evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

STATISTICKÉ CHARAKTERISTIKY

Vytvořeno s podporou projektu Průřezová inovace studijních programů Lesnické a dřevařské fakulty MENDELU v Brně (LDF) s ohledem na disciplíny společného základu (reg. č. CZ.1.07/2.2.00/28.0021) za přispění finančních prostředků EU a státního rozpočtu České republiky.

DATA → INFORMACE

Statistická analýza je založena na **zhušťování informace** – tj. jak z co nejmenšího množství vhodně zvolených údajů vytěžit maximum relevantních informací (tj. informací, které řeší studovaný praktický problém, odpovídají na položené otázky, hypotézy).

1. **prvotní zápis** – naprosto neuspořádaná data, údaje v té podobě, a v tom pořadí jak jsou naměřeny – většinou nemůžeme postřehnout žádné společné podstatné vlastnosti
2. **tříděný soubor** – jednotlivá měřená data jsou tříděna do tříd, místo všech původních dat používáme třídní reprezentanty a počty hodnot ve třídách – dnes se příliš nepoužívají, účelem třídění bylo především zjednodušení výpočtů, ale také alespoň částečně zpřehledňují data – podrobněji **teorie text I, str. 16 - 23**
3. **statistické charakteristiky** – speciální veličiny, které **podávají koncentrovanou formou informaci o podstatných statistických vlastnostech** studovaného souboru

STATISTICKÉ CHARAKTERISTIKY

statistické charakteristiky – speciální veličiny, které **podávají koncentrovanou formou informaci o podstatných statistických vlastnostech** studovaného souboru. **Správně zvolené a správným způsobem vypočítané charakteristiky** (především musí být dodrženy podmínky jejich platnosti) **obsahují v rámci jednoho nebo několika málo čísel veškerou informaci o podstatných statistických vlastnostech studovaného souboru**, která je obsažena v původních datech, tj. v prvotním zápisu.

Jsou založeny na dvou odlišných principech stanovení:

- charakteristiky **momentové**
- charakteristiky **kvantilové**

MOMENTOVÉ CHARAKTERISTIKY

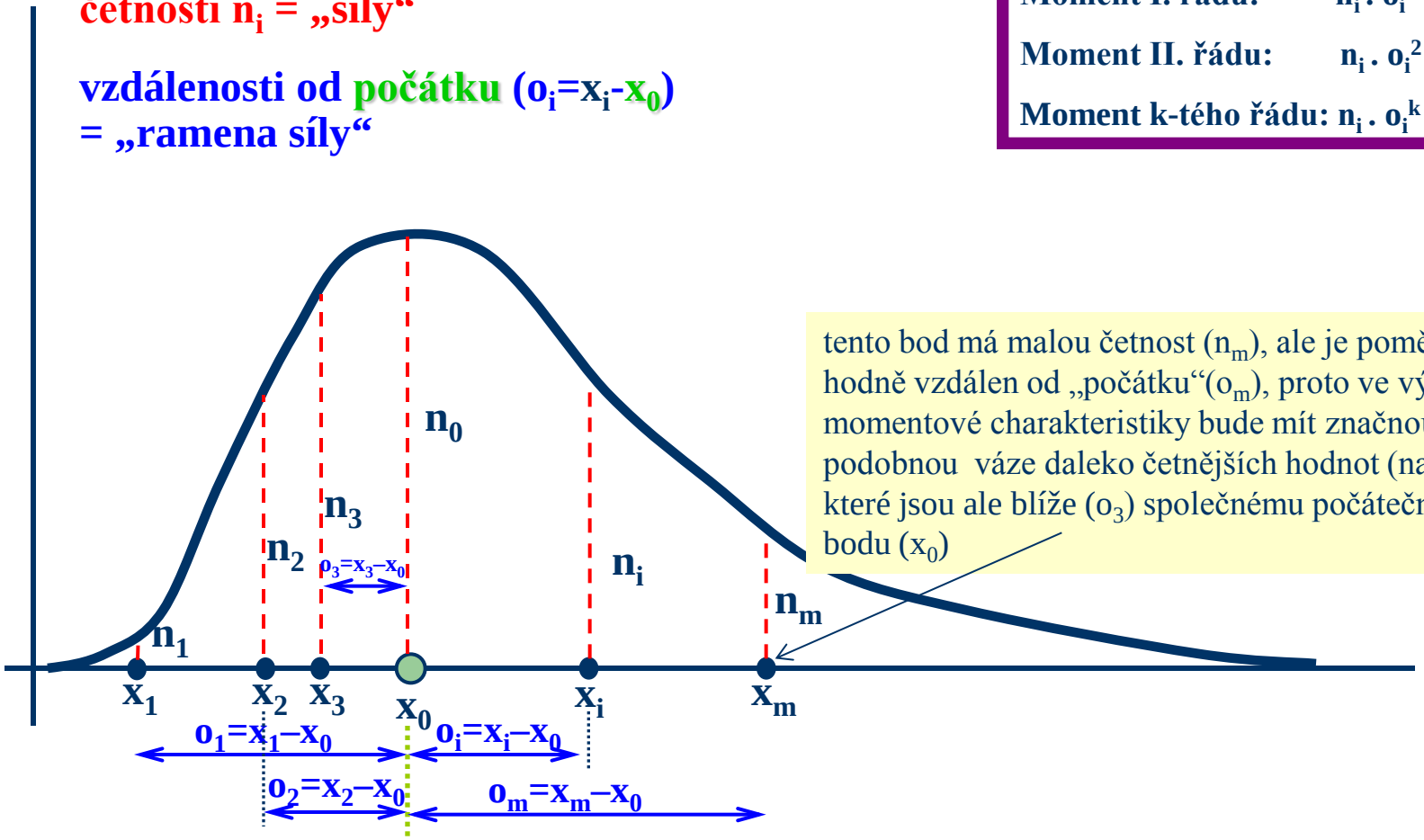
Jsou založeny na principu „statistických momentů“. Vycházíme z analogie fyzikálních momentů, např. moment síly jako součin síly a jejího ramene. Ve statistické analogii je „silou“ četnost určité hodnoty, „ramenem“ potom vzdálenost této hodnoty od určitého bodu (např. nuly, průměru nebo libovolného bodu na číslelné ose). Potom na výpočet příslušné charakteristiky mají větší vliv hodnoty, které mají vyšší „sílu“, tj. četnost nebo které mají velké „rameno síly“, tj. jsou více vzdálené od společného počátečního bodu.

MOMENTOVÉ CHARAKTERISTIKY

četnosti n_i = „síly“

vzdálenosti od počátku ($o_i = x_i - x_0$) = „ramena síly“

Moment I. řádu:	$n_i \cdot o_i$
Moment II. řádu:	$n_i \cdot o_i^2$
Moment k-tého řádu:	$n_i \cdot o_i^k$



tento bod má malou četnost (n_m), ale je poměrně hodně vzdálen od „počátku“(o_m), proto ve výpočtu momentové charakteristiky bude mít značnou váhu, podobnou váze daleko četnějších hodnot (např. n_3), které jsou ale blíže (o_3) společnému počátečnímu bodu (x_0)

MOMENTOVÉ CHARAKTERISTIKY

Statistický moment *k-tého* řádu je aritmetický průměr všech momentů *k-tého* řádu (pro všechna x_i) vztažených k hodnotě x_0.

