

# 1. PODSTATA STATISTIKY

*Původní význam* - pouhé sbírání čísel

(název z latinského "status" = stát, použití k označení vědy zabývající se sběrem informací o státu - o počtu obyvatel, ekonomice,...)

*Dnešní pojetí* - shromažďování, klasifikace a tabelování dat, ale také analýza informací za účelem formulování obecných závěrů a rozhodování.

## 1.1 Dva základní typy statistiky

- *Popisná statistika* (descriptive statistics)
- *Inferenční statistika* (inferential statistics)

*Popisná statistika* se skládá z metod pro zjišťování a sumarizaci informací.

*Inferenční statistika* se skládá z metod pro přijímání a měření spolehlivosti závěrů o základním souboru založených na informacích získaných z výběru ze ZS.

## 1.2 Základní statistické pojmy

*Hromadné jevy a procesy* - jevy a procesy vyskytující se u velkého množství prvků.

**Statistické jednotky** - elementární jednotky stat. pozorování (např. osoby, organizace, věci, události,...)

**Statistické znaky - veličiny** (variable) - vlastnosti statistických jednotek (pracovník podniku: mzda, stáří, kvalifikační třída, nejvyšší dosažené vzdělání).

**Data** - informace získané měřením hodnot statistických znaků.

## *Základní třídění statistických znaků a dat*

### **Veličiny:**

**Kvantitativní:** Hodnoty veličiny lze vyjádřit číselně.

**Kvalitativní:** Hodnoty veličiny nelze vyjádřit číselně.

**Diskrétní:** Kvantitativní veličiny, jejichž možné hodnoty tvoří konečnou nebo spočetnou množinu čísel.

**Spojitě:** Kvantitativní veličiny, jejichž možné hodnoty tvoří číselný interval.

statistické znaky

kvantitativní

kvalitativní

nespojité

spojité

**Statistický soubor** (SS) - množina všech statistických jednotek, u nichž zkoumáme příslušné statistické znaky.

*Jednorozměrný SS* - u každé statistické jednotky zjišťujeme pouze jeden statistický znak.

*Vícerozměrný SS* - u každé stat. jednotky zjišťujeme dva nebo více stat. znaků.

**Základní soubor** (ZS) (population) - SS všech jednotek, který je vlastním předmětem sledování, o němž chceme provádět závěry.

**Výběrový soubor - výběr** (sample) - část ZS vybraná určitým způsobem, z které jsou shromažďovány informace.

**Rozsah výběru** - počet jednotek vybraných ze ZS.

## 1.3 Náhodný výběr

*Prostý náhodný výběr* (simple random sample)

*Vícestupňový náhodný výběr* (multistage r.s.)

*Proč výběr?* - obecné zásady

1. Omezené zdroje
2. řídký výskyt
3. Destruktivní testování
4. Výběr může být přesnější

## **Znáhodněné pokusy**

**A** - Pokusné a kontrolní skupiny

**B** - Náhodné přiřazení

**C** - Utajení a dvojité utajení

## **Pozorovací studie versus znáhodněné pokusy**

**A** - Znáhodnění někdy není možné

**B** - Znáhodnění někdy není praktické

**C** - Znáhodnění se někdy neprovádí i když by bylo praktické

**D** - Některé etické problémy

**E** - Odstranění jednostrannosti z pozorovacích studií: *regrese*

*Na volbě statistických jednotek a vhodném výběru statistických znaků, pomocí nichž chceme sledovat vlastnosti statistického souboru, závisí úspěch i výsledky veškeré další práce.*

## 2. POPISNÁ STATISTIKA

### 2.1 Elementární zpracování statistických údajů

*Třídění* – rozdělení jednotek souboru do takových skupin, aby co nejlépe vynikly charakteristické vlastnosti zkoumaných jevů (uspořádání a zhuštění údajů)

*Jednostupňové třídění* - podle 1 stat. znaku.

*Vícestupňové třídění* - podle více stat. znaků najednou

#### 2.1.1 Statistické tabulky

*Rozdělení četností a relativních četností*

Naměřené hodnoty kvantitativního znaku nazýváme **pozorování, měření** nebo **vstupní data**.

**Absolutní četnost** (frequency) - počet příslušných jednotek, přiřazených každé hodnotě zkoum. znaku

**Poměrná (relativní) četnost** (relative frequency) - podíl jednotlivých absolutních četností a celkového rozsahu souboru  $n$  všech pozorování souboru.

