

Statistika

Zpracování informací ze statistického šetření

– Charakteristiky úrovně, variability a koncentrace kvantitativního znaku

Roman Biskup

(zapálený) statistik ve výslužbě, aktuálně analytik v praxi ;-)
roman.biskup(at)email.cz

20. února 2012



Obsah

Úvod do problému

Míry polohy (úrovně)

- Míry centrální tendence

- Ostatní míry polohy

Míry variability

- Rozptyl

- Směrodatná odchylka a variační koeficient

- Rozptyl pro ordinální data

- Rozpětí

Míry šikmosti a špičatosti

- Míry šikmosti

- Míry špičatosti

Grafická vizualizace

- Krabicový graf



Popisné (deskriptivní) charakteristiky/statistiky

- ▶ jsou číselné charakteristiky, které „koncentrovanou“ formou (jediným číslem) vyjadřují určitou vlastnost statistického znaku,
- ▶ obvykle slouží pro popis kvantitativního (kardinálního) statistického znaku, ale některé je možno použít i pro „jednodušší“ statistické znaky (např.: modus)
- ▶ někdy je problém s jejich interpretací i u diskrétního kardinálního znaku ($\bar{x}$ je 1,87 dítěte,)
- ▶ v některých případech mají jiný tvar (předpis) pro populaci a výběr (viz rozptyl)
- ▶ odlehlé hodnoty (pozorování)
 - ▶ pozorování jehož hodnota znaku vybočuje (reálná/nereálná)
 - ▶ bývají vlivnými hodnotami
- ▶ vlivné hodnoty (pozorování)
 - ▶ ovlivňují výsledek statistické analýzy (výpočet charakteristik, odhadů parametrů, ...)



Značení

n, N rozsah souboru – počet pozorování,

$x_{(i)}$ pořadové statistiky,

- ▶ seřazení hodnot podle velikosti,
- ▶ $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$,

n_i absolutní četnost, pro $i = 1, \dots, k$; $\sum_{i=1}^k n_i = N$,

‡ odděluje pojmenování respektive vzorce pro netříděná a **tříděná** data; navíc vše pro tříděná data je barevně **odlišeno**,

w_i pokud se ve tříděných datech formálně nahradí n_i symbolem w_i a k symbolem n , pak se obvykle mluví o vahách w_i jednotlivých hodnot x_i pro $i = 1, \dots, n$ a platí čím vyšší hodnota w_i , tím větší vliv hodnoty x_i pro výsledný výsledek, pochopitelně $N = \sum_{i=1}^n w_i$ – proto vážené charakteristiky,

x_i hodnota statistického znaku ($i = 1, \dots, n$) ‡ **hodnota statistického znaku při prostém třídění respektive středu intervalu pro třídění intervalové ($i = 1, \dots, k$),**



Míry polohy

- ▶ Mají být typickou hodnotou statistického znaku z daného statistického souboru,
- ▶ jsou jednoznačně definované a relativně jednoduše zjistitelné,
- ▶ slouží k porovnání úrovně různých statistických souborů, nebo vývoje statistického souboru v čase,
- ▶ mají co nejméně podléhat nahodilostem výběru respektive odlehlým hodnotám (pozorováním) – požadavek **robustnosti**.



Průměry I

$\bar{x}$ aritmetický průměr ‡ vážený aritmetický průměr

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \ddagger \quad \bar{x} = \frac{1}{N} \sum_{i=1}^k x_i n_i \quad \left(\bar{x} = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n x_i w_i \right)$$

- ▶ **nerobustní** míra ovlivněná odlehlými hodnotami,
- ▶ průměr může představovat rovnoměrnost nebo normu, která vůbec neexistuje a nemá odraz ve skutečnosti,
- ▶ jistě platí:
 - ▶ $n\bar{x} = \sum_{i=1}^n x_i$,
 - ▶ $\sum_{i=1}^n (x_i - \bar{x}) = 0$,
 - ▶ $\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2$, pro libovolné a .
- ▶ α -useknutý průměr
 - ▶ průměr vypočtený klasickým způsobem bez $\alpha/2$ % největších a nejmenších hodnot (robustnější než „obyčejný“ průměr) – méně nerobustní



Průměry II

$\bar{x}_G$ geometrický průměr ‡ vážený geometrický průměr

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \dots x_n} \quad \ddagger \quad \bar{x}_G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}},$$

kde $x_i \geq 0$, $i = 1, \dots, n$ ‡ $i = 1, \dots, k$

- ▶ **nerobustní** míra ovlivněná odlehlými hodnotami,
- ▶ ve výpočtech časových řad a některých indexů (inflace, apod.).

