

4 Statistická analýza

Obsah:

- 4.1 Testy hypotéz (Hypothesis Tests)
- 4.2 Testy hypotéz (Power and Sample Size)
- 4.3 Popisné statistiky (Descriptive Statistics)
- 4.4 Analýza rozptylu (ANOVA)
- 4.5 Neparametrické testy (Nonparametric Tests)
- 4.6 Vícerozměrná statistická analýza (Multivariate Analysis)
 - 4.6.1 Metoda hlavních komponent (Principal Component Analysis)
 - 4.6.2 Shluková analýza (Cluster Analysis)
 - 4.6.3 Diskriminační analýza (Discriminant Analysis)

4.1 Testy hypotéz (Hypothesis Tests)

Testy hypotéz jsou často používány k měření kvality výběru (vzorku) nebo ke zjištění, zda odhady daného parametru pro dva výběry jsou stejné. U parametrických metod je třeba nejprve ověřit předpoklady o rozdělení výběru, který byl vybrán ze souboru. Obvykle se vyžaduje, aby údaje byly nezávislá měření, jež vykazují normální rozdělení.

Kroky:

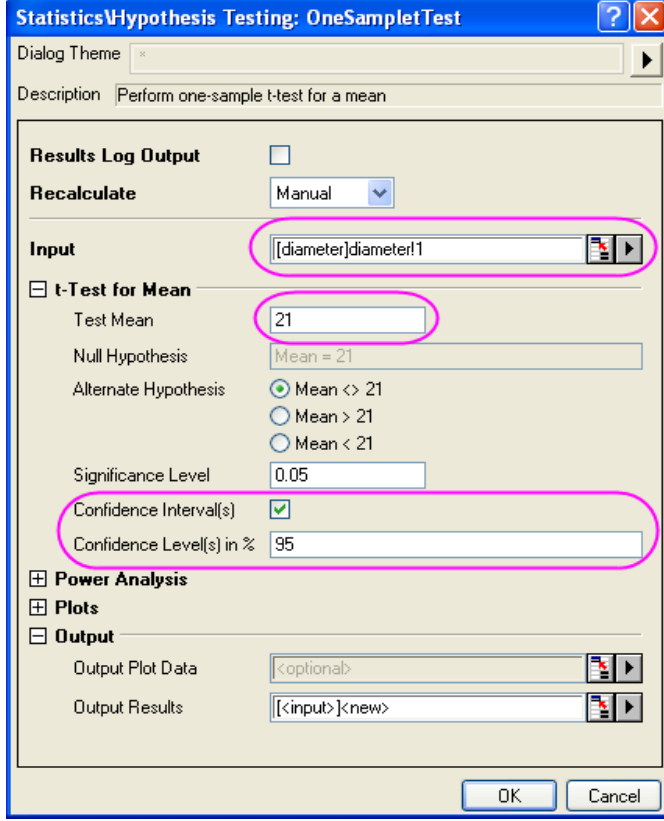
a) t-Test jednorozměrného výběru (One-Sample t-test)

Předpokládejme, že výrobce vyrábí vysoce kvalitní šroubové matice o průměru 21 mm. Oddělení kontroly jakosti náhodně odebralo 120 matic vyrobených matic, změřilo průměr u každé matice v mm a výsledky jsou **Diameters.dat**. Cílem je ověřit, zda střední hodnota (zde aritmetický průměr) matice je skutečně rovna 21 mm. O rozdělení naměřených průměrů je známo, že bývá normální, zatímco směrodatná odchylka souboru není známa. Budete používat **One-Sample t-test** podle následujících kroků:

1. Začněte s novým sešitem a naimportujte soubor **File, Import, Single ASCII \Samples\Statistics\Diameter.dat, Open, OK**.
2. V menu **Statistics, Hypothesis Testing**, otevřete **One-Sample t-test, Open dialog**. Klik na trojúhelníkovou šipku bloku **Input** vyberte sloupec **A(X): diameter** a zadejte oboustranný test a zadejte požadovanou hodnotu 21 k testování střední hodnoty pro úroveň spolehlivosti 95%.

3. Všimněte si, že ve výchozím nastavení poskytne postup popisné statistiky sledovaného průměru a výsledky testů hypotéz. Kromě toho je možné vytvořit také histogram dat a interval spolehlivosti pro střední hodnoty.

4. Klikněte na tlačítko **OK** k dokončení analýzy a generování výsledků. Tabulka **Descriptive Statistics** ukazuje velikost vzorku, průměr, směrodatnou odchylku a směrodatnou odchylku měřené proměnné. Vzorek vykazuje 21,00459 mm, což je nepatrně větší než požadovaná nulová hypotéza 21 mm a směrodatná odchylka průměru (SEM) je 0,00156 mm.



Descriptive Statistics

	N	Mean	SD	SEM
"diameter"	100	21.005	0.0156	0.00156

Z tabulky **t-test** je zřejmé, že statistika **t** (= 2,94337) a s ní související **p-hodnota Prob** (= 0,00404) prokazuje, že aritmetický průměr sledovaného průměru matic je odlišný od velikosti 21, a to na hladině významnosti $\alpha = 0,05$.

Interval spolehlivosti znamená, že s 95%ní statistickou jistotou tvrdíme, že skutečný průměr proměnné leží v intervalu [21,0015, 21,00769].

Confidence Intervals for Mean

	Conf. Levels in %	Lower Limits	Upper Limits
"diameter"	95	21.0015	21.00769

Test Statistics

	t Statistic	DF	Prob> t
"diameter"	2.9437	99	0.00404

Null Hypothesis: Mean = 21
 Alternative Hypothesis: Mean <> 21
 "diameter": At the 0.05 level, the population mean is significantly different from the test mean (21).

Output:

Book1

1 One Sample t Test (1.3.2014 07:08:34)

Notes

X-Function	One Sample t Test
User Name	mime0352
Time	1.3.2014 07:08:34

Input Data

	Data	Range
diameter	[Book1]diameter!A*diameter	[1*:100*]

Descriptive Statistics

	N	Mean	SD	SEM
"diameter"	100	21,00459	0,0156	0,00156

Test Statistics

	t Statistic	DF	Prob> t
"diameter"	2,9437	99	0,00404

Null Hypothesis: Mean = 21
Alternative Hypothesis: Mean \neq 21
"diameter": At the 0.05 level, the population mean is significantly different from the test mean (21).

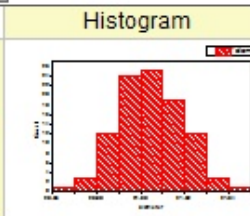
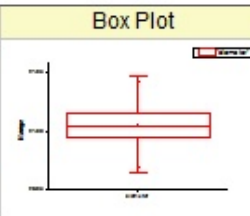
Confidence Intervals for Mean

	Conf. Levels in %	Lower Limits	Upper Limits
"diameter"	95	21,0015	21,00769

Powers

	Alpha	Sample Size	Power
"diameter"	0,05	100	0,83023
	0,05	50	0,53217
	0,05	100	0,83023
	0,05	200	0,98548

Plots

	Histogram	Box Plot
"diameter"		

PlotData1 OneSampletTest1

b) Pairový t-test

1. Začněte v novém sešitě a nainportujte **File, Import, Single ASCII, /Samples/Statistics /abrasion_raw.dat, Open, OK**. Pak zavolejte **Statistics, Hypothesis Testing, Pair-sample t-Test, Open dialog**.

2. V bloku **Input** nastavte sloupec **tireA** jako **1st Data Range** a sloupec **tireB** jako **2nd Data Range**, zadejte **0** na testovaný průměr **Test Mean**.

3. Ponechte ostatní defaultní nastavení. Klikněte na **OK** pro generování výsledků.

V tabulce **t-test** vidíte, že statistika t ($= 2,83119$) a související **p-hodnota Prob** ($= 0,02536$) ukazuje, že rozdíl mezi těmito dvěma středními hodnotami je statisticky významný, a je proto třeba říci, že oba typy pneumatik mají odlišnou odolnost proti otěru.

Descriptive Statistics

	N	Mean	SD	SEM
"tireA"	8	6145	1366.49709	483.12968
"tireB"	8	5825	1097.46461	388.01233
Difference		320		

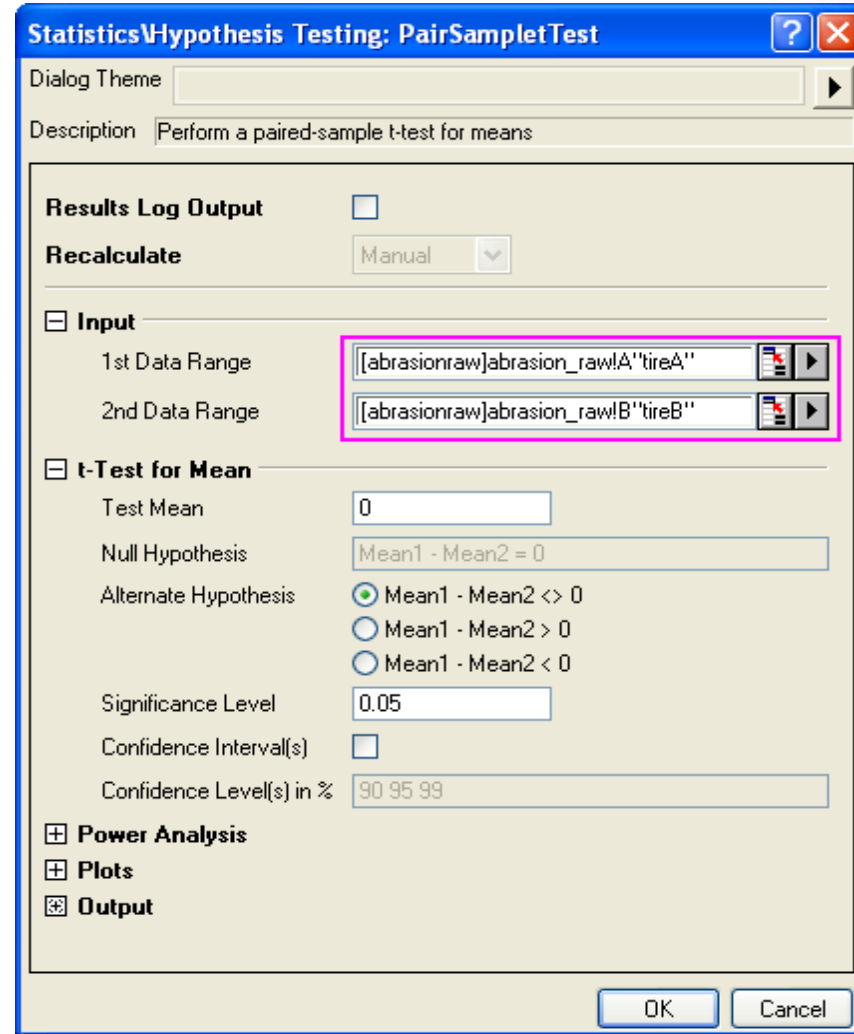
Test Statistics

t Statistic	DF	Prob> t
2.83119	7	0.02536

Null Hypothesis: mean1-mean2 = 0

Alternative Hypothesis: mean1-mean2 \neq 0

At the 0.05 level, the difference of the population means is significantly different from the test difference(0).



Output:

Book1

1 Paired Sample t Test (1.3.2014 07:21:00)

Notes

X-Function	Paired Sample t Test
User Name	mime0352
Time	1.3.2014 07:21:00

Input Data

	Data	Range
1st Data Range	[Book1]abrasion_raw!A"tireA"	[1*:8*]
2nd Data Range	[Book1]abrasion_raw!B"tireB"	[1*:8*]

Descriptive Statistics

	N	Mean	SD	SEM
"tireA"	8	6145	1366,49709	483,12968
"tireB"	8	5825	1097,46461	388,01233
Difference		320		

Test Statistics

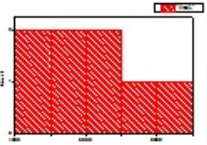
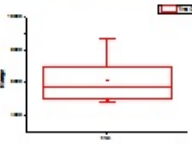
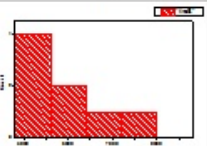
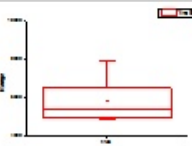
t Statistic	DF	Prob> t
2,83119	7	0,02536

Null Hypothesis: mean1-mean2 = 0
Alternative Hypothesis: mean1-mean2 <> 0
At the 0.05 level, the difference of the population means is significantly different from the test difference(0).

Powers

	Alpha	Sample Size	Power
Actual Power	0,05	8	0,68167
Hypo. Power	0,05	50	1
	0,05	100	1
	0,05	200	1

Plots

	Histogram	Box Plot
"tireA"		
"tireB"		

PlotData1 Pair SampletTest1

c) Studentův t-test shodnosti výběrů

Lékař hodnotí účinek dvou uspávacích léků. K otestování účinnosti obou léků se vybere 20 pacientů trpících nespavostí. Polovina pacientů vzala lék **A** a druhá polovina lék **B**. Byl zaznamenán prodloužený čas spaní u každého pacienta a je v datech **time_raw.dat**. Test shodnosti, zda oba léky mají různý vliv na pacienty, se provede se dvěma výběry nezávislým t-testem:

1. Začněte s nového sešitu a importovat soubor **File, Import, Single ASCII, \Samples\statistika \time_raw.dat, Open, OK**.
2. Otevřete **Statistics, Hypothesis Testing, Two-Sample t-Test, Open dialog**, a pokračujte....
3. Vyberte "**Raw**" do **Input Data Form**, nastavte sloupec **A** a sloupec **B** jako první a druhý výběr.
4. Ostatní defaultní nastavení ponechte a klikněte na tlačítko **OK** pro generování výsledků.

Statistics\Hypothesis Testing: TwoSampletTest

Dialog Theme *

Description Perform a two-sample t-test for means

Results Log Output

Recalculate Manual

Indexed: factor variable and response data are stored in separate columns.

Raw: each column contains response data from a level of the factor variable.

Input Data Form Raw

Input

1st Data Range [timeraw]time_raw!A"medicineA"

2nd Data Range [timeraw]time_raw!B"medicineB"

t-Test for Mean

Test Mean 0

Null Hypothesis Mean1 - Mean2 = 0

Alternate Hypothesis Mean1 - Mean2 <> 0
 Mean1 - Mean2 > 0
 Mean1 - Mean2 < 0

Significance Level 0.05

Confidence Interval(s)

Confidence Level(s) in % 90 95 99

OK Cancel

t-Test Statistics



	t Statistic	DF	Prob> t
Equal Variance Assumed	1.89811	18	0.07384
Equal Variance NOT Assumed (Welch Correction)	1.89811	17.8248	0.074

Null Hypothesis: mean1-mean2 = 0
Alternative Hypothesis: mean1-mean2 <> 0
At the 0.05 level, the difference of the population means is NOT significantly different from the test difference(0).

Testování poskytuje dva testy rozdílu středních hodnot. První je založen na předpokladu, že rozptyly dvou výběrů jsou shodné a druhý nejsou shodné. V této úloze oba testy ukazují, že nebyl prokázán rozdíl účinků mezi lékem A a lékem B. (p-hodnoty jsou 0,0738 a 0,074, což je větší, než je hladina významnosti 0,05.)

Output:

Book1

1 Two sample t Test (1.3.2014 07:29:22)

Notes

X-Function	Two sample t Test
User Name	mime0352
Time	1.3.2014 07:29:22

Input Data

	Data	Range
1st Data Range	[Book1]time_raw!A"medicineA"	[1*:10*]
2nd Data Range	[Book1]time_raw!B"medicineB"	[1*:10*]

Descriptive Statistics

	N	Mean	SD	SEM
"medicineA"	10	2,35	1,97611	0,6249
"medicineB"	10	0,75	1,78901	0,56573
Difference		1,6		

t-Test Statistics

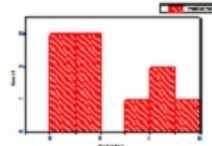
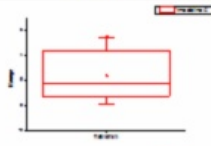
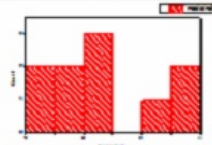
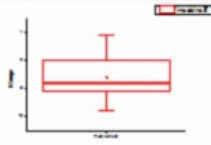
	t Statistic	DF	Prob> t
Equal Variance Assumed	1,89811	18	0,07384
Equal Variance NOT Assumed (Welch Correction)	1,89811	17,8248	0,074

Null Hypothesis: mean1-mean2 = 0
 Alternative Hypothesis: mean1-mean2 <> 0
 At the 0.05 level, the difference of the population means is NOT significantly different from the test difference(0).

Powers

	Alpha	Sample Size	Power
Actual Power	0,05	20	0,43526
	0,05	50	0,83662
Hypo. Power	0,05	100	0,98753
	0,05	200	0,99997

Plots

	Histogram	Box Plot
"medicineA"		
"medicineB"		

PlotData1 TwoSampletTest1

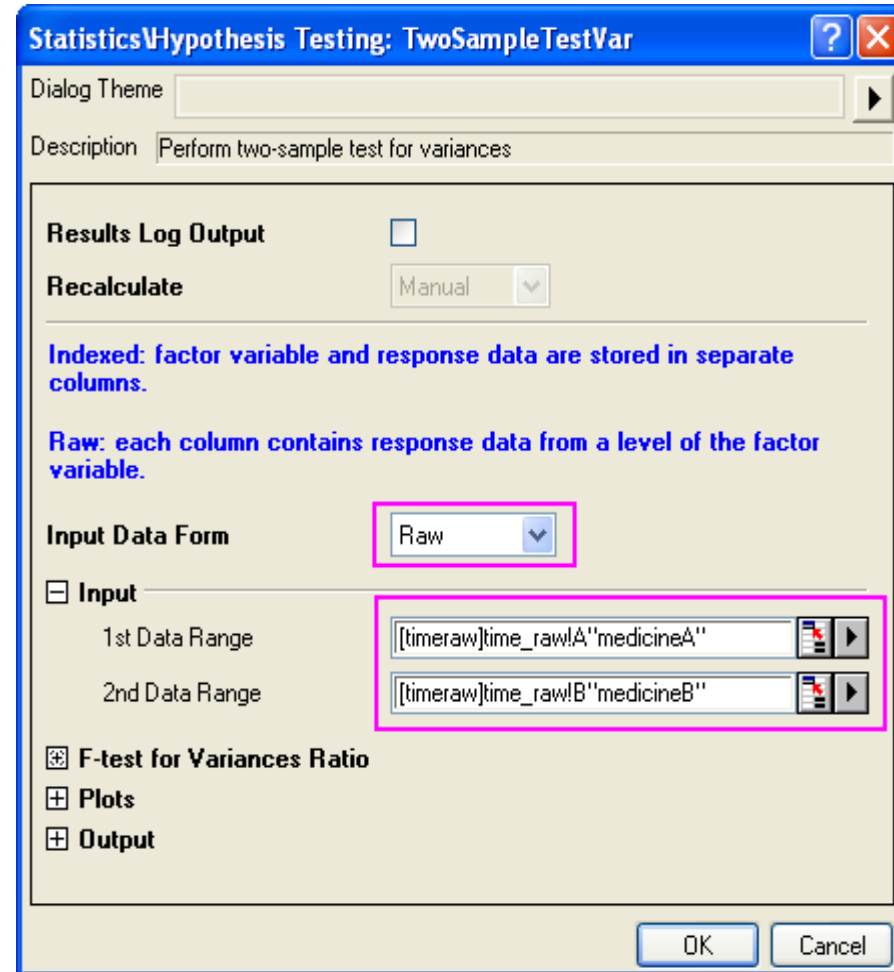
d) Test shodnosti rozptylů

1. Pokračujte v novém sešitě **File, Import, Single ASCII \Samples\Statistics\ time_raw.dat, Open, OK.**

1. Otevřete **Statistics, Hypothesis Testing, Two-Sample Test for Variance, Open dialog** a pokračujte....

3. Vyberte "**Raw**" do řádku **Input Data Form**, a v **Input** nastavíte sloupec **A:medicine A** a sloupec **B:medicine B** jako první a druhý výběr.

4. Ponecháte ostatní defaultní nastavení a kliknete na tlačítko **OK** pro generování výsledků.



Descriptive Statistics

	N	Mean	SD	Variance
"medicineA"	10	2.35	1.97611	3.905
"medicineB"	10	0.75	1.78901	3.20056

F Statistics

	F	Numer. DF	Denom. DF	Prob > F
	1.2201	9	9	0.77181

Null Hypothesis: Variance1/Variance2 = 1
Alternative Hypothesis: Variance1/Variance2 <> 1
At the 0.05 level, the two population variances are NOT significantly different.

