

Manuál pro zaokrouhlování

k předmětu Pravděpodobnost a Statistika (PS)

Michal Běreš, Martina Litschmannová

19. března 2019

Obsah

1	Úvod	2
2	Obecné poznámky	2
2.1	Typy zaokrouhlování	2
3	Bodové odhady (Explorační analýza)	2
3.1	Směrodatná odchylka s	2
3.2	Míry polohy: výběrový průměr $\bar{x}$, medián $\tilde{x}_{0,5}$, horní a dolní kvartil $\tilde{x}_{0,75}, \tilde{x}_{0,25}$	3
3.3	Minimum a maximum	3
3.4	Vnitřní hradby	3
3.5	Výběrová šikmost a špičatost	3
3.6	Variační koeficient	3
3.7	Relativní četnosti	3
3.8	$\sigma, 2\sigma, 3\sigma$ intervaly dle pravidla 3σ /Čebyševovy nerovnosti	3
3.9	Poměr rizik $\widehat{RR}$ a poměr šancí $\widehat{OR}$	3
4	Intervalové odhady	4
4.1	Způsob zaokrouhlování	4
4.2	Přesnost zaokrouhlování	4
5	Ostatní	4
5.1	p-hodnoty	4
5.2	Pravděpodobnosti (analyticky - Bayes, apod.)	4
5.3	Požadovaná velikost výběru	4
5.4	Očekávané četnosti pro χ^2 test	4
6	Přílohy - pravidlo 3σ a Čebyševova nerovnost	4
7	Přílohy - značení statistik pro populaci a výběr	5

1 Úvod

Pravidla pro zaokrouhlování prezentovaná v tomto textu jsou zaměřená především na data pocházející ze standardních statistických šetření o velikosti do cca 1000 vzorků a na prezentování základních číselných statistik. V případech simulovaných dat a řádově větších výběrů se můžou pravidla lišit (zejména požadovaná přesnost).

Zaokrouhlování má dva základní významy:

- nezpůsobovat dojem falešné přesnosti (př. naměřili jsme 15,123456, odchylka měření je 10),
- vizuální stránka a čitelnost dat.

Existuje více různých pravidel pro zaokrouhlování, vždy však mají výše zmíněné cíle.

2 Obecné poznámky

- Pokud při zaokrouhlení čísla vyjde **jedna nebo více nejnižších cifer 0**, stejně je zapíšeme. Dáváme tak čtenáři najevo, na jakou přesnost jsme zaokrouhlovali.

Ukázka – zaokrouhlení na tisíce: 21,9996 → 22,000; 1,10023 → 1,100

- Pokud zaokrouhluje **více porovnatelných hodnot** (např. směrodatné odchylky kapacity baterií více různých výrobců, intervalové odhady střední hodnoty kapacity baterií různých výrobců, relativní četnosti výskytu vadných výrobků u různých výrobců ...), volíme **stejnou přesnost** zaokrouhlování **rovnou nejvyšší přesnosti** mezi zaokrouhlovanými hodnotami.

Ukázka – společné zaokrouhlení (nahoru) směrodatné odchylky na 3 platné číslice, každé zvlášť:
11,5396 → 11,6; 9,1023 → 9,11 ⇒ zaokrouhlení při společné prezentaci dat (např. v jedné tabulce, nebo při srovnávání v textu): 11,54; 9,11

- Zaokrouhlování se týká pouze prezentace dat. Tedy **zaokrouhlená data nepoužíváme pro další výpočty**.
- Pokud po zaokrouhlení vyjde malé číslo (v absolutní hodnotě), použijeme **zápis pomocí mocniny 10**. Hranici malého čísla můžeme brát jako 10^{-5} . Nepoužíváme zápis typu 10e-5.
Ukázka: 0,000001234 → $1,234 \cdot 10^{-7}$; -0,0000054 → $-5,4 \cdot 10^{-6}$

2.1 Typy zaokrouhlování

V následujících „definicích“ budeme předpokládat zaokrouhlování na jednotky (celá čísla). Zaokrouhlování na jinou přesnost (desítky, desetiny, tisíce, ...) provádíme obdobně, pouze místo nejbližších celých čísel hledáme nejbližší čísla s předepsanou přesností.