$\bar{x}_H$ harmonický průměr ‡ vážený harmonický průměr

$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad \ddagger \quad \bar{x}_H = \frac{N}{\sum_{i=1}^k \frac{n_i}{x_i}},$$

kde $x_i \neq 0$, $i = 1, \dots, n$ ‡ $i = 1, \dots, k$

- ▶ **nerobustní** míra ovlivněná odlehlými hodnotami,
- ▶ v časových výpočtech (frekvence, ...)



Modus a medián I

$\hat{x}$ modus

- ▶ nejčetnější hodnota znaku (unimodální, bimodální, vícenásobný modus)
- ▶ modální interval – interval s největší četností (relativní či absolutní)
- ▶ odhad módu na základě intervalového třídění
 - ▶ $\hat{x} \approx x_0 + \frac{h}{2} \frac{n_1 - n_{-1}}{2n_0 - n_1 - n_{-1}}$,
 - ▶ kde x_0 – je střed modálního intervalu, n_0 – je četnost modálního intervalu, n_{-1} – je četnost intervalu, který předchází modálnímu intervalu, n_1 – je četnost intervalu, který následuje za modálním intervalem a h – je délka modálního intervalu

$\tilde{x}_{0,5}$ medián ($\tilde{x}$)

- ▶ hodnota znaku, jež dělí soubor na dvě poloviny, na ty pozorování s nižšími hodnotami znaku a ty s vyššími hodnotami znaku
 - ▶
$$\begin{aligned} \tilde{x}_{0,5} &= \frac{x_{(n/2)} + x_{(n/2+1)}}{2} && \text{pro } n \text{ sudé} \\ &= x_{((n+1)/2)} && \text{pro } n \text{ liché} \end{aligned}$$
- ▶ mediánový interval – interval obsahující medián, tj. první interval, pro který platí: $k_{p_i} \geq 0,5$
- ▶ odhad mediánu na základě intervalového třídění



Modus a medián II

- ▶ $\tilde{x} \approx x_0 + \frac{\frac{n+1}{2} - \sum_{i=1}^{j-1} n_i}{n_j} h,$
- ▶ kde x_0 – je střed mediánového intervalu, $\sum_{i=1}^{j-1} n_i$ – je kumulativní četnost intervalu, který předchází mediánovému intervalu, n_j – je četnost intervalu mediánového intervalu a h – je délka mediánového intervalu



Ostatní míry polohy I

x_{min} minimum

$$\blacktriangleright x_{min} = \min x = x_{(1)} = \tilde{x}_0$$

x_{max} maximum

$$\blacktriangleright x_{max} = \max x = x_{(n)} = \tilde{x}_1$$

$\tilde{x}_{0,25}$ dolní kvartil

- ▶ hodnota znaku, jež dělí soubor (pozorování) na dvě části – čtvrtinu a tři čtvrtiny; na čtvrtinu pozorování s nižšími hodnotami znaku a tři čtvrtiny pozorování s vyššími hodnotami znaku

$\tilde{x}_{0,75}$ horní kvartil

- ▶ hodnota znaku, jež dělí soubor na dvě části – tři čtvrtiny a čtvrtinu; na čtvrtinu pozorování s nižšími hodnotami znaku a čtvrtinu pozorování s vyššími hodnotami znaku

$\tilde{x}_p$ $p \cdot 100\%$ (výběrový) kvantil $p \in \langle 0; 1 \rangle$

- ▶ hodnota znaku, jež dělí soubor na dvě části – p -tinu a $(1 - p)$ -tinu; p -tinu pozorování s nižšími hodnotami znaku a $(1 - p)$ -tinu s vyššími hodnotami znaku
- ▶ $\tilde{x}_p = (1 - P) \cdot x_{(S)} + P \cdot x_{(H)},$



Ostatní míry polohy II

- ▶ kde $S = \lfloor p(n-1) + 1 \rfloor$, $H = \lceil p(n-1) + 1 \rceil$ a $P = p(n-1) - S + 1$
- ▶ decil $p = 0; 0,1; 0,2; \dots; 0,9; 1$
- ▶ percentil $p = 0; 0,01; 0,02; \dots; 0,99; 1$



Míry variability

- ▶ Vypovídají o variabilitě/proměnlivosti hodnot statistického znaku z daného statistického souboru,
- ▶ jsou jednoznačně definované a relativně jednoduše zjistitelné,
- ▶ slouží k porovnání variability různých statistických souborů, nebo vývoje statistického souboru v čase,
- ▶ mají co nejméně podléhat nahodilostem výběru respektive odlehlým hodnotám – požadavek **robustnosti**
- ▶ některé vycházejí v odlišných jednotkách než posuzovaný statistický znak (rozptyly), nebo jsou relativní mírou variability (variační koeficient).



