Popisné statistiky číselných dat.

Číselné charakteristiky

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

doc. Ing. Petr Seďa, Ph.D.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co se dnes dozvíte?

- Míry polohy, střední hodnoty, průměry.
- Míry variability, rozptyl a směrodatná odchylka.
- Čebyševova nerovnost.
- Normované hodnoty, pravidlo 6 sigma.
- Míry tvaru rozdělení – šikmost a špičatost.
- Kvantily, explorační analýza dat.

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co se dnes dozvíte?

- agregují informaci o statistickém znaku do několika málo hodnot,
- jsou stručnější a přehlednější než výchozí data,
- snaží se charakterizovat rozdělení hodnot znaku.

základní typy charakteristik:

míry polohy – umístění hodnot znaku (na číselné ose)
míry variability - rozptýlení hodnot kolem typické polohy
míry tvaru rozdělení – symetrie, koncentrace hodnot znaku

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

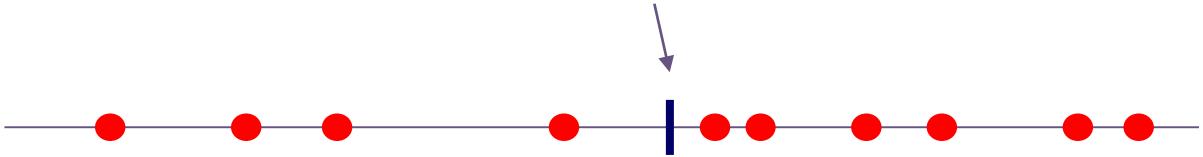
Míry polohy:

určují polohu (pomyslný střed) statistického znaku

střední hodnota – aritmetický průměr

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

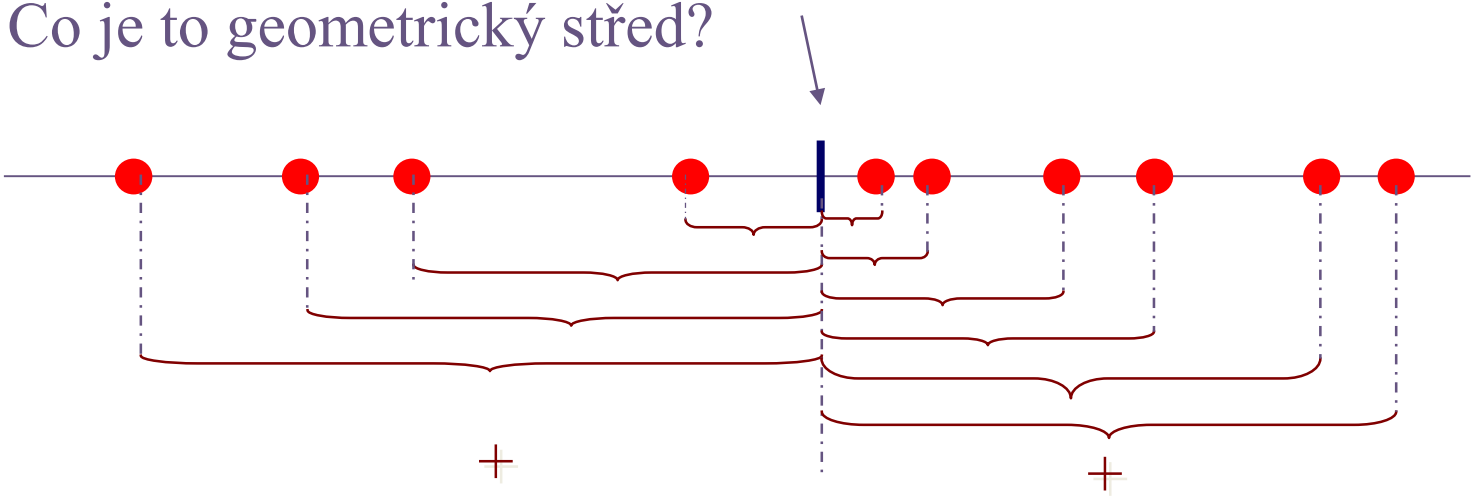
vyjadřuje geometrický střed (těžiště) statistického znaku na číselné ose



Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co je to geometrický střed?



součet vzdáleností od průměru hodnot **menších** než průměr

součet vzdáleností od průměru hodnot **větších** než průměr

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

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Nejednoznačná terminologie?

