# 1 Popisná statistika

## 1.1 Úvod

V této kapitole se omezíme na popis jednorozměrných statistických souborů. Budeme předpokládat, že jsou k dispozici reálná čísla $x_1, \ldots, x_n$. To mohou být výsledky nějakých měření nebo pokusů. Dost často se stává, že číslo $n$ je velké. Prostý výpis hodnot $x_1, \ldots, x_n$ je pak zcela nepřehledný. Je třeba proto informaci o tomto souboru zkoncentrovat do menšího počtu ukazatelů, které se dají snadno interpretovat.

Z terminologického hlediska si povšimněme, že *soubor hodnot* $x_1, \ldots, x_n$ není totéž jako *množina hodnot* $\{x_1, \ldots, x_n\}$. Důvodem je to, že v souboru se mohou některé hodnoty opakovat. To odpovídá termínu soubor užívanému při práci s počítači.

Někdy je nutné čísla $x_1, \ldots, x_n$ uspořádat do neklesající posloupnosti

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}.$$

To tedy znamená, že specielně

$$x_{(1)} = \min_i x_i, \qquad x_{(n)} = \max_i x_i.$$

V této kapitole se nejdřív zmíníme o případu, kdy skutečně je možné s celým souborem $x_1, \ldots, x_n$ pracovat najednou. To je dnes už poměrně časté díky velké paměti počítačů. I přesto se někdy z různých důvodů stane, že je nutné pracovat s daty shrnutými do nějaké tabulky. O práci s tabulkami se zmíníme v závěru této kapitoly.

## 1.2 Míry polohy

Snad nejznámější mírou polohy je *aritmetický průměr*

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

Stručně mu budeme říkat jen *průměr*. Dnes je sice proti jeho používání v některých situacích vznášena řada námitek (např. jeho citlivost na hrubé chyby — i jeden jediný údaj $x_i$ může výsledek zcela zkreslit), ale už vzhledem k jeho úzké souvislosti se *souhrnem* $\sum x_i = n\bar{x}$ stále zůstává důležitým ukazatelem. Není pochyb o tom, že třebas souhrn úspor obyvatelstva přináší podstatnou informaci o ekonomické situaci státu.

Velmi jednoduchou, ale neustále používanou vlastností průměru je

$$\sum(x_i - \bar{x}) = 0. \tag{1.1}$$

Slovy vyjádřeno to znamená, že součet odchylek od aritmetického průměru je roven nule. Důkaz vzorce (1.1) je triviální. Uveďme další jednoduchou vlastnost průměru. Nechť $a, b$ jsou daná čísla a položme $y_i = a + bx_i$, $i = 1, \ldots, n$. Pak

$$\bar{y} = a + b\bar{x}. \tag{1.2}$$

Dříve se *lineárních transformací* používalo hlavně ke zjednodušení numerického výpočtu průměru.

Jsou-li všechna čísla $x_1, \ldots, x_n$ kladná, pak se definuje *geometrický průměr* vzorcem

$$\bar{x}_G = \sqrt[n]{x_1 \ldots x_n}.$$

**Příklad 1.1** Naznačme jedno použití geometrického průměru. Nechť $x_i$ je počet výrobků prodaných v $i$-tém časovém období (např. počet automobilů prodaných v $i$-tém roce), $i = 0, \ldots, n$. Vývoj prodeje se dá charakterizovat pomocí tzv. *řetězových indexů*

$$i_1 = \frac{x_1}{x_0}, \ldots, i_n = \frac{x_n}{x_{n-1}}.$$

Pomocí nich lze množství $x_n$ vyjádřit jako

$$x_n = x_0 i_1 \ldots i_n. \tag{1.3}$$

Soubor indexů $i_1, \ldots, i_n$ je výhodné charakterizovat jejich geometrickým průměrem

$$\bar{\imath}_G = \sqrt[n]{i_1 \ldots i_n}.$$

Ten totiž může zastoupit každý z indexů $i_1, \ldots, i_n$ ve vzorci (1.3) v tom smyslu, že

