Statistika pro geografy

2. Popisná statistika

Mgr. David Fiedor
23. února 2015

Osnova

Pojmy - Bodové rozdělení četností

Absolutní četnost

Absolutní četností hodnoty x_j znaku x rozumíme počet statistických jednotek daného statistického souboru, které mají stejnou hodnotu x_j znaku x a značíme n_j.

Relativní četnost

Relativní četností v_j hodnoty x_j znaku x rozumíme podíl absolutní četnosti hodnoty x_j a rozsahu celého statistického souboru, tj. $v_j = \frac{n_j}{n}$. Součet relativních četností všech různých hodnot daného znaku je roven jedné (resp. 100 %).

Pojmy - Bodové rozdělení četností

Absolutní kumulativní četnost

Absolutní kumulativní četností N_j rozumíme součet prvních j absolutních četností.

$$N_j = n_1 + \cdots + n_j$$

Relativní kumulativní četnost

Relativní kumulativní četností V_j rozumíme součet prvních j relativních četností.

$$V_j = \frac{N_j}{n} = v_1 + \cdots + v_j$$

Příklad

1,2,3,4,2,3,1,3,4,2,2,1,3,4,4,1,2,2,1,3,4,4,4,1,1

Řešení

Hodnota znaku	n_j	N_j	v_j	V_j
1	7	7	0,28	0,28
2	6	13	0,24	0,52
3	5	18	0,20	0,72
4	7	25	0,28	1,00

Pojmy - Skupinové (intervalové) rozdělení četností

- pro spojité znaky - udáváme počet prvků s hodnotami znaku patřících do daného intervalu (třídy)
- používáme při velkém počtu různých variant hodnot znaku
- *šířku intervalu* určuje rozdíl horní a dolní hranice (meze) a je konstantní pro všechny třídy
- střed intervalu je určen jako aritmetický průměr dolní a horní meze daného intervalu

Určení počtu tříd

Sturgesovo pravidlo

- k - počet tříd
- $k \doteq 1 + 3,3 \log n$

počet variant znaku	počet třídicích intervalů
1	1
2	2
3-5	3
6-11	4
12-23	5
24-46	6
47-93	7
94-187	8
188-376	9
377-756	10

Příklad

U obcí Moravskoslezského kraje s počtem obyvatel větším než tisíc a menších než deset tisíc byl zjištěn počet narozených dětí za rok 2008. Výsledky jsou následující:

$$28, 28, 23, 51, 21, 25, 9, 6, 30, 18, 16, 15,$$

$$65, 14, 9, 40, 16, 23, 12, 21, 10, 10, 40, 38, 10,$$

$$21, 31, 48, 19, 17, 16, 16, 11, 11, 27, 19, 20, 46.$$

Pomocí Sturgesova pravidla určete počet třídicích intervalů, vytvořte tabulku skupinového rozdělení četností a relativních četností.

Řešení

Nejdříve zjistíme počet všech hodnot zkoumaného znaku, abychom pomocí Sturgesova pravidla určili počet třídicích intervalů. Uspořádejme si všechny hodnoty do řady od nejmenších po největší:

$$6, 9, 9, 10, 10, 10, 11, 11, 12, 14, 15, 16, 16,$$

$$16, 17, 18, 19, 19, 20, 21, 21, 21, 23, 23, 25,$$

$$27, 28, 28, 30, 31, 38, 40, 40, 46, 48, 51, 65.$$

Rozsah souboru je 37, odkud jsme dosazením do vzorce Sturgesova pravidla dostali, že počet třídicích intervalů je roven šesti. Nejmenší, resp. největší hodnota statistického znaku tohoto souboru je rovna 6, resp. 65. Délka jednoho intervalu se proto bude rovnat deseti. Sestrojme nyní tabulku skupinového rozdělení četností a relativních četností.

intervaly znaku x	n_j	N_j	v_j	V_j
6-15	11	11	$\frac{11}{37}$	$\frac{11}{37}$
16-25	14	25	$\frac{14}{37}$	$\frac{25}{37}$
26-35	5	30	$\frac{5}{37}$	$\frac{30}{37}$
36-45	3	33	$\frac{3}{37}$	$\frac{33}{37}$
46-55	3	36	$\frac{3}{37}$	$\frac{36}{37}$
56-65	1	37	$\frac{1}{37}$	$\frac{37}{37}$

