Deskriptivní statistika

B02907 Informační a komunikační technologie

Lubomír Štěpánek,
Ústav biofyziky a informatiky
1. LF UK

Deskriptivní statistika

- „statistika zkoumá jevy, které se projeví až na velkém souboru případů, nikoliv pouze v jednom případě“
- statistika pracuje s *daty*
- pojmy vystihující data:
  - *statistický soubor* = množina *statistických jednotek* (prvků statistického souboru)
  - každá *statistická jednotka* dána svým *znakem*

Základní zápis dat

statistický soubor	znak A	znak B	…	znak X
statistická jednotka 1	hodnota	hodnota		
statistická jednotka 2	hodnota	…		
statistická jednotka 3				
…				
statistická jednotka n				

statistický soubor	výše mzdy (Kč/měsíc)	vysoká škola	pořadí dle výše mzdy	
Jan Novák	15000	ano	5	
Jiří Novotný	18000	ne	3	
Eva Dvořáková	21000	ano	2	
…	…	…	…	
Petr Nováček	13000	ne	12	

Typy dat

typ dat	význam	typická hodnota	příklad
metrická	změřený výsledek (obdržený měřícím přístrojem či nástrojem)	reálné číslo se svou jednotkou (!)	123,4 cm, 12000 Kč, …
ordinální	pořadí	vždy přirozené číslo	1., 150., 3. stádium, …
nominální, též alternativní	ano/ne; rovnocenné neporovnatelné hodnoty	nečíselná hodnota (jev/jeho logický opak); soubor rovnocenných hodnot)	prospěl/neprospěl; karcinom prostaty přítomen/karcinom prostaty nepřítomen, muž/žena, …

Převod metrických dat na kategorie

- metrická data někdy „zbytečně“ přesná
- jsou spojitá, ale lze k nim vytvořit schodovitou škálu – osu všech hodnot (*a;b*) rozdělíme na *n* shodných menších intervalů; *i*-tý interval pokrývá hodnoty

$$\left\langle a + \frac{b-a}{n} \cdot (i-1); a + \frac{b-a}{n} \cdot i \right\rangle$$

- např. vytvoření kategorií pro věk (roky, dekády) nebo výšku (celé cm, decimetry atd.)

Jak určit správný počet kategorií

- velký počet kategorií – kategorizace byla zbytečná, nedošlo ke zjednodušení
- malý počet kategorií – na úkor přesnosti
- Sturgesovo pravidlo (počet kategorií *m*):

$$m \approx \log_2 n \approx 1 + 3{,}3 \cdot \log_{10} n$$

Intermezzo

- určete počet a jednotlivé kategorie, do kterých je vhodné rozdělit hodnoty tělesné výšky v souboru 32 mužů s nejnižší výškou 170 cm a největší výškou 195 cm

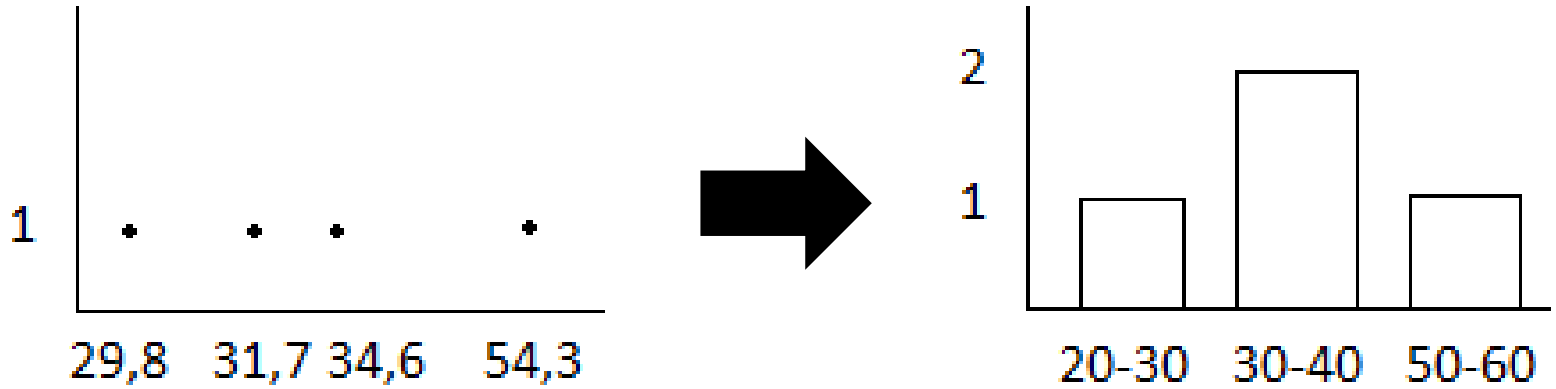
Intermezzo

- určete počet a jednotlivé kategorie, do kterých je vhodné rozdělit hodnoty tělesné výšky v souboru 32 mužů s nejnižší výškou 170 cm a největší výškou 195 cm

- $\log_2 32 = 5$ kategorií; kategorie (cm) 170-175, 175-180, 180-185, 185-190, 190-195

Převod metrických dat na kategorie

statistický soubor	věk	věk (dekády)	tělesná výška	tělesná výška (celé cm)
Jan Novák	34,6	30-40	175,4	175-176
Jiří Novotný	54,3	50-60	169,3	169-170
Eva Dvořáková	31,7	30-40	155,0	155-156
…				
Petr Nováček	29,8	20-30	183,7	183-184

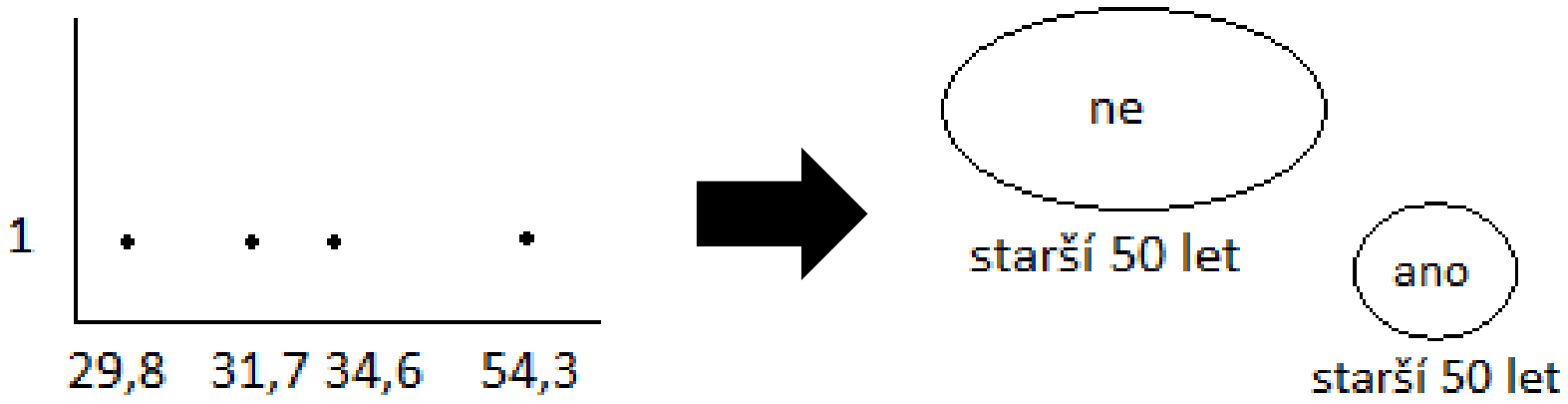


Převod na alternativní data

- metrická i ordinální data lze převést na alternativní
- výhodné při porovnávání se stanovenou mezí
- např. „je tělesná výška větší než 175 cm?“, „je pacient starší padesáti let?“, „je pořadí závodníka menší než 5 a větší než 25?“
- výběr je rozdělen na dvě skupiny:
  - s hodnotami znaku menšími než zvolená mez
  - s hodnotami znaku většími než mez či rovnými mezi
  - případně hodnoty znaku v nějakém rozmezí a mimo něj

Převod na alternativní data

statistický soubor	věk	starší 50 let	tělesná výška	vyšší než 160 cm a nižší než 180 cm?
Jan Novák	34,6	ne	175,4	ano
Jiří Novotný	54,3	ano	169,3	ano
Eva Dvořáková	31,7	ne	155,0	ne
…				
Petr Nováček	29,8	ne	183,7	ne