Místo pojmu **průměr** se v praxi často používá:

- průměrná hodnota
- střední hodnota
- prostřední hodnota
- charakteristická hodnota
- typická hodnota
- očekávaná hodnota

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vlastnosti aritmetického průměru:

Součet odchylek všech hodnot znaku od aritmetického průměru je roven nule.

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

Přičteme-li ke všem hodnotám znaku stejné číslo, zvětší se o toto číslo také aritmetický průměr.

$$\overline{x + a} = \overline{x} + a$$

Vynásobíme-li všechny hodnoty znaku stejným číslem, zvětší se stejným způsobem i aritmetický průměr.

$$\overline{ax} = a \cdot \overline{x}$$

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vážený aritmetický průměr:

střední hodnota pro tabulku rozdělení četností

$$\overline{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots + x_k \cdot n_k}{n} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$

- počet tříd (kategorií)
- četnosti jednotlivých tříd
- jednotlivé hodnoty znaku
- velikost souboru

u rozdělení relativních četností

$$\overline{x} = x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_k \cdot p_k = \sum_{i=1}^{k} x_i \cdot p_i$$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jiné typy průměrů:

Možná znáte kromě aritmetického průměru:

- harmonický průměr $\overline{x}_H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$

- geometrický průměr $\overline{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Kdy se který z těchto průměrů používá?

Harmonický průměr: Používá se, pokud potřebujeme hodnotu, která zastupuje ostatní, co se týče převrácených hodnot, například při výpočtu průměrné rychlosti na úsecích stejné délky. Dále jsou-li hodnoty znaku nerovnoměrně rozloženy kolem aritmetického průměru, nebo když jsou hodnoty extrémně nízké či vysoké.

Geometrický průměr: Používá se např. na koeficienty růstu pro výpočet průměrného tempa růstu.

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Další míry polohy – modus a medián:

modus $\widehat{x}$ – nejčetnější hodnota znaku v souboru (vyskytuje se v souboru nejčastěji)

- vhodné zejména pro nominální znaky
- nemusí být určen jednoznačně

medián $\tilde{x}$ – prostřední hodnota znaku v souboru uspořádaného podle velikosti znaku

- vhodné pro ordinální a nesymetrické znaky
- není důležitá hodnota, ale pořadí

u sudého počtu prvků souboru se medián počítá jako průměr ze dvou hodnot nejbližších středu

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jakou střední hodnotu použít ?

(aritmetický) průměr - u číselných znaků, které nevykazují extrémní hodnoty

medián - u číselných znaků s extrémy, u ordinálních nečíselných znaků

modus - u nominálních nečíselných znaků

Otázka k zamyšlení: Proč aritmetický průměr není vhodnou střední hodnotou pro znak „měsíční příjem zaměstnance“ ?

Příklad 1 – portfolio akcií

cena akcie	počet
200 Kč	3
300 Kč	5
500 Kč	2
1 000 Kč	1
1 500 Kč	1

modus (nejčetnější hodnota)
300 Kč

medián: $n = 12$

$$\tilde{x} = \frac{x_6 + x_7}{2} = \frac{300 + 300}{2} = 300$$

průměrná cena akcie:

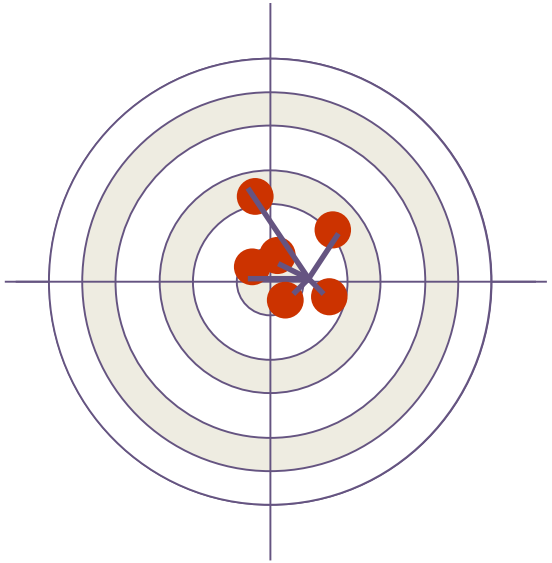
$$\overline{x} = \frac{200 \cdot 3 + 300 \cdot 5 + 500 \cdot 2 + 1000 \cdot 1 + 1500 \cdot 1}{3 + 5 + 2 + 1 + 1} = \frac{5600}{12} = 466,67$$