Podle polohy bodu x_0 rozeznáváme statistické momenty:

1. **Všeobecné** ($x_0 = 0$)

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} n_i \cdot \underbrace{\left(x_i - 0\right)}_{o_i = x_i - x_0}^{k=1} = \frac{1}{n}\sum_{i=1}^{n} n_i \cdot x_k$$

Aritm.průměr

2. **Centrální** ($x_0 = \bar{x}$)

$$m_k = \frac{1}{n}\sum_{i=1}^{n} n_i \cdot \left(x_i - \bar{x}\right)^k \leftarrow$$

k=2 – rozptyl
k=3 – koef.nesouměrnosti
k=3 – koef. špičatosti

MOMENTOVÉ CHARAKTERISTIKY

Aritmetický průměr	$= m'_1$	**všeobecný moment 1.řádu)**
Rozptyl	$= m_2$	**centrální moment**
Koeficient nesouměrnosti	$= m_3/(m_2^{3/2}) = m_3/s^3$	**centrální moment**
Koeficient špičatosti	$= m_4/(m_2^2) = m_3/s^4$	**centrální moment**

MOMENTOVÉ CHARAKTERISTIKY

Vlastnosti momentových charakteristik:

- jsou **vypočítány ze všech hodnot souboru** (z toho vyplývá, že obsahují úplnou statistickou informaci, a proto **se používají jako nejlepší charakteristiky prioritně, pokud jsou splněny níže uvedené podmínky),**
- **nejsou vhodné pro soubory s extrémními hodnotami**
- rozdělení hodnot souboru musí odpovídat **normálnímu (Gaussovu) rozdělení** (viz prezentace „rozdělení“ nebo **teorie text I, str. 71-77)**
- **nejsou vhodné pro velmi malé soubory**

KVANTILOVÉ CHARAKTERISTIKY

Kvantil je hodnota **určitým způsobem v souboru umístěná.** Zpravidla je určena svým **pořadím** ve vzestupně uspořádaném souboru a **leží pod ní (100.p) % hodnot** souboru. Hodnota p se pohybuje mezi 0 a 1.

Pořadí kvantilu se určí: $$i = \frac{N+1}{p} \cdot r$$

N	rozsah souboru
p	počet skupin dělení.
r	pořadí kvantilu

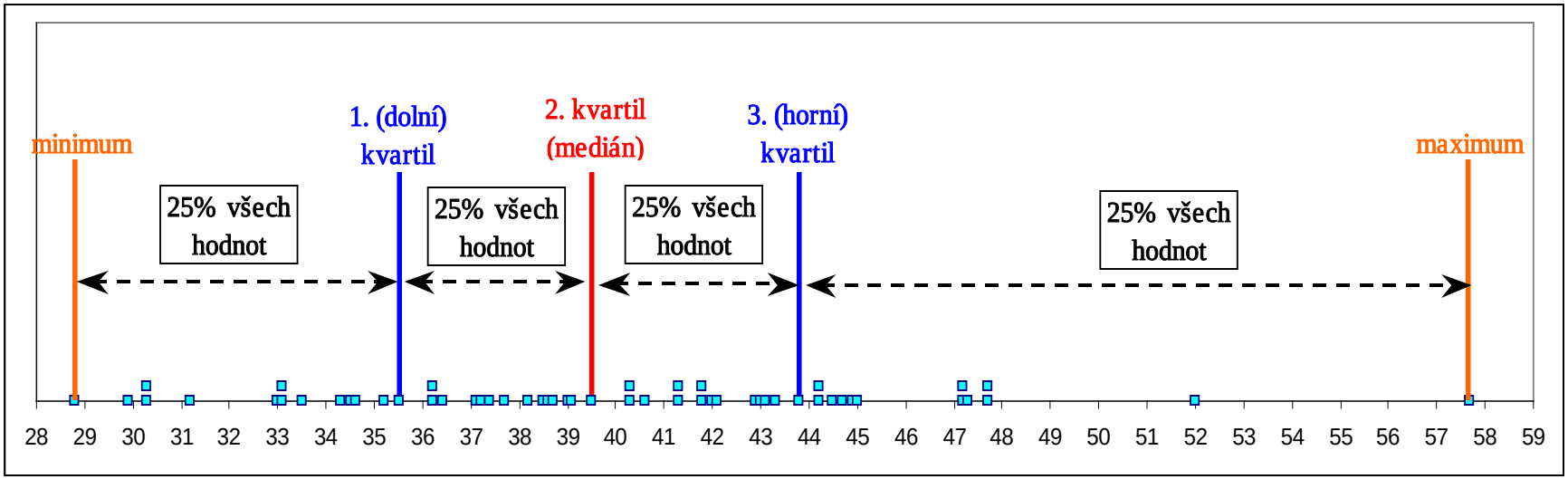
KVANTILOVÉ CHARAKTERISTIKY

Důležité kvantily:

25% kvantil – **dolní kvartil**
50% kvantil – **medián**
75% kvantil – **horní kvartil**

Další používané kvantily:

10% kvantil – decil
12,5% kvantil – oktil
6,25 % kvantil - sedecil



KVANTILOVÉ CHARAKTERISTIKY

Výhody kvantilových charakteristik:

- nejsou ovlivněny extrémními hodnotami
- jsou vhodné i pro malé soubory
- nezávisí na rozdělení veličiny
- jsou snadno zjistitelné a interpretovatelné

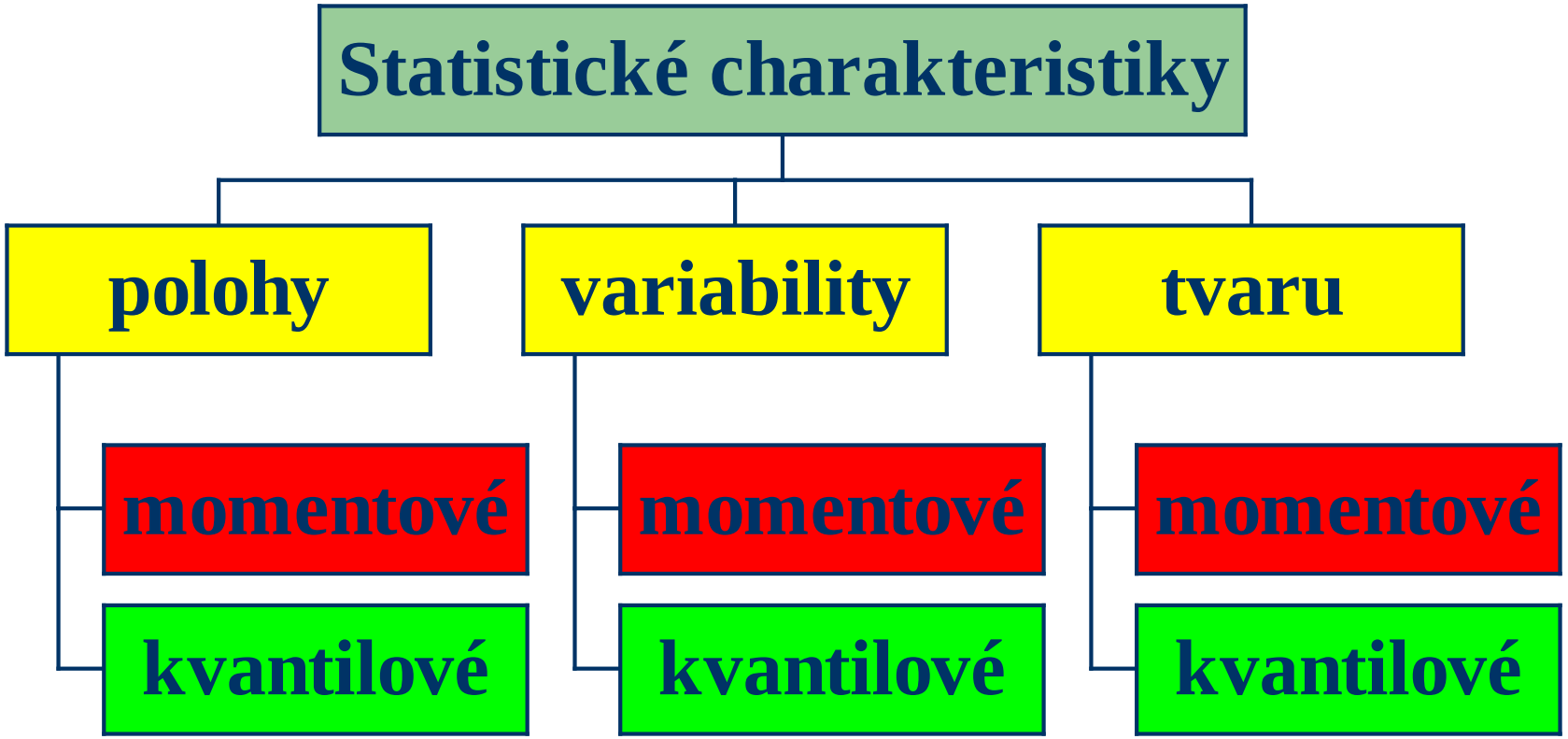
Nevýhody kvantilových charakteristik:

- nevycházejí ze všech hodnot souborů, pouze z hodnot určitého pořadí
- nelze s nimi provádět matematické operace v plném rozsahu
- nevypovídají o některých zvláštnostech statistických souborů (např. extrémy)

KVANTILOVÉ CHARAKTERISTIKY

Kvantilové charakteristiky se **používají tehdy, pokud nejsou splněny** **podmínky momentových charakteristik**, tj. pro soubory s výraznými extrémy, se silně nenormálním rozdělením dat nebo pro velmi malé soubory (a samozřejmě tím více pro jakoukoli kombinaci těchto podmínek)

STATISTICKÉ CHARAKTERISTIKY



STATISTICKÉ CHARAKTERISTIKY

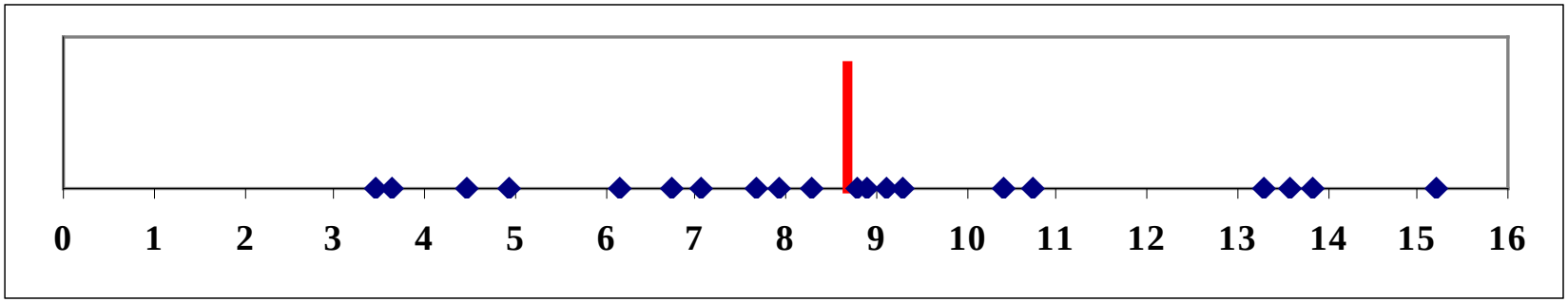
podrobněji viz **teorie text I, kap. 4 – str. 24 - 48**

Pamatujte, že **pro správné statistické zhodnocení** jakéhokoliv souboru **je nutné použít charakteristiky všech tří skupin** – polohy, variability a tvaru – protože každá z nich popisuje soubor z jiného hlediska. Je tedy zcela nesprávné používat např. „izolovaně“ jen aritmetický průměr bez dalších údajů o souboru, který reprezentuje (např. údaje v médiích o „průměrných platech“ nemají prakticky žádnou vypovídací schopnost, viz např. srovnání průměrů a jednotlivých kvantilů platů http://user.mendelu.cz/drapela/Statisticke_metody/Prezentace/ soubor „Prumerne_platy.xls“) – viz např. srovnání „průměrů“ a „mediánu“ platů – o významu jejich srovnání viz následující snímky.

STATISTICKÉ CHARAKTERISTIKY

Typy charakteristik:

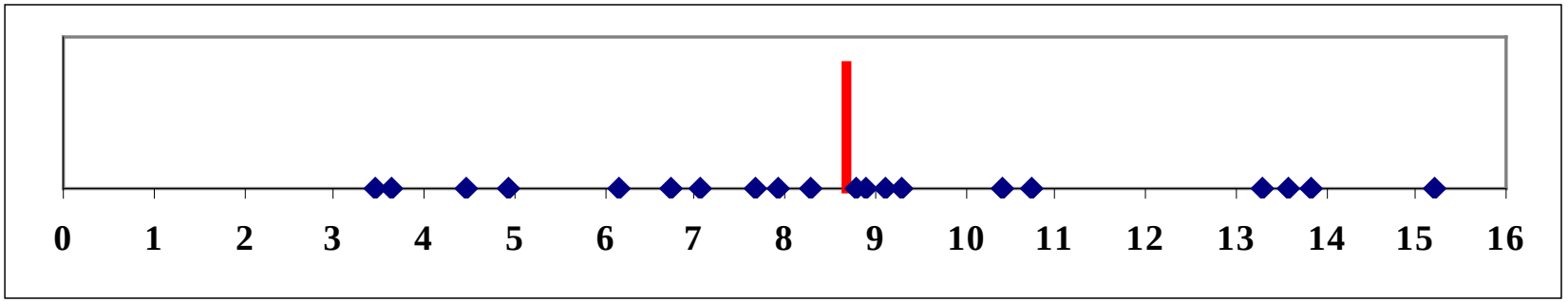
1. polohy – **reprezentace souboru na číselné ose**

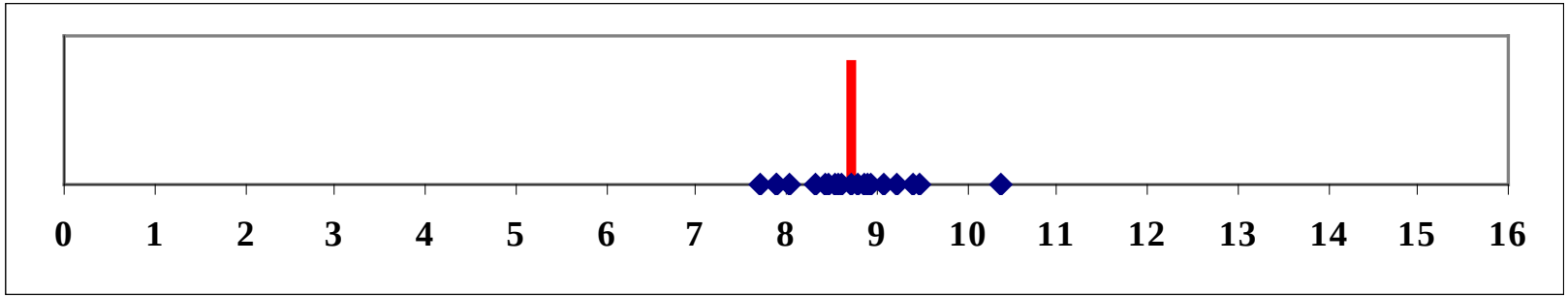


STATISTICKÉ CHARAKTERISTIKY

Typy charakteristik:

2. variability – rozptýlení hodnot po číselné ose navzájem a vůči charakteristice polohy