Četnost

- jednotlivé statistické jednotky vyjádřené stejným znakem tvoří různě početné skupiny
- počet prvků (statistických jednotek) v této skupině nazveme *(absolutní) četnost*

statistický soubor	vysoká škola	statistický soubor	vysoká škola
Jan Novák	ano	Jan Novák	ano
Jiří Novotný	ne	Eva Dvořáková	ano
Eva Dvořáková	ano	Jiří Novotný	ne
Pavel Nový	ne	Pavel Nový	ne
Petr Nováček	ne	Petr Nováček	ne

Četnost

- je-li ve statistickém souboru o *n* statistických jednotkách právě *k* z nich charakterizováno stejnou hodnotou znaku, pak (absolutní) četnost této hodnoty znaku je v daném souboru právě *k*
- relativní četnost této hodnoty znaku je *k/n*

statistický soubor	vysoká škola
Jan Novák	ano
Eva Dvořáková	ano
Jiří Novotný	ne
Pavel Nový	ne
Petr Nováček	ne

hodnota znaku	absolutní četnost	relativní četnost
vysoká škola „ano“	2	2/5
vysoká škola „ne“	3	3/5

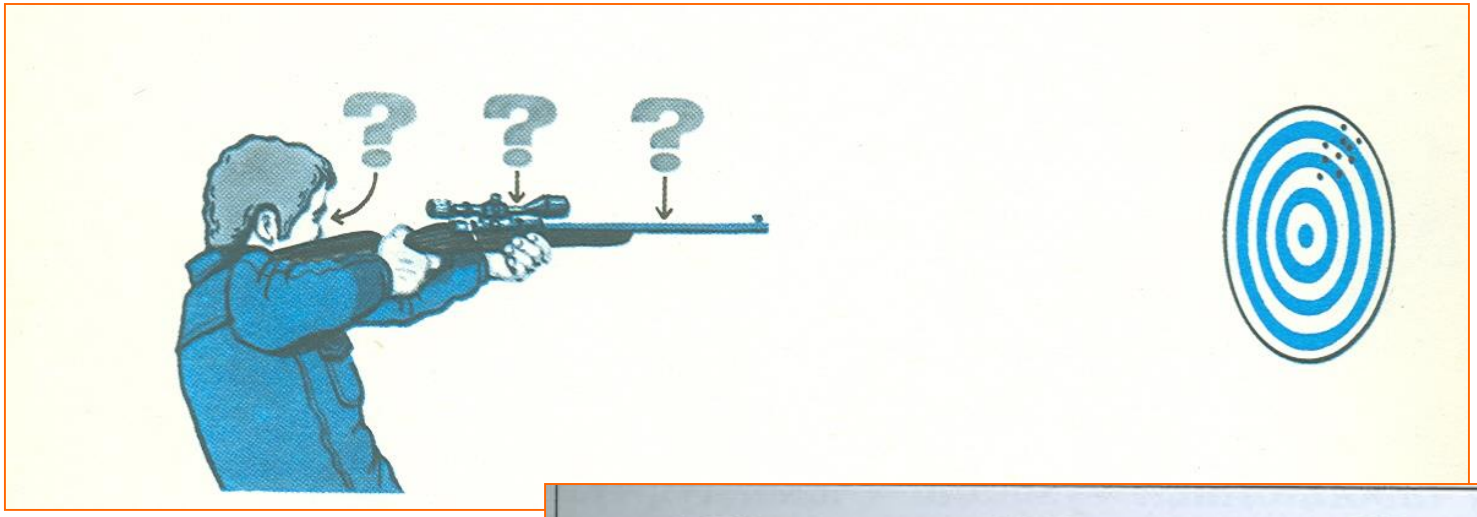
Popis statistického souboru

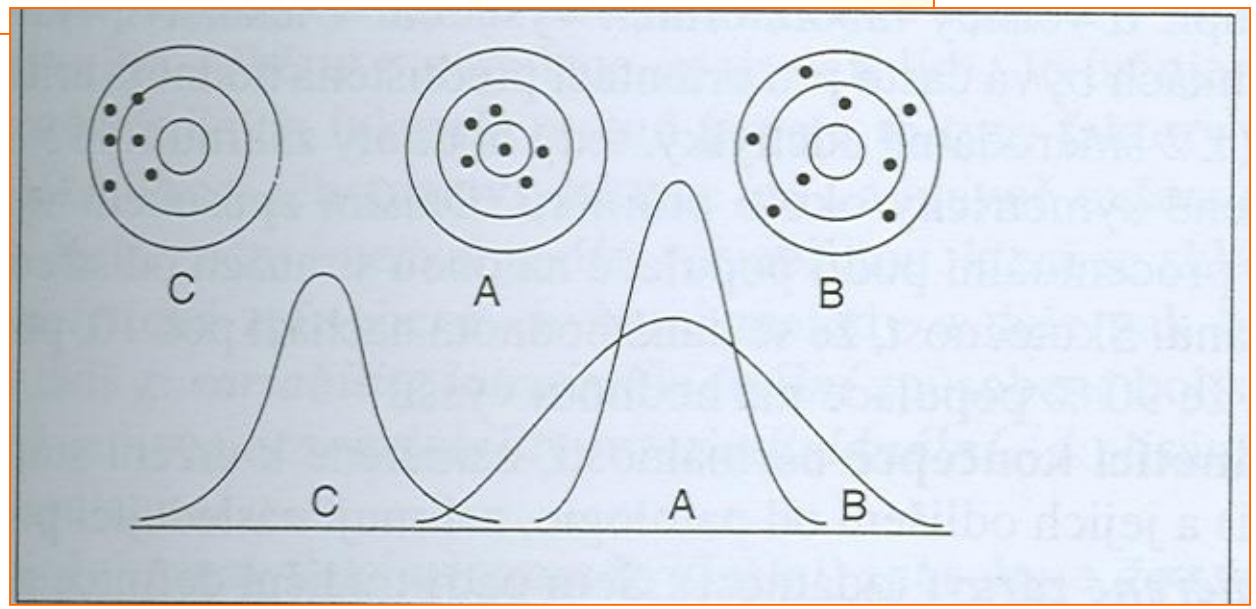
- zkoumaný znak souboru lze popsat:
  - úplným výčtem četností jednotlivých hodnot znaku (*rozdělením četností, pravděpodobnostním rozdělením*; viz dále)
  - vhodně vybranou hodnotou/hodnotami znaku, které dobře vystihují celý soubor
- vhodnými hodnotami jsou:
  - *střední hodnota* statistického souboru a zároveň
  - *variabilita* statistického souboru

Popis statistického souboru

- míry polohy/střední hodnoty (center)
- míry variability (spread)
- míry „tvaru“ (shape)

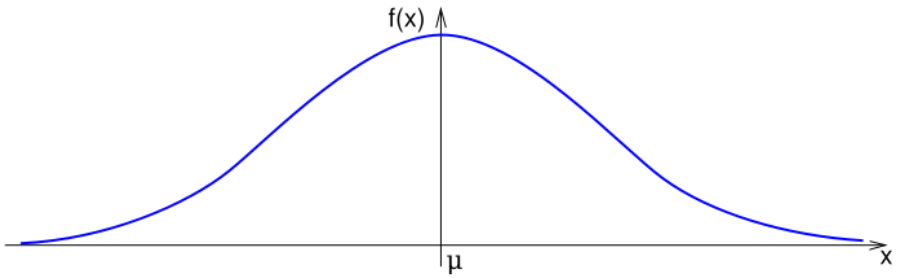
Popis statistického souboru

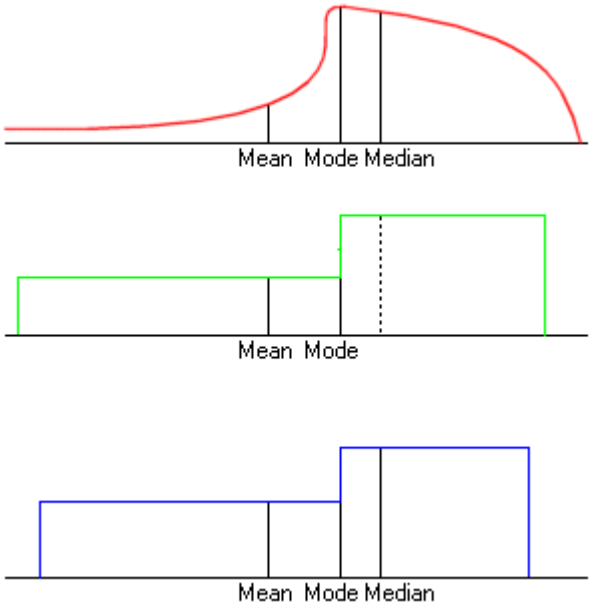




Střední hodnota

- ta hodnota znaku, která vhodně vystihuje polohu četností všech hodnot znaku na vodorovné ose
- *průměr, medián, (modus)*