Rozptyl I

s_p^2 (populační) rozptyl ‡ vážený (populační) rozptyl

$$s_p^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \ddagger \quad s_p^2 = \frac{1}{N} \sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i$$

s_v^2 výběrový rozptyl ‡ vážený výběrový rozptyl

$$s_v^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \ddagger \quad s_v^2 = \frac{1}{N-1} \sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i$$

-
- ▶ **nerobustní** míry ovlivněné odlehlými hodnotami,
 - ▶ rozptyl vychází v jednotkách na druhou!,
 - ▶ $s_p^2 = \overline{x^2} - \bar{x}^2$ – výpočetní vzorec populačního rozptylu
 - ▶ $s_p^2 < s_v^2$ – „daň“ za výběrové šetření + požadavek nestrannosti odhadu,
 - ▶ pro velká n respektive **N** není znatelný numerický rozdíl mezi populačním a výběrovým rozptylem,



Rozptyl II

- transformace (Besselova oprava)

$$s_p^2 = \frac{n-1}{n} s_v^2 \quad \dagger \quad s_p^2 = \frac{N-1}{N} s_v^2$$

- obecná míra variability mezi všemi hodnotami (nejen vůči průměru)

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



Směrodatná odchylka

s_p (populační) směrodatná odchylka

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

s_v výběrová směrodatná odchylka

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

-
- ▶ **nerobustní** míry ovlivněné odlehlými hodnotami,
 - ▶ směrodatné odchylky vychází v jednotkách analyzovaného statistického znaku,
 - ▶ $s_p < s_v$ – „daň“ za výběrové šetření + požadavek nestrannosti odhadu,
 - ▶ pro velká n respektive N není znatelný numerický rozdíl mezi populační a výběrovou směrodatnou odchylkou,
 - ▶ transformace

$$s_p = \sqrt{\frac{n-1}{n}} s_v \quad \dagger \quad s_p = \sqrt{\frac{N-1}{N}} s_v$$



Variační koeficient

V_X variační koeficient (populační a výběrový)

$$V_X = \frac{s_p}{\bar{x}} \quad (V_X \cdot 100 \%), \quad V_X = \frac{s_v}{\bar{x}} \quad (V_X \cdot 100 \%)$$

-
- ▶ **nerobustní** míry ovlivněné odlehlými hodnotami,
 - ▶ V_X je bezrozměrná charakteristika respektive procentuálně vyjádřená;
 - ▶ Interpretace procentuálního vyjádření: „ V_X udává z kolika procent se podílí směrodatná odchylka na aritmetickém průměru“,
 - ▶ Variační koeficienty jsou relativní míry variability („indexy“), což umožňuje porovnávat variabilitu statistických znaků:
 - ▶ s odlišnými jednotkami,
 - ▶ mající sice stejné jednotky, ale odlišnou míru polohy.



Rozptyl pro ordinální data

- ▶ *dorvar* Rozptyl pro ordinální data

$$dorvar = \frac{4}{k-1} \sum_{i=1}^k k_{p_i}(1 - k_{p_i}),$$

- ▶ kde k je počet uspořádatelných kategorií a k_{p_i} , pro $i = 1, \dots, k$ jsou kumulativní relativní četnosti.
- ▶ *Dorvar* je variantou rozptylu (míry variability) pro ordinální data.