Nechť  $y_i, i = 1, \dots, k, 1 \leq k \leq n$  jsou různé hodnoty diskř. znaku a  $n_i$  odpovídající četnosti,  $n$  je rozsah souboru,  $n = \sum_{i=1}^k n_i \implies$  relativní četnost  $f_i$

$$f_i = \frac{n_i}{n} \quad \text{a platí} \quad \sum_{i=1}^k f_i = 1$$

**Absolutní kumulativní č.** (cumulative frequency) hodnot znaku menších nebo rovných  $y_r: \sum_{i=1}^r n_i, 1 \leq r \leq k$ .

**Poměrná kumulativní č.** (cumulative relative f.) hodnot znaku menších nebo rovných  $y_r: \sum_{i=1}^r f_i, 1 \leq r \leq k$ .

## 1) Diskrétní veličina

### Rozdělení četností a relativních četností diskrétní veličiny

Tabulka rozdělení četností – vhodný prostředek pro zpracování diskrétního znaku, který nabývá pouze menšího počtu hodnot.

Hodnota znaku	Četnost		Kumulativní četn.	
	absolutní	relativní	absolutní	relativní
$y_i$	$n_i$	$f_i$		
$y_1$	$n_1$	$f_1$	$n_1$	$f_1$
$y_2$	$n_2$	$f_2$	$n_1 + n_2$	$f_1 + f_2$
...	...	...	...	...
$y_k$	$n_k$	$f_k$	$\sum_{i=1}^k n_i$	$\sum_{i=1}^k f_i$
Celkem	$\sum_{i=1}^k n_i = n$	$\sum_{i=1}^k f_i = 1$		

2) Spojitá veličina nebo diskrétní, která může nabývat velkého počtu různých hodnot

**Intervalové rozdělení četností** - rozdíl mezi maximální a minimální zjištěnou hodnotou znaku rozdělíme na určitý počet intervalů a pak zjistíme počty hodnot patřících do těchto intervalů.

– počet intervalů

– hranice intervalů

Nechť  $\{x_1, \dots, x_n\}$ ,  $x_i \in [a, b]$ ,  $i = 1, \dots, n$ ,

$a = a_0 < a_1 < \dots < a_k = b$ ,  $k$  disjunktních podintervalů  $(a_{i-1}, a_i]$ ,  $i = 1, 2, \dots, k$  – **třídní intervaly - třídy** (classes).

**Dolní hranice  $i$ -té třídy** (lower class limit) – nejmenší hodnota v intervalu  $(a_{i-1}, a_i]$ .

**Horní hranice  $i$ -té třídy** (upper class limit) – největší hodnota v intervalu  $(a_{i-1}, a_i]$ .

**Střed  $i$ -té třídy** (class mark) – střed intervalu  $(a_{i-1}, a_i]$ .

**Šířka  $i$ -té třídy** (class width) – rozdíl mezi horní hranicí intervalu  $(a_{i-1}, a_i]$  a horní hranicí  $(a_{i-2}, a_{i-1}]$

**Absolutní četnost  $n_i$   $i$ -té třídy** – počet pozorování  $x_j$ :  
 $a_{i-1} < x_j \leq a_i$ .

**Poměrná (relativní) četnost  $f_i$   $i$ -té třídy:**  $f_i = \frac{n_i}{n}$ .

**Absolutní kumulativní četnost  $N_i$   $i$ -té třídy** – počet pozorování  $x_j$ :  $a_{i-1} < x_j \leq a_i$ ,  $N_i = \sum_{r=1}^i n_r$ .

**Poměrná kumulativní četnost  $F_i$   $i$ -té třídy:**

$$F_i = \sum_{r=1}^i f_r.$$

Pozorování, která patří do jedné skupiny nahrazujeme při výpočtech statistických charakteristik jedinou zastupitelnou hodnotou - zpravidla **střed intervalu**.

## 2.1.2 Statistické grafy

Z hlediska konstrukce lze grafy rozdělit do různých skupin:

a) *spojnicové a sloupkové grafy*

a<sub>1</sub>) *spojnicové*

**polygon četností:** – v pravoúhlém souř. systému spojíme úsečkami body o souřadnicích  $(x_i, n_i)$ ,

**polygon relativních četností:** – v pravoúhlém souř. systému spojíme úsečkami body o souřadnicích (střed  $i$ -té třídy,  $f_i$ ).

*Tvar rozdělení četností*

**Modus** – nejčtenější hodnota znaku.

**Jednovrcholová** (unimodal) – modus leží mezi minimální a maximální hodnotou veličiny  
(**rozdělení J, obrácené rozdění J**).

