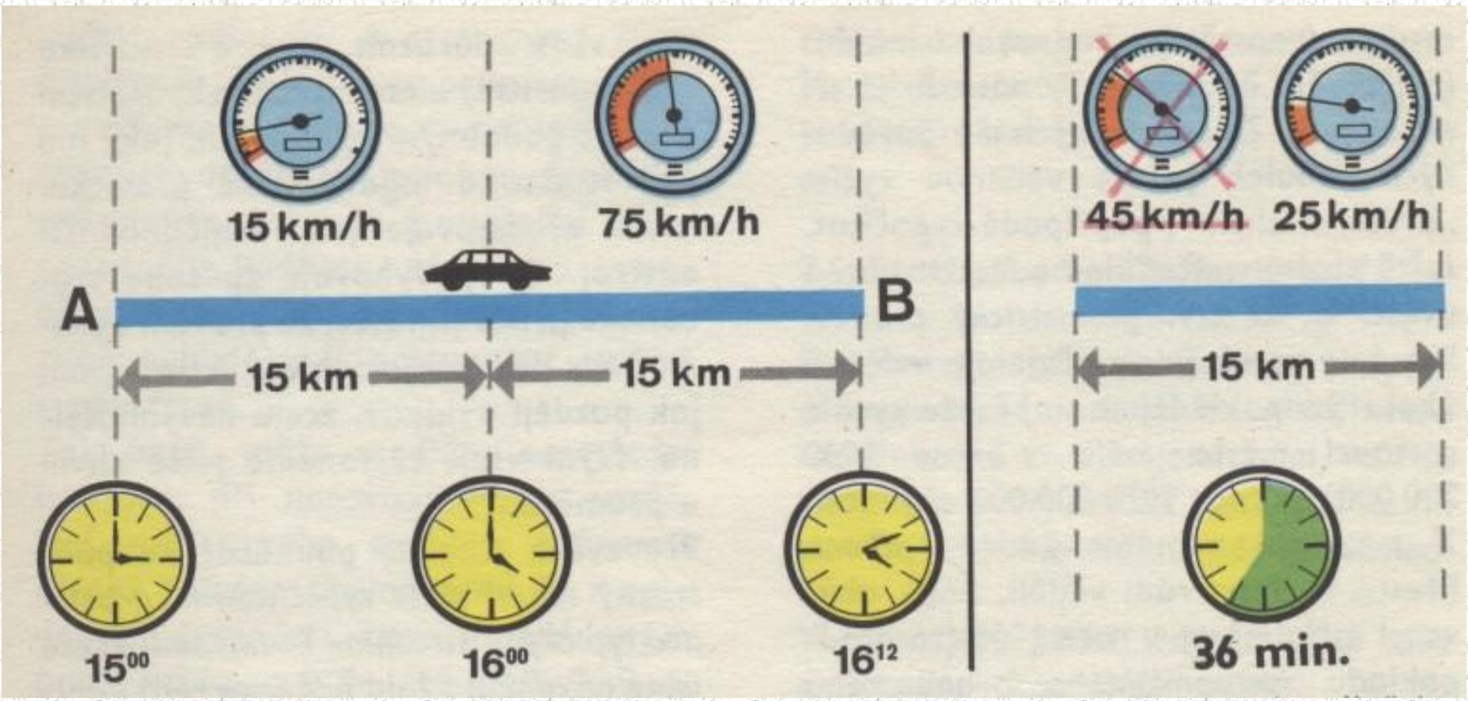
$$\bar{x}_h = \frac{\sum n_i}{\sum \frac{n_i}{x_i}}$$

Charakteristika harmonického priemeru

- Je odolný voči extrémnym hodnotám.
- Využíva sa pri stanovení priemernej veľkosti výberového súboru vychádzajúc z viacerých výberov.
- Ak hodnoty obsahujú určitú dôležitosť- váhu, používame vážený priemer.
- V praxi sa používa pri výpočte priemeru z indexov, pomerných čísel.
- Prevrátená hodnota HP je aritmetický priemer prevrátených hodnôt x.