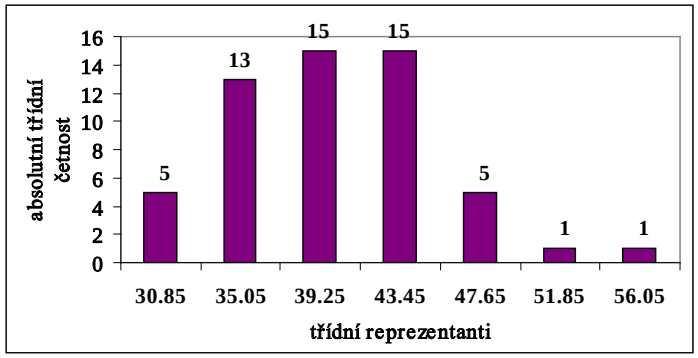


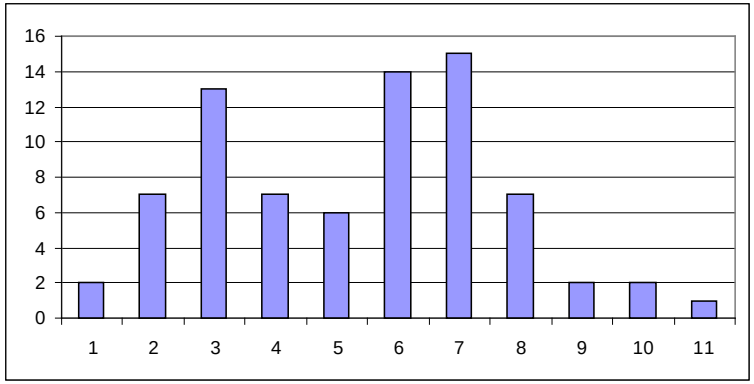


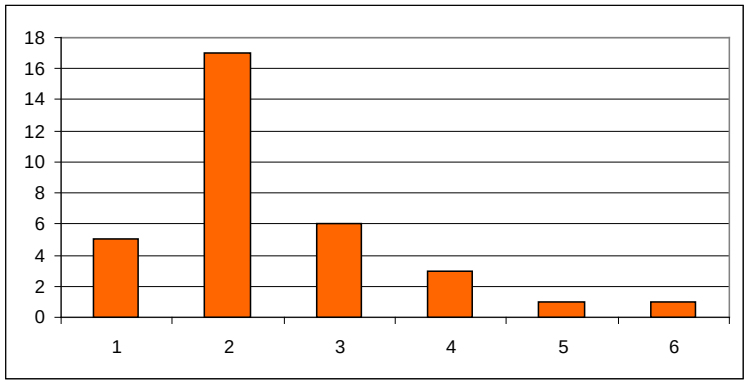
STATISTICKÉ CHARAKTERISTIKY

Typy charakteristik:

3. tvaru – rozložení četností hodnot







CHARAKTERISTIKY POLOHY

- **ARITMETICKÝ PRŮMĚR – hodnota reprezentující všechny hodnoty souboru s nejmenší chybou**
- **MEDIÁN – 50% kvantil, prostřední hodnota vzestupně uspořádaného souboru**
- **MODUS – nejčastěji se vyskytující hodnota v souboru**

ARITMETICKÝ PRŮMĚR ($\bar{x}$)

- základní statistická **MOMENTOVÁ** charakteristika polohy
- je to hodnota, která reprezentuje **VŠECHNY** hodnoty souboru s nejmenší chybou
- fyzikálně je možné jej považovat za „těžiště“ souboru

$$\bar{x}_1 = \frac{\sum_{i=1}^{N} x_i}{N}$$

vzorec pro netříděný soubor
x_i – měřené hodnoty
N – celkový počet hodnot

$$\bar{x}_2 = \frac{\sum_{i=1}^{m} n_i \cdot \bar{x}_i}{N}$$

vzorec pro tříděný soubor
ni – četnost hodnot v i-té třídě
$\bar{x}_i$ - střed třídy (třídní reprezentant)

MEDIÁN ($\tilde{x}$)

- základní statistická **KVANTILOVÁ** charakteristika polohy
- je to hodnota, která reprezentuje **PROSTŘEDNÍ PRVEK VZESTUPNĚ USPOŘÁDANÉHO SOUBORU**

$$\tilde{x} = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{pro N liché} \\ \frac{1}{2} \cdot \left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)} \right) & \text{pro N sudé} \end{cases}$$

MEDIÁN

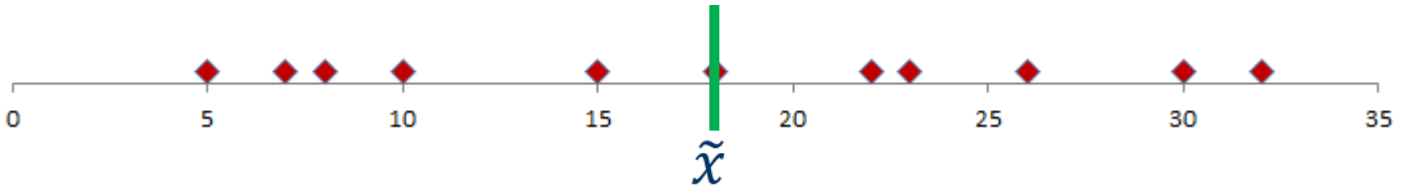
Stanovení mediánu:

1) stanovit pořadové číslo mediánu podle vzorce na předchozím snímku (závisí na tom, zda je sudý nebo lichý počet hodnot)
2) na základě pořadového čísla stanovit medián

lichý počet hodnot – N = 11

pořadové číslo mediánu: (N+1)/2 = (11+ 1)/2 = 6

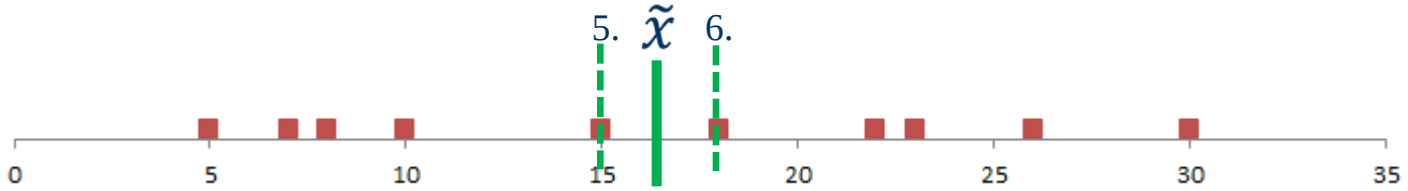
šestá hodnota je medián



sudý počet hodnot – N = 10

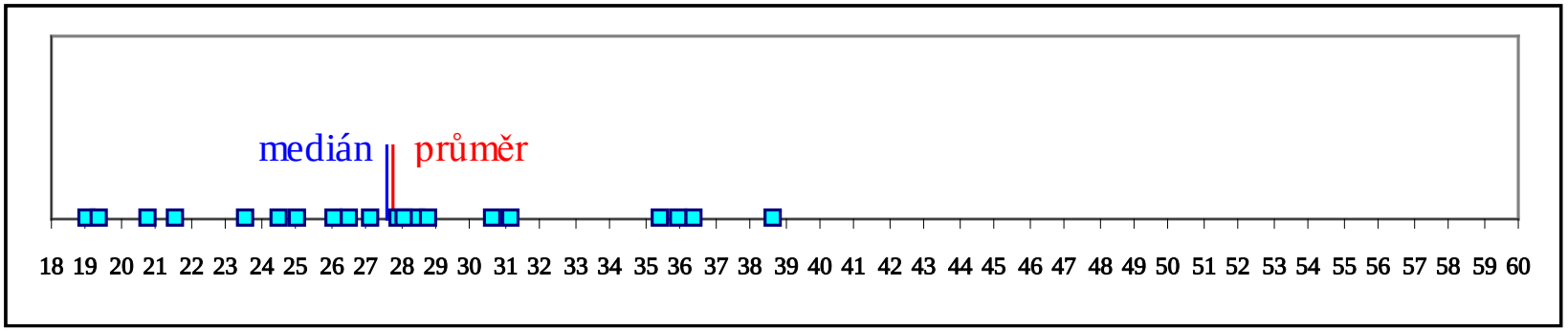
pořadové číslo mediánu: (N+1)/2 = (10+ 1)/2 = 5,5

