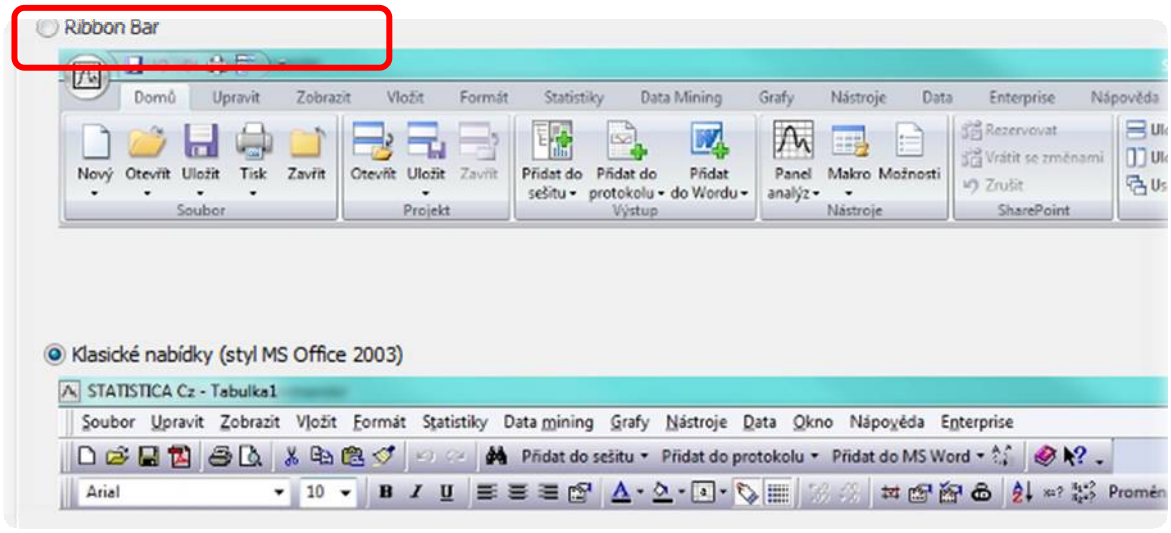
Ovládání a základy statistiky v softwaru STATISTICA

StatSoft CR s.r.o.
Ringhofferova 115/1
155 21 Praha 5 – Zličín
tel.: +420 233 325 006 • fax: +420 233 324 005 • e-mail: info@statsoft.cz • www.statsoft.cz

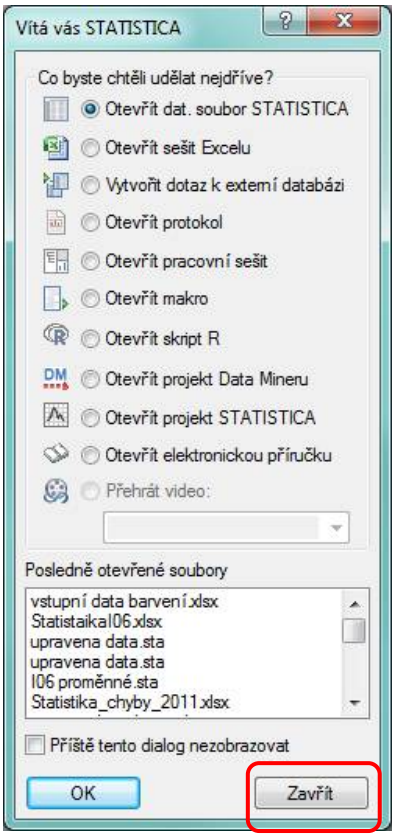
Obsah:

1 Spuštění programu *STATISTICA*

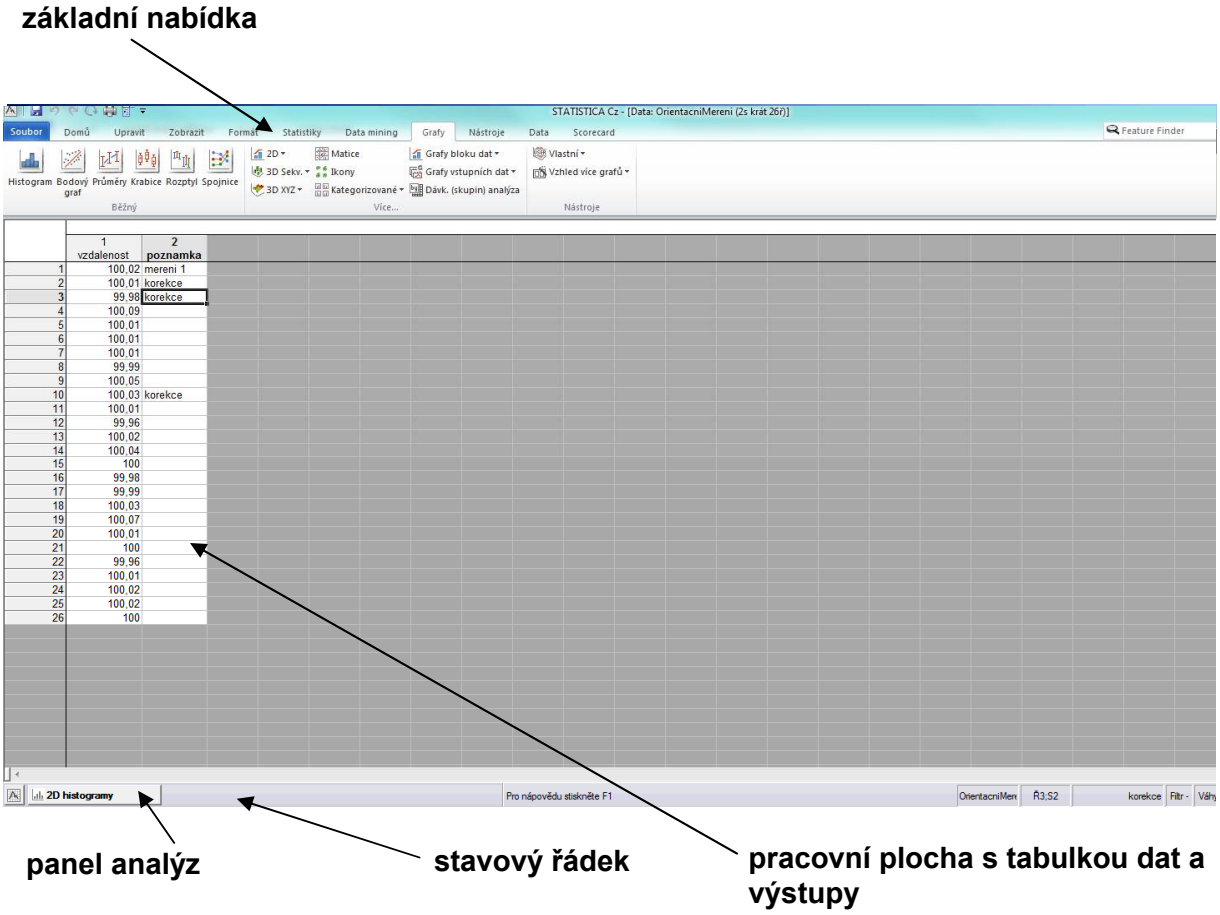
Při prvním spuštění nám dá program vybrat mezi 2 typy menu:



Vybereme Pás karet - po potvrzení ***OK*** se obrazovce se objeví rychlá navigace, kterou zavřeme a

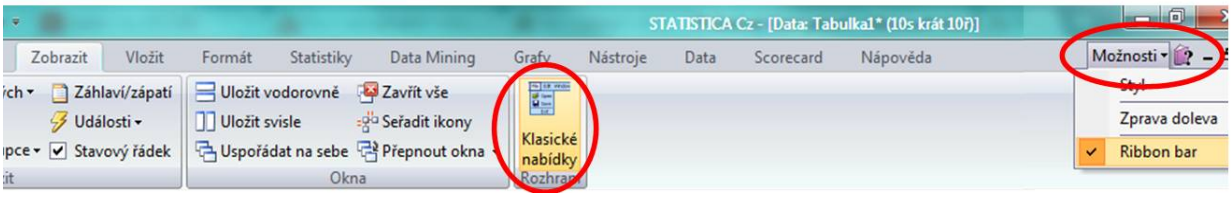


máme zde okno aplikace *STATISTICA*:



- *základní nabídka* - slouží k ovládání systému, zpřístupňuje všechny nástroje programu
- *panely nástrojů s tlačítky* - jednodušší přístup k různým příkazům
- *panel analýz* - zde minimalizována okna všech spuštěných analýz, mezi kterými se lze přepínat
- *stavový řádek* - podává zkrácenou nápovědu a základní informace o aktivním dokumentu. Můžeme odtud např. ovládat filtry či váhy.

Software *STATISTICA* umožňuje práci v zobrazení *Ribbon bar*, přepnutí do klasického zobrazení provedete přes záložku ***Možnosti*** v pravém horním rohu, nebo přes záložku ***Zobrazit.***



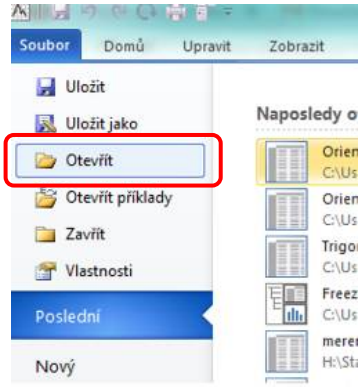
2 Načtení souboru

Data pro vlastní analýzu můžeme získat několika způsoby:

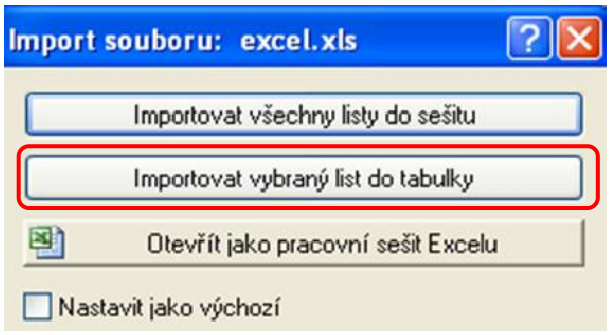
- *importem již uložených souborů různých formátů*
- *připojením k databázi* – pomocí SQL dotazů lze pracovat s daty uloženými například v databázi Oracle, MS SQL Server, Sybase atd.
- *otevřením tabulky Microsoft Excel* v programu *STATISTICA* bez importu
- *vložením dat do nové tabulky* v programu *STATISTICA*
- *sběrem dat on-line* - pokud je systém napojen na měřicí zařízení, naměřené hodnoty se dají ihned zpracovávat.

Příklad – import dat z Excelu

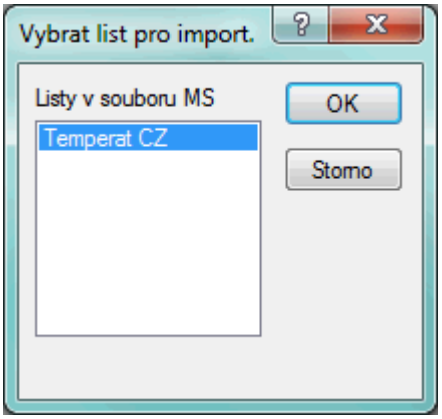
V menu ***Soubor*** a možnost ***Otevřít*** vybereme soubor **Temperat CZ.xls**



Při otevírání „Excelovských“ tabulek mámě několik možností, jak k tabulkám přistupovat:

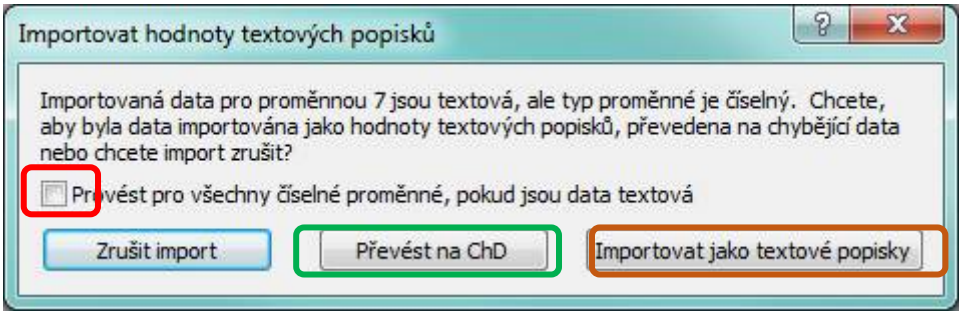


- ***Importovat vybraný list do tabulky*** – nejčastější možnost – pokud máme více listů, tak upřesníme list, který chceme importovat – **vybereme:**

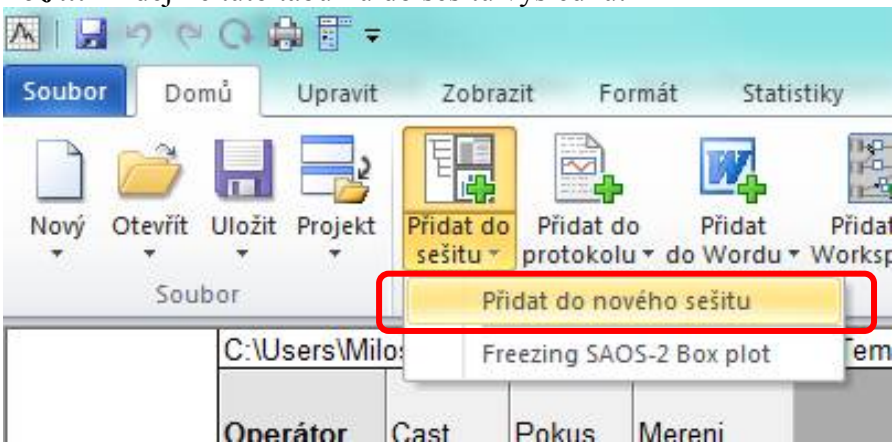


Data jsou načtena do tabulky softwaru *STATISTICA* (** **.sta***) stejně jako v případě načítání dat z textových souborů.

Pokud mám v původním souboru textové popisky, ale formát proměnné je číselný, tak mě *STATISTICA* upozorní, převedeme tyto textové popisky v číselné proměnné na chybějící hodnotu, nebo je naimportujeme jako textové a následně se v načteném souboru podíváme a smažeme je.



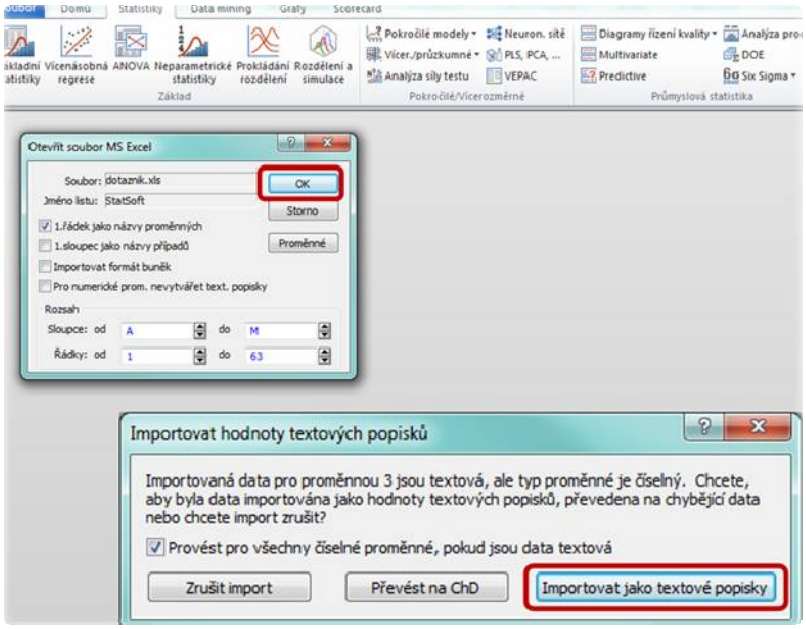
Pozn.: Přidejme tuto tabulku do sešitu výsledků:



3 Zpracování chybějících dat

Načtení souboru

Postup si představíme na kompletním příkladu, jak postupovat. Pro zopakování začneme samotným datovým souborem a jeho načtením. Máme excelovský soubor, do kterého byly ručně zadány výsledky dotazníkového šetření. Soubor obsahuje řadu chybějících hodnot a překlepů. Přes ***Soubor -> Otevřít*** načteme tento datový soubor:

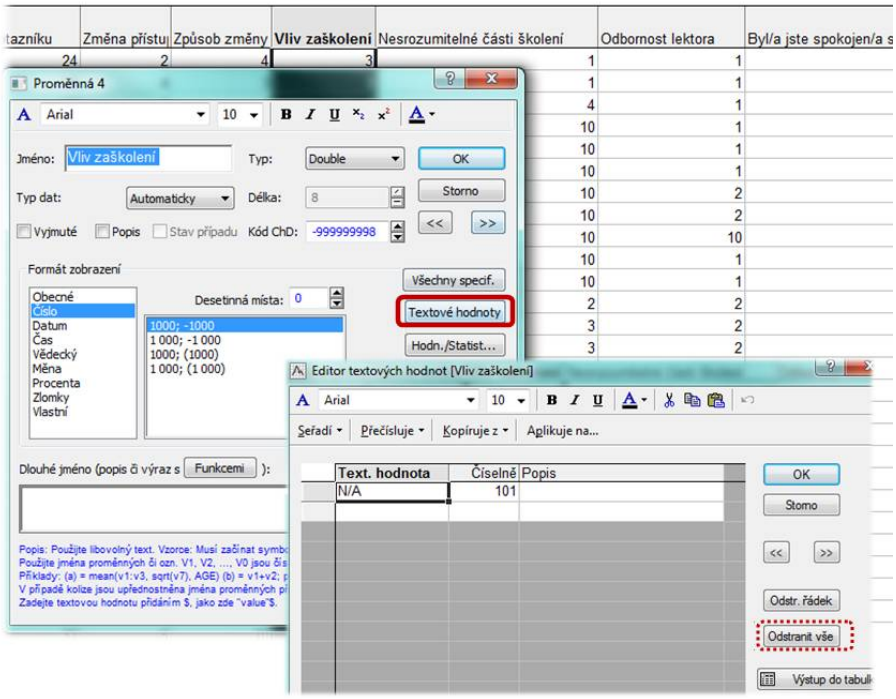


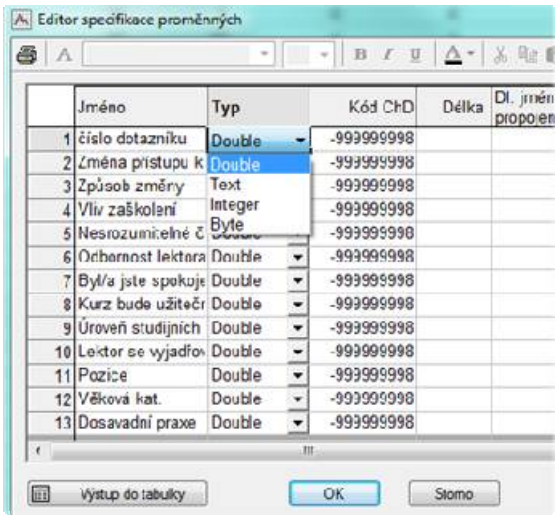
V posledním kroku mě software *STATISTICA* upozorňuje na to, že v proměnné, kterou vyhodnotil jako číselnou, se vyskytují textové popisky. Ve verzi *12* lze zaškrtnou „*Provést pro všechny…*“ a nově kliknout na ***Převést na ChD***, v tomto případě budou textové popisky v číselných proměnných (např. *N/A* apod.) převedeny na chybějící pozorování, tedy na prázdnou buňku.

Starší verze tuto možnost nemají, a proto si ukážeme případ, kde tyto textové popisky v číslených proměnných máme.

Editor textových hodnot

Dvojklikem na záhlaví každé proměnné můžeme vyvolat dialog konkrétní proměnné a v části ***Textové hodnoty*** se lze podívat, jestli se zde nějaký text (kterému by software přiřadil číselnou reprezentaci) nevyskytuje:

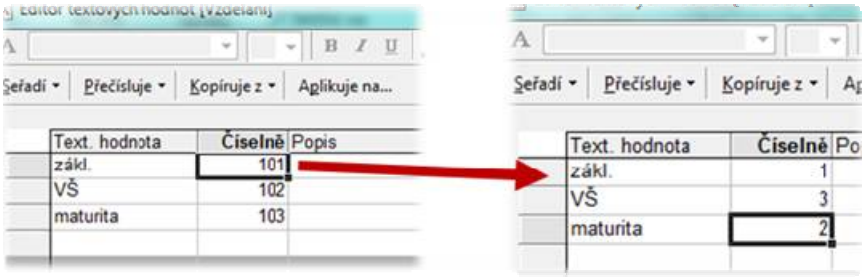




Textový popisek má od softwaru přiřazenu číselnou reprezentaci, pokud je proměnná typu *Double*, lze se na tuto reprezentaci v *Editoru textových hodnot* podívat. Pokud je proměnná typu *Text*, přiřazení čísel proběhne automaticky až v případě využití proměnné k analýze. Máte-li v softwaru kategorické proměnné, které budou vstupovat do analýz jako grupovací proměnné (faktory), doporučujeme mít všechny tyto proměnné jako číselný typ *Double* s právě zmíněnými textovými popisky. Číselnou reprezentaci si mohu libovolně překódovat (v ***Editoru textových hodnot***) na vlastní hodnoty (vhodné a využitelné například u pořadí sloupcových grafů nebo při řazení případů číselně, apod.). Změnu z *Text* na *Double* provedeme buď jednotlivě ve specifikaci jednotlivých proměnných nebo hromadně ve specifikaci všech proměnných, tedy po kliknutí na tlačítko ***Všechny specif***. v dialogu kterékoli proměnné.

Vlastní překódování bychom potom provedli individuálně, například takto:

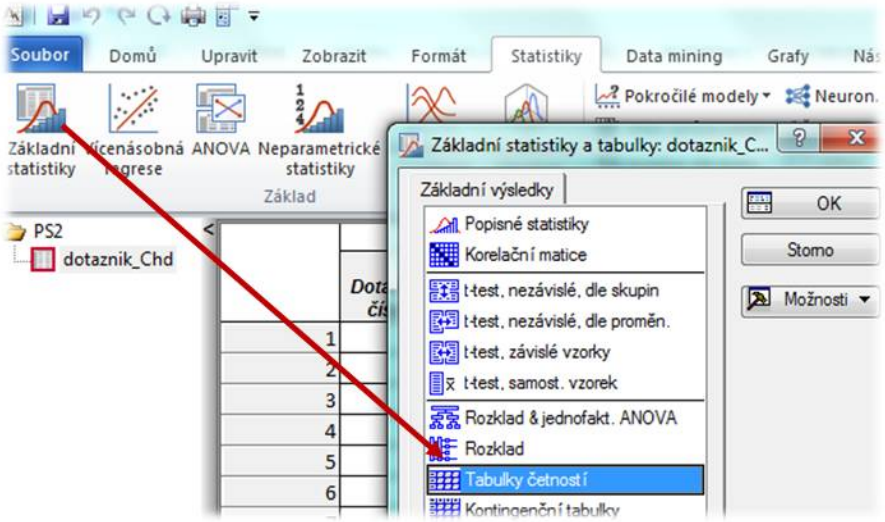
Poznámka: Textové popisky jsou vlastně přiřazení textu jakékoli číselné hodnotě, což je vhodné především pro přehlednost souboru, kde můžeme vidět buď textové popisky, nebo číselnou reprezentaci.



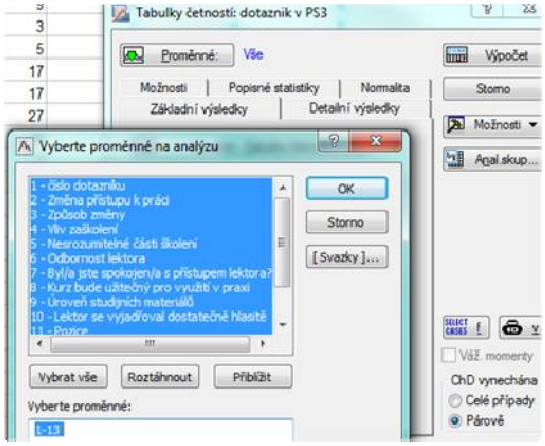
U proměnných číselných jsou samozřejmě textové popisky nežádoucí, pojďme se nyní podívat na to, jak bychom je detekovali.