Podle **p-hodnoty Prob** = 0,77181 > 0,05 plyne, že nelze odmítnout nulovou hypotézu o rozdílnosti rozptylů.

Output:

Book1

1 Two-Sample Test for Variance (1.3.2014 07:33:57)

Notes

X-Function	Two-Sample Test for Variance
User Name	mime0352
Time	1.3.2014 07:33:57

Input Data

	Data	Range
1st Data Range	[Book1]time_raw!A"medicineA"	[1*:10*]
2nd Data Range	[Book1]time_raw!B"medicineB"	[1*:10*]

Descriptive Statistics

	N	Mean	SD	Variance
"medicineA"	10	2,35	1,97611	3,905
"medicineB"	10	0,75	1,78901	3,20056

F Statistics

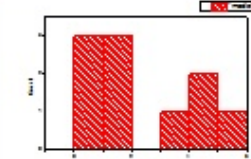
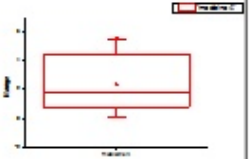
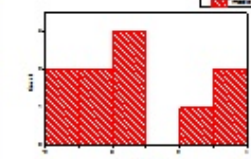
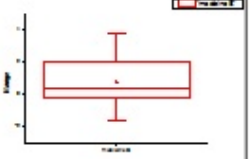
	F	Numer. DF	Denom. DF	Prob > F
	1,2201	9	9	0,77181

Null Hypothesis: Variance1/Variance2 = 1
Alternative Hypothesis: Variance1/Variance2 <> 1
At the 0.05 level, the two population variances are NOT significantly different.

Confidence Intervals for Variances

Conf. Levels in %	Lower Limits	Upper Limits
95	0,30306	4,91212

Plots

	Histogram	Box Plot
"medicineA"		
"medicineB"		

PlotData1 | TwoSampleTestVar1

4.2 Test velikosti výběru (Power and Sample Size)

Test síly a test velikosti výběru jsou užitečné pro správný návrh experimentu. Nedostatečné údaje a nedostatek síly odhadu k odmítnutí falešné nulové hypotézy může vést k chybnému závěru, stejně jako na druhé straně příliš mnoho nadbytečných dat vede ke ztrátě času a peněz. Je proto třeba určit velikost výběru před provedením experimentu. Sílu odhadu lze vypočítat pro danou velikost výběru, stejně jako lze opačně vypočítat velikost výběru pro danou sílu odhadu. Tutoriál ukáže, jak pro výpočet velikosti výběru nebo velikost síly odhadu navrhovat experimenty v různých praktických situacích.

a) PSS-Analýza velikosti a síly výběru - (PSS)One-Sample t-Test

Podstata: Sociolog chce zjistit, zda průměrná míra kojenecké úmrtnosti v USA je rovna 8. V návrhu experimentu by se neměl rozdíl lišit o více než 0,5. Z pilotních studií je známo, že směrodatná odchylka by měla být 2,1.

Otázka: Jaká bude velikost výběru při odhadu průměrné kojenecké úmrtnosti na statistické jistotě 95% ($\alpha = 0,05$) pro hodnoty síly odhadu 0,7, 0,8 a 0,9 ?

Kroky:

1. Aktivujte prázdný list, zvolte **Statistics, Power and Sample Size, (PSS)One-Sample t-test, Open dialog** a pokračujte...

2. Proveďte nastavení dle následujícího obrázku vpravo pro dialogové okno **PSS_tTest1** a klikněte na tlačítko **OK**.

Výstup: Je vygenerován výsledkový list spolu s přehledem vypočtené velikosti výběru pro hypotetické síly odhadu.

Sample Size(s) for Hypothetical Power(s)

Alpha	Power	Sample Size
0.05	0.7	111
0.05	0.8	141
0.05	0.9	188

Null Mean = 8; Alternate Mean = 8.5; SD = 2.1; 2-Sided Test

Interpretace: Podle návrhu experimentu by sociolog měl provést analýzu výběru o 111 vzorcích pro sílu odhadu 0,7, výběr 141 vzorků pro sílu 0,8 a/nebo výběr 188 vzorků pro sílu 0,9.

Statistics Power and Sample Size: PSS_tTest1

Dialog Theme *

Description Perform power and sample size analysis for one-sample t-test.

Results Log Output

Calculate Sample Size

Test Specification

Null Mean 8

Alternate Mean 8.5

Standard Deviation 2.1

Alpha 0.05

Hypothetical Power(s) 0.7 0.8 0.9

Tail 2 side

Options

Output Results <new>

OK Cancel

b) PSS-Analýza velikosti a síly dvou výběrů

- (PSS)Two-Sample t-Test

Podstata: Ordinace lékaře spolupracuje se dvěma zdravotními pojišťovnami, Healthwise a Medicare. Cílem je porovnat střední dobu úhrady pohledávek (ve dnech) obou pojišťoven. Historická data ukazují, že pro pojišťovnu Healthwise je průměrná doba 32 dnů se směrodatnou odchylkou 7,5 dne. Pro pojišťovnu Medicare je průměrná doba úhrady 42 dnů se směrodatnou odchylkou 3,5 dne.

Otázka: Bylo vybráno 5 požadavků z každé pojišťovny a byly zaznamenány odpovídající doby úhrady. Jaká je síla detekování rozdílu v průměrných časech úhrad mezi dvěma pojišťovnami o 5% nebo více?

Kroky:

1. Vypočtete sdruženou směrodatnou odchylku dle vzorce

$$\sqrt{((5 - 1) * 7.5^2 + (5 - 1) * 3.5^2) / (5 + 5 - 2)} = 5.85235$$

Všimněte si, že tato hodnota se použije jako směrodatná odchylka k pozdějším výpočtům síly.

2. Velikost vzorku z první skupiny a druhá skupiny by měla být $5 + 5 = 10$.

3. Aktivujte prázdný list, zvolte **Statistics, Power and Sample Size, (PSS)Two-Sample t-test, Open dialog** a pokračujte....

4. Proved'te nastavení dle následujícího obrázku pro dialogové okno **PSS_tTest2** a klikněte na tlačítko **OK**.

Výstup: Je vygenerován výsledkový list spolu s přehledem vypočtené velikosti výběru pro hypotetickou síly odhadu.

Power(s) for Hypothetical Sample Size(s)

Alpha	Sample Size	Power
0.05	10	0.95054

Group1 Mean = 32; Group2 Mean = 42; SD = 5.85235; 2-Sided Test

Statistics\Power and Sample Size: PSS_tTest2

Dialog Theme: *

Description: Perform power and sample size analysis for two independent sample t-test.

Results Log Output:

Calculate: Power

Test Specification:

- 1st Group Mean: 32
- 2nd Group Mean: 42
- Standard Deviation: 5.8524
- Alpha: 0.05
- Hypothetical Sample Size(s): 10
- Tail: 2 side

Options:

Output Results: <new>

OK Cancel

Interpretace:

Lze konstatovat, že ordinace má sílu 0.95054 : 1 (nebo 95%ní) zjištění rozdílu mezi oběma pojišťovny, když shromažďuje 5 nároků na každou pojišťovnu. Jinými slovy, existuje šance, že se nepodaří zamítnout nulovou hypotézu o odlišnosti obou pojišťoven, protože spočtená pravděpodobnost (hladina významnosti) je rovna 4,946% (1 - 0,95054) je menší než 5%.

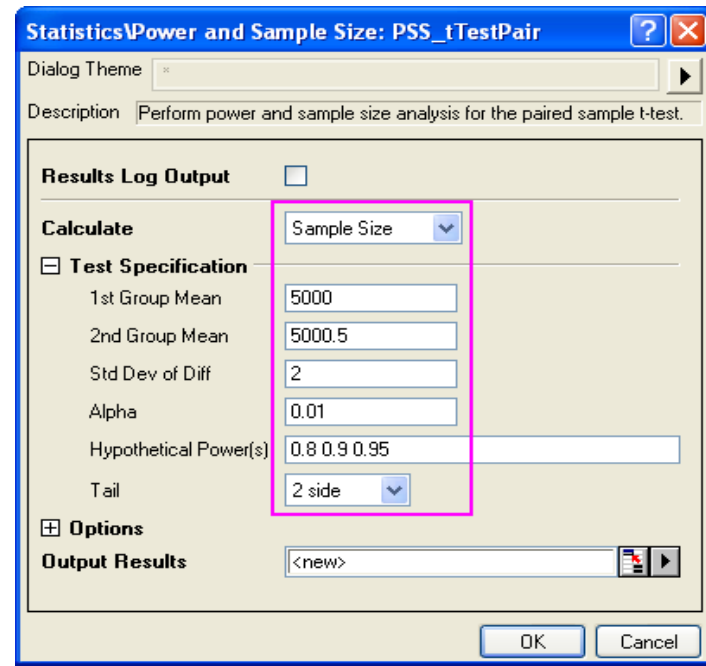
c) PSS-Analýza velikosti a síly párového výběru (PSS Paired Sample t-Test)

Podstata: Existují dva měřicí přístroje stejného typu k měření hloubky a-Si tenkého filmu. Zjistěte, zda existuje nějaký rozdíl v obou přístrojích, když je požadavek na experiment k měření hloubky a-Si tenké vrstvy na stejné pozici u obou přístrojů v různých produktech. Podle předchozí studie bylo zjištěno, že směrodatná odchylka rozdílu je $2 \mu\text{m}$. Ta bude sloužit za odhad směrodatné odchylky rozdílů při plánování experimentu. Rozdíl v měření u dvou přístrojů nemůže být více než $0,5 \mu\text{m}$, a průměrná hloubka naměřená prvním přístrojem je $5000 \mu\text{m}$.

Otázka: Kolik vzorků musí být naměřeno na úrovni statistické jistoty 99% pro sílu 0,8, 0,9, 0,95?

Kroky:

1. Podle informace plyne, že průměr u prvního přístroje je $5000 \mu\text{m}$ a průměr u druhého je $5000,5 \mu\text{m}$.
2. Aktivujte prázdný list a v menu vyberte **Statistics, Power and Sample Size, (PSS)Paired t-Test** a pokračujte...
3. Dle obrázku vpravo nastavte data v okně **PSS_tTestPair** a klikněte na **OK**.



Výstup:

Výsledný list ukazuje vypočítané velikosti výběr (tj. počet vzorků) pro různé síly.

Alpha	Power	Sample Size
0.01	0.8	191
0.01	0.9	242
0.01	0.95	289

Group1 Mean = 5000; Group2 Mean = 5000.5; SD = 2; 2-Sided Test

Interpretace:

Z výsledků lze usuzovat, že technik má ke zjištění rozdílu typu užitého přístroje 80% ní šanci, pokud bude měřit 191krát hloubku a-Si tenké vrstvy filmu, 90% ní šanci, pokud bude měřit 242krát hloubku a-Si tenké vrstvy filmu a 95% ní šance, pokud bude měřit 289krát hloubku a-Si tenké vrstvy filmu.

4.3 Popisné statistiky (Descriptive Statistics)

Origin poskytuje komplexní popisnou statistiku včetně základní statistiky (průměr, medián, rozptyl, apod.), je vyčíslena frekvence a korelační koeficienty dat. Kromě silné grafické vlastnosti, statistické nástroje zde pomáhají sumarizovat a analyzovat data.

A. Nalezení informace o frekvenci skupin

Můžeme použít nástroj **Discrete Frequency** pro rychlé získání informací o frekvenci skupin dat.

1. Začněte s novým projektem. Importujte data souboru **File, Import, Single ASCII, \Samples \Statistics\automobile.dat, Open, OK.**
1. Zvýrazněte první dva sloupce. Vyberte **Statistics, Descriptive Statistics, Discrete Frequency, Open dialog** a otevře se dialogové okno. Sloupec **A** a sloupec **B** jsou automaticky vybrány za vstupní data. Klikněte na **OK.**

B. Výpočet popisných statistik seskupených dat

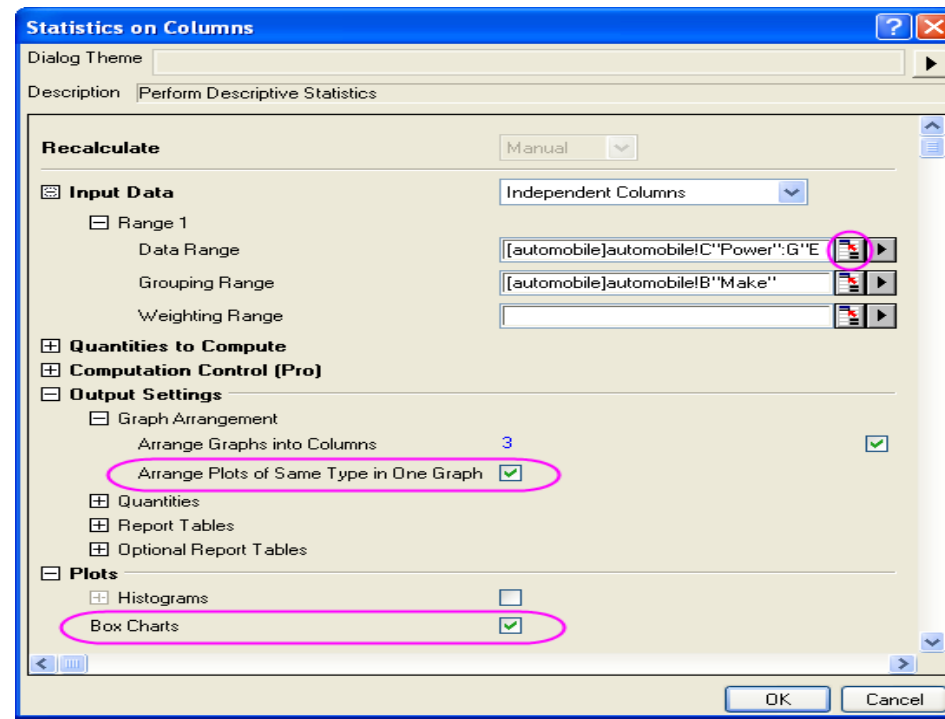
Pomocí nástroje **Statistics on Columns** lze najít základní statistické údaje každé skupiny dat.

Kroky:

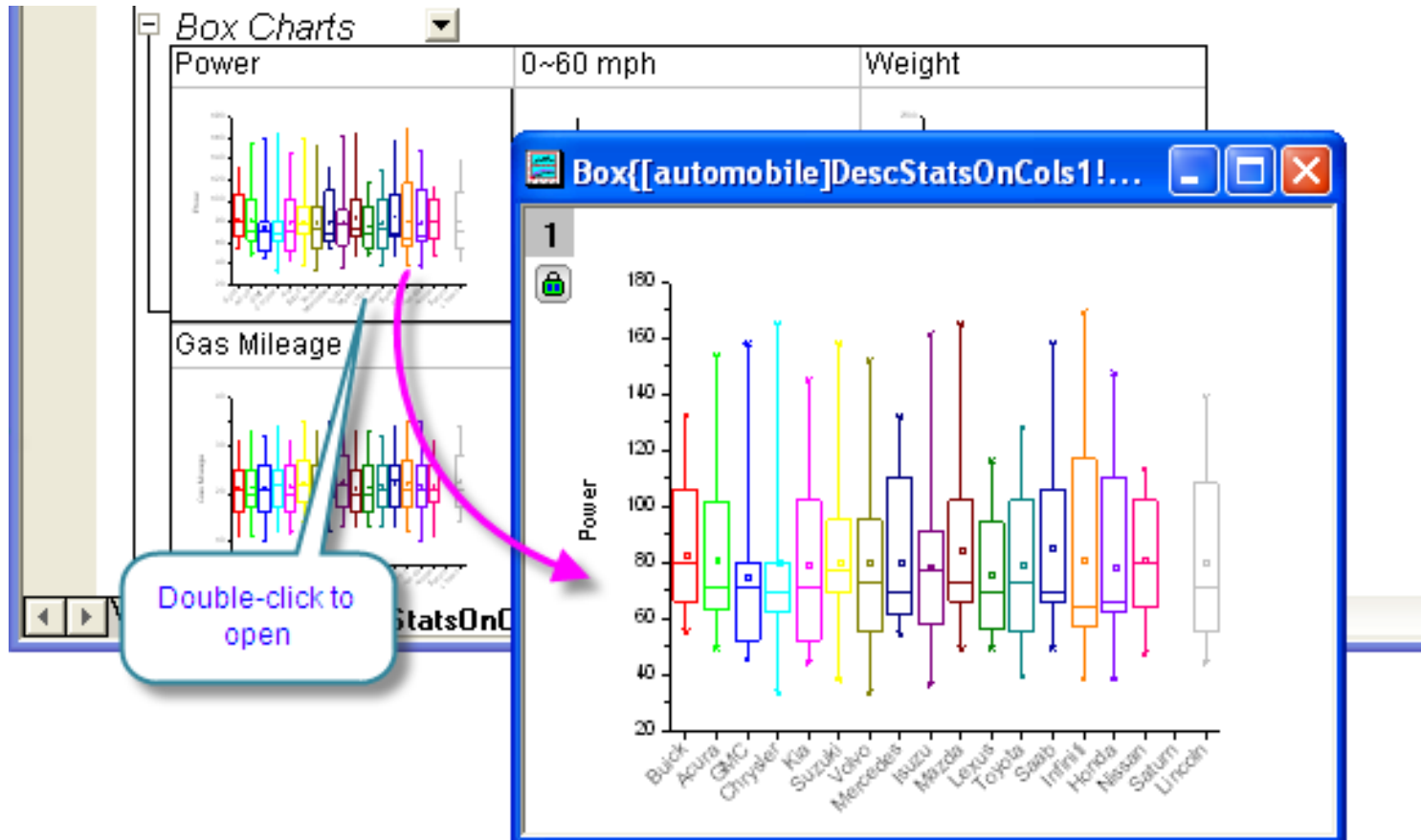
1. Přepněte zpět na první list **automobile** a vyberte **Statistics: Descriptive Statistics: Statistics on Columns** se otevře dialog **Statistics on Column, Open dialog**.
2. Otevřete uzel **Range 1** a klikněte na **interactive button**. Dialogové okno "srolovat" a můžete nastavit rozsah dat **Data Range** jako sloupec **C** až sloupec **G** volbou sloupců **C(Y)** a tažením až do sloupce **G(Y)** v listu. Klikněte na tlačítko v dialogovém okně roletky až do obnovení dialogu. Klikněte na tlačítko **triangle button**, umístěného vedle **Grouping Range** a zvolte **B(Y): Make**.

3. Zde ukážeme, jak se dělá krabicový graf pro seskupená data, aby bylo možné porovnat všechny skupiny v grafu k rychlému porovnání. Proveďte proto následující kroky:

- 1) Rozbalte **Output Settings** a **Graph Arrangement**. Vyberte **Arrange Plots of Same Type in One Graph** zaškrtnutím políčka.
- 2) Rozbalte **Plots** a zaškrtněte políčko **Box Charts**.



