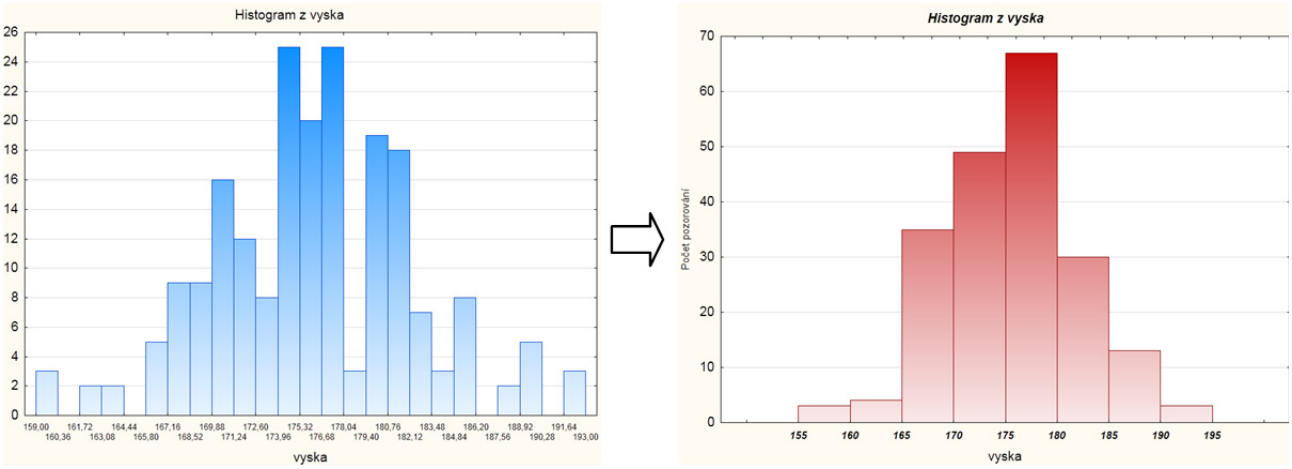
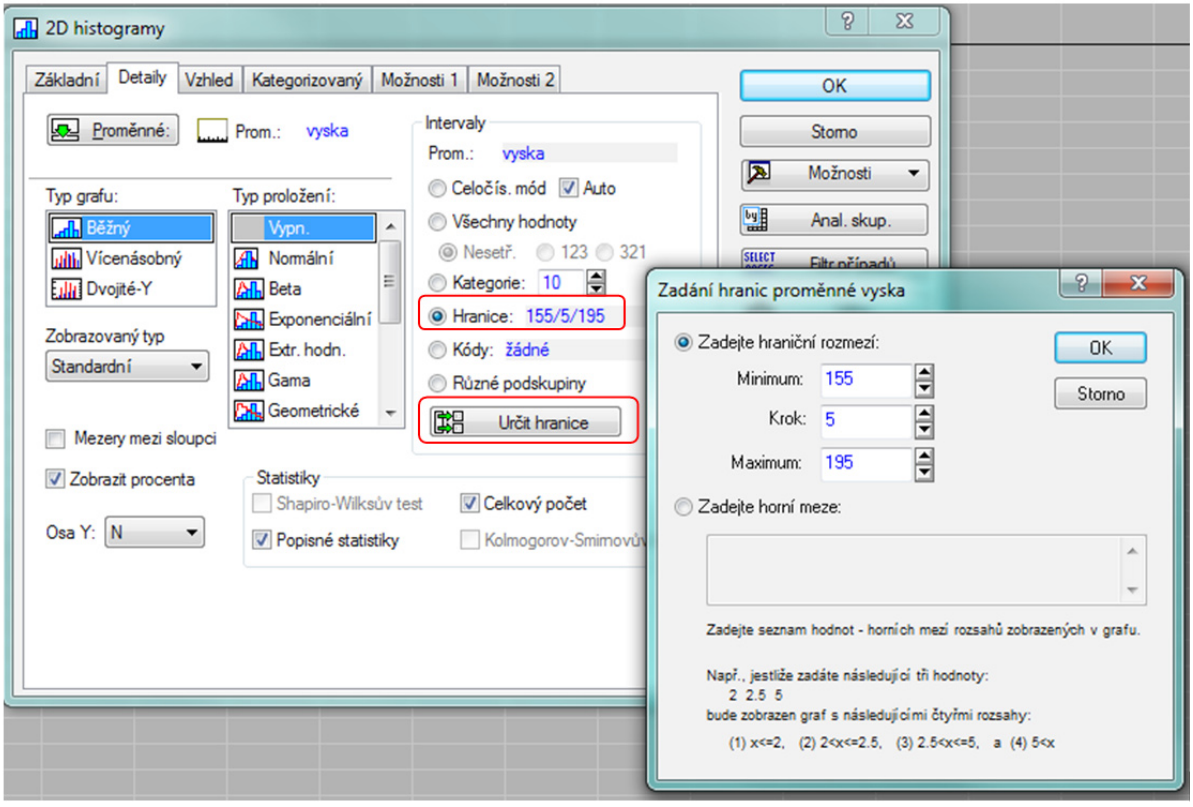
StatSoft

