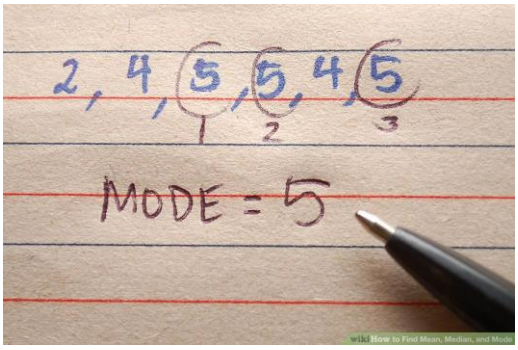
ANALYTICKÁ CHEMOMETRIE

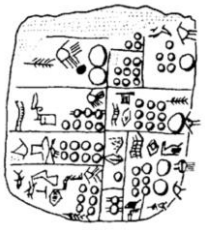
číselné charakteristiky rozdělení



Patrik Kania
Ústav analytické chemie, VŠCHT Praha

23. 09. 2025

- 16. století období rozvoje popisné statistiky (popis stavu země)
- Girolamo Ghilini použil prvně termín *statistika* (*status* = stav, stato = stát)



- v 17. století vzniká v Anglii odvětví statistiky - *politická aritmetika*
  (narození × úmrtí a zkoumání vývoje obyvatelstva v časových obdobích)

Adolphe Lambert Quételet (1796–1874)

- belgický matematik, statistik a sociolog
- vypracoval moderní metody sčítání lidu
- náznaky pojmů jako průměr, rozptyl a rozdělení
- v roce 1853 zorganizoval první statistickou konferenci

ČÍSELNÉ CHARAKTERISTIKY ROZDĚLENÍ

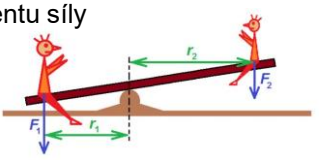
- speciální veličiny informující o podstatných statistických vlastnostech studovaného souboru
- zhuštěná informace o vlastnostech rozdělení našich dat
- náhrada za složité funkce rozdělení

MOMENTOVÉ
(angl. *moments*)

KVANTILOVÉ
(angl. *quantiles*)

Momentové charakteristiky rozdělení

- jsou vypočítány ze všech hodnot souboru
  (obsahují úplnou statistickou informaci → prioritní použití)
- nejsou vhodné pro soubory s extrémními hodnotami
- analogie k fyzikálním momentům, např. momentu síly



obecný moment *r*-tého stupně

diskrétní veličiny

$$m_r = \sum_i x_i^r f(x_i)$$

spojité veličiny

$$m_r = \int x^r f(x)dx$$

centrální moment *r*-tého stupně

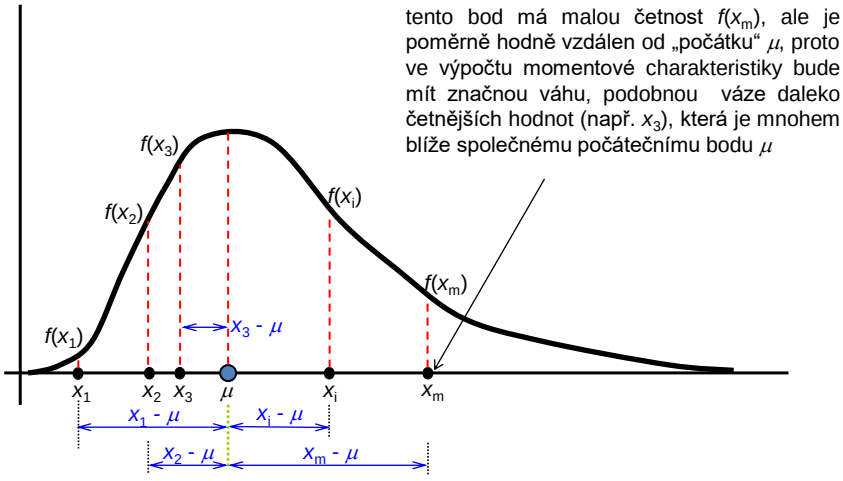
diskrétní veličiny

$$m(\mu)_r = \sum_i (x_i - \mu)^r f(x_i)$$

spojité veličiny

$$m(\mu)_r = \int (x-\mu)^r f(x)dx$$

četnost určité hodnoty $f(x_i)$ = „síla“
vzdálenost od určitého bodu (např. nuly nebo průměru) = „rameno síly“



