

Pravděpodobnost a statistika

Úvod do popisné statistiky

Vilém Vychodil

KMI/PRAS, Přednáška 1

Vytvořeno v rámci projektu 2963/2011 FRVŠ

Přednáška 1: Přehled

1 Úvodní pojmy:

- co je a čím se zabývají pravděpodobnost a statistika,
- náhodné pokusy, elementární jevy, náhodné veličiny,
- populace, výběrové soubory.

2 Grafické metody popisu rozložení dat:

- absolutní a relativní četnost,
- frekvenční tabulky,
- histogramy,
- intervalové rozdělení četností.







3 Míry centrální tendence a míry rozptýlenosti:

- výběrový průměr,
- výběrový rozptyl a směrodatná odchylka,
- zkeslené a nezkreslené odhady populačního rozptylu,
- empirické pravidlo.

Přehled kursu

- 1 Úvod do popisné statistiky
- 2 Úvod do analýzy závislostí
- 3 Pravděpodobnost
- 4 Podmíněná pravděpodobnost, nezávislost jevů, Bayesova věta
- 5 Diskrétní náhodné veličiny
- 6 Diskrétní rozdělení: binomické, geometrické, Poissonovo
- 7 Spojité náhodné veličiny
- 8 Vícerozměrné náhodné veličiny
- 9 Normální rozdělení, centrální limitní věta
- 10 Bodové odhady a intervaly spolehlivosti
- 11 Testování statistických hypotéz

Literatura

-  Capinski M., Zastawniak T. J.: *Probability Through Problems*
Springer 2001, ISBN 978-0-387-95063-1.
-  Devore J. L.: *Probability and Statistics for Engineering and the Sciences*
Duxbury Press, 7. vydání 2008, ISBN 978-0-495-55744-9.
-  Gentle J. E.: *Random Number Generation and Monte Carlo Methods*
Springer 2004, ISBN 978-0-387-00178-4.
-  Hendl, J.: *Přehled statistických metod zpracování dat*
Portál, Praha 2006, ISBN 978-80-7367-123-5
-  Hogg R. V., Tanis E. A.: *Probability and Statistical Inference*
Prentice Hall; 7. vydání 2005, ISBN 978-0-13-146413-1.
-  Johnson J. L.: *Probability and Statistics for Computer Science*
Wiley-Interscience 2008, ISBN 978-0-470-38342-1.

Pravděpodobnost a statistika

Mathematické disciplíny, zabývající se následujícími problémy:

- **Pravděpodobnostní modely**

- modely experimentů jejichž výsledky nemohou být s určitostí predikovány
- zjednodušené modely komplexních systémů

- **Teorie odhadu**

- odhady hodnot veličin na základě opakovaných měření

- **Statistická inference**

- predikce, stanovení intervalů spolehlivosti
- testování statistických hypotéz

- **Analyza závislostí v datech**

- metody rozpoznávání častých/obvyklých/zajímavých vzorů v datech
- data mining (faktorová analýza, dekompozice dat)

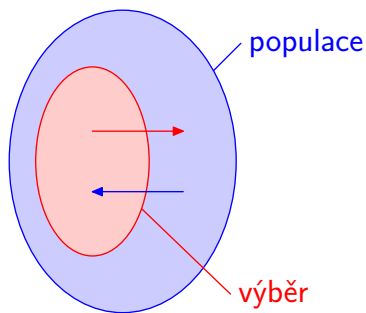
- a další . . .

Statistická inference

Statistická inference = usuzování o populacích na základě výběrů:

- **populace** – obecný pojem (rodiny v ČR, vyrobené elektronické součástky)
- **výběr** – několik vybraných prvků populace (několik rodin / součástek)

Cíl statistické inference: udělat věrohodný závěr o populaci na základě výběru.



populace: rodiny v ČR

populační parametr: počet dětí

cenzus: průměrný počet dětí v populaci

výběr: 100 vybraných rodin

výběrová statistika: průměrný počet dětí (100)

statistická analýza dat (popisná statistika)

inference (hodnoty neznámého parametru)

Náhodný pokus

Definice (Náhodný pokus a jeho výsledek)

Náhodný pokus je činnost probíhající pod vlivem náhody a jehož výsledek není plně určen podmínkami, za kterých je prováděn. Každý **náhodný pokus** (angl.: *random experiment*) končí výsledkem, který je nazýván **elementární jev** (angl.: *outcome*).

Dále předpokládáme, že

- náhodný pokus může být libovolně *opakován*,
- výsledek náhodného pokusu je *nejistý dokud není pokus dokončen*,
- předpokládáme, že všechny možné výsledky náhodného pokusu jde vymezit:

Definice (Prostor elementárních jevů Ω)

Množina všech elementární jevů náhodného pokusu, o který se zajímáme, se označuje Ω a nazývá se **prostor (elementárních jevů)**, angl.: *outcome space*.

Příklady

- Jsou vrženy dvě kostky; zajímáme se o součet teček na obou kostkách.

$$\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\} .$$

- Každý ze šesti studentů zvolí číslo od nuly do 52; zajímá nás, jestli aspoň dvě z těchto čísel jsou shodná (označíme A) či nikoliv (označíme B).

$$\Omega = \{A, B\} .$$

- Házíme mincí tak dlouho, dokud neuvidíme orla; zajímáme se o počet hodů.

$$\Omega = \{1, 2, 3, 4, \dots\} = \mathbb{N} .$$

- Zkoumáme hmotnost produktů, rychlost jízdy, čas, a pod.

$$\Omega \subseteq \mathbb{R} , \text{ například } \Omega = (50, 400) .$$

Příklad

Náhodný pokus: Jsou vrženy dvě různě barevné kostky; zajímáme se počty teček, které padnou na obou kostkách.

$$\Omega = \{\langle x, y \rangle \mid x, y \in \{1, 2, 3, 4, 5, 6\}\} = \{\langle 1, 1 \rangle, \langle 1, 2 \rangle, \dots, \langle 6, 6 \rangle\}.$$

Ω je druhá kartézská mocnina $\{1, 2, 3, 4, 5, 6\}$, to jest $|\Omega| = 6 \cdot 6 = 36$. Uspořádaná dvojice $\langle x, y \rangle \in \Omega$ reprezentuje elementární jev: „první kostka má na horní straně x teček a druhá kostka má na horní straně y teček“.

Po provedení experimentu můžeme provést dodatečná měření na Ω :

- 1 součet teček na obou kostkách,
- 2 větší hodnota z hodnot na obou kostkách,
- 3 průměr hodnot na obou kostkách, ...