Úloha:

Predpokladajme, že ideme 30 km ďaleko a prvých 15 km prejdeme rýchlosťou 15 km za hod. a druhých 15 km rýchlosťou 75 km za hod. Akú priemernú rýchlosť sme dosiahli za hodinu?



Prvú trať ideme rýchlosťou 15km/hod… k jej prejdeniu potrebujeme práve 1hod. - 60 minút (15/15*60)
Druhú trať (15 km) ideme rýchlosťou 75 km/hod….
K jej prejdeniu potrebujeme len 12 minút (15/75*60)
→ celková doba jazdy je teda 72 minút.
→ Aritmetický priemer
Nás zmýli výslekom (15+75)/2=45km za hodinu.
K zisteniu priemernej doby jazdy pre oba úseky potrebujeme 60min+12min= 72/2 = 36 minút pre každý úsek jazdy, čo predstavuje priemernú rýchlosť 25 km / hod.

$$\bar{x} = \frac{n}{\sum(\frac{1}{x_j})} = \frac{2}{\frac{1}{15} + \frac{1}{75}} = \frac{2}{\frac{6}{75}} = \frac{150}{6} = 25$$

Auto prešlo 50 km. Prvých 10 km išlo rýchlosťou 50 km/h, ďalších 10 km rýchlosťou 80 km/h, potom 10 km rýchlosťou 90 km/h, ďalších 10 km rýchlosťou 60 km/h a posledných 10 km rýchlosťou 80 km/h. Akou priemernou rýchlosťou išlo auto?

Geometrický priemer (Geomean)

používa sa pri časových radoch -rast objemu produkcie

jednoduchý

$$\overline{x_g} = \sqrt[n]{x_1 . x_2 . \dots . x_n} = \sqrt[n]{\prod_{j=1}^{n} x_j}$$

vážený

$$\overline{x_g} = \sqrt[n]{x_1^{n} . x_2^{n} . \dots . x_m^{n_m}} = \sqrt[n]{\prod_{i=1}^{m} x_i^{n_i}}$$

Charakteristika GP

- GP predstavuje n-tú odmocninu zo súčinu hodnôt znaku X.
- Znižuje vplyv extrémnych hodnôt súboru dát.
- Ak hodnota súboru dát bude rovná nule, GP bude tiež rovný nule – nevýhoda GP.
- GP sa nepočíta ak je hodnota súboru menšia ako nula.
- V praxi sa využíva pri stanovení priemernej hodnoty koeficientov rastu...
- Pre priemery tých istých kladných hodnôt by mala platiť podmienka:

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x}$$

Vývoj HNP SR za rr.95-99 v US$ na obyv a rok.

Rok	GNPSR (US$)	koeficient rastu	koeficient prirastku v %	tempo rastu	tempo prirastku v %
1995	3110				
1996	3570	1.148	114.79	0.15	14.79
1997	3860	1.081	108.12	0.08	8.12
1998	3870	1.003	100.26	0.00	0.26
1999	3770	0.974	97.42	-0.03	-2.58

V roku 1997 oproti r. 96 vzrástol HNP na obyv. na 108,12%

V roku 1997 oproti r. 96 vzrástol HNP na obyv. o 8,12%

Z jednotlivých koeficientov rastu možno vypočítať:

priemerný koeficient rastu

$$\bar{k} = \sqrt[T-1]{k_1 . k_2 \; ....k_{T-1}}$$

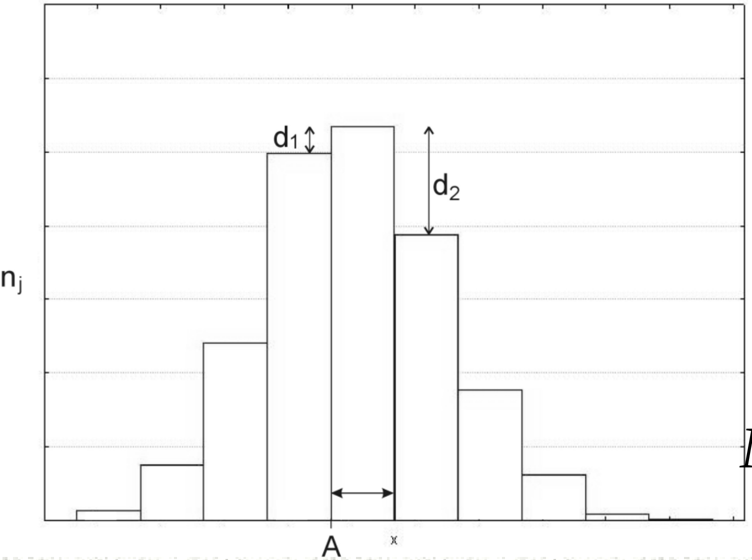
$$\mathbf{\bar{k} = \sqrt[4]{(1,148.1,081.\ 1,003\ .\ 0,974)} = 1.0493}$$