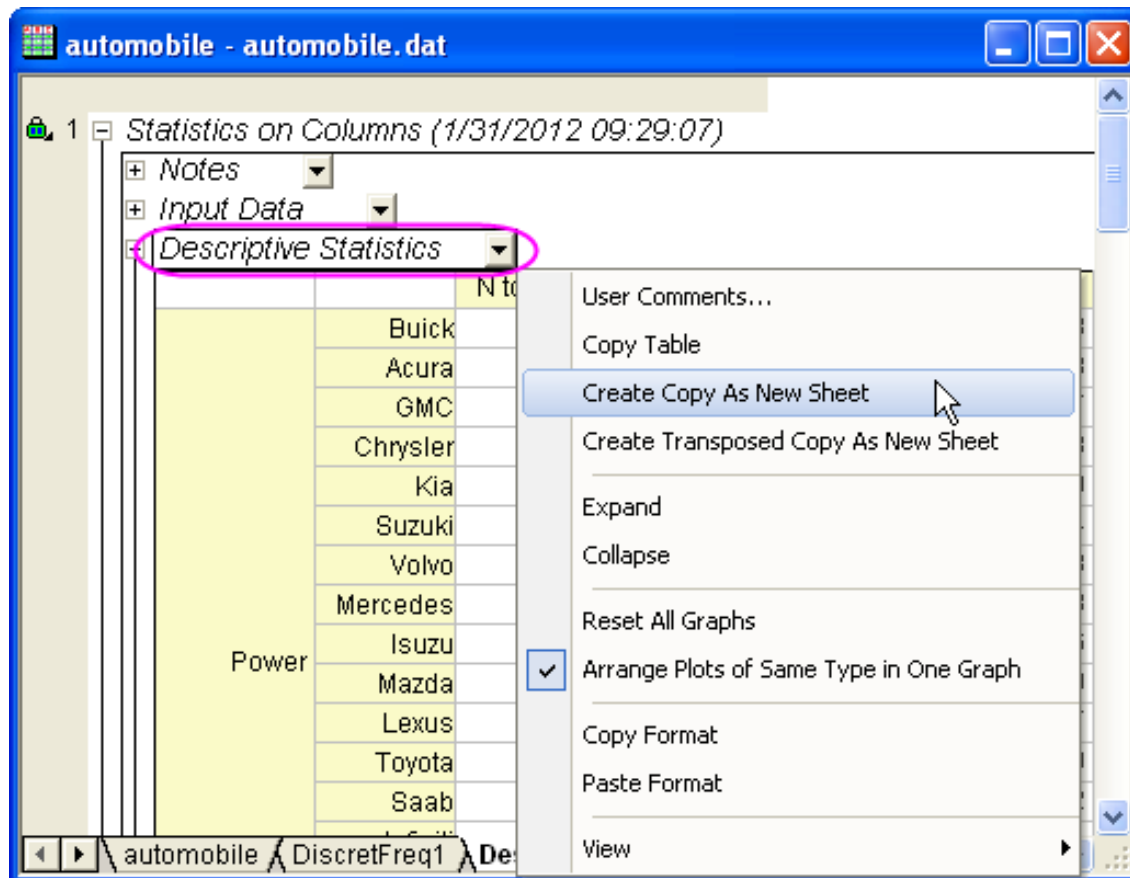
4. Klikněte na **OK**, aby se výsledky ve zprávě listu.



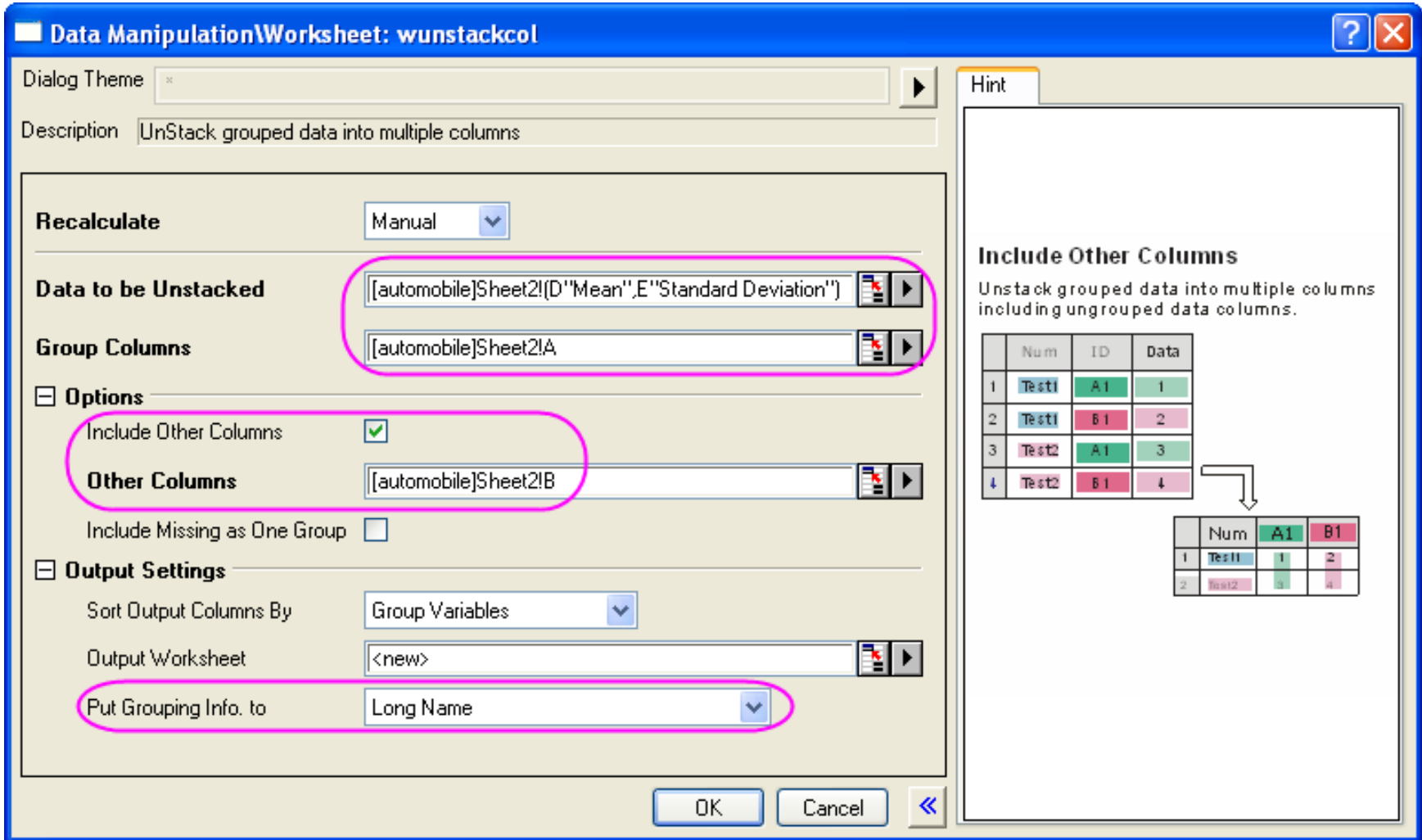
C. Použití statistických výsledků pro další operace

Po použití dialogu **Statistics on Columns** k vypracování zprávy, bývá požadována další analýza a grafické vykreslení statistických výsledků. Například, k prezentování průměrných hodnot parametrů vozidel vyrobených v letech 1992 - 2004 (tj. koní, 0-60 mph čas, hmotnost, stav tachometru) , proveďte následující kroky:

1. Ve zprávě listu, klikněte pravou myší na název tabulky **Descriptive Statistics** a vyberte **Create Copy as New Sheet** jako nový list ze zkrácené nabídky.



2. Je-li aktivní nový list, vyberte **Worksheet, Unstack Columns**.
3. V otevřeném dialogovém okně vyberte sloupce **D** a **E** jako **Data to be Unstacked**. Protože trojúhelníkové tlačítko podporuje pouze jednu volbu, musíte použít tlačítko **interactive button**.
4. Nastavte sloupec **A** jako **Group Variables**.
5. Zaškrtněte **Include Other Columns** a nastavte **Other Columns** na sloupec **B**.
6. Nastavte **Put Grouping Info. To** na **Long Name**. Klikněte na **OK**.

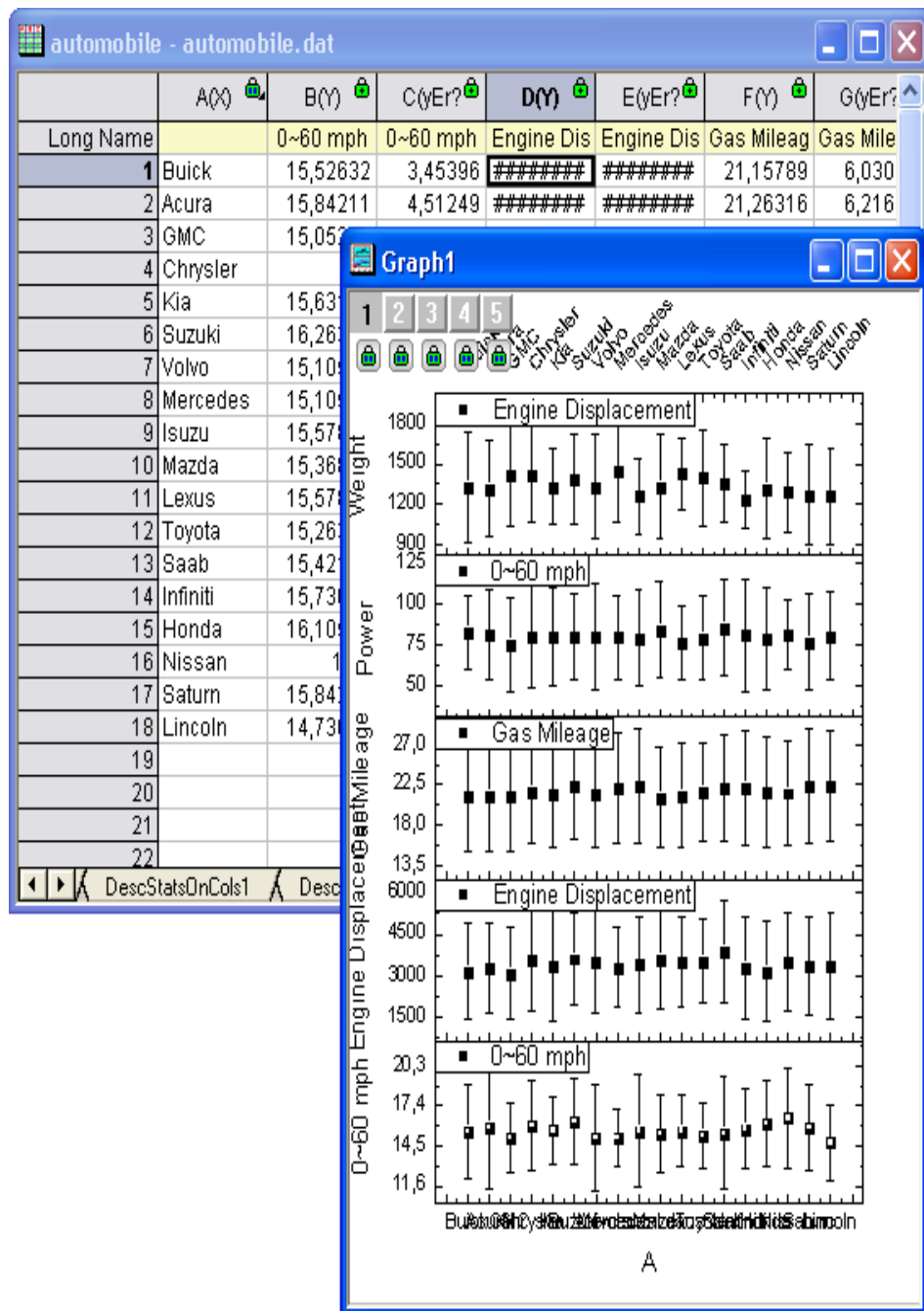


7. V důsledku **unstack** sloupce, dostaneme střední hodnotu a směrodatnou odchylku Power, 0 ~ 60 mph času, hmotnost, plynu najetých kilometrů a motoru na 18 různé značky vozidel.

8. Zvýrazněte celý výsledek listu. Vyberte **Plot, Multi-Curve, Stack** z hlavního menu.

9. V dialogovém okně pop-up, všechny sloupce v listu jsou automaticky nastaveny jako vstup. Nastavte **Plot Type** na **Scatter** a klikněte na tlačítko **OK**.

V uvedeném obrázku jsou horní popisky osy **X-Axis Tick** pro přehlednost otočeny o 45 stupňů. Chcete-li to provést, dvoj-klikem na popisky ticků a otevře se dialog **X-Axis**. Nastavte **Rotation** na záložce **Custom Tick Labels**.



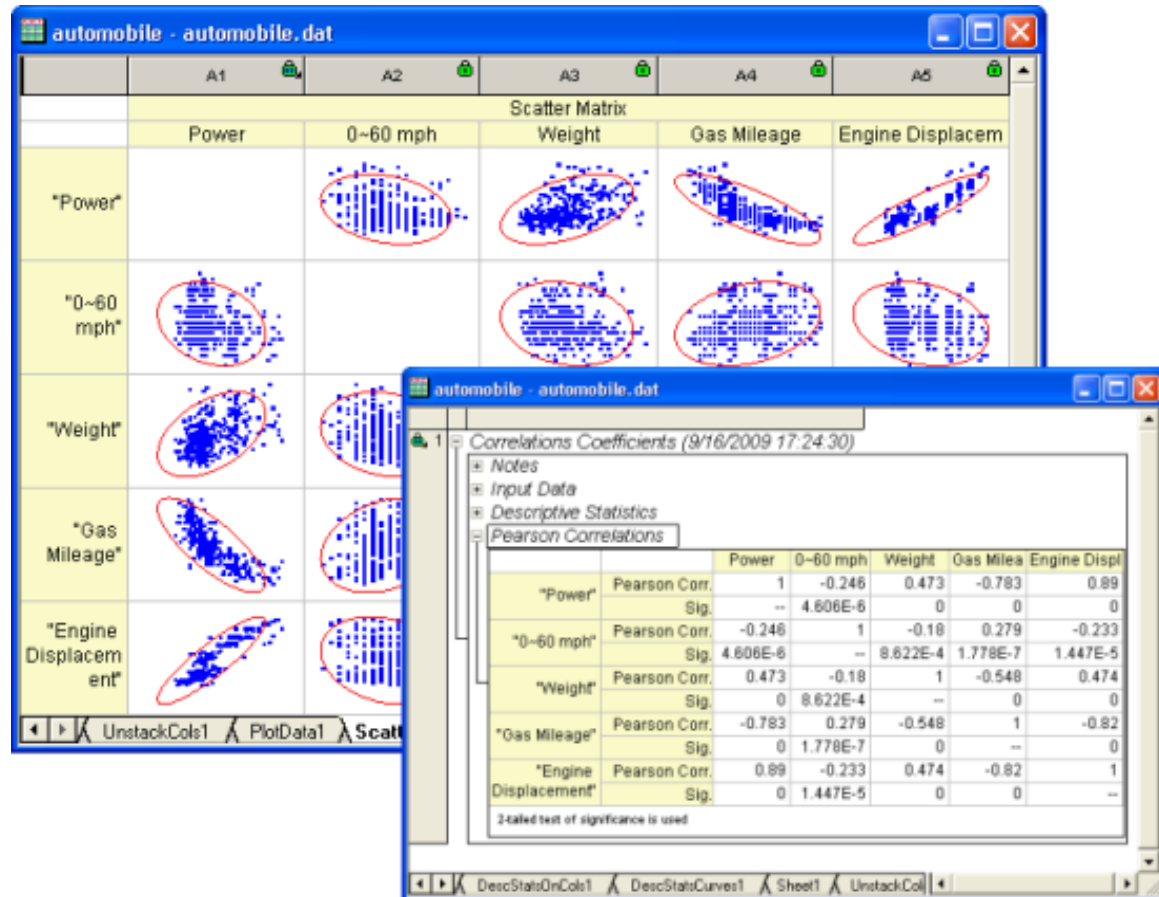
D. Analýza vztahu mezi různými indikátory

Můžeme použít korelační koeficient k prozkoumání vztahu mezi sloupci našich automobilových údajů. Kromě toho můžeme vykreslit maticový graf s konfidenční elipsou ke grafickému znázornění korelace.

1. Přejděte do původního listu zdrojových dat. Zvýrazněte posledních pět sloupců.
2. Zvolte **Statistics, Descriptive Statistics, Correlation Coefficient** a otevřete nástroj **Correlation Coefficient**. Všimněte si, že Pearsonův korelační koeficient je zvolen defaultně. Tato metoda je vhodná pro kvantitativní data.

3. V oddíle **Plots** zaškrtněte **Add Confidence Ellipse**. Políčko **Scatter Plot** by pak mělo být vybráno automaticky. Klikněte na **OK**.

Poznámka: Vysokou pozitivní korelaci mezi **Engine Displacement** a **Power** a vysokou negativní korelaci mezi **Gas Mileage** a **Engine Displacement** čili počtem najetých kilometrů a zdvihovým objemem motoru.



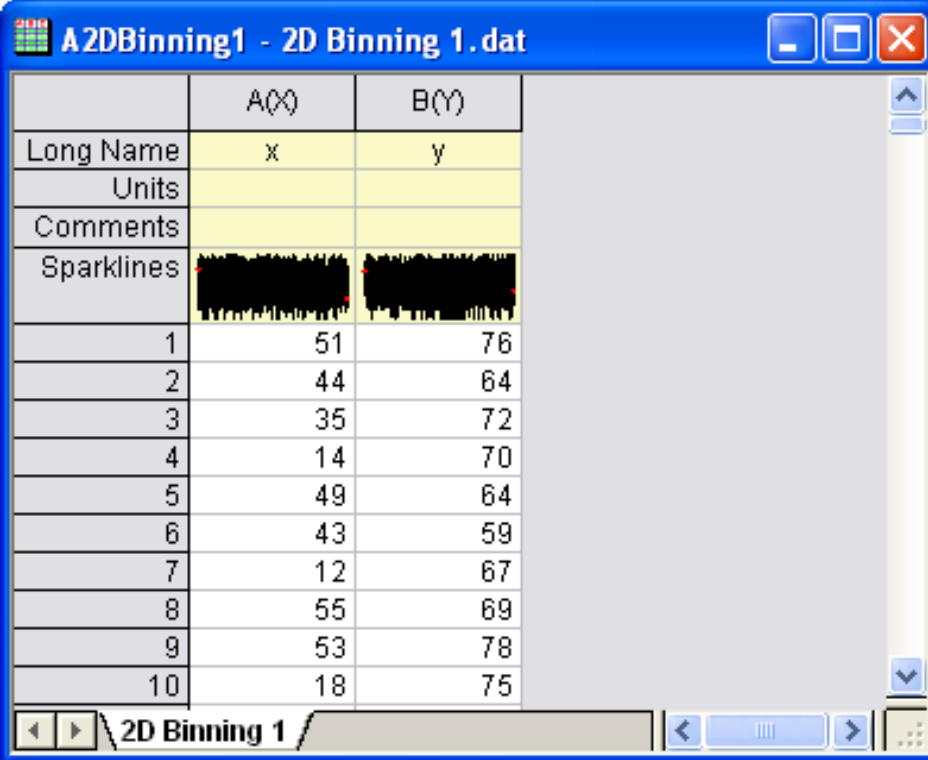
E. 2D-čítač plošného rozdění (Binning)



Operace **2D Frequency Count/Binning** počítá frekvence pro data se dvěma proměnnými. V případě potřeby se sestojí 3D-sloupcový graf a/nebo obrazový diagram k intuitivní demonstraci rozdění datových bodů.

Kroky:

1. Vytvořte nový projekt a importujte data **File, Import, Single ASCII, \Samples\Statistics\ 2D Binning 1. dat, OK.**

2. Zvýrazněte sloupec **A** a sloupec **B**, zvolte **Statistics, Descriptive Statistics, 2D Frequency Count/Binning** vyvolejte dialog **TwoDBinning** a pokračujte....



	A(X)	B(Y)
Long Name	x	y
Units		
Comments		
Sparklines		
1	51	76
2	44	64
3	35	72
4	14	70
5	49	64
6	43	59
7	12	67
8	55	69
9	53	78
10	18	75

3. Zadejte následující nastavení v dialogu dle obrázku vpravo:

3.1 Vyberte **Auto** ze **Recalculate**.

3.2 V oddíle **X** zrušte zaškrtnutí políčka **Auto** pro **Minimum Bin Beginning**, **Maximum Bin End** a **Bin Size** zadejte **40**, **60** a **5** do tří textových polí. Stejné parametry v oddíle **Y** nastavte na **50**, **70**, a **10** v tomto pořadí.

3.3 Vyberte **Sum** z **Quantity to Compute**.
Zaškrtněte políčko **Output Matrix**. V oboru **Matrix Plots** zkontrolujte obě **3D Bars** a **Image Plot**.