Popisná statistika – kvantitativní veličiny

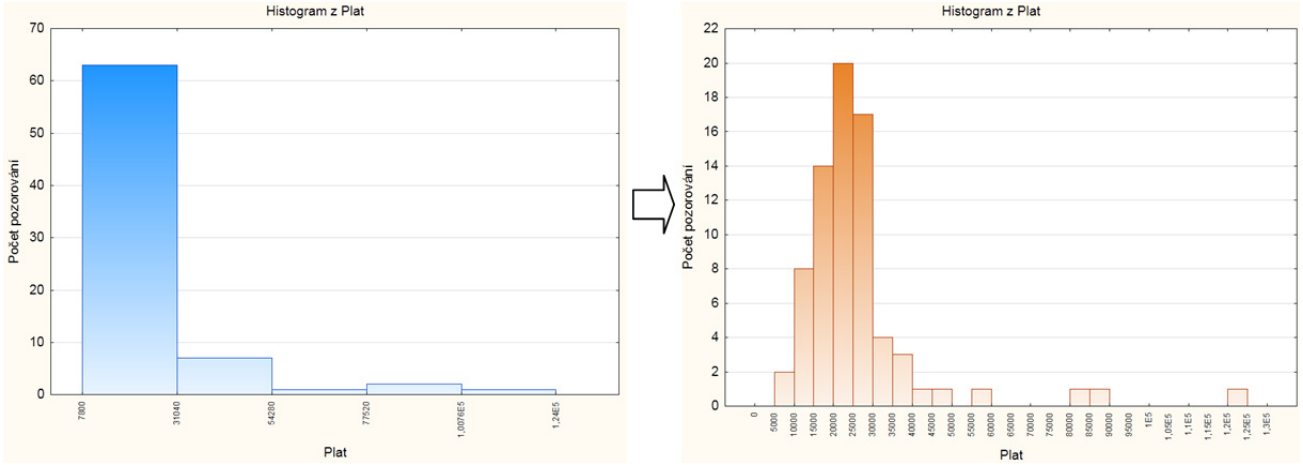
Protože nám surová data obvykle žádnou smysluplnou informaci neposkytnou, je žádoucí vyjádřit tyto ve zhuštěnější formě. V předchozím dílu jsme začali tabulkou četností, u dalšího příkladu si ukážeme, jak tabulku četností, resp. histogram využít v případě spojitých dat. Chceme-li tabulku četností aplikovat na spojitá data, bude nutné stanovit vhodný počet intervalů, na které rozdělíme variační rozpětí (maximální mínus minimální hodnota). K tomu slouží různá pravidla. Základním přibližným kritériem je $\sqrt{n}$, používanějším je pak například Sturgesovo pravidlo: $K \approx 1+3{,}3\log_{10} n$. Ukázku nevhodně zvoleného intervalu pro proměnnou „výška" na histogramu zachycuje následující obrázek (modrý graf):



V druhém případě (červený graf) jsou intervaly zvoleny expertně na základě potřeb zadavatele. Proměnná „výška" se pravděpodobně blíží normálnímu rozdělení, jehož identifikace bude probrána v některém z příštích dílů. Minimální výška v souboru je 159 a maximální je 193, zvolené nastavení hranic a kroku zachycuje následující obrázek: (***V softwaru STATISTICA: Grafy —> Histogramy —> karta Detaily***)