Za obdobie rr. 95-99 HNP v SR rástol ročne približne o 4,9%

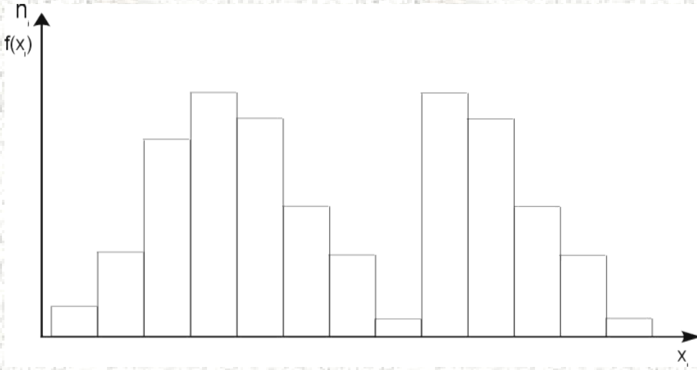
Miery polohy - Modus

- Charakteristikou polohy , tj. najpočetnejší variant, alebo stred najpočetnejšieho intervalu. Modus je tá hodnota štatistického znaku, ktorá sa v príslušnom štatistickom súbore vyskytuje najčastejšie. Označuje sa viacerými spôsobmi napr. M_o, x_{mod}, alebo . Modus je vždy robustný, nie je citlivý na odľahlé hodnoty vo výberovom súbore dát.

- A – počiatok modálneho intervalu,
- c – šírka (veľkosť) intervalu,
- d_1 – rozdiel medzi početnosťou modálneho a predchádzajúceho intervalu,
- d_2 – rozdiel medzi početnosťou modálneho a nasledujúceho intervalu.



$$M_o = A + c.\frac{d_1}{d_1 + d_2}$$



$$M_o = A + c.\frac{d_1}{d_1 + d_2} = 95 + 30.\frac{27-16}{(27-16)+(27-14)} = 108{,}75$$

Trieda	x_d	x_h	x_j	n_j	f_j	N_j	F_j
1	30,4	30,6	30,5	2	2/100	2	2/100
2	30,6	30,8	30,7	4	4/100	6	6/100
3	30,8	31,0	30,9	5	5/100	11	11/100
4	31,0	31,2	31,1	10	10/100	21	21/100
5	31,2	31,4	31,3	27	27/100	48	48/100
6	31,4	31,6	31,5	29	29/100	77	77/100
7	31,6	31,8	31,7	13	13/100	90	90/100
8	31,8	32,0	31,9	4	4/100	94	94/100
9	32,0	32,2	32,1	5	4/100	99	99/100
10	32,2	32,4	32,3	0	0/100	99	99/100
11	32,4	32,6	32,5	1	1/100	100	1

30,4	30,6	30,7	30,8	30,8	30,8	30,9	30,9	31	31
31	31,1	31,1	31,2	31,2	31,2	31,2	31,2	31,2	31,2
31,2	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3
31,3	31,3	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4
31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,6	31,6	31,6	31,6
31,6	31,6	31,6	31,6	31,6	31,6	31,6	31,7	31,7	31,7
31,7	31,7	31,7	31,7	31,7	31,8	31,8	31,8	31,8	31,8
31,9	31,9	31,9	32	32,1	32,1	32,1	32,2	32,2	32,5

Medián $\tilde{x}$

Hodnota, ktorá rozdeľuje súbor vzostupne usporiadaných údajov na dve rovnako početné časti.