nejužívanější momenty:

první moment, tj. střední (očekávaná) hodnota rozdělení, E(ξ)
- charakterizuje polohu rozdělení na ose reálných čísel (těžiště)

diskrétní veličiny

$$m_1 = \mathrm{E}(\xi) = \mu = \sum_i x_i f(x_i)$$

spojité veličiny

$$m_1 = \mathrm{E}(\xi) = \mu = \int x\, f(x) dx$$

druhý centrální moment, tj. rozptyl rozdělení, D(ξ)
- charakterizuje variabilitu (kolísavost) rozdělení

diskrétní veličiny

$$m(\mu)_2 = \sigma^2 = \sum_i (x_i - \mu)^2 f(x_i)$$

spojité veličiny

$$m(\mu)_2 = \sigma^2 = \int (x-\mu)^2 f(x) dx$$

důležité vlastnosti střední hodnoty a rozptylu:

$E(\mathrm{k}) = \mathrm{k}$	$D(\mathrm{k}) = 0$	k je konstanta
$E(\mathrm{k} \cdot \xi) = \mathrm{k} \cdot E(\xi)$	$D(\mathrm{k} \cdot \xi) = \mathrm{k}^2 \cdot D(\xi)$	
$E(\xi \pm \eta) = E(\xi) \pm E(\eta)$	$D(\xi \pm \eta) = D(\xi) + D(\eta)$	
$E(\xi \pm \mathrm{k}) = E(\xi) \pm \mathrm{k}$	$D(\xi \pm \mathrm{k}) = D(\xi)$	
$E(\xi \cdot \eta) = E(\xi) \cdot E(\eta)$	$D(\xi \cdot \eta) = E(\xi^2) \cdot E(\eta^2) - (E(\xi) \cdot E(\eta))^2$	

Příklad:

Diferenční metoda vážení spočívá ve dvou postupných vážení, jednak váženky se vzorkem, jednak váženky se zbytkem. Obě hmotnosti můžeme považovat za nezávislé náhodné veličiny ξ_1 a ξ_2. Zjistěte střední hodnotu a rozptyl rozdílu obou hmotností.

$\mathrm{E}(\xi_1 - \xi_2) = \mathrm{E}(\xi_1) - \mathrm{E}(\xi_2) = \mu_1 - \mu_2$

$\mathrm{D}(\xi_1 - \xi_2) = \mathrm{D}(\xi_1) + \mathrm{D}(\xi_2)$,

je-li $\mathrm{D}(\xi_1) = \mathrm{D}(\xi_2) = \sigma^2$, potom $\mathrm{D}(\xi_1 - \xi_2) = 2\sigma^2$.

další používané momenty:

třetí centrální moment $m(\mu)_3$ definující koeficient šikmosti *A* rozdělení

- šikmost charakterizuje asymetrii v rozložení četností hodnot rozdělení

$$A = \frac{m(\mu)_3}{(m(\mu)_2)^{3/2}} = \frac{m(\mu)_3}{\sigma^3}$$

čtvrtý centrální moment $m(\mu)_4$ definující koeficient špičatosti *E* rozdělení

- špičatost charakterizuje koncentraci četností hodnot rozdělení okolo střední hodnoty

$$E = \frac{m(\mu)_4}{(m(\mu)_2)^2} = \frac{m(\mu)_4}{\sigma^4}$$

Kvantilové charakteristiky rozdělení

- jsou vypočítány pouze z určitých hodnot souboru - infoztráta
- nejsou ovlivněny extrémními hodnotami
- nezávisí na rozdělení veličiny
- snadno zjistitelné a interpretovatelné
- jsou vhodné i pro malé soubory dat
- používají se hlavně pro soubory s extrémy nebo se silně nenormálním rozdělením

