NEVYCHÝLENOST VÝBĚROVÉHO PRŮMĚRU

*Josef Bukač**

Na základní škole

O průměru jsme se učili všichni již na základní škole. Děti si počítají průměrné známky již v tom okamžiku, kdy se naučili dělit, a od té doby mu věří. Pravým důvodem k používání průměru jsou ale jeho dobré vlastnosti.

Nejdůležitější je vlastnost zvaná nevychýlenost výběrového průměru, která nás opravňuje užívat jej jako odhad průměru v populaci, když k dispozici je jen výběr.

Průměr jako nevychýlený odhad

Snadno si představíme případ, kdy se populace skládá z velkého počtu prvků a nelze změřit hodnoty všech prvků populace, abychom stanovili jejich průměr, neboli populační průměr. Statistika doporučuje, abychom provedli náhodný výběr. Z tohoto výběru pak stanovíme výběrový průměr a pomocí tohoto výběrového průměru odhadneme populační průměr. Říká se, že výběrový průměr je odhadem populačního průměru. Přesný populační průměr nikdy nezjistíme, to je pro nás jakási teoretická hodnota, kterou se snažíme odhadnout pomocí výběrového průměru, protože víme, že tento výběrový průměr bude přibližně roven populačnímu. Populační průměr přesně neznáme, máme jeho odhad.

Co to ale znamená a proč si máme myslet, že výběrový průměr bude přibližně roven populačnímu? Není to jen proto, že nám nic jiného nezbývá. Chceme také vědět v jakém smyslu přibližně a co nás k tomu opravňuje.

Začít se dá příkladem. Populace se skládá ze 4 prvků nabývajících hodnot 1, 3, 5, 6. Populační průměr je $\mu=(1+3+5+6)/4=3{,}75$.

Provádíme výběry bez vracení, neboli bez opakování, aby počítání bylo co nejjednodušší a počet výběrů malý. Jako velikost výběru volíme 3.

Výběr	Průměr
1, 3, 5	(1+3+5)/3=3
1, 3, 6	(1+3+6)/3=3,333
1, 5, 6	(1+5+6)/3=4
3, 5, 6	(3+5+6)/3=4,667

Ke každému výběru byl výběrový průměr zapsán do druhého sloupce, čili pro celkem čtyři výběry máme čtyři výběrové průměry. Vypočítáme nyní průměr těchto průměrů

$$(3+3{,}333+4+4{,}667)/4=3{,}75.$$

Je důležité si všimnout, že je přesně roven populačnímu průměru. Tento jev je to, co nás opravňuje odhadovat populační průměr pomocí výběrového průměru.

V našem příkladě se průměr z průměrů počítá jako

$$\frac{1}{4}\left(\frac{1+3+5}{3}+\frac{1+3+6}{3}+\frac{1+5+6}{3}+\frac{3+5+6}{3}\right)=\frac{1+3+5+1+3+6+1+5+6+3+5+6}{3.4}.$$

Důležité také je, že se číslo 1 opakuje třikrát a další čísla také třikrát a součet v závorce je

$$\frac{1.3+3.3+5.3+6.3}{3.4}=3.\frac{1+3+5+6}{3.4}=\frac{1+3+5+6}{4}=3{,}75$$

což je populační průměr, jen počítaný z výběrů trochu jinak.

---

* Ústav biofyziky, Lékařská fakulta, Šimkova 870, 50000 Hradec Králové

Tato práce byla podporována grantem FRVŠ 1277/99.

Bylo ukázáno, co můžeme očekávat obecně. Pokusíme se důkaz provést pro libovolnou velikost populace $N_P > 1$ a libovolnou velikost výběru N, kde $0 < N < N_P$. Prvky populace označme $y_1, y_2, \ldots, y_{Np}$. Při výběru bez vracení je celkový počet výběrů N_V roven

$$N_V = \binom{N_P}{N}.$$

Jak je vidět na příkladě, každý prvek populace se vyskytuje ve stejném počtu výběrů. Obecně můžeme uvažovat tak, že si zvolíme libovolný prvek populace, tím v populaci zbývá $N_P - 1$ prvků a do výběru je třeba přidat jen $N - 1$ prvků. Ty vybíráme opět bez opakování a počet způsobů jakými je to možné provést je roven

$$\binom{N_P - 1}{N - 1}.$$

Označme $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{N_V}$ výběrové průměry, které jsou vypočtené z jednotlivých výběrů. Jejich počet je N_V. K označení hodnot z výběrů použijeme dvojité indexování, x_{ij} značí j-tou hodnotu v i-tém výběru.

Výběry je možné si představit jako

$$\begin{array}{c} x_{11}, x_{12}, \ldots, x_{1N}, \\ x_{21}, x_{22}, \ldots, x_{2N}, \\ \ldots\ldots\ldots\ldots\ldots \\ X_{Nv1}, X_{Nv2}, \ldots, X_{NvN}. \end{array}$$

Je jich celkem N_V na prvním řádku je první výběr, na druhém řádku je druhý výběr a tak by to pokračovalo až do N_V-tého řádku.

Ke každému řádku si představíme výběrový průměr $\bar{x}_i$ a naším úkolem je nyní vypočítat průměr ze všech těchto výběrových průměrů, což je

$$\frac{1}{N_V}\sum_{i=1}^{N_V}\bar{x}_i = \frac{1}{N_V}\sum_{i=1}^{N_V}\frac{1}{N}\sum_{j=1}^{N}x_{ij}.$$

Je možné vytknout 1/N a dostaneme vzorec

$$\frac{1}{N_V}\sum_{i=1}^{N_V}\bar{x}_i = \frac{1}{N.N_V}\sum_{i=1}^{N_V}\sum_{j=1}^{N}x_{ij}.$$

Víme ale, že každý prvek y_K z populace se vyskytuje ve výběrech stejně-krát, je tedy možné psát tento výraz jako

$$\frac{1}{N.N_V}\sum_{K=1}^{N_P} y_K \binom{N_P - 1}{N - 1} = \binom{N_P - 1}{N - 1}\frac{1}{N.N_V}\sum_{K=1}^{N_P} y_K$$

Nyní je třeba ukázat, čemu se rovná

$$\binom{N_P - 1}{N - 1}\frac{1}{N.N_V} = \binom{N_P - 1}{N - 1}\frac{1}{N\binom{N_P}{N}} = \frac{(N_P - 1)!N!(N_P - N)!}{N(N-1)!(N_P - 1 - N + 1)!N_P!} = \frac{(N_P - 1)!}{N_P!} = \frac{1}{N_P}$$

Tím je důkaz proveden. Nepočítá ale s tím, že by mohla být některá čísla v populaci stejná. Pak by neplatilo, že se jich nachází ve výběrech stejný počet. Proto raději hovoříme o prvcích populace, u kterých je vše jasné. Je ale jen trochu nepřesné smazat rozdíl mezi prvkem populace a hodnotou na něm naměřenou.

Výběrové průměry dávají v průměru populační průměr. To je důležitý pojem a ve statistice se mu říká nevychýlenost odhadu. Na ilustračním příkladě si můžeme objasnit a prověřit i další pojmy bez znalosti počtu pravděpodobnosti. Zajímavé by jistě bylo obecně ukázat, že součet čtverců odchylek od průměru dělený N-1 je nevychýleným odhadem populačního rozptylu. To už se tak snadno nedaří.