Medián je poradovým ukazovateľom stredu – reprezentuje hodnotu, ktorá sa nachádza uprostred súboru údajov.

Polovica hodnôt súboru je menšia ako medián a polovica hodnôt je väčšia ako medián.

Hodnoty v súbore sa musia vždy usporiadať vzostupne.

Medián nie je citlivý na extrémne hodnoty.

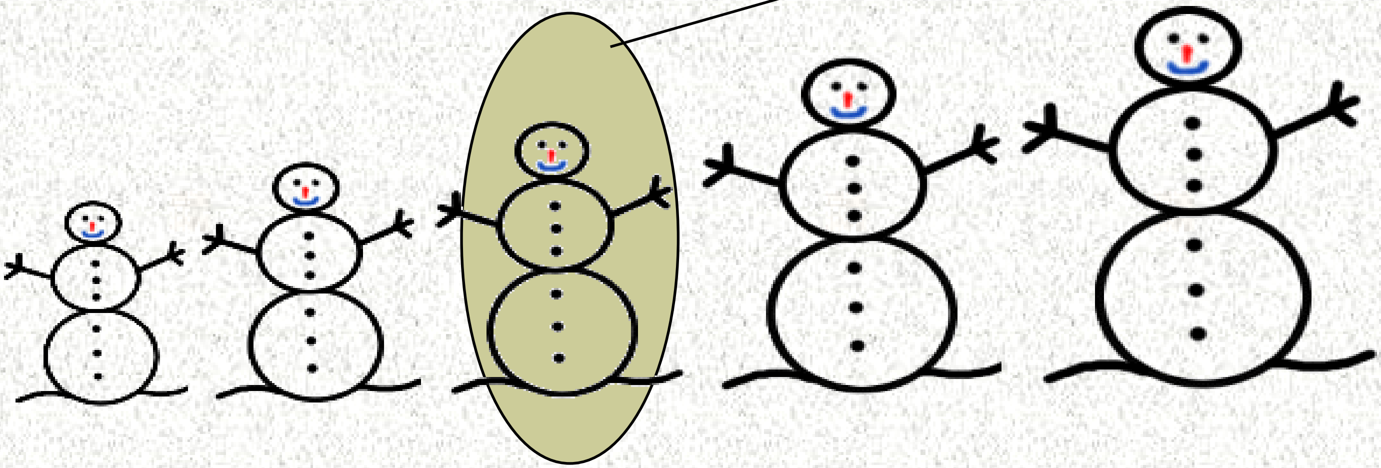
Medián – nepárny počet hodnôt

a) určovanie poradia mediánu v štatistickom súbore, v ktorom je nepárny počet štatistických jednotiek (r)