Statistics\Descriptive Statistics: twoDBinning

Dialog Theme [x]

Description Calculate frequencies on bivariate data

Recalculate Auto

Input [A2DBinning1]!"2D Binning 1"!"(A"x",B"y")"

"x"(X) [5 , 65]

Specify Binning Range by Bin Ends

Minimum Bin Beginning 40 Auto

Maximum Bin End 60 Auto

Step by Bin Size Number of Bins

Bin Size 5 Auto

Number of Bins 4

Periodical

Border Options

Output Binning Order Ascending

"y"(Y) [45 , 86]

Specify Binning Range by Bin Ends

Minimum Bin Beginning 50 Auto

Maximum Bin End 70 Auto

Step by Bin Size Number of Bins

Bin Size 10 Auto

Number of Bins 2

Periodical

Border Options

Output Binning Order Ascending

Quantity to Compute Sum

Column to Compute Quantity Y

Output Worksheet <new>

Subtotal Count for Each Binned Y

Output Matrix <new>

Matrix Plots

3D Bars

Image Plot

OK Cancel

4. Klepněte na **OK** a budete mít následující výstupy:

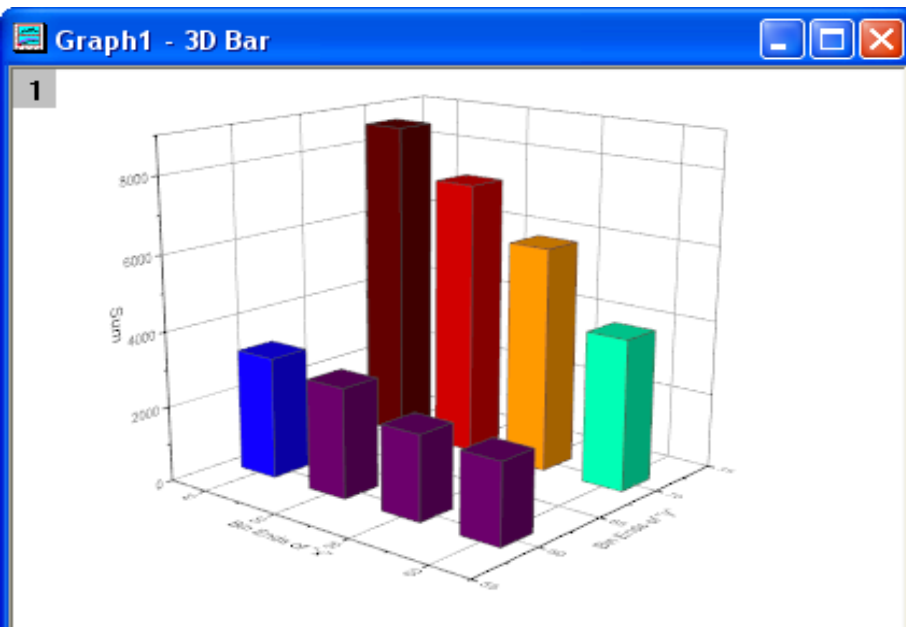
List

	A(X)	B(Y)	C(Y)
Long Name	Bin Ends of "x"	Sum	Sum
Bin Ends of "y"		60	70
Comments		50 - 60	60 - 70
1	45	3228	8596
2	50	2929	7324
3	55	2271	5985
4	60	2164	4033
5			
6			
7			
8			

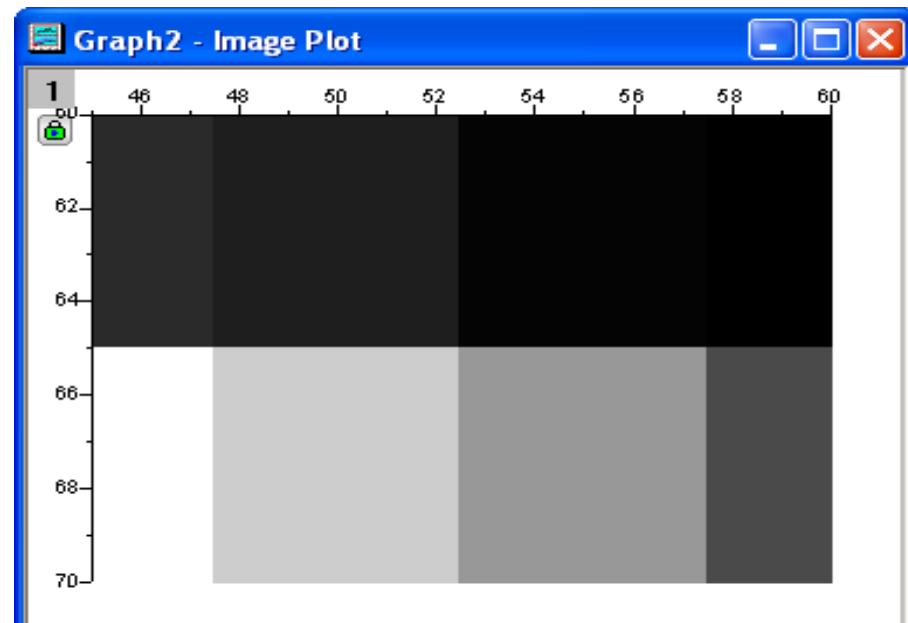
Matice

	1	2	3	4
1	3228	2929	2271	2164
2	8596	7324	5985	4033

3D-Sloupce



Obrázkový graf



5. Chcete-li dát odlehle hodnoty pro proměnnou Y do koše, klepněte na ikonu zámku v listu **TwoDBin1** a zvolte **Change Parameters**.

6. Ve větvi **Y** rozbalte uzel **Border Options**, pak zaškrtněte oba, **Include Outliers <Minimum** a **Include Outliers >= Maximum**.

7. Klikněte na **OK**. Dva sloupce pro odlehle hodnoty jsou přidány do listu **TwoDBin1**.

	A(X)	B(Y)	C(Y)	D(Y)	E(Y)
Long Name	Bin Ends of "x"	Sum	Sum	Sum	Sum
Bin Ends of "y"		50	60	70	80
Comments		< 50	50 - 60	60 - 70	>= 70
1	45	225	3228	8596	13440
2	50	315	2929	7324	13298
3	55	135	2271	5985	11193
4	60	135	2164	4033	9367
5					

Statistics\Descriptive Statistics: twoDBinning

Dialog Theme

Description Calculate frequencies on bivariate data

"y"(Y) [45 , 86]

Specify Binning Range by Bin Ends

Minimum Bin Beginning 50 Auto

Maximum Bin End 70 Auto

Step by Bin Size Number of Bins

Bin Size 10 Auto

Number of Bins 2

Periodical

Border Options

Include Outliers < Minimum

Include Outliers >= Maximum

Separately Count Minimum

Separately Count Maximum

Output Binning Order Ascending

Quantity to Compute Sum

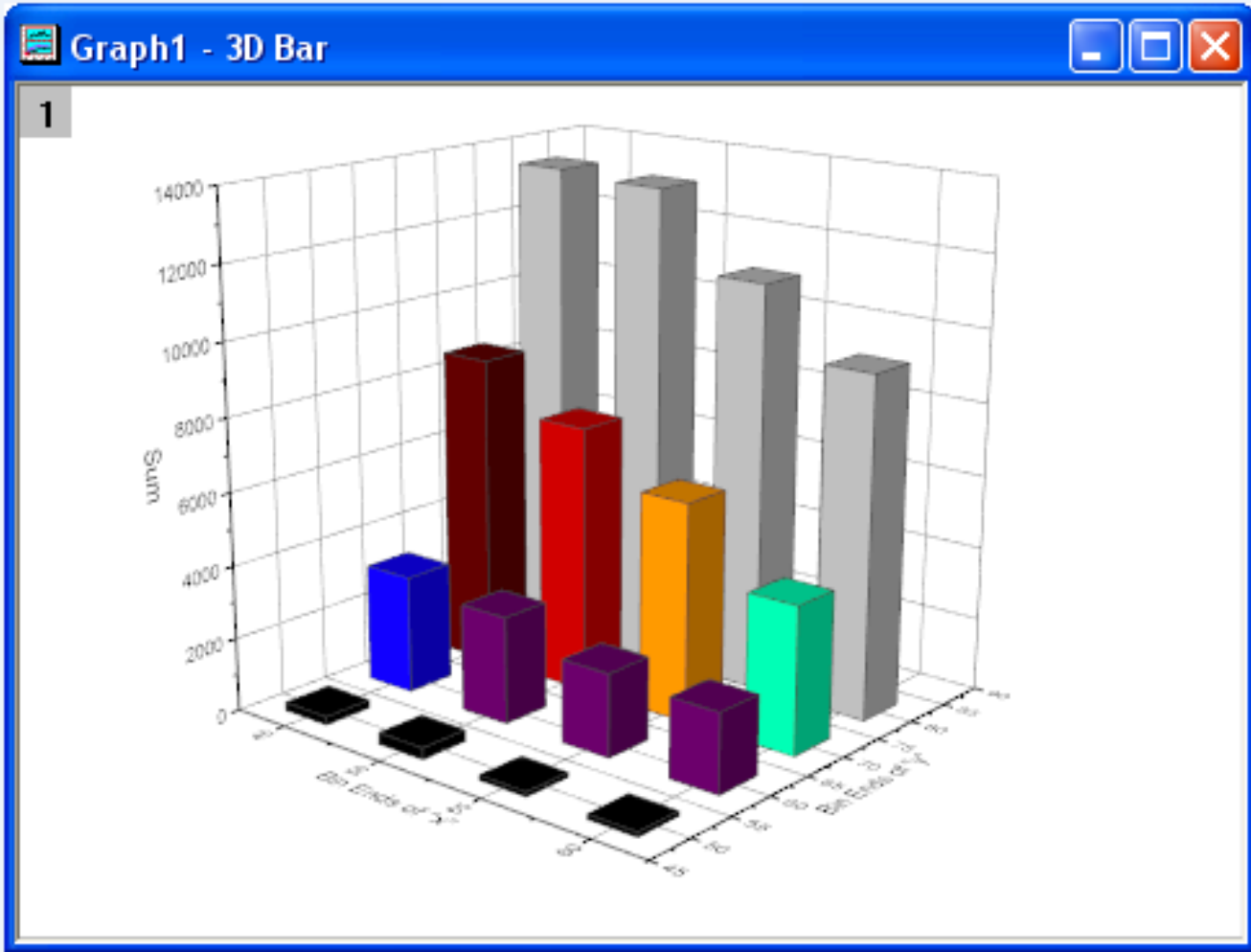
Column to Compute Quantity Y

Output Worksheet [A2DBinning1]TwoDBin1

Subtotal Count for Each Binned Y

OK Cancel

8. Dvojitým kliknutím na osu Y se otevře **Axis Dialog**, změňte měřítko **Scale From** na **To**, sice **45** a **90**. Klikněte na měřítko osy **Z** a změňte **To** na **14000**.



4.4 Analýza rozptylu (ANOVA)

4.4.1 One Way ANOVA

Existují dva druhy datových souborů v analýze rozptylu ANOVA – **data indexovaná** a **data původní**. Při provádění analýzy dat, není třeba používat celý datový soubor, Origin proto nabízí několik způsobů, jak účelně vybrat data. Například, lze použít tlačítko interaktivní **Regional Data Selector**, který vybere data, nebo lze použít dialogové okno **Column Browser**. Pomocí analýzy rozptylu se dozvíte, jak používat oba druhy dat, které mají provádět analýzy a jak vybrat data pomocí prohlížeče sloupce. ANOVA je druh parametrické metody k porovnání a rozšíření t-testu. Pokud existují více než dvě skupiny, které mají být porovnány a použití t-testu pro dvojice zde není vhodné a použije se ANOVA. Ta ale vyžaduje normalitu dat a shodné rozptyly. V opačném případě by měly být použity neparametrické metody ANOVA. Pro jednorozměrnou analýzu rozptylu **One-Way ANOVA** a použití indexovaného režimu dat jsou data organizována vždy ve dvou sloupcích: první sloupec pro faktor a druhý pro data.

Data indexovaná: První sloupec pro faktor a druhý pro data:

	A(Y)	B(Y)
Long Name	plant	nitrogen
Comments	Factor	Data
1	PLANT3	18.15473
2	PLANT3	12.90409
3	PLANT2	18.61197
4	PLANT1	17.7111
5	PLANT4	11.81661
6	PLANT3	11.68327
7	PLANT2	23.43165
8	PLANT2	14.01454

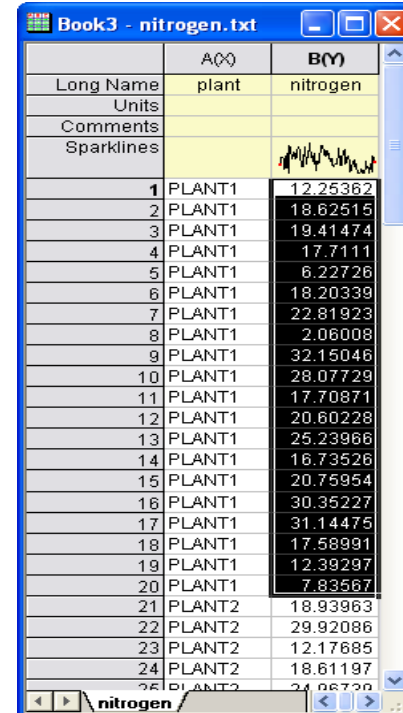
Data původní: data různé faktory jsou v různém sloupci.


	A(Y)	B(Y)	C(Y)	D(Y)
Long Name	Plant1	Plant2	Plant3	Plant4
Comments	Level1	Level2	Level3	Level4
1	17.7111	18.61197	18.15473	11.81661
2	32.15046	23.43165	12.90409	2.39438
3	17.70871	14.01454	11.68327	1.09914
4	28.07729	12.17685	23.52293	16.00756
5	7.83567	4.86902	16.00594	13.85077
6	2.06008	18.93963	3.04056	9.22245
7	22.81923	29.92086	14.29516	14.86523

Data indexovaná

Byl naměřen obsah dusíku v miligramech pro 4 druhy rostlin a je třeba vyšetřit, zda různé rostliny se liší v obsahu dusíku. Užijte se **One-Way ANOVA** v režimu indexovaných dat pro následující příklad:

1. Začněte s novým sešitem a importovat soubor **File, Import, Single ASCII, \Samples\Statistics\Nitrogen.txt, Open, OK**. Ujistěte se, že jste vybrali data s koncovkou ***.txt**. Za prvé je třeba provést test normality pro každý sloupec dat, zda jsou data normálního rozdělení.
2. Zvýrazněte první sloupec kliknutím pravé myši a vyberte **Worksheet, Sort Worksheet, Ascending**.
3. Zvýrazněte druhý sloupec od řádku 1 do řádku 20, které patří do **PLANT1**, otevřete **Statistics, Descriptive Statistics, Normality Test, OK**.
4. Použijte výchozí nastavení v dialogu a klikněte na **OK**. Z **p-hodnota Prob = 0,58545**, můžeme vidět "PLANT1" a následuje normální rozdělení.
5. Podobným způsobem si můžete zvýraznit dat "PLANT2", "PLANT3" a "PLANT4" a test normality. Náš výběr dat vykazuje normální rozdělení pro všechny rostliny (čili sloupce).
6. S aktivním listem otevřete dialog **Statistics, ANOVA, One-Way ANOVA, Open dialog**. Nastavte **Input Data** jako **Indexed**, přiřad'te sloupce „plant" a „nitrogen" jako **Factor** a **Data** pomocí tlačítek pravostranných šipek. Klepnutím na + rozbalte uzel **Means Comparison**, nastavte **Significance Level** na **0,05** a zvolte **Tukey Means Comparison Method**. Zvolte **Levene** | | z testů **Tests for Equal Variance**. Klikněte na **OK** a provede se **One-Way ANOVA**.



	A(0)	B(1)
Long Name	plant	nitrogen
Units		
Comments Sparklines		
1	PLANT1	12.25362
2	PLANT1	18.62515
3	PLANT1	19.41474
4	PLANT1	17.7111
5	PLANT1	6.22726
6	PLANT1	18.20339
7	PLANT1	22.81923
8	PLANT1	2.06008
9	PLANT1	32.15046
10	PLANT1	28.07729
11	PLANT1	17.70871
12	PLANT1	20.60228
13	PLANT1	25.23966
14	PLANT1	16.73526
15	PLANT1	20.75954
16	PLANT1	30.35227
17	PLANT1	31.14475
18	PLANT1	17.58991
19	PLANT1	12.39297
20	PLANT1	7.83567
21	PLANT2	18.93963
22	PLANT2	29.92086
23	PLANT2	12.17685
24	PLANT2	18.61197
25	PLANT2	21.06799

Vysvětlení: Z tabulky testu Homogeneity of Variance jednorozměrné ANOVA lze vidět, že čtyři skupiny mají stejný rozptyl, protože **p-hodnota Prob** je větší než **0,05**.

Homogeneity of Variance Test

Levene's Test(Absolute Deviations)

	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	18.06843	6.02281	0.34578	0.79229
Error	76	1323.76846	17.41801		

At the 0.05 level, the population variances are not significantly different.

Z výsledku celkové analýzy rozptylu můžeme konstatovat, že nejméně dvě skupiny čtyř mít významný různé prostředky, protože **p-hodnota Prob** je menší než 0,05.

Overall ANOVA

	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	1996.36652	665.45551	12.86214	6.99338E-7
Error	76	3932.05317	51.73754		
Total	79	5928.41969			

Null Hypothesis: The means of all levels are equal.
 Alternative Hypothesis: The means of one or more levels are different.
 At the 0.05 level, the population means are significantly different.

Pro další výzkum, je třeba rozšířit výsledky **Means Comparison**.

Means Comparisons

Tukey Test

	MeanDiff	SEM	q Value	Prob	Alpha	Sig	LCL	UCL
PLANT2 PLANT1	2.26308	2.27459	1.40706	0.75274	0.05	0	-3.71181	8.23796
PLANT3 PLANT1	-2.46538	2.27459	1.53284	0.70039	0.05	0	-8.44027	3.5095
PLANT3 PLANT2	-4.72846	2.27459	2.93989	0.16935	0.05	0	-10.70334	1.24643
PLANT4 PLANT1	-10.93833	2.27459	6.80085	4.38499E-5	0.05	1	-16.91322	-4.96345
PLANT4 PLANT2	-13.20141	2.27459	8.20791	8.24355E-7	0.05	1	-19.1763	-7.22653
PLANT4 PLANT3	-8.47295	2.27459	5.26801	0.00207	0.05	1	-14.44784	-2.49807

Sig equals 1 indicates that the means difference is significant at the 0.05 level.
 Sig equals 0 indicates that the means difference is not significant at the 0.05 level.

Zde je vidět, že má PLANT4 se liší ve srovnání s každým z ostatních tří.

Data původní

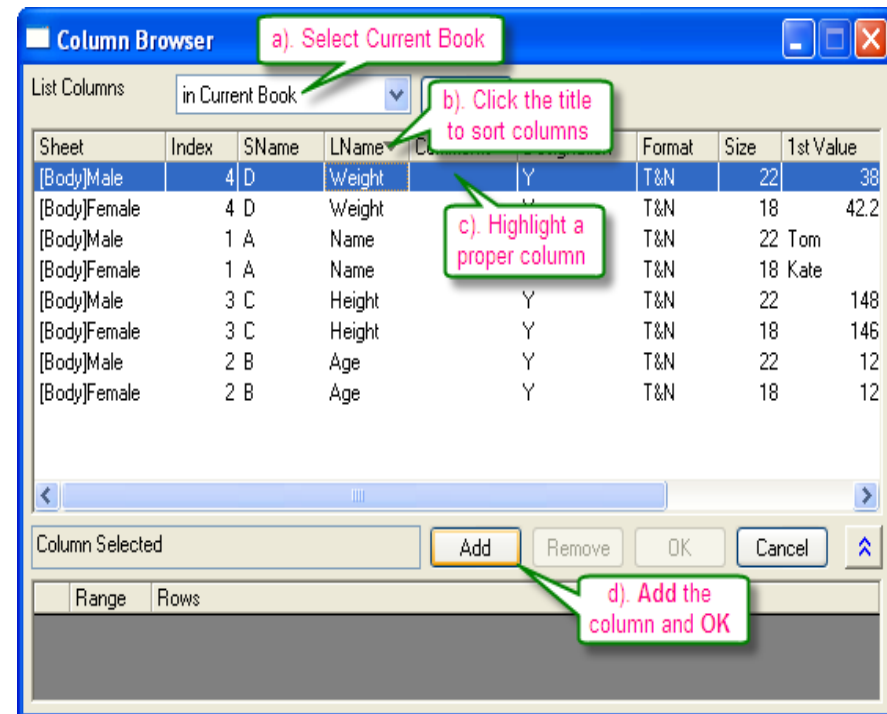
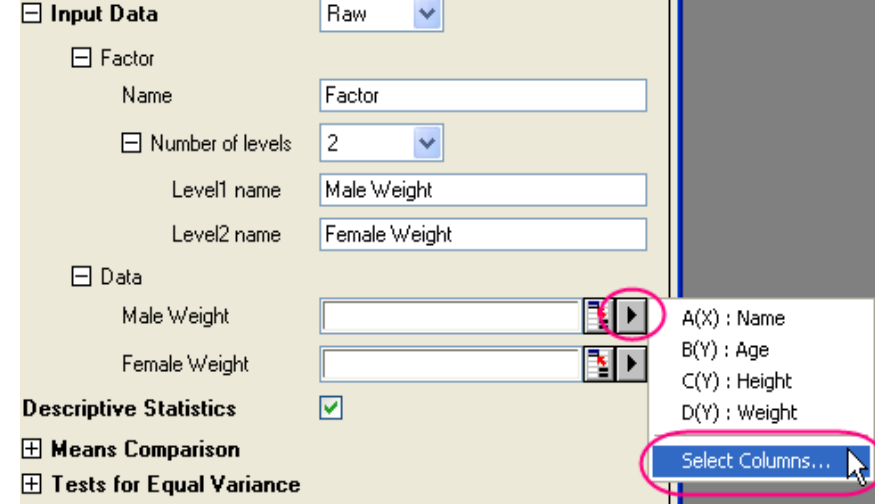
1. Vyberte **File, Open** a vyberte sešity **\Samples\Statistics\Body.ogw, Open**.