Výsledek měření = číselná hodnota z \mathbb{R} .

Náhodný pokus \implies výsledek (elementární jev) \implies výsledek měření (nejistý)

Náhodné veličiny (neformálně)

Náhodná veličina / proměnná (angl.: *random variable*)

- zobrazení $X: \Omega \rightarrow \mathbb{R}$ (náhodné veličiny označujeme X, Y, Z, \dots)
- formalizuje měření na výsledku náhodného pokusu (PŘEDNÁŠKA 5)

Základní problém statistického usuzování:

Sledujeme hodnoty, kterých nabývá náhodná veličina po opakování náhodného pokusu. Některé z hodnot se vyskytnou častěji než jiné – chceme tento fenomén formalizovat (kvantifikovat).

Základní cíle statistické inference:

- Popsat *rozdělení pravděpodobnosti náhodné veličiny*: popsat poměrné množství případů, kdy náhodný pokus a následné měření skončí danými hodnotami.
- Rozdělení pravděpodobnosti jsou obvykle odhadovány pomocí *výběrových souborů* obsahujících pozorované hodnoty náhodné veličiny.

Základní populace a výběrový soubor

Definice (Základní populace)

Množina všech hodnot, které mohou být teoreticky zaznamenány jako výsledek náhodného pokusu nebo výsledek následného měření, se nazývá (**základní**) **populace**, *angl.: population*.

(*Základní*) *populace* = statistický protějšek termínu *prostor elementárních jevů*.

Definice (Výběrový soubor / výběr / statistický výběr / vzorek)

Uvažujme náhodný pokus s prostorem Ω . Pokud je náhodný pokus opakován n -krát, pak se posloupnost zaznamenaných výsledků x_1, \dots, x_n nazývá **výběrový soubor** (nebo jen **výběr**) velikosti n z populace Ω , *angl.: (statistical) sample*.

Zjednodušení: Prostory a výběrové soubory uvažujeme *číselné*, tedy $\Omega \subseteq \mathbb{R}$.

Absolutní četnost a relativní četnost

Definice (Absolutní četnost, relativní četnost)

Uvažujme náhodný pokus s prostorem Ω . Pokud je náhodný pokus opakován n -krát a f je počet výskytů výsledku $x \in \Omega$, pak se f nazývá (**absolutní**) **četnost** x , angl.: *frequency*. Zlomek $\frac{f}{n}$ se nazývá **relativní četnost** x , angl.: *relative frequency*.

Poznámka: „Frekventistická interpretace pravděpodobnosti“

Relativní četnost x je (obvykle) nestabilní při malém počtu opakování náhodného pokusu, ale má tendenci se stabilizovat a blížit hodnotě p , pokud n roste. Hodnota p se interpretuje jako *pravděpodobnost elementárního jevu* (PŘEDNÁŠKA 3).

Tabulky absolutní / relativní četnosti

- standardní metoda zápisu (relativních) četností hodnot ve výběrech

Příklad (Výběrový soubor a tabulka absolutních/relativních četností)

Uvažujme výběrový soubor obsahující počty dětí ve 100 vybraných rodinách:

4 6 2 7 2 9 3 4 2 1 5 4 1 3 2 5 2 2 3 6 3 3 5 2 3
1 5 2 2 3 4 0 4 2 0 4 2 4 3 5 0 3 4 5 1 3 7 4 2 2
4 3 5 3 6 2 3 3 2 9 4 4 2 5 2 2 4 2 2 3 1 4 3 3 2
5 6 3 2 2 3 3 3 2 2 4 2 4 8 2 2 5 2 4 3 6 2 3 1 5

počet dětí	tabelace	četnost	relativní četnost
0			
1			
2			
⋮			
8			
9			

Příklad (Výběrový soubor a tabulka absolutních/relativních četností)

Uvažujme výběrový soubor obsahující počty dětí ve 100 vybraných rodinách:

4 6 2 7 2 9 3 4 2 1 5 4 1 3 2 5 2 2 3 6 3 3 5 2 3
1 5 2 2 3 4 0 4 2 0 4 2 4 3 5 0 3 4 5 1 3 7 4 2 2
4 3 5 3 6 2 3 3 2 9 4 4 2 5 2 2 4 2 2 3 1 4 3 3 2
5 6 3 2 2 3 3 3 2 2 4 2 4 8 2 2 5 2 4 3 6 2 3 1 5

počet dětí	tabelace	četnost	relativní četnost
0			
1			
2			
:			
8			
9			

Příklad (Výběrový soubor a tabulka absolutních/relativních četností)

Uvažujme výběrový soubor obsahující počty dětí ve 100 vybraných rodinách:

4 6 2 7 2 9 3 4 2 1 5 4 1 3 2 5 2 2 3 6 3 3 5 2 3
1 5 2 2 3 4 0 4 2 0 4 2 4 3 5 0 3 4 5 1 3 7 4 2 2
4 3 5 3 6 2 3 3 2 9 4 4 2 5 2 2 4 2 2 3 1 4 3 3 2
5 6 3 2 2 3 3 3 2 2 4 2 4 8 2 2 5 2 4 3 6 2 3 1 5

počet dětí	tabelace	četnost	relativní četnost
0		3	$\frac{3}{100} = 0.03$
1			
2			
⋮			
8			
9			

Příklad (Výběrový soubor a tabulka absolutních/relativních četností)

Uvažujme výběrový soubor obsahující počty dětí ve 100 vybraných rodinách:

4 6 2 7 2 9 3 4 2 ~~1~~ 5 4 ~~1~~ 3 2 5 2 2 3 6 3 3 5 2 3
~~1~~ 5 2 2 3 4 ~~0~~ 4 2 ~~0~~ 4 2 4 3 5 ~~0~~ 3 4 5 ~~1~~ 3 7 4 2 2
 4 3 5 3 6 2 3 3 2 9 4 4 2 5 2 2 4 2 2 3 ~~1~~ 4 3 3 2
 5 6 3 2 2 3 3 3 2 2 4 2 4 8 2 2 5 2 4 3 6 2 3 ~~1~~ 5

počet dětí	tabelace	četnost	relativní četnost
0		3	0.03
1	 		
2			
:			
8			
9			

Příklad (Výběrový soubor a tabulka absolutních/relativních četností)

Uvažujme výběrový soubor obsahující počty dětí ve 100 vybraných rodinách:

4 6 2 7 2 9 3 4 2 ~~X~~ 5 4 ~~X~~ 3 2 5 2 2 3 6 3 3 5 2 3
~~X~~ 5 2 2 3 4 ~~Ø~~ 4 2 ~~Ø~~ 4 2 4 3 5 ~~Ø~~ 3 4 5 ~~X~~ 3 7 4 2 2
4 3 5 3 6 2 3 3 2 9 4 4 2 5 2 2 4 2 2 3 ~~X~~ 4 3 3 2
5 6 3 2 2 3 3 3 2 2 4 2 4 8 2 2 5 2 4 3 6 2 3 ~~X~~ 5

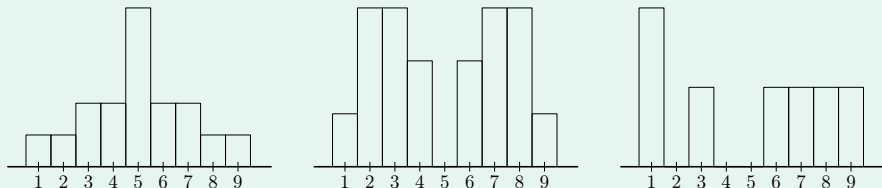
počet dětí	tabelace	četnost	relativní četnost
0		3	0.03
1	 	6	$\frac{6}{100} = 0.06$
2			
⋮			
8			
9			

Grafická reprezentace dat z tabulek četností

Histogram absolutní četnosti / histogram relativní četnosti:

- diagram zakreslený do kartézské roviny,
- pro každý $x \in \Omega$: obdélník s výškou rovnou absolutní/relativní četnosti x ,
- šířka všech obdélníků je konstantní.

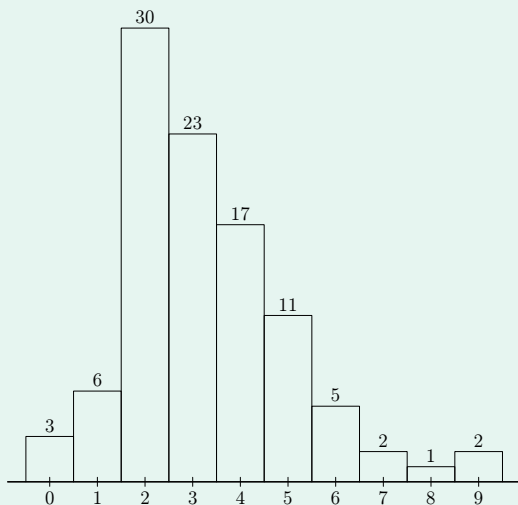
Příklad



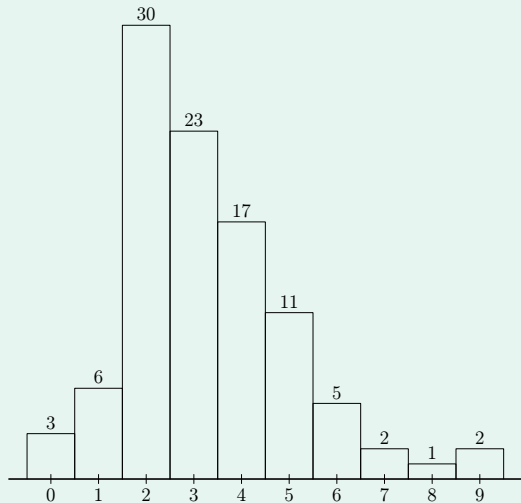
- možné znázorňovat i v jiném tvaru (koláče a podobně)
- tvar histogramu – vztah k hustotě pravděpodobnosti (PŘEDNÁŠKA 7)

Příklad (Tabulka a histogram četnosti)

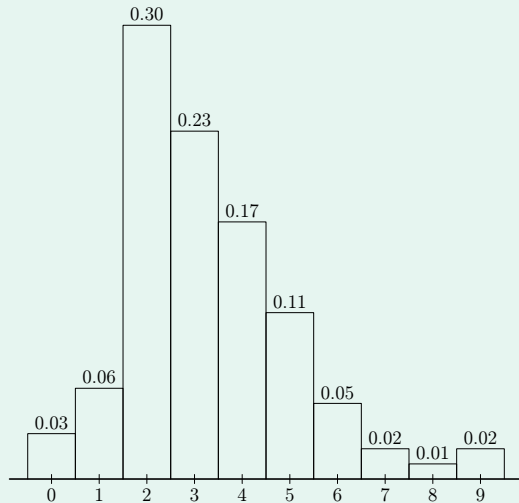
počet dětí	absolutní četnost	relativní četnost
0	3	0.03
1	6	0.06
2	30	0.30
3	23	0.23
4	17	0.17
5	11	0.11
6	5	0.05
7	2	0.02
8	1	0.01
9	2	0.02
$\Sigma :$	100	1.00



Příklad (Histogramy absolutních/relativních četností)



histogram absolutních četností



histogram relativních četností

Příklad (Skupinové rozdělení četností: motivační příklad)

Uvažujme výběrový soubor obsahující hmotnosti 40 produktů stejného typu:

22.38 21.55 22.20 22.55 22.87 24.40 21.65 23.30 23.58 22.37
19.39 23.86 23.28 24.63 22.35 24.23 23.95 23.21 22.11 21.42
25.92 24.00 22.97 23.43 22.72 22.90 23.32 23.58 21.37 22.67
19.84 21.97 20.11 21.06 20.57 22.48 23.60 22.45 21.00 23.82

hmotnost	četnost	relativní četnost
19.39	1	0.025
19.84	1	0.025
20.11	1	0.025
⋮	⋮	⋮
24.63	1	0.025
25.92	1	0.025

K NIČEMU!

Prostor hodnot
je spojitý!

Skupinové (intervalové) rozdělení četností

Pokud je Ω spojitá (např. reálný interval) nebo obsahuje-li Ω mnoho hodnot, rozdělíme Ω na disjunktní podmnožiny (skupiny) a uvažujeme četnosti celých skupin.

Postup:

- 1 Určíme největší (max) a nejmenší (min) hodnotu ve výběru.
Rozdíl $r = max - min$ se nazývá **variační rozpětí**, angl.: *range*.
- 2 Zvolíme **k intervalů**, angl.: *class intervals*, které tvoří rozklad na (min, max) :
$$(c_0, c_1), (c_1, c_2), \dots, (c_{k-1}, c_k)$$

Číslo k volíme obvykle $5 \leq k \leq 20$ (Sturgessovo pravidlo: $k \approx 1 + 3.3 \log n$).
- 3 Čísla c_{i-1} a c_i se nazývají **hranice intervalu** (c_{i-1}, c_i) , angl.: *class boundaries*.
- 4 Číslo $\frac{c_{i-1} + c_i}{2}$ se nazývá **střed intervalu** (c_{i-1}, c_i) , angl.: *class mark*.