$$r_{\tilde{x}} = \frac{n+1}{2}$$

b) Určenie hodnoty mediánu $\tilde{x}$

$$\tilde{x} = x_{\frac{n+1}{2}}$$

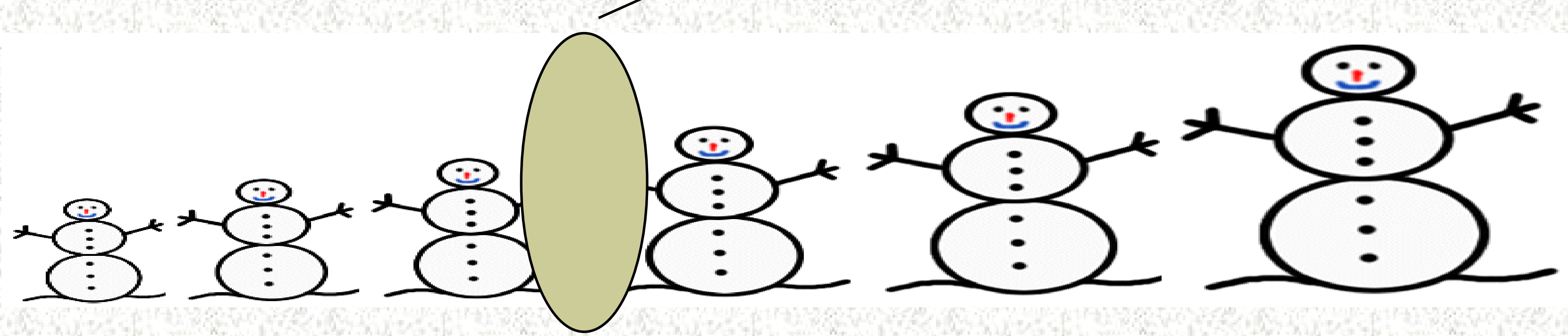


Medián – párny počet hodnôt

- určovanie poradia mediánu
- Určenie hodnoty mediánu ako aritmetický priemer dvoch hodnôt určených na základe poradia

$$r_{\tilde{x}} = \frac{n}{2}$$

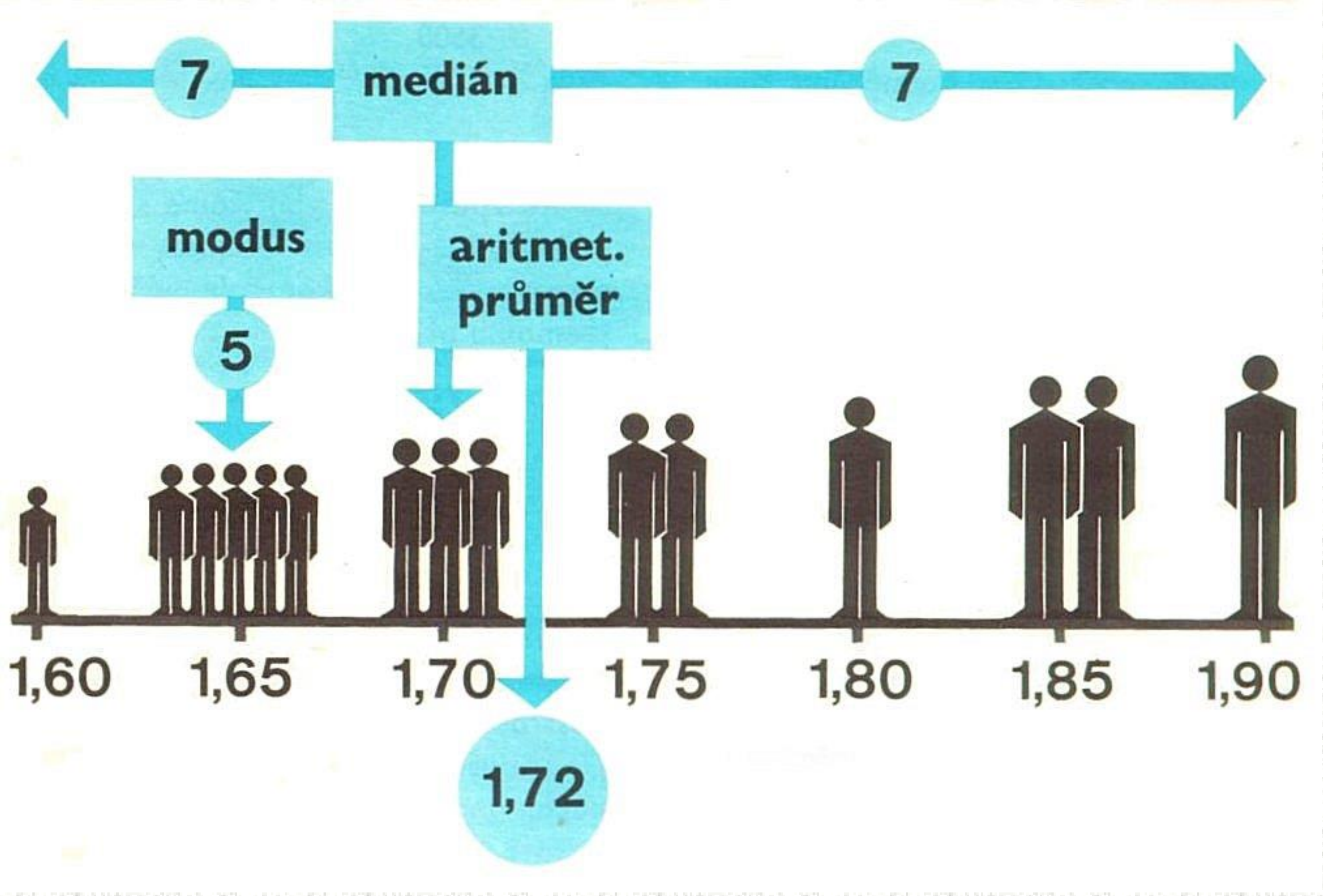
$$\tilde{x} = \left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}}\right) / 2$$