Aritmetický průměr

- je-li v souboru *n* jednotek, u nichž hodnotíme znak *x*, a jsou-li x_1, x_2, x_3, …, x_n hodnoty znaku *x* postupně 1., 2., 3., …, *n*-té jednotky, je *aritmetický průměr* hodnot znaku *x* v souboru

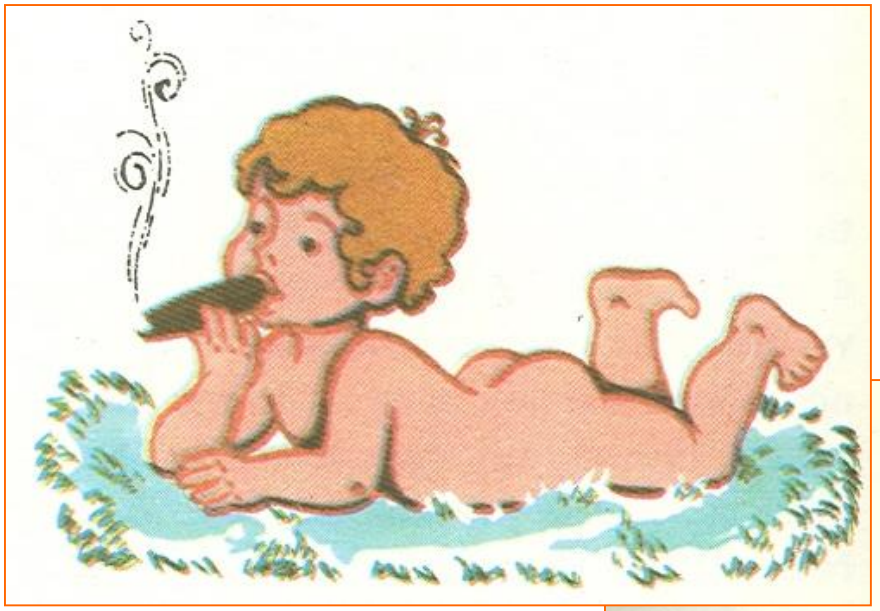
$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

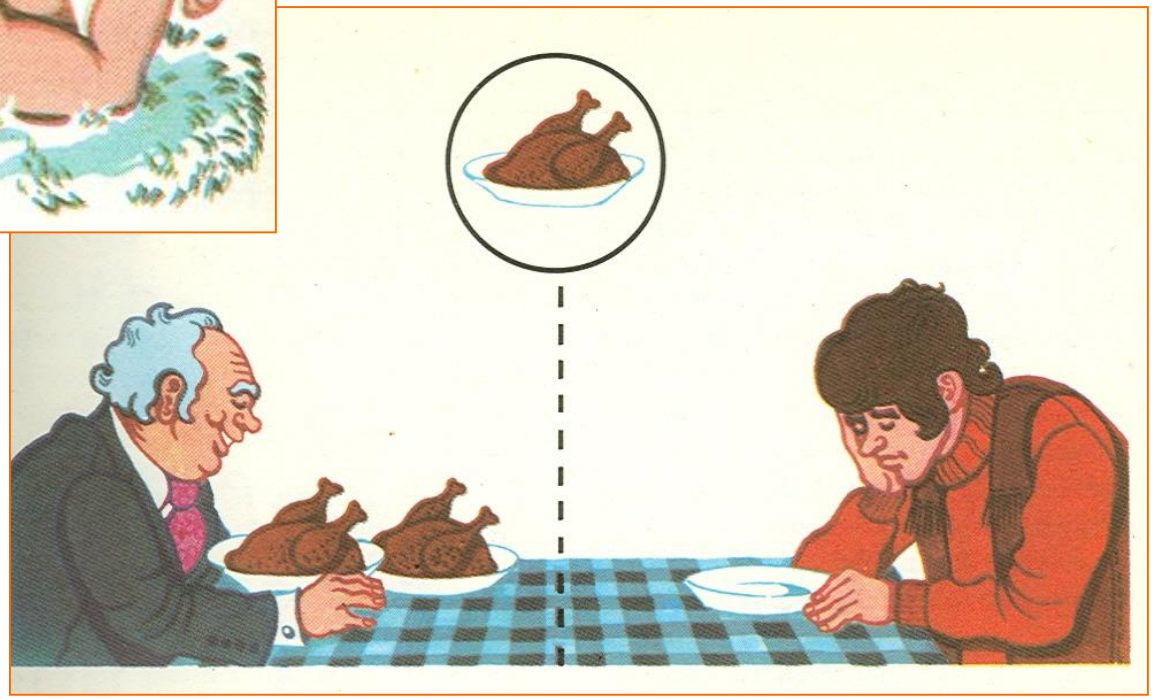
Aritmetický průměr

- pokud se hodnoty znaku *x* v souboru u jednotlivých prvků opakují, je výhodné zavést četnosti těchto hodnot znaku (vytvořit třídy těchto hodnot)
  - je-li v souboru *n* jednotek a *m* různých hodnot znaku *x*, dále jsou-li x_1, x_2, x_3, …, x_m tyto navzájem různé hodnoty znaku *x* a c_1, c_2, c_3, …, c_m (absolutní) četnosti po řadě 1., 2., 3., …, *m*-té hodnoty znaku, je *aritmetický průměr* hodnot znaku *x* v souboru

$$\overline{x} = \frac{c_1 x_1 + c_2 x_2 + \ldots + c_m x_m}{n} = \frac{1}{n} \sum_{j=1}^{m} c_j x_j = \sum_{j=1}^{m} \left(\frac{c_j}{n} \right) x_j$$

Ošidnost průměru





Medián a modus

- *medián* je hodnota znaku prostřední statistické jednotky v souboru, v němž jsou jednotky uspořádány vzestupně (sestupně) podle hodnot znaku
  - v souboru o n jednotkách, u nichž hodnotíme znak x, uspořádáme jednotky vzestupně (sestupně) podle hodnot: $x_1, x_2, x_3, \ldots, x_n$, kde x_i je menší (větší) nebo rovno než x_{i+1} pro všechna x, potom *medián* je hodnota $x_{(n+1)/2}$ pro lichá n, resp. $(x_{n/2}+x_{(n/2)+1})/2$ pro sudá n
- *modus* je hodnota znaku s největší četností v souboru

Intermezzo

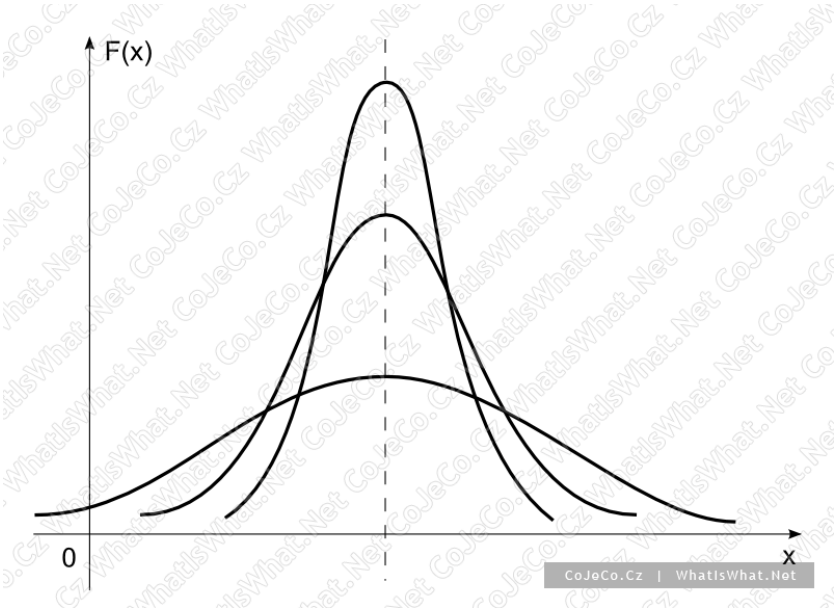
- dopočítejte aritmetický průměr a medián u následujících souborů:
- 1, 2, 2, 2, 3
- 1, 2, 2, 2, 493