Velikost a krok intervalu (šířka sloupce) často vychází z potřeb konkrétního procesu a velikost kroku je dána samotným procesem, kdy znalec procesu sám určí, které intervaly pro něj mají praktický význam a které ne. Opačný případ oproti příkladu s „výškou" nastane, pokud jsou data zhuštěna příliš a dochází ke ztrátě informace, tento příklad ilustruje následující obrázek na proměnné „plat":

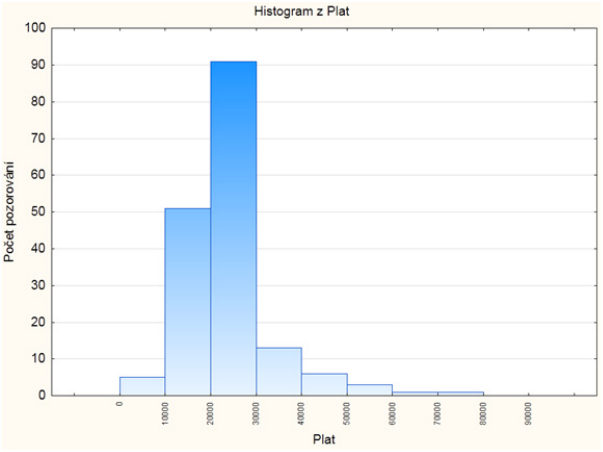


Při použití tabulky četností, jsou možnosti pro určení intervalů dostupné na kartě ***Detailní výsledky***. (*V softwaru STATISTICA*: ***Statistiky —> Základní statistiky/tabulky —> Tabulky četností***).

V konkrétním vzorku (opět proměnná „plat") převažují respondenti, kteří mají hrubý plat mezi 20–30 tisíci (54,3 %) a také předcházející skupina s platy mezi 10–20 tisíci je výraznější (28,6 % vzorku). Tyto informace můžeme jednoduše vyčíst z tabulek četností či histogramu (viz následující graf a obrázek).

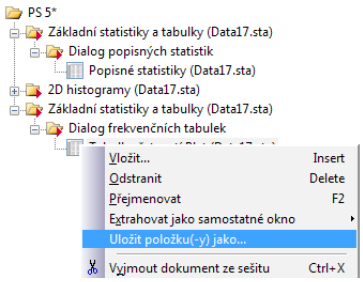
Tabulka četností:Plat (Data_hlavni_171.sta)

OD DO	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
0,000000<=x<10000,00	5	5	2,92398	2,9240
10000,00<=x<20000,00	49	54	28,65497	31,5789
20000,00<=x<30000,00	93	147	54,38596	85,9649
30000,00<=x<40000,00	13	160	7,60234	93,5673
40000,00<=x<50000,00	6	166	3,50877	97,0760
50000,00<=x<60000,00	3	169	1,75439	98,8304
60000,00<=x<70000,00	1	170	0,58480	99,4152
70000,00<=x<80000,00	1	171	0,58480	100,0000
80000,00<=x<90000,00	0	171	0,00000	100,0000
ChD	0	171	0,00000	100,0000

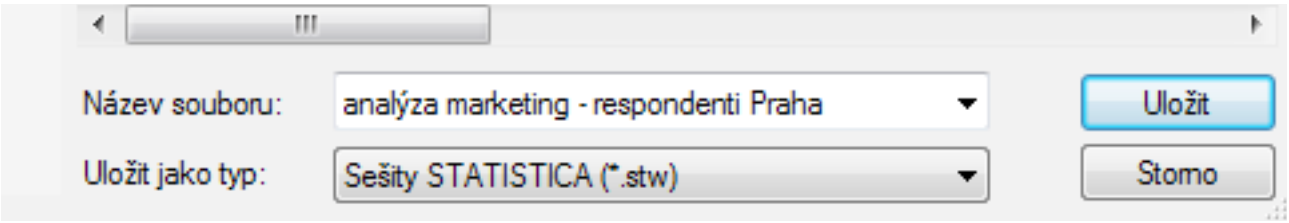


Všimněme si, že histogram i tabulka četností poskytují stejnou informaci o datovém souboru.

Obrázek, který následuje, ukazuje, jak výslednou tabulku četností uložit. V dialogu příslušné analýzy v stromové struktuře sešitu *STATISTICA* přes pravé tlačítko na myši vyvoláme roletku, ve které volíme „Uložit položku(-y) jako…", případně rovnou přes záložku ***Soubor —> Uložit položku(-y) jako…***



Volbou ***Soubor —> Uložit jako*** se uloží sešit *STATISTICA (*.stw)*, tedy celá stromová struktura se všemi výstupy analýz.