Trieda	x_d	x_h	x_j	n_j	f_j	N_j	F_j
1	30,4	30,6	30,5	2	2/100	2	2/100
2	30,6	30,8	30,7	4	4/100	6	6/100
3	30,8	31,0	30,9	5	5/100	11	11/100
4	31,0	31,2	31,1	10	10/100	21	21/100
5	31,2	31,4	31,3	27	27/100	48	48/100
6	31,4	31,6	31,5	29	29/100	77	77/100
7	31,6	31,8	31,7	13	13/100	90	90/100
8	31,8	32,0	31,9	4	4/100	94	94/100
9	32,0	32,2	32,1	5	4/100	99	99/100
10	32,2	32,4	32,3	0	0/100	99	99/100
11	32,4	32,6	32,5	1	1/100	100	1

30,4	30,6	30,7	30,8	30,8	30,8	30,9	30,9	31	31
31	31,1	31,1	31,2	31,2	31,2	31,2	31,2	31,2	31,2
31,2	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3	31,3
31,3	31,3	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4
31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,4	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5	31,5
31,5	31,5	31,5	31,5	31,5	31,5	31,6	31,6	31,6	31,6
31,6	31,6	31,6	31,6	31,6	31,6	31,6	31,7	31,7	31,7
31,7	31,7	31,7	31,7	31,7	31,8	31,8	31,8	31,8	31,8
31,9	31,9	31,9	32	32,1	32,1	32,1	32,2	32,2	32,5

Porovnanie modusu, mediánu a strednej hodnoty



Miery polohy - Kvantily

- Charakteristikou polohy je α-kvantil. Ak je $\alpha \in (0;1)$, potom α-kvantil x_α je číslo, ktoré delí usporiadaný dátový súbor na dve časti. Spodný úsek, obsahujúci aspoň α % všetkých dát a na horný úsek obsahujúci aspoň (1 – α) % všetkých dát.
- $x_{0,50}$ – medián M
- $x_{0,25}$ – dolný kvartil F_D
- $x_{0,75}$ – horný kvartil, F_H
- $x_{0,125}$ – dolný oktil E_D
- $x_{0,875}$ – horný oktil E_H
- $x_{0,1}$ – dolný decil,
- $x_{0,9}$ – horný decil,
- $x_{0,01}$ – dolný percentil.
- $x_{0,99}$ – horný percentil.

Miery variability – kvartilová odchýlka

- Charakteristikou variability je **kvartilová odchýlka**: $q = x_{0,75} - x_{0,25}$. Kvartilová odchýlka, alebo aj **medzikvartilové rozpätie** špecifikuje ako sa v intervale vyskytuje 50 % dát.
- **Pr.**: V priebehu semestra sa študenti podrobili písomnému testu z matematiky, v ktorom bolo možné získať 0 až 10 bodov. Výsledky sú uvedené v tabuľke:

Počet bodov	0	1	2	3	4	5	6	7	8	9	10
Počet študentov	1	4	6	7	11	15	19	17	12	6	3

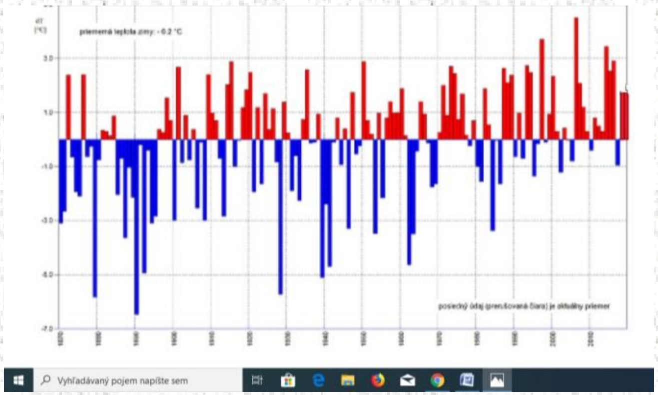
α	$n\alpha$	c	$x_\alpha = x_{(c)}$
0,50	50,5	51	6
0,10	10,1	11	2
0,90	90,9	91	8
0,25	25,25	26	4
0,75	75,75	76	7