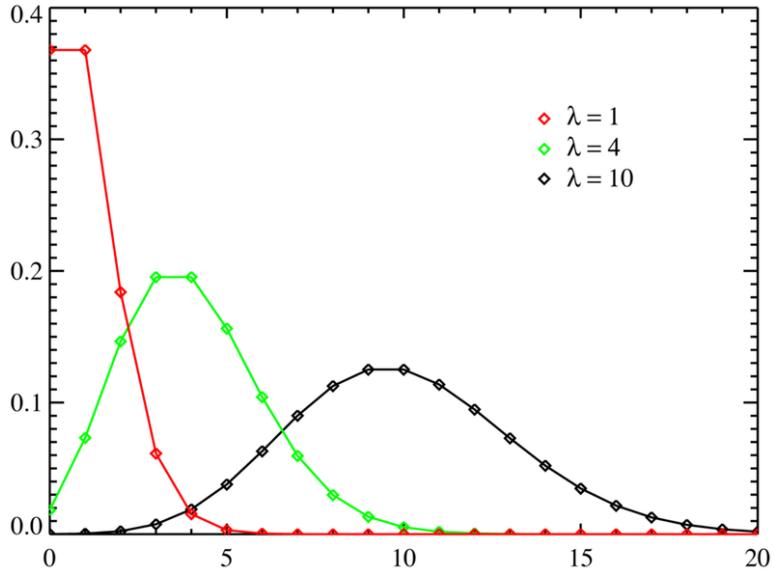
Intermezzo

- dopočítejte aritmetický průměr a medián u následujících souborů:
- 1, 2, 2, 2, 3
- průměr: 2, medián 2
- 1, 2, 2, 2, 493
- průměr: 100, medián 2

Variabilita

- míra „kolísání“, proměnlivosti hodnot znaku kolem střední hodnoty
- *rozptyl*, *kvantily*





Rozptyl

- používá se tehdy, když střední hodnotu charakterizujeme aritmetickým průměrem
  - je-li v souboru *n* jednotek, u nichž hodnotíme znak *x*, a jsou-li x_1, x_2, x_3, …, x_n hodnoty znaku *x* postupně 1., 2., 3., …, *n*-té jednotky, pak *rozptyl* je dán:

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

  - a *(výběrová) směrodatná odchylka* je dána vztahem:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \qquad \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Kvantily

- používají se tehdy, když místo aritmetického průměru za střední hodnotu volíme medián
  - kvantil Q_p je hodnota znaku, která dělí statistický soubor uspořádaný vzestupně dle hodnot znaku na (100p) % prvků s hodnotou nižší nebo rovnou hodnotě Q_p a na zbylých (100(1-p)) % prvků s hodnotami znaku vyššími, než je hodnota Q_p
- důležité jsou *kvartily* a *percentily* (a *medián*)

Kvartily a percentily

- kvartily dělí statistický soubor početně na čtvrtiny
  - 1. kvartil je hodnota znaku jednotky, která odděluje dolních 25 % souboru s nižšími hodnotami znaku, než je hodnota 1. kvartilu
  - 3. kvartil je hodnota znaku jednotky, která od zbylých 75 % odděluje horních 25 % souboru s vyššími hodnotami znaku, než je hodnota 3. kvartilu
  - 2. kvartil = medián
- percentily dělí statistický soubor početně na setiny
  - např. čtyřicátý pátý percentil odděluje dolních 45 % souboru, které mají menší nebo stejnou hodnotu znaku, než má 45. percentil
  - 100*p*-percentil (*k* je celé):

$$x_p = x_{k+1} \text{ pro } k \neq np$$

$$x_p = 0{,}5 \cdot (x_k + x_{k+1}) \text{ pro } k = np$$

Intermezzo

- spočítejte medián, 1. a 3. kvartil a 80. percentil v následujícím souboru:
- 1, 1, 2, 3, 3, 4, 5, 7

$$x_p = x_{\lfloor np \rfloor + 1} \text{ pro } k \neq np$$

$$x_p = 0{,}5 \cdot \left(x_k + x_{k+1} \right) \text{pro } k = np$$

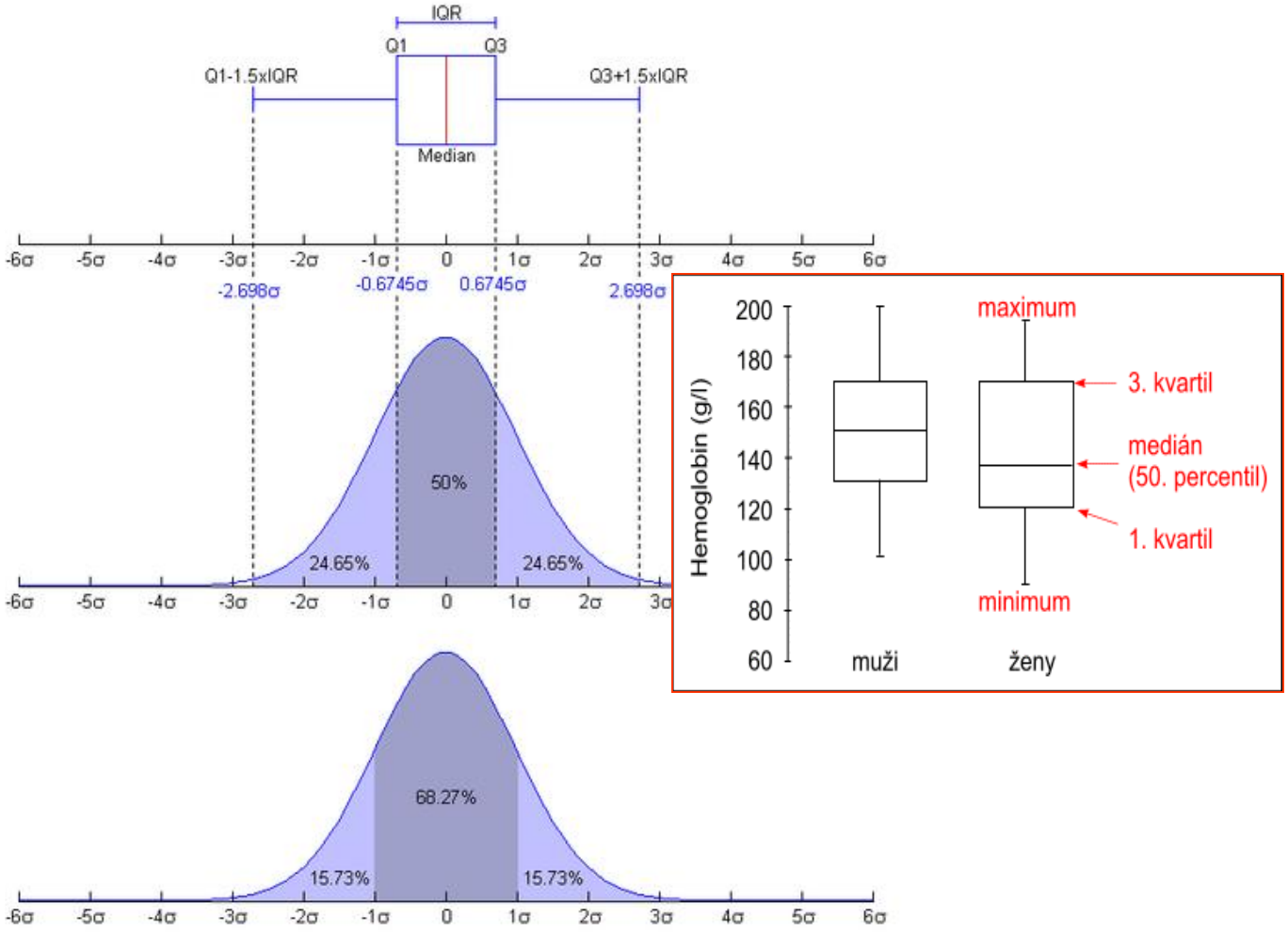
Intermezzo

- spočítejte medián, 1. a 3. kvartil a 80. percentil v následujícím souboru:
- 1, 1, 2, 3, 3, 4, 5, 7
- medián: 3; 1. kvartil: 1,5; 3. kvartil: 4,5; 80. percentil: 5

$$x_p = x_{\lfloor np \rfloor + 1} \text{ pro } k \neq np$$

$$x_p = 0{,}5 \cdot \left(x_k + x_{k+1} \right) \text{pro } k = np$$

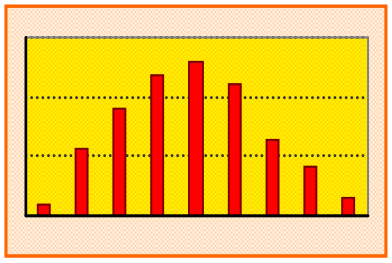
Kvartily a percentily



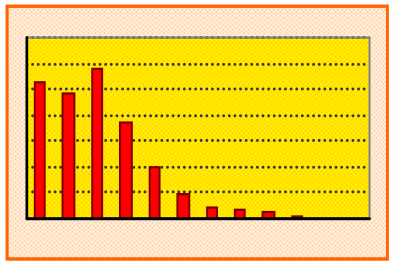
Symetrie dat

- lze posoudit pouhým „pohledem“
  - pokud existuje svislá osa souměrnosti diagramu, data jsou symetrická, jinak asymetrická
- exaktně výpočtem veličiny *šikmost* (*skewness*) a špičatost (*kurtosis*), poměrem *průměr/medián*
  - data jsou symetrická, pokud je splněna podmínka (n>30)