Popisné statistiky číselných dat

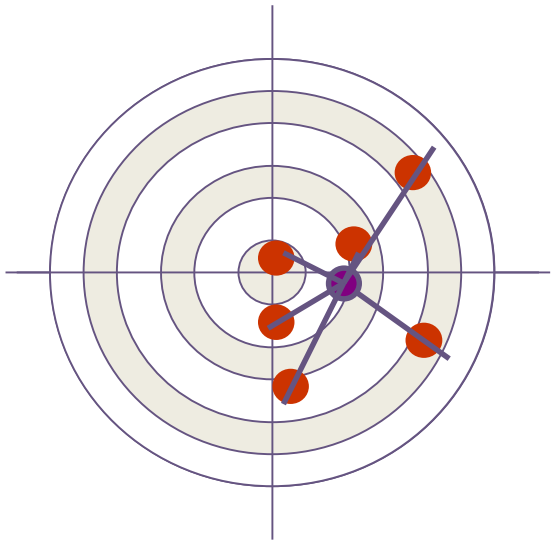
SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Variabilita znaku:

variabilita určuje, jak se hodnoty znaku liší od průměru



malý rozptyl



velký rozptyl

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Rozptyl – ukazatel variability:

rozptyl - variabilita znaku v souboru

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

průměrný čtverec odchylek od průměru

nezáleží na znaménku odchylky

vzorec vhodnější pro ruční výpočet:

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

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vlastnosti rozptylu:

Rozptyl konstanty (znaku, který nemění svou hodnotu) je roven nule.

$$s^2(a) = 0$$

Přičteme-li ke všem hodnotám statistického znaku stejné číslo, rozptyl se nezmění.

$$s^2(x + a) = s^2(x)$$

Vynásobíme-li všechny hodnoty statistického znaku stejným číslem (např. k-krát), zvětší se rozptyl znaku dvojmocí této hodnoty (tj. k^2-krát).

$$s^2(ax) = a^2 \cdot s^2(x)$$

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Další ukazatele variability:

směrodatná odchylka

$$s = \sqrt{s^2}$$

← průměrná odchylka od průměru (tzv. kvadratický průměr)

variační koeficient

$$V_x = \frac{s}{\overline{x}}$$

- použití pro znaky s nezápornými hodnotami
- srovnání znaků s různou velikostí hodnot
- obvykle se vyjadřuje v % (x 100)

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Jak chápat směrodatnou odchylku?

Čebyševova nerovnost:

V intervalu $\left(\overline{x} - k \cdot s \,;\, \overline{x} + k \cdot s\right)$ se nachází nejméně

$1 - \frac{1}{k^2}$ hodnot znaku. $(k > 1)$

pravidlo 6 sigma:

Všechny hodnoty znaku, které se nacházejí ve vzdálenosti větší než 3 směrodatné odchylky od průměru, se považují za **extrémní.**

Příklad 2 – portfolio akcií

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

rozptyl ceny akcie:

$$s^2 = \frac{200^2 \cdot 3 + 300^2 \cdot 5 + 500^2 \cdot 2 + 1000^2 \cdot 1 + 1500^2 \cdot 1}{12} - 466,67^2 = 142219$$

směrodatná odchylka a variační koeficient:

$$s = \sqrt{s^2} = \sqrt{142219} = 377$$

$$V_x = \frac{s}{\overline{x}} = \frac{377}{466,67} = 0,808 = 80,8\%$$

závěr: vysoká variabilita znamená, že střední hodnota (průměr) není dobrým reprezentantem znaku

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Normovaná hodnota z:

určuje vzdálenost hodnoty znaku od střední hodnoty (v násobcích směrodatné odchylky)