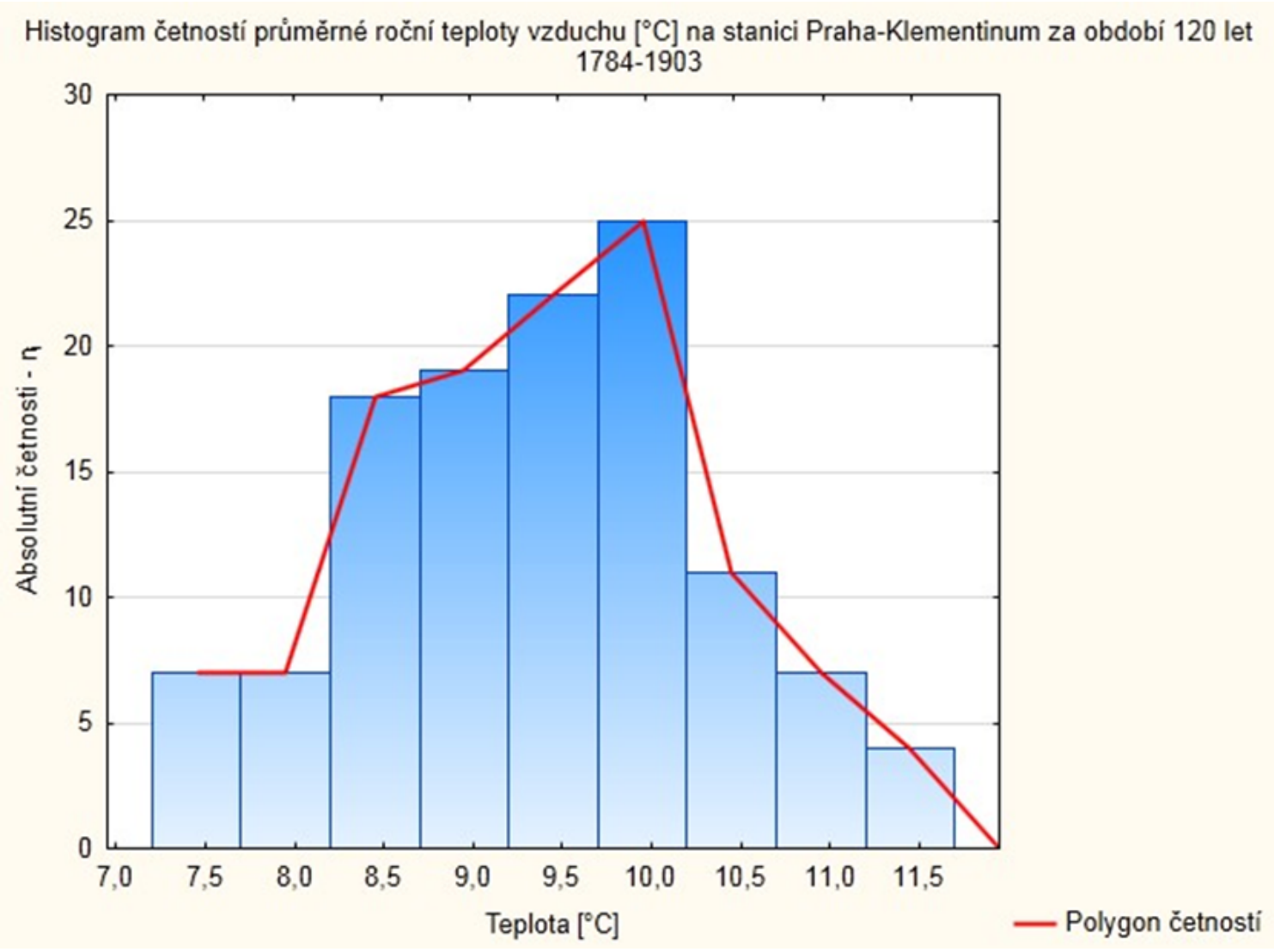
Polygon četností

- typ spojnicového grafu
- slouží k znázornění četností kvantitativních znaků
- používá se pro intervalové i bodové rozdělení četností
- propojuje všechny body v pravoúhlé soustavě, kde osa *x* vyjadřuje hodnotu znaku a osa *y* znázorňuje odpovídající četnost