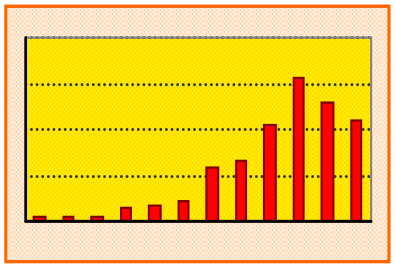
$$-2\sqrt{\frac{6}{n}} \le skewness \le +2\sqrt{\frac{6}{n}} \qquad -4\sqrt{\frac{6}{n}} \le kurtosis \le +4\sqrt{\frac{6}{n}}$$



1. šikmost = 0,02



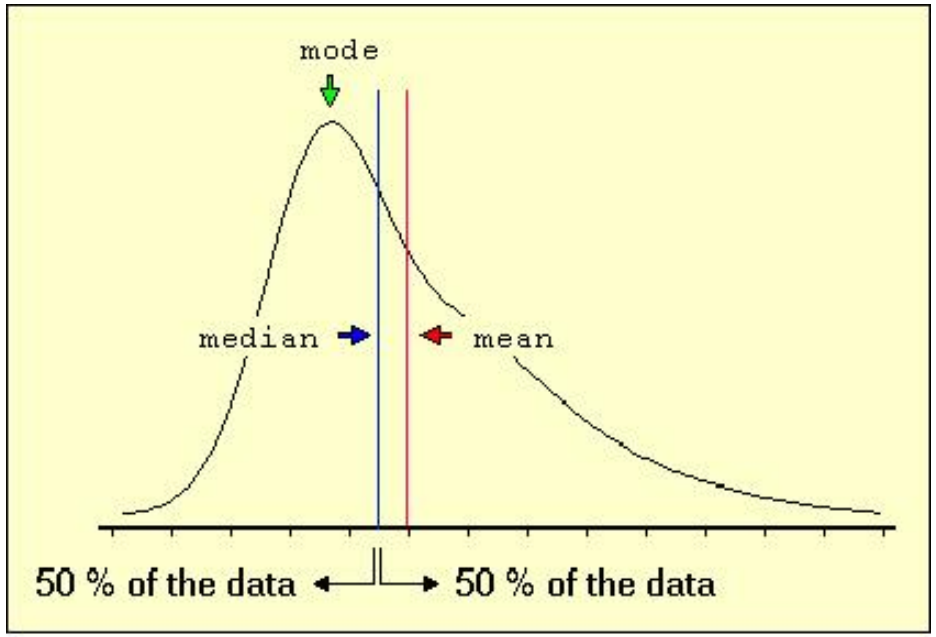
2. šikmost = 2,1



3. šikmost = -1,8

Kdy co použít?

typ dat	míra střední hodnoty	míra variability
symetrická	aritmetický průměr (je ovlivněn extrémními hodnotami)	rozptyl, směrodatná odchylka
asymetrická (často obsahují extrémní hodnoty)	medián (nezávisí na extrémních hodnotách)	kvartily, percentily



Odlehlé hodnoty

- v souboru metrických dat mohou existovat tzv. odlehlé a extrémní hodnoty
- x je odlehlá (outlier value), pokud:

$$x < Q_1 - 1{,}5 \cdot (Q_3 - Q_1) \text{ nebo } x > Q_3 + 1{,}5 \cdot (Q_3 - Q_1)$$

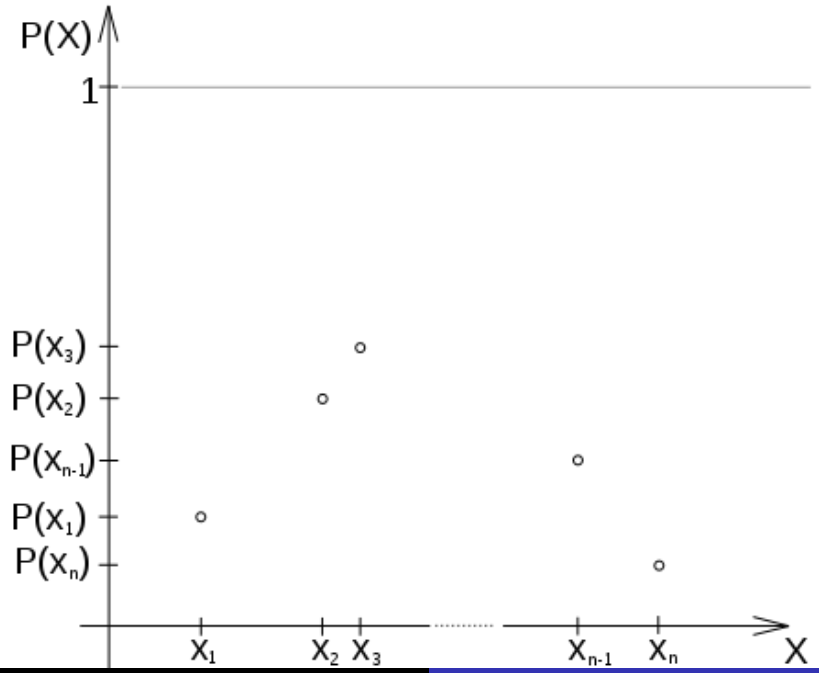
- x je extrémní hodnota (extreme value), pokud:

$$x < Q_1 - 3{,}0 \cdot (Q_3 - Q_1) \text{ nebo } x > Q_3 + 3{,}0 \cdot (Q_3 - Q_1)$$

- s extrémními hodnotami dále nepočítáme, odlehlé je dobré přeměřit, případně také vyřadit