Detekce neexistujících kategorií

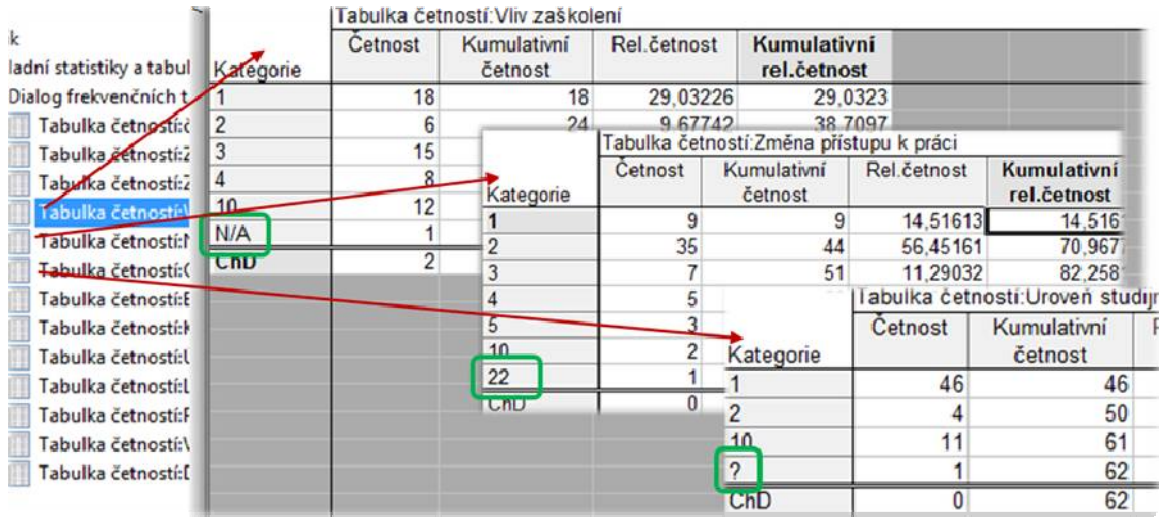
Jednou z možností, jak se podívat na jednotlivé proměnné je tabulka četností. V základních statistikách vybere ***Tabulku četností***:



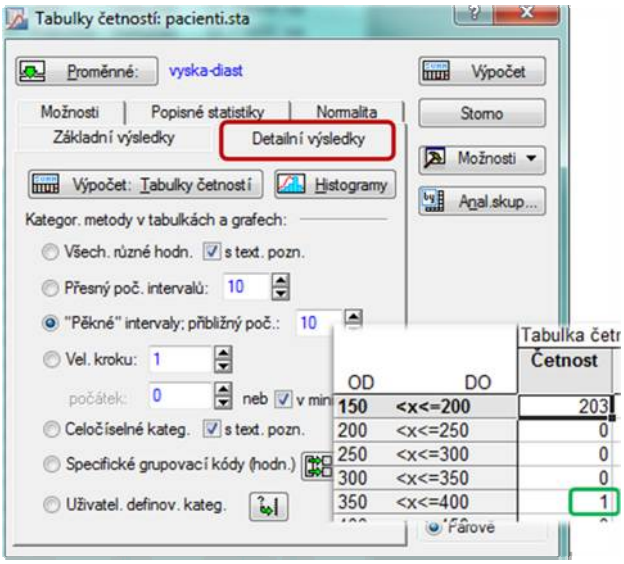
V případě našeho datového souboru (výsledky dotazníkového šetření) vybereme všechny proměnné a klikneme na ***Výpočet***.



Postupně se proklikám jednotlivými tabulkami četností v sešitu výsledků a snadno identifikuji, jestli se v datech nevyskytují jiné kategorie, než mají, kolik je chybějících hodnot, atd.



V případě, že v datovém souboru máme i spojité proměnné, tak tyto proměnné načteme zvlášť v druhém kroku, v dialogu tabulky četností přepneme na ***Detaily*** a zvolíme například ***Pěkné intervaly:***

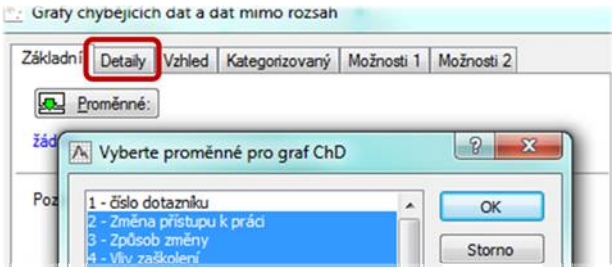


Takto můžeme například identifikovat hodnoty, které jsou například mimo reálně možné meze.

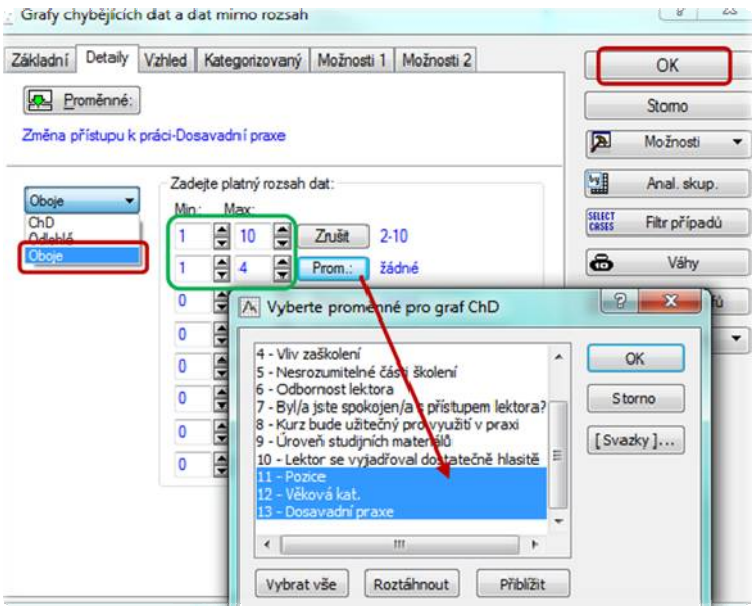
Berme tuto metodu pouze jakousi základní hrubou detekci nevhodných dat, rozsahy intervalů bychom pro potřeby popisné statistiky optimalizovali samozřejmě pro každou proměnnou zvlášť.

V hlavní roli průzkumník

Nyní bychom chtěli identifikovat případy (řádky), ve kterých se „škodlivá“ data vyskytují, to bude dalším krokem v naší analýze. Využijeme grafickou metodu, kterou je *Graf chybějících hodnot.* V záložce ***Grafy -> 2D -> Grafy chybějících hodnot dat nebo dat mimo rozsah*** otevřeme dialog tohoto grafu a vybereme proměnné.

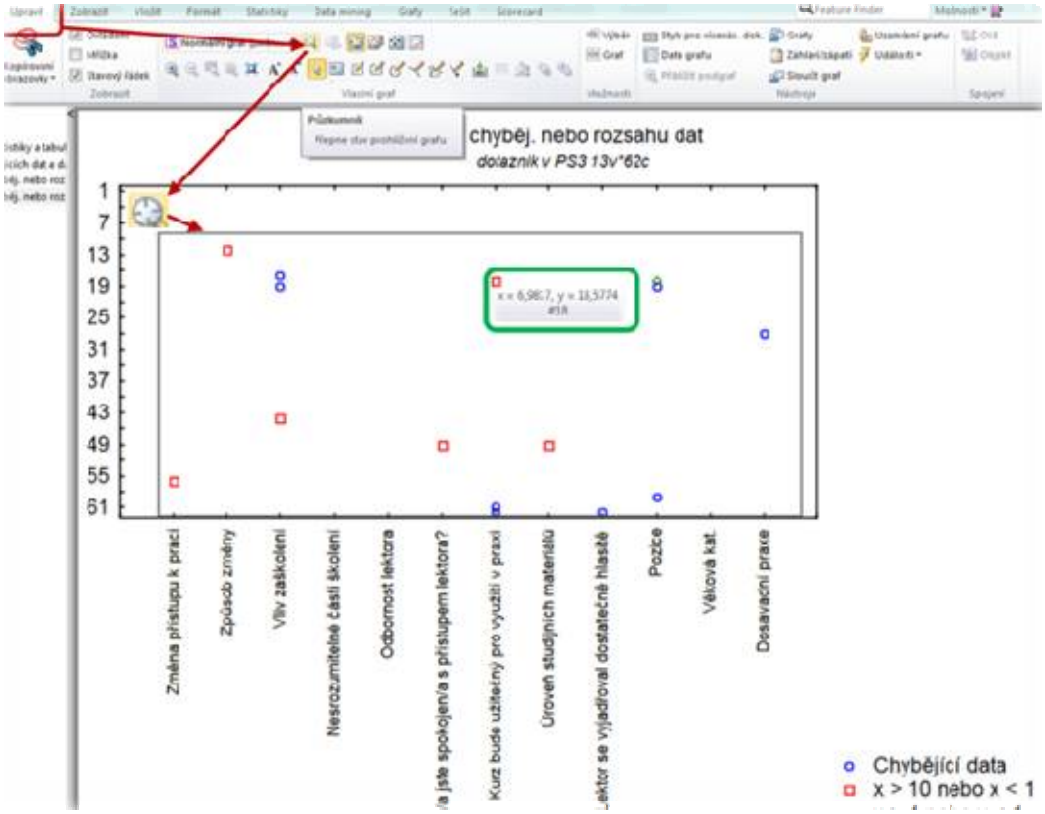


Přepneme na kartu ***Detaily*** a v roletce zvolíme *Oboje* (tedy detekci dat mimo rozsah i ChD).



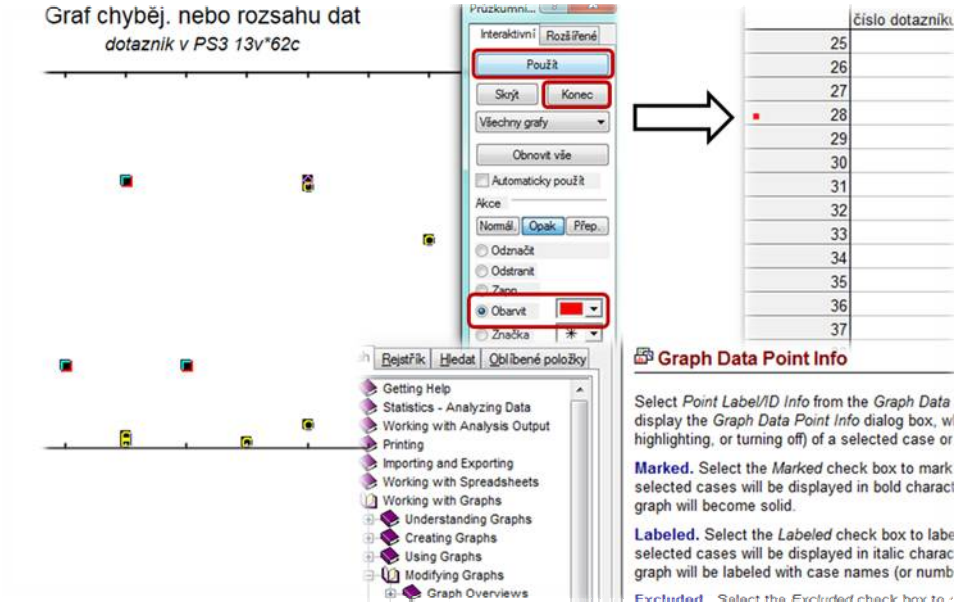
V části ***Zadejte platný rozsah dat*** je možné zvolit rozmezí hodnot, které jsou platné. V našich datech máme dva možné typy rozsahů, rozdělíme tedy proměnné na dvě skupiny a určíme pro ně rozsahy. Zvolíme první a druhou sadu proměnných a upřesníme jejich rozsah (to je výhodné především u dotazníků, kdy víme předem, jaké jsou možné výsledky otázky, které otázky jsou například na škále 1-10, atd.), po té klikneme ***Ok*** a získáme graf. Jedná se o graf, který vykresluje místa, kde v souboru chybí pozorování nebo je zde pozorování mimo stanovený rozsah. Jsou tedy vyobrazeny jen problémové místa souboru.

Najedeme-li kurzorem na **konkrétní označené pozorování**, získáme informaci o čísle případu (*v obrázku jde o pozorování č. 18).* Naším cílem je identifikovat všechna tato pozorování v datovém souboru. Jednou z možností je využít interaktivního průzkumníka grafu. V záložce ***Upravit*** vyberme ***Průzkumníka*** (to platí pro nabídky typu *Pás karet* nebo klikneme do grafu pravým tlačítkem – například vedle nadpisu - a vybereme ***Průzkumník)***.

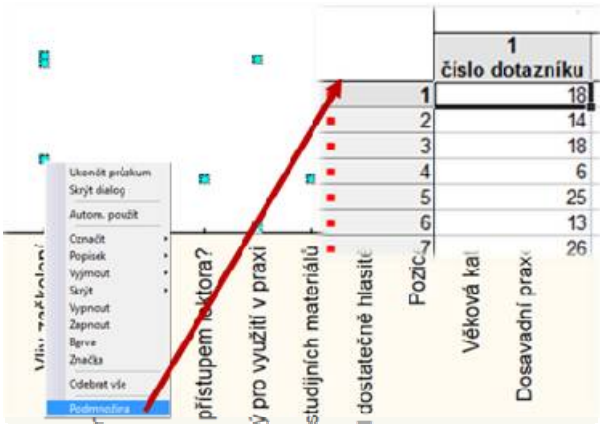


Poté obdélníkovým výběrem vyberte označte body grafu – při zapnutém Průzkumníku dáte kurzor do plochy grafu, následně stiskněme levé tlačítko myši a označme (roztáhněme čtverec) celou plochu grafu.

V dialogu ***Průzkumníka*** zvolme potom např. ***Obarvit*** a klikněme na ***Použít*** a následně na ***Konec***:

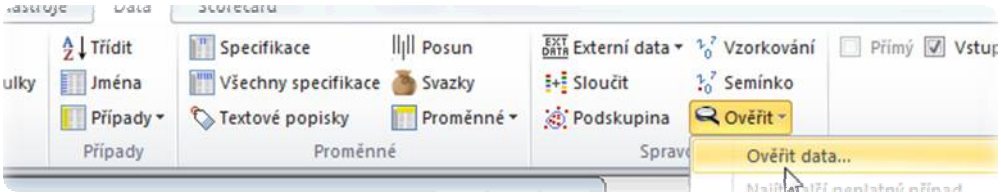


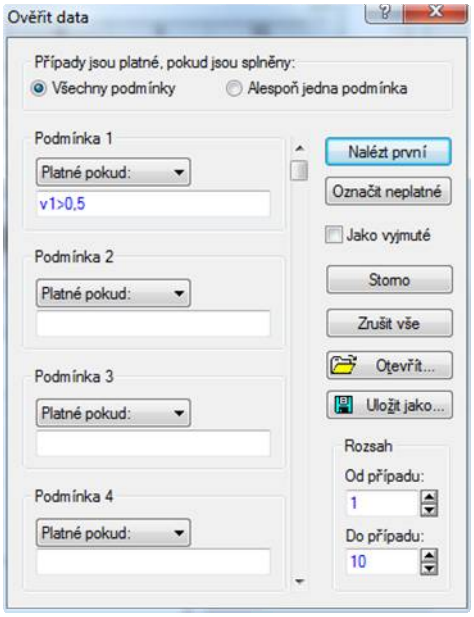
Případy, které přísluší označeným bodům v grafu, byly obarveny přímo v datovém souboru. Klávesou ***F1*** v dialogu ***Průzkumník*** vyvoláme nápovědu, kde je popsán význam jednotlivých možností. Novinkou ve verzi *STATISTICA 12* je možnost (přes pravé tlačítko myši) označená data nechat vygenerovat jako podmnožinu do nové tabulky. Na takovéto podmnožině se poté přehledně podíváme na jednotlivá vadná pozorování.



Možnost Ověřit data

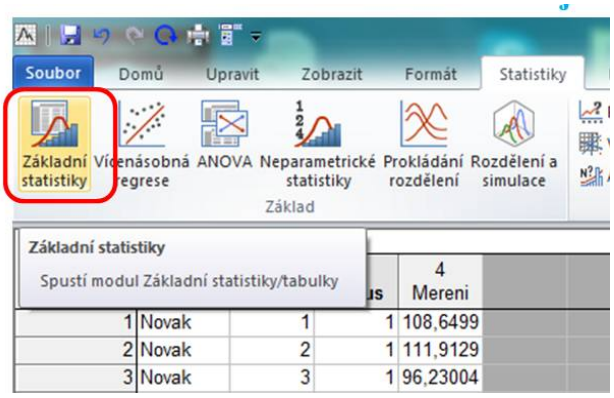
Velmi obdobně, jako jsme využili před chvílí graf hodnot mimo rozsah, můžeme najít data mimo rozsah i jinak. Stačí použít funkcionalitu ***Ověřit data***, kterou najdeme v záložce ***Data-Ověřit-Ověřit data…*** Zde si můžeme zadat velký počet podmínek a omezení, které mají data splňovat (tyto podmínky lze pomocí tlačítka ***Uložit jako*** uložit pro pozdější použití, taktéž lze pomocí ***Otevřít*** podmínky nahrát). Data, která nejsou platná poté můžeme označit pomocí tlačítka ***Označit neplatné*** nebo jít jedno neplatné pozorování po druhém, podobně jako funguje vyhledávání textu v souborech (tlačítko ***Nalézt první*** a poté přejít na další pomocí klávesové zkratky ***ctrl+F3***). Takto je možné neplatné pozorování v souboru postupně kontrolovat a případně přímo manuálně opravovat.

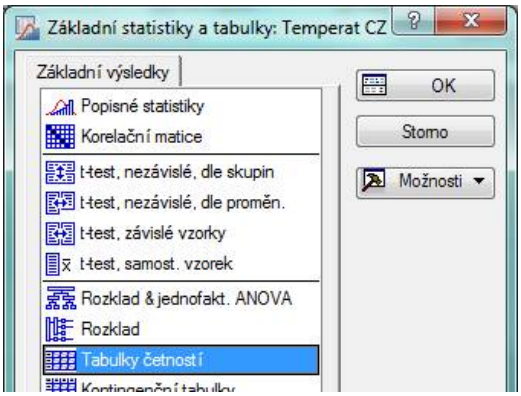




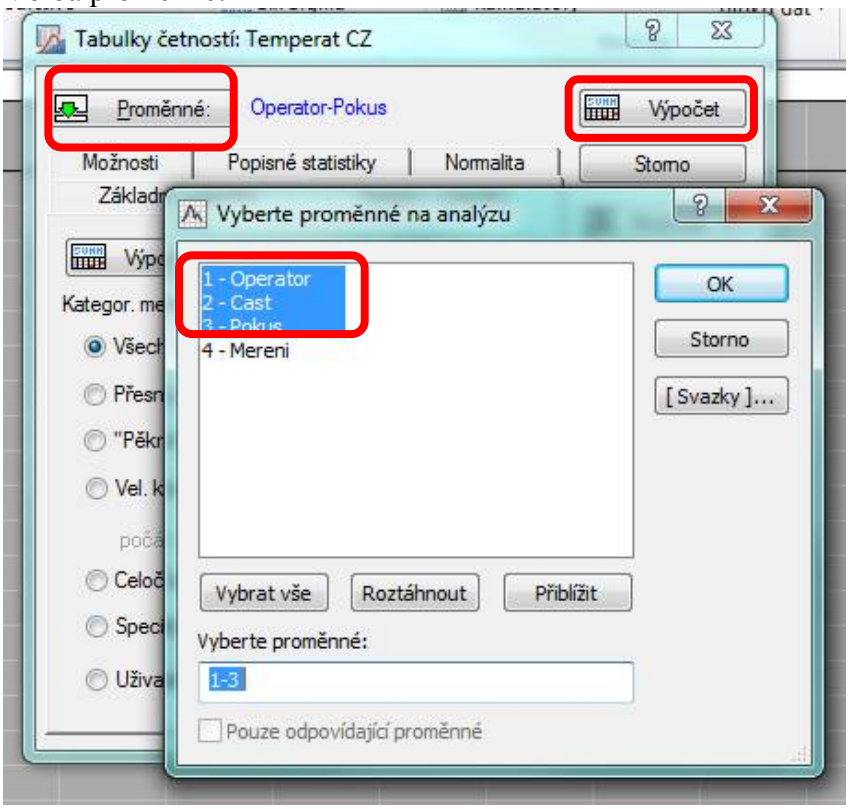
4 Vytvoření základní výpočtů

4.1 Tabulka četností

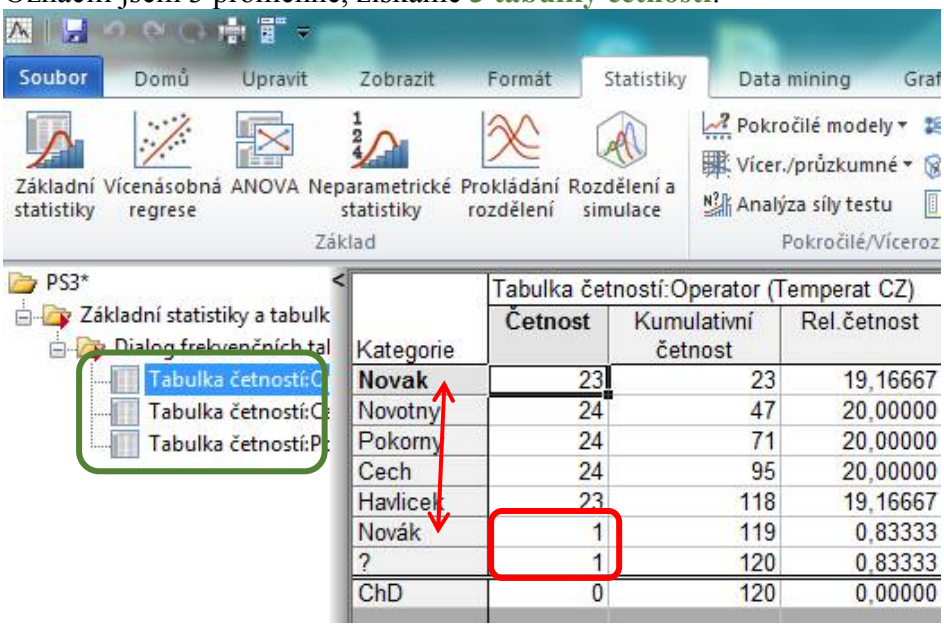




Volba proměnné:

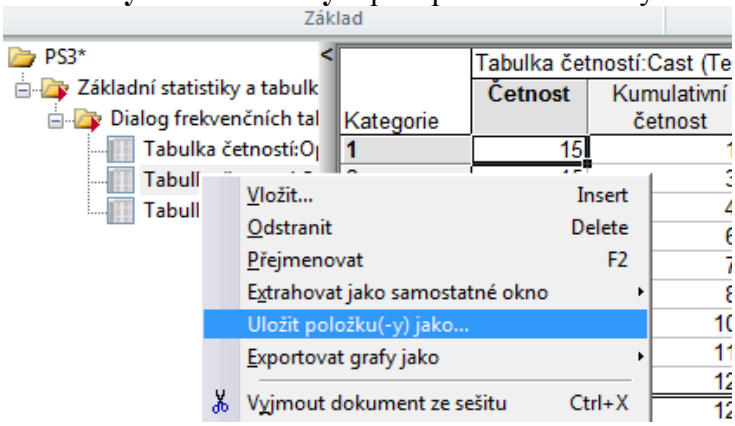


Označili jsem 3 proměnné, získáme **3 tabulky četností**:

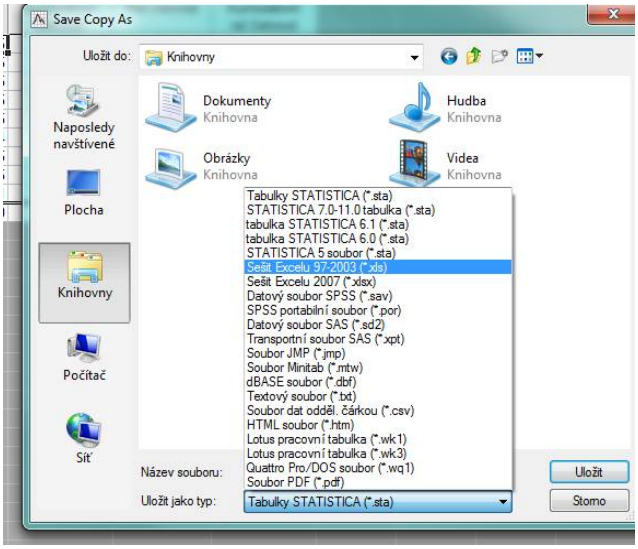


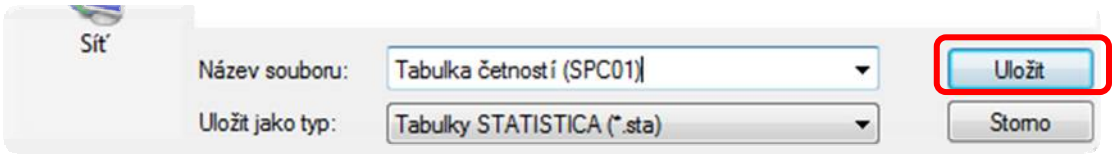
Detekujeme chybně napsané/duplicitní kategorie….

