StatSoft

Popisná statistika – míry variability

V **předchozím dílu** jsme si ukázali základní míry polohy. V tomto článečku si ukážeme, jak aplikovat charakteristiky variability, neb náhodně proměnlivá data nestačí popsat pouze střední hodnotou. Znalost středních hodnot nám dává užitečnou informaci o tom, kde jsou data „centrována" (průměr, medián), případně která data jsou nejčetnější (modus). Míra rozptýlenosti hodnot různých souborů se stejnou střední hodnotou se však může velmi lišit, a proto je důležité s popisem charakteristik polohy uvádět v rámci popisné statistiky také charakteristiky variability, které nám řeknou, jak moc naše charakteristiky polohy daný soubor vystihují.

Charakteristiky variability

Rozpětí - první jednoduchou charakteristikou variability, jíž si popíšeme, je variační rozpětí R, které definujeme jako rozdíl mezi maximální a minimální hodnotou řady, tedy

$$R = x_{max} - x_{min}.$$

Variační rozpětí je velice hrubou charakteristikou variability, protože neříká nic o proměnlivosti jednotlivých hodnot v souboru. Maximální a minimální hodnoty mohou být navíc zkresleny odlehlými pozorováními. Nicméně, jistě uznáte, že i jednoduchá informace o rozpětí dat, je přínosná.

Rozptyl - další charakteristikou variability je základní a nejpoužívanější statistika a tou je bezpochyby rozptyl. Následující vzorec popisuje výběrový rozptyl, kterým z dostupných dat odhadujeme hodnotu populačního rozptylu:

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

Směrodatná odchylka – výběrový rozptyl, který se počítá pomocí čtverců odchylek dat od průměru, nemá stejný rozměr jako původní data. Do měřítka původních dat nás vrací odmocnina z rozptylu – výběrová směrodatná odchylka se definuje jako:

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

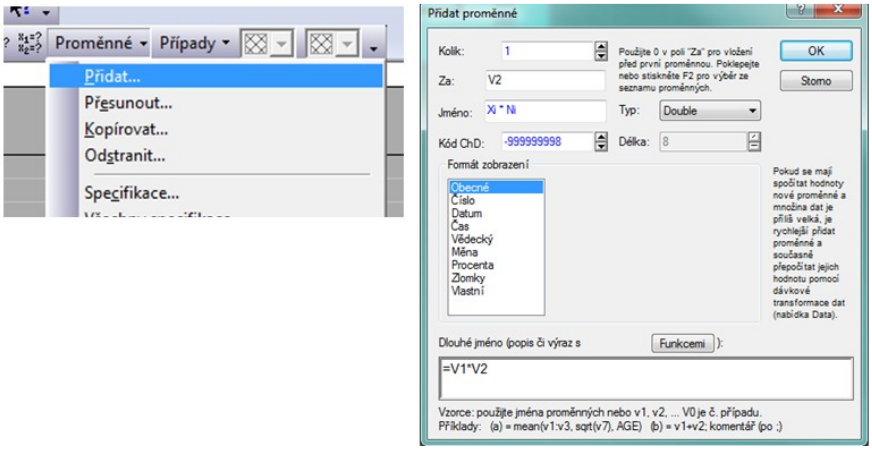
Výpočtem směrodatné odchylky změříme rozptýlenost kolem průměru. Je-li $s = 0$, soubor má nulovou variabilitu a všechna data jsou stejná.

Aplikace

Aplikaci si ukážeme na následujícím příkladu. Tabulka ukazuje *ha* výnos dvou plodin a plochu, na které byl výnos dosažen. Naším úkolem je vypočítat charakteristiky polohy a variability a zjistit kolísavost *ha* výnosu u obou plodin. Vzhledem k závislosti na velikost osevní plochy je potřeba využít vážené charakteristiky.