Rozpětí

R variační rozpětí

$$R = x_{\max} - x_{\min}$$

IQR (inter)kvartilové rozpětí

$$IQR = \tilde{x}_{0,75} - \tilde{x}_{0,25}$$



Koeficient šikmosti I

$m_{t,3}$ koeficient (populační) šikmosti ‡ vážený koeficient (populační) šikmosti

$$m_{t,3} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_p} \right)^3 \quad \ddagger \quad m_{t,3} = \frac{1}{N} \sum_{i=1}^k \left(\frac{x_i - \bar{x}}{s_p} \right)^3 \cdot n_i$$

► třetí centrální moment

$m_{t,3}$ koeficient výběrové šikmosti ‡ vážený koeficient výběrové šikmosti
(dle MS Excel 2000)

$$m_{t,3} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_v} \right)^3 \quad \ddagger$$

$$m_{t,3} = \frac{N}{(N-1)(N-2)} \sum_{i=1}^k \left(\frac{x_i - \bar{x}}{s_v} \right)^3 \cdot n_i$$

► Je-li



Koeficient šikmosti II

- $m_{t,3} > 0$, pak mluvíme o kladném zešikmení – vyšší koncentraci podprůměrných hodnot v porovnání s koncentrací hodnot nadprůměrných,
- $m_{t,3} = 0$, pak mluvíme symetrickém zešikmení – stejné koncentraci podprůměrných a nadprůměrných hodnot,
- $m_{t,3} < 0$, pak mluvíme o záporném zešikmení – vyšší koncentraci nadprůměrných hodnot v porovnání s koncentrací hodnot podprůměrných.



Pearsonova míra šikmosti

τ Pearsonova míra šikmosti

$$\tau = \frac{\bar{x} - \hat{x}}{s_x}$$

- ▶ míra šikmosti založená na $\bar{x}$ a $\hat{x}$ (přibližná míra)
- ▶ Je-li
 - $\tau > 0$, pak mluvíme o kladném zešikmení – koncentrace některých podprůměrných hodnot je vyšší v porovnání s koncentrací hodnot nadprůměrných,
 - $\tau = 0$, pak mluvíme symetrickém zešikmení – „průměrné“ hodnoty jsou nejčastější,
 - $\tau < 0$, pak mluvíme o záporném zešikmení – koncentrace některých nadprůměrných hodnot je vyšší v porovnání s koncentrací hodnot podprůměrných.



Míra špičatosti I

$m_{t,4}$ koeficient (populační) špičatosti ‡ vážený koeficient (populační) špičatosti

$$m_{t,4} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_p} \right)^4 \quad \ddagger \quad m_{t,4} = \frac{1}{N} \sum_{i=1}^k \left(\frac{x_i - \bar{x}}{s_p} \right)^4 \cdot n_i$$

- ▶ čtvrtý centrální moment

kurt modifikovaný koeficient (populační) špičatosti

$$kurt = m_{t,4} - 3$$

- ▶ pro lepší srovnávání se od čtvrtého centrálního momentu odečítá 3, což je hodnota koeficientu špičatosti normálního rozdělení; „větší“ respektive „menší“ špičatost pak určíme ve srovnání se špičatostí normálního rozdělení



Míra špičatosti II

$m_{t,4}$ koeficient výběrové špičatosti ‡ vážený koeficient výběrové špičatosti
(dle MS Excel 2000)

$$m_{t,4} = \frac{(n+1)n}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_v} \right)^4 \quad \ddagger$$

$$m_{t,4} = \frac{(N+1)N}{(N-1)(N-2)(N-3)} \sum_{i=1}^k \left(\frac{x_i - \bar{x}}{s_v} \right)^4 \cdot n_i$$

$kurt$ modifikovaný koeficient výběrové špičatosti ‡ vážený modifikovaný
koeficient výběrové špičatosti (dle MS Excel 2000)

$$kurt = m_{t,4} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad \ddagger \quad kurt = m_{t,4} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

► stejné jak pro koeficient (populační) špičatosti

► Je-li

$kurt > 0$, pak mluvíme o kladné špičatosti – koncentrace průměrných hodnot je vyšší,
než bývá u normálního rozdělení,

$kurt = 0$, pak mluvíme o normální špičatosti – koncentrace průměrných hodnot je
právě taková jako u normálního rozdělení,