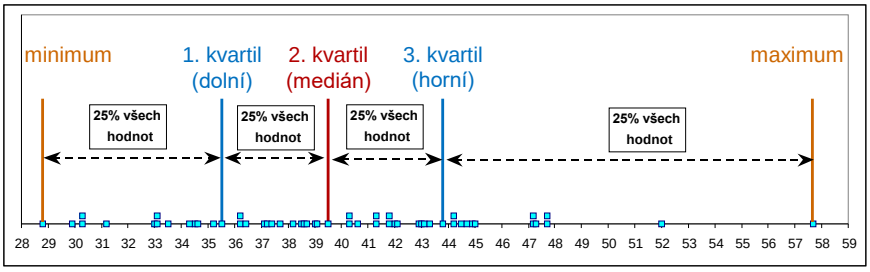
Kvantil
- je hodnota určitým způsobem v souboru umístěná
- je určen svým pořadím v uspořádaném souboru
- leží pod ním (100 · *p*) % hodnot souboru

důležité kvantily:

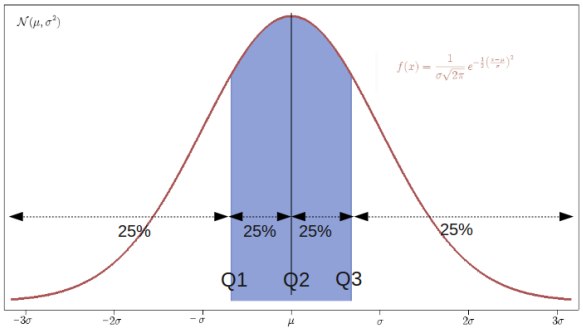
25% kvantil - dolní kvartil
50% kvantil - medián
75% kvantil - horní kvartil

další používané kvantily:

33,3% kvantil - tercil
12,5% kvantil - oktil
10% kvantil - decil



Kvantil - je hodnota určitým způsobem v souboru umístěná
- je určen svým pořadím v uspořádaném souboru
- leží pod ním (100 · *p*) % hodnot souboru



Kvantil - je hodnota určitým způsobem v souboru umístěná
- je určen svým pořadím v uspořádaném souboru
- leží pod ním (100 · *p*) % hodnot souboru

Jak se určí kvantilová hodnota v datovém souboru?

Pořadové číslo hodnoty odpovídající hledanému kvantilu $\tilde{x}_{100p}$ je určena součinem $n \cdot p$ (*n* je celkový počet prvků v souboru). Pokud je tento součin:

desetinným číslem $\Rightarrow$ kvantilem je hodnota s pořadovým číslem *m* splňující nerovnost $n \cdot p \leq m \leq n \cdot p + 1$

celým číslem $\Rightarrow$ kvantilem je aritmetický průměr hodnot s pořadovým číslem $n \cdot p$ a $n \cdot p + 1$

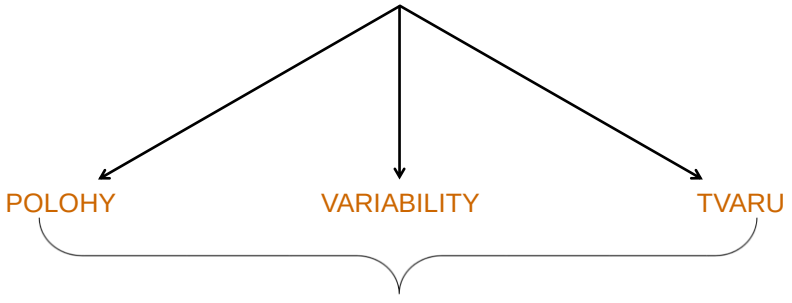
<u>Příklad:</u> Pro předchozí soubor na straně 42 určete všechny tři kvartily

$\tilde{x}_{25} =$ 14. hodnota protože $n \cdot p = 55 \cdot 0{,}25 = 13{,}75$ a $n \cdot p + 1 = 14{,}75$

$\tilde{x}_{50} =$ 28. hodnota protože $n \cdot p = 55 \cdot 0{,}50 = 27{,}50$ a $n \cdot p + 1 = 28{,}50$

$\tilde{x}_{75} =$ 42. hodnota protože $n \cdot p = 55 \cdot 0{,}75 = 41{,}25$ a $n \cdot p + 1 = 42{,}25$

DĚLENÍ ČÍSELNÝCH CHARAKTERISTIK



momentové / kvantilové / jiné

Pro správné statistické zhodnocení jakéhokoliv souboru je třeba použít charakteristiky všech tří skupin - polohy, variability a tvaru (každá z nich popisuje soubor z jiného hlediska).