	1 Ječmen	2 plocha (Ha)	3 Brambory	4 plocha (Ha)
1	3,35	600	18	25
2	3,65	1000	19	255
3	4,32	1100	22	450
4	4,75	700	24	200
5	5,3	600	24,5	250
6	5,55	750	25	300

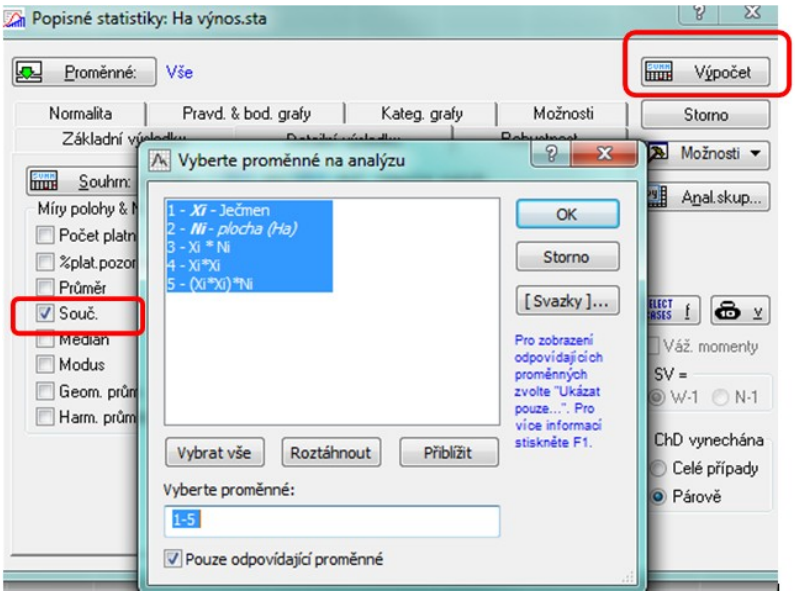
K výsledku se nejprve dostaneme zkratkou, **je to sice dál, ale zato horší cesta**. Tento postup nám však ilustruje výpočet bez použití funkcionality *Váhy* v softwaru *STATISTICA* a také ukazuje možnosti záložky ***Data.*** Soustřeďme se nyní pouze na plodinu Ječmen, do otevřené tabulky postupně přidáme 3 nové proměnné. V softwaru *STATISTICA* přes tlačítko ***Proměnné*** a v dialogu ***Přidat proměnné*** napíšeme příslušné vzorce, které později využijeme pro dosazení do vzorce pro rozptyl.



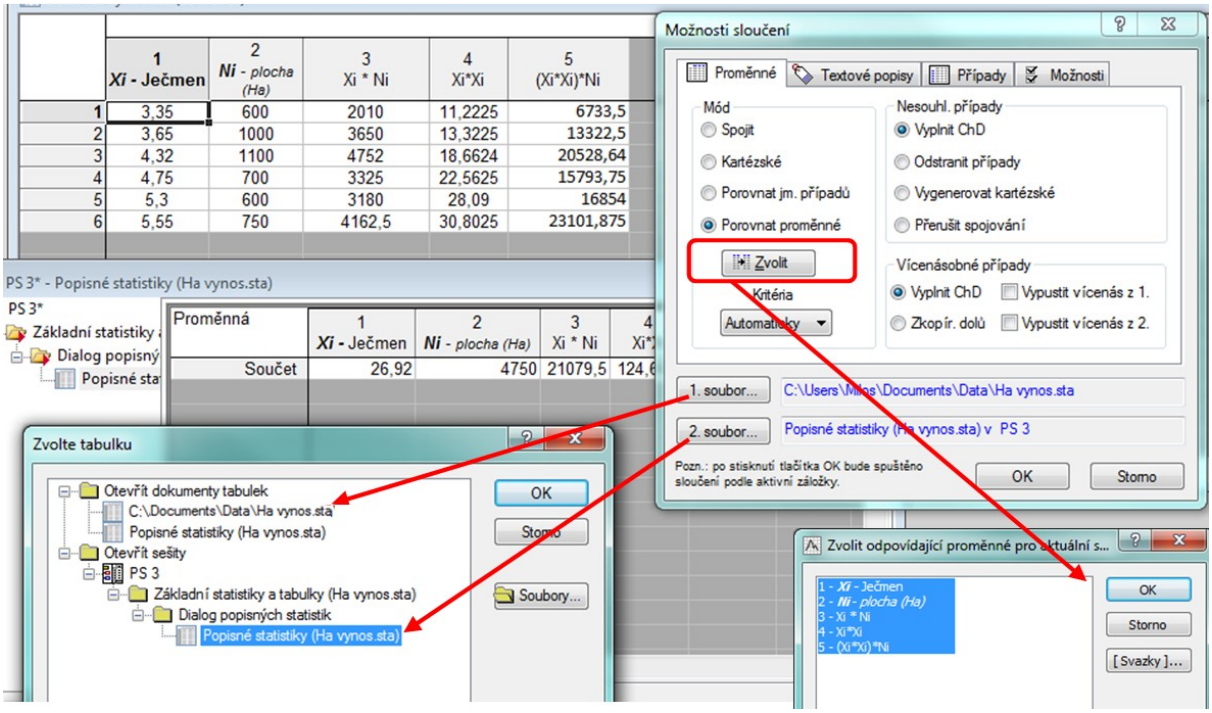
Výsledná tabulka má tuto podobu:

	1 ***Xi*** **-** Ječmen	2 ***Ni*** *- plocha (Ha)*	3 Xi * Ni	4 Xi*Xi	5 (Xi*Xi)*Ni
1	3,35	600	2010	11,2225	6733,5
2	3,65	1000	3650	13,3225	13322,5
3	4,32	1100	4752	18,6624	20528,64
4	4,75	700	3325	22,5625	15793,75
5	5,3	600	3180	28,09	16854
6	5,55	750	4162,5	30,8025	23101,875

Přes záložku *Statistiky -> Základní statistiky/tabulky -> Popisné statistiky* spočteme sumu všech proměnných (lze i přes *Statistiky -> Statistiky bloku dat...*):



Pokud bychom chtěli mít tyto nové výsledky v jediné tabulce, pak to můžeme provést například následujícím způsobem: Výslednou tabulku transponujeme přes záložku *Data -> Transponovat (Soubor)* a sloučíme s předcházející tabulkou přes záložku *Data -> Sloučit (Porovnat proměnné)*:



Výstupní tabulka má tento tvar:

Data: Tabulka11* (5s krát 7ř)

Proměnná	Popisné statistiky (Ha vynos.sta)				
	1 *Xi* - Ječmen	2 *Ni* - plocha (Ha)	3 Xi * Ni	4 Xi*Xi	5 (Xi*Xi)*Ni
1	3,35	600	2010	11,2225	6733,5
2	3,65	1000	3650	13,3225	13322,5
3	4,32	1100	4752	18,6624	20528,64
4	4,75	700	3325	22,5625	15793,75
5	5,3	600	3180	28,09	16854
6	5,55	750	4162,5	30,8025	23101,875
Součet	***26,92***	***4750***	***21079,5***	***124,6624***	***96334,265***

Tento postup jsme si ukazovali hlavně proto, abychom nastínili široké možnosti **ovládání softwaru *STATISTICA***, kterým bude věnováno některé z příštích čísel. Výslednou tabulku však využijeme na dosazení do vzorců pro výpočet charakteristiky polohy, resp. variability.