medián je průměr mezi pátou a šestou hodnotou

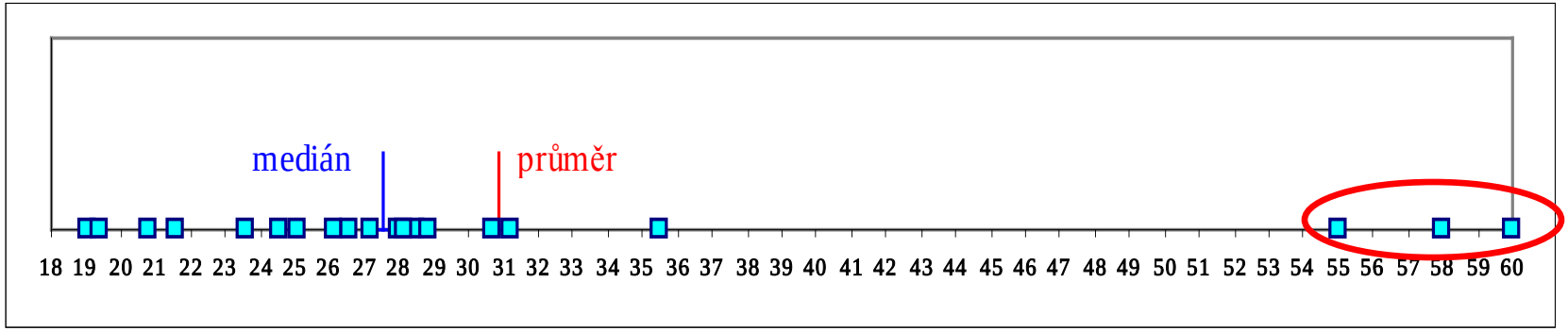


POUŽITÍ PRŮMĚRU A MEDIÁNU

Soubor bez extrémních hodnot:



Soubor s extrémními hodnotami:



POUŽITÍ PRŮMĚRU A MEDIÁNU

Z předchozího obrázku vyplývá , že průměr je vždy „vytahován“ za extrémy, tedy platí, že

- pokud je **průměr výrazně vyšší než medián**, jsou v souboru **extrémy nejvyšších hodnot**
- pokud je **průměr výrazně menší než medián**, jsou v souboru **extrémy nejmenších hodnot**

MODUS

- **nejčastěji se vyskytující hodnota souboru**
- **existují soubory:**
  - **amodální** – bez modu (všechny prvky souboru mají stejnou četnost)
  - **unimodální** – jeden modus
  - **polymodální** – dva a více modů
- **nemá příliš velkou vypovídací schopnost**

CHARAKTERISTIKY VARIABILITY

- informují o tom, jak jsou jednotlivé hodnoty souboru **rozptýleny**, tj. jak se jednotlivé hodnoty znaku **liší vzhledem k sobě navzájem** nebo **vzhledem ke střední hodnotě**
- existují dva typy:
  - **absolutní** - mají rozměr studované veličiny
  - **relativní (poměrné)** - bez rozměru nebo v procentech.

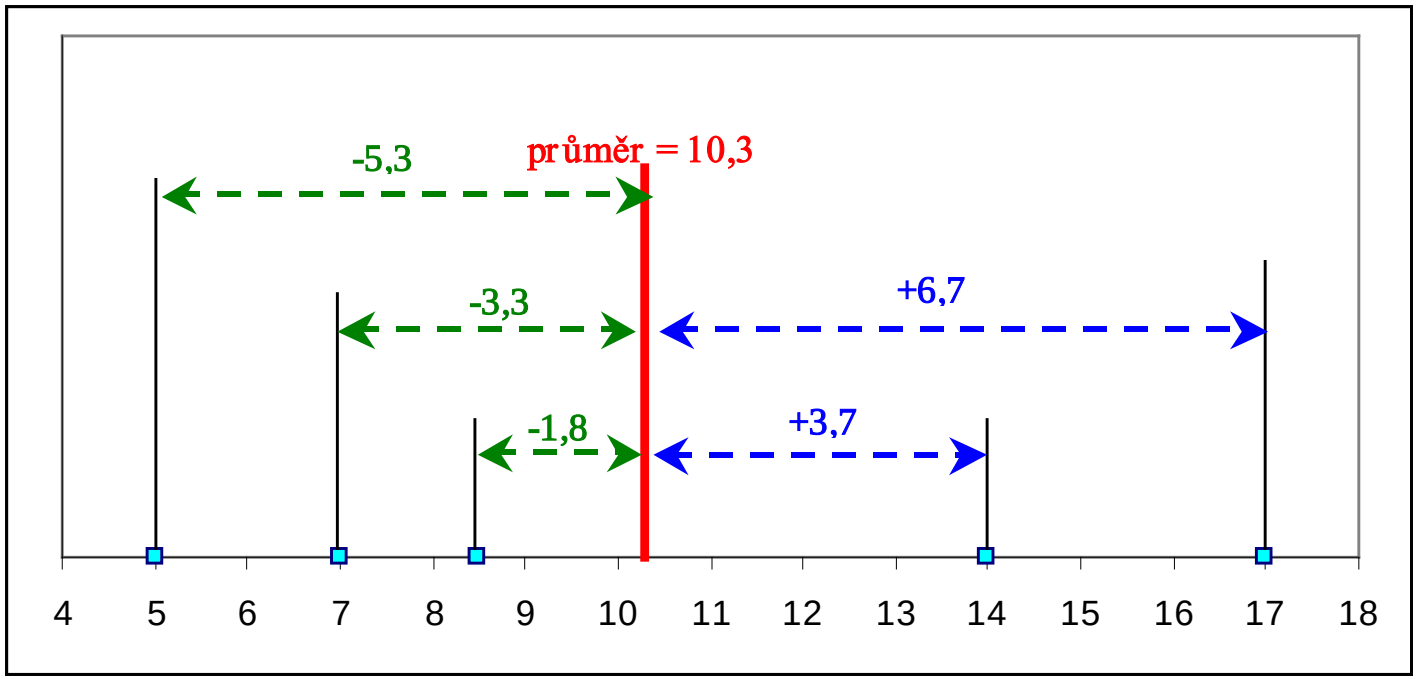
  Jsou vhodné pro porovnání variability různých souborů

CHARAKTERISTIKY VARIABILITY

- **variační rozpětí – rozdíl maximální a minimální hodnoty**
- **rozptyl – základní momentová míra variability, průměr čtverců odchylek od průměru**
- **směrodatná odchylka – odmocnina z rozptylu, využívaná hlavně pro popis souborů**
- **variační koeficient – relativní míra variability užívaná ke srovnání variability různých souborů**
- **kvantilové odchylky – kvantilová míra variability počítaná obvykle z kvartilů nebo decilů**
- **interkvartilové rozpětí – rozdíl horního a dolního kvartilu**

ROZPTYL

Rozptyl je základní mírou variability. Je to **aritmetický průměr čtverců odchylek od průměru** a je tedy konstruován k vyjádření variability hodnot kolem průměru, ale vyjadřuje i vzájemnou odlišnost hodnot znaku (Druhé mocniny odchylek jsou zde proto, aby se při výpočtu průměrné odchylky nevyrovnávaly kladné a záporné odchylky).



ROZPTYL

pro základní soubor:

$$\sigma^2 = \text{var}\ X = \frac{\sum_{j=1}^{N}\left(x_j - \mu\right)^2}{N}$$

pro výběrový soubor:

$$S^2 = \text{var}\ X = \frac{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}{n-1}$$

pro tříděný soubor:

$$S^2 = \frac{\sum_{i=1}^{m} n_i\left(\bar{x}_i - \bar{\bar{x}}\right)^2}{N}$$

SMĚRODATNÁ ODCHYLKA

je odmocnina z rozptylu. **Rozměr** směrodatné odchylky **je stejný jako rozměr veličiny**, což je její hlavní výhodou oproti rozptylu pro účely popisné statistiky, jinak směrodatná odchylka poskytuje stejnou informaci o variabilitě souboru jako rozptyl – průměrnou odchylku hodnot od střední hodnoty.

VARIAČNÍ KOEFICIENT

je **relativní mírou variability** a používá se **k vzájemnému porovnávání variability** různých souborů.

$$S\% = \frac{S}{\bar{x}} \cdot 100$$

K porovnávání variability různých souborů je vždy nutné použít variační koeficient, především pro soubory používající různé jednotky nebo mající hodnoty v různých řádech (např. jednotky a tisíce))!!