Nejpoužívanější charakteristiky polohy

PRŮMĚR $\bar{x}$ (angl. *mean*)

- hodnota reprezentující všechny hodnoty souboru
- základní momentová charakteristika zastupující střed souboru

aritmetický průměr (angl. *arithmetic mean*)
=PRŮMĚR(A1:A10)

$$\bar{x} = \frac{1}{n}\sum_i x_i$$

vážený aritmetický průměr (angl. *weighted arithmetic mean*)
=SOUČIN.SKALÁRNÍ(A1:A10;B1:B10)/SUMA(A1:A10)

$$\bar{x}_w = \frac{\sum_i w_i \cdot x_i}{\sum_i w_i}$$, kde w_i je statistická váha čísla

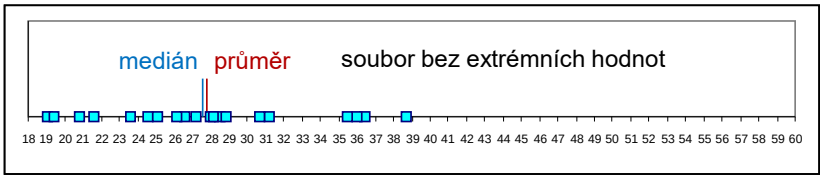
geometrický průměr (angl. *geometric mean*)
=GEOMEAN(A1:A10)

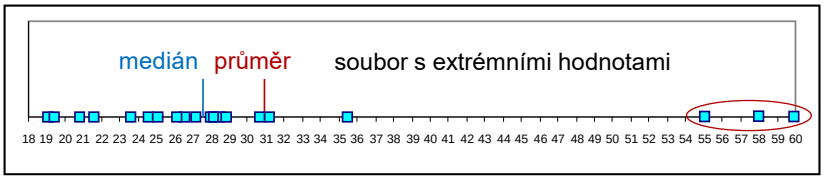
$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

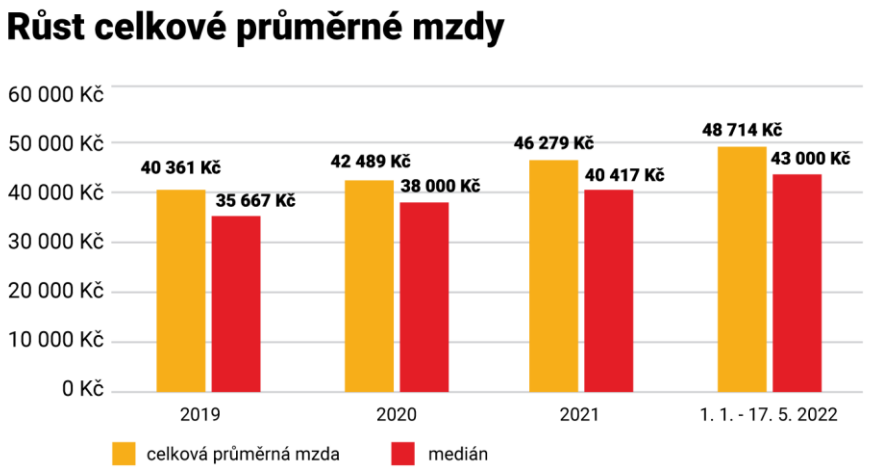
použití pro výpočet průměrného růstu či poklesu

MEDIÁN $\tilde{x}$ (angl. *median*)
=MEDIAN(A1:A10)

- hodnota prostředního prvku vzestupně uspořádaného souboru
- základní kvantilová charakteristika



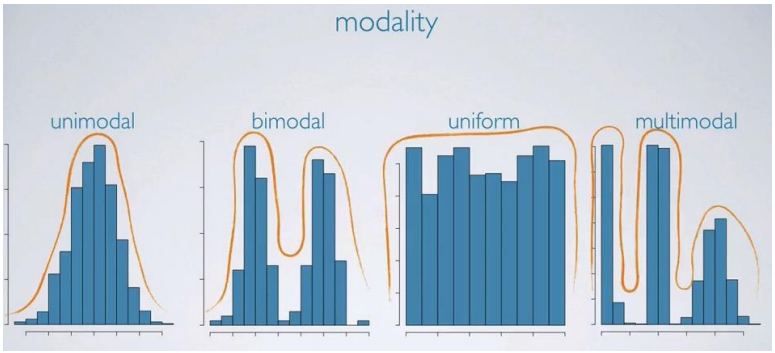




MODUS $\hat{x}$ nebo Mod(X) (angl. *mode*)

=MODE.MULT(A1:A10) – pozor, zadat maticově!