Tabulky intervalového rozdělení četností

Meze intervalu (c_{i-1}, c_i) , angl.: *class limits*, jsou hodnoty d_{i-1} a d_i z výběru x_1, \dots, x_n , zapisované (d_{i-1}, d_i) , pro které platí

① $d_{i-1} \leq d_i$ a dále

② $\{x_1, \dots, x_n\} \cap (c_{i-1}, d_{i-1}) = \emptyset$ a $\{x_1, \dots, x_n\} \cap (d_i, c_i) = \emptyset$.

Četnost intervalu (c_{i-1}, c_i) , angl.: *class frequency*, je číslo f_i označující počet hodnot z výběru patřících do (c_{i-1}, c_i) , to jest $f_i = |(c_{i-1}, c_i) \cap \{x_1, \dots, x_n\}|$.

Tabulka absolutních četností:

- řádky: korespondují s *interval* (c_{i-1}, c_i)
- sloupce: *interval*, *meze intervalu*, *střed intervalu*, *četnost intervalu*

Tabulka relativních četností: jako tabulka absolutních četností + sloupec h_i :

$$h_i = \frac{f_i}{n \cdot (c_i - c_{i-1})}, \quad \text{hodnota } \frac{f_i}{n} \text{ je } \textbf{relativní četnost intervalu } (c_{i-1}, c_i).$$

Histogramy intervalového rozdělení četností

Histogram int. rozdělení absolutních četností: pro každý interval (c_{i-1}, c_i) zakreslíme obdélník daný body $[c_{i-1}, 0]$, $[c_i, 0]$, $[c_i, f_i]$ a $[c_{i-1}, f_i]$.

Histogram int. rozdělení relativních četností: pro každý interval (c_{i-1}, c_i) zakreslíme obdélník daný body $[c_{i-1}, 0]$, $[c_i, 0]$, $[c_i, h_i]$ a $[c_{i-1}, h_i]$.

Věta

Obsah všech obdélníků v histogramu int. rozdělení relativních četností je rovna 1.

Důkaz.

$$\sum_{i=1}^k ((c_i - c_{i-1}) \cdot h_i) = \sum_{i=1}^k \frac{(c_i - c_{i-1}) \cdot f_i}{n \cdot (c_i - c_{i-1})} = \sum_{i=1}^k \frac{f_i}{n} = \frac{1}{n} \sum_{i=1}^k f_i = \frac{1}{n} \cdot n = 1.$$



Příklad (Intervalové rozdělení četností s intervaly stejných délek)

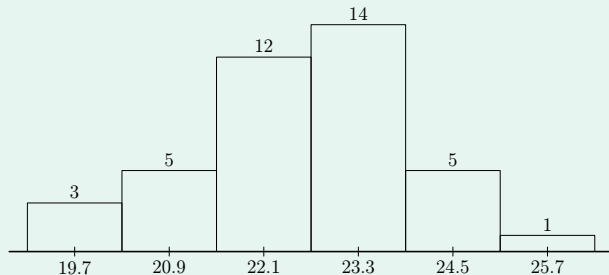
Uvažujme výběrový soubor obsahující hmotnosti 40 produktů stejného typu:

22.38 21.55 22.20 22.55 22.87 24.40 21.65 23.30 23.58 22.37
19.39 23.86 23.28 24.63 22.35 24.23 23.95 23.21 22.11 21.42
25.92 24.00 22.97 23.43 22.72 22.90 23.32 23.58 21.37 22.67
19.84 21.97 20.11 21.06 20.57 22.48 23.60 22.45 21.00 23.82

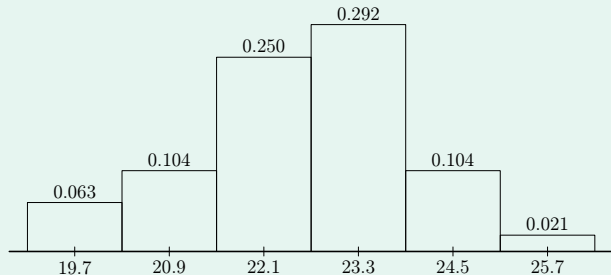
interval	meze	f_i	h_i	střed
(19.1, 20.3)	(19.39, 20.11)	3	0.063	19.7
(20.3, 21.5)	(20.57, 21.42)	5	0.104	20.9
(21.5, 22.7)	(21.55, 22.67)	12	0.250	22.1
(22.7, 23.9)	(22.72, 23.86)	14	0.292	23.3
(23.9, 25.1)	(23.95, 24.63)	5	0.104	24.5
(25.1, 26.3)	(25.92, 25.92)	1	0.021	25.7

Příklad (Intervalové rozdělení četností s intervaly stejných délek)

interval	f_i	střed
(19.1, 20.3)	3	19.7
(20.3, 21.5)	5	20.9
(21.5, 22.7)	12	22.1
(22.7, 23.9)	14	23.3
(23.9, 25.1)	5	24.5
(25.1, 26.3)	1	25.7



interval	h_i	střed
(19.1, 20.3)	0.063	19.7
(20.3, 21.5)	0.104	20.9
(21.5, 22.7)	0.250	22.1
(22.7, 23.9)	0.292	23.3
(23.9, 25.1)	0.104	24.5
(25.1, 26.3)	0.021	25.7



Příklad (Intervalové rozdělení četností s intervaly různých délek)

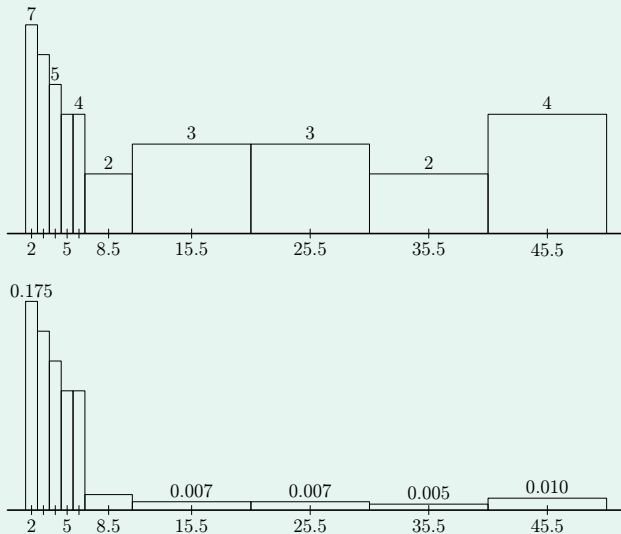
Uvažujme výběrový soubor obsahující počty mrtvých při 40 živelních pohromách:

2 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 5 5
5 5 6 6 6 6 8 9 12 16 18 22 25 29 32 39 41 47 48 50

interval	meze	f_i	h_i
(1.5, 2.5)	(2, 2)	7	0.175
(2.5, 3.5)	(3, 3)	6	0.150
(3.5, 4.5)	(4, 4)	5	0.125
(4.5, 5.5)	(5, 5)	4	0.100
(5.5, 6.5)	(6, 6)	4	0.100
(6.5, 10.5)	(8, 9)	2	0.013
(10.5, 20.5)	(12, 18)	3	0.007
(20.5, 30.5)	(22, 29)	3	0.007
(30.5, 40.5)	(32, 39)	2	0.005
(40.5, 50.5)	(41, 50)	4	0.010

Příklad (Intervalové rozdělení četností s intervaly různých délek)

interval	f_i	h_i
(1.5, 2.5)	7	0.175
(2.5, 3.5)	6	0.150
(3.5, 4.5)	5	0.125
(4.5, 5.5)	4	0.100
(5.5, 6.5)	4	0.100
(6.5, 10.5)	2	0.013
(10.5, 20.5)	3	0.007
(20.5, 30.5)	3	0.007
(30.5, 40.5)	2	0.005
(40.5, 50.5)	4	0.010



Míry centrální tendence: výběrový průměr

Definice (Výběrový (aritmetický) průměr)

Výběrový (aritmetický) průměr \bar{x} výběru x_1, \dots, x_n je číslo definované

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

Poznámky:

- interpretace: \bar{x} = „střední hodnota z výběru“
- \bar{x} nemusí být jednou z hodnot x_1, \dots, x_n
- triviální případy:
 - $n = 1$ (jednoprvkový výběr): $\bar{x} = x_1$;
 - $x_1 = x_2 = \dots = x_n$ (uniformní výběr): $\bar{x} = x_1 = x_2 = \dots = x_n$.
- další typy průměru: geometrický, harmonický (požívané zřídka)

Vlastnosti výběrového průměru

Věta (o výběrovém průměru)

Pro každý výběr x_1, \dots, x_n a jeho průměr \bar{x} platí:

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

❷
$$\sum_{i=1}^n (x_i - \bar{x}) = 0,$$

❸ funkce $f: \mathbb{R} \rightarrow \mathbb{R}$ definovaná $f(x) = \sum_{i=1}^n (x_i - x)^2$

nabývá v bodě \bar{x} svého globálního minima.

Důsledek: \bar{x} minimalizuje hodnotu $\sum_{i=1}^n (x_i - x)^2$.

Důkaz.

První dvě tvrzení jsou zřejmá:

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

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = \sum_{i=1}^n x_i - n\bar{x} = \frac{n}{n} \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

Stačí ověřit, že $f'(\bar{x}) = 0$ a že $f''(\bar{x}) > 0$.

$$f'(x) = \sum_{i=1}^n ((x_i - x)^2)' = -2 \sum_{i=1}^n (x_i - x), \text{ užitím předchozího: } f'(\bar{x}) = -2 \cdot 0 = 0.$$

$$f''(x) = -2 \sum_{i=1}^n (x_i - x)' = -2 \sum_{i=1}^n -1 = -2 \cdot -n = 2n > 0, \text{ to jest: } f''(x) > 0.$$



Příklady (Výběrový průměr)

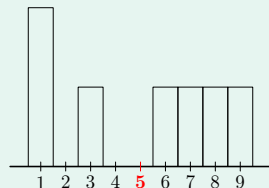
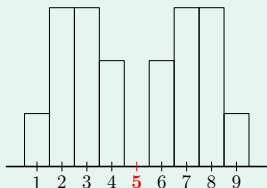
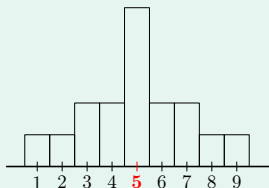
❶ Pro výběr obsahující $n = 5$ hodnot 3, 7, 2, 5 a 3 máme:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{3 + 7 + 2 + 5 + 3}{5} = \frac{20}{5} = 4.$$

❷ Předchozí příklady:

- průměrný počet dětí v rodinách: $\bar{x} = 3.28 \approx 3$,
- průměrná hmotnost produktů: $\bar{x} = 22.6265$,
- průměrné ztráty na životech: $\bar{x} = 12.3 \approx 12$.

❸ Výběrový průměr zaznačený v histogramech:



Míry rozptýlenosti: výběrový rozptyl a směrodatná odchylka

Definice (výběrový rozptyl / výběrová variance / výběrová disperze)

Výběrový rozptyl s^2 výběru x_1, \dots, x_n je číslo definované

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

Definice (výběrová směrodatná odchylka)

Výběrová směrodatná odchylka s výběru x_1, \dots, x_n je číslo definované

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

angl.: *sample variance, sample standard deviation*

Vlastnosti výběrového rozptylu a směrodatné odchylky

Výběrový rozptyl s^2

- míra rozptýlenosti hodnot ve výběru od jeho průměru,
- $s^2 \geq 0$,
- pokud je hodnota s^2 malá, pak je většina hodnot ve výběru blízko \bar{x} ,
- triviální případy:
 - pro $n = 1$ (jednoprvkový výběr) není s^2 definovaná,
 - $s^2 = 0$ právě když $x_1 = x_2 = \dots = x_n = \bar{x}$ (uniformní výběr).

Výběrová směrodatná odchylka s

- podobné vlastnosti a význam jako výběrový rozptyl,
- používá stejné jednotky jako data ve výběru.

Například: data ve výběru v m (metrech) $\implies s^2$ v $m^2 \implies s$ v m .

Interpretace výběrové směrodatné odchylky

Interpretace hodnoty s

Výběrová směrodatná odchylka $s = \sqrt{s^2} \geq 0$ je mírou disperze dat od jejich středu. Hodnota s může být chápána jako horní aproximace „průměrné vzdálenosti hodnot x_1, \dots, x_n od \bar{x} “. Symbolicky:

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

Poznámka

Lze ukázat, že vždy platí $s \geq \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$, ale obecně $s \neq \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$.