VARIAČNÍ KOEFICIENT

Příklad: Který ze dvou zadaných souborů má vyšší variabilitu?

1. soubor

$\overline{x} = 3\,cm, \quad S = 3{,}1\,cm$

2. soubor

$\overline{x} = 150\,cm, \quad S = 75\,cm$

Pouhým srovnáním směrodatných odchylek (S) dospějeme k závěru, že vyšší variabilitu má 2.soubor, protože jeho S je výrazně vyšší

Porovnání pomocí variačního koeficientu:

$$S\% = \frac{S}{\overline{x}} \cdot 100 = \frac{3{,}1}{3} \cdot 100 = 103,\overline{3}\%$$

$$S\% = \frac{S}{\overline{x}} \cdot 100 = \frac{75}{150} \cdot 100 = 50\%$$

Využitím S% zjistíme, že vyšší variabilitu (tj. více rozptýlené hodnoty souboru) má 1. soubor, protože průměrná odchylka měřené hodnoty od průměru je více než 100 % hodnoty průměru, zatímco u 2. souboru je to pouze 50 % jeho hodnoty

KVANTILOVÉ MÍRY VARIABILITY

Kvantilové odchylky jsou horší mírou variability než momentové charakteristiky. **Používají se tam, kde nelze použít momentové charakteristiky** (silně nenormální rozdělení, výskyt extrémních hodnot, apod.)

Kvartilová odchylka:

$$Q = \frac{\left(\tilde{x}_{75} - \tilde{x}\right) + \left(\tilde{x} - \tilde{x}_{25}\right)}{2} = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2}$$

Interkvartilové rozpětí:

$$R_F = \tilde{x}_{75} - \tilde{x}_{25}$$

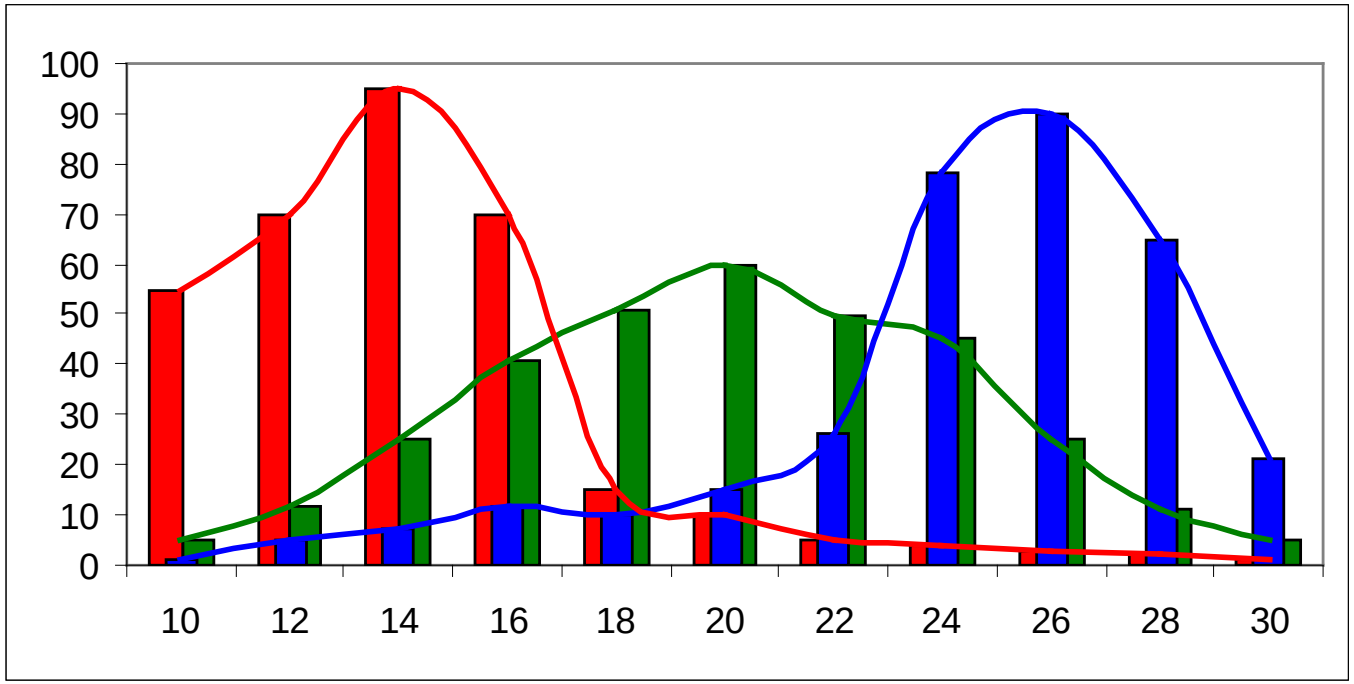
CHARAKTERISTIKY TVARU

měří odchylku v rozložení četností hodnot oproti danému referenčnímu rozdělení četností (obvykle normálnímu): Skládá se ze dvou složek:

- **nesouměrnosti (šikmosti, asymetrie)**
- **špičatosti (zahrocenosti, excesu)**

NESOUMĚRNOST

se projevuje tím, že v souboru je **více hodnot menších než větších ve srovnání se střední hodnotou (levostranná nesouměrnost)** nebo **více hodnot větších než menších ve srovnání se střední hodnotou (pravostranná nesouměrnost).**