Aritmetické zaokrouhlení (round) - výsledkem je nejbližší celé číslo. Problém může nastat, pokud existují taková čísla dvě = na pozici desetin je 5 a další menší cifry jsou nulové, v takovémto případě volíme sudou z těchto dvou nejbližších čísel (IEC 60559, IEEE 754).

Ukázka: 15,4351 → 15; 15,5001 → 16; 15,5 → 16; -15,4351 → -15; -15,5001 → -16; -15,5 → -16; 0,5 → 0; -0,5 → 0; 1,5 → 2

Nahoru (ceil, ceiling) - výsledkem je nejbližší celé číslo, které je větší nebo rovno zaokrouhlovanému číslu.

Ukázka: 1,0001 → 2; 5,9999 → 6; -1,0001 → -1; -5,9999 → -5

Dolů (floor) - výsledkem je nejbližší celé číslo, které je menší nebo rovno zaokrouhlovanému číslu.

Ukázka: 1,0001 → 1; 5,9999 → 5; -1,0001 → -2; -5,9999 → -6

V následujícím textu, pokud nebude řečeno jinak, volíme aritmetické zaokrouhlování.

3 Bodové odhady (Explorační analýza)

3.1 Směrodatná odchylka s

Zaokrouhluje vždy **nahoru** a na stanovený počet **platných cifer**. Počet platných cifer značí počet cifer, které chceme nechat počínaje nejvyšší (vzhledem k pozici - ..., setiny, desetiny, jednotky, desítky, ...) nenulovou cifrou. Počet platných cifer volíme především vzhledem k velikosti výběru, viz Tabulka 1.

Ukázka – na 2 platné cifry: 123,468 → 130; 0,01234 → 0,013; 1,6589 → 1,7; 1,00123 → 1,1
Ukázka – na 3 platné cifry: 123,468 → 124; 0,01234 → 0,0124; 1,6589 → 1,66; 1,00123 → 1,01

velikost výběru	$\langle 2; 10 \rangle$	$\langle 10; 30 \rangle$	$\langle 30; 2000 \rangle$	$\langle 2000; \dots \rangle$
počet platných cifer	1	2	3	4

Tabulka 1: Volba počtu platných cifer pro zaokrouhlení s

3.2 Míry polohy: výběrový průměr $\bar{x}$, medián $\tilde{x}_{0,5}$, horní a dolní kvartil $\tilde{x}_{0,75}, \tilde{x}_{0,25}$

Zaokrouhlujeme klasicky, **dle přesnosti směrodatné odchylky** (je-li s na setiny → na setiny apod.). Toto pravidlo se samozřejmě vztahuje obecně na libovolný kvantil $\tilde{x}_{q,q} \in (0; 1)$.

Ukázka – s zaokrouhlena na 1,34 (setiny): 123,468 → 123,47; 0,01234 → 0,01; 1,6589 → 1,66; 1,00123 → 1,00

Ukázka – s zaokrouhlena na 21 (jednotky): 123,468 → 123; 0,01234 → 0; 1,6589 → 2; 1,00123 → 1

3.3 Minimum a maximum

Pokud je to možné, přepisujeme **hodnoty přímo z datového souboru**. Pokud jsou data v datovém souboru s příliš velkou přesností (a tedy je nepraktické je opisovat), volíme zaokrouhlení dle pravidel pro míry polohy. Hranice příliš velké a nepraktické přesnosti může být subjektivní a závislá na kontextu, pro jednoduchost však můžeme jako takovou hranici volit o dvě cifry více než přesnost pro míry polohy.

3.4 Vnitřní hradby

Zaokrouhlujeme na **o jednu cifru vyšší přesnost, než data v datovém souboru**. Tak, abychom mohli u každého záznamu jednoznačně určit polohu vzhledem k vnitřním hradbám. Pokud je taková přesnost příliš velká a nepraktická, můžeme volit stejnou přesnost, jakou volíme při zaokrouhlení minima a maxima.