Máme k dispozici ha výnos dané plodiny a velikost osevní plochy pro každý výnos, kterou je třeba ve výpočtu zohlednit, použijeme proto vážený průměr, který jsme si ukázali v minulém dílu:

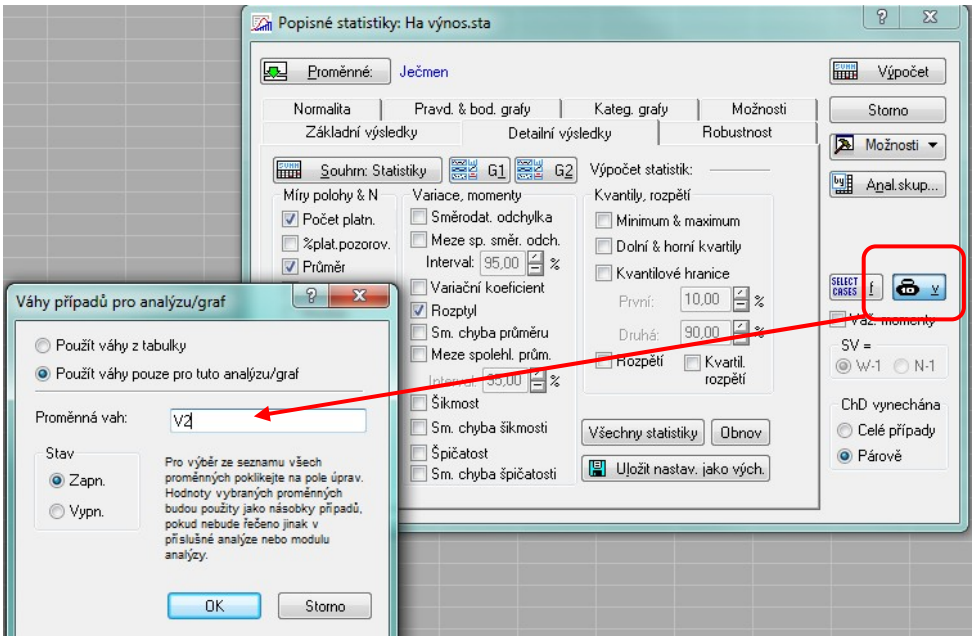
$$\bar{x}_v = \frac{\sum_{i=1}^{n} x_i n_i}{\sum_{i=1}^{n} n_i} = \frac{21079{,}5}{4750} = 4{,}4377$$

a vážený rozptyl:

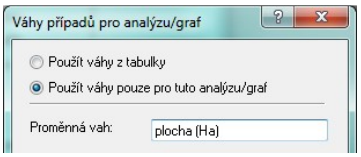
$$s_v^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_v)^2 \cdot n_i}{\sum_{i=1}^{n} n_i - 1} = \frac{\sum_{i=1}^{n} x_i^2 \cdot n_i - \bar{x}_v \cdot \sum_{i=1}^{n} x_i \cdot n_i}{\sum_{i=1}^{n} n_i - 1} = \frac{96334{,}265 - 4{,}4377 \cdot 21079{,}5}{4750 - 1} = 0{,}587$$

Po odmocnění odhadu rozptylu získáme směrodatnou odchylku s_v.

Celý výše uvedený postup v softwaru *STATISTICA* řeší jednoduše několika kliknutími funkcionalita *Váhy případů*.



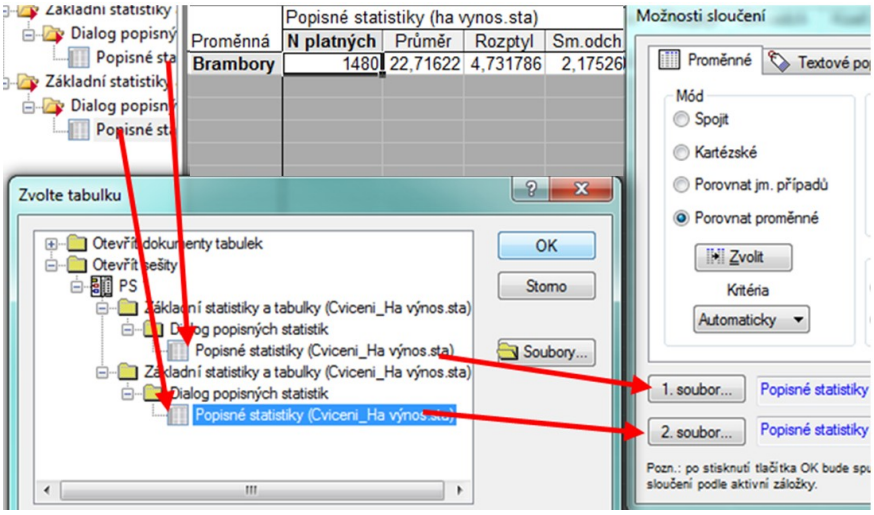
Do okna ***Proměnné vah*** vepíšeme číslo proměnné (zde V2), nebo její celý název „plocha (Ha)", resp. po poklikání do okna můžeme proměnnou vybrat ze seznamu, který se Vám zobrazí v dialogu ***Zvolit proměnnou.***