**Vícevrcholová** (multimodal) – více než jeden modus  
(**bimodální – rozdění U**)

*a<sub>2</sub>) sloupkové*

**histogram četností:** – sloupkový graf tvořený pravidelnými rovnoběžníky, jejichž základny mají délku zvolených intervalů a jejichž výšky mají velikost příslušných relativních četností;

*b) výsečové grafy - kruhové diagramy:* (pie chart) – relativní četnosti hodnot znaku znázorňujeme pomocí výsečí kruhu, které získáme rozdělením středového úhlu úměrně k podílu jednotlivých částí zobrazovaného jevu vyjádřených v procentech;

*c) stonek s listy* (stem and leaf diagram or stemplot) – grafická obdoba četnostního histogramu;

*d) krabicový graf* (box and whiskers plot) – slouží k znázornění extrémních hodnot souboru a kvartilů.

## 2.2 Popis jednorozměrných SS

Nechť  $x_i$ ,  $i = 1, \dots, n$  jsou pozorování diskrétního stat. znaku a necht'  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  je uspořádaná posloupnost těchto pozorování.

### 2.2.1 Kvantily

Kvantil je taková hodnota, která rozděluje soubor hodnot určitého znaku na dvě části - jedna obsahuje ty hodnoty, které jsou menší (nebo stejné) než tento kvantil, druhá část naopak obsahuje ty hodnoty, které jsou větší (nebo stejné) než tento kvantil.

– **100p% kvantil - percentil** ( $0 \leq p \leq 1$ ): kvantil, který odděluje zhruba 100p% malých hodnot znaku ( $p$  je relativní četnost malých hodnot) od 100(1 - p)% velkých hodnot znaku.

$$\tilde{x}_{100p} = \begin{cases} x_{([np]+1)} & \text{pokud není } np \text{ celé číslo} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}) & \text{pro } np \text{ celé.} \end{cases}$$

– **Dolní kvartil** ( $\tilde{x}_{25}$ ) (lower quartile): odděluje zhruba 25% nejnižších hodnot znaku od ostatních.

$$\tilde{x}_{25} = \begin{cases} x_{([\frac{n}{4}]+1)}, & \text{pokud je } n \text{ nedělitelné } 4 \\ \frac{1}{2}(x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)}) & \text{pro } n \text{ dělitelné } 4 \end{cases}$$

– **Prostřední kvartil - medián** ( $\tilde{x}_{50}$ ): hodnota, která odděluje 50% hodnot znaku menších nebo rovných této hodnotě od ostatních.

$$\tilde{x}_{50} = \begin{cases} x_{([\frac{n}{2}]+1)}, & \text{pokud je } n \text{ liché číslo} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{pro } n \text{ sudé.} \end{cases}$$

– **Horní kvartil** ( $\tilde{x}_{75}$ ) (upper quartile): odděluje zhruba 75% nejnižších hodnot znaku od zbývajících 25%

$$\tilde{x}_{75} = \begin{cases} x_{([\frac{3}{4}n]+1)}, & \text{pokud je } n \text{ nedělitelné } 4 \\ \frac{1}{2}(x_{(\frac{3}{4}n)} + x_{(\frac{3}{4}n+1)}) & \text{pro } n \text{ dělitelné } 4 \end{cases}$$

### 2.2.2 Míry polohy (measures of central tendency)

#### **A) ARITMETICKÝ PRŮMĚR** ( $\bar{x}$ ) (mean)

Nechť  $x_1, x_2, \dots, x_n$  jsou pozorované hodnoty znaku  $x$ ,  $n$  je celkový počet pozorování

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

Jsou-li zjištěné hodnoty znaku uspořádány do tabulky rozdělení četností, pak

$$\bar{x} = \frac{\sum_{i=1}^k y_i n_i}{\sum_{i=1}^n n_i} = \frac{1}{n} \sum_{i=1}^k y_i n_i = \sum_{i=1}^k y_i f_i$$

Četnosti  $f_i$  udávají váhu, která je přisuzována jednotlivým  $k$  různým hodnotám  $y_i$  znaku.

#### **B) MEDIÁN** ( $\tilde{x}_{50}$ ) (median)

$$\tilde{x}_{50} = \begin{cases} x_{([\frac{n}{2}]+1)}, & \text{pokud je } n \text{ liché číslo} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{pro } n \text{ sudé.} \end{cases}$$

#### **C) MODUS** ( $\hat{x}$ ) (mode)

Modus – nejčastěji se vyskytující hodnota (každá hodnota, jejíž četnost je větší než jedna a je stejně velká nebo větší než četnost jiných hodnot).