$$x_n = x_0 \left(\bar{\imath}_G\right)^n. \qquad \diamondsuit$$

**Věta 1.2** *Nechť $x_1, \ldots, x_n$ jsou kladná čísla. Pak platí*

$$\bar{x}_G \leq \bar{x} \tag{1.4}$$

*a rovnosti je dosaženo právě tehdy, platí-li $x_1 = \cdots = x_n$.*

*Důkaz.* Položme $f(x) = \ln x$ pro $x > 0$. Podle Taylorova vzorce pro libovolné pevné $\mu > 0$ platí

$$f(x) = f(\mu) + (x-\mu)f'(\mu) + \frac{1}{2}(x-\mu)^2 f''(\theta), \tag{1.5}$$

kde bod $\theta$ leží mezi $x$ a $\mu$. Protože $f''(\theta) = -\theta^{-2} < 0$ pro každé $\theta > 0$, dostáváme

$$f(x) \leq f(\mu) + (x-\mu)f'(\mu). \tag{1.6}$$

Je jasné, že rovnost v (1.6) platí právě tehdy, je-li $x = \mu$. Proto pro každé $x_i$ máme

$$\ln x_i \leq \ln \mu + (x_i - \mu)f'(\mu), \qquad i = 1, \ldots, n. \tag{1.7}$$

Sečtením těchto nerovností se dostane

$$\sum \ln x_i \leq n \ln \mu + f'(\mu) \sum (x_i - \mu).$$

Položme $\mu = \bar{x}$. Užitím (1.1) se dostane

$$\sum \ln x_i \leq n \ln \bar{x}, \tag{1.8}$$

což je ekvivalentní s (1.4). Přitom rovnost v (1.8) platí právě tehdy, platí-li rovnost v (1.6) pro každé $i$, tj. když $x_1 = \mu, \ldots, x_n = \mu$. Tedy rovnost v (1.8) nastane, právě když si jsou všechna čísla $x_1, \ldots, x_n$ rovna. □

Jsou-li všechna čísla $x_1, \ldots, x_n$ kladná, definuje se *harmonický průměr* $\bar{x}_H$ vztahem

$$\bar{x}_H = \frac{n}{x_1^{-1} + \cdots + x_n^{-1}}.$$

**Věta 1.3** *Jsou-li všechna čísla $x_1, \ldots, x_n$ kladná, pak platí*

$$\bar{x}_H \leq \bar{x}_G. \tag{1.9}$$

*Přitom rovnosti v (1.9) je dosaženo právě tehdy, platí-li $x_1 = \cdots = x_n$.*

*Důkaz.* Užitím věty 1.2 dostaneme

$$\frac{1}{\bar{x}_H} = \frac{1}{n} \sum x_i^{-1} \geq \sqrt[n]{x_1^{-1} \ldots x_n^{-1}},$$

což je ekvivalentní s (1.9). Podmínka pro rovnost se rovněž přenáší z věty 1.2. □

**Příklad 1.4** Význam harmonického průměru může osvětlit následující úloha. Auto jede do kopce rychlostí $v_1$ a pak po stejné trati zpět z kopce rychlostí $v_2$. Jaká je jeho průměrná rychlost?

Označme $d$ délku tratě do kopce, $t_1$ dobu jízdy do kopce a $t_2$ dobu jízdy z kopce. Máme $t_1 = d/v_1$, $t_2 = d/v_2$. Celková dráha do kopce i zpět je $2d$, celkový čas potřebný k jejímu projetí je $t = t_1 + t_2$. Proto je průměrná rychlost

$$v = \frac{2d}{t} = \frac{2}{v_1^{-1} + v_2^{-1}} \tag{1.10}$$

a je to tedy harmonický průměr rychlostí $v_1$ a $v_2$. ♢

Někdy se definuje *kvadratický průměr* $\bar{x}_K$ vztahem

$$\bar{x}_K = \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}}.$$

**Věta 1.5** *Pro libovolná reálná čísla $x_1, \ldots, x_n$ platí*

$$\bar{x} \le \bar{x}_K. \tag{1.11}$$

*Přitom rovnost nastává tehdy a jen tehdy, je-li $x_1 = \cdots = x_n$.*

*Důkaz.* Položme $f(x) = x^2$. Podle Taylorova vzorce platí pro každé pevně dané reálné $\mu$ vzorec (1.5), kde nyní $f''(\theta) = 2$. Máme tedy

$$f(x) \ge f(\mu) + (x - \mu) f'(\mu),$$

tj.