2. Zvolte **Statistics, ANOVA, One-Way ANOVA, Open Dialog**. V **Input Data** zvolte **Raw**. Zapište do **Level1 name** a **Level2 name** název **Male Weight** a **Female Weight**.

3. Nyní budeme používat **Data Browser** k zadání dat do **Data** bloku. Klikněte na trojúhelníkovou ikonku vedle **Male Weight** a zvolte **Select Columns...** k otevření dialogu **Column Browser**.

V dialogovém okně **Column Browser** zvolíte **in Current Book** ze **List Columns** listu, aby byly vidět všechny sloupce listu v aktuální knize.

Vyberte **Weight** v listu **[Body]Male** a klikněte na tlačítko **Add** a **OK** a přidejte ji do **Male Weight** editačního pole. Podobně, přiřadit **Weight** z **[Body]Female** do **Female Weight** editačního pole.



Column Browser

List Columns: in Current Book

a). Select Current Book

b). Click the title to sort columns

c). Highlight a proper column

d). Add the column and OK

Sheet	Index	SName	LName	Column	Format	Size	1st Value
[Body]Male	4	D	Weight	Y	T&N	22	38
[Body]Female	4	D	Weight	Y	T&N	18	42.2
[Body]Male	1	A	Name		T&N	22	Tom
[Body]Female	1	A	Name		T&N	18	Kate
[Body]Male	3	C	Height	Y	T&N	22	148
[Body]Female	3	C	Height	Y	T&N	18	146
[Body]Male	2	B	Age	Y	T&N	22	12
[Body]Female	2	B	Age	Y	T&N	18	12

Column Selected

Add Remove OK Cancel

Range	Rows

4. Přijmout další defaultní nastavení v dialogovém okně **ANOVAOneWay** a klikněte na **OK**. Z výstupní sestavy v poznámce pod čarou lze konstatovat, že na úrovni 0,05, hmotnost souboru mezi mužem a ženou se nijak výrazně neliší.

4.5 Neparametrické testy (Nonparametric Tests)

Neparametrické testy jsou používány, když není známo, zda data vykazují normální rozdělení, nebo se potvrdilo, že data nevykazují normální rozdělení. Neparametrické testy nevyžadují předpoklad normality v následujících úlohách:

- Malá velikost vzorku,
- Kategoriální (binární) pořadové údaje,
- Normální rozdělení nelze předpokládat.

4.5.1 Testy nezávislosti jednorozměrného výběru

Jednorozměrný Wilcoxonův znaménkový test je navržen k analýze mediánu souboru relativně k zadané hodnotě. Lze si vybrat jedno- nebo dvou-stranný test. Wilcoxonův znaménkový test hypotézy H_0 : medián = předpokládaný medián *versus* H_1 : Medián nepředpokládaný.

Příklad: Inženýr kvality ve výrobní hale se zajímá, zda střední hodnota (třeba průměr) hmotnosti produktu se rovná 166. U náhodného výběru 10 výrobků se bude měřit hmotnost. Naměřená data jsou 151.5 152.4 153.2 156.3 179.1 180.2 160.5 180.8 149.2 188.0 a zadejte je do **sloupce A**. Bude následovat test normality se zjištěním, zda rozdělení dat je normální.

Kroky:

1. Otevřete nový list a zadejte uvedené údaje ve **Column A**. Vyberte **Statistics, Descriptive Statistics, Normality Test...** a otevřete dialog **Normality Test**.
2. Otevřete uzel **Input Data** název sloupce **A(x)** v řádce **Data range**.
3. Klikněte na **OK** a dostanete výsledky:

Shapiro-Wilk				
	DF	Statistic	p-value	Decision at level(5%)
B	10	0.83472	0.03814	Reject normality

B: At the 0.05 level, the data was not significantly drawn from a normally distributed population.

Podle výsledku, **P-hodnota Prob = 0,03814**, distribuce dat není normálního rozdělení na hladině významnosti 0,05, a proto je třeba použít neparametrický test **One-Sample Wilcoxon Signed Rank test**.

Normality Test

Dialog Theme []

Description Perform Normality Test

Recalculate Manual [v]

Input Data

Range 1

Data Range [Book2]Sheet1!A

Grouping Range []

Quantities to Compute

Shapiro-Wilk

Kolmogorov-Smirnov

Lilliefors

Anderson-Darling

D'Agostino-K squared

Chen-Shapiro

Significance Level 0.05

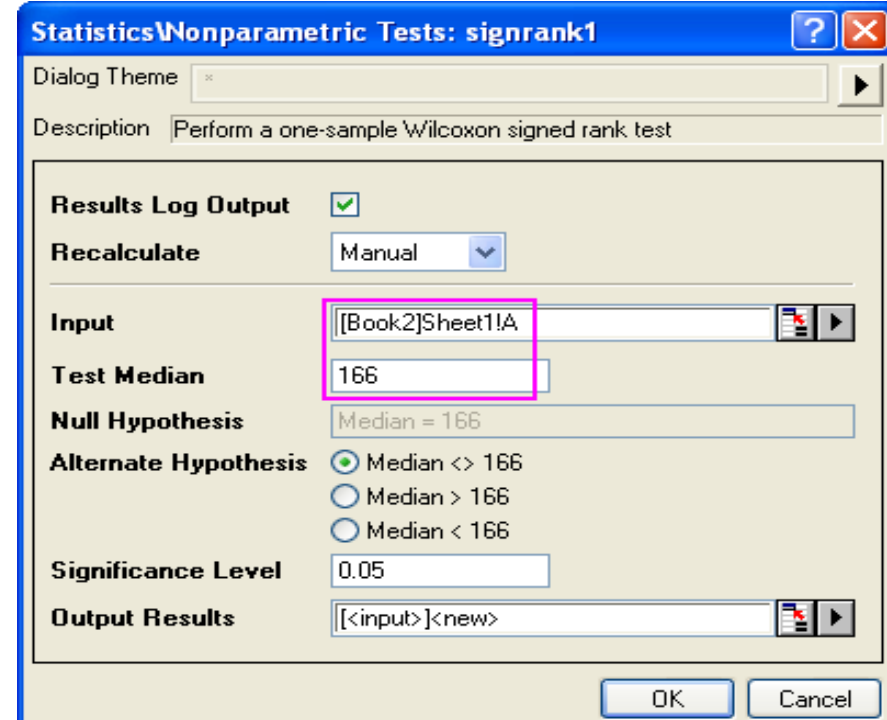
Output Settings

Plots

OK Cancel

Kroky:

1. Vyberte **Statistics, Nonparametric Tests, One-Sample Wilcoxon Signed Rank Test...** a otevře se dialogové okno.
2. Nastavte sloupec **A(x)** jako rozsah dat **Data Range**.
3. Zadejte **166** v textovém řádku **Test Median**.
3. Klikněte na **OK** a dostanete výsledky:



Descriptive Statistics

	N	Min	Q1	Median	Q3	Max
A	10	149.2	152.175	158.4	180.35	188

Test Statistics

	W	Z	Exact Prob > W	Asymp. Prob > W
A	28	0	1	1

Null Hypothesis: Median = 166

Alternative Hypothesis: Median <> 166

A: At the 0.05 level, the population median is NOT significantly different from the test median (166).

Podle modře psaného závěru v outputu výsledků nelze zamítnout nulovou hypotézu H_0 o rovnosti mediánu 166 na hladině významnosti 0,05.

4.5.2 Testy nezávislosti dvou výběrů

Existují dva neparametrické testy dvou výběrů nezávislého systému: **Mann-Whitney test** a **Two Sample Kolmogorov-Smirnov test**.

Příklad: praktické využití Mann-Whitney testu. Obroušený materiál z pneu (v mg) se měří pro dva typy pneumatik (A a B) a 8 experimentů bylo zde provedeno pro každý typ pneumatiky. Data jsou v **abrasion_indexed.dat**.

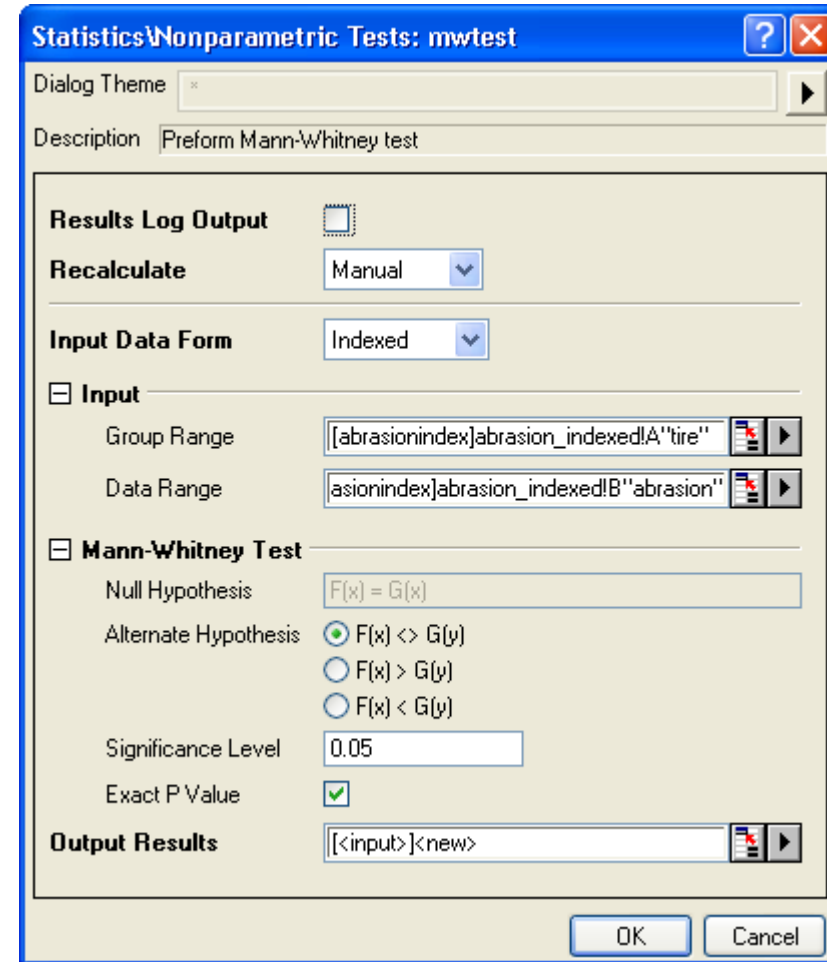
1. Importujte **File, Import, Single ASCII \Samples \Statistics\ abrasion_indexed.dat** Open, OK.

2. Vyberte **Statistics, Nonparametric Tests, Mann-Whitney Test**, abyste otevřeli dialogové okno.

3. Do **Input Data Form** zadejte **Indexed**.

4. Nastavte sloupec **A(x)** v **Group Range**, nastavte sloupec **B(y)** jako **Data Range**.

5. Zaškrtněte políčko **Exact P Value**.



6. Klikněte na **OK** pro spuštění výpočtu, což by mělo být v listu MannWhitney1.

Test Statistics

	U	Z	Exact Prob> U	Asymp. Prob> U
	34.5	0.2102	0.82191	0.83351

Null Hypothesis: $F(x) = G(y)$

Alternative Hypothesis: $F(x) \neq G(y)$

At the 0.05 level, the two distributions are NOT significantly different.

U: statistiku U lze jednoduše vypočítat z hodnoty dvou skupin. Jde o číslo, vyjadřující kolikrát je skóre ve 2. skupině je větší než skóre v 1. skupině.

Z: přibližná statistika testu normality. Poskytuje vynikající přiblížení s růstem velikosti výběru.

Exact Prob: přesná hodnota **p-value**, je k dispozici pouze tehdy, když **Exact P Value** je v dialogu zadána. Nicméně, může to být velmi CPU-časově náročné pro nadměrné velikosti výběru.

Asymp.Prob: asymptotická hodnota **p-value**, vypočtená z přibližné statistiky testu normality **Z**.

Book1

Mann-Whitney Test (3.3.2014 11:52:18)

Notes

X-Function	Mann-Whitney Test
User Name	mime0352
Time	3.3.2014 11:52:18

Input Data

	Data	Range
Group Range	[Book1]abrasion_indexed!A"tire"	[1*:16*]
Data Range	[Book1]abrasion_indexed!B"abrasion"	[1*:16*]

Descriptive Statistics

	N	Min	Q1	Median	Q3	Max
A	8	4870	4980	5760	7330	8650
B	8	4900	4950	5420	6687,5	7930

Ranks

	N	Mean Rank	Sum Rank
A	8	8,8125	70,5
B	8	8,1875	65,5

Test Statistics

	U	Z	Exact Prob> U	Asymp. Prob> U
	34,5	0,2102	0,82191	0,83351

Null Hypothesis: $F(x) = G(y)$
Alternative Hypothesis: $F(x) \neq G(y)$
At the 0.05 level, the two distributions are NOT significantly different.

abrasion_indexed \ MannWhitney1 /

4.5.3 Neparametrické testy korelace

Korelační koeficient je používán jako měřítko síly lineárního vztahu mezi dvěma proměnnými.

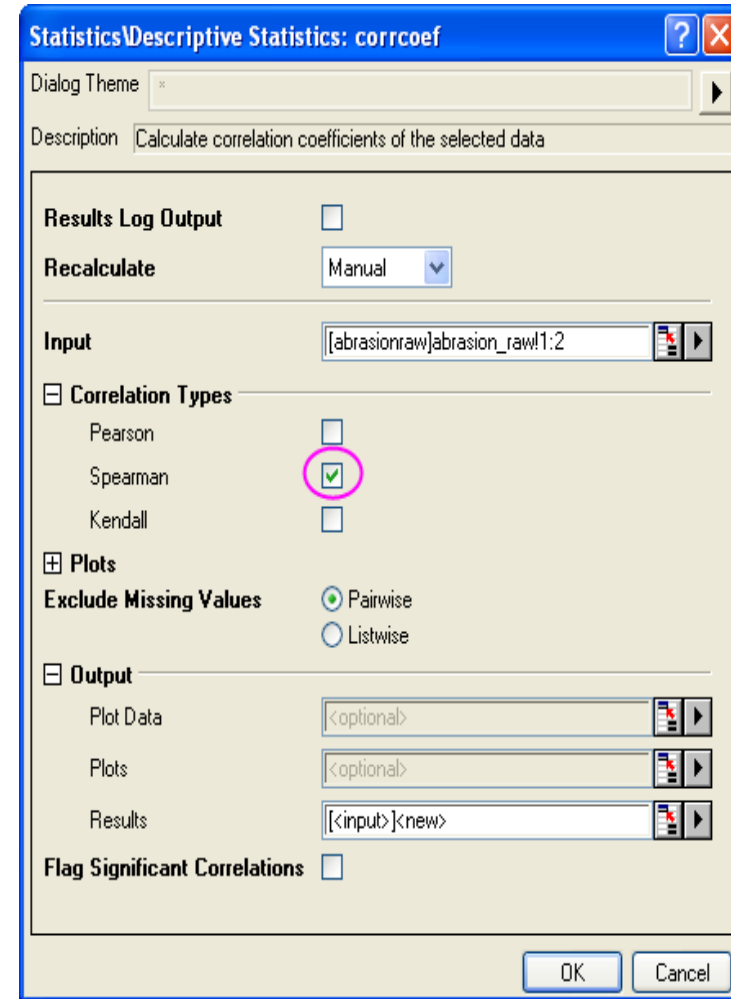
Existují dvě neparametrické metody výpočtu korelace mezi proměnnými :

- **Spearman:** společná náhrada Pearsonova korelačního koeficientu, Spearmaným koeficientem pokud obě proměnné závisle proměnná a nezávisle proměnná jsou pořadová čísla, nebo když jedna proměnná je pořadové číslo a druhá proměnná je spojitě číslo. Lze použít Spearmanovu korelaci také kdy obě proměnné jsou spojitě.

- **Kendall:** Používá se pro ordinální proměnné k posuzování shody mezi hodnotiteli.

Kroky:

1. Importujte **File, Import, Single ASCII \Samples \Statistics\ abrasion_raw.dat, Open, OK.**
2. Zvýrazněte sloupec sloupec **B**. Vyberte **Statistics, Descriptive Statistics, Correlation Coefficient** k otevření dialogu **corrcoef**.
3. Zaškrtněte **Spearman** a zrušte zaškrtnutí **Pearson**.



4. Klikněte na tlačítko **OK** pro provedení výpočtu a generování výsledků v listu **CorrCoef1**.

Z hodnoty **Spearman Corr.**, lze dojít k závěru, že oděr mezi pneumatikami A a B spolu silně souvisí.

Spearman Correlations

		tireA	tireB
"tireA"	Spearman Corr.	1	0.90476
"tireA"	Sig.	--	0.00201
"tireB"	Spearman Corr.	0.90476	1
"tireB"	Sig.	0.00201	--

2-tailed test of significance is used

Book1

1 Correlations Coefficients (3.3.2014 12:00:39)

Notes

X-Function	Correlations Coefficients
User Name	mime0352
Time	3.3.2014 12:00:39

Input Data

	Data	Range
tireA	[Book1]abrasion_raw!A"tireA"	[1*:8*]
tireB	[Book1]abrasion_raw!B"tireB"	[1*:8*]

Descriptive Statistics

	N	Mean	SD	Sum	Min	Max
"tireA"	8	6145	1366,49709	49160	4870	8650
"tireB"	8	5825	1097,46461	46600	4900	7930

Pearson Correlations

		tireA	tireB
"tireA"	Pearson Corr.	1	0,99006
"tireA"	Sig.	--	2,43892E-6
"tireB"	Pearson Corr.	0,99006	1
"tireB"	Sig.	2,43892E-6	--

2-tailed test of significance is used

Spearman Correlations

		tireA	tireB
"tireA"	Spearman Corr.	1	0,90476
"tireA"	Sig.	--	0,00201
"tireB"	Spearman Corr.	0,90476	1
"tireB"	Sig.	0,00201	--

2-tailed test of significance is used

Kendall Correlations

		tireA	tireB
"tireA"	Kendall Corr.	1	0,78571
"tireA"	Sig.	--	0,00649
"tireB"	Kendall Corr.	0,78571	1
"tireB"	Sig.	0,00649	--

2-tailed test of significance is used

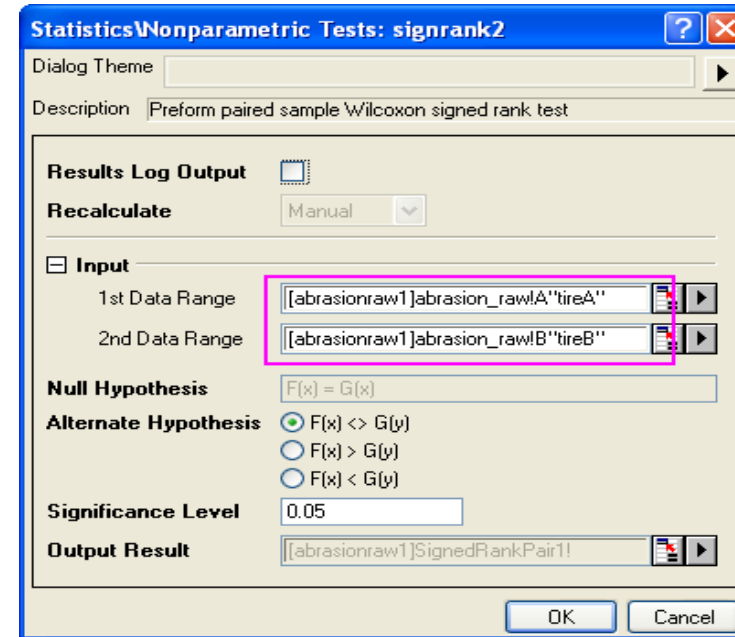
ScatterMatrix1 CorrCoef1

4.5.4 Párový Wilcoxonův znaménkový test

Budeme porovnávat dva mediány pneumatik A a pneumatiky B z předešlého příkladu.