- nejčastěji se vyskytující hodnota souboru (např. nejpravděpodobnější velikost chyby stanovení)
- amodální / unimodální / polymodální
- nemá velkou vypovídací schopnost – je to nepar. charakteristika



Nejpoužívanější charakteristiky variability

ROZPTYL σ^2 (angl. *variance*)
=VAR(A1:A10)

- základní momentová charakteristika variability
- aritmetický průměr čtverců odchylek měření od střední hodnoty

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2$$

SMĚRODATNÁ ODCHYLKA σ (angl. *standard deviation*)
=SMODCH(A1:A10)

- odmocnina z rozptylu
- rozměr je stejný (oproti rozptylu) jako rozměr studované veličiny
- nejlepší a nejpoužívanější charakteristikou variability

RELATIVNÍ SMĚRODATNÁ ODCHYLKA *RSD* (angl. *relative standard deviation*)

- relativní míra variability
- použití k vzájemnému porovnání variability různých souborů (soubory mající různé jednotky či různé řády hodnot)

$$RSD = \frac{\sigma}{\mu}$$

VARIAČNÍ KOEFICIENT *CV* (angl. *coefficient of variation*)

- relativní míra variability v procentech $CV = RSD \cdot 100$

VARIAČNÍ ROZPĚTÍ *R* (angl. *range*)
=MAX(A1:A10)-MIN(A1:A10)

- rozdíl maximální a minimální hodnoty $R = x_{max} - x_{min}$

KVARTILOVÁ ODCHYLKA *Q* (angl. *quartile deviation*)
=QUARTIL(A1:A10;číslo kvartilu)

- kvantilová míra variability
- použití u silně nenormálních souborů či při výskytu extrémů

$$Q = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2}$$

INTERKVARTILOVÉ ROZPĚTÍ *IQR* nebo R_F (angl. *interquartile range*)

- rozdíl horního a dolního kvartilu

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

Nejpoužívanější charakteristiky tvaru

- vyjadřují odchylku v rozložení četností hodnot (vůči normálnímu r.)

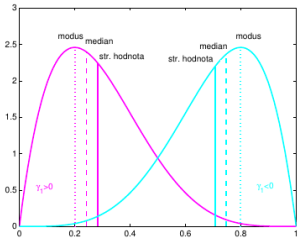
ŠIKMOST (angl. *skewness*) nastává pokud je v souboru

více hodnot větších než menších ve srovnání se střední hodnotou
- levostranná šikmost

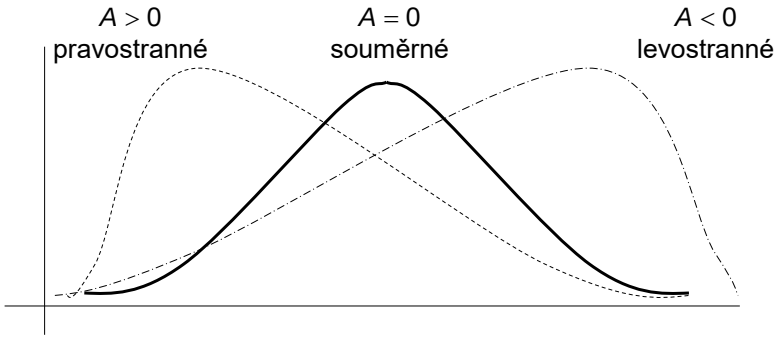
více hodnot menších než větších ve srovnání se střední hodnotou
- pravostranná šikmost