Uložení výsledné tabulky – přes pravé tlačítko myši na vybrané tabulce v sešitu:



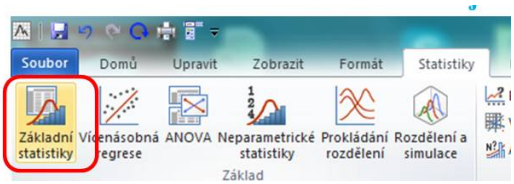
Voba názvu a formátu:





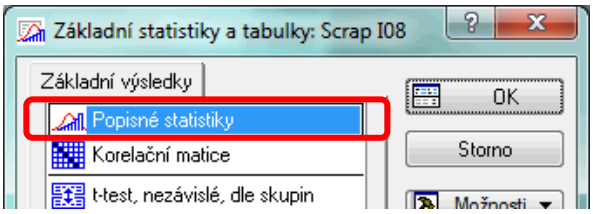
4.2 Popisné statistiky

Na popisnou statistiku si vyzkoušejme 2 příklady:

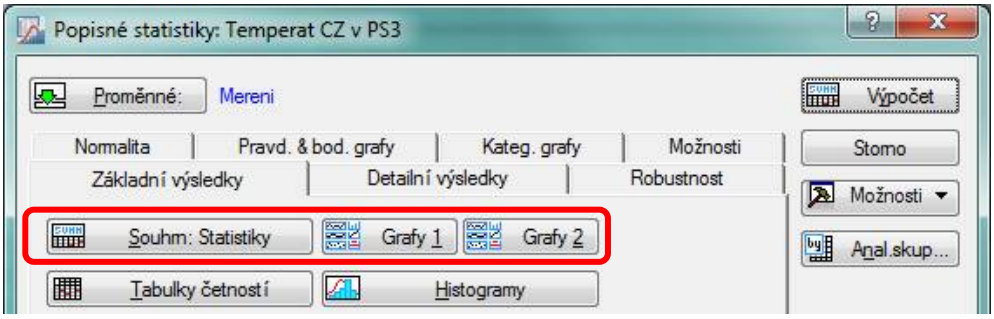


4.2.1 Soubor Temperat CZ.sta - měření dílů jednotlivými operátory

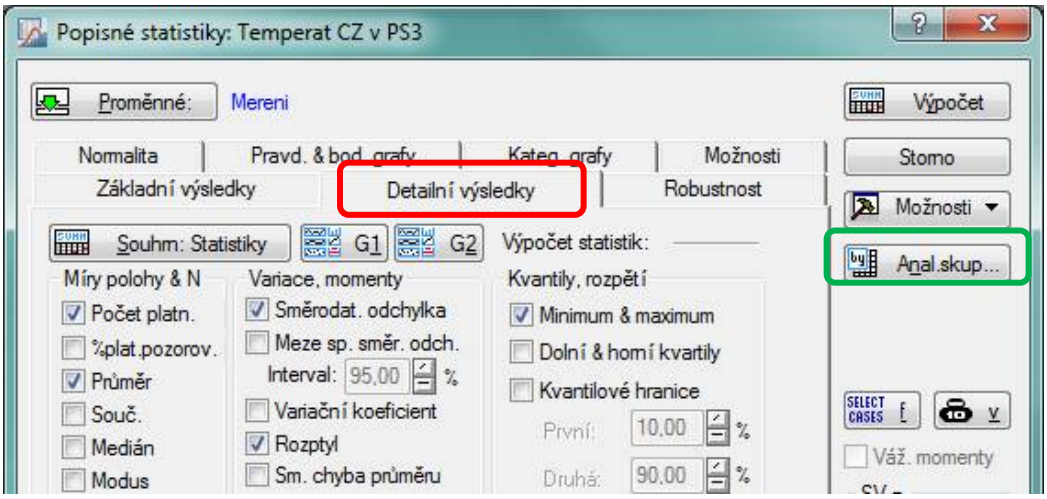
Přes záložku ***Statistiky -> Základní statistiky -> Popisné statistiky***



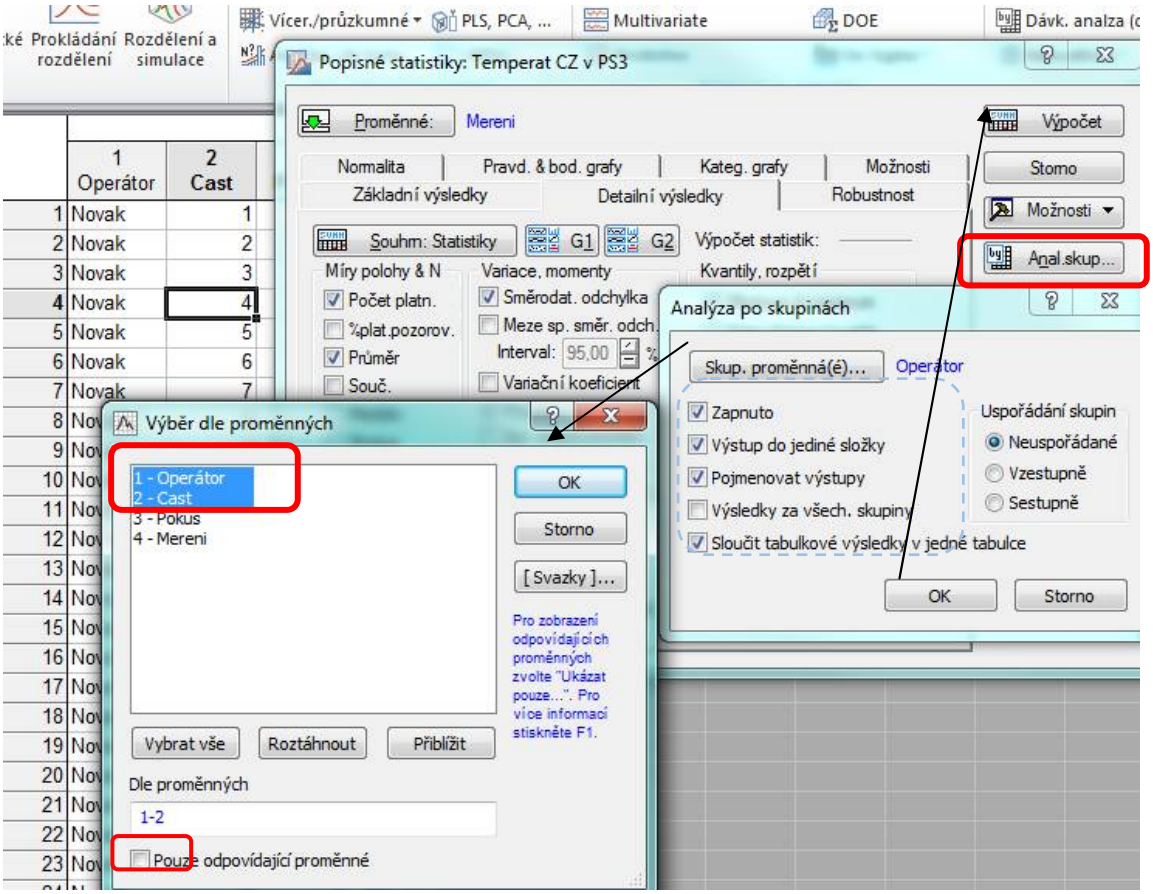
Vybereme proměnné (pro více proměnných držím při výběru myši **Ctrl**) – vybrali jsme spojitou (měřenou) proměnnou „měření“. V softwaru je několik možností, jak popisné charakteristiky získat:



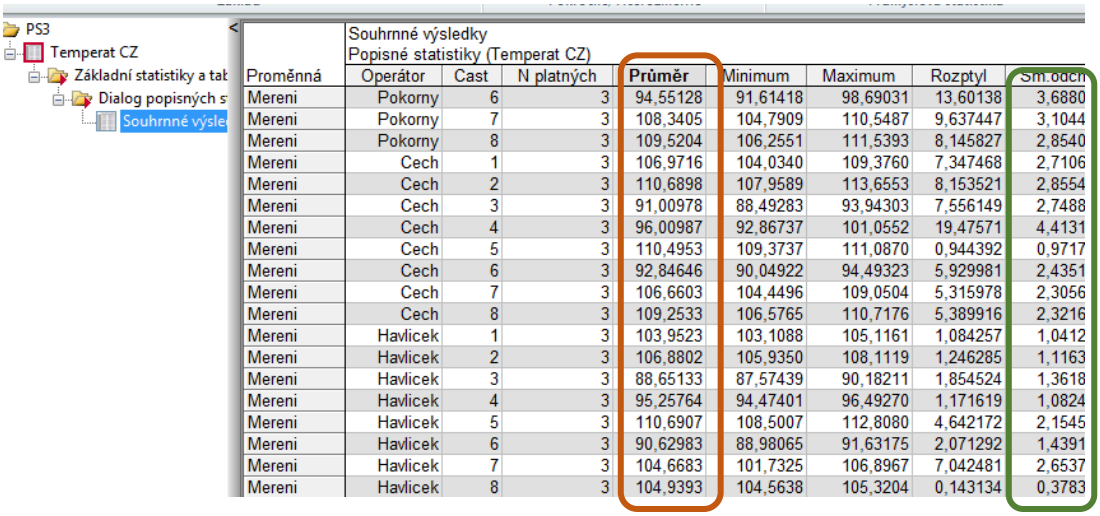
Na části ***Detailní výsledky*** si lze vybrat přesně to, co potřebuji (průměry, medián apod.):



Tlačítkem ***Anal.skupiny*** si rozdělíme výpočet podle jednotlivých operátorů a podle jednotlivých dílů:



Výsledkem je tabulka popisných statistik (**průměrně změřené hodnoty** rozdělené dle typu dílu a operátora, **variabilita (kolísavost)** měření dle operátorů a dílů):



Souhrnné výsledky
Popisné statistiky (Temperat CZ)

Proměnná	Operátor	Cast	N platných	Průměr	Minimum	Maximum	Rozptyl	Sm.odch
Mereni	Pokorny	6	3	94,55128	91,61418	98,69031	13,60138	3,6880
Mereni	Pokorny	7	3	108,3405	104,7909	110,5487	9,637447	3,1044
Mereni	Pokorny	8	3	109,5204	106,2551	111,5393	8,145827	2,8540
Mereni	Cech	1	3	106,9716	104,0340	109,3760	7,347468	2,7106
Mereni	Cech	2	3	110,6898	107,9589	113,6553	8,153521	2,8554
Mereni	Cech	3	3	91,00978	88,49283	93,94303	7,556149	2,7488
Mereni	Cech	4	3	96,00987	92,86737	101,0552	19,47571	4,4131
Mereni	Cech	5	3	110,4953	109,3737	111,0870	0,944392	0,9717
Mereni	Cech	6	3	92,84646	90,04922	94,49323	5,929981	2,4351
Mereni	Cech	7	3	106,6603	104,4496	109,0504	5,315978	2,3056
Mereni	Cech	8	3	109,2533	106,5765	110,7176	5,389916	2,3216
Mereni	Havlicek	1	3	103,9523	103,1088	105,1161	1,084257	1,0412
Mereni	Havlicek	2	3	106,8802	105,9350	108,1119	1,246285	1,1163
Mereni	Havlicek	3	3	88,65133	87,57439	90,18211	1,854524	1,3618
Mereni	Havlicek	4	3	95,25764	94,47401	96,49270	1,171619	1,0824
Mereni	Havlicek	5	3	110,6907	108,5007	112,8080	4,642172	2,1545
Mereni	Havlicek	6	3	90,62983	88,98065	91,63175	2,071292	1,4391
Mereni	Havlicek	7	3	104,6683	101,7325	106,8967	7,042481	2,6537
Mereni	Havlicek	8	3	104,9393	104,5638	105,3204	0,143134	0,3783

V zápětí se podíváme na možnosti editace tabulky.Více o průměrech a mírách kolísavosti se dočtete v:

17/09/2012 StatSoft ACADEMY – charakteristiky polohy

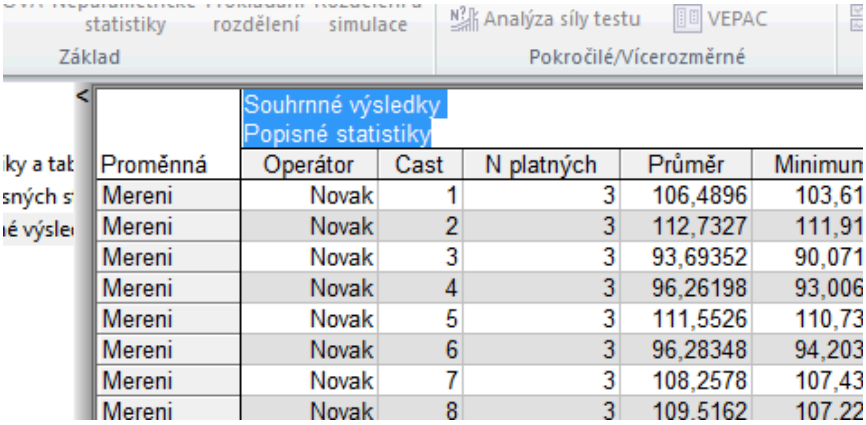
http://www.statsoft.cz/file1/PDF/newsletter/2012_09_17_StatSoft_popisna_statistika.pdf

15/10/2012 StatSoft ACADEMY - charakteristiky variability

http://www.statsoft.cz/file1/PDF/newsletter/2012_10_15_StatSoft_Popisne_statistiky_-_miry_variabily.pdf

4.2.2 Editace tabulky (Anglické popisky apod.)

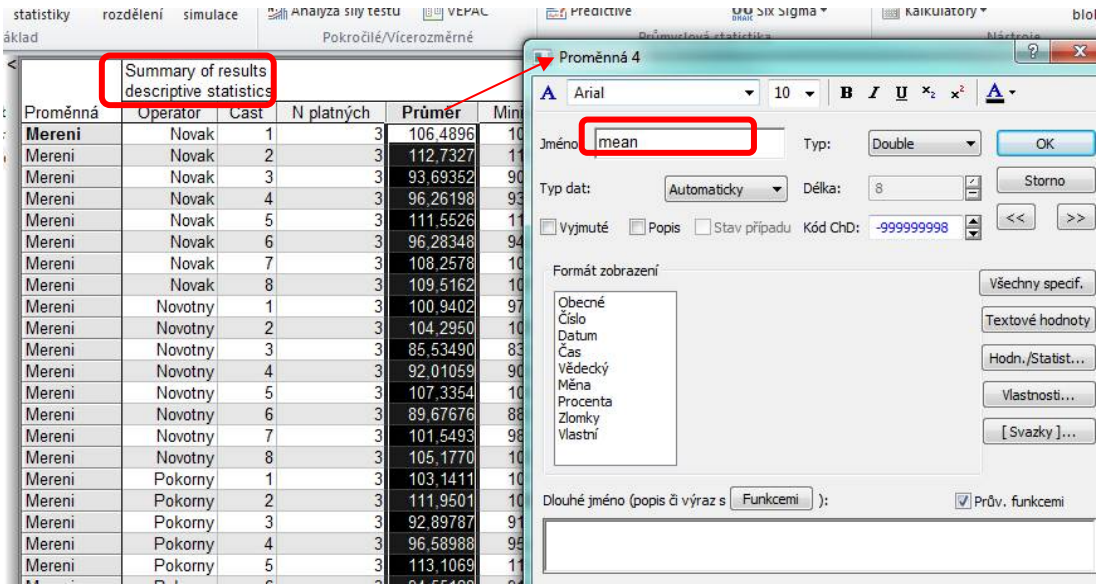
Dvojklikem do např. záhlaví tabulky



Souhrnné výsledky
Popisné statistiky

Proměnná	Operátor	Cast	N platných	Průměr	Minimun
Mereni	Novak	1	3	106,4896	103,61
Mereni	Novak	2	3	112,7327	111,91
Mereni	Novak	3	3	93,69352	90,071
Mereni	Novak	4	3	96,26198	93,006
Mereni	Novak	5	3	111,5526	110,73
Mereni	Novak	6	3	96,28348	94,203
Mereni	Novak	7	3	108,2578	107,43
Mereni	Novak	8	3	109,5162	107,22

- *CTRL+A* označíme celý text, následně *CTRL+C* zkopírujeme a vložíme např. do překladače Google apod.
- Nový text zkratkou *CTRL+V* vložíme do záhlaví:

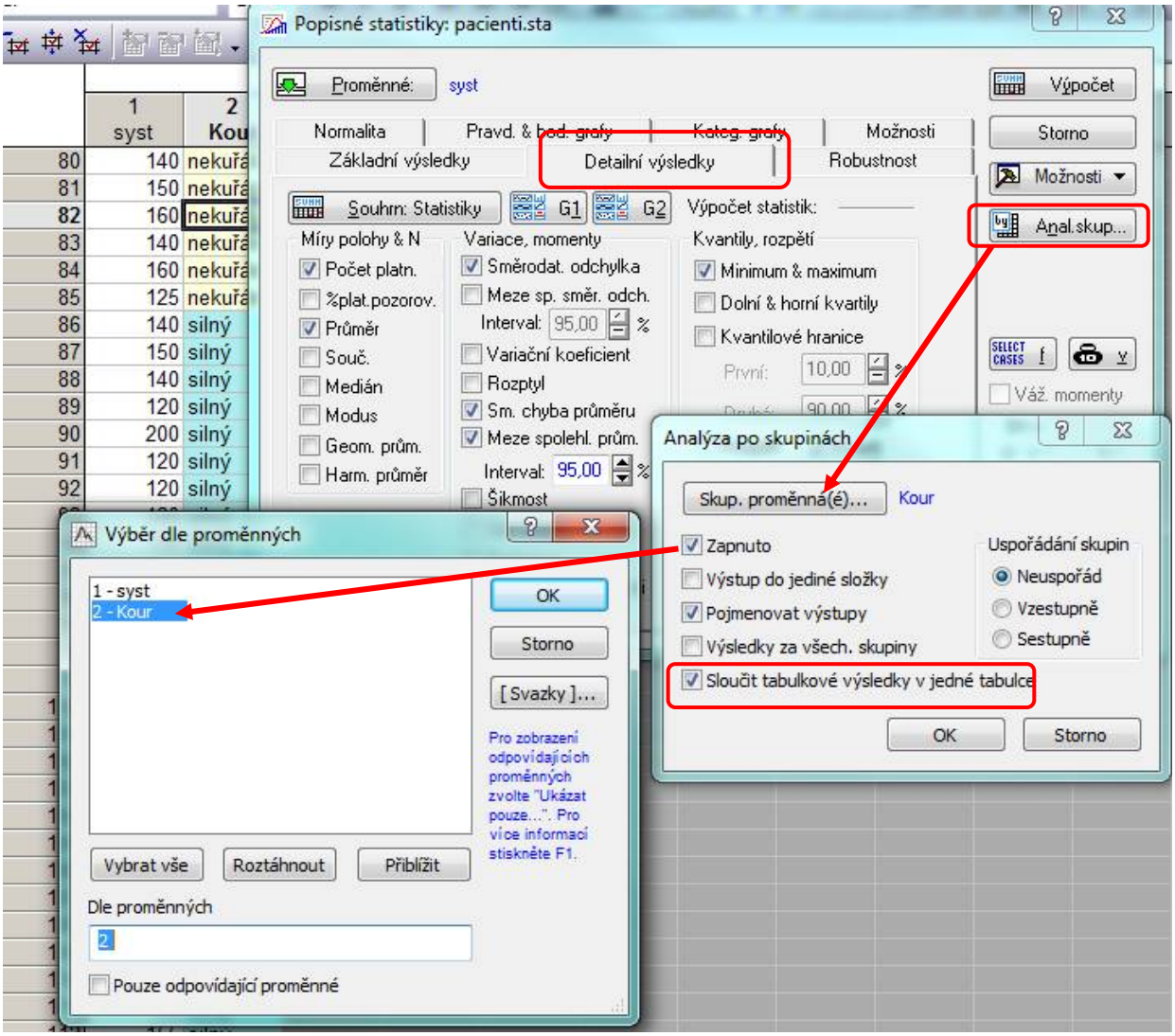


Proměnná	Operator	Cast	N platných	Průměr	Mini
Mereni	Novak	1	3	106,4896	10
Mereni	Novak	2	3	112,7327	11
Mereni	Novak	3	3	93,69352	90
Mereni	Novak	4	3	96,26198	93
Mereni	Novak	5	3	111,5526	11
Mereni	Novak	6	3	96,28348	94
Mereni	Novak	7	3	108,2578	10
Mereni	Novak	8	3	109,5162	10
Mereni	Novotny	1	3	100,9402	97
Mereni	Novotny	2	3	104,2950	10
Mereni	Novotny	3	3	85,53490	83
Mereni	Novotny	4	3	92,01059	90
Mereni	Novotny	5	3	107,3354	10
Mereni	Novotny	6	3	89,67676	88
Mereni	Novotny	7	3	101,5493	98
Mereni	Novotny	8	3	105,1770	10
Mereni	Pokorny	1	3	103,1411	10
Mereni	Pokorny	2	3	111,9501	10
Mereni	Pokorny	3	3	92,89787	91
Mereni	Pokorny	4	3	96,58988	95
Mereni	Pokorny	5	3	113,1069	11

Dvojklikem na proměnnou ***Průměr*** vyvoláme dialog proměnné a změníme její název. Nyní se podíváme, jak si stojí jednotlivý operátoři v grafickém výstupu:

4.2.3 Rozdělení spojité proměnné dle kategorie

Na kartě ***Detailní výsledky*** v dialogu ***Popisné statistiky: Statistiky —> Základní statistiky —> Popisné statistiky –> anal. Skupiny:***



Kde vybereme proměnnou stupeň kouření (kouř) a charakteristiky polohy a variability tak vypočteme zvlášť pro jednotlivé kategorie.

Kompletní řešené příklady na char. variability a polohy, které ukážou další možnosti softwaru *STATISTICA* v této oblasti lze najít v našich newsletterech:

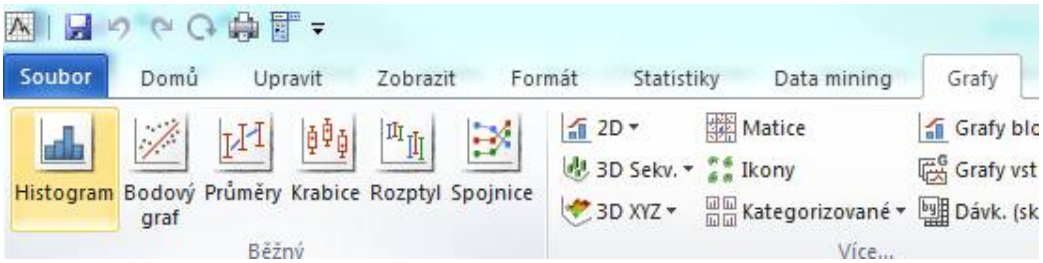
- *Newsletter 20/08/2012*
- *Newsletter 17/09/2012*
- *Newsletter 15/10/2012*

http://www.statsoft.cz/o-firme/archiv-newsletteru/

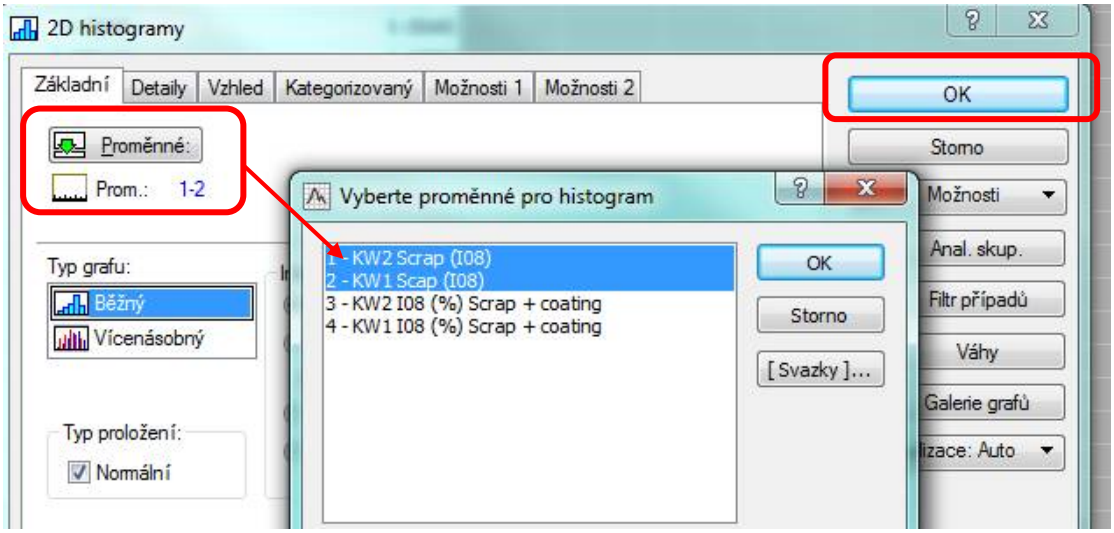
5 Vytvoření grafu

5.1 Histogram

Přes ***Grafy -> 2D grafy -> Histogramy***

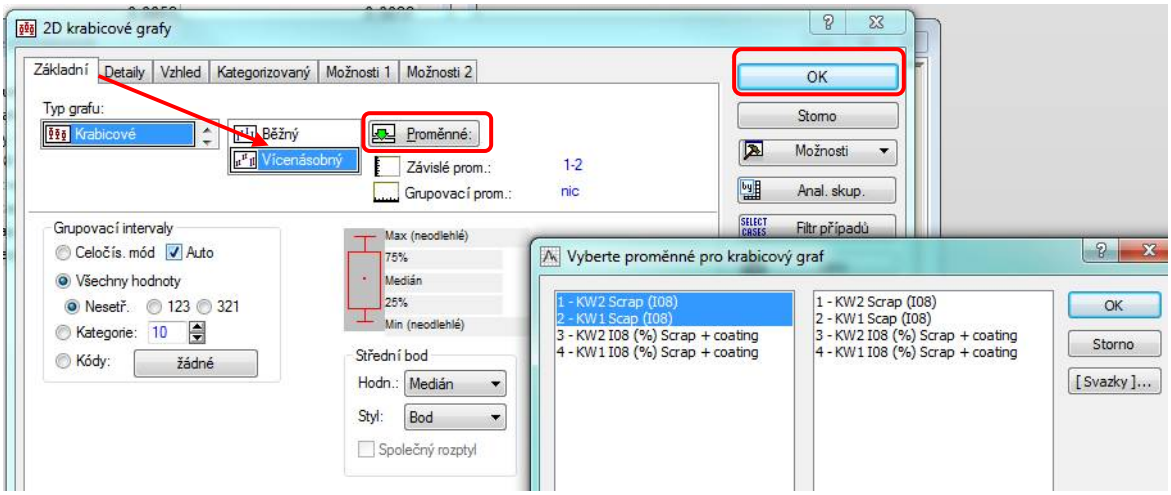


Vybereme proměnné pro obě období:

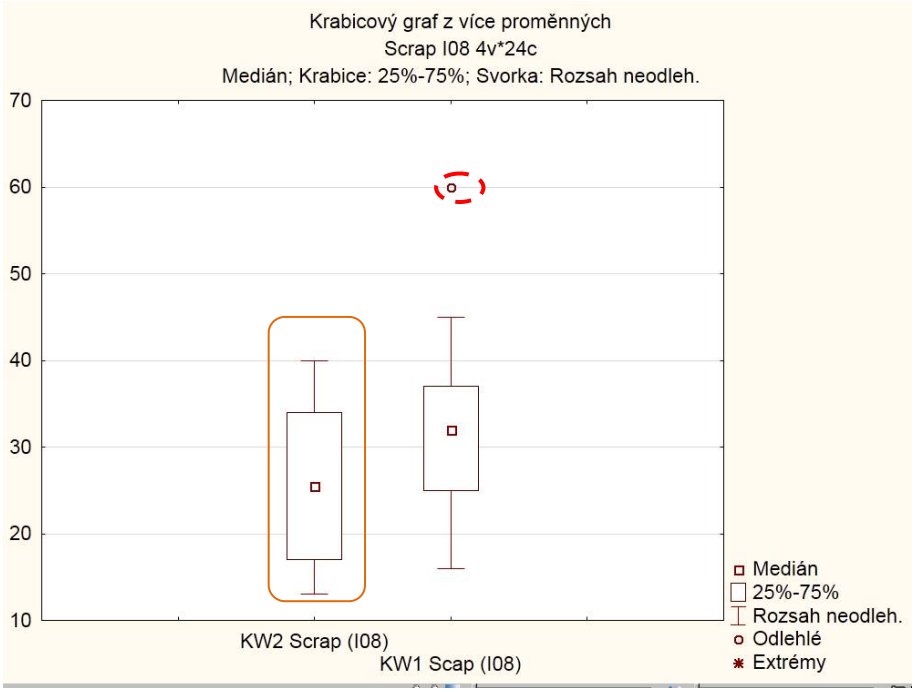


5.2 Krabicový graf (Box Plot)

Tímto grafem si vizuálně porovnáme oba naše vzorky, tedy před vyčištěním a po vyčištění stroje. **Přes** *Grafy -> 2D grafy -> Krabicové grafy*. Zvolíme ***Vícenásobný*** a opět vybereme proměnnou:



Z grafu je vidět, že v období po vyčištění stroje (**KW2**) došlo k celkovému poklesu variability souboru (krabička je níž):



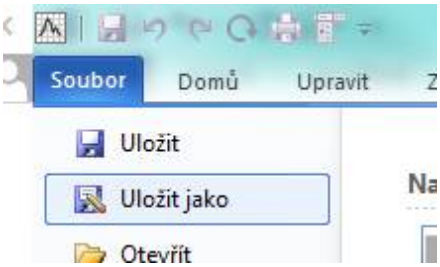
V souboru je jedno **odlehlé pozorování**, které bylo naměřené v období KW1, je třeba zkontrolovat, jestli nejde o chybnou hodnotu operátora.

