Deskriptivní statistika [1]

[1]Tyto materiály byly vytvořeny za pomoci grantu FRVŠ číslo 1145/2004.

Základní charakteristiky souboru

Pro lepší představu používáme k popisu vlastností zkoumaného jevu určité charakteristiky - statistiky. Statistikami zde rozumíme jistá čísla, která jsou nositeli důležitých informací o zkoumaných jevech. Způsob jejich zjišťování je jednoznačně dán.

Pro veličinu v měřítku alespoň ordinálním, lze vytvořit uspořádaný soubor z původního neuspořádaného souboru o velikosti n takto:

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

Pak lze snadno zjistit **maximum** či **minimum**
Minimum:

$$x_{min} = x_{(1)} \tag{1}$$

Maximum:

$$x_{max} = x_{(n)} \tag{2}$$

Charakteristiky polohy - úrovně

Míry polohy charakterizují obecnou úroveň (polohu) hodnot statistického znaku. Tyto statistiky lze dělit na průměry a ostatní střední hodnoty. První a nejčastěji používanou charakteristikou je všeobecně známý aritmetický průměr. Ten je definován takto:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ , \tag{3}$$

jeho vážená varianta

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i \ , \tag{4}$$

kde n_i jsou absolutní četnosti v jednotlivých k třídách. Jednotlivé hodnoty x_i, jsou buď hodnoty znaku (v případě prostého třídění) nebo středy intervalů (v případě intervalového třídění).

Další míry polohy, řadící se mezi průměry, jsou harmonický a geometrický průměr. Ty jsou definovány po řadě takto:

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{5}$$

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} \ . \tag{6}$$

Další charakteristikou je např. kvadratický průměr definovaný jako

$$\bar{x}_K = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2} \ . \tag{7}$$

Kvantil

Velmi důležitým pojmem ve statistické teorii je pojem kvantilu. Je definován následovně: $100P\%$-ním kvantilem $\tilde{x}_P$ statistického znaku X je takové vhodně zvolené číslo, pro které platí že $100P\%$ hodnot znaku je menších než $\tilde{x}_P$ a $100(1-P)\%$ hodnot znaku je větších než toto číslo.

Mezi nejpoužívanější kvantily patří: **dolní kvartil** $\tilde{x}_{25}$, **medián** $\tilde{x}_{50}$ a **horní kvartil** $\tilde{x}_{75}$. Tyto tři kvantily rozdělují uspořádanou řadu dat na zhruba čtyři části s přibližně stejnými rozsahy. Ve statistické praxi se lze setkat i s **decily** nebo **percentily**.

Ostatní střední hodnoty

Při charakterizování souboru se někdy s výhodou používá tzv. **medián**, který udává *prostřední* hodnotu souboru. Jde o tzv. robustní charakteristiku. V uspořádaném souboru $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(l)} \leq \cdots \leq x_{(n)}$ musí počet menších nebo stejných hodnot jako medián činit alespoň tolik, jako počet hodnot větších či stejných jako medián. Použití mediánu přichází v úvahu již u ordinální stupnice. Medián lze definovat takto:

$$\tilde{x}_{50} = \begin{cases} x_{(\frac{n+1}{2})} & \text{liché } n, \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{sudé } n. \end{cases} \tag{8}$$

V případě intervalového třídění dat nelze stanovit medián přesně. V takovém případě lze s jistotou stanovit pouze mediánový interval, tj. interval ve kterém medián leží. Hodnotu mediánu pak stanovíme lineární interpolací.

$$\tilde{x}_{50} = x_0 + \frac{\frac{n+1}{2} - \sum_{i=1}^{n_{j-1}} n_i}{n_j} h, \tag{9}$$

kde x_0 je dolní mez mediánového intervalu, n_j je četnost mediánového intervalu, h délka mediánového intervalu a $\sum_{i=1}^{n_{j-1}} n_i$ je kumulativní četnost intervalů, předcházející mediánový interval.