Histogram

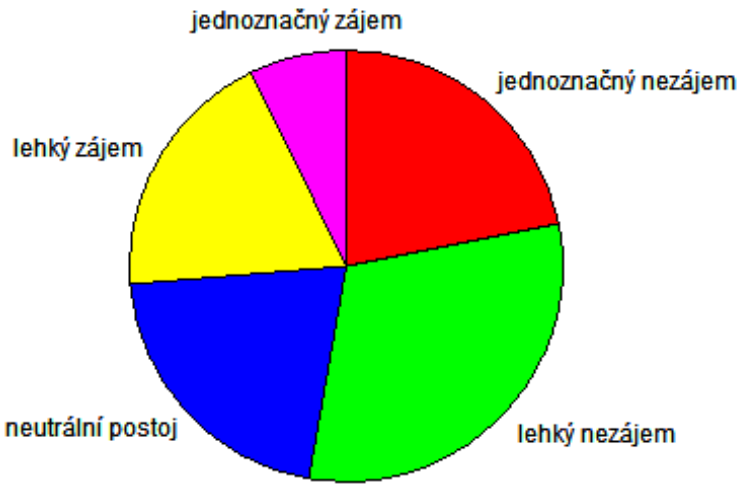
- typ sloupcového diagramu
- slouží k znázornění četností kvantitativních znaků
- převážně se používá pro intervalové rozdělení četností
- graf je tvořen pravidelnými rovnoběžníky, jejichž základny mají délku zvolených intervalů a jejichž výšky mají velikost příslušných intervalových četností

Vztah histogramu a polygonu četností



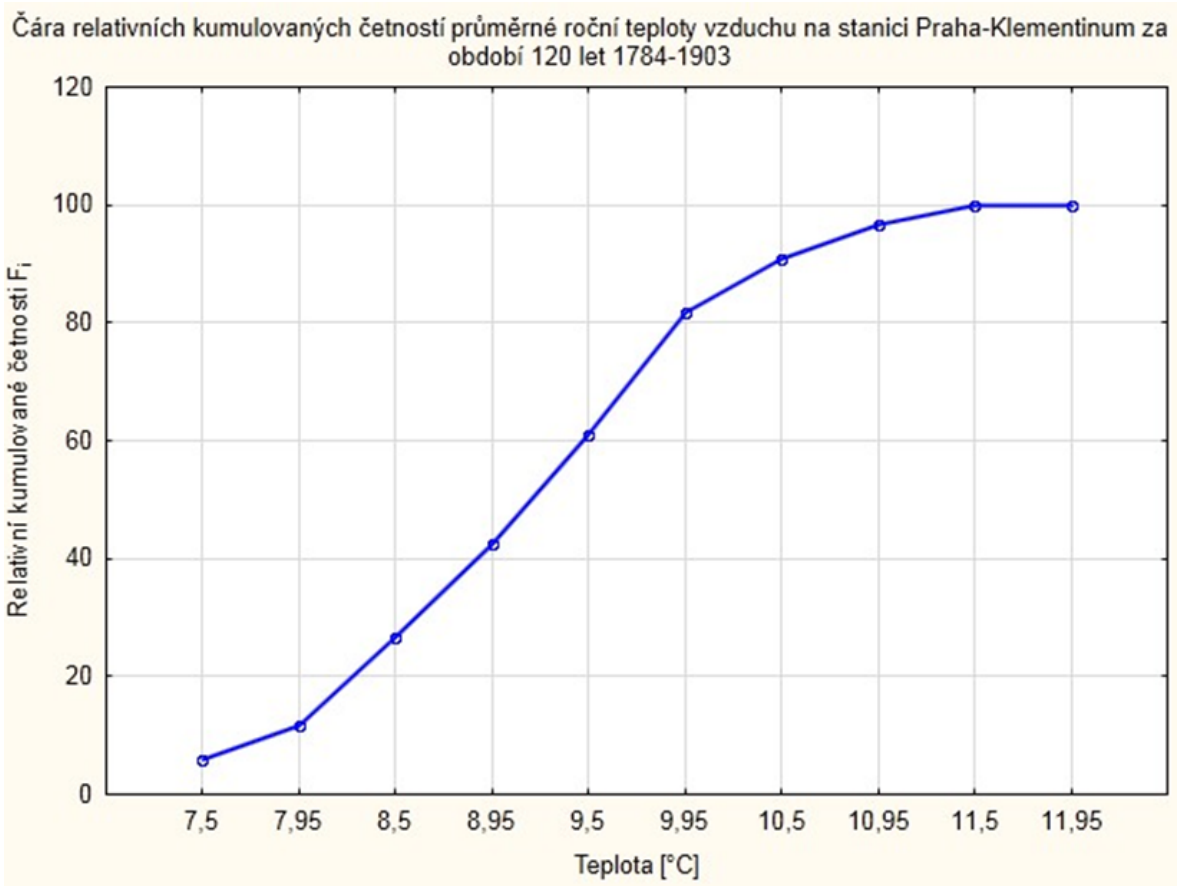
Kruhový diagram

- slouží k znázornění četností kvalitativních znaků
- obsahy kruhových výsečí znázorňující jednotlivé hodnoty statistického znaku jsou přímo úměrné relativním četnostem v procentech