Příklad (Výběrová směrodatná odchylka \times průměrná vzdálenost od \bar{x})

Pro výběr obsahující $n = 5$ hodnot 3, 7, 2, 5 a 3 máme $\bar{x} = 4$ a dále:

$$\begin{aligned}s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\&= \frac{(3-4)^2 + (7-4)^2 + (2-4)^2 + (5-4)^2 + (3-4)^2}{5-1} \\&= \frac{(-1)^2 + 3^2 + (-2)^2 + 1^2 + (-1)^2}{4} \\&= \frac{1+9+4+1+1}{4} = \frac{16}{4} = 4; \quad s = \sqrt{s^2} = 2.\end{aligned}$$

Průměrná vzdálenost od výběrového průměru je

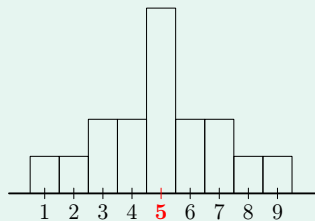
$$\frac{|3-4| + |7-4| + |2-4| + |5-4| + |3-4|}{5} = \frac{1+3+2+1+1}{5} = \frac{8}{5} = 1.6 < 2.$$

Příklad (Výběrový rozptyl a výběrová směrodatná odchylka)

- Předchozí příklady:

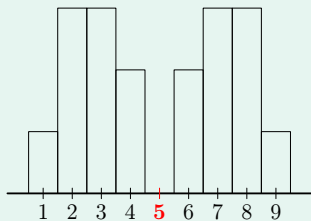
- počet dětí v rodinách: $\bar{x} = 3.28$, $s^2 = 3.113$, $s = 1.764$.
- hmotnost produktů: $\bar{x} = 22.6265$, $s^2 = 1.865$, $s = 1.366$.
- ztráty na životech: $\bar{x} = 12.3$, $s^2 = 215.703$, $s = 14.687$.

- Předchozí histogramy:



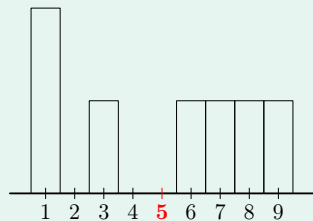
$$s^2 = 4.375$$

$$s = 2.092$$



$$s^2 = 6.706$$

$$s = 2.950$$



$$s^2 = 10.154$$

$$s = 3.187$$

Další vlastnosti rozptylu

Rozptyl definovaný $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ lze rovněž vyjádřit:

Věta (o výpočtovém tvaru s^2)

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

Poznámka:

- $\sum_{i=1}^n x_i^2 - n\bar{x}^2$ v předchozí Větě znamená $\left(\sum_{i=1}^n x_i^2 \right) - n\bar{x}^2$, nikoliv $\sum_{i=1}^n (x_i^2 - n\bar{x}^2)$

Důkaz.

Nejprve ukážeme rovnost $\sum_{i=1}^n (x_i - \bar{x})^2$ a $\sum_{i=1}^n x_i^2 - n\bar{x}^2$. Platí, že

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 \\&= \sum_{i=1}^n x_i^2 - 2\bar{x} \left(n \frac{1}{n} \sum_{i=1}^n x_i \right) + \sum_{i=1}^n \bar{x}^2 \\&= \sum_{i=1}^n x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\&= \sum_{i=1}^n x_i^2 - n\bar{x}^2.\end{aligned}$$

Druhá část tvrzení plyne z následující rovnosti:

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



Příklad

výběr: 3, 7, 2, 5, 3 ($\bar{x} = 4$)

- **výpočet s^2 podle definice:**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1 + 9 + 4 + 1 + 1}{4} = \frac{16}{4} = 4.$$

- **výpočet s^2 užitím předchozí Věty:**

$$n\bar{x}^2 = 5 \cdot 4^2 = 80,$$

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{(3^2 + 7^2 + 2^2 + 5^2 + 3^2) - 80}{4} = \frac{96 - 80}{4} = \frac{16}{4} = 4.$$

Výpočetní složitost algoritmů pro výpočet rozptylu

Algoritmus 1

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

- 1 výpočet \bar{x} (n operací),
- 2 výpočet všech $(x_i - \bar{x})^2$ ($2n$ oper.),
- 3 součet všech $(x_i - \bar{x})^2$ (n oper.),
- 4 podíl výsledku $n-1$ (1 oper.).

Celkem: $4n + 3$ operací

Složitost: $O(n)$

Algoritmus 2

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

- 1 součet čtverců všech x_i ($2n$ operací),
- 2 výpočet \bar{x} (n oper.),
- 3 čtverec předchozího (1 oper.),
- 4 násobení výsledku n (1 oper.),
- 5 odečet dvou výsledků (1 oper.),
- 6 podíl výsledku $n-1$ (1 oper.).

Celkem: $3n + 4$ operací (**zlepšení**)

Složitost: $O(n)$ (**řádově stejná**)

Výběrový rozptyl: otázky o tvaru definice

Diskuse

Výběrový rozptyl jsme zavedli jako míru rozptýlenosti hodnot ve výběru od jeho průměrné hodnoty. Nabízí se otázky:

- Je definice přirozená?
- Proč dělíme sumu čtverců číslem $n - 1$ a ne n ?
- Můžeme nadefinovat rozptyl jako míru *vzájemné rozptýlenosti hodnot* bez explicitního použití výběrového průměru?

Ukážeme, že

- 1 s^2 lze definovat bez explicitního použití \bar{x} ,
- 2 existuje rozumná alternativní definice pro s^2 , která používá dělitel n místo $n - 1$ (obě varianty mají smysl, PŘEDNÁŠKA 10).

Příklad (Motivační příklad pro definici s^2 nepoužívající \bar{x})

Vezmeme předchozí výběr $x_1 = 3$, $x_2 = 7$, $x_3 = 2$, $x_4 = 5$ a $x_5 = 3$.

Tabulka rozdílů hodnot z výběru:

	3	7	2	5	3
3	0	-4	1	-2	0
7	4	0	5	2	4
2	-1	-5	0	-3	-1
5	2	-2	3	0	2
3	0	-4	1	-2	0

Tabulka druhých mocnin rozdílů:

	3	7	2	5	3
3	0	16	1	4	0
7	16	0	25	4	16
2	1	25	0	9	1
5	4	4	9	0	4
3	0	16	1	4	0

- 1 Vypočteme průměr všech hodnot mimo diagonálu z tabulky (napravo).
- 2 Podělíme výsledný průměr dvěma (důvod objasníme později).