6 Uložení práce

6.1 Uložení celého sešitu výstupů

Výstupy v souboru lze ukládat několika způsoby, začneme sešitem, který je dobré použít, pokud chci uložit kompletní práci v softwaru *STATISTICA:*

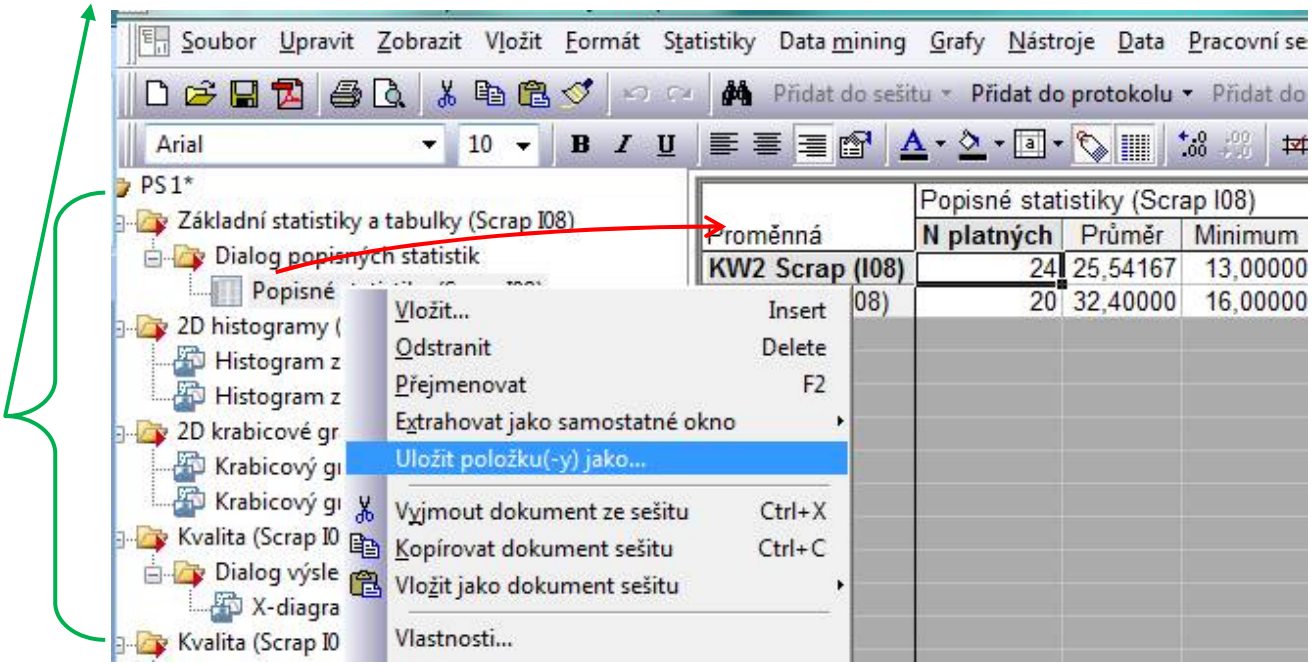
Přes ***Soubor -> Uložit jako…***



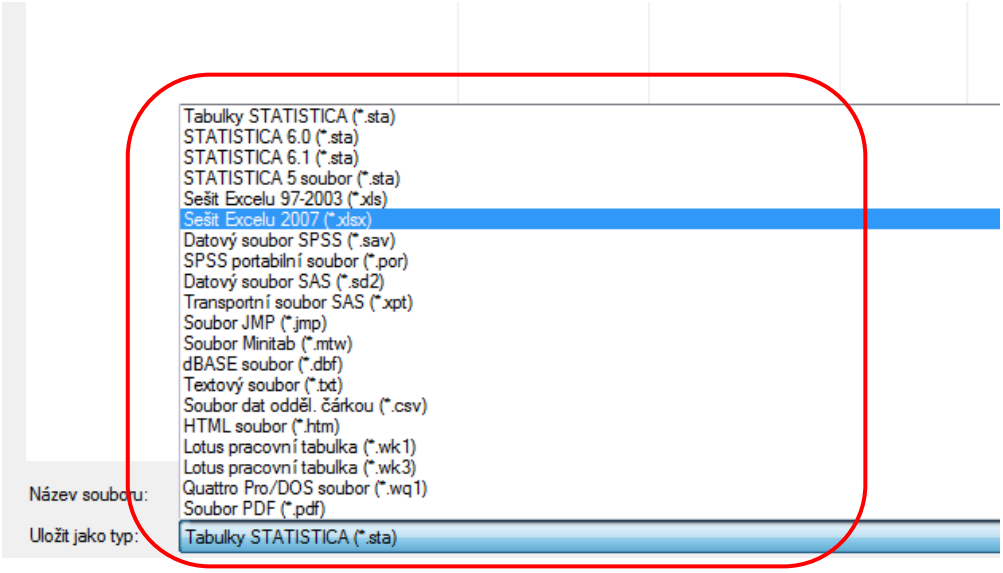
Uložíme soubor, ve kterém je všechno, co jsme vygenerovali. Tento soubor následně otevřeme přes ***Soubor -> Otevřít*** nebo dvojklik přímo na soubor.

6.2 Uložení tabulky v softwaru

– ve **stromu sešitu *STATISTICA*** klikneme přes pravé tlačítko na tabulku Popisných statistik



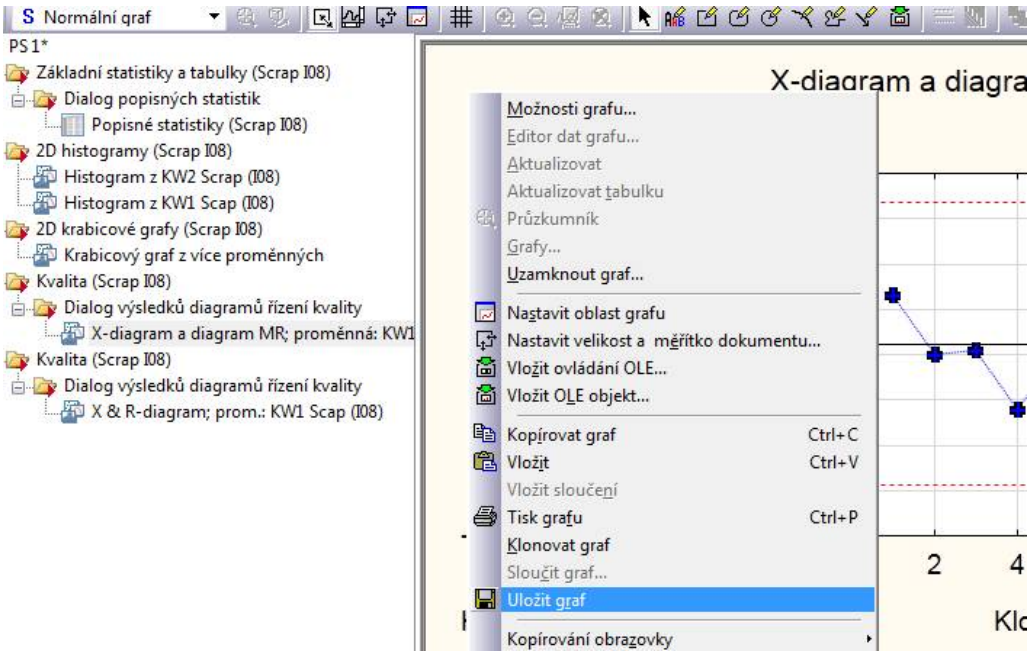
a zvolíme ***Ulož položku(-y) jako…***



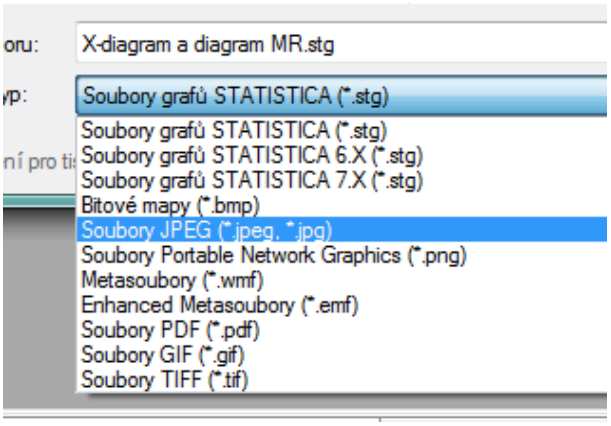
A tabulku si uložíme třeba ve formátu Excelu.

6.3 Uložení grafu

V příslušném grafu kliknu pravím tlačítkem

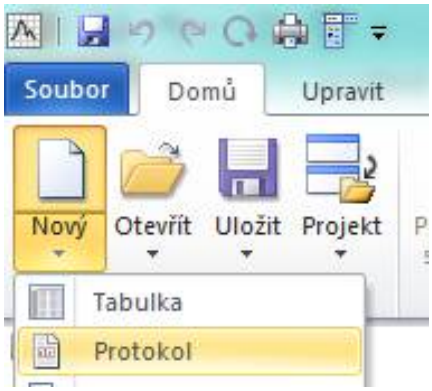


A opět vyberu formát pro uložení:



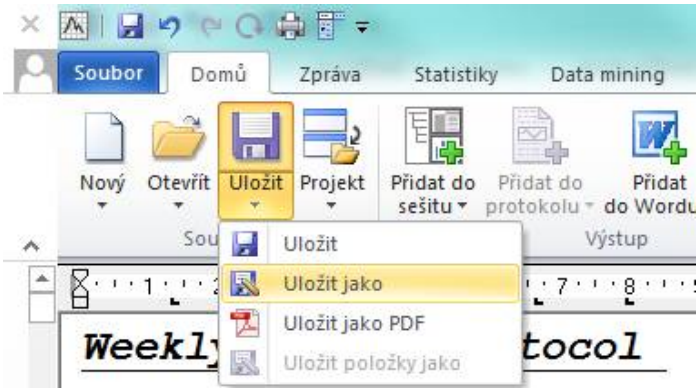
6.4 Přidání výstupů do Protokolu/Microsoft Wordu

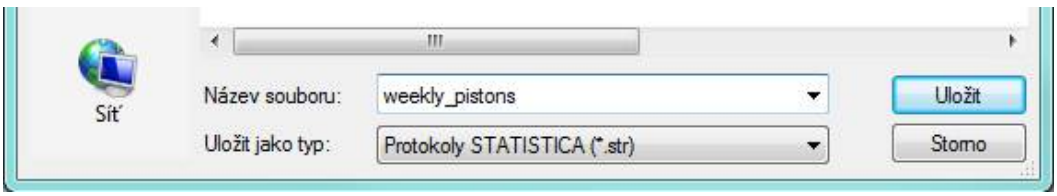
1. Založme si nový protokol:



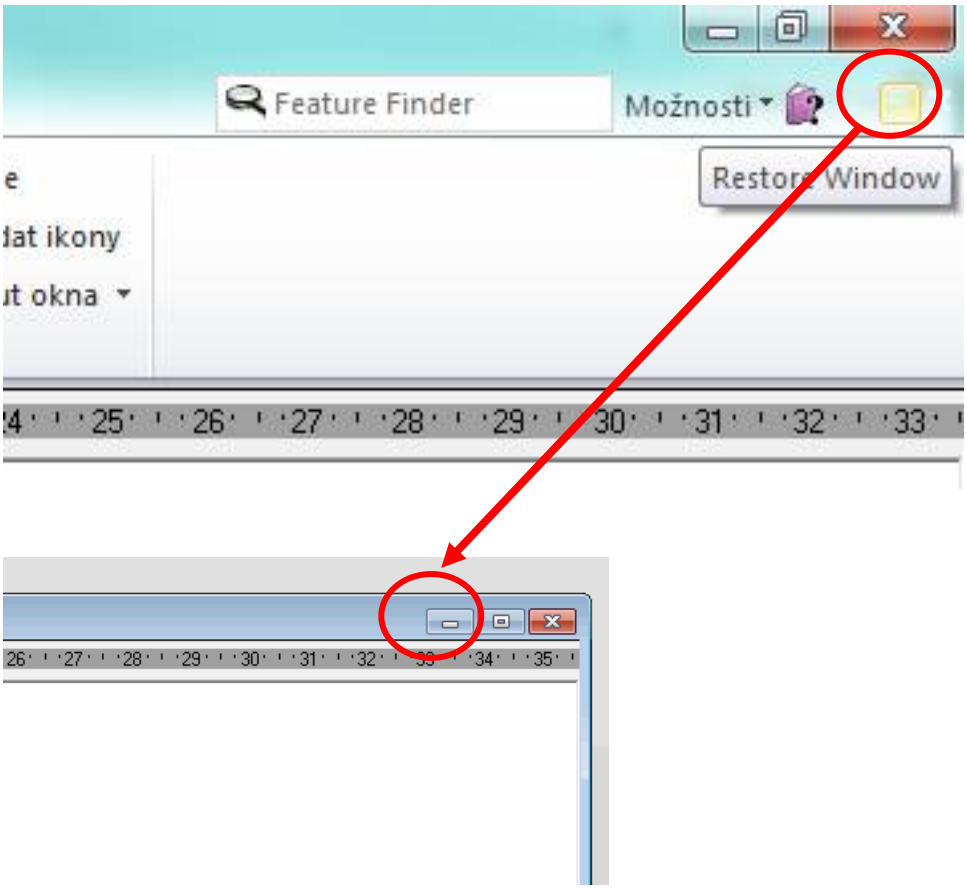
a přidejme popis protokolu:

A protokol uložme:

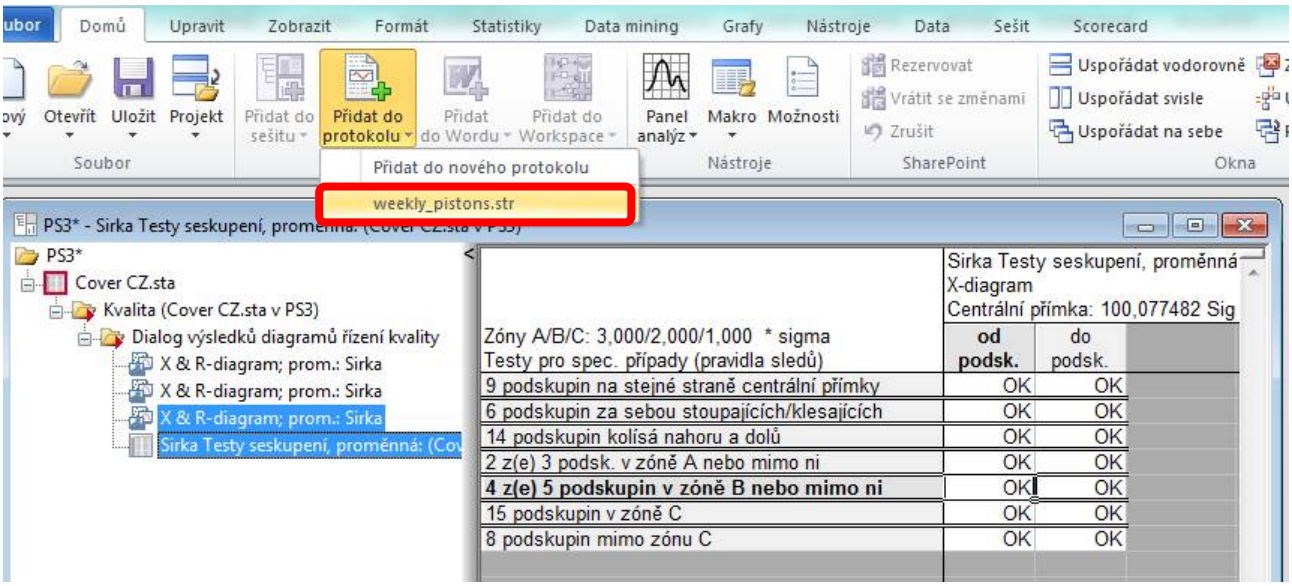




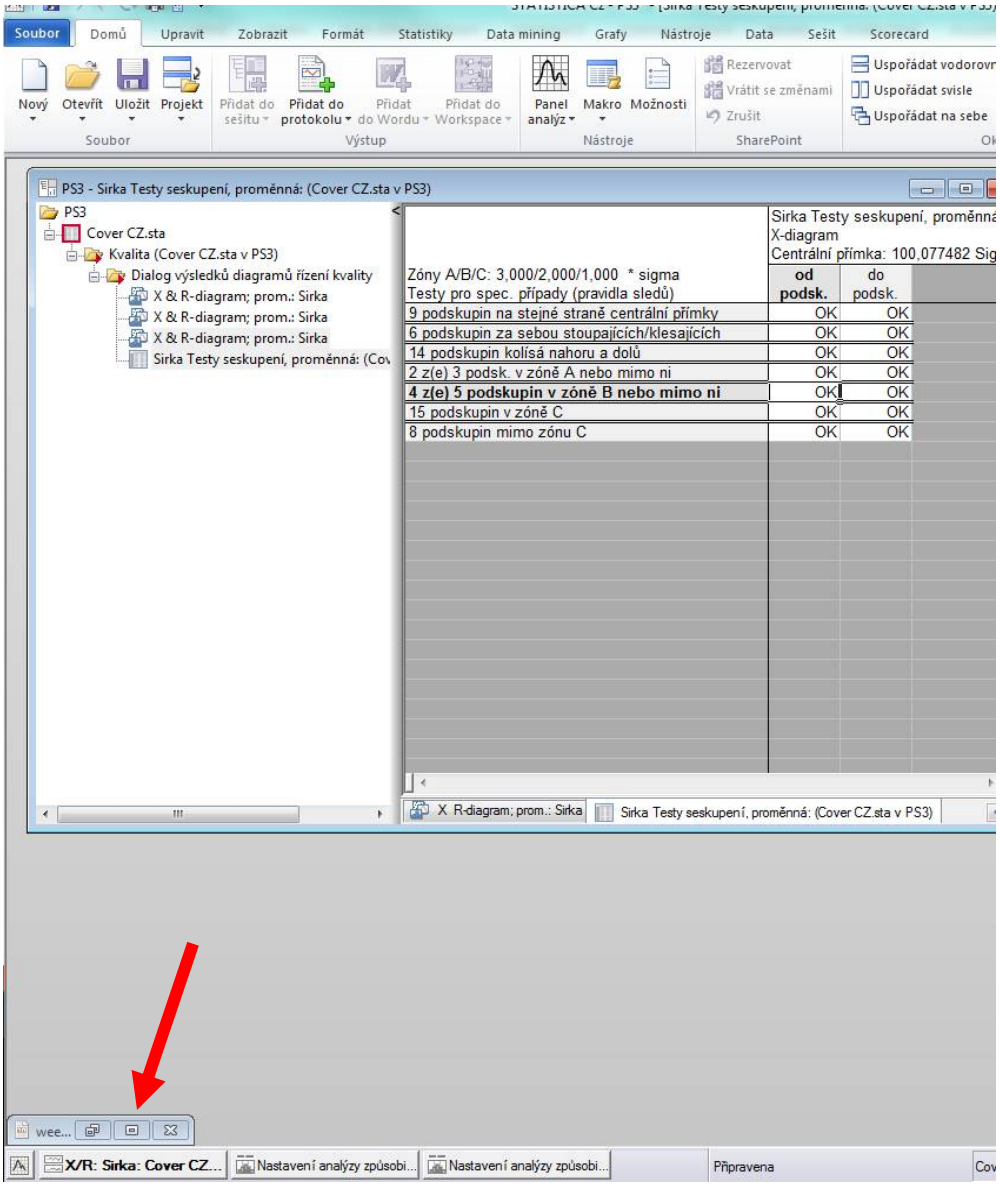
2. Minimalizujme si v pravém horním rohu protokol:



3. V sešitě výsledků označíme (*při stisknutém **Ctrl***) výstupy, které chceme přidat do reportu a klikneme na *Přidat do protokolu*:

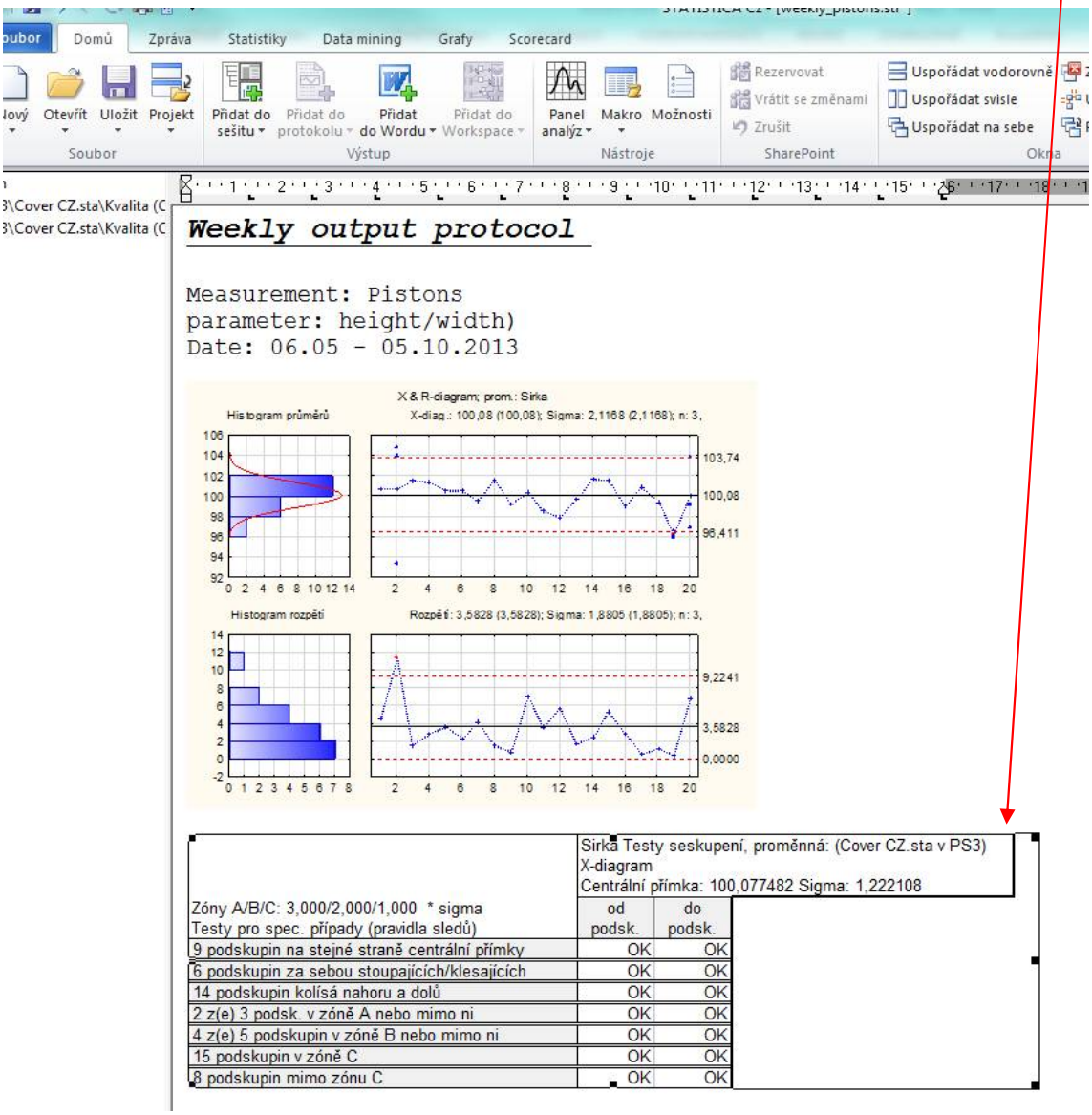


Nyní si podíváme na náš protokol:



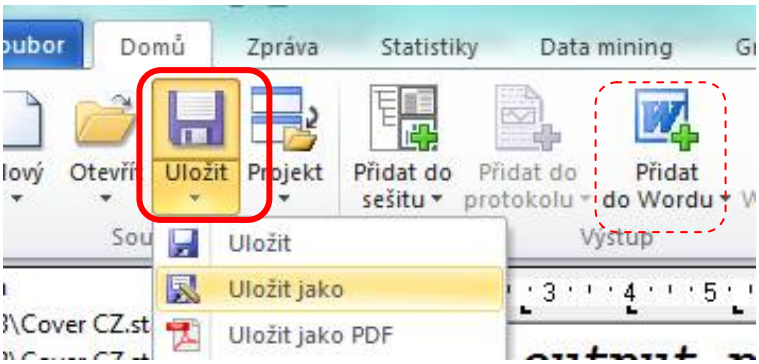
Pokud mám všechna okna maximalizovaná (sešit výsledků i protokol), tak přepínáme pomocí zkratky ***CTRL + TAB***

Výsledný protokol lze dále editovat, pokud se některé tabulky nezobrazili celé, tak je roztáhnu myší.