Kroky:

1. Importujte **File, Import, Single ASCII \Samples \Statistics\ abrasion_raw.dat, Open, OK.**
2. Vyberte **Statistics, Nonparametric Tests, Paired Sample Wilcoxon Signed Rank Tests**
3. Zvolte **Column A** jako **1st Range Data** a **Column B** jako **2nd Range Data.**
4. Klikněte na tlačítko **OK** pro provedení výpočtu a generování výsledků.
5. Můžeme konstatovat, že dva mediány jsou výrazně odlišné. Je zřejmé, že medián skupiny A je větší než medián skupiny B.



Descriptive Statistics

	N	Min	Q1	Median	Q3	Max
"tireA"	8	4870	4980	5760	7330	8650
"tireB"	8	4900	4950	5420	6687.5	7930

Ranks

		N	Mean Rank	Sum Rank
"tireB"- "tireA"	Positive Ranks	2	1.5	3
	Negative Ranks	6	5.5	33

Test Statistics

	W	Z	Exact Prob> W	Asymp. Prob> W
	33	2.0329	0.03906	0.04206

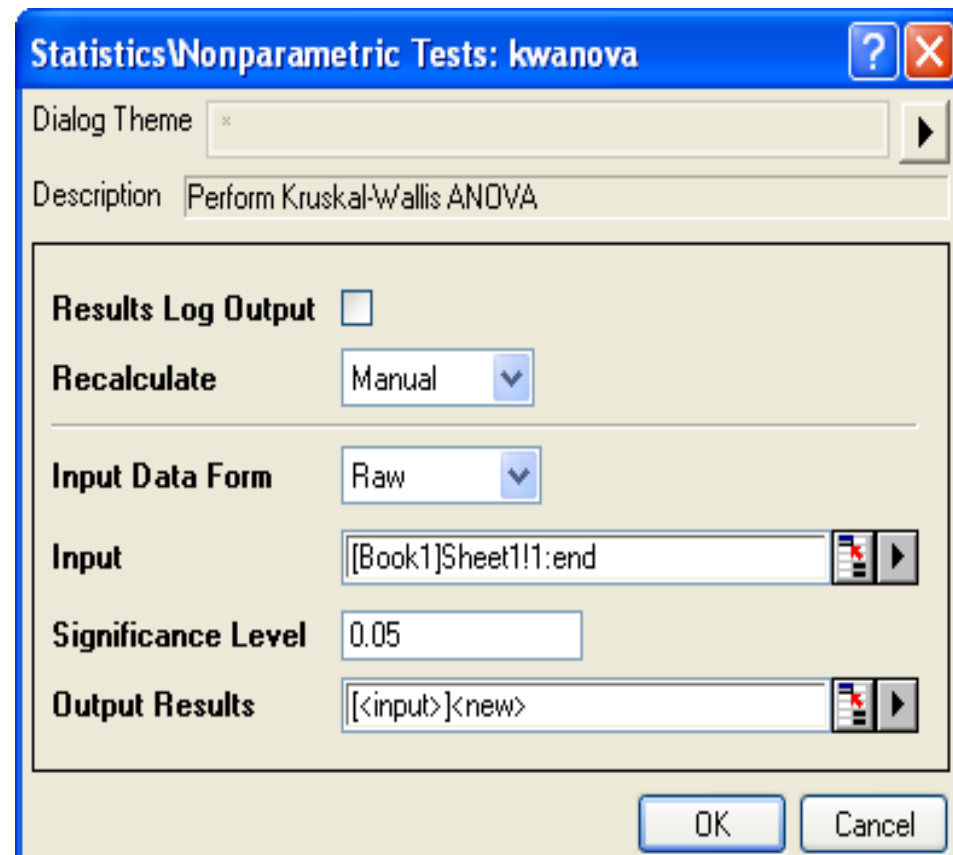
Null Hypothesis: F(x) = G(y)
Alternative Hypothesis: F(x) <> G(y)
At the 0.05 level, the two distributions are significantly different.

4.5.5 Test nezávislosti vícenásobných výběrů

Příklad: Je sledována spotřeba benzínu mpg u 4 aut. Některé experimenty jsou opakovány. K vyhodnocení počtu najetých kilometrů u čtyř aut na 1 galon benzínu a testu zda je tento počet stejný pro všechny auta, a které auto je nejúčinnější se použije Kruskal-Wallisův test rozptylu.

GMC/mpg	Infinity/mpg	Saab/mpg	Kia/mpg
26.1	32.2	24.5	28.4
28.4	34.3	23.5	34.2
24.3	29.5	26.4	29.5
26.2	35.6	27.1	32.2
27.8	32.5	29.9	
30.6	30.2		
28.1			

1. Vytvořte nový sešit v původu, zkopírujte ukázková data.
2. Vyberte **Statistics, Nonparametric Tests, Kruskal-Wallis ANOVA**,
3. Zadejte **Raw** jako **Input Data Form**.
4. Klepnutím na **trojúhelníkové tlačítko** vedle **Input** vyberte **All Columns** v menu.
5. Klikněte na **OK** ke generování výsledků, které jsou uloženy v listu **KWANOVA1**.



Test Statistics

Chi-Square	DF	Prob>Chi-Square
12.596	3	0.0055958

Null Hypothesis: The samples come from the same population
Alternative Hypothesis: The samples come from different populations
: At the 0.05 level, the populations are significantly different

Z p-hodnoty **Prob** můžeme konstatovat, že počet najetých kilometrů ze čtyř aut je významně odlišný.

Ranks

	N	Mean Rank	Sum Rank
"GMC/mpg"	7	7.7857	54.5
"Infinity/mpg"	6	17.833	107
"Saab/mpg"	5	6.2	31
"Kia/mpg"	4	15.125	60.5

Z hodnoty tabulky můžeme konstatovat, že auto **Infinity** je nejučinnější vůz.

Book1

Kruskal-Wallis ANOVA (3.3.2014 12:06:38)

Notes

X-Function	Kruskal-Wallis ANOVA
User Name	mime0352
Time	3.3.2014 12:06:38

Input Data

	Data	Range
A	[Book1]Sheet1!A	[1*:7*]
B	[Book1]Sheet1!B	[1*:7*]
C	[Book1]Sheet1!C	[1*:6*]
D	[Book1]Sheet1!D	[1*:5*]
E	[Book1]Sheet1!E	[1*:4*]
F	[Book1]Sheet1!F	[1*:0*]

Descriptive Statistics

	N	Min	Q1	Median	Q3	Max
A	7	24,3	26,1	27,8	28,4	30,6
B	6	29,5	30,025	32,35	34,625	35,6
C	5	23,5	24	26,4	28,5	29,9
D	4	28,4	28,675	30,85	33,7	34,2

Ranks

	N	Mean Rank	Sum Rank
A	7	7,78571	54,5
B	6	17,83333	107
C	5	6,2	31
D	4	15,125	60,5

Test Statistics

Chi-Square	DF	Prob>Chi-Square
12,59645	3	0,0056

Null Hypothesis: The samples come from the same population.
Alternative Hypothesis: The samples come from different populations.
: At the 0.05 level, the populations are significantly different.

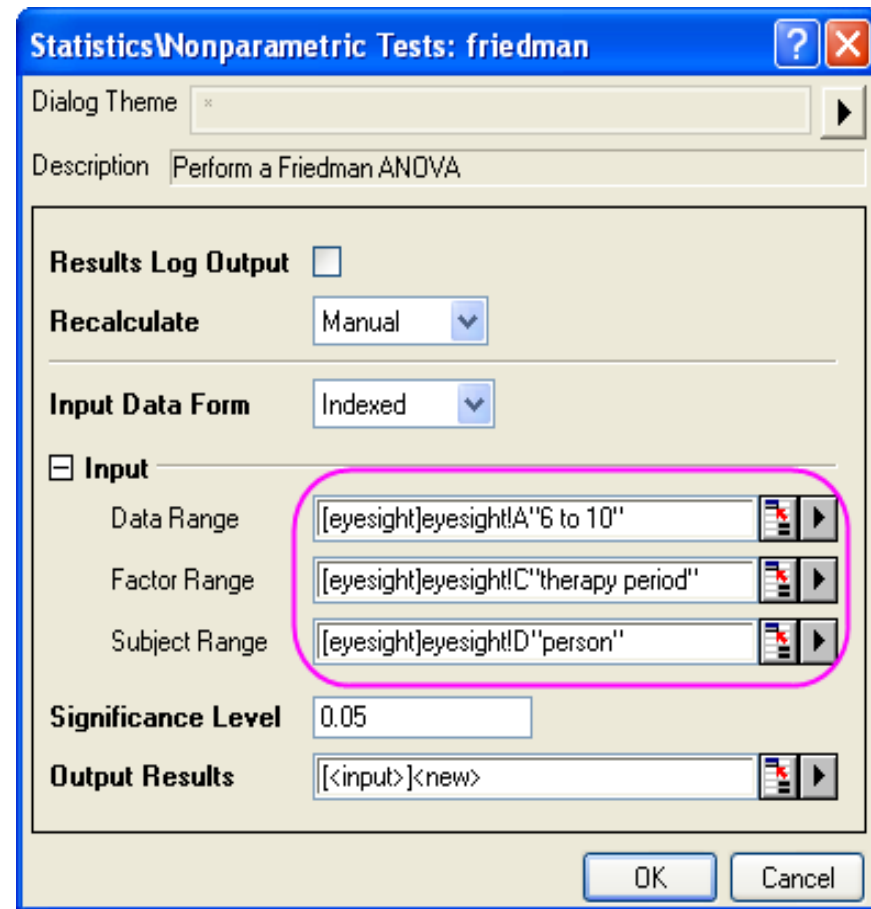
Sheet1 KWANOVA1

4.5.5 Test vícenásobných výběrů

Ophthalmolog vyšetřuje, zda léčba He-Ne laserem je použitelná také u dětí. Data obsahují 2 skupiny dětí ve věku 6-10 let a 11-16 let. Každý datový výběr obsahuje studii rozdílu pouhým okem 5ti osob a vždy po 3 období léčby. Výsledky jsou **eyesight.dat**. Vzhledem k malé velikosti vzorku byly použity neparametrické statistiky v analýze:

1. Importovat **File, Import, Single ASCII**, z **\Samples\Statistics\eyesight.dat**, **Open, OK**.
2. Vyberte **Statistics, Nonparametric Tests, Friedman ANOVA**.
3. Vyberte **Column A** za **Data Range**, **Column C** za **Factor Range** a **Column D** za **Subject Range**.
4. Klikněte na **OK** pro generování výsledků.

P-hodnota 0.0067379 je menší než 0,05. Soubor se proto značně liší, což naznačuje, že léčba je účinná pro věkovou skupinu 6-10.



Test Statistics

Chi-Square	DF	Prob>Chi-Square
10	2	0.0067379

Null Hypothesis: The samples come from the same population
Alternative Hypothesis: The samples come from different populations
At the 0.05 level, the populations are significantly different

1 Friedman ANOVA (3.3.2014 12:11:23)

Notes

X-Function	Friedman ANOVA
User Name	mime0352
Time	3.3.2014 12:11:23

Input Data

	Data	Range
Data Range	[Book1]eyesight!A"6 to 10"	[1*:15*]
Factor Range	[Book1]eyesight!C"therapy period"	[1*:15*]
Subject Range	[Book1]eyesight!D"person"	[1*:15*]

Descriptive Statistics

	N	Min	Q1	Median	Q3	Max
1	5	0,072	0,075	0,088	0,1375	0,165
2	5	0,038	0,0415	0,057	0,084	0,093
3	5	0,045	0,0525	0,078	0,095	0,098

Ranks

	N	Mean Rank	Sum Rank
1	5	3	15
2	5	1	5
3	5	2	10

Test Statistics

	Chi-Square	DF	Prob>Chi-Square
	10	2	0,00674

Null Hypothesis: The samples come from the same population.

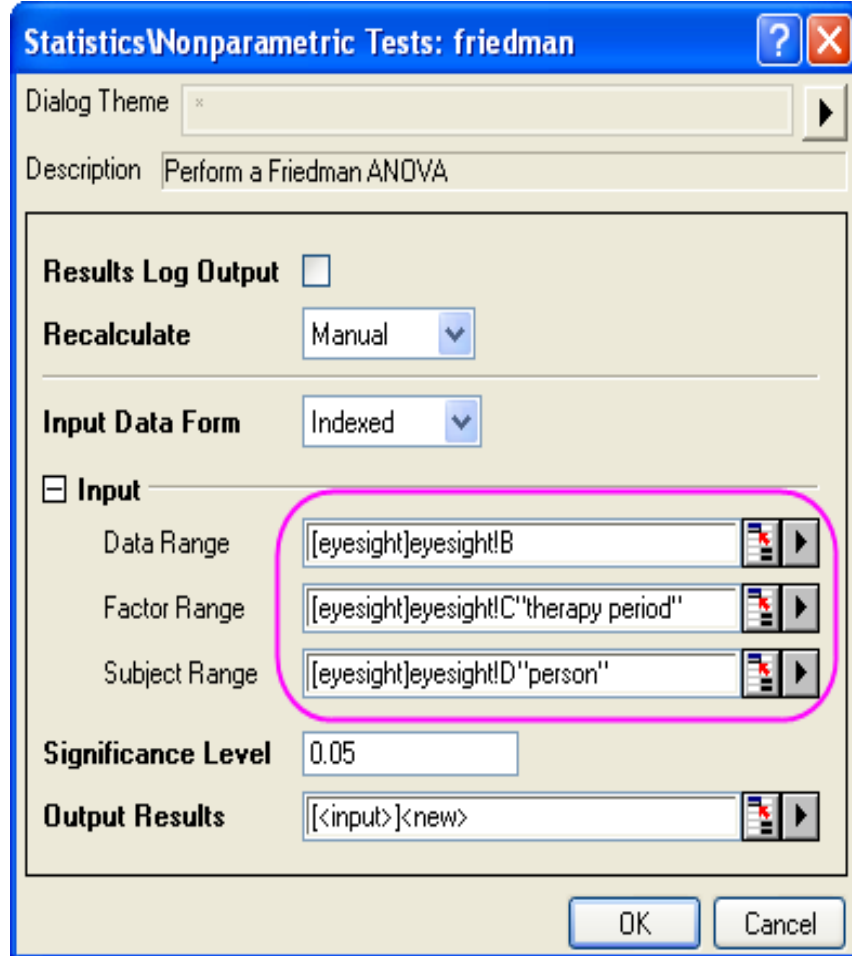
Alternative Hypothesis: The samples come from different populations.

At the 0.05 level, the populations are significantly different

Podobným způsobem vyberte **Column B** za **Data Range** a zbytek nastavení bude stejný jako předešle.

Na výsledek je vidět, že **p-hodnota Prob** 0,02599 je menší než 0,05 nebo 0,10. Lze také konstatovat, že zrak na 11-16 roků starých dětí je lepší po 3 obdobích léčby.

Z hodnot Chi-Square je zřejmé, že léčba laserem He-Ne funguje lépe u 6-10 roků starých dětí. Starší děti musí být více zapojeny v terapii, aby se jejich zrak zlepšil.



Test Statistics

	Chi-Square	DF	Prob>Chi-Square
	7.3	2	0.025991

Null Hypothesis: The samples come from the same population

Alternative Hypothesis: The samples come from different populations

At the 0.05 level, the populations are significantly different

1 Friedman ANOVA (3.3.2014 12:13:54)

Notes

X-Function	Friedman ANOVA
User Name	mime0352
Time	3.3.2014 12:13:54

Input Data

	Data	Range
Data Range	[Book1]eyesight!B"11-16"	[1*:15*]
Factor Range	[Book1]eyesight!C"therapy period"	[1*:15*]
Subject Range	[Book1]eyesight!D"person"	[1*:15*]

Descriptive Statistics

	N	Min	Q1	Median	Q3	Max
1	5	0,054	0,054	0,064	0,1375	0,172
2	5	0,033	0,0345	0,054	0,073	0,079
3	5	0,005	0,018	0,046	0,0785	0,095

Ranks

	N	Mean Rank	Sum Rank
1	5	2,9	14,5
2	5	1,9	9,5
3	5	1,2	6

Test Statistics

	Chi-Square	DF	Prob>Chi-Square
	7,3	2	0,02599

Null Hypothesis: The samples come from the same population.

Alternative Hypothesis: The samples come from different populations.

At the 0.05 level, the populations are significantly different

4.6 Vícerozměrná statistická analýza (Multivariate Analysis)

4.6.1 Metoda hlavních komponent (Principal Component Analysis)

Analýza hlavních komponent je užitečná pro snížení rozměrů a interpretaci velkých vícerozměrných datových souborů o lineární struktuře a pro objevování dosud netušených skrytých vztahů. Začnete s daty spotřeby bílkovin v pětadvaceti zemích Evropy v 9 druzích potravin. Použití Principal Component Analysis budete zkoumat vztah mezi zdroji bílkovin a těmito evropskými zeměmi.

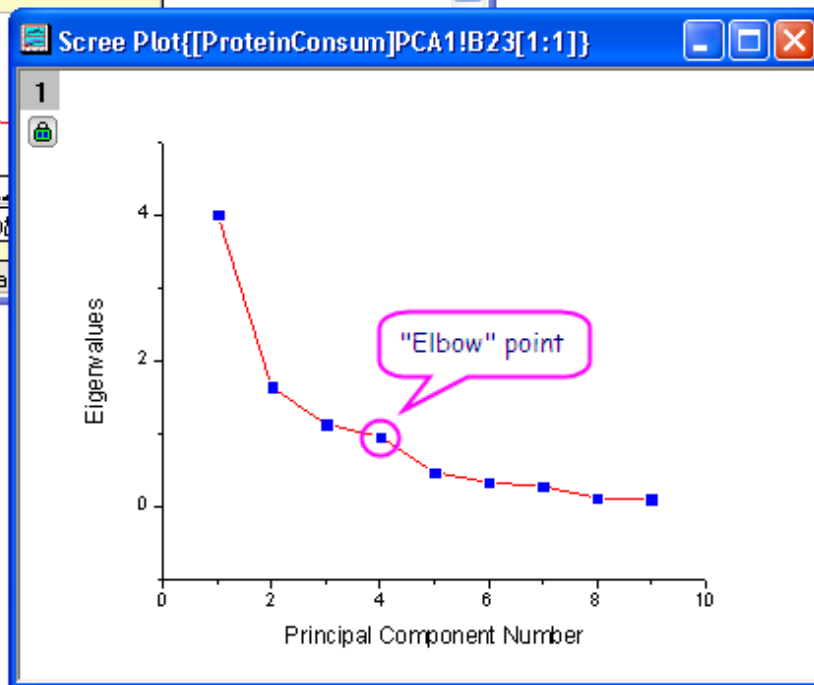
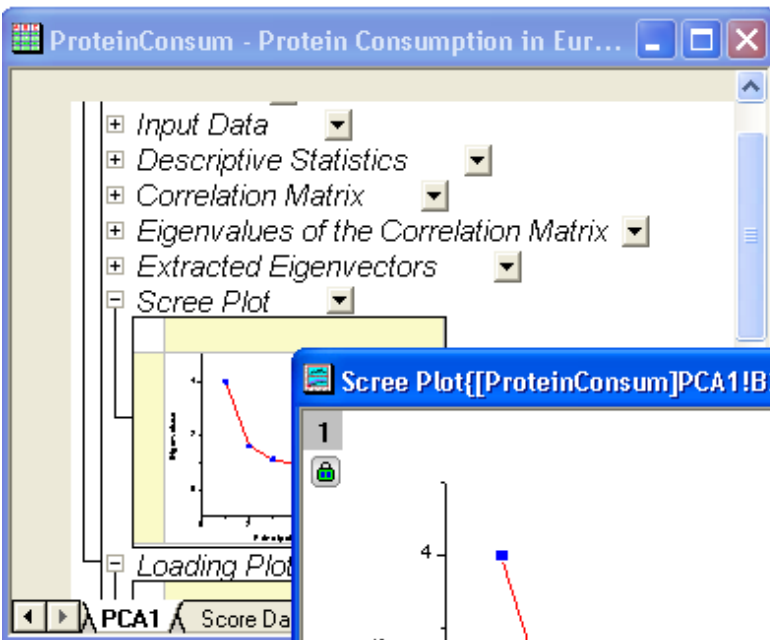
Kroky:

1. Otevřít nový projekt a importovat **File, Import, Single ASCII, \Samples\Statistics\Protein Consumption in Europe.dat, Open, OK** .
2. Vyberte celý list a pak **Statistics: Multivariate Analysis: Principal Component Analysis**.
3. Přijměte výchozí nastavení v otevřeném dialogovém okně a klepněte na **OK**.
4. Vyberte list **PCA1**.
5. Ve **Eigenvalues of the Correlation Matrix** lze vidět, že první čtyři hlavní komponenty vysvětlují 86% rozptylu a zbývající komponenty přispívají do něho 5% nebo méně. Budeme proto sledovat první čtyři hlavní komponenty.