$$x^2 \ge \mu^2 + (x - \mu) f'(\mu).$$

Přitom zde rovnost platí tehdy a jen tehdy, je-li $x = \mu$. Tudíž

$$x_i^2 \ge \mu^2 + (x_i - \mu) f'(\mu), \qquad i = 1, \ldots, n$$

a sečtením dostaneme

$$\sum x_i^2 \ge n\mu^2 + f'(\mu) \sum (x_i - \mu). \tag{1.12}$$

Zvolíme-li $\mu = \bar{x}$, pak užitím (1.1) máme

$$\sum x_i^2 \ge n\bar{x}^2, \tag{1.13}$$

což je ekvivalentní s (1.11). Rovnost v (1.13) platí právě tehdy, platí-li v (1.12) rovnost pro každé $i = 1, \ldots, n$. To znamená, že si musí být všechna $x_i$ rovna. □

Nechť všechna čísla $x_1, \dots, x_n$ jsou kladná. Snadno se lze přesvědčit, že platí

$$x_{(1)} \le \bar{x}_H, \qquad \bar{x}_K \le x_{(n)}.$$

Proto z dosavadních výsledků vyplývá, že

$$x_{(1)} \le \bar{x}_H \le \bar{x}_G \le \bar{x} \le \bar{x}_K \le x_{(n)}. \tag{1.14}$$

V případě $x_1 = \cdots = x_n$ platí všude v (1.14) rovnost. Nejsou-li si všechna $x_i$ rovna, pak v (1.14) jsou všude ostré nerovnosti.

Zhruba se dá říci, že $\bar{x}$ slouží k náhradě individuálních $x_i$ při sčítání, neboť

$$x_1 + \cdots + x_n = \bar{x} + \cdots + \bar{x}.$$

Obdobně $\bar{x}_G$ nahrazuje jednotlivé hodnoty $x_i$ při násobení, $\bar{x}_H$ při počítání s převrácenými hodnotami a $\bar{x}_K$ při sčítání čtverců v tom smyslu, že

$$x_1 \dots x_n = \bar{x}_G \dots \bar{x}_G,$$
$$\frac{1}{x_1} + \cdots + \frac{1}{x_n} = \frac{1}{\bar{x}_H} + \cdots + \frac{1}{\bar{x}_H},$$
$$x_1^2 + \cdots + x_n^2 = \bar{x}_K^2 + \cdots + \bar{x}_K^2.$$

Je evidentní, že by se daly definovat jiné typy průměrů, které by nahrazovaly původní hodnoty při dalších operacích. Zde rozebereme jednu takovou možnost. Zřejmě platí

$$\bar{x}_H = \left(\frac{1}{n}\sum x_i^{-1}\right)^{-1}, \qquad \bar{x} = \frac{1}{n}\sum x_i, \qquad \bar{x}_K = \left(\frac{1}{n}\sum x_i^2\right)^{\frac{1}{2}}.$$

Zaveďme tedy v případě kladných $x_1, \dots, x_n$ *průměr stupně* $\alpha$ vzorcem

$$\bar{x}_\alpha = \left(\frac{1}{n}\sum x_i^\alpha\right)^{1/\alpha}, \qquad \alpha \ne 0. \tag{1.15}$$

Vidíme, že

$$\bar{x}_{-1} = \bar{x}_H, \qquad \bar{x}_1 = \bar{x}, \qquad \bar{x}_2 = \bar{x}_K.$$

Ve vzorci (1.15) bylo nutné vyloučit hodnotu $\alpha = 0$. Nyní vyšetříme limitu $\bar{x}_\alpha$ pro $\alpha \to 0$.

**Věta 1.6** *Nechť $x_1, \dots, x_n$ jsou vesměs kladná čísla. Pak platí*

$$\lim_{\alpha \to 0} \bar{x}_\alpha = \bar{x}_G.$$