Charakteristiky variability

Charakteristiky variability (proměnlivosti souboru)

Statistické znaky jako číselné proměnné jsou vždy různě variabilní (proměnlivé). Malý stupeň variability znamená malou vzájemnou různost (velkou podobnost) hodnot dané proměnné, což zároveň signalizuje, že střední hodnota (průměr), medián a případně i modus jsou v tomto případě dobrými charakteristikami obecné velikosti hodnot dané proměnné v daném souboru. Naopak vysoká variabilita značí velkou vzájemnou odlišnost hodnot dané proměnné, což zároveň signalizuje, že vypočítané parametry středu souboru nejsou v tomto případě dobrými charakteristikami obecné výše hodnot dané proměnné v daném souboru.

Charakteristiky středu souboru (střední hodnoty) udávají pouze informaci o poloze statistického souboru na číselné ose, ale neudávají, jak jsou hodnoty v souboru rozptýleny kolem středu, případně, zda existují v souboru tzv. extrémní hodnoty. Tuto informaci poskytují tzv. **míry variability** (charakteristiky variability), které vyjadřují rozmístění hodnot dané proměnné okolo střední hodnoty celého souboru.

1. Variační rozpětí (The Range)

(označení: R)

Variační rozpětí $\boldsymbol{R}$ řady n čísel můžeme definovat jako rozdíl mezi největší a nejmenší hodnotou řady (rozdíl mezi největší a nejmenší hodnotou znaku v souboru):

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

Variační rozpětí není příliš přesnou charakteristikou variability hodnot sledované numerické proměnné, neboť je ovlivněno velikostí extrémních hodnot a zároveň neříká nic o tom, jak se chovají hodnoty uvnitř souboru.

Tento nedostatek R překonávají rozpětí kvantilů, z nichž nejpoužívanější je *kvartilové rozpětí* R_q:

$$\boldsymbol{R_q = x_{0,75} - x_{0,25}}$$

Je zřejmé, že variační rozpětí ani kvantilová rozpětí neberou při charakterizování variability v úvahu velikost všech hodnot sledované numerické proměnné, což je mnohdy pociťováno jako závažný nedostatek.

2. Rozptyl (variance, The variance)

s^2 (základní soubor), s^2 (výběrový soubor)

Rozptyl můžeme definovat (jako přesný parametr populace) jako aritmetický průměr čtverců odchylek jednotlivých hodnot sledované proměnné x_i od průměru celého souboru:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$ (rozptyl **základního souboru**)

Pohlíží-li se na daný soubor jako na výběrový, potom mluvíme o *výběrovém rozptylu* s^2, který slouží jako odhad skutečného rozptylu populace a jeho výpočet se poněkud liší. U výpočtu výběrového rozptylu je ve jmenovateli výraz (n-1), který označujeme jako *počet stupňů volnosti* výběrového souboru (blíže viz Předn.2: Odhady parametrů základního souboru). Použitím tohoto výrazu (n-1) místo prosté velikosti souboru n docílíme přesnějšího odhadu skutečné hodnoty populačního rozptylu, především při výpočtu na základě malých výběrových souborů:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}$$ (rozptyl **výběrového souboru**)

Z obou výpočtů je zřejmé, že rozdíl mezi rozptylem s^2 na jedné straně a výběrovým rozptylem s^2 na druhé straně je při velkém rozsahu souboru (n > 30) prakticky zanedbatelný.

Při praktických výpočtech podle výše uvedeného vzorce výběrového rozptylu by byl postup příliš zdlouhavý (především u velkých výběrových souborů), proto je možno pro usnadnění použít ještě jinou variantu vzorce pro výpočet výběrového rozptylu:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

Vlastnosti rozptylu:

- jestliže jsou všechny hodnoty souboru stejné, potom je variabilita hodnot sledované proměnné v souboru nulová a výběrový rozptyl $s^2 = 0$

- velikost rozptylu se zvyšuje při zvětšující se variabilitě hodnot sledované proměnné

- rozptyl je odvozen od součtu čtverců odchylek jednotlivých hodnot od průměru souboru, proto nemůže nikdy nebývat záporných hodnot

- přičte-li se ke všem hodnotám (odečte-li se od všech hodnot) proměnné X libovolná kladná konstanta a, potom se rozptyl nezmění:

$$\sigma^2(X \pm a) = \frac{1}{n}\sum_{i=1}^{n}(x_i \pm a - x \pm a)^2 = \sigma^2$$

- násobí-li se (dělí-li se) všechny hodnoty proměnné nenulovou konstantou g, potom je rozptyl znásoben (vydělen) čtvercem této konstanty:

$$\sigma^2(g.X) = \frac{1}{n}\sum_{i=1}^{n}(g.x_i - g.X)^2 = g^2.\sigma^2$$

- rozptyl je uveden ve čtvercích měrných jednotek hodnot sledovaných číselných proměnných (vyplývá to z definice rozptylu). Např.: jestliže budou hodnoty sledované proměnné vyjádřeny v gramech, jejich rozptyl bude v g^2. Jestliže hodnoty sledované proměnné budou vyjádřeny v cm^2, rozptyl těchto hodnot bude vyjádřen v $(cm^2)^2$, bez ohledu na to, že takové jednotky nemají žádný fyzikální význam.