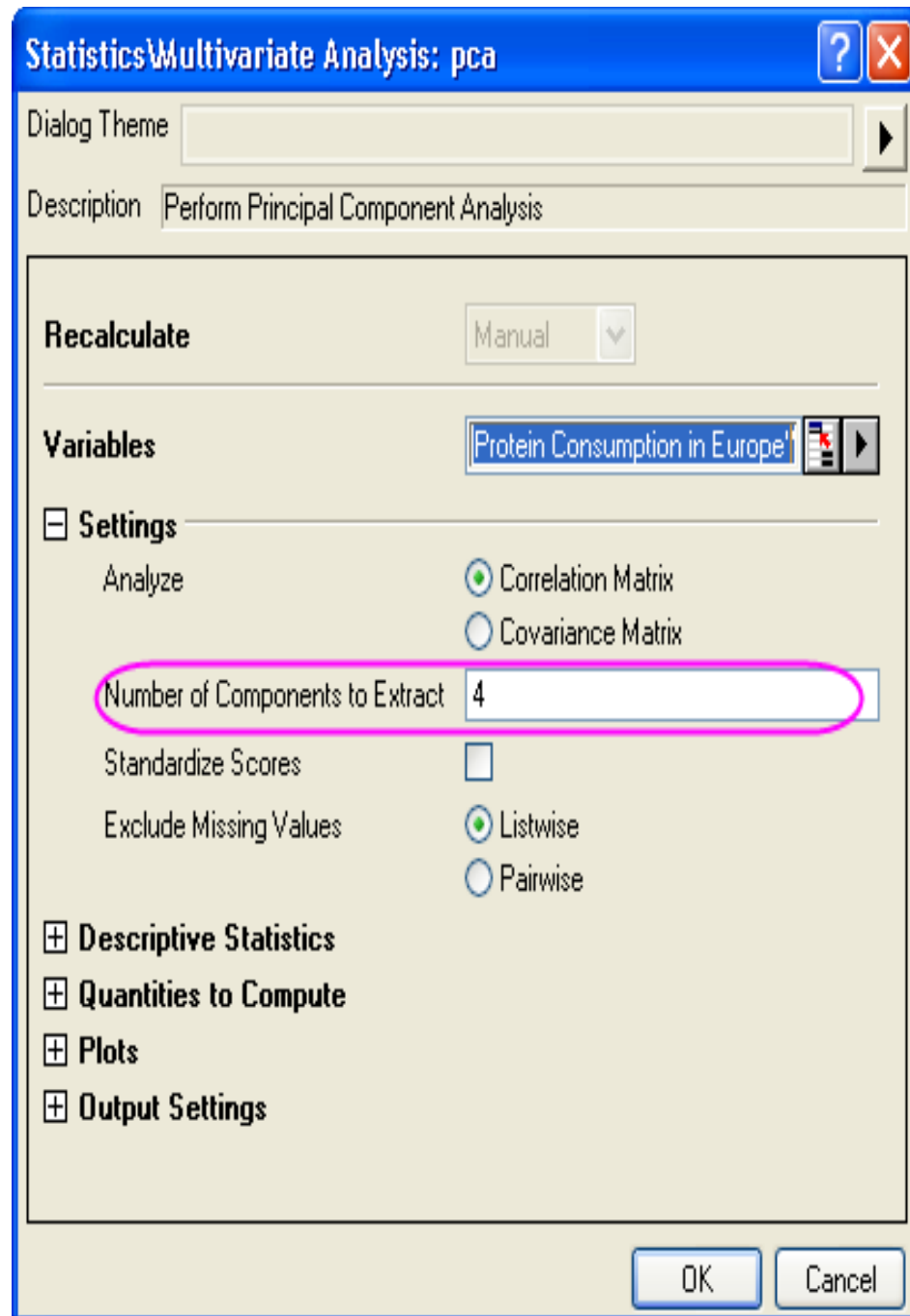
6. Scree plot je užitečnou vizuální pomůckou k určení vhodného počtu hlavních komponent. Počet složek závisí na zlomovém bodě, ve kterém jsou zbývající vlastní čísla relativně malá a o stejné velikosti. Tento bod není sice příliš zřetelný, ale přesto lze říci, že čtvrtý bod je zlomový bod.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Percentage of Variance	Cumulative
1	4.00644	44.52%	44.52%
2	1.635	18.17%	62.68%
3	1.12792	12.53%	75.22%
4	0.95466	10.61%	85.82%
5	0.46384	5.15%	90.98%
	0.32513	3.61%	94.59%
	0.27161	3.02%	97.61%
	0.11629	1.29%	98.90%
	0.09911	1.10%	100.00%



7. Klikněte na ikonu zámku ve výsledcích stromu a zvolte **Change Parameters** v menu. Nastavení počtu složek k extrahování **Number of Components to Extract** na **4**. Nezávírejte dialogové okno, v dalších krocích budeme načítat diagramy komponent.



Požadavek metody hlavních komponent

Ve větvi **Plots** si uživatel může vybrat, zda chce vytvořit sutinový graf nebo diagram hlavních komponent.

- **Sutinový graf (Scree Plot):** Sutinový graf je užitečná vizuální pomůcka pro určení vhodného počtu hlavních komponent.

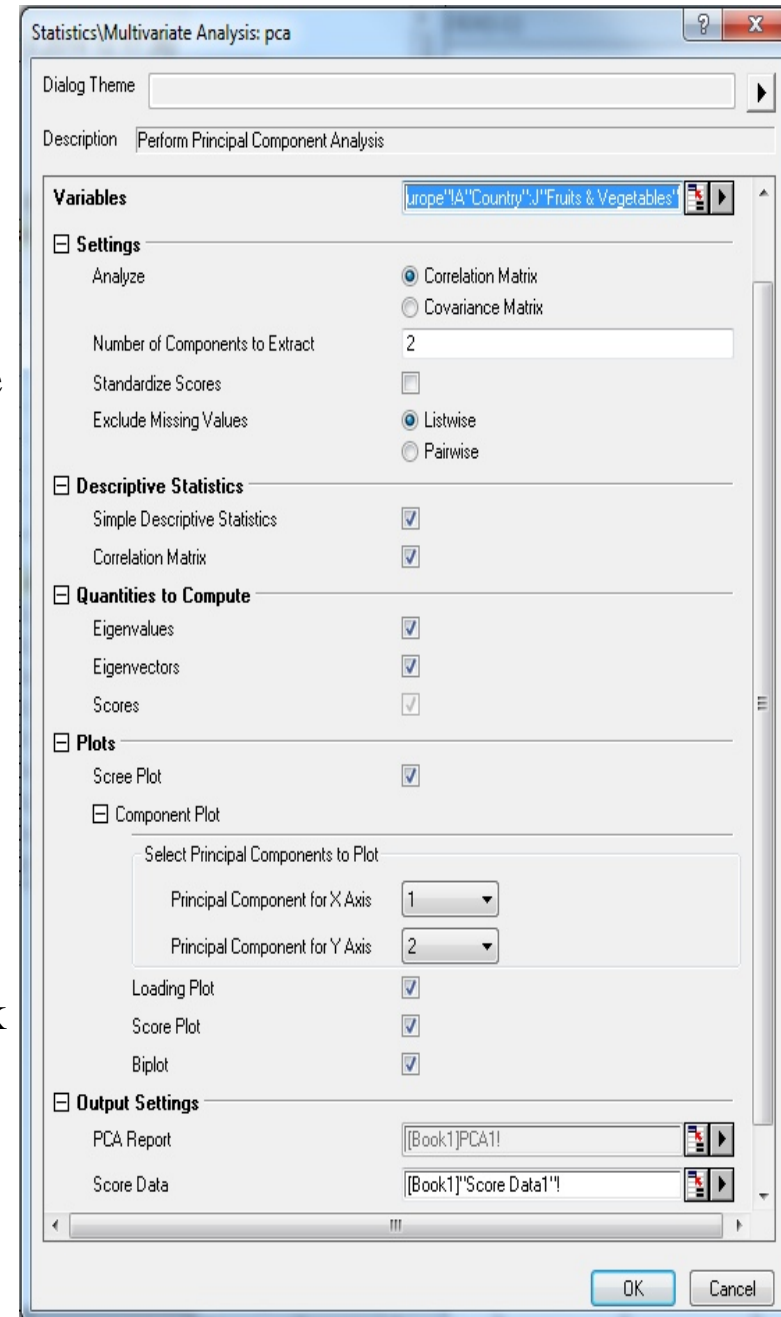
- **Komponentní grafy (Component Plots):** Vyberte **Select Principal Components to Plot**, kde lze zadat, který pár hlavních komponent se vynese do grafu.

Komponentní grafy zahrnují:

1. **Graf komponentních vah (Loading Plot):** jde o graf vztahu mezi původními proměnnými a dimenzemi podprostoru. Používá se k interpretaci vztahů mezi proměnnými.

2. **Graf komponentního skóre (Score Plot):** jde o graf projekce původních dat do subprostoru. Používá se k interpretaci vztahů mezi pozorováními.

- **Dvojný graf (BiPlot):** ukazuje, jak na komponentní váhy, tak i na komponentní skóre dvou vybraných komponent.



1. V dialogu předcházejících kroků otevřete větev **Plots**. Ujistěte se, že jsou vybrány **Scree Plot**, **Loading Plot** a **Biplot**.

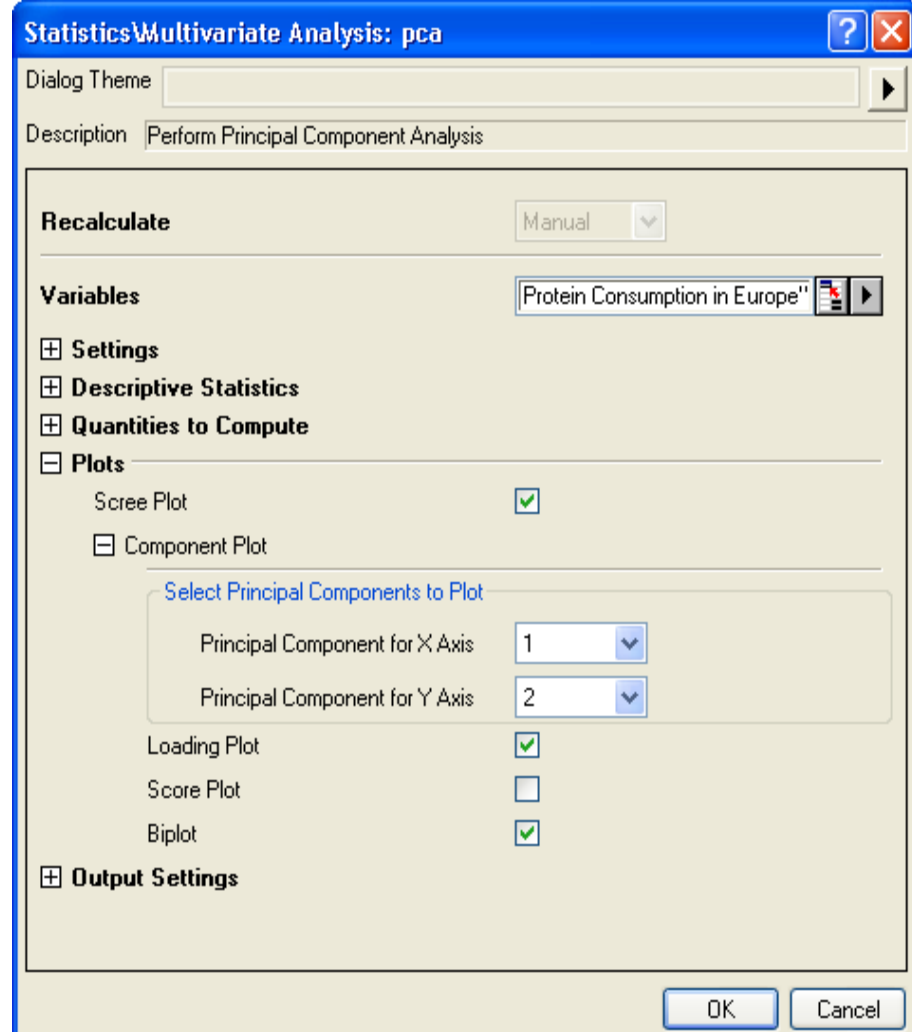
2. První dvě složky jsou obvykle zodpovědné za velkou část rozptylu. To je důvod, proč se vynáší graf v prostoru prvních dvou hlavních komponent. Vyberte **Select Principal Components to Plot**, nastavte 1. hlavní komponentu na osu X a 2. hlavní komponentu na osu Y. Klikněte na **OK**.

Vysvětlení výsledků:

1. Z korelační matice je vidět, že proměnné jsou vysoce korelované. Mnoho hodnot je větších než 0,3. Analýza hlavních komponent je proto vhodným nástrojem k odstranění kolinearity.

Correlation Matrix

	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fruits & Vegetables
Red Meat	1	0.153	0.58561	0.50293	0.06096	-0.49988	0.13543	-0.34945	-0.07422
White Meat	0.153	1	0.62041	0.28148	-0.23401	-0.4138	0.31377	-0.63496	-0.06132
Eggs	0.58561	0.62041	1	0.57553	0.06557	-0.71244	0.45223	-0.55978	-0.04552
Milk	0.50293	0.28148	0.57553	1	0.13788	-0.59274	0.22241	-0.62109	-0.40836
Fish	0.06096	-0.23401	0.06557	0.13788	1	-0.52423	0.40385	-0.14715	0.26614
Cereals	-0.49988	-0.4138	-0.71244	-0.59274	-0.52423	1	-0.53326	0.651	0.04655
Starch	0.13543	0.31377	0.45223	0.22241	0.40385	-0.53326	1	-0.47431	0.08441
Nuts	-0.34945	-0.63496	-0.55978	-0.62109	-0.14715	0.651	-0.47431	1	0.37497
Fruits & Vegetables	-0.07422	-0.06132	-0.04552	-0.40836	0.26614	0.04655	0.08441	0.37497	1



2. Hlavní komponenty jsou definovány jako lineární kombinace původních proměnných. Tabulka **Extracted Eigenvectors** poskytuje koeficienty pro definiční rovnice hlavních komponent.

	Coefficients of PC1	Coefficients of PC2	Coefficients of PC3	Coefficients of PC4
Red Meat	0.30261	-0.05625	-0.29758	0.64648
White Meat	0.31056	-0.23685	0.6239	-0.03699
Eggs	0.42668	-0.03534	0.18153	0.31316
Milk	0.37773	-0.18459	-0.38566	-0.00332
Fish	0.13565	0.64682	-0.32127	-0.21596
Cereals	-0.43774	-0.23349	0.09592	-0.0062
Starch	0.29725	0.35283	0.24298	-0.33668
Nuts	-0.42033	0.14331	-0.05439	0.33029
Fruits & Vegetables	-0.11042	0.53619	0.40756	0.46206

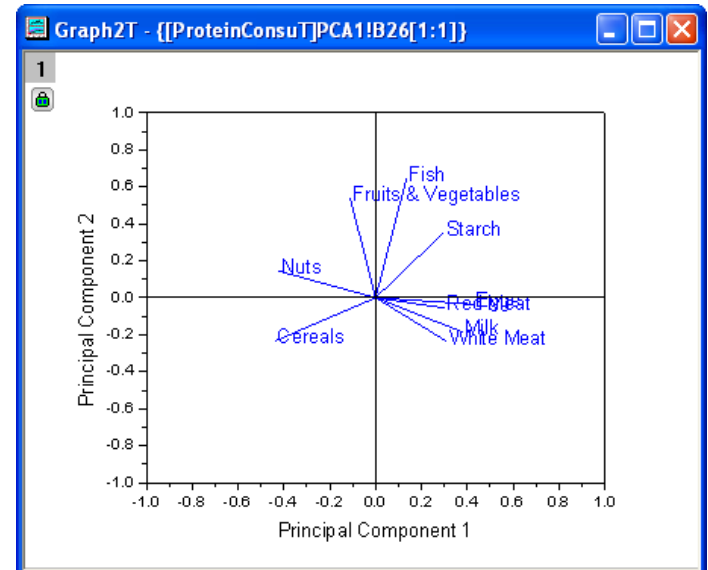
$PC1 = 0.30261 * RedMeat + 0.31056 * WhiteMeat + 0.42668 * Eggs + 0.37773 * Milk + 0.13565 * Fish - 0.43774 * Cereals + 0.29725 * Starch - 0.42033 * Nuts - 0.11042 * FruitsVegetables$

$PC2 = -0.05625 * RedMeat - 0.23685 * WhiteMeat - 0.03534 * Eggs - 0.18459 * Milk + 0.64682 * Fish - 0.23349 * Cereals + 0.35283 * Starch + 0.14331 * Nuts + 0.53619 * FruitsVegetables$

$PC3 = -0.29758 * RedMeat + 0.6239 * WhiteMeat + 0.18153 * Eggs - 0.38566 * Milk - 0.32127 * Fish + 0.09592 * Cereals + 0.24298 * Starch - 0.05439 * Nuts + 0.40756 * FruitsVegetables$

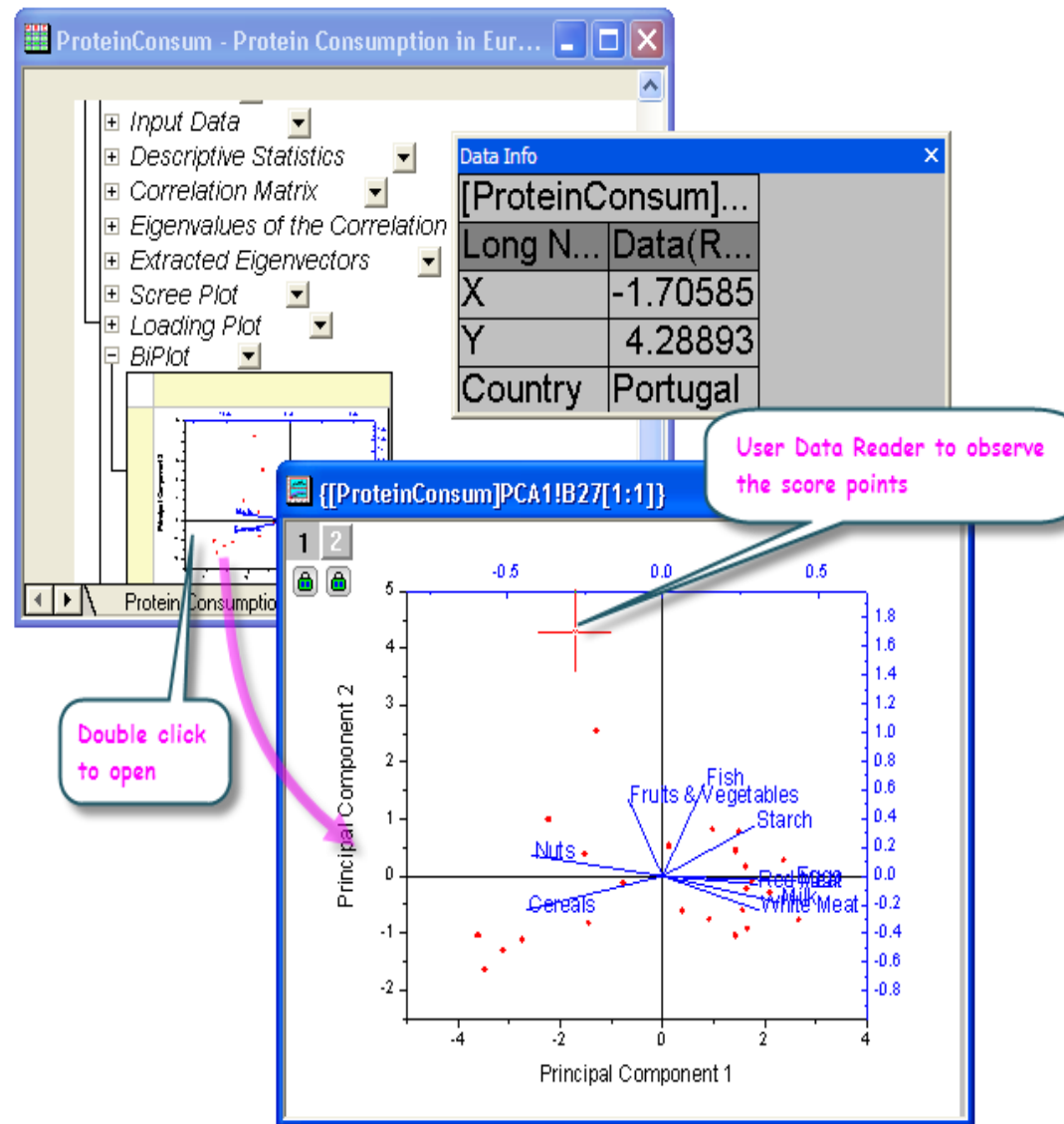
$PC4 = 0.64648 * RedMeat - 0.03699 * WhiteMeat + 0.31316 * Eggs - 0.00332 * Milk - 0.21596 * Fish - 0.0062 * Cereals - 0.33668 * Starch + 0.33029 * Nuts + 0.46206 * FruitsVegetables$

3. **Loading Plot** odhaluje vztahy mezi proměnnými v prostoru prvních dvou hlavních komponent. V grafu komponentních vah lze vidět, že **Red Meat, Eggs, Milk, a White Meat** mají podobné velké zátěže pro **PC1**. **Fish, Fruits&Vegetables** mají podobné velké zátěže pro **PC2**.