vyjadřuje se pomocí koeficientu šikmosti *A* nebo γ

$$A = \frac{\sum_i (x_i - \mu)^3}{n \cdot \sigma^3}$$



ŠIKMOST



pokud je koeficient šikmosti A

mezi −0,5 a 0,5 $\Rightarrow$ data jsou poměrně symetrická

mezi −1 a −0,5 nebo 0,5 a 1 $\Rightarrow$ data jsou mírně zešikmená

menší než −1 nebo větší než 1 $\Rightarrow$ data jsou velmi zešikmená

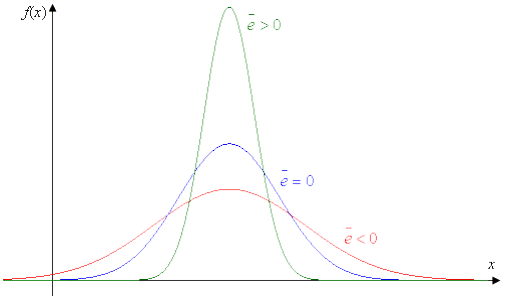
ŠPIČATOST (angl. *kurtosis*) nastává pokud jsou v souboru

hodnoty koncentrovány dominantně u jedné hodnoty
- špičatý soubor

hodnoty více rozptýlené než u normálního rozdělení
- plochý soubor

vyjadřuje se pomocí koeficientu špičatosti E

$$E = \frac{\sum_i (x_i - \mu)^4}{n \cdot \sigma^4} - 3$$



doplněk č. 1

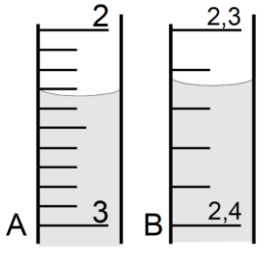
Zaokrouhlování výsledků měření

Zaokrouhlování výsledků

Každý výsledek by měl obsahovat pouze tzv. platné číslice, měl by tedy mít vztah k realitě!

Platnými číslicemi výsledku jsou všechny jisté číslice a jedna nejistá.

- jisté číslice lze odečíst přímo z kalibrované stupnice
- nejistá (odhadnutá) číslice leží mezi dvěma kalibračními značkami



V = 2,33 ml V = 2,328 ml

A co nula?

pokud je nula obklopena jinými číslicemi, je platnou číslicí (50,22 ml)

pokud pouze indikuje řád čísla, je číslicí neplatnou (0,05022 ml)

Koncová nula je platnou číslicí udává-li přesnost údaje, např. 7,0000 g

Zaokrouhlování výsledků

Aby měl výsledek vztah k realitě, musí být řádně zaokrouhlen, aby obsahoval pouze platné číslice!

pravidla určení platných či desetinných míst:

1) při sčítání a odečítání výsledek zaokrouhlujeme na stejný počet desetinných míst, kolik jich má ve výpočtu číslo s nejmenším počtem desetinných míst

$$2{,}005 + 7{,}1 + 0{,}02 = 9{,}125 = 9{,}1$$

2) při násobení či dělení zaokrouhlujeme výsledek tak, aby obsahoval stejný počet platných číslic jako číslo ve výpočtu s nejmenším počtem platných číslic

$$12 \cdot 8{,}03/10{,}00 = 9{,}648 = 9{,}6$$

53

Zaokrouhlování výsledků

3) při kombinacích (sčítání se například kombinuje s dělením) postupujeme dle výše uvedených pravidel, ale v mezivýsledcích používáme vždy o jednu platnou číslici navíc (tzv. ochranná číslice), abychom zabránili chybám při postupném zaokrouhlování. Oba mezivýsledky sice zapíšeme s jednou (ochrannou) číslicí navíc, ale finální zaokrouhlení je provedeno na základě počtu platných číslic, či desetinných míst mezivýsledků bez ochranných číslic.

$$(17{,}7/22{,}123) \cdot (321{,}3 - 12{,}789) = 0{,}8001 \cdot 308{,}51 = 247$$

4) Pokud logaritmujeme číslo mající *N* platných číslic, pak by výsledek měl obsahovat *N* desetinných míst.

$$\log 0{,}00056 = -3{,}25 \quad \text{či} \quad \ln 4{,}756 \cdot 10^5 = 13{,}0723$$

5) Pokud umocňujeme desítkový či exponenciální základ exponentem majícím *N* desetinných míst, pak by výsledek měl obsahovat *N* platných míst.

$$10^{2{,}891} = 778 \quad \text{či} \quad e^{-2{,}57} = 0{,}077$$

54

Pozor!

4,59/1,02 = 4,5 (toto ukáže kalkulačka)

Výsledek však musí být uveden jako 4,50, aby byl vyjádřen správným počtem platných číslic.