Dalším možnostem uložení bude věnován nějaký z dalších dílů našeho seriálu.

Dalším krokem pro zpřehlednění datového souboru je výpočet statistických charakteristik polohy a variability, pokud je daná proměnná pro tyto výpočty vhodná. Statistické charakteristiky shrnují informace obsažené v datech, jde o vyjádření v tzv. koncentrované formě. Díky těmto mírám lze také jednotlivé soubory porovnávat. Základními charakteristikami, které bychom měli u každého souboru vypočítat, jsou míry polohy a variability.

Rozlišujeme mezi dvěma typy charakteristik, a to sice mezi charakteristikami, které jsou vypočteny ze všech hodnot dané proměnné (průměry, rozptyl, atd.) a jsou tedy ovlivněny například odlehlými hodnotami, a charakteristiky, které nejsou počítány ze všech hodnot dané proměnné (modus, medián a další kvantily).

Charakteristiky polohy

Popisné statistiky shrnují různé charakteristiky vzorku, získáme tak odlišné úhly pohledu na vzorek. Charakteristiky polohy, resp. míry centrální tendence, by měly charakterizovat typickou hodnotu pro daný datový soubor. Slovo „měly" je použito záměrně, protože tyto míry jsou počítány z celého datového souboru a extrémní hodnoty, kterých může být velmi malé množství, potom posunou výsledek do jiné než typické polohy. Ukázkovým příkladem je výpočet průměrného platu v populaci.

Základní charakteristikou polohy je průměr, který rozlišujeme na:

Aritmetický průměr – základní míra polohy, kterou využijeme tam, kde má smysl sčítat, například součet odchylek měření, suma vzdáleností. Průměr bychom neměli používat pro data kategoriální a rozdělení hodnot by mělo být symetrické. Pokud máme několik průměrů z různých podmnožin a známe velikost těchto podmnožin, lze vypočítat celkový průměr ze všech hodnot všech podmnožin jako vážený průměr. Příkladem na vhodné použití aritmetického průměru může být třeba odhad průměrné výšky mužů/žen v populaci, průměrný počet dětí v určité oblasti atd.