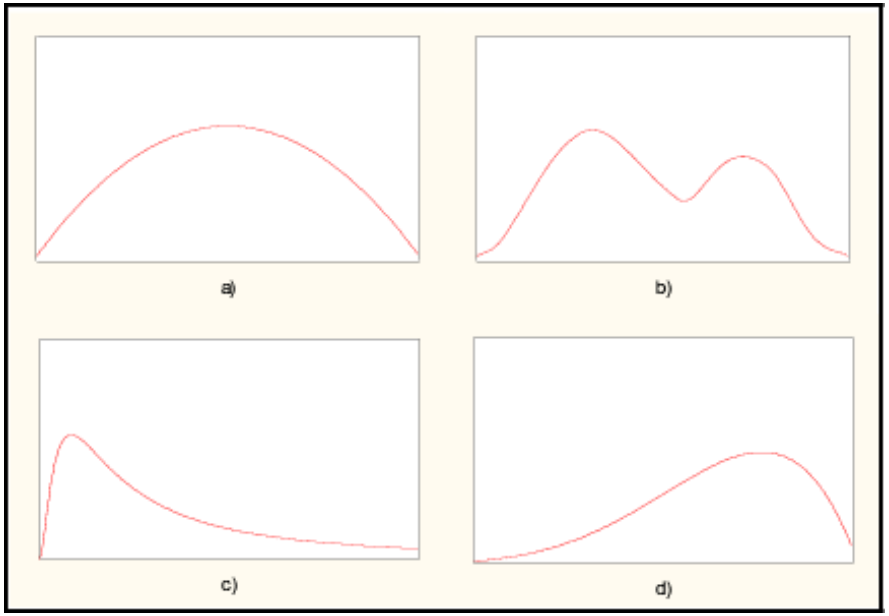
Pravděpodobnostní rozdělení

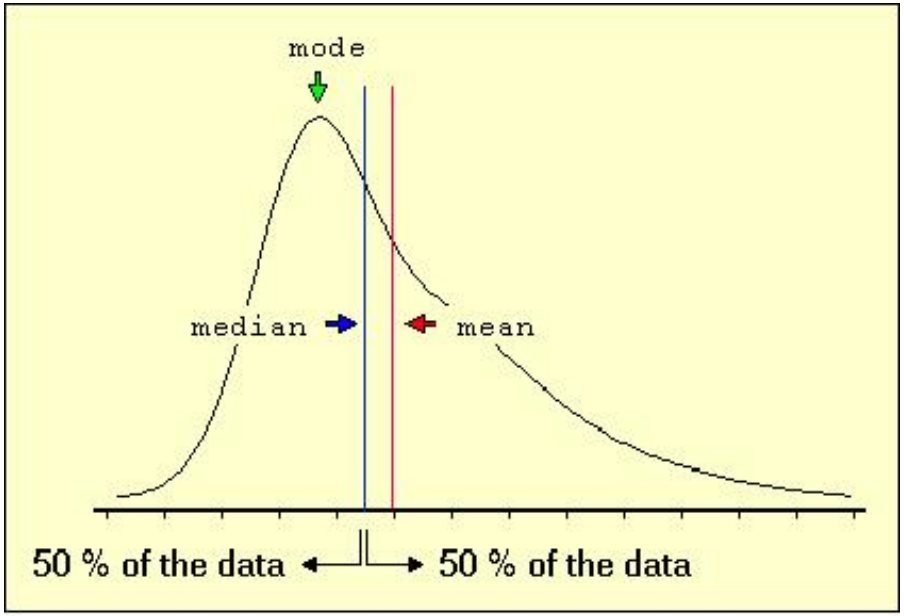
- jde o funkci, která každé hodnotě znaku přiřazuje pravděpodobnost jejího výskytu (předpokládanou četnost)



Typy pravděpodobnostních rozdělení

- *symetrická* (průměr, odchylka, ale i kvartily) X *asymetrická* (pouze kvartily)
- *jednovrcholová* X *vícevrcholová*

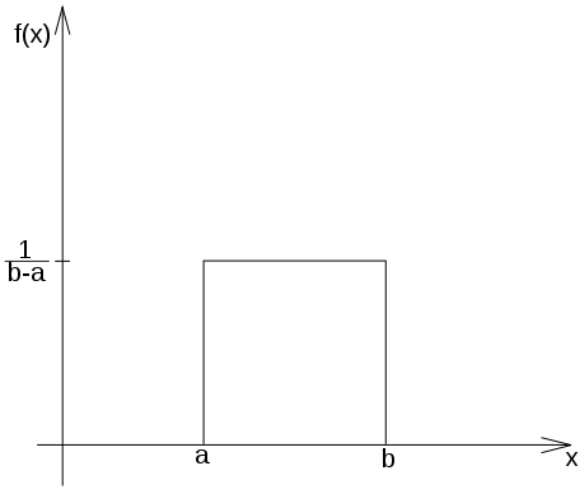




http://new.euromise.org/czech/tajne/ucebnice/html/html/img160.gif

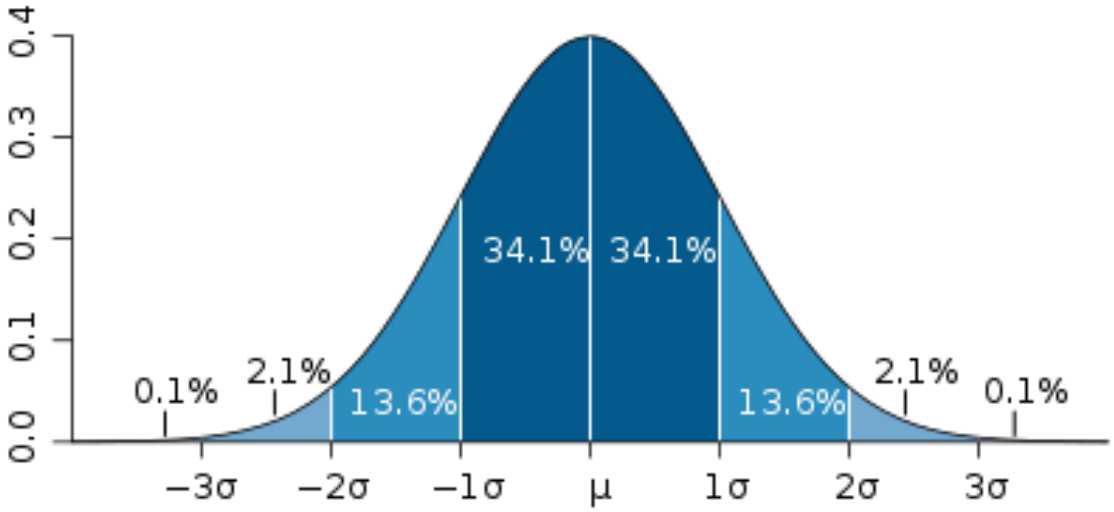
Rovnoměrné rozdělení

- jednoduché diskrétní symetrické rozdělení, každé z *n* hodnot znaku přiřazuje pravděpodobnost 1/*n*
  - např. při hodu kostkou má každá ze šesti stran pravděpodobnost padnutí 1/6
  - může být i spojité, pak každé hodnotě znaku z intervalu [*a*; *b*] přiřazuje pravděpodobnost 1/(*b*-*a*)



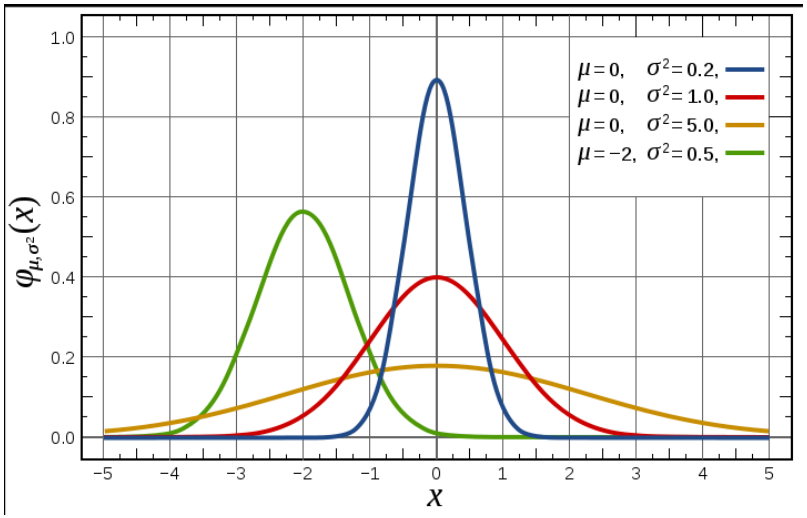
Normální (Gaussovo) rozdělení

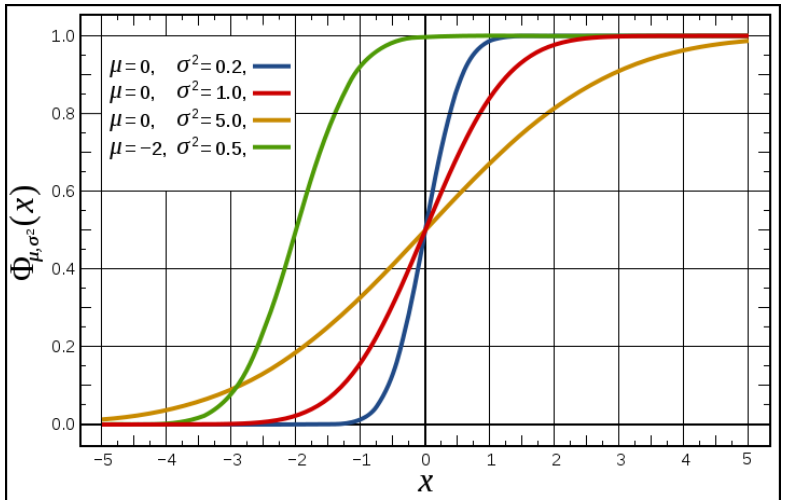
- spojité symetrické rozdělení, popisuje pravděpodobnosti (= relativní četnosti) většiny jevů v biologii, medicíně, ekonomii, sociometrii atd.
  - např. rozložení IQ v populaci, rozložení referenčních mezí fyziologických hodnot, tělesné výšky (nikoliv hmotnosti) atd.