Součtová čára

- grafické znázornění relativních kumulativních četností



Typy statistických znaků

Typy statistických znaků podle stupně kvantifikace:

- kvalitativní
  - nominální znaky
  - ordinální znaky
- kvantitativní
  - intervalové znaky
  - poměrové znaky

Základní statistické charakteristiky

1. Charakteristiky polohy (úrovně)
   - střední hodnoty, míry polohy, míry centrální tendence
   - charakterizují obecnou velikost hodnot statistického znaku
2. Charakteristiky variability
   - směrodatná odchylka, rozptyl, variační koeficient
   - popisují stupeň proměnlivosti hodnot daného znaku
3. Charakteristiky asymetrie
   - koeficient asymetrie (šikmosti)
   - umožňuje objektivně posoudit tvary histogramů - rozdělení četností hodnot daného znaku
4. Charakteristiky špičatosti
   - koeficient špičatosti
   - vyjadřuje koncentraci hodnot kolem určité hodnoty - rozdělení špičaté x ploché

Aritmetický průměr

Aritmetický průměr $\bar{x}$ hodnot $x_1, x_2, \ldots, x_n$ znaku x je definován jako podíl součtu hodnot znaku a jejich počtu (rozsahu souboru) n, tj. je určen vzorcem:

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

- vhodný pouze pro znaky intervalového a poměrového typu
- měl by být typickou hodnotou daného znaku - ostatní hodnoty by se neměly příliš lišit a měl by se blížit také nejčetnější hodnotě

Vlastnosti aritmetického průměru

a) Součet všech rozdílů $x_i - x$ jednotlivých hodnot znaku x_i a jejich aritmetického průměru x se rovná nule:

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

b) Přičteme-li ke všem hodnotám znaku konstantu, aritmetický průměr se také zvětší právě o danou konstantu.

c) Vynásobíme-li všechny hodnoty znaku konstantou k, aritmetický průměr se k-krát zvětší.

d) Průměr součtu dvou proměnných se rovná součtu obou průměrů.

e) Aritmetický průměr si lze geometricky představit jako těžiště.

Příklad

Na meteorologické stanici Brno-Tuřany (241 m n.m.) byly za rok 2008 naměřeny a stanoveny průměrné měsíční teploty vzduchu (°C). Bez ohledu na počet dní v jednotlivých měsících stanovte z těchto teplot průměrnou roční teplotu vzduchu (°C).

Měsíc	1.	2.	3.	4.	5.	6.
Teploty	1,7	3,1	4,6	10,1	15,5	19,9

Měsíc	7.	8.	9.	10.	11.	12.
Teploty	20,3	19,9	14,4	9,9	6,5	2,1

Řešení

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1,7 + 3,1 + \cdots + 2,1}{12} = 10,7$$

Vážený aritmetický průměr

- každé hodnotě zkoumaného znaku přiřazujeme „váhu“, tedy důležitost
- jako váhu lze vnímat i absolutní četnosti hodnot datového souboru
- $$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i,$$

  přičemž $n_1, n_2, \ldots, n_k$ značí váhy příslušných hodnot statistických znaků $x_1, x_2, \ldots, x_k$ a platí:
  $n_1 + n_2 + \cdots + n_k = n$.

Příklad

Vypočtěte průměrnou denní teplotu vzduchu, jestliže znáte teploty: $t_7 = 5\,^\circ\mathrm{C}$, $t_{14} = 15\,^\circ\mathrm{C}$, $t_{21} = 8\,^\circ\mathrm{C}$.

Řešení

$$\bar{t} = \frac{t_7 + t_{14} + 2 \cdot t_{21}}{4} = \frac{5 + 15 + 2 \cdot 8}{4} = 9$$

Průměrná denní teplota je $9\,^\circ\mathrm{C}$.