Prostý: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

Vážený: $\bar{X} = \frac{\sum_{i=1}^{n} x_i n_i}{\sum_{i=1}^{k} n_i}$

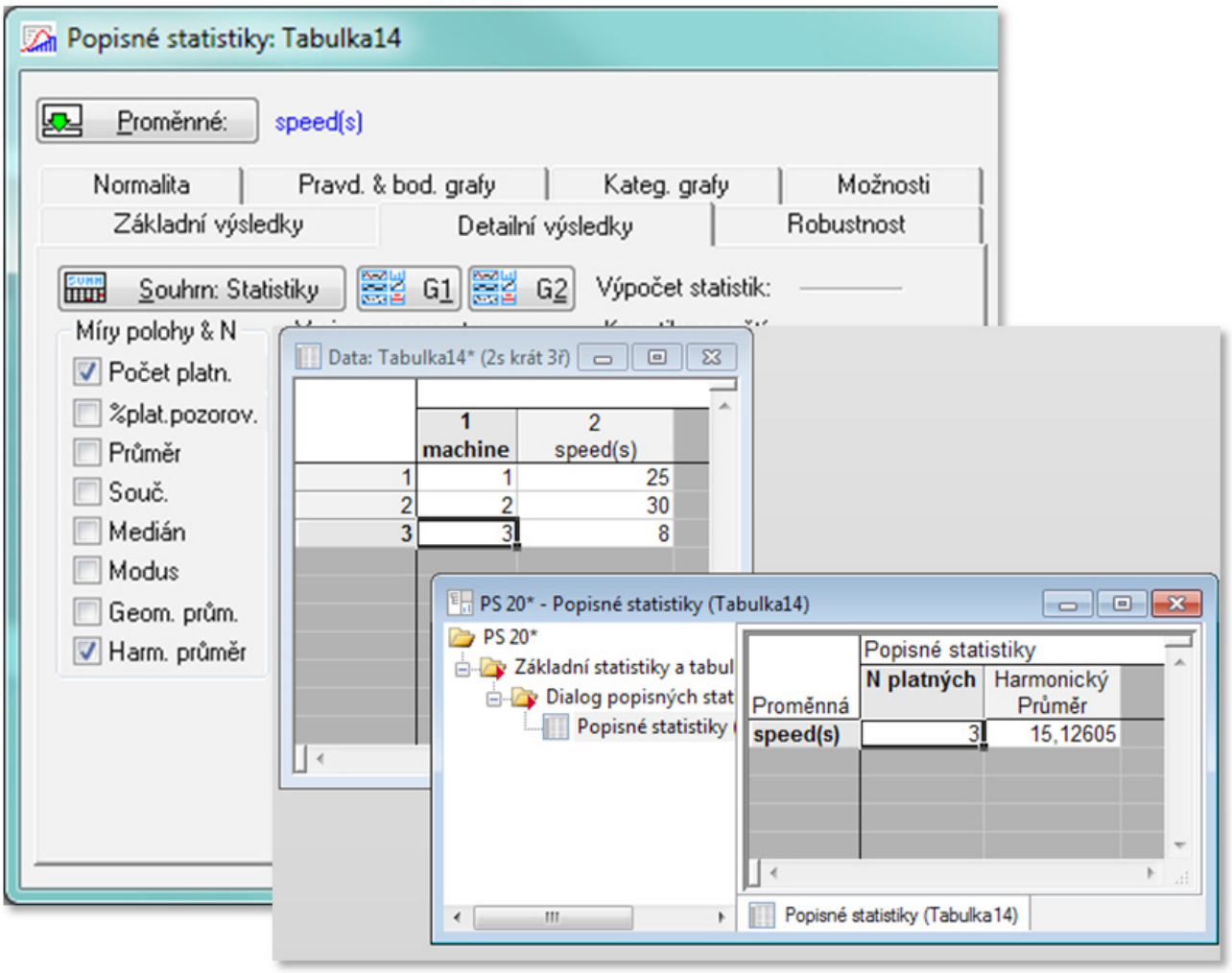
Harmonický průměr – harmonický průměr je rozsah souboru vydělený sumou převrácených hodnot znaků. Využijeme ho zejména v případě, kdy jsou znaky měřeny jako čas na jednotku výkonu, případně tam, kde počítáme průměr z poměrných čísel.

Prostý: $\bar{X}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

Vážený: $\bar{X}_h = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Aplikaci harmonického průměru ukazuje následující obrázek: linka má 3 výrobní stroje, kdy každý zpracuje polotovar za odlišnou dobu v sekundách. Harmonickým průměrem jsme vypočetli celkovou průměrnou rychlost výroby:

$$\bar{X}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{3}{\frac{1}{25}+\frac{1}{30}+\frac{1}{8}} = 15{,}12\ s$$



Geometrický průměr – je vhodný používat tam, kde má věcný význam součin znaků, což je především při analýze znaků, které jsou odvozeny z podílu dvou veličin a tvoří posloupnost. Používá se zejména k charakterizování průměrného tempa růstu.