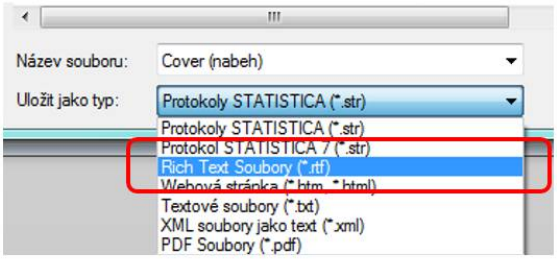


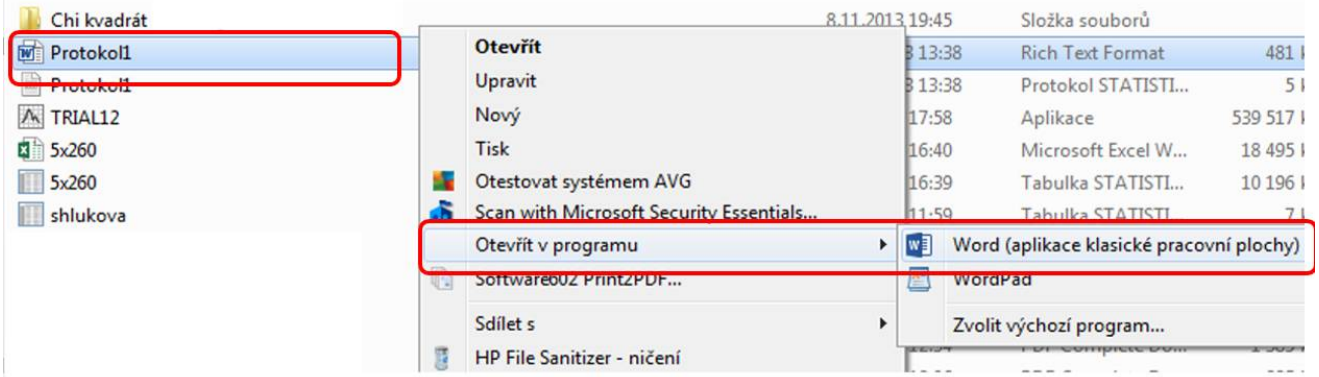
	Sirka Testy seskupení, proměnná: (Cover CZ.sta v PS3) X-diagram Centrální přímka: 100,077482 Sigma: 1,222108	
Zóny A/B/C: 3,000/2,000/1,000 * sigma Testy pro spec. případy (pravidla sledů)	od podsk.	do podsk.
9 podskupin na stejné straně centrální přímky	OK	OK
6 podskupin za sebou stoupajících/klesajících	OK	OK
14 podskupin kolísá nahoru a dolů	OK	OK
2 z(e) 3 podsk. v zóně A nebo mimo ni	OK	OK
4 z(e) 5 podskupin v zóně B nebo mimo ni	OK	OK
15 podskupin v zóně C	OK	OK
8 podskupin mimo zónu C	OK	OK

Výsledný protokol uložím jako PDF, nebo jako ***RTF*** (formát, který lze otevřít ve Wordu a přeuložit jako *.docx)



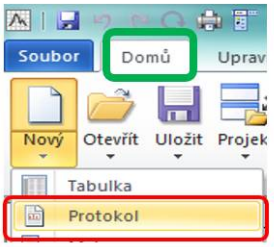
Protokol ***STATISTICA:***



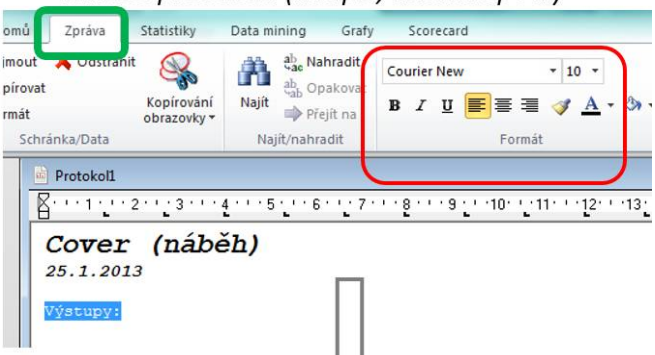


Nebo lze importovat přímo do MS Word (záleží na verzi Office, vždy lze uložit jako RTF a přeuložit). Postup shrnuje obrázek níže:

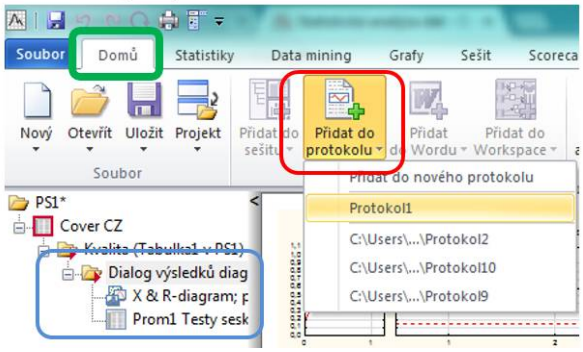
Protokol ***STATISTICA:***

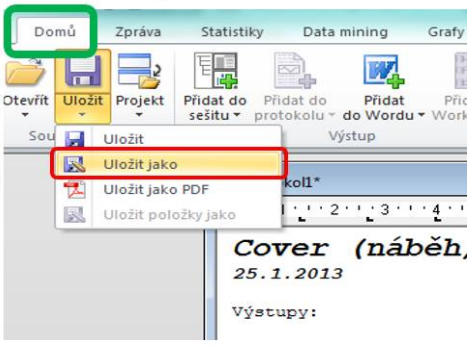


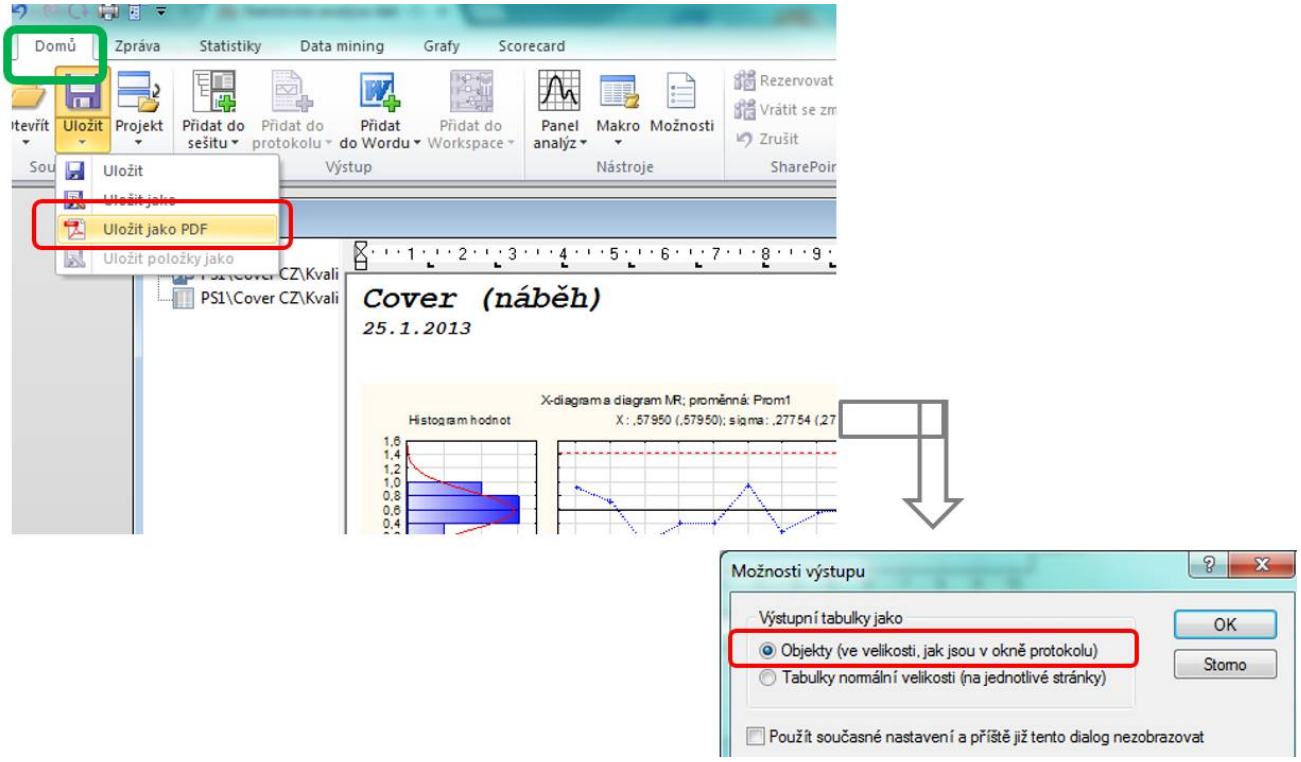
Editace protokolu (nadpis, datum apod.)



Přidání výstupů



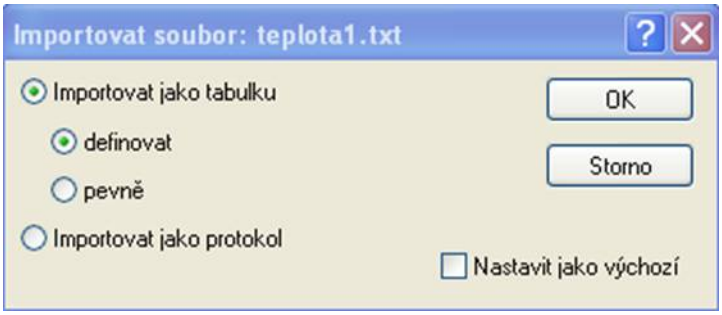




7 Další možnosti načtení souborů

7.1 Otevření textového souboru

V menu ***Soubor*** zvolíme možnost ***Otevřít…*** a pomocí procházení úložišť osobního počítače nadefinujeme cestu k textovému souboru (např. s koncovkou .txt nebo .csv). Potvrdíme ***OK*** a zobrazí se následující dialog:



Ten necháme beze změny a opět potvrdíme ***OK***. Definici, jak přesně chceme k obsahu textového souboru přistupovat, upřesníme prostřednictvím následujícího dialogu:
V horní části dialogu nastavíme oddělovač proměnných (defaultní nastavení je tabelátor nebo středník, podle typu dokumentu). Máme možnost nadefinovat i vlastní oddělovač – volba ***Jiný*** umožňuje vepsat vlastní typ oddělovače. Pokud je oddělovač tvořen celou skupinou znaků, je nutné zaškrtnout možnost ***Užít vše***.

V dolním okně dialogu se automaticky zobrazuje náhled souboru tak, jak bude vypadat po načtení do *STATISTICA*, jednotlivé proměnné (sloupce) jsou odděleny svislými čarami.

Pokud je textový soubor tvořen automaticky – jde například o výstup z nějakého programu – a na úvod dokumentu se zobrazuje hlavička identifikačních údajů a potom teprve samostatná data, máme možnost nastavit přeskočení prvních ***n*** řádků souboru (volba ***Počet případů k přeskočení***). Dále je důležité si uvědomit, zda proměnné mají nějaký název – většinou chceme načíst tyto názvy jako záhlaví tabulky, proto i defaultní volba pro načtení souboru je ***Vzít jména proměnných z prvního řádku***.

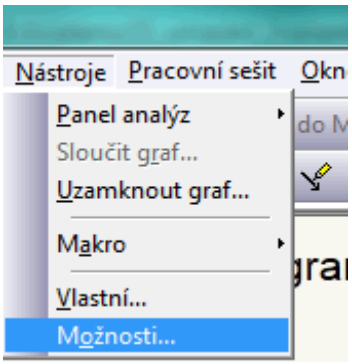
Zkontrolujeme také oddělovač desetinných míst, *STATISTICA* používá nastavení oddělovače pro Windows, tj. pokud otevíraný soubor vznikl například ve skriptu pro Linux systém, může být kódování desetinných míst tohoto souboru odlišné.

V tabulce náhledu můžeme myší vybrat konkrétní sloupec – proměnnou. Tím aktivujeme střední část menu ***Možnosti proměnné***. Nyní lze nastavit jméno proměnné, nastavit datový typ anebo zvolený sloupec vyloučit z načítání.

Po nastavení všech parametrů potvrdíme ***OK***. Výsledkem je otevření tabulky formátu *.sta* ve *STATISTICA*:

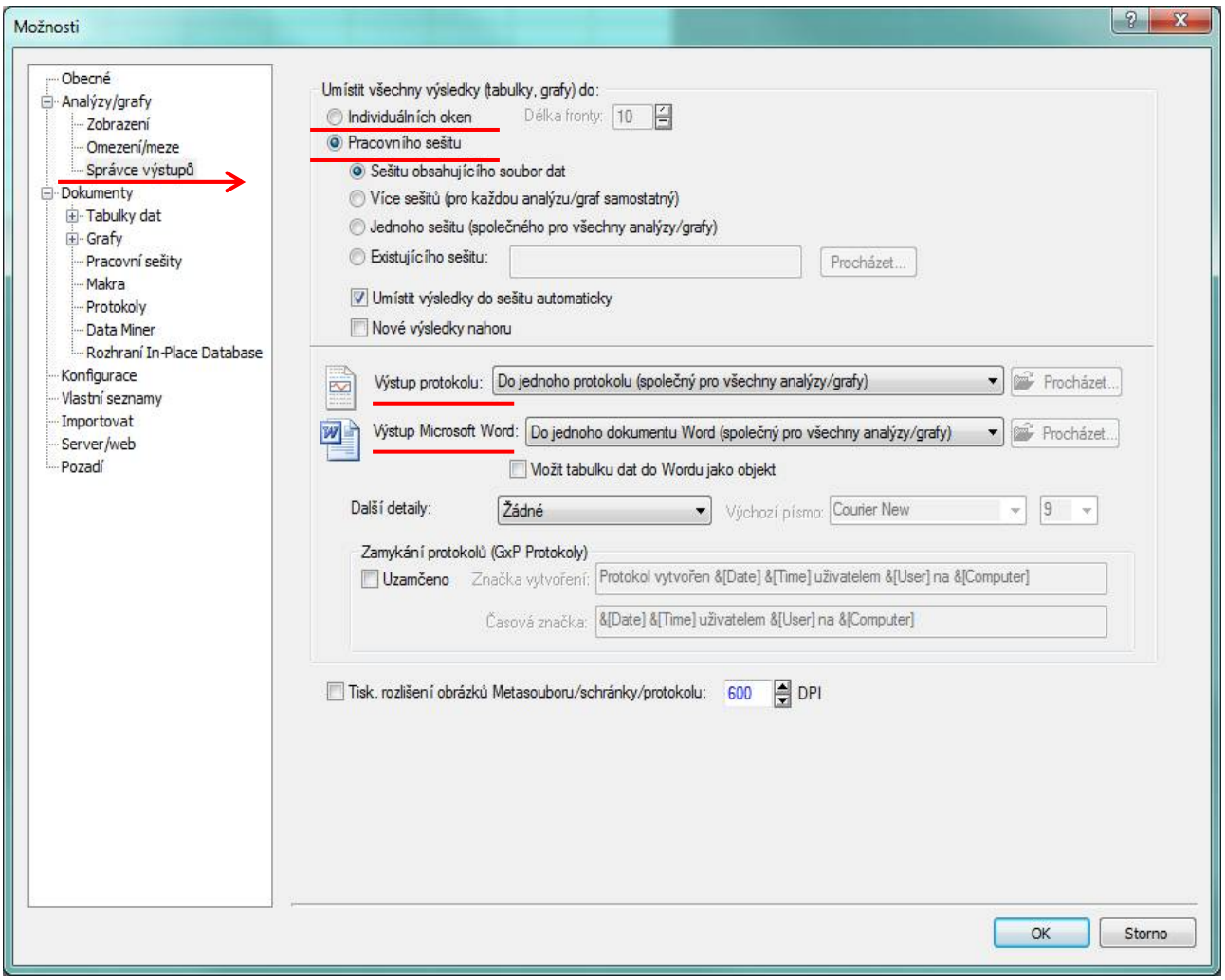
	1 datum	2 teplota
1	1.1.2010	-5.10
2	2.1.2010	-3.30
3	3.1.2010	-7.20
4	4.1.2010	-2.20
5	5.1.2010	1.20
6	6.1.2010	2.30
7	7.1.2010	2.00
8	8.1.2010	1.50
9	9.1.2010	0.30
10	10.1.201	4.00

8 Správce výstupů



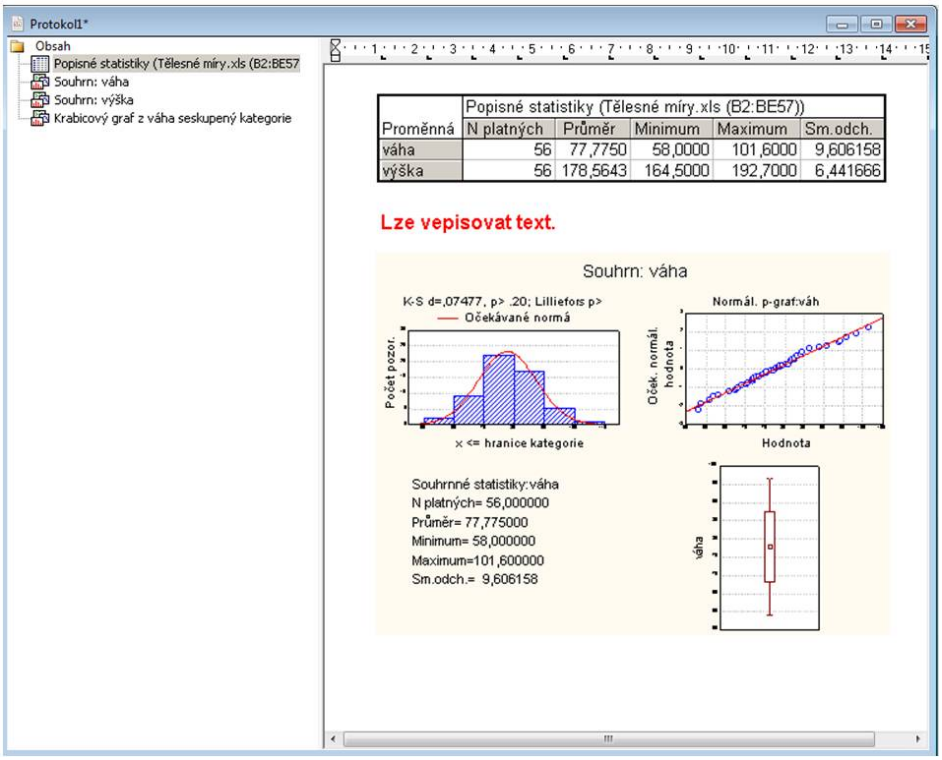
8.1 Výstup do Microsoft Word / do protokolu *STATISTICA*

V programu *STATISTICA* můžeme nastavit, v jakém formátu se budou ukládat výstupy. Ze základní nabídky vybereme ***Nástroje - Možnosti…***. Otevře se dialog ***Možnosti***, ve kterém přejdeme na záložku ***Správce výstupů***:



Můžeme zvolit některé z těchto možností:

- *individuální okna* - každá tabulka či graf se zobrazuje v samostatném oknu. Jednotlivá okna pak lze uložit ve formátu programu *STATISTICA* nebo v jiném formátu podle toho, zda se jedná o tabulku nebo graf. Pomocí nabídky ***Soubor – Uložit*** můžeme vybrat formáty **.xls*, **.txt*, **.htm*, **.pdf*, **.wmf*, **.jpg*, **.gif* atd.
- *pracovní sešit* - standardní formát výstupů v programu *STATISTICA* s příponou **.stw*. Právě v tomto formátu máme nyní výstupy z výše uvedených příkladů (pokud jsme neměnili výchozí nastavení). Okno pracovního sešitu je rozděleno na dvě části. Levá část zobrazuje stromovou strukturu (obdoba Průzkumníka). Pravá část je editorem vybraných dokumentů.
- *protokol* - má podobný vzhled jako pracovní sešit. V jeho levé části se zobrazuje seznam objektů protokolu. Pravá část je obdobou textového editoru. Na rozdíl od pracovního sešitu lze do protokolu mezi jednotlivé výstupy vepisovat text (viz následující ilustrační obrázek).



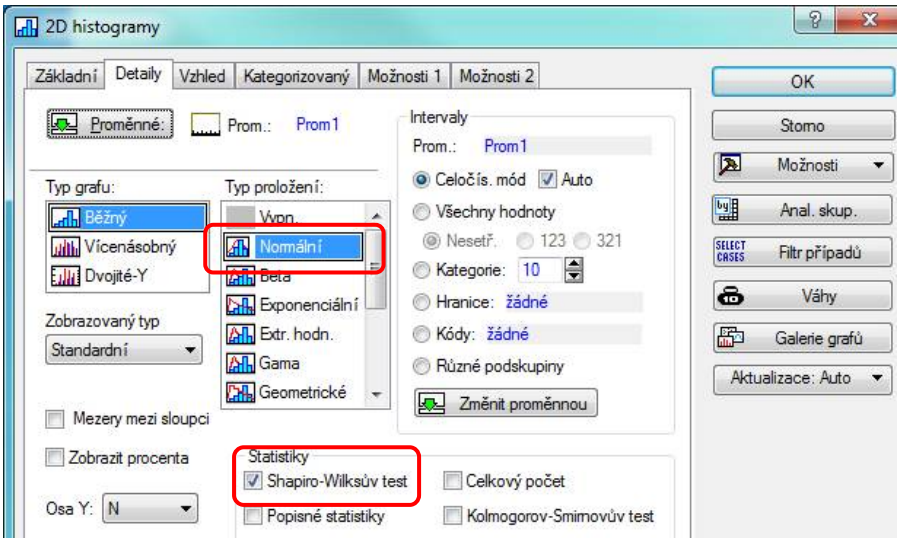
- *výstup do Microsoft Word* – výstupy se vkládají do dokumentu Microsoft Word, a mohou tak být jednoduše sdíleny s dalšími spolupracovníky.

9 Ověření normality v softwaru *STATISTICA*

Jedním ze základních předpokladů mnoha statistických analýz je normalita. Pokud některý test či metoda normální rozdělení předpokládá, je nutné to nejprve ověřit. K ověření lze použít mj. i statistické testy. Než však k testování normality přistoupíme, je dobré se zamyslet, zda se vůbec dá očekávat, že data jsou výběrem z normálního rozdělení. Pokud např. sledujeme platy obyvatelstva, víme, že nejsou omezené shora, zato jsou zdola omezené minimální mzdou, a rozhodně nejsou symetricky rozdělené kolem průměru. Takže prostou úvahou vyloučíme normalitu, aniž by bylo třeba provádět jakékoliv testy. Naopak u mnoha veličin, jako třeba byla v předchozím případě výška, je už z předchozích zkušeností známo, že se normálním rozdělením řídí. Potom testování také není nezbytné.