V případě našeho příkladu:
$$\frac{160}{2n(n-1)} = \frac{160}{10 \cdot 4} = \frac{160}{40} = 4 = s^2.$$

Zobecnění předchozího příkladu

Postup z předchozího příkladu vede na obecný vzorec:

$$\frac{1}{2n(n-1)} \sum_{i,j=1}^n (x_i - x_j)^2$$

- dále prokážeme, že tato hodnota je rovna s^2
- ve vztahu není použito \bar{x}
- nehodí se na výpočet, asymptotická časová složitost: $O(n^2)$

Lemma (pomocná věta, bude použita později)

$$\sum_{i,j=1}^n (x_i - x_j)^2 = 2n \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).$$

Důkaz.

$$\begin{aligned}\sum_{i,j=1}^n (x_i - x_j)^2 &= \sum_{i,j=1}^n (x_i^2 - 2x_i x_j + x_j^2) = \sum_{i,j=1}^n x_i^2 - 2 \sum_{i,j=1}^n x_i x_j + \sum_{i,j=1}^n x_j^2 \\&= \sum_{i=1}^n n x_i^2 - 2 \sum_{i,j=1}^n x_i x_j + \sum_{j=1}^n n x_j^2 = 2 \sum_{i=1}^n n x_i^2 - 2 \sum_{i,j=1}^n x_i x_j \\&= 2n \sum_{i=1}^n x_i^2 - 2 \sum_{i,j=1}^n x_i x_j = 2n \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n \left(x_i \sum_{j=1}^n x_j \right) \\&= 2n \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n (x_i \cdot n\bar{x}) = 2n \sum_{i=1}^n x_i^2 - 2n\bar{x} \sum_{i=1}^n x_i \\&= 2n \sum_{i=1}^n x_i^2 - 2n\bar{x}n\bar{x} = 2n \sum_{i=1}^n x_i^2 - 2n^2\bar{x}^2 = 2n \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).\end{aligned}$$



Výběrový rozptyl vyjádřený bez použití \bar{x}

Věta (o významu dvojnásobku s^2)

$$s^2 = \frac{1}{2n(n-1)} \sum_{i,j=1}^n (x_i - x_j)^2$$

Důkaz.

Tvrzení dokážeme použitím předchozích pozorování:

$$\begin{aligned} \frac{1}{2n(n-1)} \sum_{i,j=1}^n (x_i - x_j)^2 &= \frac{1}{2n(n-1)} \cdot 2n \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = s^2. \end{aligned}$$



Příklad (Motivace pro „zkreslený odhad rozptylu“)

Zvolme výběr $x_1 = 3$, $x_2 = 7$, $x_3 = 2$, $x_4 = 5$ a $x_5 = 3$.

Tabulka rozdílů hodnot z výběru:

	3	7	2	5	3
3	0	-4	1	-2	0
7	4	0	5	2	4
2	-1	-5	0	-3	-1
5	2	-2	3	0	2
3	0	-4	1	-2	0

Tabulka druhých mocnin rozdílů:

	3	7	2	5	3
3	0	16	1	4	0
7	16	0	25	4	16
2	1	25	0	9	1
5	4	4	9	0	4
3	0	16	1	4	0

- 1 Vypočteme průměr všech hodnot z tabulky (napravo) včetně diagonály.
- 2 Opět podělíme výsledný průměr dvěma.

Zopakováním předchozí úvahy dostaneme:
$$\frac{1}{2n^2} \cdot \sum_{i,j=1}^n (x_i - x_j)^2 .$$

Vlastnosti „zkresleného odhadu rozptylu“

Věta

$$\frac{1}{2n^2} \cdot \sum_{i,j=1}^n (x_i - x_j)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Důkaz.

Dokážeme použitím předchozích tvrzení:

$$\begin{aligned} \frac{1}{2n^2} \cdot \sum_{i,j=1}^n (x_i - x_j)^2 &= \frac{1}{2n^2} \cdot 2n \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \\ &= \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned}$$



Nezkreslený \times zkreslený odhad rozptylu populace

Odhady rozptylu populace:

- uvažujeme (velkou) populaci a (nepoměrně menší) výběr z populace,
- výběrový rozptyl s^2 slouží jako **nezkreslený odhad** rozptylu celé populace;
- hodnota s_n^2 daná vztahem

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

se nazývá **zkreslený odhad** rozptylu populace (PŘEDNÁŠKA 10);

- rozdíl $s^2 - s_n^2 \geq 0$ se nazývá **zkreslení**.

Terminologie (používají se různé názvy)

- nestranný / nezkreslený / nevychýlený odhad, *angl.: unbiased estimate*
- stranný / zkreslený / vychýlený odhad, *angl.: biased estimate*
- zkreslení / vychýlení, *angl.: bias*

Vzájemný vztah s^2 a s_n^2

Zřejmě platí:

$$s^2 = \frac{n}{n-1} \cdot s_n^2,$$

$$s_n^2 = \frac{n-1}{n} \cdot s^2.$$

Vždy platí:

$$s^2 > s_n^2,$$

$$\text{pokud } \sum_{i=1}^n (x_i - \bar{x})^2 > 0.$$

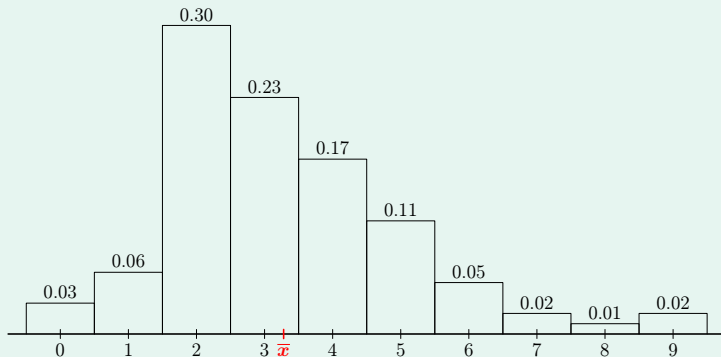
Poznámky o použití s_n^2

- S rostoucím n (obvykle $n > 30$) je rozdíl mezi hodnotami s^2 a s_n^2 zanedbatelný.
- s_n^2 má tendenci zkreslovat hodnotu rozptylu celé populace (PŘEDNÁŠKA 10).
- V případě, že počítáme rozptyl z celé populace, používáme s^2 .