3.5 Výběrová šikmost a špičatost

Tyto hodnoty používáme především k exploračnímu ověření normality, tedy porovnáváme jejich polohu vzhledem k intervalu $(-2; 2)$. Tedy stačí **zaokrouhlení na desetiny**. Poznámka: pozor na alternativní definice špičatosti, např. R balíček „moments“ vrací nenormované hodnoty špičatosti $(+3$ oproti intervalu $(-2; 2)$).

3.6 Variační koeficient

Zapisujeme **v procentech** a zaokrouhlujeme **na desetiny**.

3.7 Relativní četnosti

Zaokrouhlujeme **na tisíce**. Pokud by výsledek byl touto cestou zaokrouhlen na 0, zaokrouhlíme na jednu platnou cifru. Poznámka: v některých případech můžeme výsledky vycházející 0 při zaokrouhlení na tisíce zapisovat jako $< 0,001$, toto použijeme například v případech kdy „přesně zapsaná“ relativní četnost (menší než tisícina) nedává vzhledem k dosažitelné přesnosti význam.

Ukázka: 0,46812453 → 0,468; 0,0000123 → 0,00001 = $1 \cdot 10^{-5}$; 0,0000000658 → 0,00000007 = $7 \cdot 10^{-8}$

3.8 $\sigma, 2\sigma, 3\sigma$ intervaly dle pravidla 3σ /Čebyševovy nerovnosti

Zaokrouhlujeme **stejně jako míry polohy** (průměr, medián, ...). $\sigma, 2\sigma, 3\sigma$ intervaly a pravděpodobnosti příslušnosti viz Tabulka 2.

3.9 Poměr rizik $\widehat{RR}$ a poměr šancí $\widehat{OR}$

Zaokrouhlujeme **na setiny**.

4 Intervalové odhady

4.1 Způsob zaokrouhlování

Při zaokrouhlování mezi intervalového odhadu nechceme narušit „statistickou pravdivost odhadu“. Například intervalový odhad pro střední hodnotu (např. dle t-testu, pokud jsou data z normálního rozdělení) $P(a < \mu < b) = 1 - \alpha$ ve skutečnosti říká $P(a < \mu < b) \geq 1 - \alpha$, pokud zaokrouhlíme a na $\tilde{a}$ a b na $\tilde{b}$ musí stále platit $P(\tilde{a} < \mu < \tilde{b}) \geq 1 - \alpha$. Tedy chceme aby $\tilde{a} \leq a$ a $b \leq \tilde{b}$.

- **Dolní mez** intervalového intervalu **zaokrouhlujeme dolů** a
- **horní mez** intervalového odhadu **zaokrouhlujeme nahoru**.

Ukázka (zaokrouhlení na setiny): $(7,36999; 8,42001) \rightarrow (7,36; 8,43); (7,00999; 9,99001) \rightarrow (7,00; 10,00)$

4.2 Přesnost zaokrouhlování

Zaokrouhlujeme na stejnou **přesnost jako přesnost bodových odhadů** (viz sekce 3). V případě, že pracujeme s velkými datovými soubory, můžeme narazit na intervalové odhady s výrazně vyšší přesností, která přesahuje původní přesnost bodových odhadů. Toto se projeví pouze drobným/žádným rozdílem mezi horní a dolní mezí intervalového odhadu. V těchto případech volíme vyšší přesnost zaokrouhlení intervalových odhadů a to tak, ať je **rozdíl mezi horní a dolní mezí intervalového odhadu alespoň o jeden řád vyšší než přesnost zaokrouhlování**.

Ukázky (v případě, že by přesnost převzatá z bodového odhadu nebyla dostatečná): $(1,123456789; 1,1235567) \rightarrow$ rozdíl je 0,0001000 $\rightarrow$ řád rozdílu jsou deseti-tisíciny $\rightarrow$ zaokrouhlujeme na sto-tisíciny $(1,12345; 1,12356); (956132; 956567) \rightarrow$ rozdíl je 435 $\rightarrow$ řád rozdílu jsou stovky $\rightarrow$ zaokrouhlujeme na desítky $(956130; 956570)$

5 Ostatní

5.1 p-hodnoty

Výsledek testu hypotéz vyjádřený p-hodnotou zaokrouhlujeme **na tisíciny**. Pokud by hodnota po zaokrouhlení na tisíciny byla 0, píšeme p-hodnota $<< 0,001$. Dále pokud by při zaokrouhlení vyšla p-hodnota 1,000 zapíšeme výsledek jako p-hodnota $> 0,999$.