K ověřování normality systém *STATISTICA* poskytuje následující nástroje:

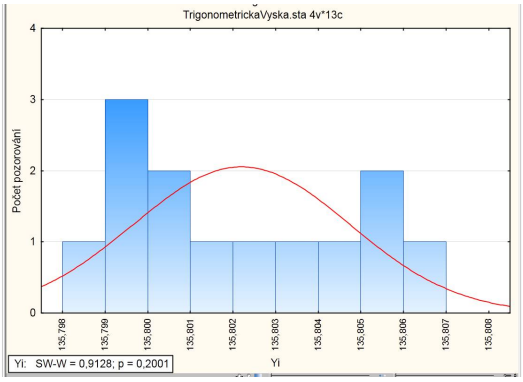
1. ***Histogram*** – vytvoříme histogram sledované proměnné a vizuálně ho porovnáme s normálním proložením:



Zajímavý článek o tomto tématu naleznete zde:

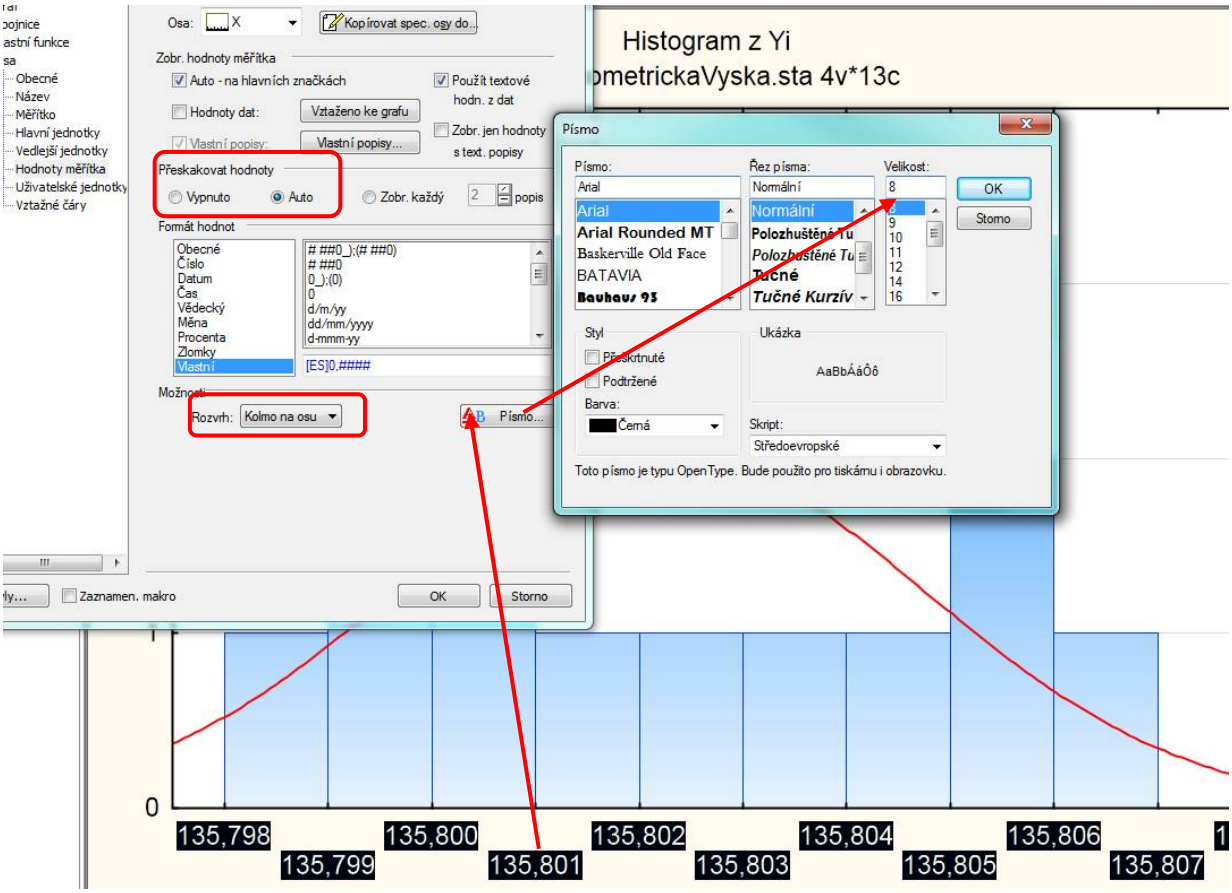
http://www.statsoft.cz/file1/PDF/newsletter/2013_10_09_StatSoft_Jak_se_pozna_normalita_pomoci_grafu.pdf

Doplňkově si lze zaškrtnou Shapiro-Wilkův test pro otestování normality, v tomto konkrétním případě jsme nezamítli nulovou hypotézu o normalitě (P (0,2>0,05):



Pozn.: Pokud použijeme K-S test, P hodnota se zobrazuje intervalem, pro přesnou P hodnotu využijte modul Rozdělení a simulace (viz níže modul Rozdělení a simulace).

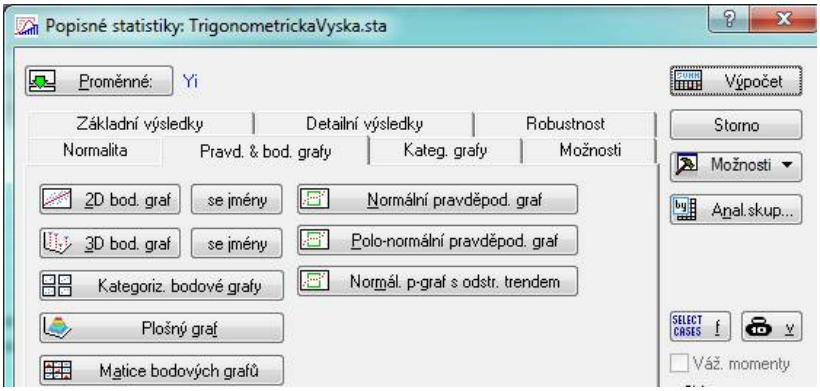
Dvojklikem do grafu vyvoláme dialog **Možnosti grafu** a graf si upravíme:



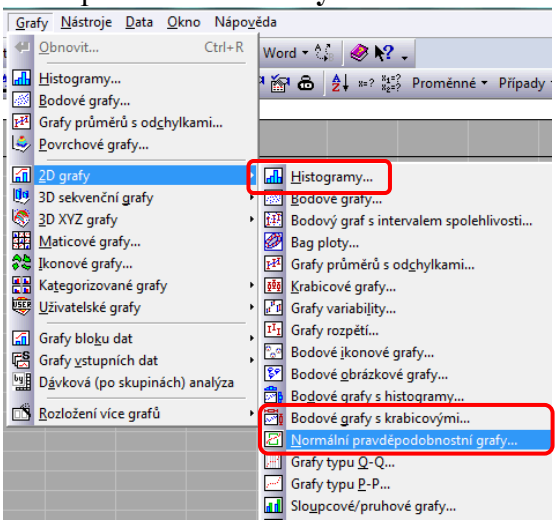
Nevyváženy počet dat v jednotlivých intervalech nemusí nutně znamenat významné odchylky od normality, a proto je vhodnější použít **kvantilové grafy:**

2. ***Normální pravděpodobnostní graf*** – jde o bodový graf, který porovnává kvantily spočtené z dat (osa x) s kvantily standardizovaného normálního rozdělení (osa y). Pokud veličina má normální rozdělení, leží body grafu na přímce. Tyto grafy lze vytvořit z nabídky ***Statistika - Základní statistiky/tabulky - Popisné statistiky - Pravděpodobnostní & bodové grafy***. Kromě ***Normálního pravděpodobnostního grafu***

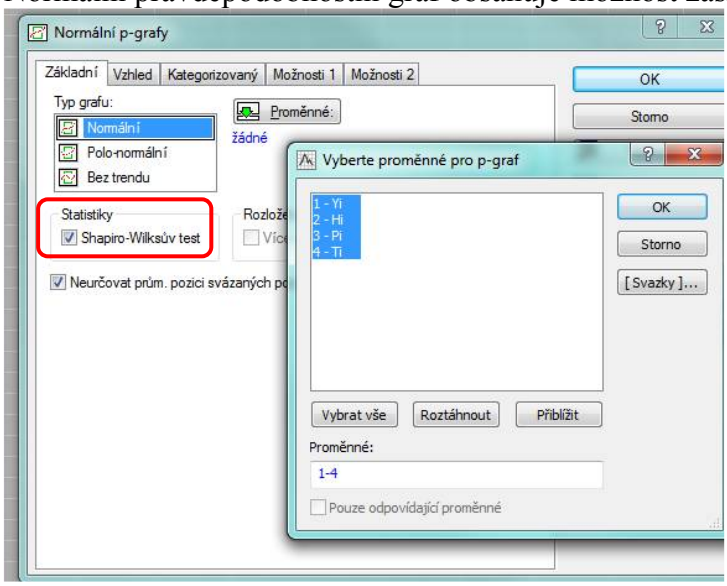
STATISTICA nabízí ještě ***Polo-normální pravděpodobnostní graf*** (obsahuje jen kladné hodnoty normálního rozdělení) a ***Normální pravděpodobnostní graf s odstraněným trendem*** (odstraněn lineární trend).



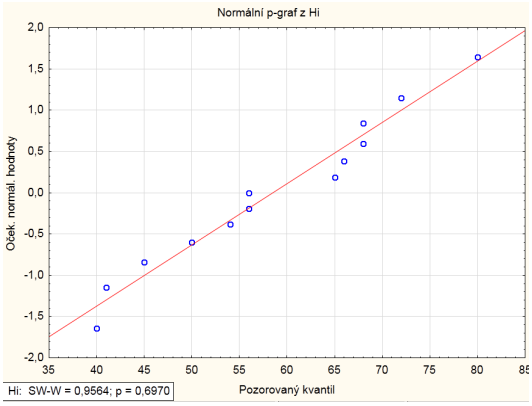
Nebo přes záložku **Grafy:**



Normální pravděpodobnostní graf obsahuje možnost zaškrtnout také Shapiro-Wilkův test:



Výsledný graf se statistikou SW testu:

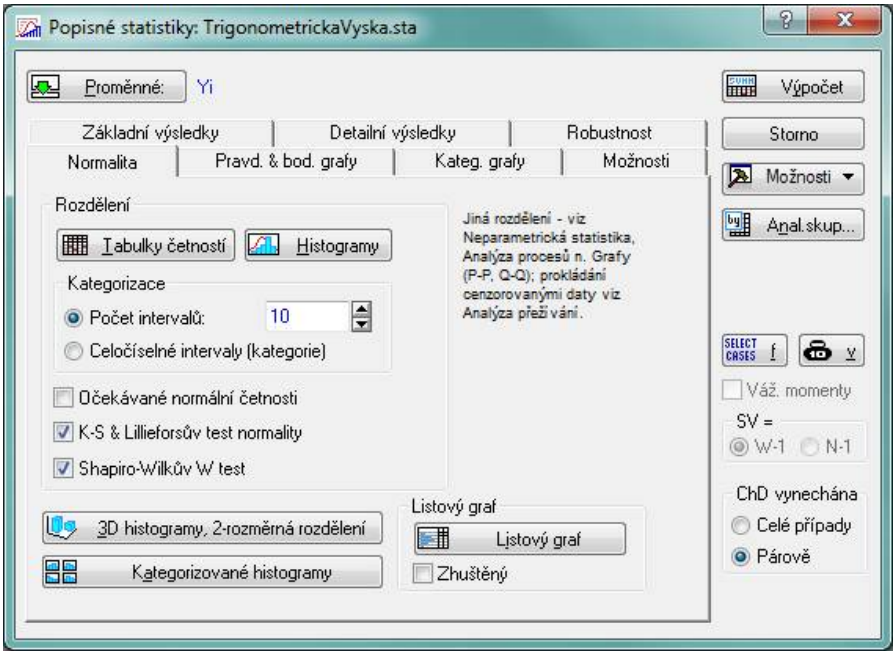


Zde nezamítáme nulovou hypotézu o normalitě *P (0,69) > 0,05*.

3. ***Testy***

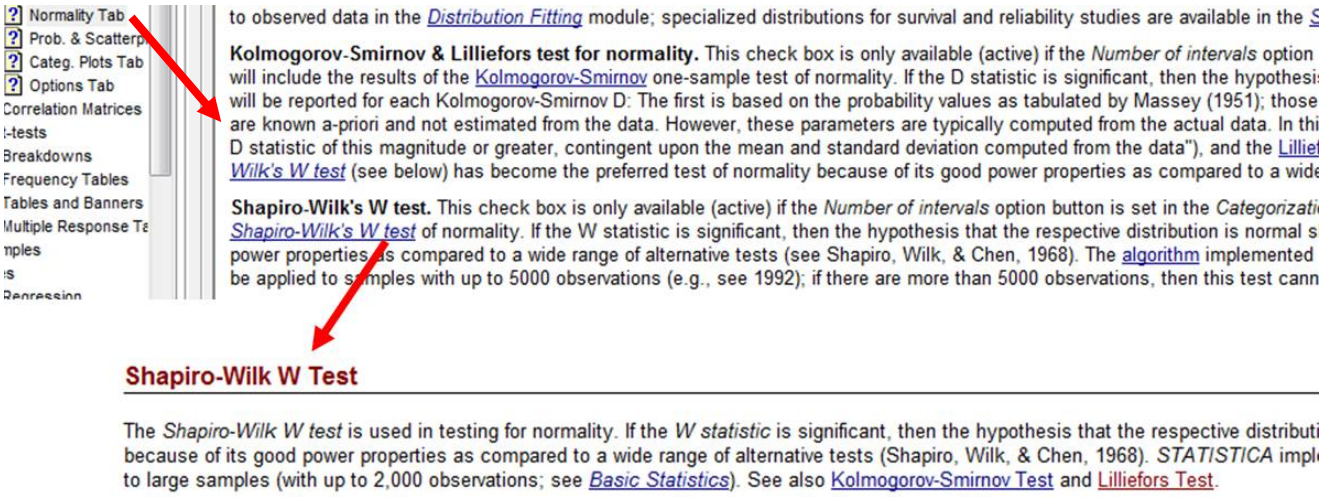
Kromě vizuálního ohodnocení jsou k dispozici také testy, které přímo s určitou pravděpodobností otestují, zda jsou data výběrem z normálního rozdělení, či nikoli. *STATISTICA* nabízí testy, např. Shapirův –Wilksův, Kolmogorovův – Smirnovův a Lillieforsův, Anderson – Darling atd.

Přes ***Statistiky -> Základní statistiky a tabulky -> Popisné statistiky ->*** karta ***Normalita:***

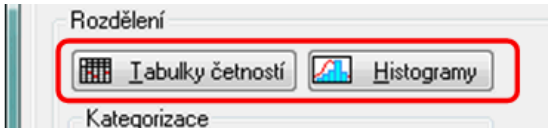


Jako nejjednodušší se doporučuje používat test Shapirův – Wilksův. Kolmogorovův – Smirnovův test se nedá použít přímo, protože předpokládá, že ověřujeme shodu našich dat s rozdělením, u kterého známe střední hodnotu a rozptyl. Ty se však většinou odhadují z dat samotných. Pro tento případ lze použít Lillieforsův test, který je modifikací Kolmogorovova – Smirnovova testu.

Klávesa ***F1*** v políčku pro zaškrtnutí příslušného testu vyvolá nápovědu k tématu a doporučení k jednotlivým testů:



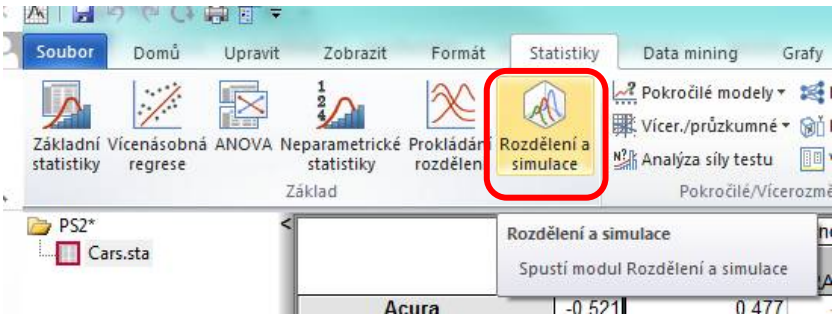
Shapiro-Wilkův test je zde upraven i pro relativně velké vzorky (5tis.). Po zaškrtnutí testu mám na výběr dvě možnosti reprezentace výsledku testu:

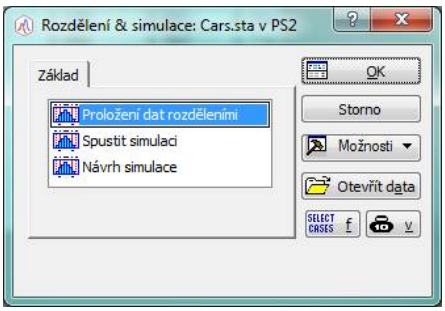


V modulu ***Statistika - Prokládání rozdělení*** se počítá test chí-kvadrát. Oboustranný či jednostranný T-test pro dva výběry pouze na základě statistik (průměry, směrodatné odchylky a rozsahy výběrů) je dostupný přes volbu ***Základní statistiky a tabulky – Testy rozdílů: r, %, průměry***.

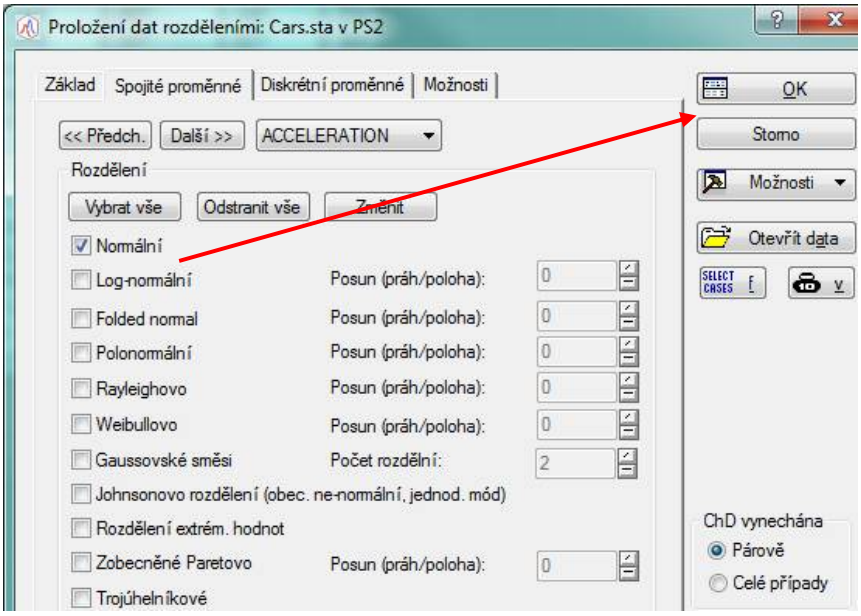
Modul Rozdělení a simulace

- Modul, který slouží přímo pro testování různých rozdělení je modul Rozdělení a simulace:

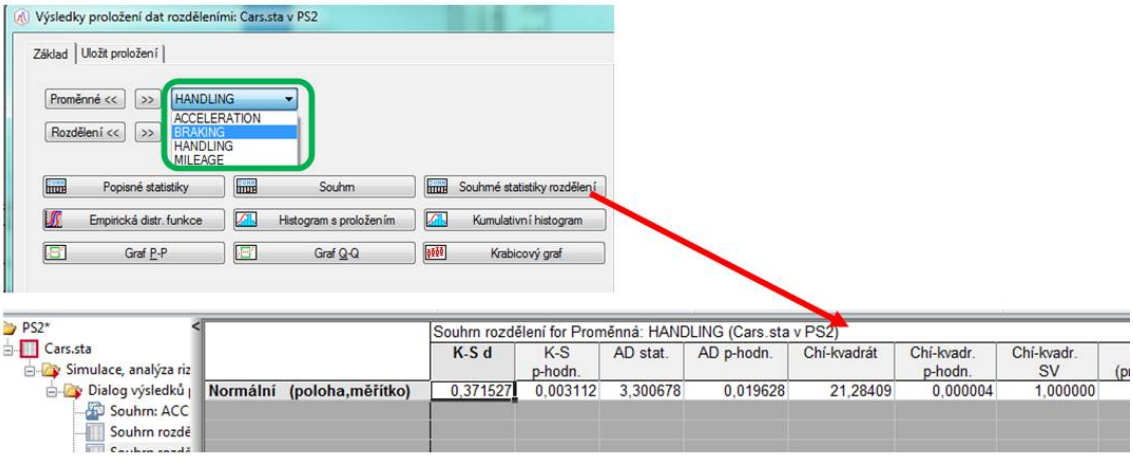




Na kartě ***Základ*** vybereme proměnné a přepneme na kartu ***Spojité proměnné***. Zde vybereme Normální rozdělení.



Volba konkrétního výstupu pro dané proměnné. Tlačítkem ***Souhrnné statistiky rozdělení*** získáme výstupy z testů normality:



Následující příklad slouží k ověření normality vybraných veličin:

Příklad - Normalita a důležitost náhodného výběru

Úkol: Vytvoříme novou tabulku s proměnnou, která bude mít normální rozdělení. Ověříme její vlastnosti a otestujeme, zda jde skutečně o normální rozdělení. Vytvoříme náhodný a nenáhodný výběr a porovnáme výsledky. Poté v souboru ***SpotřebaAut.sta*** ověříme normalitu u proměnných ***Zrychlení*** a ***Hmotnost***.

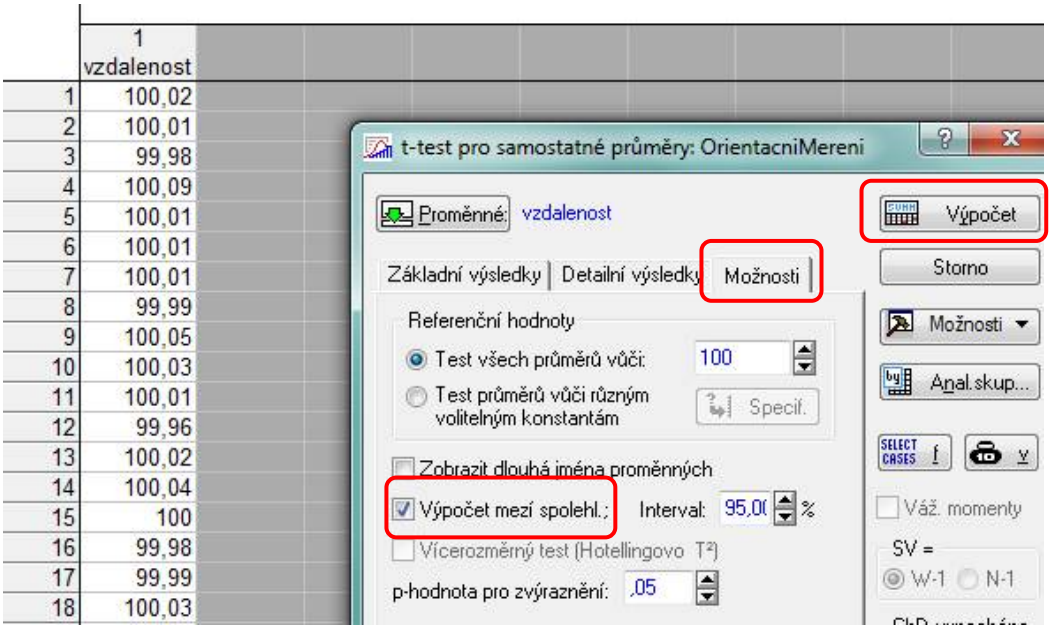
1. Vytvoříme novou tabulku o rozměrech *1s krát 1000 ř.* Zvolíme ***Soubor - Nový - Tabulka***. ***Počet proměnných*** *1* a ***Počet případů*** *1000*.
2. Poklepáním na záhlaví se otevře dialog ***Proměnná 1***, kam zadáme informace o proměnné: nazvěme ji Normální a do pole ***Dlouhé jméno*** vepíšeme funkci, která proměnnou vyplní. (Viz př. 2, bod 3.) *STATISTICA* disponuje funkcí ***RndNormal*** s parametrem *x*, který znamená směrodatnou odchylku. Pokud je zaškrtnut Průvodce funkcemi, po napsání = a počátečního písmene funkce program nabízí různé možnosti. Můžeme poklepat na zvolenou funkci a ta se sama vepíše do pole. Poté si můžeme zvolit směrodatnou odchylku

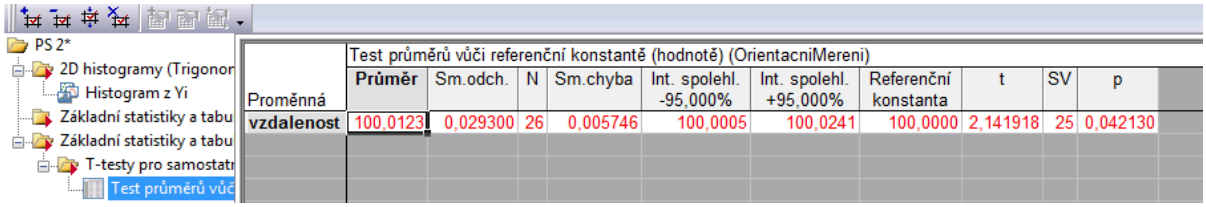
a po kliknutí na ***OK*** se vygeneruje 1000 náhodných čísel z normálního rozdělení o střední hodnotě *0* a zadané směrodatné odchylce.

3. Nyní můžeme provést příslušné testy: Spustíme ***Základní statistiky a tabulky - Tabulky četností***. Nejprve se podíváme na ***Histogramy*** na záložce ***Detaily***, kde zadáme, že chceme ***Přesný počet intervalů***, a to *10*. Vidíme, že rozdělení v histogramu odpovídá očekávanému normálnímu. Na záložce ***Normalita*** zadáme, že chceme ***Shapirův-Wilksův W test***. Ve výsledné tabulce máme vysokou hodnotu *p*, takže nemůžeme zamítnout, že by data nepocházela z normálního rozdělení. Na záložce ***Popisné*** zvolme ***Normální pravděpodobnostní grafy***. Na něm se body vyskytují na přímce.
4. Na záložce ***Detaily*** dialogu ***Základní statistiky a tabulky - Popisné statistiky*** kromě nabídnutých možností zaškrtněme ještě ***Šikmost*** a ***Špičatost***. a volme ***Výpočet: Popisné statistiky***. V tabulce vidíme, že rozdělení je *symetrické* (šikmost je přibližně 0) a *normálně špičaté* (špičatost také přibližně 0).
5. Soubor vygenerovaných náhodných čísel z normálního rozdělení budeme považovat za celou populaci. Známe její průměr a směrodatnou odchylku. Nyní vytvoříme podsoubor čítající přibližně 50 hodnot z této populace. Volíme ***Data - Náhodné vzorkování***. V záložce ***Možnosti*** vybereme ***Výpočet pomocí přibližného počtu***. Na kartě ***Jednoduché vzorkování*** zvolíme *50* jako ***Přibližný počet případů***. Tím se vytvoří nová tabulka s výběrem. Pokud porovnáme popisné statistiky u populace a výběru, shledáváme, že náš výběr slouží jako dobrý odhad pro celou populaci.
6. Nyní původní data setřídíme podle velikosti. Volíme ***Data - Setřídit***. Tím se data po ***OK*** setřídí. Pomocí funkce ***Data – Podmnožina*** vytvoříme filtr, který vybere prvních 50 případů (klikneme na ***Případy***, povolíme filtr a v části ***Zahrnout*** zadáme čísla případů 1-50). Tím jsme provedli nenáhodný výběr z dat. Pokud nyní porovnáme popisné statistiky u výběru i populace, vidíme, že by naše závěry byly silně zkreslené. Při zkoumání normality výběru se totiž ukáže, že výběr není výběrem z normálního rozdělení.
7. Otevřeme soubor ***SpotřebaAut.sta***.
8. Spustíme ***Statistika - Prokládání rozdělení***. Zvolíme ***Normální***. Nastavíme ***Proměnnou Zrychlení***. Pak už jen dáme ***Graf pozorovaného a normálního rozdělení***. Na histogramu vidíme shodu s normálním rozdělením, stejně tak chí-kvadrát test ji nezamítá. Ještě by nás zajímal pravděpodobnostní graf. Ten je např. v modulu ***Základní statistiky a tabulky – Tabulky četností - Popisné***. I na něm je vidět jasná shoda.
9. V případě ***Hmotnosti*** vidíme, že histogram neodpovídá normálnímu rozdělení. chí-kvadrát test ji také zamítá. Podíváme-li se na pravděpodobnostní graf, vidíme esovité zakřivení, stejně tak šikmost (0,53) naznačuje pravostranné zešikmení. Tato data nemůžeme považovat za výběr z normálního rozdělení.