Normální (Gaussovo) rozdělení

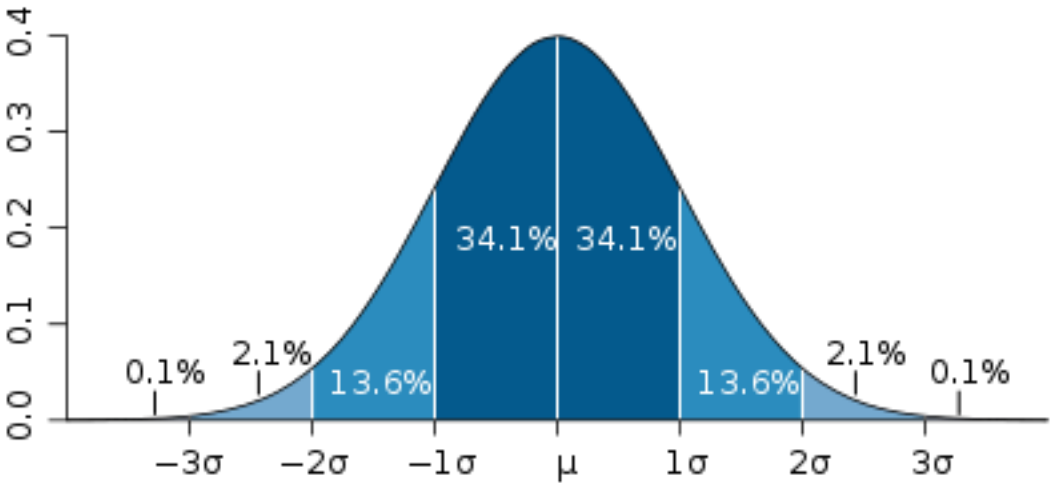
- jednoznačně popsáno průměrem μ a rozptylem σ^2/odchylkou σ
- průměr = medián = modus – vždy hodnota „uprostřed“
- rozptyl udává „šířku“ křivky
- celková plocha pod křivkou je 1, plocha pod křivkou zvoleného intervalu udává pravděpodobnost, s jakou hodnoty z intervalu při pokusu nastanou





Normální (Gaussovo) rozdělení

- 68 % hodnot leží v intervalu (μ-1σ, μ+1σ)
- **95 % hodnot leží v intervalu (μ-2σ, μ+2σ)**
- jinak též: náhodně vybraná hodnota leží s pravděpodobností 0,95 v intervalu $\mu \pm 2\sigma$
- 99 % hodnot leží v intervalu (μ-3σ, μ+3σ)



Normální (Gaussovo) rozdělení

- některé veličiny nemají normální rozdělení, ale lze je na něj převést transformací – logaritmováním, odmocněním, reciprokou hodnotou atd.
- logaritmus „znormalizuje“ např. tělesnou hmotnost, doba přežití po jedné dávce ozáření atd.
- když ani to nepomůže, použijeme *neparametrické metody* (viz další přednášky)

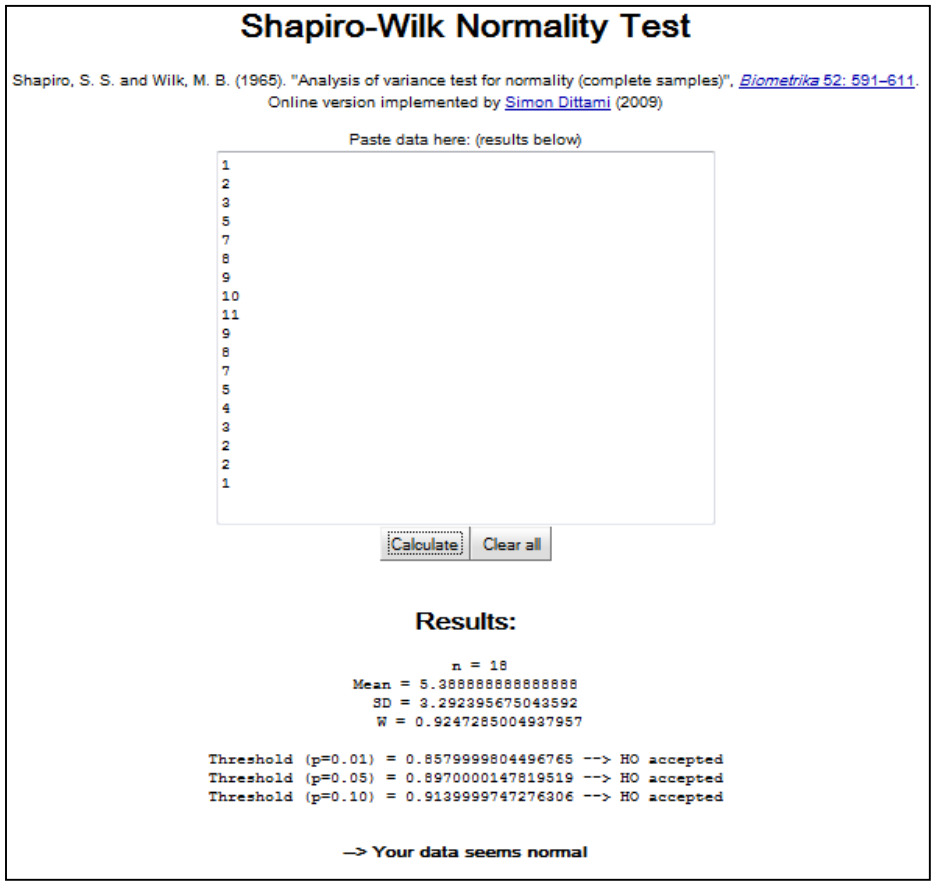
Jak prokážu normálnost dat

- 0) vizuální posouzení v histogramu
- 1) šikmost a špičatost (např. MS Excel)

$$-2\sqrt{\frac{6}{n}} \leq \mathrm{SKEW} \leq +2\sqrt{\frac{6}{n}}$$

$$-4\sqrt{\frac{6}{n}} \leq \mathrm{KURT} \leq +4\sqrt{\frac{6}{n}}$$

- 2) Shapiro-Wilkův test
  - http://dittami.gmxhome.de/shapiro/



Binomické rozdělení

- může být symetrické i asymetrické, nespojité (= diskrétní)
- pro popis veličin, kdy v každém pokusu mohou nastat právě dva výsledky (s neměnnými pravděpodobnostmi)
  - např. hod mincí (hlava/orel), narození dítěte (syn/dcera), multifaktoriální dědičnost (faktor přítomen/nepřítomen), aditivní dědičnost (např. dědičnost barvy vlasů – alela pro tmavé vlasy přítomna/ nepřítomna) atd.

Binomické rozdělení

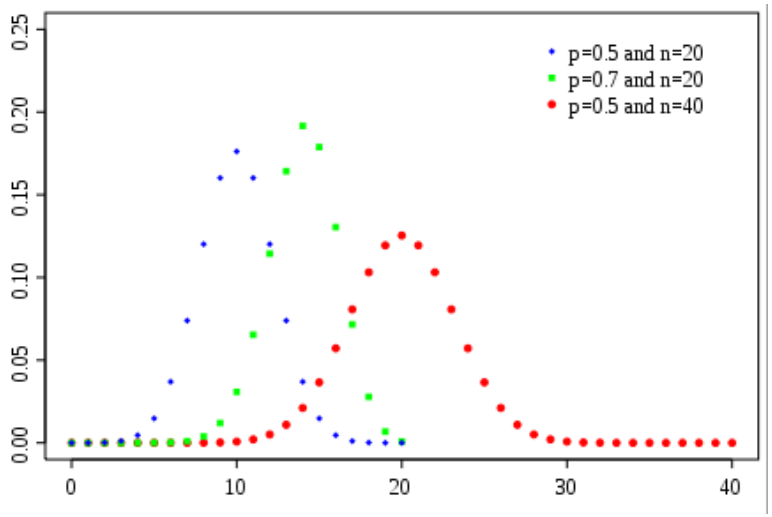
- základem je binomická věta (π je pravděpodobnost „úspěchu“ v každém pokusu; *P(k)* je pravděpodobnost jevu, že „úspěch“ nastane právě *k*-krát mezi *n* nezávislými pokusy)

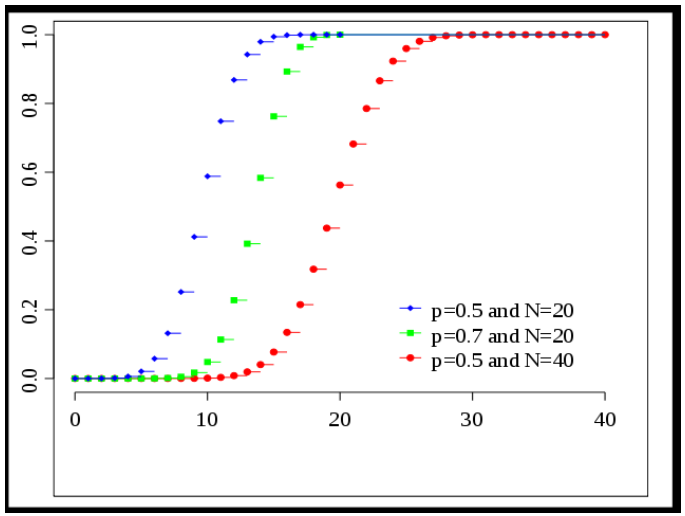
$$P(k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