Prostý: $\bar{X}_g = \sqrt[n]{x_1 x_2 \dots x_n}$

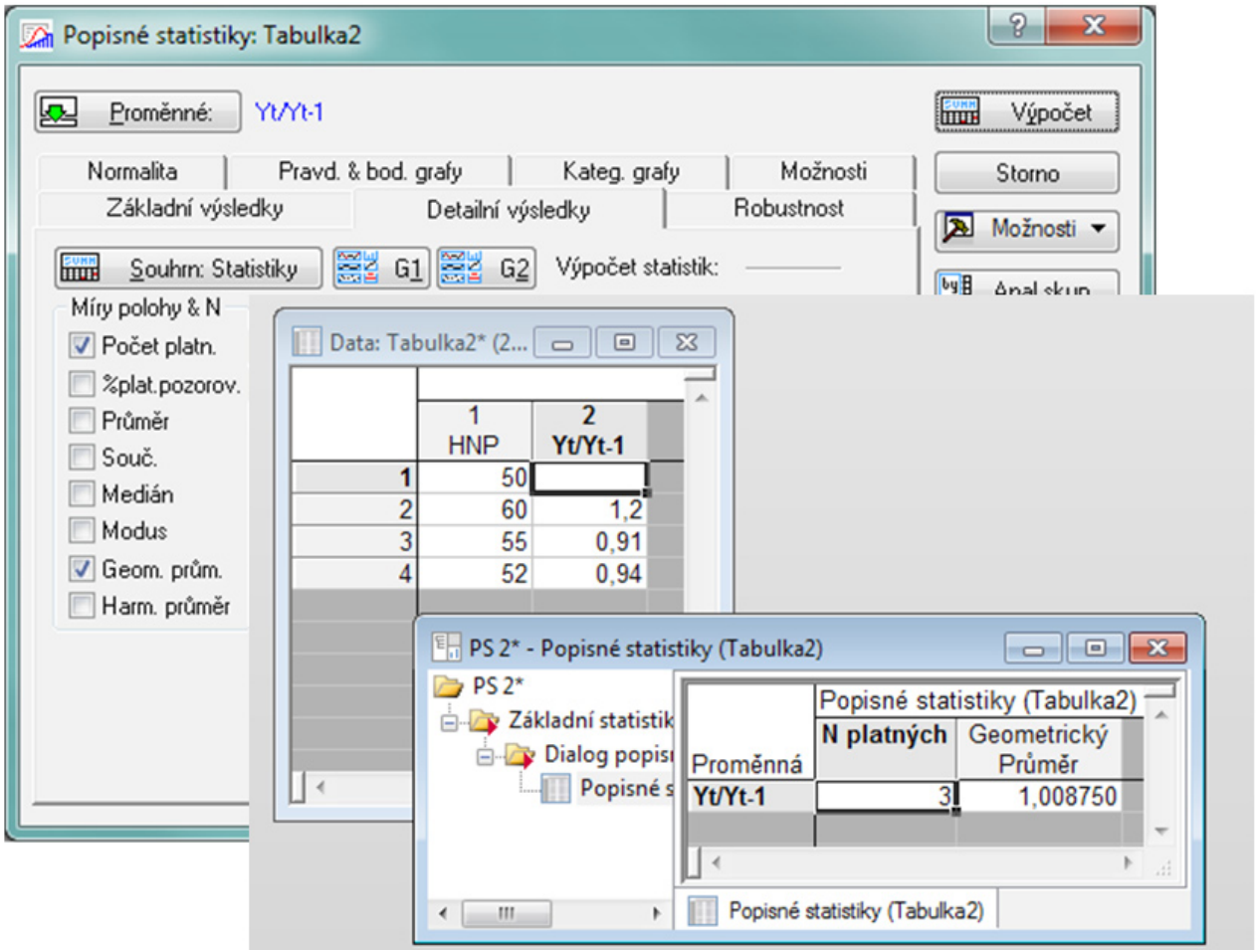
Vážený: $\bar{X}_g = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$

Aplikaci toho průměru ukazuje následující příklad, který zobrazuje vývoj HNP v 4 obdobích.

HNP	50	60	55	52
rok	1	2	3	4
Tempo růstu	1	(60/50)	(55/60)	(52/55)

Prostý: $\bar{X}_g = \sqrt[3]{1{,}2 \cdot 0{,}91 \cdot 0{,}94} = 1{,}0087$

Po odečtení čísla 1 od průměru získáme tzv. čistý růst, v tomto případě došlo k nárůstu o 0,87 %. Postup v softwaru ukazuje obrázek níže. Jako proměnou uvažujeme tzv. tempo růstu, tedy podíl aktuálního období a předchozího období daného znaku.



Medián – je další základní charakteristikou a běžně se značí jako $\tilde{x}$. Je definován jako prostřední hodnota v souboru setříděném podle velikosti. Medián dělí řadu podle velikosti uspořádaných hodnot na 2 shodně početné poloviny. V případě lichého čísla jde o prostřední hodnotu v setříděném souboru, v případě sudého souboru jde o průměr ze dvou prostředních hodnot, viz následující příklad:

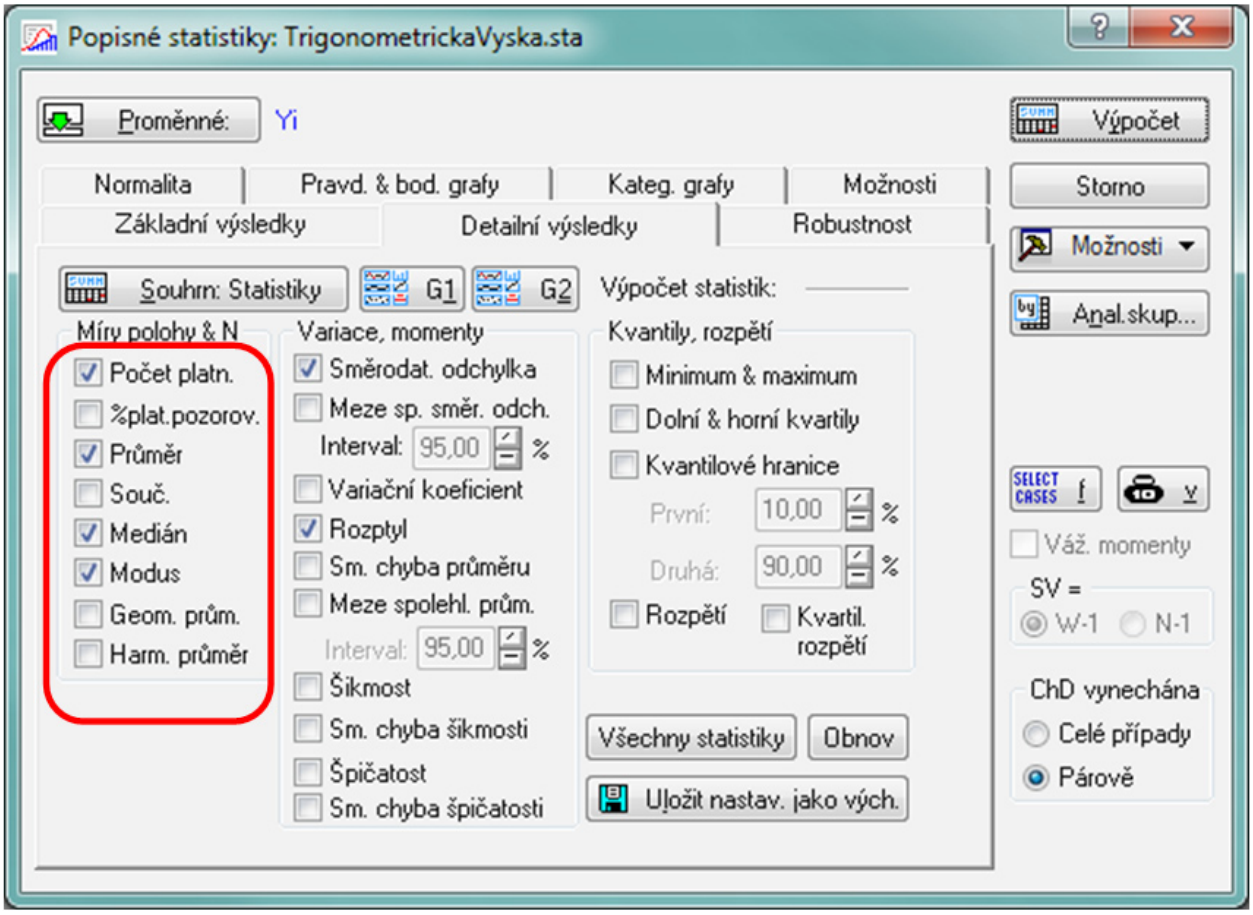
jméno	počet dětí
člověk 1	1
člověk 2	2
člověk 3	3
člověk 4	3

$$\tilde{x} = \frac{2+3}{2} = 2{,}5$$

V poslední době je v některých seriózních denících občas také uváděn mediánový plat, který se reálnému „obvyklému" platu přibližuje více než populistický, manažerskými platy zkreslený průměr. Medián použijeme u dat, která jsou min. v ordinálním měřítku, a aplikujeme ho tam, kde data obsahují odlehlé hodnoty nebo je rozložení dat zešikmené a průměr nebude určitě typickou hodnotou souboru. Dále pokud chceme znát střed daného rozdělení.

Modus – běžně značený stříškou $\hat{x}$, zachycuje nejčastější hodnotu v datech (nejčastější počet dětí, nečastější známka atd.). Modus má smysl počítat zejména u kategorických, resp. obecně u všech kvalitativních dat. Modus použijeme tam, kde je žádoucí znát nejčastější hodnotu znaku.

V softwaru *STATISTICA* získáme tyto i další charakteristiky na kartě Detailní výsledky v dialogu ***Popisné statistiky: Statistiky —> Základní statistiky/tabulky —> Popisné statistiky***



V příští části našeho seriálu budeme pokračovat s aplikací popisné statistiky na spojitá data.

StatSoft®
STATISTICA

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