Harmonický průměr

Harmonickým průměrem x_H hodnot znaku $x_1, x_2, \ldots, x_n$ rozumíme podíl rozsahu souboru a součtu převrácených hodnot znaku, tj. platí:

$$\bar{x_H} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = n : \sum_{i=1}^{n} \frac{1}{x_i}.$$

- vhodný pouze pro znaky intervalového a poměrového typu
- používá se pro charakterizování průměrné rychlosti změny - k popisu intenzitních ukazatelů

Příklad

Vzhledem k rozdílné dopravní propustnosti se na jednotlivých stejně dlouhých úsecích cesty do centra výrazně mění průměrná rychlost vozidla. Tyto úseky jsme při pokusu postupně zvládli překonat za 20 minut, 30 minut a poslední za pouhých 6 minut. Vypočítejte průměrný čas nutný k překonání jednoho úseku.

Řešení

$$\bar{x_H} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{3}{\frac{1}{20} + \frac{1}{30} + \frac{1}{6}} = 12$$

Geometrický průměr

Geometrickým průměrem x_G hodnot zkoumaného znaku $x_1, x_2, \ldots, x_n$ rozumíme $n - tou$ odmocninu ze součinu hodnot $x_1, x_2, \ldots, x_n$, proto:

$$\bar{x_G} = \sqrt[n]{x_1 x_2 \ldots x_n}.$$

- vhodný pouze pro znaky poměrového typu
- slouží zpravidla pouze k určení průměrného tempa růstu za jedno období (v časových řadách)

Příklad

Průměrné koeficienty růstu produkce určitého podniku za období posledních pěti let byly postupně: 4 %; 3,5 %; 7 %; 5 %, 2,7 %. Určete průměrný koeficient růstu produkce za dané období.

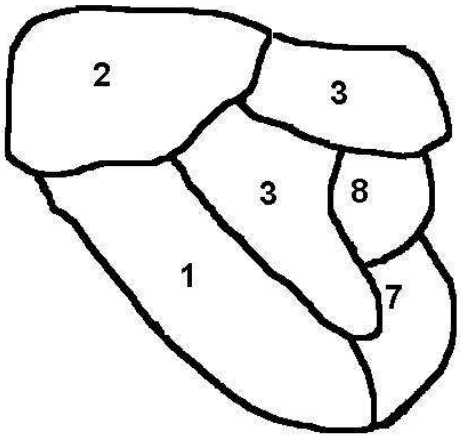
Řešení

$$\bar{x_G} = \sqrt[n]{x_1 x_2 \dots x_n} = \sqrt[5]{4 \cdot 3,5 \cdot 7 \cdot 5 \cdot 2,7} = \sqrt[5]{1323} = 4,21$$

Modus

Modus znaku x je jeho hodnota, která má největší četnost. Modus značíme symbolem $Mod(x)$.

- vhodný pro znaky jakéhokoliv typu - tedy nominálního, ordinálního, intervalového i poměrového typu
- slouží k určení dominantní třídy