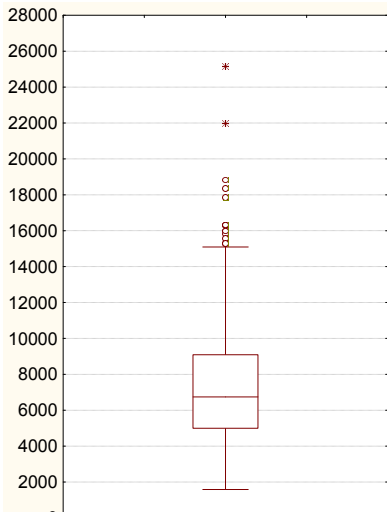
Krabicový graf – Box-plot

- ▶ vizualizace popisných statistik – vybrané míry polohy a vybraných variabilit
- ▶ plným názvem Box-and-whisker(s) plot – krabicový graf s vousy
 - ▶ celý graf je složen z boxu (krabice), „vousů“, příčné čárky respektive čtverečku
 - ▶ jaké hodnoty volit pro nastavení krabicového grafu záleží povaze dat a záměru analýzy – principiálně lze pro nastavení volit parametry $\bar{x}$ a s , to pro „ilustraci“ statistické indukce, nebo $\tilde{x}_{0,50}$ a IQR , to pro analýzu odlehklých pozorování



Box-plot

Cena zaplacená za celkový spotřebitelský úvěr



— $\tilde{x}_{0,50} = 6\,741$

□ $\tilde{x}_{0,25}-\tilde{x}_{0,75}$ tj. 4 995–9 086

┌
└ Rozsah neodlehých hodnot
= (1 584; 15 093)

○ Odlehlé hodnoty

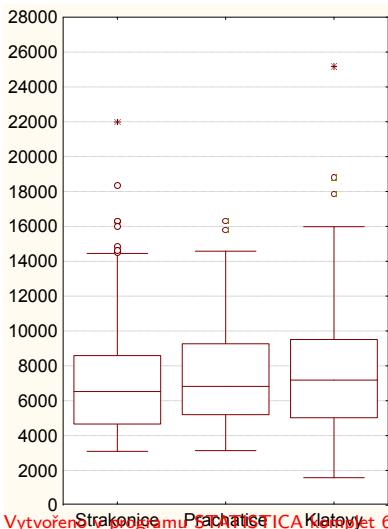
* Extrémní hodnoty

Vytvořeno v programu STATISTICA komplet 6.1 Cz



Box-plot

Cena zaplacená za celkový spotřebitelský úvěr – dle provozoven I



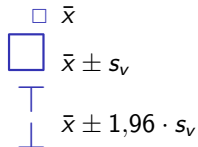
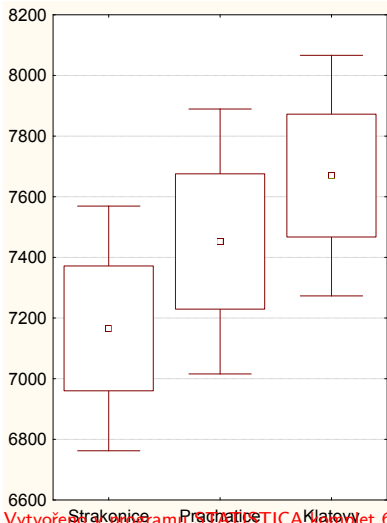
- $\tilde{x}_{0,50}$
- $\tilde{x}_{0,25} - \tilde{x}_{0,75}$
- | — Rozsah neodlehých hodnot
- Odlehlé hodnoty
- * Extrémní hodnoty

Vytvořeno v programu STATISTICA komp 6.1 Cz



Box-plot

Cena zaplacená za celkový spotřebitelský úvěr – dle provozoven II

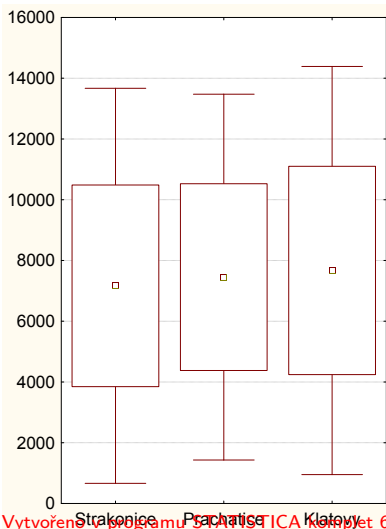


Vytvořeno v programu STATISTICA komp 6.1 Cz



Box-plot

Cena zaplacená za celkový spotřebitelský úvěr – dle provozoven III



$\square \bar{x}$
 $\square \bar{x} \pm \frac{s_v}{\sqrt{n}}$
 $\text{---} \bar{x} \pm 1,96 \cdot \frac{s_v}{\sqrt{n}}$

Vytvořeno v programu STATISTICA komplet 6.1 Cz