Modem souboru je hodnota $\hat{x}$, která se v souboru nejčastěji opakuje, tj. má největší četnost. Z tohoto hlediska lze rozeznávat unimodální, bimodální a multimodální soubory. Pokud je soubor intervalově tříděn, pak nelze určit modus přesně. Přesně lze stanovit pouze modální, tj. nejčetnější interval. Přibližnou hodnotu modu určíme v tomto případě dle vzorce

$$\hat{x} = \hat{x}_0 + \frac{h}{2} \frac{n_1 - n_{-1}}{2n_0 - n_1 - n_{-1}}, \tag{10}$$

kde n_{-1} a n_1 jsou četnosti intervalu který předchází resp. následuje za modálním intervalem. Délka a četnost modálního intervalu je označena po řadě symboly h a n_0. Střed modálního intervalu je označen symbolem $\hat{x}_0$.

Pro získání základní představy o rozložení studovaného souboru zpravidla stačí uvést $\bar{x}, \hat{x}, \tilde{x}_{25}, \tilde{x}_{75}$ a hodnotu max a min, v případě multimodálního rozdělení pak i jednotlivá maxima souboru. Pro úplnost lze dodat, že hodnota modu je

značně ovlivněna variabilitou znaku a to zejména při menších rozsazích výběrů. U jednovrcholových rozdělení platí přibližně vztah

$$\hat{x} = 3\tilde{x}_{50} - 2\bar{x} \quad . \tag{11}$$

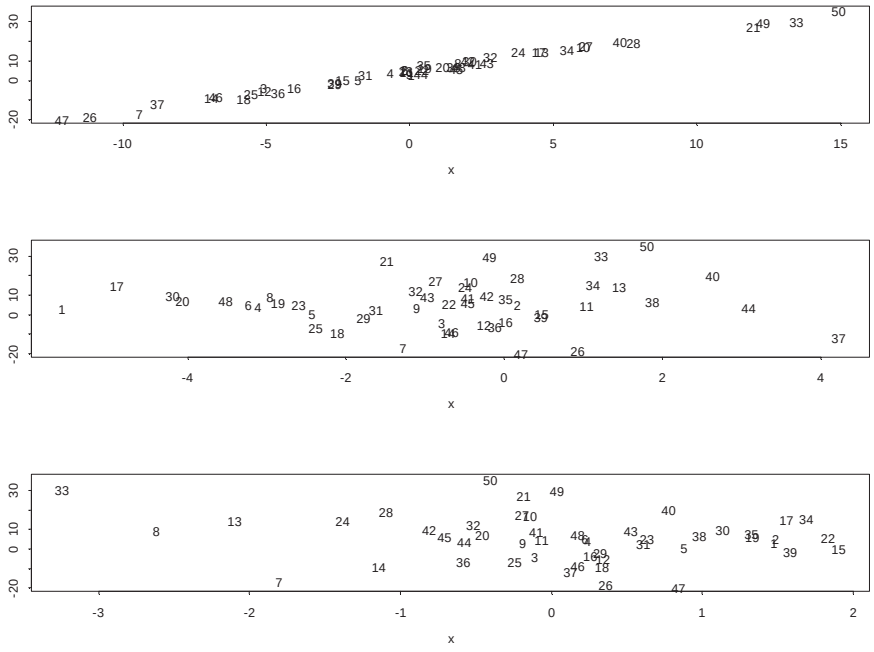
Průměrná chyba

Průměrná chyba byla zavedena jako protiklad směrodatné odchylky na základě přesvědčení, že je vhodnější měřit variabilitu hodnot na základě aritmetického průměru odchylek spíše než na základě kvadratického průměru. Průměrná chyba $\bar{d}$ vypočtená z řady n hodnot $x_1, x_2, \cdots, x_n$ je definována jako

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad . \tag{12}$$

Míry variability

Další důležitou vlastností, kterou je třeba umět charakterizovat, je variabilita dat. Míry variability určitým způsobem charakterizují proměnlivost hodnot. Míry variability jsou v podstatě dvojího typu. První z nich se počítají pouze z některých hodnot Druhá skupina naopak vychází ze všech hodnot obsažených ve studovaném souboru.