## D) Vzájemná poloha modu, mediánu, průměru

*Symetrické* rozdělení:  $\hat{x} = \bar{x} = \tilde{x}_{50}$

*Nesymetrické* rozdělení: vzhledem k  $\hat{x}$  leží  $\tilde{x}_{50}$  ve směru delší části rozdělení a  $\bar{x}$  dále v tomto směru.

Popisná míra se nazývá **resistentní**, jestliže *není citlivá* na vliv malého počtu extrémních hodnot.

*Která charakteristika je nejvhodnější?*

*Shrnutí:*

- Modus je char., kterou lze nejsnadněji nalézt, ale která nemá velký význam při hledání polohy rozdělení.
- Medián užitečnější, představuje typičtější hodnotu.
- Průměr zahrnuje všechna pozorování.

## USEKNUTÉ PRŮMĚRY

Nechť  $x_1, x_2, \dots, x_n$  je posloupnost pozorovaných hodnot statistického znaku

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  je uspořádaná posloupnost

$0 < \alpha < 0.5$ ,

$[n\alpha]$  je největší celé číslo  $k$  splňující  $k \leq [n\alpha]$

$\alpha$ -usekнутý průměr ( $\alpha$ -trimmed mean)

$$\bar{x}_\alpha = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}$$

$\alpha$ -winsorizovaný průměr ( $\alpha$ -winsorized mean)

$$\bar{x}_{\alpha w} = \frac{1}{n} \left\{ [n\alpha]x_{([n\alpha])} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)} + [n\alpha]x_{(n-[n\alpha]+1)} \right\}$$

### 2.2.3 Míry rozptýlenosti (measures of dispersion)

Rozdělení čet. mohou mít shodnou polohu, ale přesto se od sebe výrazně liší.

Význam při posuzování vypovídací schopnosti průměru: vypovídací schopnost je tím větší, čím je rozptýlenost sledovaného znaku menší.

#### Míry absolutní rozptýlenosti

##### A) Variační rozpětí ( $R$ ) (range)

$$R = x_{(max)} - x_{(min)}$$

##### B) Mezikvartilové rozpětí ( $IQR$ ) (interquartile range)

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

$IQR$  je rezistentní.

##### C) Střední absolutní odchylka ( $MAD$ ) (mean of absolute deviation)

absolutní odchylka od průměru =  $|x_i - \bar{x}|$

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

##### D) Střední kvadratická odchylka ( $MSD$ ) (mean of squared deviation)

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

## E) Rozptyl a směrodatná odchylka

– Rozptyl ( $s^2$ ) (dispersion)

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

– Směrodatná odchylka ( $s$ ) (standard deviation)

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

Směrodatná odchylka není rezistentní.

*Interpretace směrodatné odchylky:*

**Čebyševova pravidla:**

Pro každou množinu dat platí:

*Vlastnost 1:*

Alespoň 75% dat leží mezi  $\bar{x} - 2s$  a  $\bar{x} + 2s$

*Vlastnost 2:*

Alespoň 89% dat leží mezi  $\bar{x} - 3s$  a  $\bar{x} + 3s$

*Vlastnost 3:*

Obecně, pro každé  $k > 1$

alespoň  $1 - \frac{1}{k^2}$  dat leží mezi  $\bar{x} - k.s$  a  $\bar{x} + k.s$

## F) Kvartilová odchylka (Q)

$$Q = \frac{\tilde{x}_{75} - \tilde{x}_{50}}{2} + \frac{\tilde{x}_{50} - \tilde{x}_{25}}{2} = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2}$$

*Nevýhoda:* nezachycuje rozptýlenost všech hodnot.

## Míry relativní rozptýlenosti

### A) Variační koeficient ( $V_x$ )

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

$100V_x$  udává rozptýlenost v procentech. Hrubý odhad:  $V_x$  vyšší než 50% je příznakem značné nesooudosti SS.

### B) Relativní kvartilová odchylka ( $Q_r$ )

$$Q_r = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{\tilde{x}_{75} + \tilde{x}_{25}}$$

## 5-ti číselná charakteristika (five-number summary):

$$x_{min}, \tilde{x}_{25}, \tilde{x}_{50}, \tilde{x}_{75}, x_{max}$$

## 2.3 Lineární transformace

### A) Změna počátku

$$x' = x + a \implies \bar{x}' = \bar{x} + a, \quad s_{x'} = s_x$$

### B) Změna měřítka

$$x^* = bx \implies \bar{x}^* = b\bar{x}, \quad s_{x^*} = |b|s_x$$

### C) Lineární transformace obecně

$$y = a + bx \implies \bar{y} = a + b\bar{x}, \quad s_y = |b|s_x$$