10 Jednovýběrový t test

Přes ***Statistiky -> Základní statistiky/tabulky -> t-test, samost. vzorek*** se pak dostaneme k jednovýběrovému t-testu, kde definujeme referenční konstantu a klikneme na ***Výpočet:***





Test je signifikantní, zamítáme nulovou hypotézu: *H0:* $\mu = \mu_0 = 100m$

Skutečná průměrná naměřená vzdálenost přístroje se s 95% P nachází v intervalu:

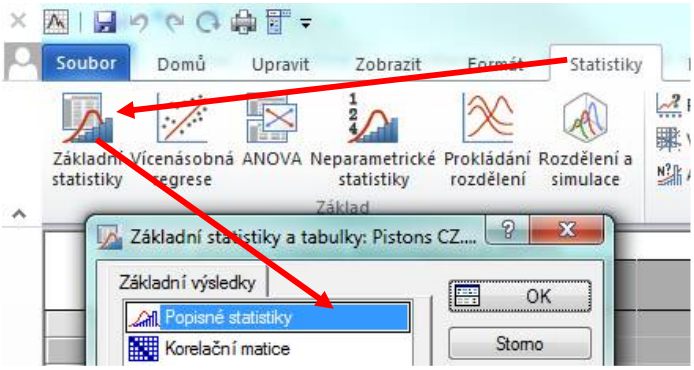
Test průměrů vůči referenční konstantě (hodnotě) (OrientacniMereni)		
	Int. spolehl. - -95,000%	**Int. spolehl. - +95,000%**
vzdalenost	100,0005	100,0241

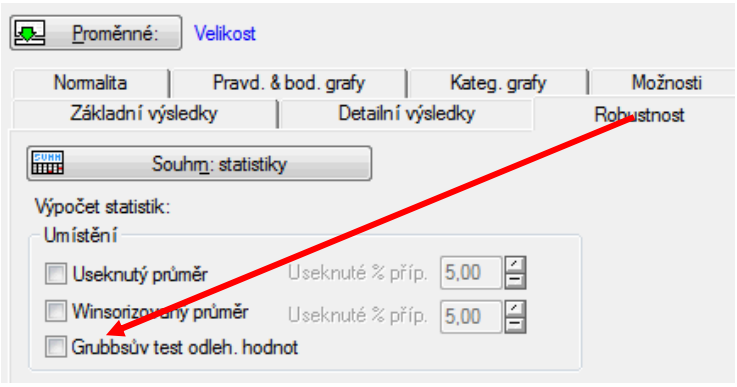
Je rozdíl také prakticky významný? Má přístroj sys. chybu?

Kompletní řešený příklad na tento test lze najít v našem newsletteru z 08/01/2013 *StatSoft ACADEMY:*
http://www.statsoft.cz/file1/PDF/newsletter/2013_01_08_StatSoft_Test.pdf

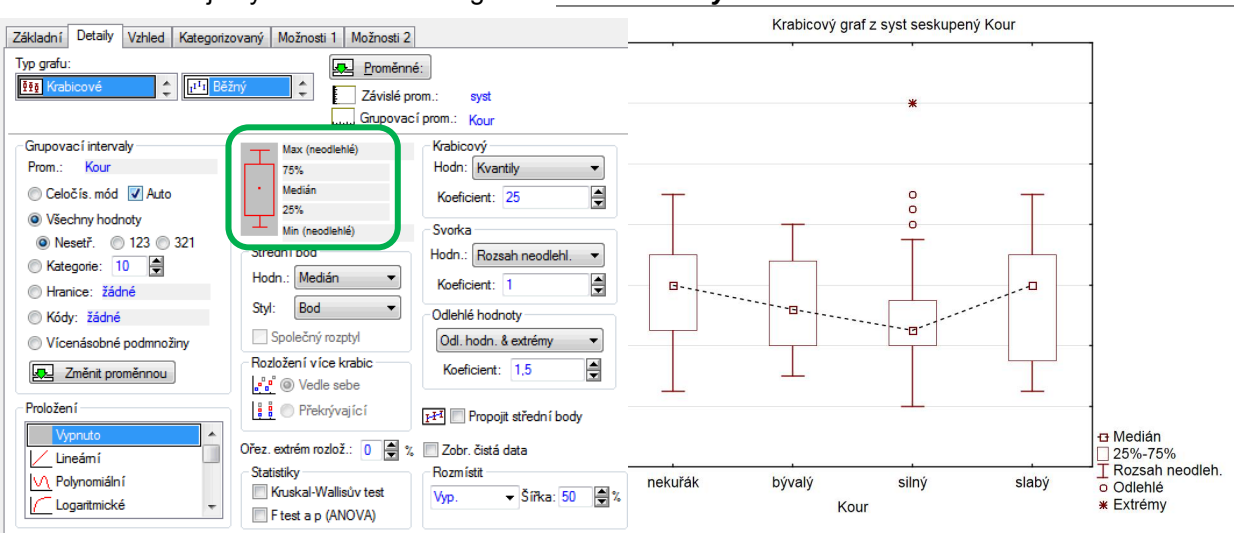
11 Testy odlehlých hodnot

Pro objektivní vylučování extrémních hodnot na základě vypočteného testovacího kritéria u souborů dat, které odpovídají Normálnímu rozdělení náhodné veličiny, je v softwaru implementován Grubbsův test





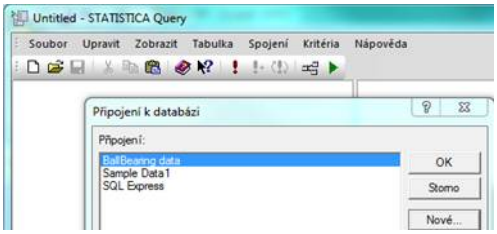
Další možností je využití krabicového grafu v záložce ***Grafy***:



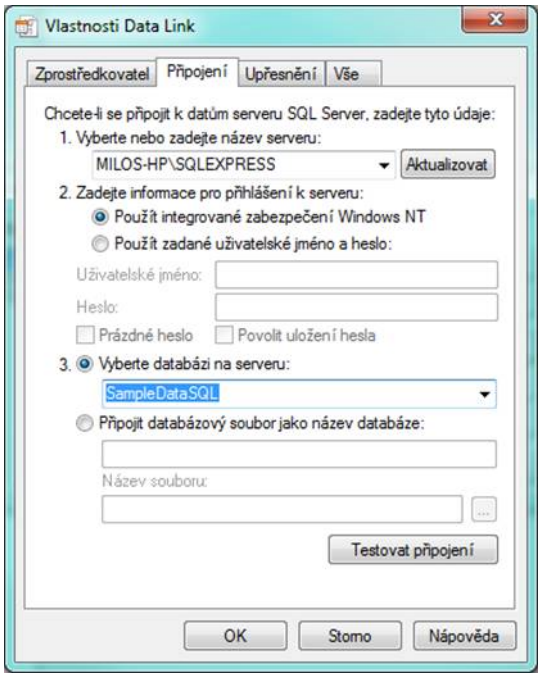
12 Připojení do databází pomocí *STATISTICA Query*

STATISTICA umožňuje přímé připojení do všech standardních databází přes konvence OLE DB a ODBC. Připojení probíhá v několika fázích:

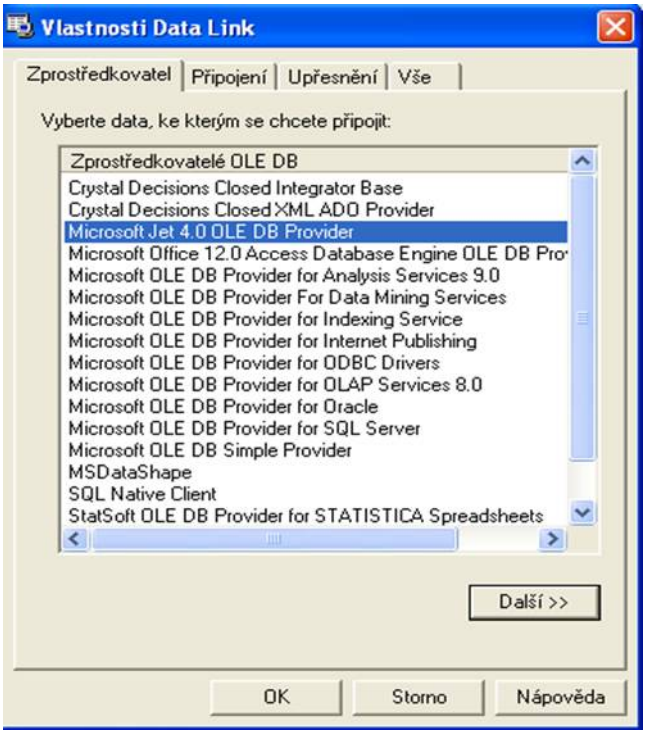
Přes ***Soubor*** - ***Získat externí data*** - ***Vytvořit dotaz*** se dostaneme do okna rozhraní ***STATISTICA Query***:



typů ovladačů, resp. databází, musíme cestu zadat ručně (např. *Access - Jet.OLEDB*.4.0). Dále zvolíme typ zabezpečení pro přístup do databáze a v rolovacím menu vybereme konkrétní databázi na serveru, který jsme definovali předchozím kroku. Vhodné je také otestovat připojení a v dalším kroku zvolíme název pro nové připojení, máme možnost zobrazit náhled připojovacího řetězce.



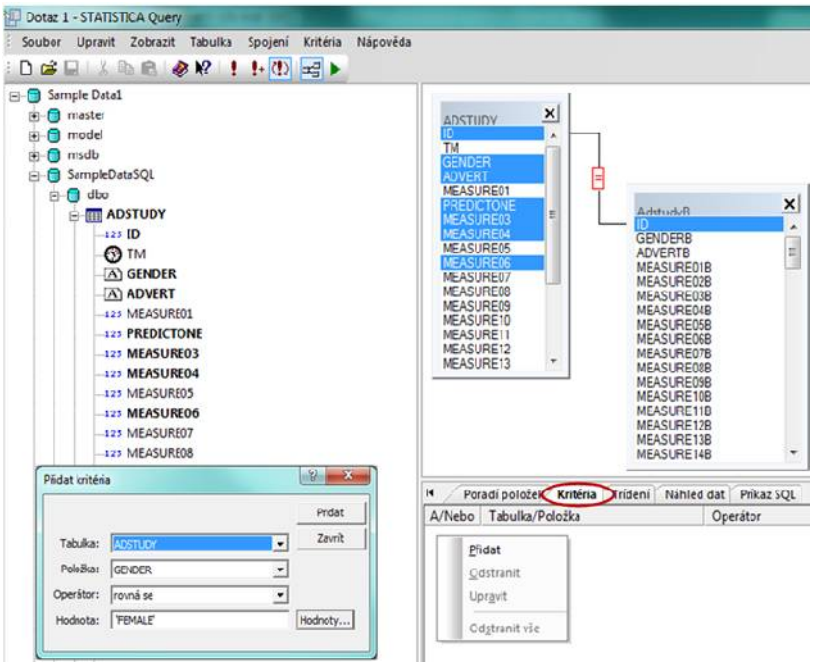
Zde tlačítkem ***Nové*** zvolíme možnost definovat nové připojení. V okně ***Vlastnosti Data Link*** vybereme vhodnou možnost z dostupných ovladačů pro připojovanou databázi:
V dalším kroku vybereme server, u některých



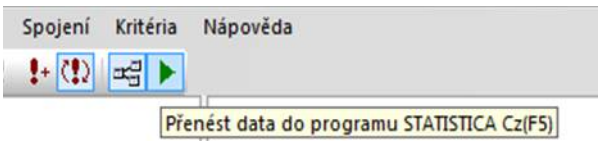
Práce v rozhraní *STATISTICA Query*

V rozhraní *STATISTICA Query* lze pracovat dvěma způsoby. První způsob využívá **grafický režim** a umožňuje práci i těm, kteří potřebují z databáze získávat konkrétní data, ale nemají potřebné znalosti dotazovacího jazyka SQL. Grafický režim funguje na principu „Táhni a pusť“. V levé části hlavního okna vidíme jednotlivé tabulky v databázi (na obrázku je to např. *ADSTUDY*), které lze přetáhnout do hlavního okna v pravé části menu. Kliknutím na jednotlivé názvy polí tabulky v hlavním okně (*ID*, *GENDER*…) vybereme, která pole z databáze chceme nahrát a automaticky tak již vytváříme SQL dotaz, který můžeme ve spodní části okna také nechat zobrazit (***Příkaz SQL***). Tlačítko ***Náhled dat*** umožňuje sledovat vybraná data.

Spojení tabulek je převzato z databáze, anebo jej lze nadefinovat přímo v prostředí *STATISTICA Query*, a to přetažením kurzoru z jedné tabulky na druhou (na konkrétním parametru, který slouží jako primární klíč), nebo přes záložku ***Spojení*** – ***Přidat***. Možnost přidat spojení vyvoláme také kliknutím pravého tlačítka myši ve volném prostoru hlavního okna. Kliknutí ve spodní části rozhraní *STATISTICA Query* (viz následující obrázek) vyvoláme možnost přidání doplňkových **omezení** pro jednotlivé parametry.



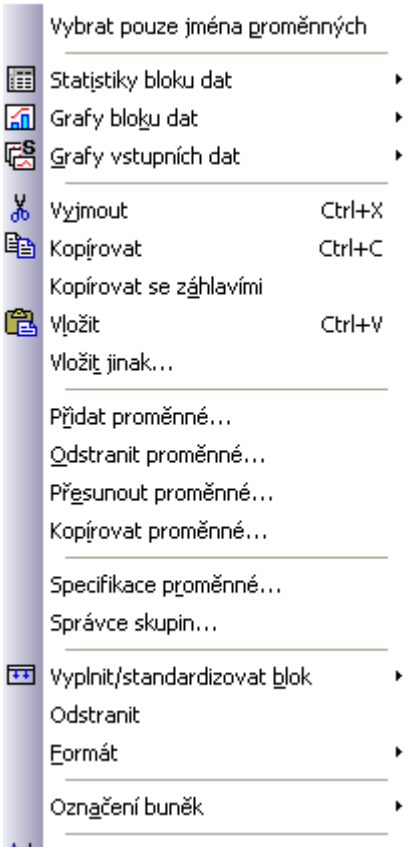
Chceme-li upřesnit již vygenerovaný SQL dotaz či napsat nový bez využití grafického módu, přes záložku ***Zobrazit*** přepneme ***grafický režim*** na skriptovací.
Přes záložku ***Soubor*** – ***Uložit jako/Otevřít*** lze hotové dotazy ukládat a načítat. Samotné spuštění dotazu probíhá přes zelenou ikonu v horní liště, nebo přes klávesu ***F5.***



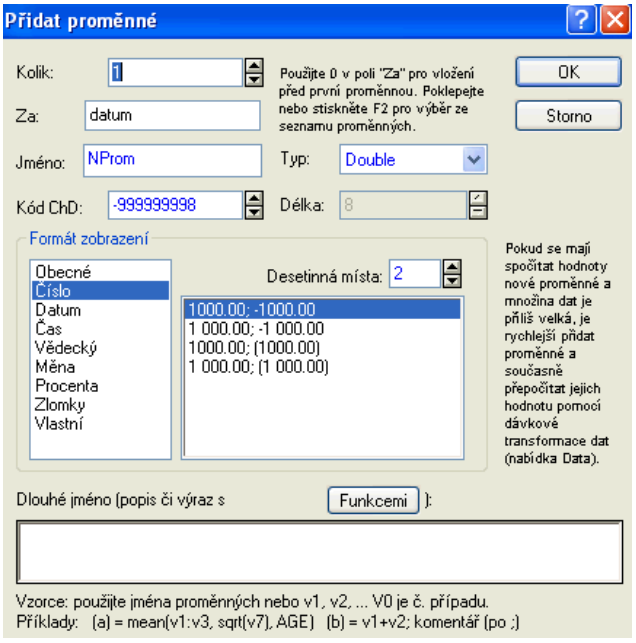
Defaultní nastavení *STATSITICA* je načítat data do aktivní tabulky dat, pokud chcete načíst data do nové prázdné tabulky, vyberte tuto možnost v následujícím dialogu:

13 Úprava načtených dat

Proměnné a případy



Přidání a odebrání proměnných provedeme následujícím způsobem: V záhlaví tabulky klikneme pravým uchem myši a zobrazíme dialog, v němž můžeme vybrat možnost ***Odebrat proměnné…*** nebo ***Přidat proměnné***.
Při **přidávání proměnných** se zobrazí dialog, v němž



uživatel specifikuje počet přidávaných proměnných, název proměnné, ze kterou se mají nové proměnné vložit, jméno proměnné (Pokud přidáváme více než jednu proměnnou, bude zadaný název použit u všech těchto proměnných – pro odlišení bude ukončen pořadovým číslem přidávané proměnné. Přejmenování proměnných můžeme nicméně provést následně.), typ hodnot proměnné a způsob zobrazení jejích hodnot. Rozlišujeme čtyři typy hodnot proměnných, a sice:

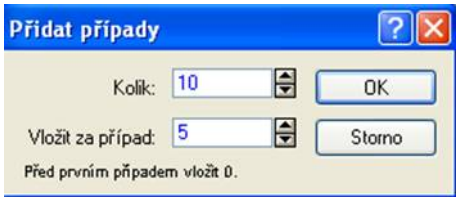
- **Double**
  Defaultní typ. Využívá se pro numerické hodnoty a umožňuje ukládat 64 bitová reálná čísla s přesností na 15 desetinných míst. Rozsah přibližně od $-1,7 * 10^{308}$ do $1,7 * 10^{308}$. Kód *chybějících dat* je -999999998.

- **Integer**
  Celá čísla v rozmezí -2 147 483 648 a 2 147 483 647. Každé číselné hodnotě lze přiřadit textový popisek. Velikost *4 byty*.

- **Byte**
  Celá čísla v rozmezí 0 až 255, nelze vložit desetinná čísla, každé číselné hodnotě lze přiřadit textový popisek. Velikost *1 byte*.

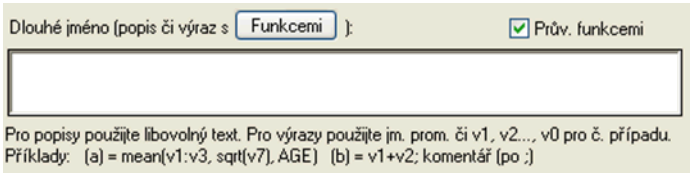
- **Text**
  Textové řetězce s neomezenou délkou bez číselné reprezentace. Pro účely numerických výpočtů jsou různým řetězcům přiřazeny *ad-hoc* různé číselné hodnoty. Kód *chybějících dat* je *prázdný řetězec*.

Přiřazením vhodného typu můžeme šetřit místo nutné pro uložení datové tabulky v paměti počítače.
Do okna dialogu pro přidání proměnných s názvem ***Dlouhé jméno*** je možné vkládat matematické, statistické, logické, textové ale i jiné funkce, jejichž vstupem jsou ostatní proměnné tabulky, nicméně vkládání těchto funkcí doporučujeme provádět až po přidání proměnných. Pokud se funkce odkazují na proměnné, které se v tabulce vyskytují až za přidávanými proměnnými, nejsou odkazy pomocí písmene *v* a čísla sloupce proměnné jednoznačné.
Při přidávání případů je potřeba zadat, jen kolik řádků chceme do tabulky přidat a za který řádek se mají vložit:



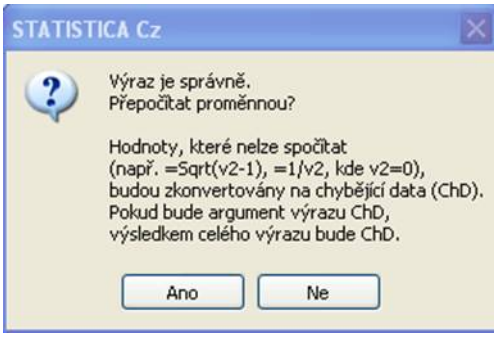
Transformace dat

Pro transformaci dat je ideální nadefinovat novou proměnnou, která bude funkcí proměnných původních. V záhlaví tabulky klikneme dvakrát na název nové proměnné a v dialogu podobném dialogu pro přidávání proměnných klikneme v dolní části na tlačítko Funkcemi.



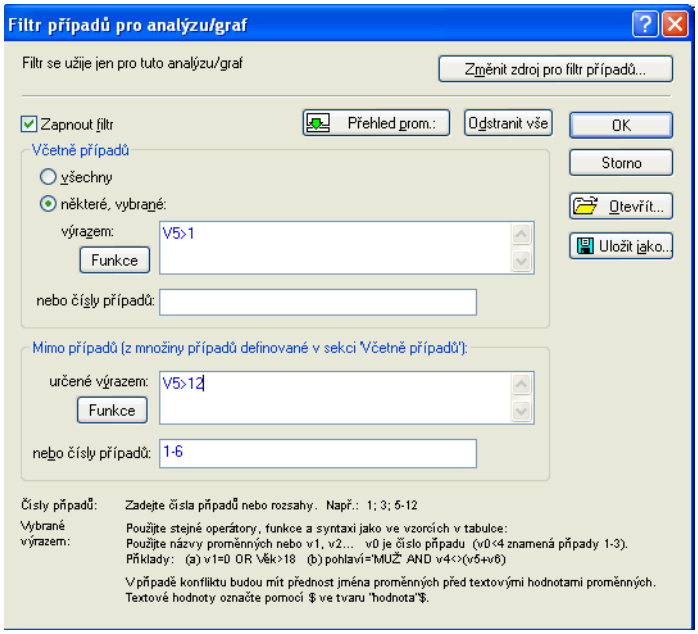
Zobrazí se ***Prohlížeč funkcí***, kde jsou dostupné všechny funkce, které jsou ve *STATISTICA* definovány. Můžeme je vybírat v levé části okna prohlížeče podle jejich typu, v pravé části okna potom vybereme konkrétní funkci a v dolní části okna se zobrazí nápověda k vybrané funkci (popis toho, co funkce dělá a jaké má vstupní parametry).
Odkaz na jiné proměnné tabulky se tvoří buď použitím názvu proměnné (pokud název obsahuje mezery, je třeba ho uvádět v uvozovkách) anebo užitím písmena *v* a čísla sloupce proměnné (například *v8* odkazuje na proměnnou v osmém sloupci tabulky). Výraz *v0* označuje pořadová čísla řádků (případů).
Zápis transformace pro novou proměnnou může vypadat například takto:

Potvrdíme volbu tlačítkem *OK*, *STATISTICA* zobrazí ještě dialog, kde odsouhlasíme přepočítání hodnot nové proměnné:



Použití filtru

Nejpohodlnější je nejspíš použití filtru při samotném volání analýzy nebo tvorbě grafu. V pravé části některého z úvodních dialogů je umístěno tlačítko ***SELECT CASES***.



Pomocí něj zobrazíme dialog, v němž je třeba zatrhnout možnost ***Zapnout filtr***. Tím se zpřístupní pole pro zadání podmínek pro zahrnutí nebo vyloučení některých řádků tabulky. Pro názornost uvádíme následující příklad zadání podmínek filtru.

Do analýzy budou zahrnuty případy, pro které je splněna podmínka, že hodnota proměnné v pátém sloupci je větší než 1, a vyloučeny budou řádky 1 až 6 a dále ty případy, které sice splňují podmínku

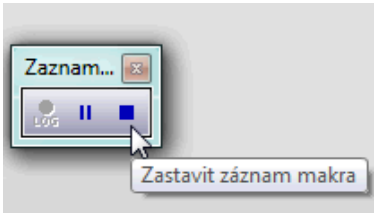
V5>1, ale u nichž je hodnota páté proměnné větší než 12.