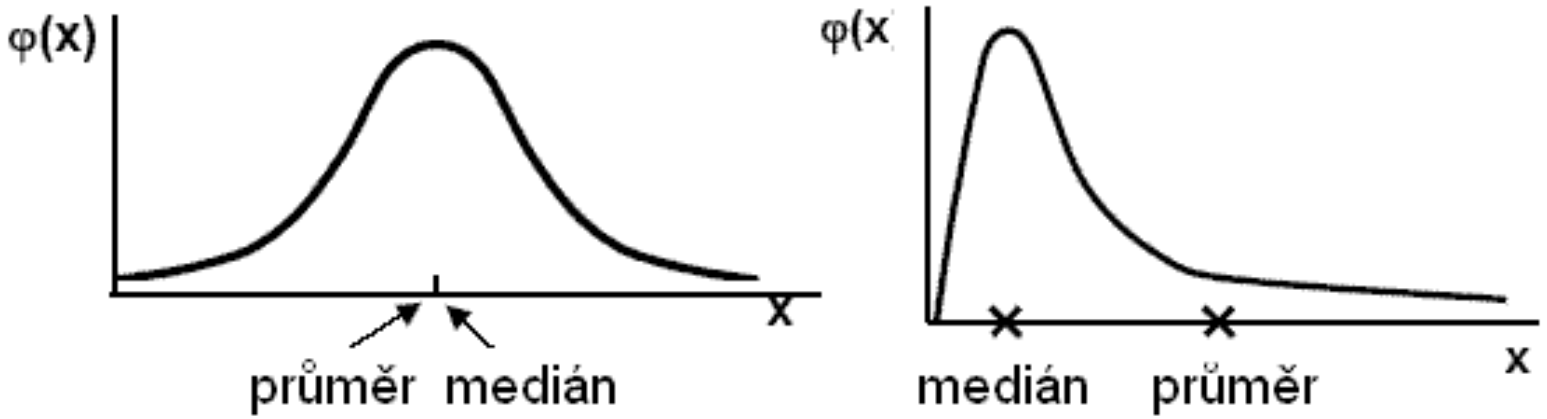


Medián

Jsou-li hodnoty $x_1, x_2, \ldots, x_n$ uspořádány podle velikosti ($x_1 \leq x_2 \leq \cdots \leq x_n$), pak *mediánem* znaku x rozumíme hodnotu znaku x, pro kterou platí:

$$Med(x) = \begin{cases} x_{\frac{n+1}{2}} & \text{je-li } n \text{ liché,} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{je-li } n \text{ sudé.} \end{cases}$$

- vhodný pro znaky takové, které lze uspořádat do pořadí - ordinální, intervalové a poměrové
- používá se v situacích, kdy je nevhodné použít aritmetický průměr


φ(x)
x
průměr
medián
φ(x
x
medián
průměr

Kvantily

Kvantilem rozumíme hodnotu statistického znaku x_ϑ, která rozděluje uspořádaná data na dva úseky – dolní a horní, přičemž dolní úsek obsahuje alespoň podíl ϑ všech dat a horní úsek alespoň podíl $1-\vartheta$ všech dat:

$$\underbrace{x_1 \leq x_2 \leq \ldots \leq x_c \leq x_\vartheta}_{\geq \vartheta} \overbrace{\leq x_{c+1} \leq \ldots \leq x_n}^{\geq 1-\vartheta}$$

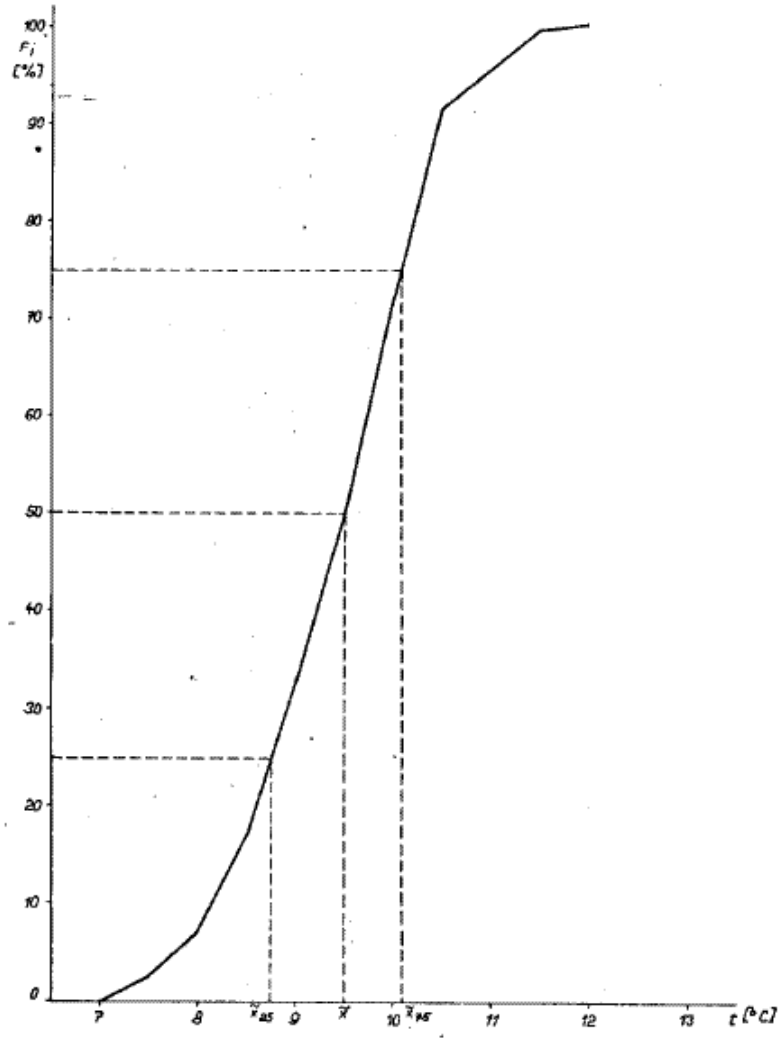
- vhodný pro znaky takové, které lze uspořádat do pořadí - ordinální, intervalové a poměrové
- vhodné především pro znaky mající velký počet možných variant