Čísla definovaná jako přesná čísla neovlivňují počet platných číslic ve výsledku. Pokud je poloměr kruhu například r = 2,22 cm, pak je průměr kruhu.

$$d = 2 \cdot r = 4{,}44 \text{ cm a nikoliv 4 cm}$$

pravidla zaokrouhlování:

1) poslední uváděná číslice výsledku se zvýší o jednotku, pokud za touto číslicí následuje číslice > 5 nebo číslice 5 následovaná dalšími číslicemi
2) poslední uváděná číslice výsledku se nemění, pokud za touto číslicí následuje číslice menší než 5
3) končí-li zaokrouhlované číslo číslicí 5, zaokrouhlí se tak, aby se poslední uváděná číslice zvýšila o jednotku (standard) nebo zaokrouhlením s preferencí sudé číslice, tj. snažíme se zaokrouhlovat tak, aby poslední uváděná číslice byla vždy sudá – norma IEEE 754

<u>Příklad:</u> 1,5; 1,5; 4,5; 2,5; 1,5 → 2; 2; 5; 3; 2

1,5; 1,5; 4,5; 2,5; 1,5 → 2; 2; 4; 2; 2

<u>Příklad:</u> Zaokrouhlete uvedená čísla na celá: 1,6; 1,4; 1,5; 2,5; 2,501; 1,50

Řešení: 2; 1; 2; 2(3); 3; 2

Konec

doplněk č. 2

Calculating percentile - Excel vs online

Asked 2 years, 7 months ago Active 2 years, 7 months ago Viewed 583 times

1

I have a set of data say {4,7,7,10,10,12,12,14,15,67} and i want to know the 95th Percentile. I used Excel and Online calculator.

Both gave different answers.

In Excel, formula i used : `=PERCENTILE.INC(A1:A10,0.95)` and result = `43.6`

But this online percentile calculator yielded a result of `67`

Which one is right?

excel excel-formula percentile vba

ukázka správného zaokrouhlování

Následující hodnoty v tabulce představují životnost spektrofotometrických přístrojů stejného výrobce a typu v tisících hodin. Vypočtěte všechny kvartily a aritmetický průměr.

12,6	12,9	13,3	13,4	13,6	13,7	13,9	14,3	14,3	14,7
13,1	14,2	14,1	14,1	13,2	15,1	14,2	12,6	13,9	13,7
13,4	13,9	13,4	13,1	13,1	13,6	12,6	13,3	13,8	14,1
13,5	13,0	12,2	13,8	13,5	13,5	14,1	14,5	13,2	13,0
13,7	11,9	11,6	11,6	13,7	15,0	14,5	14,4	13,2	13,1

výpočet aritmetického průměru:

1) provedeme součet vstupních čísel, výsledek musí být uveden na jedno desetinné místo, tedy 676,2 – viz dané pravidlo

2) získaný součet dělíme číslem 50 (přesným číslem), výsledek musí být uveden na čtyři platná místa, tedy 13,52 a nikoliv 13,5 či 13,524

ukázka správného zaokrouhlování u funkce LOG(x) a 10^x

	0,0343		log(0,0343) = -1,46471		-1,465
exp. údaj	0,0344	⇒	log(0,0344) = -1,46344	⇒ zaokrouhlení	-1,463
	0,0345		log(0,0345) = -1,46218		-1,462
	0,0346		log(0,0346) = -1,46092		-1,461

nejistota údaje nejistota údaje

opačná operace - operace 10^x

-1,465		$10^{-1,465}$ = 0,03428		0,0343
-1,463	⇒	$10^{-1,463}$ = 0,03443	⇒ zaokrouhlení	0,0344
-1,462		$10^{-1,462}$ = 0,03451		0,0345
-1,461		$10^{-1,461}$ = 0,03459		0,0346

ukázka správného zaokrouhlování u funkce LN(x) a EXP(x)

	3		
exp. údaj	4	žádné desetinné místo	⇒ žádná platná číslice pro EXP(x)
	5		

Co s tím?

	3		0,3 × 10	
exp. údaj	4	lze přepsat do tvaru	0,4 × 10	⇒ 1 platná číslice pro EXP(x)
	5		0,5 × 10	

exp(3) = 20,0855		2 × 10
exp(4) = 54,5982	⇒ zaokrouhlení	5 × 10
exp(5) = 148,413		1 × 10^2