$z > 0$ hodnota je větší než průměr
$z < 0$ hodnota je menší než průměr

$$z_i = \frac{x_i - \overline{x}}{s}$$

pravidlo 6 sigma:
- hodnoty z větší než 3 (menší než -3) značí **extrémní hodnoty**

- někdy se normovaná hodnota označuje též jako u

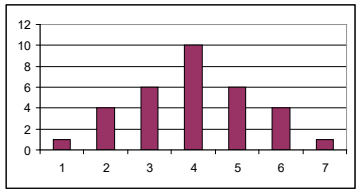
Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

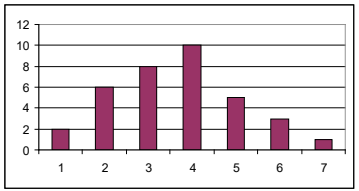
Míry tvaru rozdělení:

šikmost - vyjadřuje asymetrii rozložení hodnot znaku

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} z_i^3$$

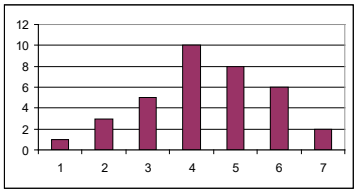


$\alpha = 0$



$\alpha > 0$

kladné sešikmení

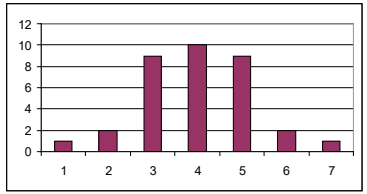


$\alpha < 0$

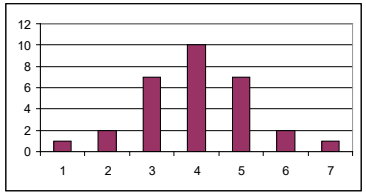
záporné sešikmení

špičatost - vyjadřuje koncentraci hodnot znaku

$$\beta = \frac{1}{n}\sum_{i=1}^{n} z_i^4 - 3$$

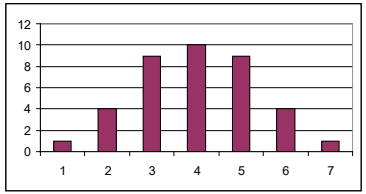


$\beta = 0$



$\beta > 0$

špičaté rozdělení



$\beta < 0$

ploché rozdělení

Příklad 3 – portfolio akcií

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

šikmost:

$$\alpha = \frac{(-0,68)^3 \cdot 3 + (-0,42)^3 \cdot 5 + 0,08^3 \cdot 2 + 1,35^3 \cdot 1 + 2,62^3 \cdot 1}{12} = 1,59$$

kladné sešikmení – *vyšší koncentrace menších hodnot*

špičatost:

$$\beta = \frac{(-0,68)^4 \cdot 3 + (-0,42)^4 \cdot 5 + 0,08^4 \cdot 2 + 1,35^4 \cdot 1 + 2,62^4 \cdot 1}{12} - 3 = 1,27$$

kladná špičatost – *vyšší koncentrace hodnot kolem průměru*

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Kvantily – specifické míry polohy:

kvantil $x_p\%$
odděluje $p\%$ nejnižších hodnot od zbytku souboru

medián						$x_{50\%} = x_{0,5}$						
kvartily				$x_{25\%}$			$x_{50\%}$		$x_{75\%}$			
decily			$x_{10\%}$		$x_{20\%}$			...			$x_{90\%}$	
Percentil	$x_{1\%}$	$x_{2\%}$				...						$x_{99\%}$

$z_p\%$ pořadí kvantilu v rámci uspořádaného znaku

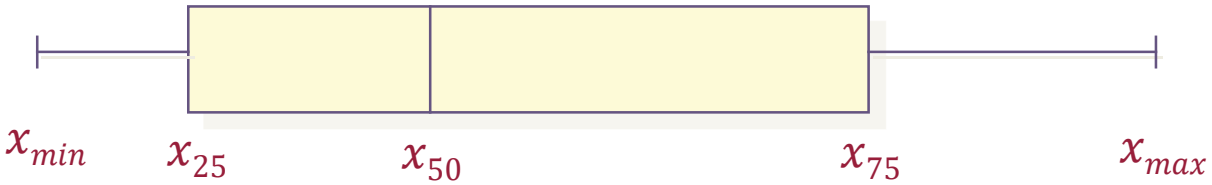
$$z_p = \frac{n \cdot p\%}{100} + 0,5$$

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Kvartily a box plot:

kvartily rozdělují uspořádaný soubor na 4 stejně početné části

box plot – graf kvartilů

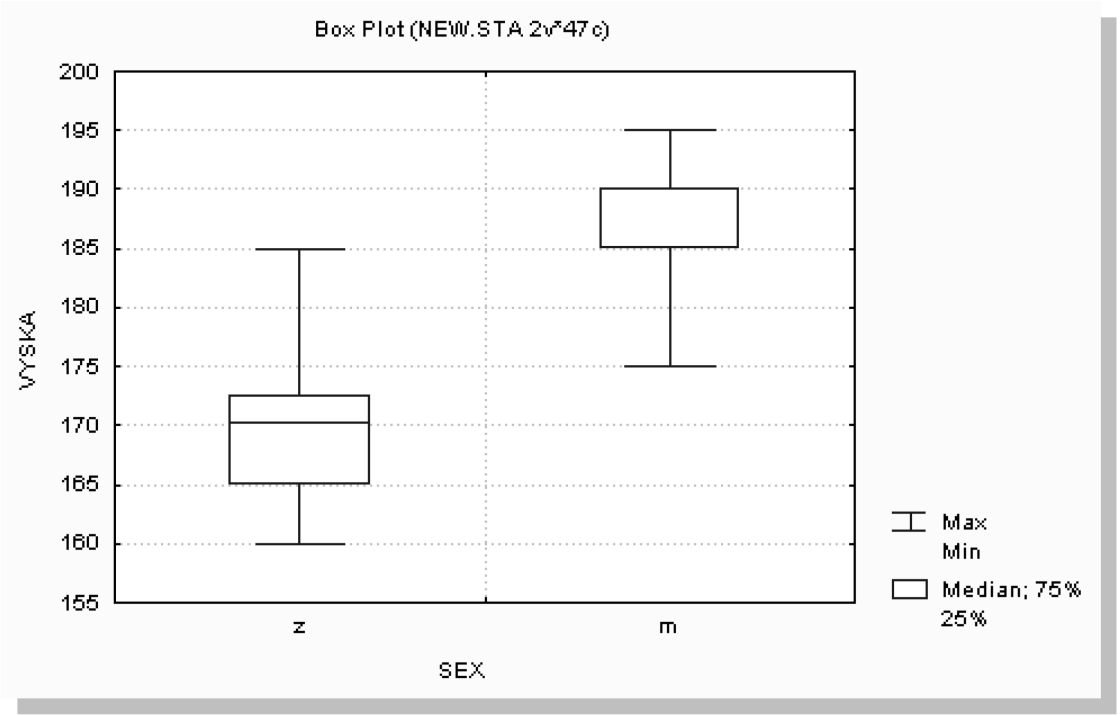


box plot slouží k porovnávání rozdělení různých znaků

Popisné statistiky číselných dat

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Ukázka použití box plotu v praxi:



SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Střední hodnota intervalového rozdělení:

střední hodnota – vážený aritmetický průměr

$$\overline{x} = \frac{\overline{x}_1 \cdot n_1 + \overline{x}_2 \cdot n_2 + \ldots + \overline{x}_k \cdot n_k}{n} = \frac{\sum_{i=1}^{k} \overline{x}_i \cdot n_i}{n}$$

neznáme-li průměry tříd, nahradíme je středy intervalů

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Rozptyl intervalového rozdělení:

známe-li rozptyly jednotlivých tříd:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot \left(\overline{x}_i - \overline{x}\right)^2 + \frac{1}{n}\sum_{i=1}^{k} n_i \cdot s_i^2$$

meziskupinový rozptyl

vnitroskupinový rozptyl

nahradíme-li průměr třídy jejím středem – Sheppardova korekce:

$$s^2 = \ldots - \frac{1}{12}h^2$$

kompenzuje nadhodnocení rozptylu při náhradě průměru středem

Příklad 4 – počet dětí

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Tabulka rozdělení četností popisuje rozdělení počtu dětí pracovníků jedné počítačové firmy.