- Dvaja strelci strieľajú do terča a zásahy si zapisujú.
- Prvý strelec zasiahol 7, 8 a 9, druhý strelec 1, 10 a 13. Vypočítajte priemerné odchýlky počtu nastrieľaných bodov oboch strelcov.

Miery variability - Rozptyl

- Momentové charakteristiky rozptylu slúžia ako odhad variability základného súboru. Rozptylom (disperziou) s^2 (σ^2) sa nazýva stredná hodnota štvorcov odchýlok všetkých hodnôt náhodnej veličiny od strednej hodnoty. Vyjadruje rozptyl hodnôt okolo strednej hodnoty. Čím je rozptyl väčší, tým sa údaje viac odchyľujú od priemeru

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



$$\sigma^2 = \frac{\sum_{j=1}^{k} n_j (x_j - \bar{x})^2}{n}$$

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

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

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

$$\sqrt{s^2} \qquad \sqrt{\sigma^2}$$

Miery variability – Koeficient variácie

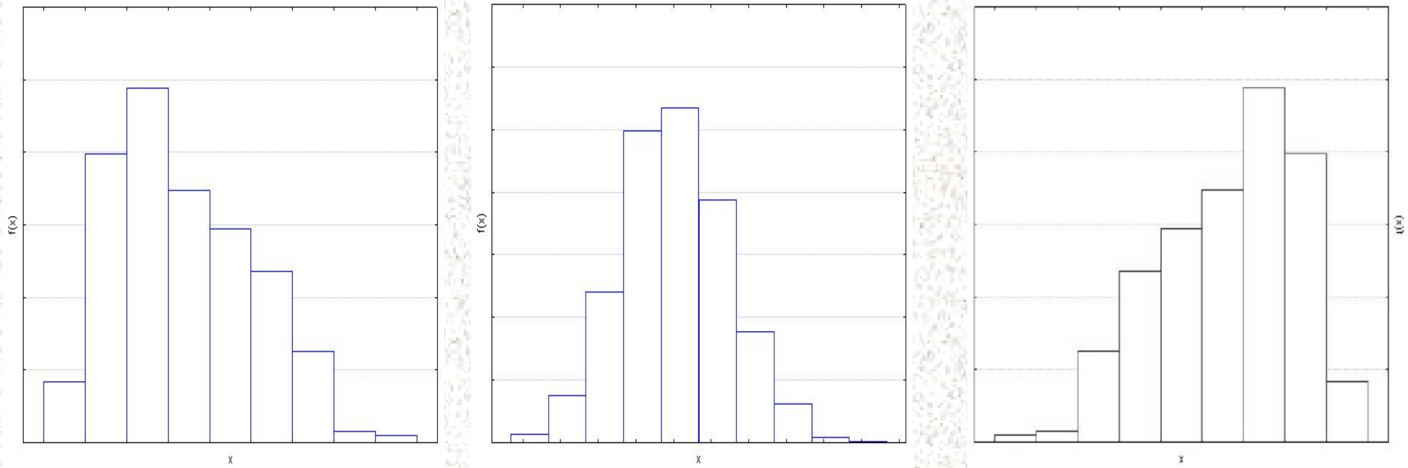
- V prípade pomerových znakov sa ako charakteristika variability používa koeficient variácie:
- **Koeficient variácie** (variačný koeficient) je relatívnou mierou variability. Je to pomer smerodajnej odchýlky ku priemeru vyjadrený v percentách. Vzhľadom na to je jeho použitie vhodné len vtedy, ak je priemer kladný a dostatočne väčší ako nula. Variabilita do 5 % sa považuje za malú, do 20 % za strednú a nad 20 % za veľkú.