- průměr binomicky rozložené veličiny je $n\pi$, rozptyl je $n\pi(1-\pi)$

Binomické rozdělení

- je-li π=0,5 a n dost vysoké, lze binomickým rozdělením proložit křivku a dostáváme normální rozdělení
- je-li π velmi malé (blízké nule), dostáváme *Poissonovo rozdělení*





Poissonovo rozdělení

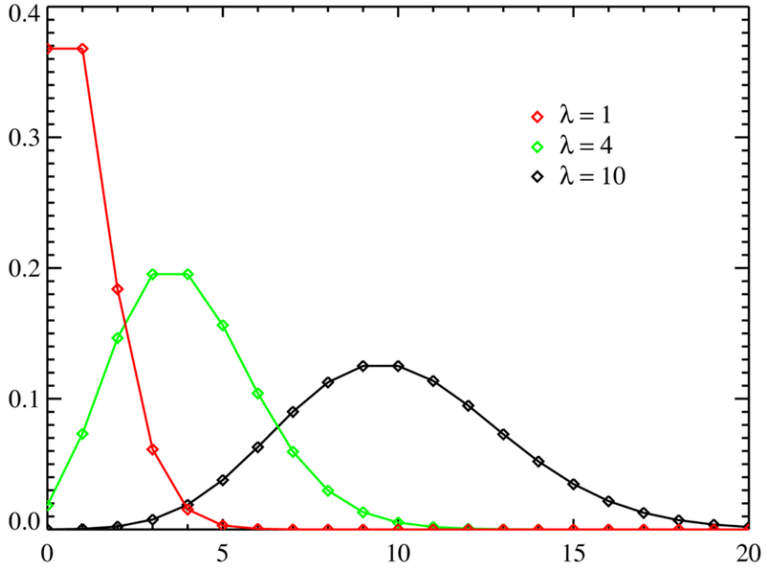
- diskrétní či spojité asymetrické rozdělení (charakterizované tedy mediánem a kvantily), které vyjadřuje četnost málo pravděpodobných jevů v časovém či objemovém intervalu, resp. vzniká jako krajní případ binomického rozložení $\pi \downarrow 0$
- vzácný jev se v daném časovém či prostorovém intervalu vyskytne *k*-krát s pravděpodobností *P(k)*

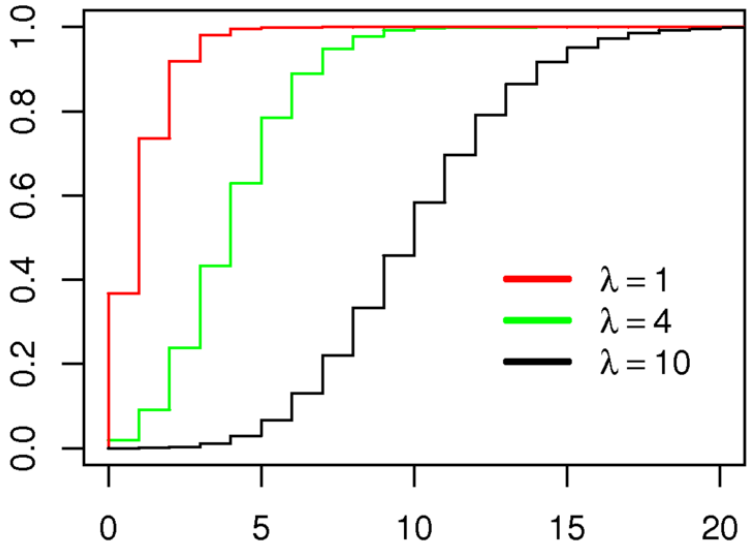
$$P(k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

- průměr i rozptyl jsou totožné a rovny λ, což je průměrný počet vzácných jevů na interval

Poissonovo rozdělení

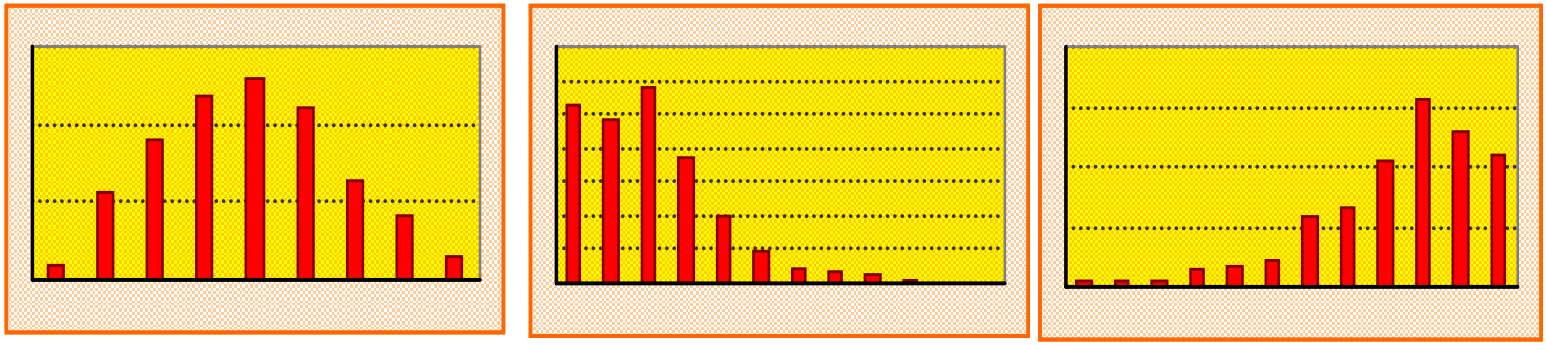
- např. „pravděpodobnost daného počtu záznamů záření radioizotopu Geiger-Müllerovým počítačem“, „četnost výdeje daného počtu léčiva za dané období pro danou oblast“, „mutace genu“ atd.





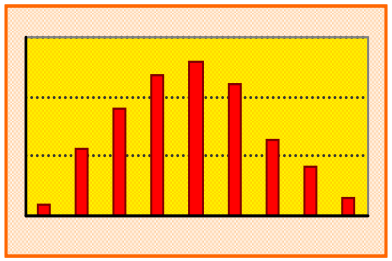
Intermezzo

- určete v následujících diagramech vždy průměr, medián, modus

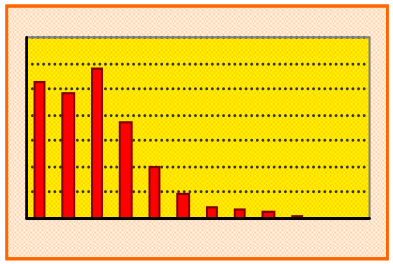


Závěrečné intermezzo

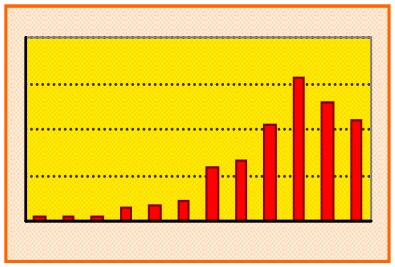
- k následujícím histogramům přiřaďte odpovídající box grafy



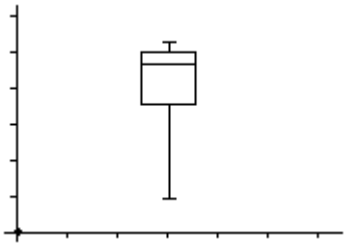
A



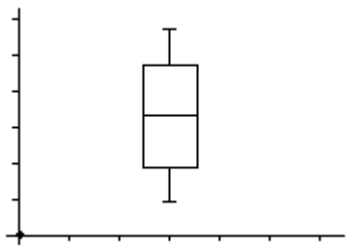
B



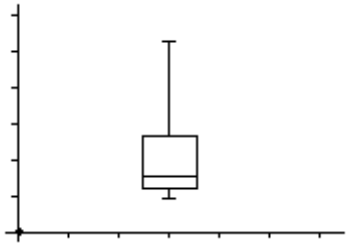
C



1



2



3