DĚTI	**četnosti**		**kumulativní četnosti**	
	absolutní	relativní	absolutní	relativní
0	6	25,0%	6	25,0%
1	8	33,3%	14	58,3%
2	7	29,2%	21	87,5%
3	2	8,3%	23	95,8%
4	1	4,2%	24	100,0%
CELKEM	**24**	**100%**		

a) Vyjádřete variabilitu počtu dětí pomocí směrodatné odchylky a variačního koeficientu.

b) Porovnejte průměrný počet dětí, medián a modus. O čem to vypovídá?

c) Vypočítejte a vysvětlete šikmost a špičatost rozdělení počtu dětí.

d) Určete jednotlivé kvartily rozdělení počtu dětí.

Příklad 4 – počet dětí

ad a) Rozptyl a směrodatnou odchylku vypočteme jako vážené:

$$s_x^2 = \frac{6 \cdot 0^2 + 8 \cdot 1^2 + 7 \cdot 2^2 + 2 \cdot 3^2 + 1 \cdot 4^2}{24} - 1{,}33^2 = 1{,}139$$

$$s_x = \sqrt{1{,}139} = 1{,}07$$

Nyní můžeme spočítat variační koeficient:

$$V_x = \frac{s_x}{\bar{x}} = \frac{1{,}07}{1{,}33} = 0{,}805 = 80{,}5\%$$

Vysoká hodnota variačního koeficientu značí značnou rozptýlenost počtu dětí. Střední hodnota (průměr) $\bar{x} = 1{,}33$ tedy není výstižným ukazatelem polohy znaku na číselné ose.

Příklad 4 – počet dětí

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

ad b) Průměr = 1,33, medián = modus = 1. Medián leží mezi 12. a 13. prvkem, modus je nejčetnější obměna, lze očekávat rozdělení mírně sešikmené doprava → převažují rodiny s malým počtem dětí.

ad c) K výpočtu šikmosti a špičatosti potřebujeme normované hodnoty. Sestavíme si tabulku, ze které lze spočítat všechny základní ukazatele:

x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	z_i	$z_i^3 \cdot n_i$	$z_i^4 \cdot n_i$
0	6	0	0	-1,246	-11,607	14,462
1	8	8	8	-0,311	-0,241	0,075
2	7	14	28	0,623	1,693	1,055
3	2	6	18	1,558	7,564	11,784
4	1	4	16	2,493	15,494	38,627
∑	24	32	70	x	12,903	66,003
∅	**x**	**1,33**	**2,917**	**x**	**0,538**	**2,750**

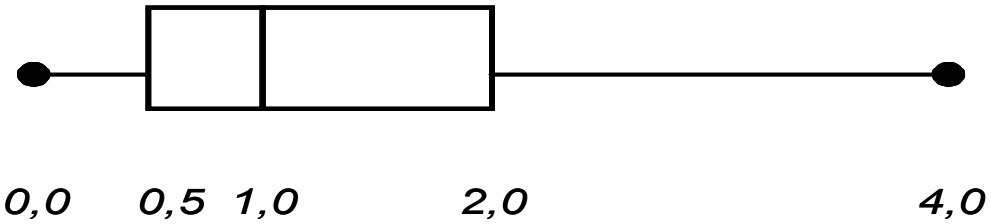
šikmost $\alpha = 0{,}538$ potvrzuje mírné sešikmení doprava
špičatost $\beta = 2{,}75 - 3 = -0{,}25$ mírně plošší rozdělení, ale ne příliš

Příklad 4 – počet dětí

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

ad d) Kvartily najdeme mezi 6. a 7. prvkem, 12. a 13. a 18. a 19. prvkem: $x_{0,25} = 0{,}5, x_{0,5} = 1, x_{0,75} = 2$. Menší vzdálenost mezi dolním kvartilem a mediánem opět signalizuje sešikmení doprava.

Box plot:



Literatura

1. Janáček J. *Statistika jednoduše*. Praha: Grada, 2022. **(kapitola 1).**

2. Ramík J. a Čemerková Š. *Statistika A.* Opava, Karviná: SLU, 2000. **(kapitola 2).**

Děkuji za pozornost.