3. Směrodatná odchylka (standardní deviace, Standard Deviation - SD)

σ (základní soubor), s (výběrový soubor)

Směrodatná odchylka je definována jako (kladná) druhá odmocnina z rozptylu, tj.

$\sigma = \sqrt{\sigma^2}$ pro základní soubor,

případně

$s = \sqrt{s^2}$ pro výběrový soubor.

Výpočet **směrodatné odchylky pro základní soubor**:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad \sigma = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}}{N}}$$

nebo

Výpočet **směrodatné odchylky pro výběrový soubor**:

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

nebo

$$s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$$

Vlastnosti směrodatné odchylky:

- směrodatná odchylka má stejné měrné jednotky jako sledovaná číselná proměnná ve statistickém souboru

- směrodatná odchylka může nabývat vždy pouze kladných hodnot (vyplývá z definice)

4. Variační koeficient (The Coefficient of Variation)

(„relativní směrodatná odchylka“)

Variační koeficient je vhodný pro vzájemné srovnávání variability dvou nebo více souborů s podstatně odlišnou úrovní hodnot (např. variabilitu váhy kuřat v gramech a variabilitu váhy skotu v kg nebo metrických centech). V těchto případech musíme odstranit vliv obecné úrovně daných hodnot. Děláme to tak, že směrodatnou odchylku dělíme střední hodnou, od které byly počítány odchylky pro součet čtverců, obvykle tedy při praktických výpočtech aritmetickým průměrem výběrového souboru. Výsledek se obyčejně vyjadřuje v procentech (po vynásobení 100).

Variační koeficient je tedy definován pro **základní soubor**:

$$V = \frac{\sigma * 100}{\mu} \quad [\%]$$

Pro **výběrový soubor** vypočteme prakticky variační koeficient podle vzorce:

$$V = \frac{s * 100}{\bar{x}} \quad [\%]$$

Vlastnosti variačního koeficientu:

- variační koeficient je relativní mírou variability a není ovlivněn absolutními hodnotami sledovaného statistického znaku

- variační koeficient udává, z kolika procent se podílí směrodatná odchylka na aritmetickém průměru

- přičte-li se ke všem hodnotám (odečte-li se od všech hodnot) dané proměnné libovolná kladná konstanta *a*, potom se variační koeficient zmenší (zvětší:

$$V(X+a) = \frac{\sigma}{\bar{x}+a} < \frac{\sigma}{\bar{x}} = V$$

resp.:

$$V(X-a) = \frac{\sigma}{\bar{x}-a} > \frac{\sigma}{\bar{x}} = V$$

- násobí-li (dělí-li) se všechny hodnoty proměnné *X* nenulovou konstantou *g*, potom se variační koeficient nezmění:

$$V(g.X) = \frac{g.\sigma}{g.x} = \frac{\sigma}{x} = V$$

5. Střední chyba průměru (Standard Error of Mean - SE, SEM)

$\sigma_{\bar{x}}$ (základní soubor), $s_{\bar{x}}$ (výběrový soubor)

Střední (směrodatná) chyba průměru patří mezi často používané relativní míry variability. Často je označována (především v zahraniční odborné literatuře) zkratkou SE (příp. SEM) – z anglického výrazu Standard Error of Mean. Střední chyba průměru neměří rozptýlenost původní náhodné proměnné, ale rozptýlenost vypočítaného aritmetického průměru v různých výběrových souborech vybraných z jednoho základního souboru.

Střední (směrodatná) chyba průměru je teoreticky definována jako směrodatná odchylka všech možných výběrových průměrů z jedné populace, vypočítaných pro výběry o rozsahu *n* členů. Vyjadřuje tedy kolísání výběrových průměrů kolem teoretické (skutečné) střední hodnoty *m* v celém základním souboru.

Střední chyba průměru závisí jednak na rozptylu základního souboru (s^2), jednak na rozsahu výběrového souboru (n):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Protože v praxi obvykle neznáme skutečnou hodnotu rozptylu s celého základního souboru, používáme prakticky výpočet pro ***výběrovou střední chybu průměru*** podle vzorce:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Výběrová střední chyba průměru ($s_{\bar{x}}$) může být použita jako míra přesnosti, s jakou výběrový aritmetický průměr $\bar{x}$ odhaduje skutečnou střední hodnotu *m*. Prakticky se používá pro výpočet intervalů spolehlivosti aritmetického průměru u výběrových souborů (blíže viz Odhady parametrů základního souboru).

Zpět