14 Automatizace rutinních analýz

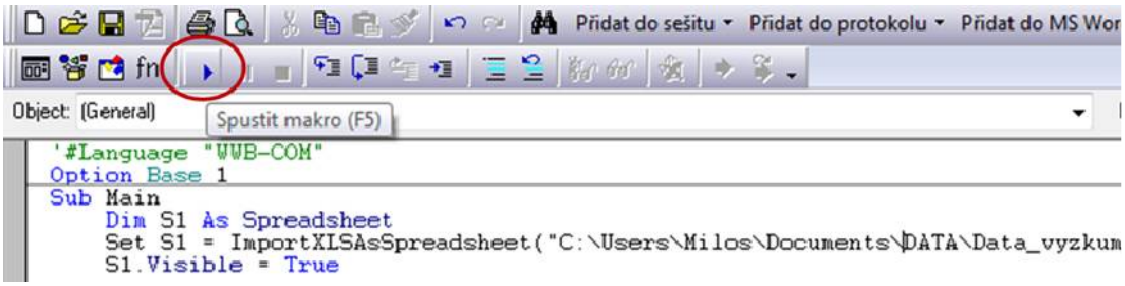
Následující postup ukazuje tvorbu jednoduchého makra pro automatizaci rutinních činností. Software *STATISTICA* umožňuje vytvářet různé dávkové analýzy pomocí integrovaného jazyka *STATISTICA Visual Basic* (SVB), který lze využít ke zjednodušení prováděných úloh různé obtížnosti, od jednoduchých maker až po pokročilé projekty. Pomocí jazyka SVB může uživatel přistupovat prakticky ke každému funkčnímu prvku systémů a tedy i využívat vlastní rozšíření systému.

Všechny postupně prováděné analýzy lze snadno automaticky zaznamenávat pomocí záznamu makra. Tímto jednoduchým záznamem potom zcela automatizujeme často se opakující analýzy, a to i bez znalosti programování. ***Postup tvorby záznamu makra je následující:***

Před vlastním spuštěním záznamu je třeba zvážit, zdali chceme provádět automatizovanou analýzu vždy nad již načtenou aktivní tabulkou *STATISTICA*, anebo bude načtení aktuálních dat také součástí kódu. V druhém ze zmíněných případů začneme nahrávat nejprve samotné otevírání příslušné tabulky.



Dále vybereme menu ***Nástroje - Makro - Spustit záznam průběhu analýzy (hlavní makro)***. Nyní provedeme požadovanou posloupnost analýz nebo vytvoříme grafy, které dále upravujeme a podobně. Záznam ukončíme kliknutím na tlačítko **Zastavit záznam makra** na minipanelu, který se otevřel v okamžiku spuštění nahrávání makra, anebo v menu ***Nástroje - Makro - Zastavit záznam.*** V následujícím dialogu si makro pojmenujeme a potvrdíme ***OK***. Nyní máme k dispozici zaznamenaný kód, který můžeme upravit a následně uložit prostřednictvím nabídky ***Soubor -> Uložit/Uložit jako***… Makro spustíme pomocí tlačítka ***Spustit makro***, které je dostupné na hlavním panelu v okamžiku, kdy je aktivní okno s kódem makra, případně můžeme použít klávesu ***F5***.



Všimněme si, že v příkladu zobrazeném na obrázku, je v kódu uložena cesta k souboru *Data_vyzkum*. Při spuštění makra proto bude vždy načtena aktuální verze tohoto souboru a analýzy se provedou nad aktuálními daty. Pokud bychom makro spustili již nad otevřenou tabulkou (Spreadsheet), v záznamu byl tento kód:

```
Dim S1 as Spreadsheet
Set S1 = ActiveDataSet
```

Makro by pak využívalo (a vyžadovalo) nějakou již otevřenou aktivní tabulku v aplikaci *STATISTICA*.

15 Analýza rozptylu

ANOVA

Analýza rozptylu je užitečná v situacích, kdy nás zajímá vliv jedné nebo více ***nominálních*** proměnných (též zvaných ***faktory***) na proměnnou ***kvantitativní***. Příkladem může být analýza velikosti tržeb v závislosti na ročním období, analýza účinků určitého léku u různých skupin pacientů, analýza mezd podle dle dosaženého vzdělání atd.
Zkoumáme-li závislost pouze na jednom faktoru, hovoříme o jednofaktorové analýze rozptylu. Celý soubor se rozčlení do příslušného počtu skupin (podle počtu úrovní faktoru) a předmětem zkoumání jsou potom střední hodnoty těchto skupin – jejich shoda či rozdílnost. Faktor může obecně nabývat libovolného počtu hodnot a testová hypotéza má pak tvar H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, čímž v podstatě říká, že sledovaná proměnná není závislá na úrovni faktoru a že při všech jeho úrovních nabývá zhruba stejných hodnot, přičemž rozdíly jsou způsobeny pouze náhodným kolísáním. Alternativní hypotéza tvrdí, že alespoň jedna z uvedených rovností neplatí.
Podstatou je, jak už název napovídá, rozklad rozptylu zkoumané (závislé) proměnné, a to jednak na část, která vzniká v důsledku skutečné rozdílnosti jednotlivých skupin, tzv. ***meziskupinový rozptyl***, a jednak na část zapříčiněnou náhodným kolísáním, tzv. ***vnitroskupinový (reziduální) rozptyl***. Testovým kritériem je pak podíl těchto složek. Pokud je meziskupinová variabilita dostatečně velká oproti reziduální, test vede k zamítnutí hypotézy o rovnosti středních hodnot.
Stejně jako regrese i analýza rozptylu je založena na obecném lineárním modelu. ANOVA je v podstatě součástí (speciálním případem) regrese.

Další návodné články k tomuto tématu naleznete v archivu newsletterů StatSoft Academy: http://www.statsoft.cz/o-firme/archiv-newsletteru/

- ✓ Anova dvojného třídění: http://www.statsoft.cz/file1/PDF/newsletter/2012_11_12_StatSoft_Analyza_rozptylu.pdf
- ✓ Neparametrická Anova: http://www.statsoft.cz/file1/PDF/newsletter/2013_06_04_StatSoft_Neparametricka_anova.pdf

Příklad – jednofaktorová ANOVA: Patnáct pozemků bylo náhodně rozděleno do tří skupin. Na dvou z nich byla použita hnojiva A a B, třetí skupina byla kontrolní bez hnojení. Určete, zda použité hnojivo má vliv na výnos obilí.
Stanovení hypotézy:

H_0: Použité hnojivo nemá vliv na výnos obilí.

H_1: Použité hnojivo má vliv na výnos obilí.

Test provedeme na 5% hladině významnosti.

1. Otevřeme datový soubor *Hnojiva.sta*. V prvních dvou sloupcích jsou uvedeny výnosy při použití hnojiv, ve třetím jsou výnosy z pozemků nehnojených. Poněvadž takto uspořádaný soubor neobsahuje žádnou proměnnou, která by označovala úroveň faktoru (tyto úrovně jsou uvedeny pouze v záhlaví), je potřeba data převést do vhodnějšího tvaru, který program *STATISTICA* očekává. To lze provést dvěma způsoby.
   První způsob:
   Vytvoříme nový soubor *Hnojiva (upraveny).sta* –› *Soubor – Nový - Tabulka dat*. Počet proměnných nastavíme na 2 a počet případů na 15. V nové tabulce vytvoříme proměnnou *Hnojivo* (faktor) a proměnnou *Výnos*. Kopírováním vložíme data. Původní a upravenou tabulku ukazují následující obrázky.

Data: Hnojiva.sta (3s krát 5ř)

Výnosy obilí	Hnojivo A	Hnojivo B	Bez hnojeni
1	71	69	65
2	68	74	62
3	73	72	57
4	73	71	63
5	69	74	60

Data: Hnojiva_uprav (2s krát

	1 Hnojivo	2 Výnos
1	A	71
2	A	68
3	A	73
4	A	73
5	A	69
6	B	69
7	B	74
8	B	72
9	B	71
10	B	74
11	zadne	65
12	zadne	62
13	zadne	57
14	zadne	63
15	zadne	60

Druhý způsob:
StejnéhoVhodnějšího tvaru dat lze dosáhnout rychleji a jednodušeji seskupením dat. *Data – Přeskupování…* - záložka *Seskupování*. Proměnné vybere všechny. Jméno cílové proměnné bude *Výnos* a jméno kódové proměnné *Hnojivo*. Potvrdíme tlačítkem *OK*. Upravenou tabulku ukazuje následující obrázek.

Data: Tabulka3* (2s krát 15ř)

Výnosy obilí	1 Výnos	2 Hnojivo
1	71	Hnojivo A
2	69	Hnojivo B
3	65	Bez hnojeni
4	68	Hnojivo A
5	74	Hnojivo B
6	62	Bez hnojeni
7	73	Hnojivo A
8	72	Hnojivo B
9	57	Bez hnojeni
10	73	Hnojivo A
11	71	Hnojivo B
12	63	Bez hnojeni
13	69	Hnojivo A
14	74	Hnojivo B
15	60	Bez hnojeni

2. Ověříme předpoklad normality dat.

K ověření normality zvolíme Shapirův-Wilkův test, který najdeme v záložce *Statistiky – Základní statistiky/tabulky - Popisné statistiky – Normalita.* Proměnná, kterou testujeme, je Výnos. Po stisknutí tlačítka *Tabulky četností* se nám spolu s tabulkou četností objeví i výsledky testu normality dat.

Kategorie	Tabulka četností:Výnos (Hnojiva (upravena)) **K-S d=,17211, p> .20; Lilliefors p> .20** **Shapiro-Wilk W=,90249, p=,10388** Četnost	Kumulativní četnost	Rel.četn. (platných)	Kumul. % (platných)	Rel.četn. všech	Kumul. % všech
55,00000<x<=60,00000	2	2	13,33333	13,3333	13,33333	13,3333
60,00000<x<=65,00000	3	5	20,00000	33,3333	20,00000	33,3333
65,00000<x<=70,00000	3	8	20,00000	53,3333	20,00000	53,3333
70,00000<x<=75,00000	7	15	46,66667	100,0000	46,66667	100,0000
ChD	0	15	0,00000		0,00000	100,0000

3. Předpoklady homogenity rozptylů.

Analýzu spustíme volbou *Statistiky - Základní statistiky/tabulky – Rozklad & Jednofakt. ANOVA*. Zvolíme *Proměnné*: *Závislá proměnná* je *Výnos*, *Grupovací proměnná* je *Hnojivo*. Přepneme na záložku *Skupiny tabulek* a v oddělení *Výstupní tabulky* vybereme *Celková tabulka průměrů*, *Analýza rozptylu*, *Leveneův test a Brown & Forsythe (HOV)*. Po stisknutí *Výpočet* je třeba znovu zadat *grupovací proměnnou Hnojivo*. Znovu klikneme na tlačítko *Výpočet* a dostaneme požadované výstupy. Výše zmiňované testy homogenity rozptylů ani zde neprokázaly rozdíl rozptylů mezi jednotlivými skupinami (jejich p hodnoty přesahují 0,05). Předpoklad homogenity je tedy splněn.

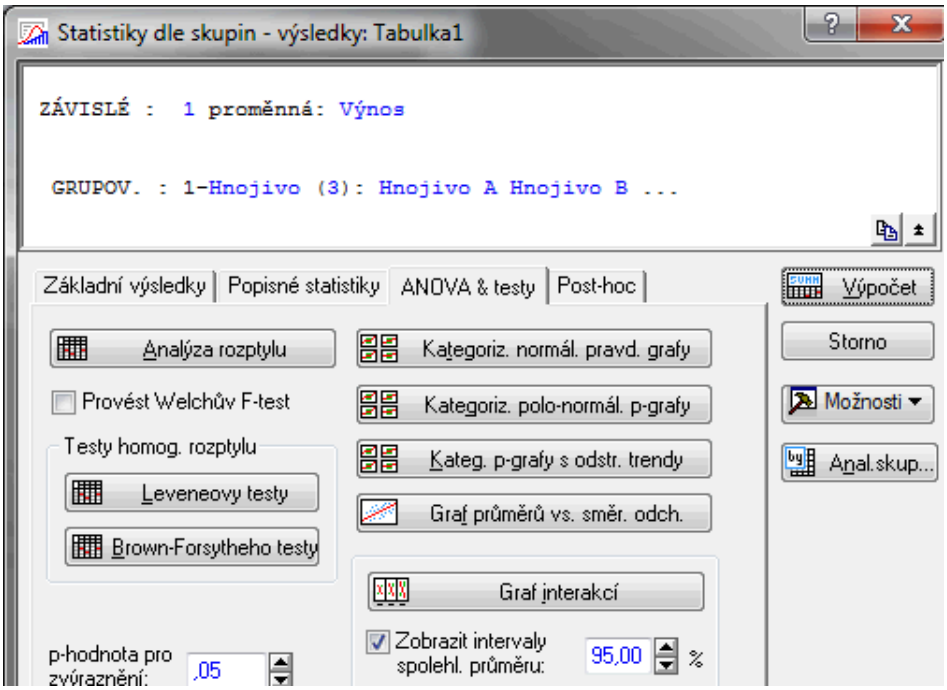
4. *Tabulka průměrů* ukazuje průměry a rozptyly v jednotlivých skupinách. Výstup vlastní analýzy rozptylu ukazuje následující obrázek.

Analýza rozptylu (Hnojiva_uprav)
Označ. efekty jsou význ. na hlad. p < ,05000

Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Výnos	336,9333	2	168,4667	76,00000	12	6,333333	26,60000	0,000039

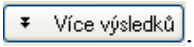
Na základě této tabulky můžeme tvrdit, že zamítáme hypotézu o stejných středních hodnotách. Mezi skupinami je statisticky významný rozdíl a výnos tedy závisí na použitém hnojivu.

Alternativním způsobem, jak spustit předešlé analýzy je i následující postup: *Statistiky - Základní statistiky/tabulky – Rozklad & Jednofakt. ANOVA*. Zvolíme *Proměnné*: *Závislá proměnná* je *Výnos*, *Grupovací proměnná* je *Hnojivo*. Místo zvolení záložky *Skupiny tabulek* potvrdíme výběr proměnných tlačítkem OK. Zobrazí se následující tabulka:

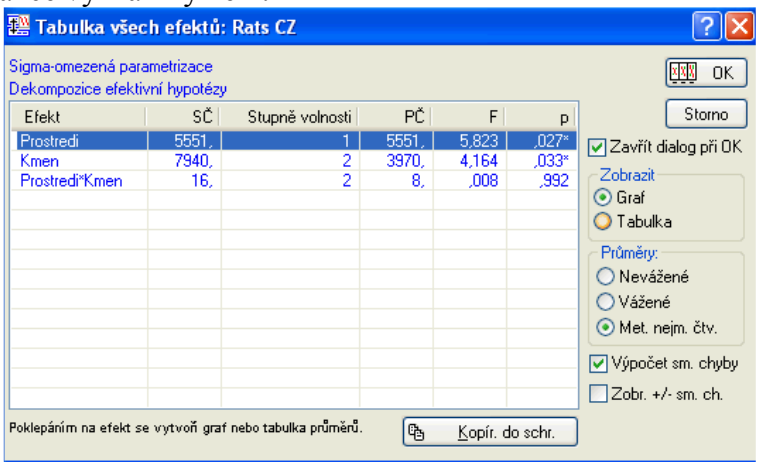


Touhle cestou se dostaneme nejen k výběru výpočtů analýzy rozptylu a testů homogenity rozptylů, ale také k různým možnostem vizualizace dat.

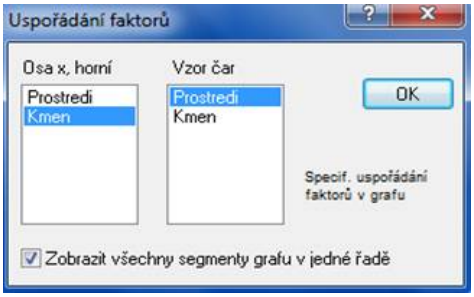
Příklad – dvojné třídění: Máme tři kmeny krys, jejichž obecná schopnost úspěšně se pohybovat v bludišti by se dala popsat jako dobrá, nestálá, nebo špatná. Čtyři krysy z každého kmenu byly vychovávány ve stimulujícím prostředí, čtyři v prostředí omezeném. Cílem je určit, zda kmen, prostředí nebo obojí má vliv na počet chyb, které krysa v bludišti udělá.

1. Otevřeme soubor *Rats CZ.sta*.
2. Z nabídky *Statistiky* vybereme položku *ANOVA* a zobrazí se úvodní panel *Obecná ANOVA/MANOVA*. Zvolíme položku *Vícefaktorová ANOVA* jako *Typ analýzy*, v poli *Metoda specifikace* ponecháme *Rychlé nastavení*. Klikneme na *OK*. Zobrazí se dialog *ANOVA/MANOVA Vícefaktorová ANOVA*. Klikneme na tlačítko *Proměnné* a zvolíme *Chyby* jako závislou proměnnou a *Kmen* a *Prostředí* jako kategoriální prediktory.
3. Klikneme dvakrát na *OK* a dostaneme se tak do dialogu *ANOVA Výsledky*. Tento dialog poskytuje spoustu možností pro volbu nejrůznějších výsledků. Jsou uspořádány na osmi záložkách. Pokud by nám tyto výsledky nestačily, je možné se přepnout do ještě obsáhlejšího výsledkového dialogu stiskem tlačítka *Více výsledků* 

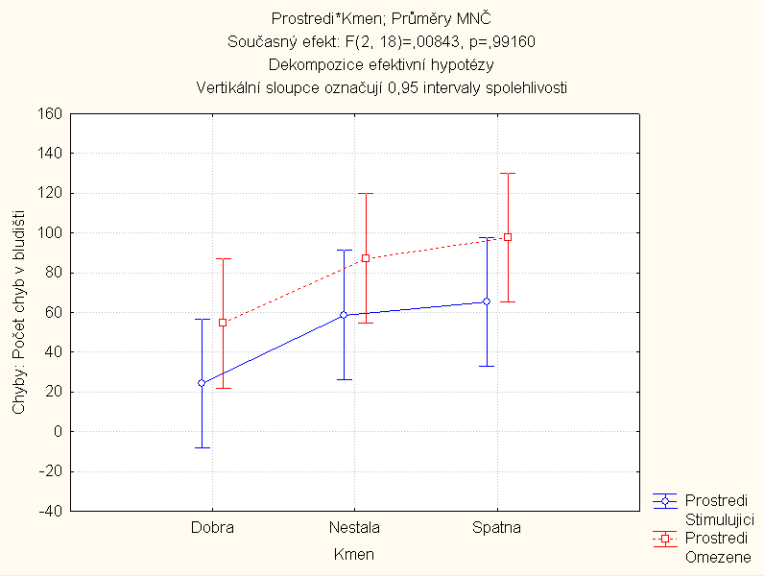
. Zpět do původního výsledkového dialogu se vrátíte stiskem tlačítka *Méně*. Nyní kliknutím na tlačítko *Všechny efekty/grafy* na záložce *Základ* zobrazíme dialog *Tabulka všech efektů*. Oba efekty (Prostředí i Kmen) jsou označeny jako významné (označeny hvězdičkou *), ale efekt interakce významný není.



4. Marginální průměry je možné vypočítat a zobrazit v grafu tak, že efekt interakce v tabulce vybereme (kliknutím) a klikneme na *OK*. Zobrazí se dialog *Uspořádání faktorů*, v němž určíme, jak bude vypadat vytvořený graf. Pro účely tohoto příkladu nastavíme Kmen v seznamu *Osa x, horní* a Prostředí v seznamu *Vzor čáry*. Kliknutím na *OK* vytvoříme příslušný graf.



Vidíme, že krysy vychované v omezeném prostředí, dělaly více chyb než krysy vychované ve stimulujícím prostředí nezávisle na kmeni. Současně krysy se špatnou schopností orientovat se v bludišti dělaly nejvíc chyb, nejméně jich dělaly chytré krysy.



5. Výsledky ANOVA lze zobrazit také ve formě tabulky kliknutím na tlačítko *Všechny efekty* na záložce *Základ*. Významné efekty jsou zvýrazněny červeně.

Jednorozměrné testy významnosti pro Chyby (Rats
Sigma-omezená parametrizace
Dekompozice efektivní hypotézy

Efekt	SČ	Stupně volnosti	PČ	F	p
Abs. člen	100233,4	1	100233,4	105,1353	0,000000
Prostredi	5551,0	1	5551,0	5,8225	0,026705
Kmen	7939,8	2	3969,9	4,1640	0,032635
Prostredi*Kmen	16,1	2	8,0	0,0084	0,991604
Chyba	17160,8	18	953,4		

Jak již bylo řečeno, v této analýze se projevil významný efekt faktorů *Prostředí* a *Kmen*. Zdůrazněme ovšem, že test významnosti nám neříká nic o tom, která (nebo které) ze skupin krys se od ostatních v počtu chyb významně liší. Abychom to zjistili, můžeme provést *Post-hoc testy*.

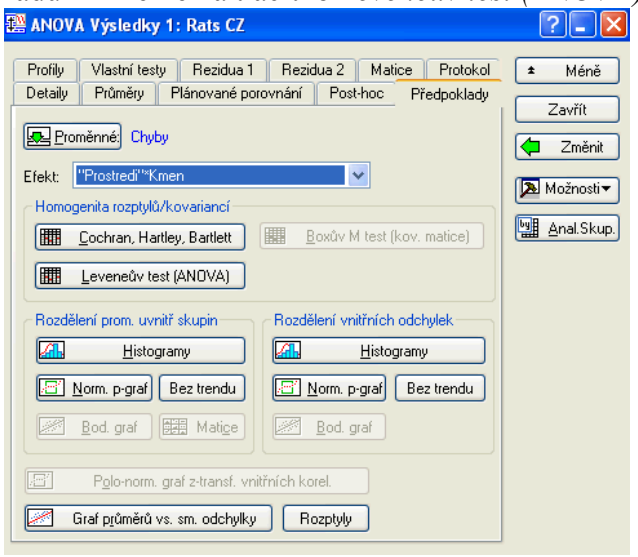
6. Klikneme na tlačítko *Více výsledků* a následně na záložku *Post-hoc*. V poli *Efekt* zvolíme *Kmen*, abychom mohli provést porovnání marginálních průměrů pro tento efekt. Kliknutím na tlačítko ***Schefféův*** se v tabulce zobrazí výsledky Schefféova testu:

Scheffeho test; proměnná Chyby (Rats CZ)
Pravděpodobnosti pro post-hoc testy
Chyba: meziskup. PČ = 953,38, sv = 18,000

Č. buňky	**Kmen**	{1} 39,375	{2} 73,000	{3} 81,500
1	Dobra		0,121814	0,044361
2	Nestala	0,121814		0,860446
3	Spatna	0,044361	0,860446	

Tato tabulka zobrazuje statistickou významnost rozdílů průměrů pro všechny páry skupin krys. Jak je vidět, pouze rozdíl mezi 1. a 3. skupinou, tj. mezi hloupými a chytrými krysami, je statisticky významný na hladině významnosti 0,05. Lze tedy utvořit závěr, že pouze hloupé krysy dělaly významně více chyb než krysy chytré, zatímco průměrné krysy se od zbývajících dvou skupin nijak významně neliší.

7. Samozřejmě je třeba otestovat předpoklady, za kterých lze metodu ANOVA uplatňovat. Přepneme se proto na záložku *Předpoklady*.
Jedním z předpokladů je homogenita rozptylů. *STATISTICA* poskytuje několik testů tohoto předpokladu ve skupině *Homogenita rozptylů/kovariancí* na záložce *Předpoklady*. Pro účely tohoto příkladu klikneme na tlačítko *Leveneův test (ANOVA)*.

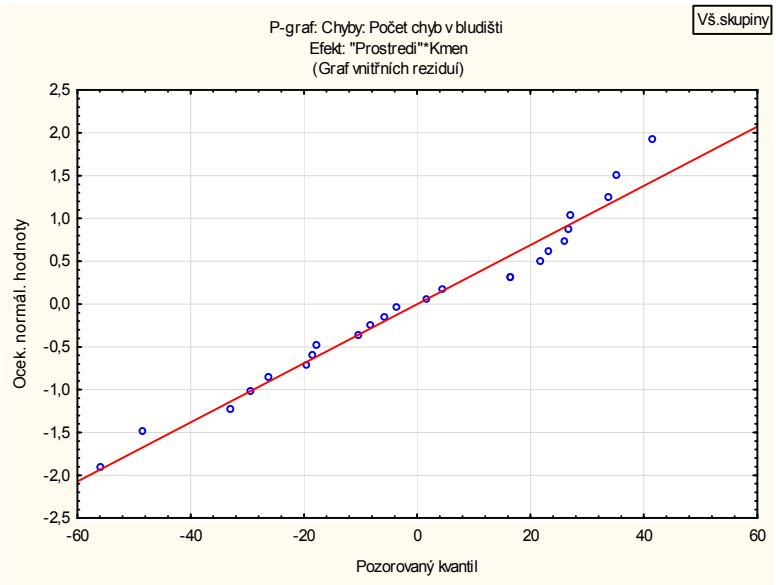


Níže uvedená tabulka s výsledky tohoto testu nevykazuje žádné údaje indikující, že by rozptyl v jednotlivých skupinách byl významně odlišný (tj. podmínka homogenity rozptylů je splněna).

Leveneův test homogenity rozptylů (Rats CZ
Efekt: "Prostredi"*Kmen
Stupně volnosti pro všechna F: 5, 18

	PČ Efekt	PČ Chyba	F	p
Chyby	226,4604	186,4132	1,214830	0,342166

ANOVA předpokládá, že rozdělení závislé proměnné v jednotlivých skupinách je normální. I přesto je ANOVA velice robustní vzhledem k porušení tohoto předpokladu. Pro posouzení typu rozdělení závislé proměnné je možné využít několika grafů, které jsou ve skupině *Rozdělení vnitřních odchylek* nebo *Rozdělení prom. uvnitř skupin* na záložce *Předpoklady***.** Pro přesnější ověření je možné použít např. ***Shapirův-Wilkův*** test normality v modulu *Základní statistiky/tabulky*.



8. Jak se zdá z výsledků analýzy, můžeme s velkou pravděpodobností říci, že faktory genetických dispozic i prostředí výchovy mají významný efekt na schopnost krys pohybovat se v bludišti.