$$V = \frac{s}{\bar{x}}.100\%$$

Miery tvaru - Šikmosť

- **Šikmosť** vyjadruje nesúmernosť rozdelenia početností okolo priemeru. Rozdelenie početností v súbore môže byť súmerné – symetrické, alebo nesúmerné – zošikmené doľava alebo doprava.
- Miera šikmosti sa bežne vyjadruje koeficientom šikmosti A
- Ak A = 0, rozdelenie početností okolo priemeru je symetrické. Ak A > 0 $\Rightarrow$ ľavostranná asymetria. Ak A < 0 $\Rightarrow$ pravostranná asymetria

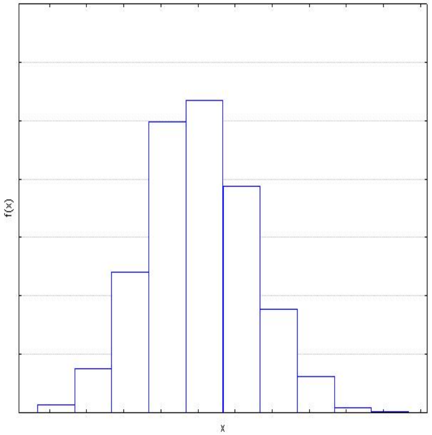
$$A = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^3}{n.s^3}$$

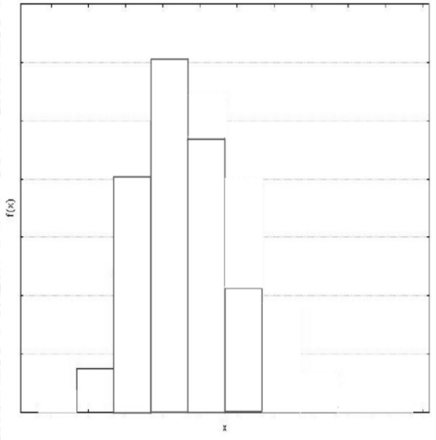


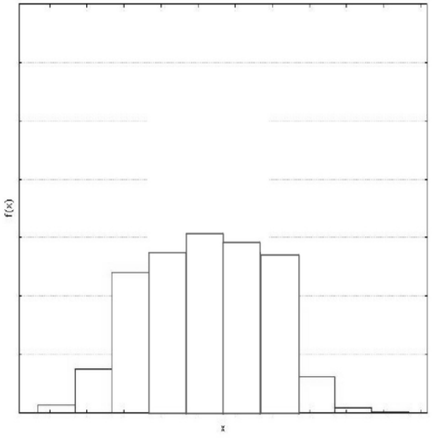
Miery tvaru - Špicatosť

- Špicatosť vyjadruje koncentráciu rozdelenia početností okolo hodnoty znaku. Čím sú početnosti sústredené viac v okolí nejakej hodnoty znaku, tým má rozdelenie početností výraznejší vrchol (je špicatejší). Špicatosť sa v štatistickom súbore porovnáva s tzv. normálnym rozdelením početností (Laplaceovo – Gaussovo rozdelenie).
- Mierou vyjadrenia špicatosti je koeficient špicatosti E

$$E = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^4}{n.s^4}$$







Histogram

- Vlastnú variabilitu hodnôt je možné vyjadriť minimálnou a maximálnou hodnotou. Pri pohľade na histogram je možné povedať, že hodnoty sa menia od ... do... .
- rozdelenie symetrické alebo asymetrické. Miera zošikmenia na jednu alebo druhú stranu upozorňuje užívateľa, že menšie (väčšie) hodnoty sa vyskytujú častejšie v určitom intervale na jednej strane rozdelenia a na druhej strane väčšie (menšie) hodnoty sa vyskytujú zriedkavejšie v širšom intervale.
- Dôležitá je aj špicatosť rozdelenia. Čím je rozdelenie špicatejšie, tým je variabilita nižšia a, naopak, plochejšie rozdelenia majú väčšiu variabilitu ako špicatejšie.