5.2 Pravděpodobnosti (analyticky - Bayes, apod.)

Při počítání pravděpodobnosti analyticky můžeme výsledek zapsat jak pomocí zlomku, tak pomocí desetinného zápisu.

- V případě zlomku musíme použít zlomek v základním tvaru. Ideální je tento zápis použít, pokud jmenovatel umožňuje dobrou představu o velikosti čísla (tj. 2,4,5,10,100,...).
- V případě desetinného zápisu zaokrouhlujeme na tisíciny, případně tak, aby byly alespoň 2 nenulové cifry.

5.3 Požadovaná velikost výběru

Pokud vypočítáváme požadovanou velikost výběru (např. při ověření předpokladů binomického testu) zaokrouhlujeme **nahoru na jednotky**.

5.4 Očekávané četnosti pro χ^2 test

Zaokrouhlujeme **na desetiny**.

6 Přílohy - pravidlo 3σ a Čebyševova nerovnost

Poznámka: tyto intervaly se formálně konstruují pomocí střední hodnoty a směrodatné odchylky: $(\mu - k\sigma; \mu + k\sigma)$, $k = 1, 2, 3$. Ty však běžně nemáme k dispozici a proto používáme jejich odhady (průměr a výběrovou směrodatnou odchylku).

interval	3σ pravidlo	Čebyševova nerovnost
$\sigma: \langle \bar{x} - s; \bar{x} + s \rangle$	68,27 %	0,00 %
$2\sigma: \langle \bar{x} - 2s; \bar{x} + 2s \rangle$	95,45 %	75,00 %
$3\sigma: \langle \bar{x} - 3s; \bar{x} + 3s \rangle$	99,73 %	88,89 %

Tabulka 2: Dolní odhady pravděpodobnosti (v procentech), že naměřená data leží v $\sigma, 2\sigma, 3\sigma$ intervalu

7 Přílohy - značení statistik pro populaci a výběr

	pro celou populaci	výběrová alternativa
střední hodnota/průměr	μ	$\bar{x}$
směrodatná odchylka/výběrová směrodatná odchylka	σ	s
rozptyl/výběrový rozptyl	σ^2	s^2
kvantil/výběrový kvantil	$x_{q,q} \in (0; 1)$	$\tilde{x}_{q,q} \in (0; 1)$
šikmost/výběrová šikmost	α	a
špičatost/výběrová špičatost	β	b
- /vnitřní hradby - dolní mez	-	h_D
- /vnitřní hradby - horní mez	-	h_h
- /maximum	-	x_{max}
- /minimum	-	x_{min}
- /variační koeficient	-	V_x
pravděpodobnost/relativní četnost	π	p
poměr rizik/výběrový poměr rizik	RR	$\widehat{RR}$
poměr šancí/výběrový poměr šancí	OR	$\widehat{OR}$

Tabulka 3: Značení parametrů populace a jejich výběrových alternativ

Reference

- [1] M. Litschmannová: *Úvod do statistiky*. URL: <http://mi21.vsb.cz/modul/uvod-do-statistiky>
- [2] M. Litschmannová: *Přednášky k předmětu Pravděpodobnost a statistika*. URL: <http://k470.vsb.cz/litschmannova/vyuka/statistika/harmonogram/>
- [3] T. J. Cole: *Too many digits: the presentation of numerical data*. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4483789/pdf/archdischild-2014-307149.pdf>
- [4] Bluman, McGraw-Hill: *Rounding Rules for Statistics1, Elementary Statistics: A Step By Step Approach 9th edition*. URL: <https://www.saddleback.edu/faculty/pquigley/math10/rounding.pdf>
- [5] A. Thompson, B. N. Taylor: *Guide for the Use of the International System of Units (SI)*. URL: <https://physics.nist.gov/cuu/pdf/sp811.pdf>