Příklad (Rozdíly mezi nezkrasleným a zkrasleným odhadem rozptylu)

Výběrový soubor počtu dětí ve 100 rodinách:

4 6 2 7 2 9 3 4 2 1 5 4 1 3 2 5 2 2 3 6 3 3 5 2 3
1 5 2 2 3 4 0 4 2 0 4 2 4 3 5 0 3 4 5 1 3 7 4 2 2
4 3 5 3 6 2 3 3 2 9 4 4 2 5 2 2 4 2 2 3 1 4 3 3 2
5 6 3 2 2 3 3 3 2 2 4 2 4 8 2 2 5 2 4 3 6 2 3 1 5



$$s^2 = 3.113$$

$$s_n^2 = 3.082,$$

$$s = 1.764$$

$$s_n = \sqrt{s_n^2} = 1.755.$$

Využití absolutních četností při výpočtu \bar{x} a s^2

Mějme výběr x_1, \dots, x_n v němž se vyskytuje k vzájemně různých hodnot u_1, \dots, u_k jejichž absolutní četnosti jsou f_1, \dots, f_k . Pak platí

Výběrový průměr

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^k f_i u_i .$$

Příklad:
$$\frac{4 + 2 + 5 + 6 + 2 + 2 + 4}{7} = \frac{3 \cdot 2 + 2 \cdot 4 + 1 \cdot 5 + 1 \cdot 6}{7} .$$

Výběrový rozptyl

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

Využití intervalového rozdělení četností při výpočtu \bar{x} a s^2

Mějme výběr x_1, \dots, x_n jehož hodnoty jsou rozděleny do k intervalů $(c_0, c_1), (c_1, c_2), \dots, (c_{k-1}, c_k)$ s absolutními četnostmi f_1, \dots, f_k a středy intervalů u_1, \dots, u_k . Pak můžeme uvažovat následující odhady pro hodnoty \bar{x} a s^2 .

Odhad \bar{u} pro výběrový průměr \bar{x}

$$\bar{u} = \frac{1}{n} \sum_{i=1}^k f_i u_i \quad \approx \quad \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} .$$

Odhad u^2 pro výběrový rozptyl s^2

$$u^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (u_i - \bar{u})^2 \quad \approx \quad \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2 .$$

Příklad (Přesné hodnoty \bar{x} a s^2 a jejich odhady)

interval	meze	f_i	h_i	střed
(19.1, 20.3)	(19.39, 20.11)	3	0.063	19.7
(20.3, 21.5)	(20.57, 21.42)	5	0.104	20.9
(21.5, 22.7)	(21.55, 22.67)	12	0.250	22.1
(22.7, 23.9)	(22.72, 23.86)	14	0.292	23.3
(23.9, 25.1)	(23.95, 24.63)	5	0.104	24.5
(25.1, 26.3)	(25.92, 25.92)	1	0.021	25.7

$$\bar{x} = 22.6265$$

$$s^2 = 1.865$$

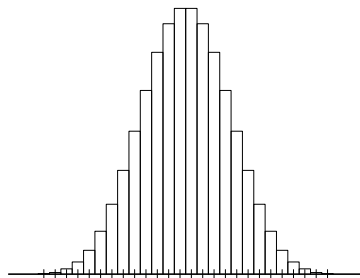
$$\bar{u} = \frac{1}{n} \sum_{i=1}^k f_i u_i = \frac{3 \cdot 19.7 + 5 \cdot 20.9 + \dots + 1 \cdot 25.7}{40} = \frac{903.2}{40} = 22.58.$$

$$u^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (u_i - \bar{u})^2 = 1.979. \quad \text{V tomto případě: } \bar{x} \neq \bar{u} \text{ a } s^2 \neq u^2.$$

Empirické pravidlo

Odhad intervalu hodnot

- aplikace výběrové směrodatné odchylky
- přibližné určení intervalů hodnot
- lze použít, když má histogram tvar „zvonu“
- dokážeme později (PŘEDNÁŠKA 9)

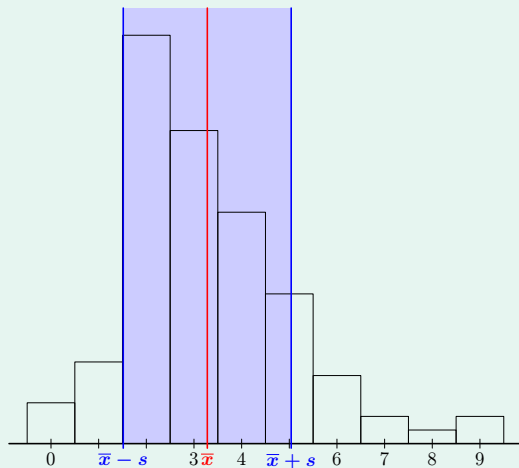


Empirické pravidlo

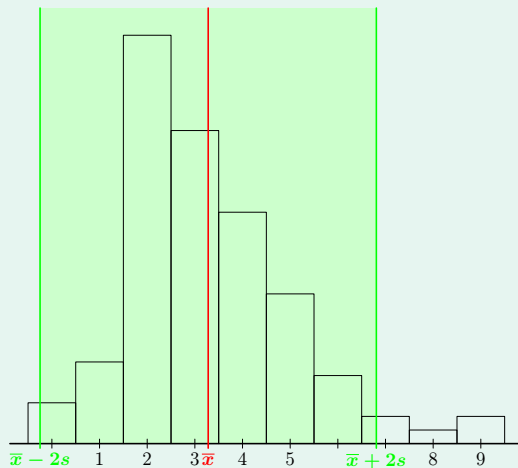
Uvažujme výběr x_1, \dots, x_n s výběrovým průměrem \bar{x} a výběrovou směrodatnou odchylkou s . Pokud má histogram tvar „zvonu“ pak

- přibližně 68 % dat z výběru se nachází v intervalu $(\bar{x} - s, \bar{x} + s)$,
- přibližně 95 % dat z výběru se nachází v intervalu $(\bar{x} - 2s, \bar{x} + 2s)$,
- přibližně 99.7 % dat z výběru se nachází v intervalu $(\bar{x} - 3s, \bar{x} + 3s)$.

Příklad (Počty dětí ve výběru 100 rodin)



přibližně 68 % dat



přibližně 95 % dat

Přednáška 1: Závěr

Pojmy k zapamatování:

- náhodný pokus, elementární jev, prostor elementárních jevů
- základní populace, výběr, relativní/absolutní četnost
- výběrový průměr (míra centrální tendence dat)
- výběrový rozptyl, směrodatná odchylka (míry rozptýlenosti dat)

Použité zdroje:



Devore J. L.: *Probability and Statistics for Engineering and the Sciences*
Duxbury Press, 7. vydání 2008, ISBN 978-0-495-55744-9.



Hendl, J.: *Přehled statistických metod zpracování dat*
Portál, Praha 2006, ISBN 978-80-7367-123-5



Hogg R. V., Tanis E. A.: *Probability and Statistical Inference*
Prentice Hall; 7. vydání 2005, ISBN 978-0-13-146413-1.