Rozpětí

Je nejjednodušší mírou variability. Jde o první typ měr variability.

$$R = x_{max} - x_{min} \tag{13}$$

Kvartilové rozpětí

Je definováno jako rozdíl mezi horním a dolním kvartilem tj.:

$$R_q = \tilde{x}_{75} - \tilde{x}_{25} \ . \tag{14}$$

Takto definované rozpětí vychází z cca 50% typických znaků sledovaného souboru.

Rozptyl

Je jednou z nejdůležitějších charakteristik variability dat. Je definován jako **aritmetický průměr čtverců odchylek od aritmetického průměru**. Z hlediska jeho konstrukce pozeznáváme rozptyl prostý a vážený. Dále rozptyl prostý výběrový a rozptyl vážený výběrový.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{15}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 n_i \tag{16}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{17}$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \tag{18}$$

Směrodatná odchylka

Vzhledem k tomu, že je rozptyl špatně interpretovatelný, používá se při charakterizování rozptýlenosti dat spíše **směrodatná odchylka**. Ta je definována jako **druhá odmocnina rozptylu**, tj.:

$$\sigma = \sqrt{\sigma^2} \tag{19}$$

a výběrová směrodatná odchylka

$$s = \sqrt{s^2} \ . \tag{20}$$

Variační koeficient

Je relativní mírou variability a vyjadřuje se nejčastěji v procentech. Používáme jej **při porovnávání variability statistických znaků které se liší z hlediska míry polohy nebo mají odlišné měrné jednotky**. Variační koeficient udává z kolika procent se podílí směrodatná odchylka na aritmetickém průměru.

$$V_X = \frac{\sigma}{\mu} \ . \tag{21}$$

Obdobně pak i pro výběrovou formu variačního koeficientu jako

$$V_X = \frac{s}{\bar{x}} \ . \tag{22}$$

Entropie

U veličin s nominálním měřítkem **nelze** použít klasických charakteristik k posouzení variability dat. V takovém případě lze použít například tzv. entropii definovanou vzorcem

$$H = -\sum_{i=1}^{m} \frac{n_i}{n} \ln \frac{n_i}{n} \ . \tag{23}$$

Entropie dosahuje vysokých hodnot, pokud jsme napozorovali mnoho různých hodnot (maximálních hodnot pak, pokud jsme pozorovali m různých hodnot a četnosti jsou pro jednotlivé kategorie stejné). Naopak nulové hodnoty nabývá entropie v případě, že $n_1 = n$, tj. všechna pozorování jsou stejná, není mezi nimi žádná variabilita.

Míry šikmosti a špičatosti

Šikmost

Pokud pozorovaná data znormalizujeme tj. provedeme jejich transformaci tak, že mají nulovou střední hodnotu a rozptyl rovný jedné, pak je lze využít k výpočtu třetího a čtvrtého centrálního momentu. Ty se nazývají šikmosti a špičatostí.

$$\mu_3 = \frac{1}{n} \sum_{i=1}^{n} z_i^3 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^3 \tag{24}$$

Šikmost vyjadřuje symetričnost sledovaného rozdělení kolem průměrné hodnoty. Je-li pozorováno více malých hodnot v porovnání s vysokými hodnotami, pak je šikmost kladná. Je-li naopak převaha vysokých hodnot v porovnání s malými hodnotami, tj. po znázornění histogramu má rozdělení souboru protáhlý levý konec, je šikmost záporná.

Špičatost

Jde o čtvrtý centrální moment. Tato statistika představuje relativní strmost či plochost rozdělení četností v porovnání s normálním rozdělením četností.

Kladná špičatost znamená, že se ve sledovaném souboru vyskytují spíše data koncentrovaná kolem střední hodnoty.

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{n} z_i^4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4 \tag{25}$$

Šičatost je občas definována různě. Například MS Excel ji počítá následovně:

$$\left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n} z_i^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}\ . \tag{26}$$