A klikneme na ***Výpočet*** popisných statistik:

Proměnná	Popisné statistiky (Ha výnos.sta)					
	N platných	Průměr	Minimum	Maximum	Rozptyl	Sm.odch.
Ječmen	4750	4,437789	3,350000	5,550000	0,587046	0,766189

Stejným způsobem vypočteme také charakteristiky variability pro druhou proměnnou. Pokud by byly váhy pro obě proměnné stejné, výslednou tabulku bychom získali v jednom kroku výběrem obou proměnných najednou. Protože jsou ale váhy odlišné, potřebujeme pro sloučení výsledků funkcionalitu ***Data -> Sloučit ->*** *Tlačítko* ***Porovnat proměnné***



sloučíme obě výsledné tabulky v sešitu *STATISTICA* a dostaneme výslednou tabulku:

Proměnná	Popisné statistiky (ha vynos.sta)			
	1 N platných	2 Průměr	3 Rozptyl	4 Sm.odch.
Brambory	1480	22,716216	4,731786	2,175267
Ječmen	4750	4,437789	0,587046	0,766189

Variační koeficient

- další mírou variability, kterou lze v softwaru *STATISTICA* vypočítat, je variační koeficient. Jde o poměr výběrové směrodatné odchylky a průměru, který slouží pro posouzení relativní míry rozptýlenosti dat vzhledem k průměru. Použijeme ho tehdy, pokud budeme porovnávat variabilitu dat jednoho parametru měřeného v různých dávkách

$$\mathrm{Vk} = \frac{s}{\bar{x}} \cdot 100\ (\%).$$

Při použití variačního koeficientu je potřeba ale dávat pozor na to, jaká máme data. Jeho použití není univerzální! Například použití na datech se zápornými hodnotami může dávat zavádějící výsledky. Více informací najdete například na **zde** (případně na **wikipedii**).

Mezikvartilové rozpětí (*Interquartile range IQR*) – poslední charakteristikou rozptýlenosti, kterou si představíme je mezikvartilové rozpětí. Vypočítáme ji jako rozdíl mezi horním kvartilem Q_{III} (75 % kvantil) a dolním kvartilem Q_I (25 % kvantil)

$$IRQ = Q_{III} - Q_I.$$

Ačkoli tuto statistiku uvádíme jako poslední, neznamená to, že by nebyla důležitá, právě naopak. Mezikvartilové rozpětí je nejpoužívanější neparametrickou mírou variability. Je totiž odolné vůči přítomnosti odlehlých hodnot v datech, což například nejznámější a nejpoužívanější rozptyl v žádném případě není. Pokud tedy máte podezření, že se Vám v datech vyskytují odlehlé hodnoty, je mezikvartilové rozpětí doporučenou volbou.

Všechny tyto i další charakteristiky naleznete na kartě *Detailní výsledky* v dialogu *Popisné statistiky: Statistiky —> Základní statistiky/tabulky —> Popisné statistiky*

StatSoft CR s. r. o.
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
t +420 233 325 006, **f** +420 233 324 005
e info@statsoft.cz