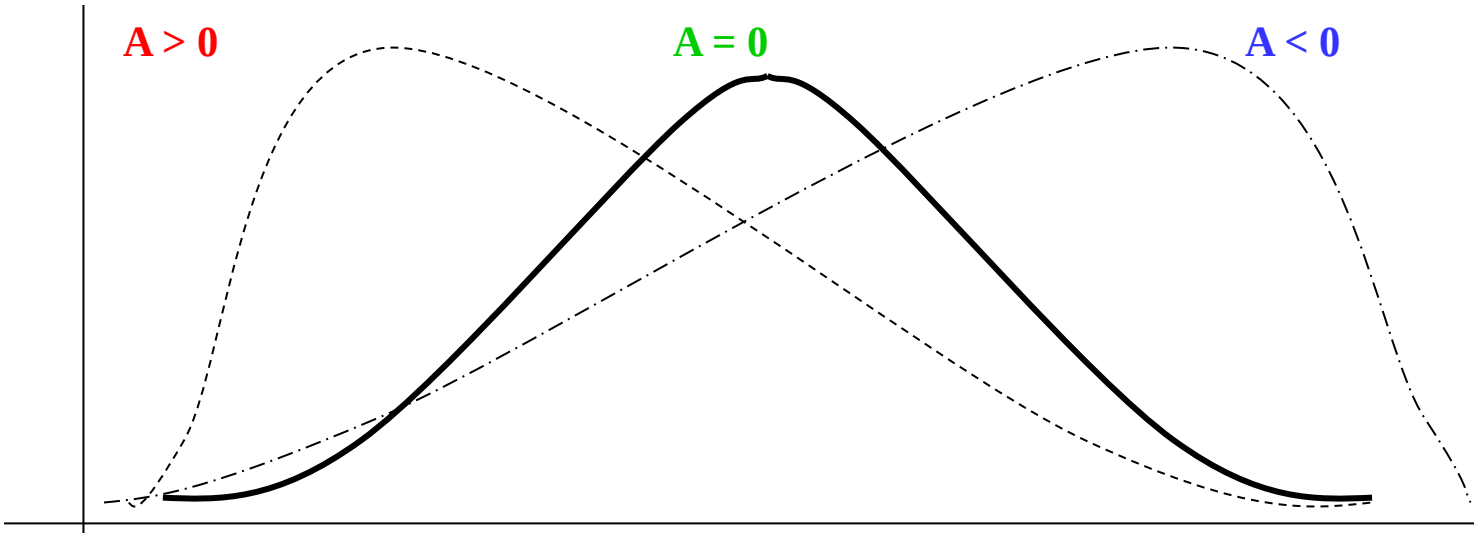


NESOUMĚRNOST

měříme **koeficientem nesouměrnosti**

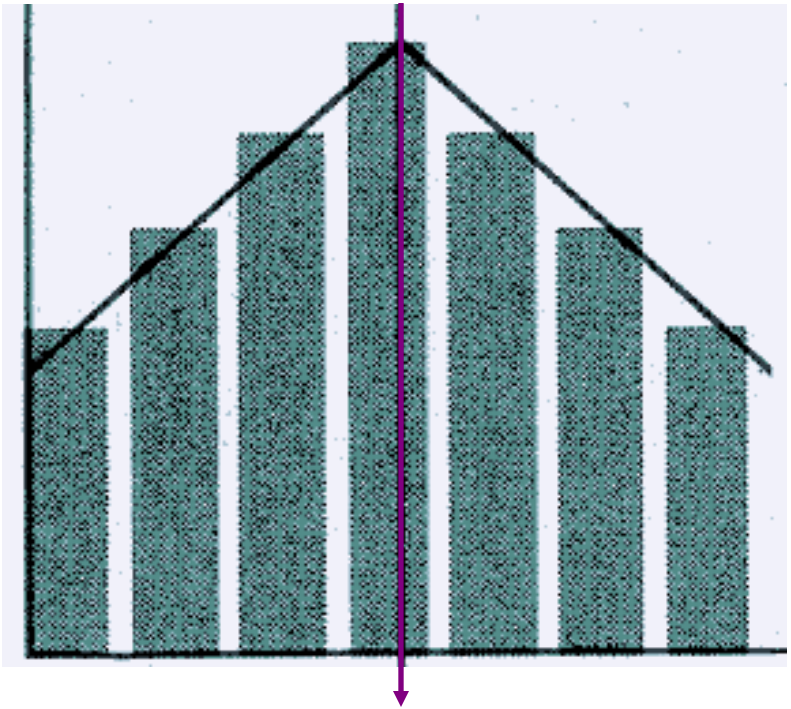
$$A = \frac{\sum_{j=1}^{N} \left(x_j - \overline{x} \right)^3}{n \cdot S^3}$$

$$A = \frac{\sum_{i=1}^{m} n_i \left(\overline{x}_i - \overline{x} \right)^3}{n \cdot S^3}$$



NESOUMĚRNOST

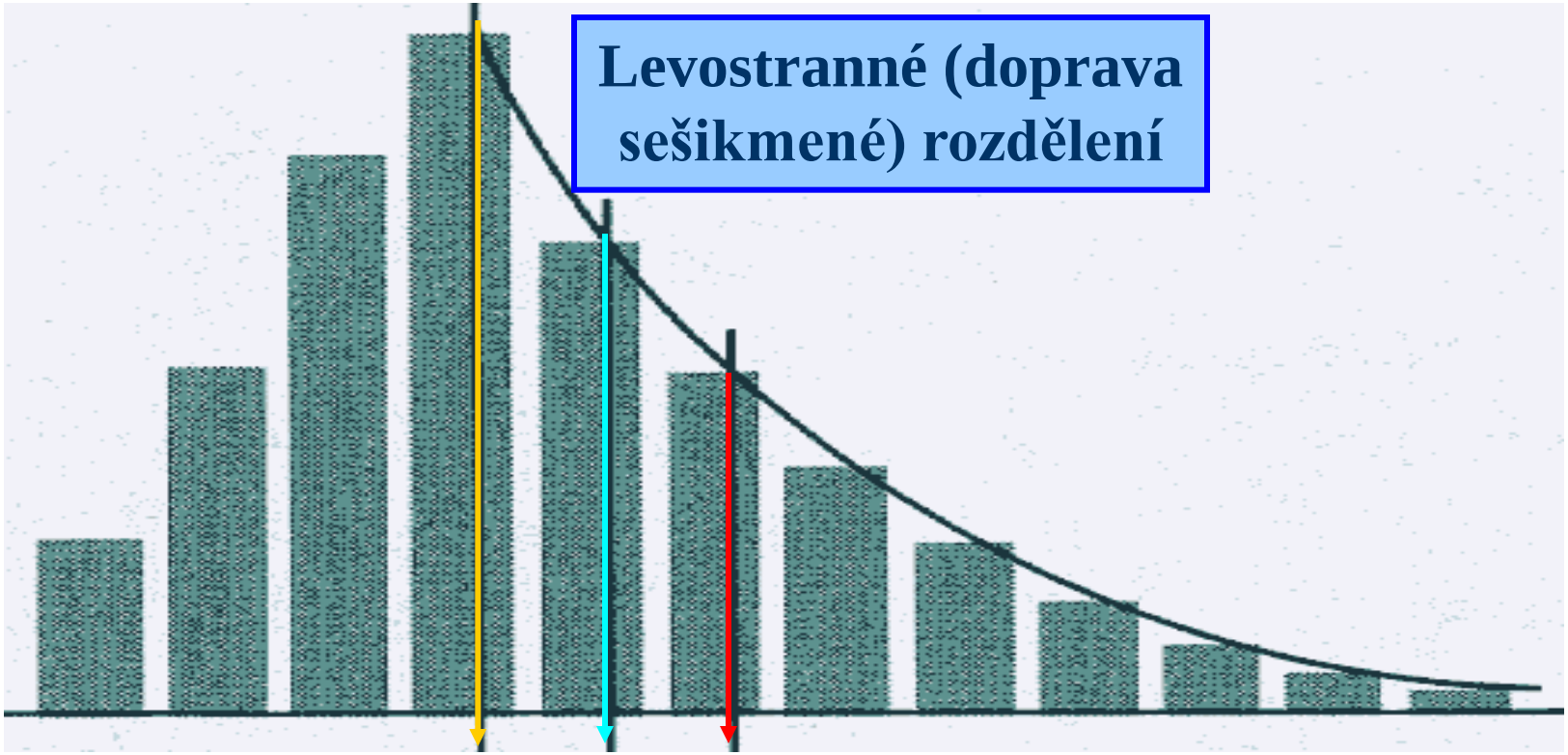
Souměrné rozdělení:



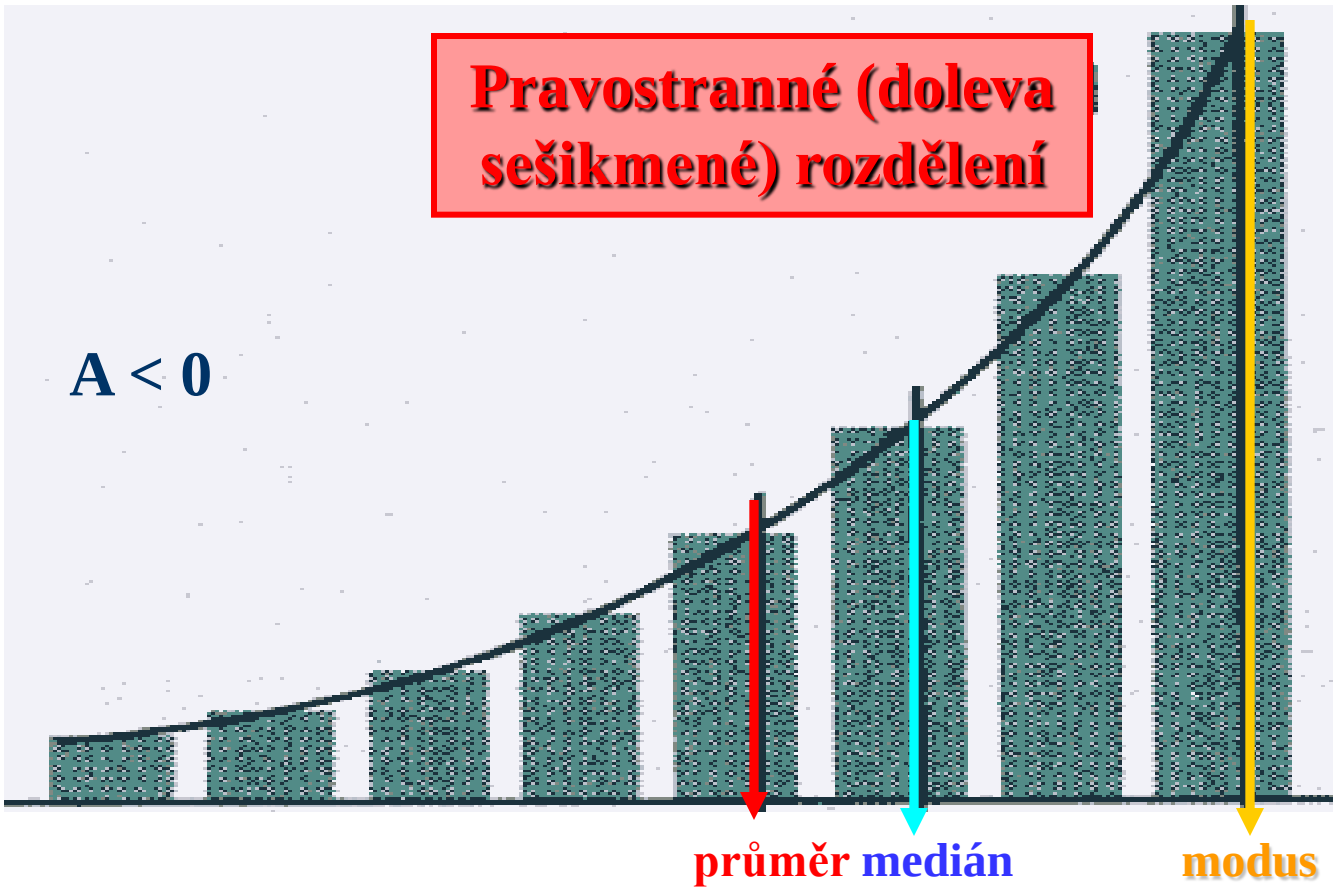
A = 0

Průměr = medián = modus

NESOUMĚRNOST



NESOUMĚRNOST

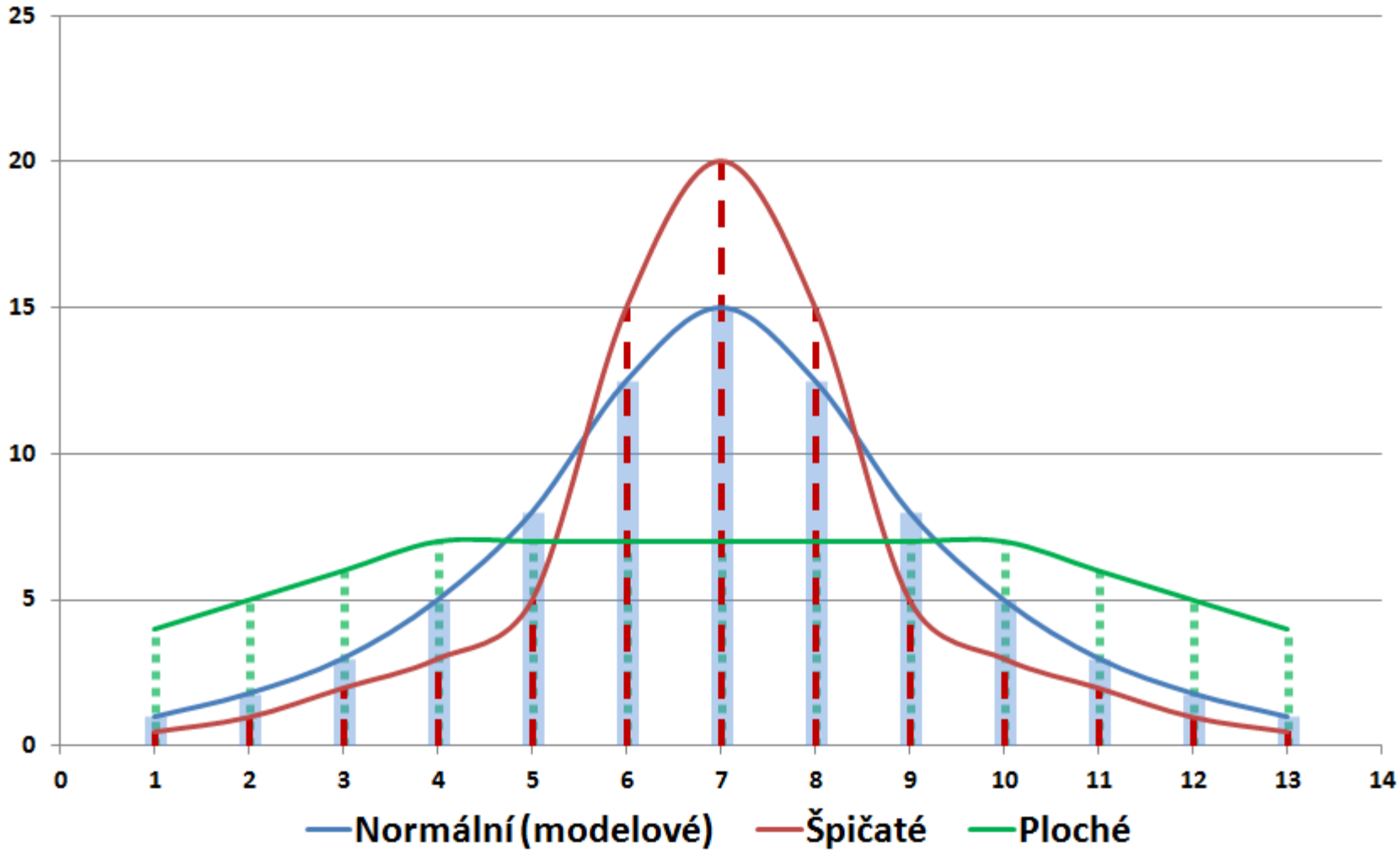


ŠPIČATOST

je mírou **koncentrace dat** kolem určité hodnoty nebo skupiny hodnot ve srovnání s určitým definovaným rozdělením veličiny (např. normálním). Rozlišujeme rozdělení:

- **ploché** – **koncentrace dat** kolem určité hodnoty **je NIŽŠÍ** než odpovídá definovanému rozdělení (tedy četnosti kolem této hodnoty jsou nižší)
- **špičaté - koncentrace dat** kolem určité hodnoty je **VYŠŠÍ** než odpovídá definovanému rozdělení(tedy četnosti kolem této hodnoty jsou vyšší)
- odpovídající danému definovanému rozdělení (např. normální)

ŠPIČATOST


25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Normální (modelové)
Špičaté
Ploché

ŠPIČATOST

Mírou špičatosti je **koeficient špičatosti**:

$$E = \frac{\sum_{j=1}^{N}\left(x_j - \mu\right)^4}{N \cdot \sigma^4} \left[-3\right]$$

vzorec pro netříděný soubor

$$E = \frac{\sum_{i=1}^{m} n_i \left(\overline{x}_i - \overline{x}\right)^4}{n \cdot S^4} \left[-3\right]$$

vzorec pro tříděný soubor

Pro normální rozdělení platí:

E = 0 (3) **normálně zahrocené**

E < 0 (3) **ploché**

E > 0 (3) **špičaté**

Každé modelové (matematicky definované) rozdělení má vlastní hodnotu špičatosti. Normální rozdělení má hodnotu 3. Pokud srovnáváme špičatost experimentálního rozdělení s rozdělením normálním a pro výpočet E použijeme pouze černou část vzorce, potom se výsledná hodnota srovnává s hodnotou 3. Pokud se ještě odečte tato hodnota, která je pro každé modelové rozdělení jiná – červené číslo v hranaté závorce - potom se hodnota E srovnává s hodnotou 0 – to je častější případ a platí v Excelu i v programu Statistika.