Kvantily

- medián je speciální případ kvantilu ($x_{0,50} = Med(x)$)
- $x_{0,25}$ - *dolní kvartil*
  $x_{0,75}$ - *horní kvartil*
  $x_{0,01}, x_{0,02}, \ldots, x_{0,98}, x_{0,99}$ - *percentily*

Korektní určení kvantilů

$$x_\vartheta = \begin{cases} \frac{(x_{(c)} + x_{(c+1)})}{2} & \text{je-li součin } n\vartheta \text{ celé číslo,} \\ x_{(c)} & \text{je-li součin } n\vartheta \text{ necelé číslo,} \\ & \text{zaokrouhlujeme nahoru na nejbližší} \\ & \text{celé číslo } c \end{cases}$$


100
F_i [%]
90
80
70
60
50
40
30
20
10
0
7
8
9
10
11
12
13
t [°C]

Shrnutí charakteristik polohy

Aritmetický průměr

- aspoň data intervalového typu
- symetrické rozdělení (symetrický tvar histogramu)
- užití ve statistických testech

Medián

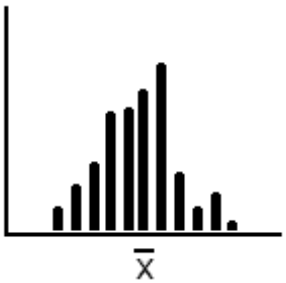
- aspoň data ordinálního typu
- chceme-li znát střed - vhodný i pro nesymetrická rozdělení (zešikmená)
- mohou obsahovat odlehlé hodnoty

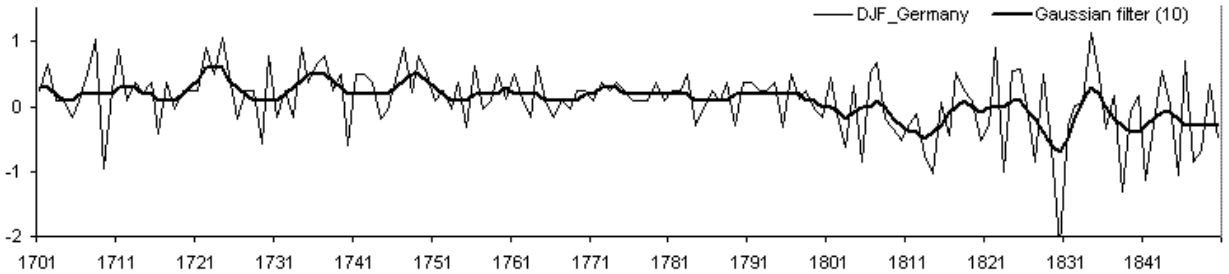
Modus

- data jakéhokoliv typu
- vícevrcholové rozdělení

Charakteristiky variability

- popis stupně proměnlivosti znaku
- vypovídají i o vhodnosti použití charakteristiky polohy





Variační rozpětí

- $R = x_{max} - x_{min}$
- není to ukazatel založený na všech hodnotách, proto nebere v úvahu rozdělení hodnot zkoumaného znaku

Rozptyl

Rozptylem s^2 hodnot znaku x rozumíme aritmetický průměr druhých mocnin odchylek hodnot znaku od aritmetického průměru, tj.:

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

- vhodný pro znaky intervalové a poměrové
- měří velikost proměnlivosti v jednotkách čtverců odchylek

Směrodatná odchylka

Směrodatná odchylka s_x je definována jako druhá odmocnina z rozptylu, tj.:

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

- vhodná pro znaky intervalové a poměrové
- absolutní míra variability - vyjádřena v jednotkách původních dat

Vlastnosti rozptylu a směrodatné odchylky

Již ze vzorců pro výpočet obou charakteristik plyne následující:

- přidáním konstanty k jednotlivým znakům souboru se ani jedna z těchto charakteristik nezmění
- vynásobením jednotlivých znaků konstantou se jejich směrodatná odchylka i rozptyl změní a to tak, že směrodatná odchylka je násobkem původní hodnoty a rozptyl je vynásoben druhou mocninou této konstanty

Variační koeficient

Variační koeficient v_x je definován jako podíl směrodatné odchylky a aritmetického průměru sledovaného znaku x, přičemž je často udáván v procentech:

$$v_x = \frac{s_x}{\bar{x}} \cdot 100\,\%$$

- vhodný pouze pro poměrová data, přičemž hodnoty by neměly být záporné
- nejpoužívanější relativní míra variability souboru
- v praxi slouží k porovnání variability více souborů

Příklad

Charakteristiky naměřené na dvou objektech (viz tabulka) mají stejnou směrodatnou odchylku, avšak výrazně se liší jejich aritmetické průměry a také variační koeficienty:

Charakteristiky	Objekt 1	Objekt 2
X1	6	56
X2	8	58
X3	10	60
X4	12	62
X5	16	66
X6	18	68
Aritmetický průměr	11,67	61,67
Směrodatná odchylka	4,23	4,23
Variační koeficient	39,5	7,5

Mezikvartilová odchylka

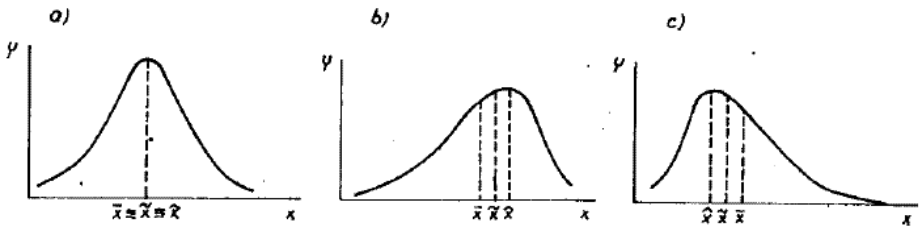
Mezikvartilovou odchylkou znaku x rozumíme hodnotu q, pro kterou platí:

$$q = \frac{(x_{0,75} - x_{0,25})}{2}.$$

- vhodný pro intervalová a poměrová data (na rozdíl od kvantilů musí mít význam i rozdíl)

Koeficient šikmosti

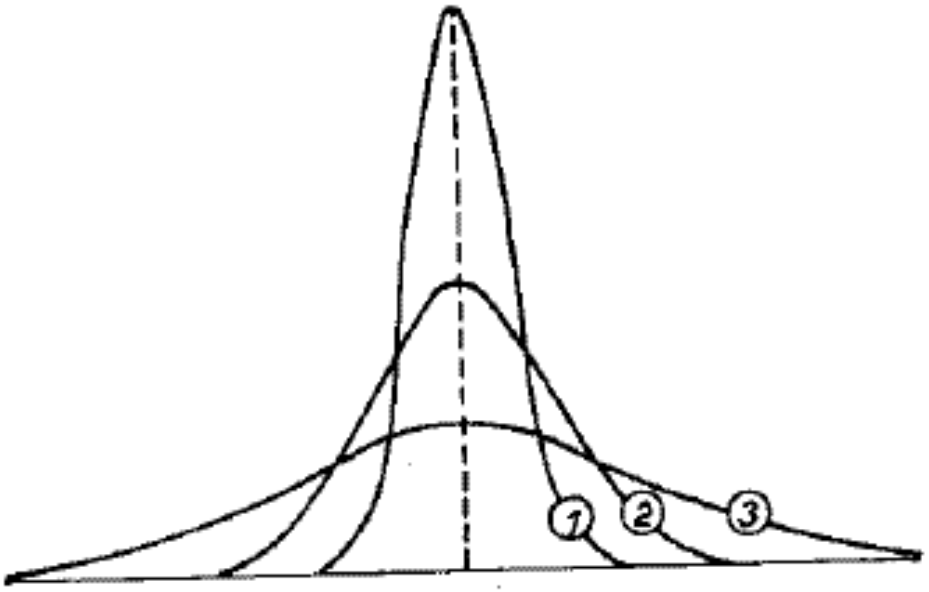
- $\alpha_3 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$
- je-li rozdělení dat symetrické, pak $\alpha_3 = 0$
- má-li prodloužený pravý konec, mluvíme o kladně zešikmeném rozdělení
- má-li prodloužený levý konec, mluvíme o záporně zešikmeném rozdělení



Koeficient špičatosti

- $\epsilon = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$
- popisuje koncentraci prvků souboru kolem určité hodnoty znaku - průměru
- špičatost (plochost) rozdělení je tím větší, čím více se hodnota ϵ odlišuje od nuly

- kladně zaspičatělé (špičaté) pro $\epsilon > 0$
- normálně zaspičatělé pro $\epsilon = 0$
- záporně zaspičatělé (ploché) pro $\epsilon < 0$



Děkuji za pozornost...