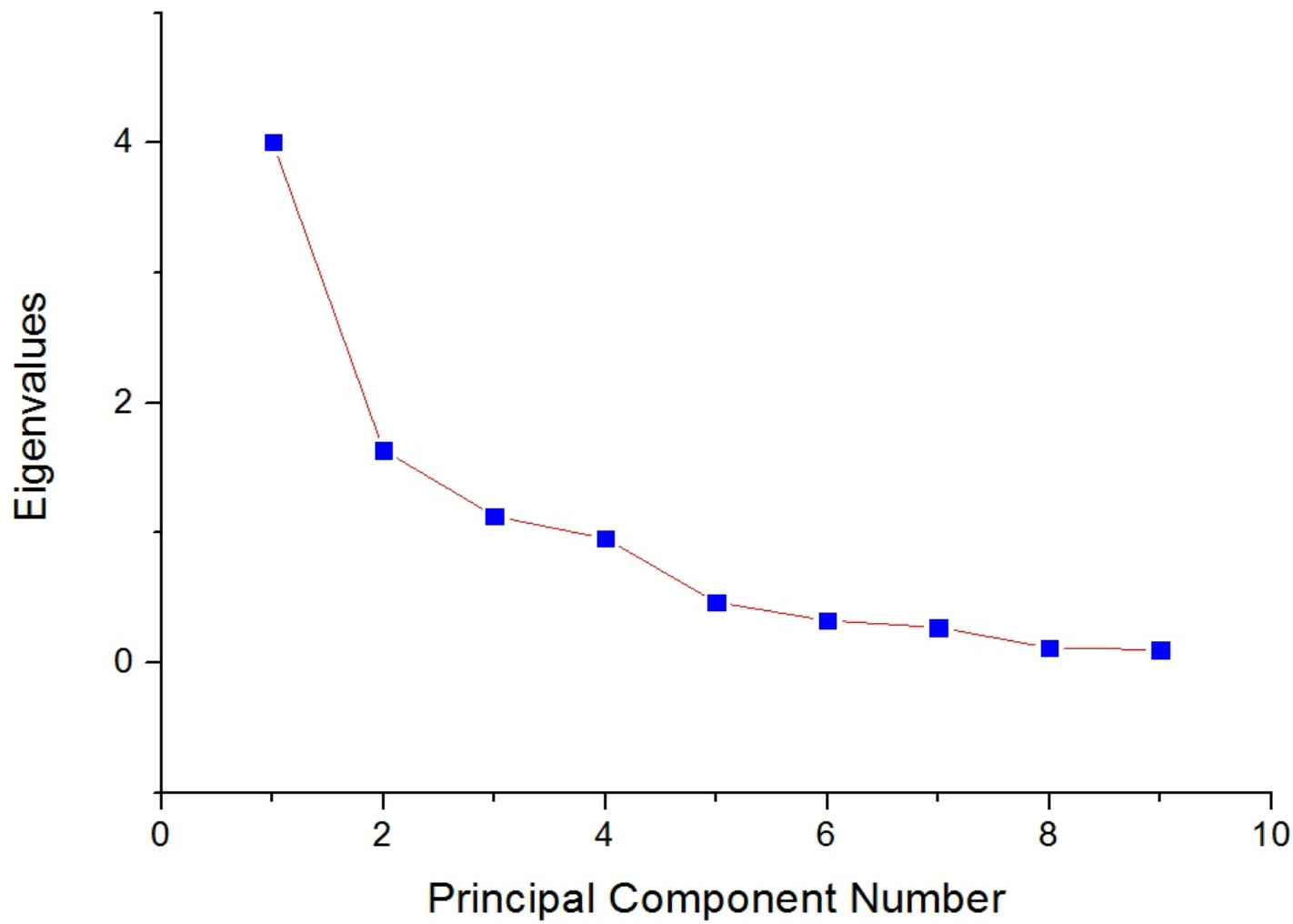


4. **Biplot** ukazuje zatížení a skóre dvou vybraných komponent současně. To odhaluje vztah mezi pozorováními a proměnnými v prostoru prvních dvou PC. (**Poznámka:** Dvoj-klik na graf se tento otevře a upraví.)

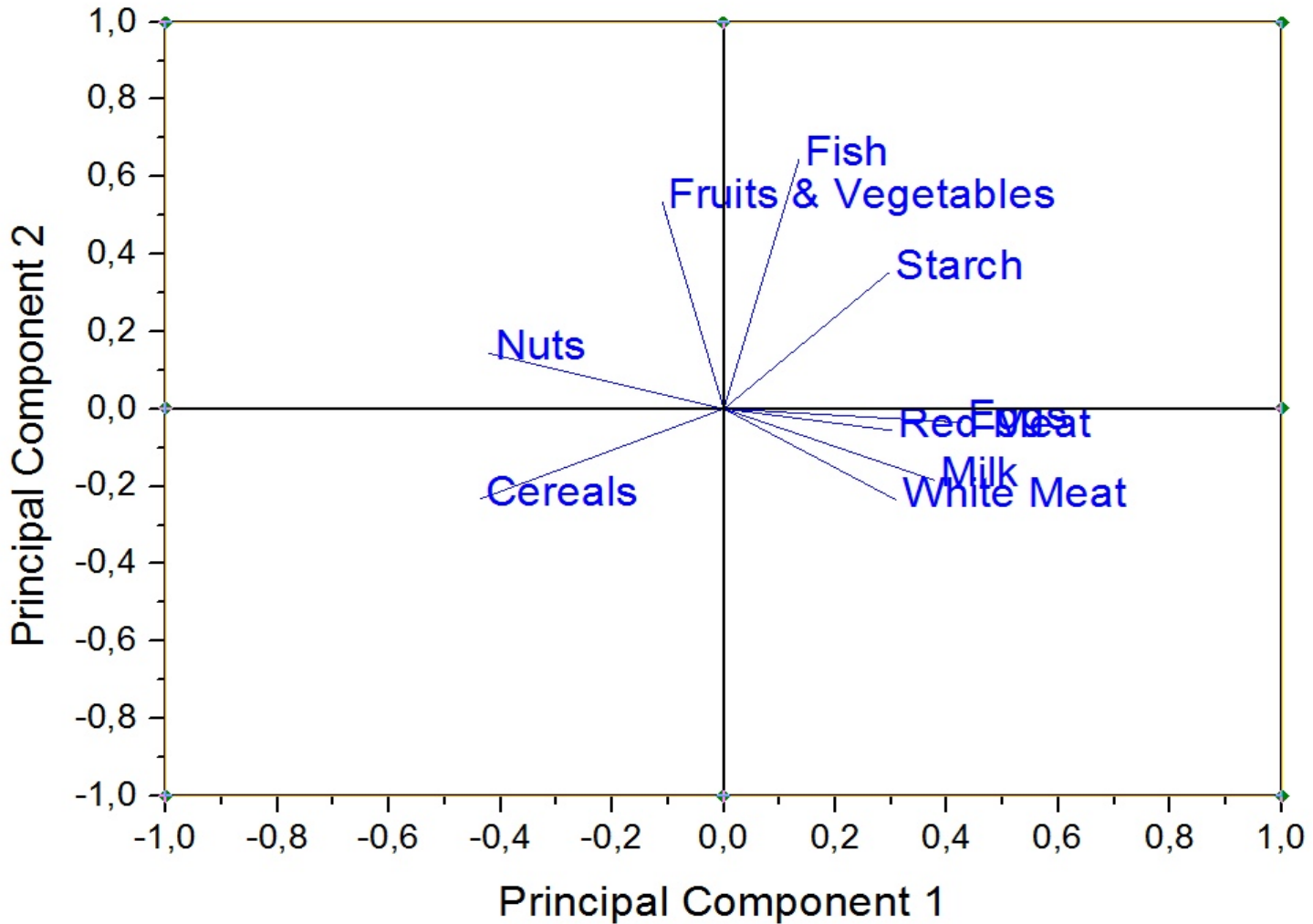
5. Pomocí **Data Reader** se otevře okno **Data Info** a lze zkoumat graf podrobněji. Je vidět, že ve Spain a Portugal se zdroj bílkovin liší od ostatních evropských zemí. Spain a Portugal spoléhají na ovoce a zeleninu, zatímco východoevropské země jako Albania, Bulgaria, Jugoslavia, a Romania raději obiloviny a ořechy.



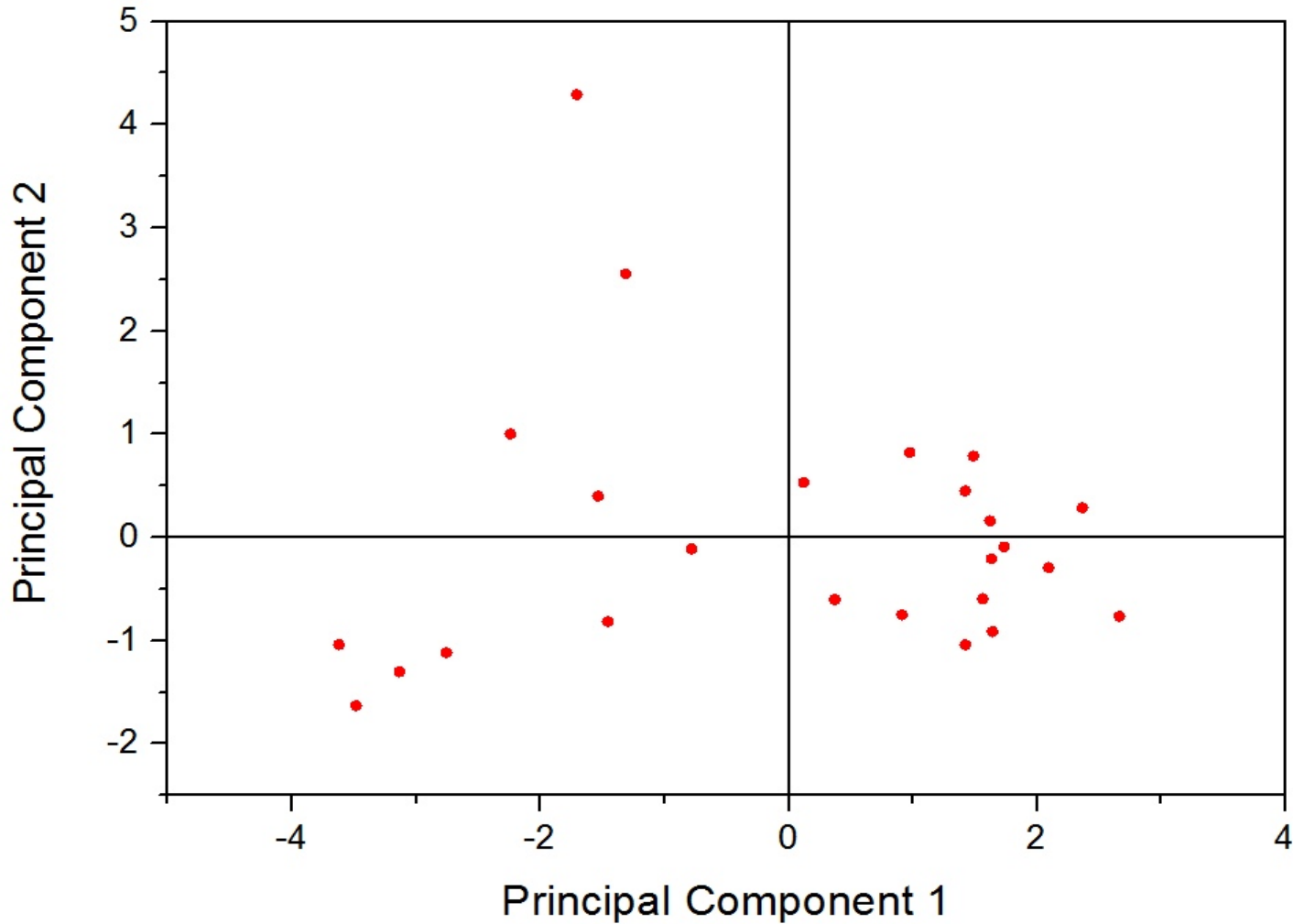
Cattelův indexový graf vlastních čísel (Scree Plot)

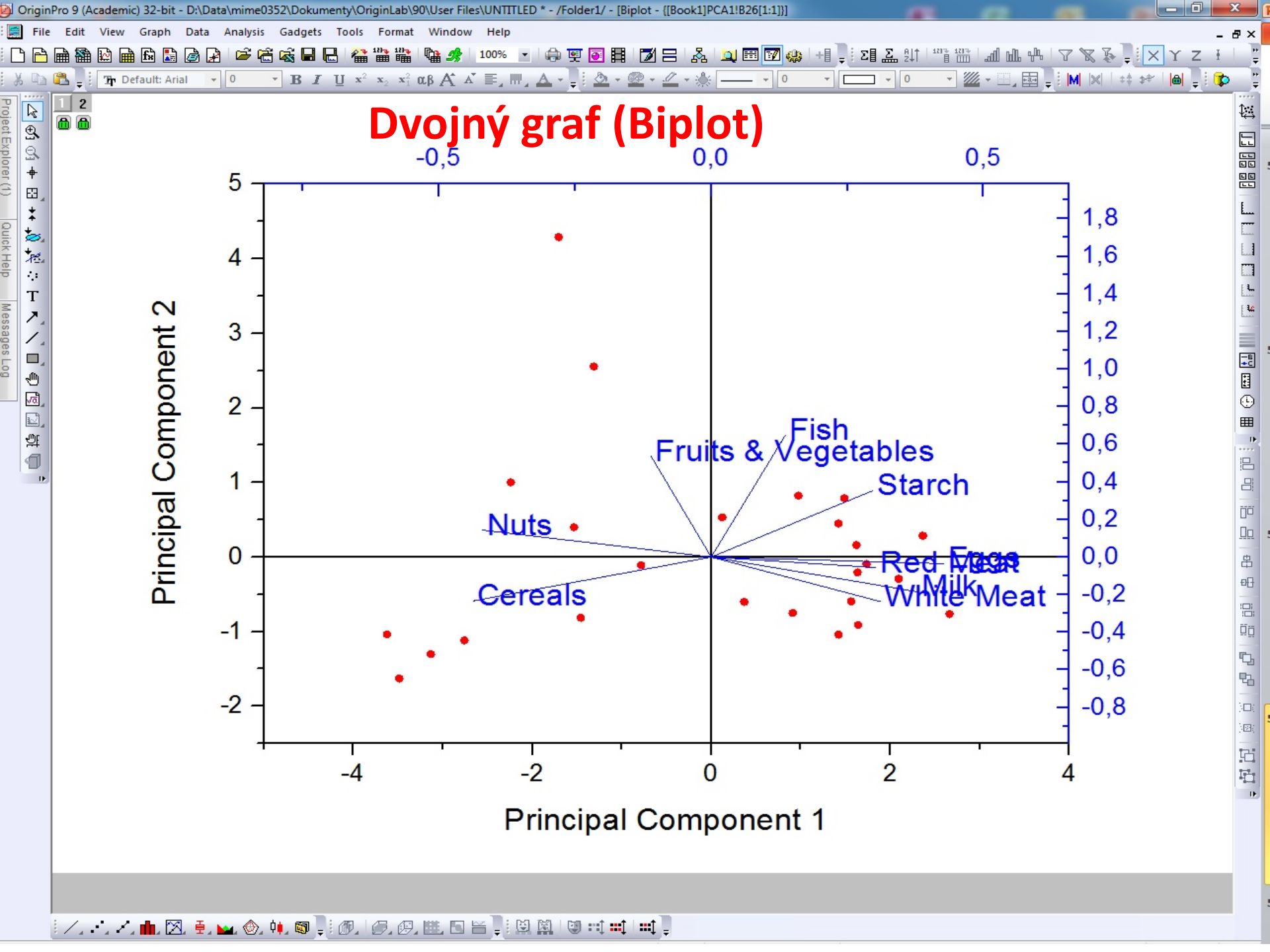


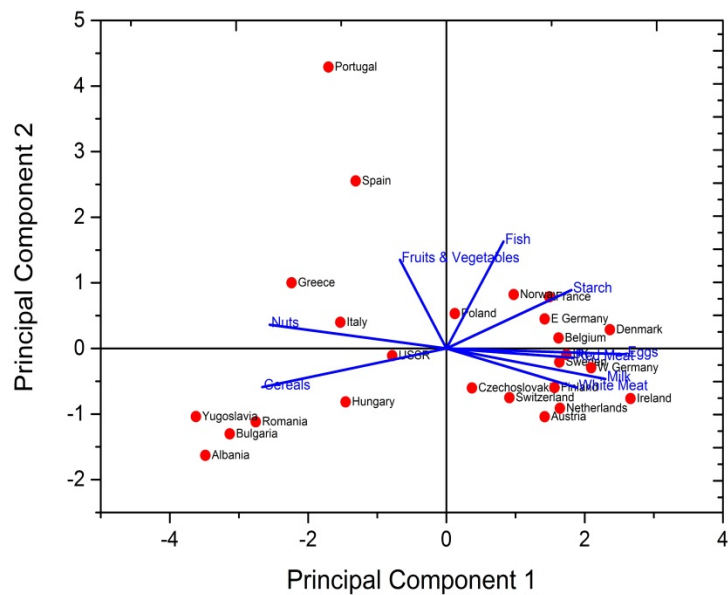
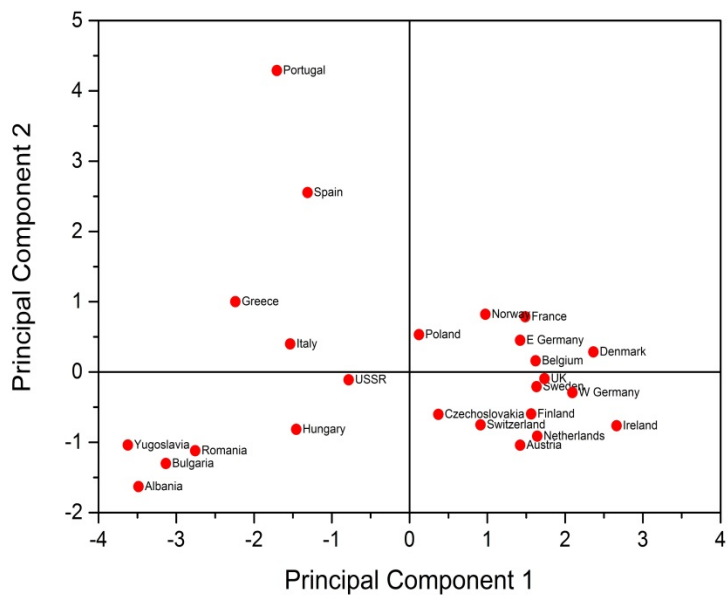
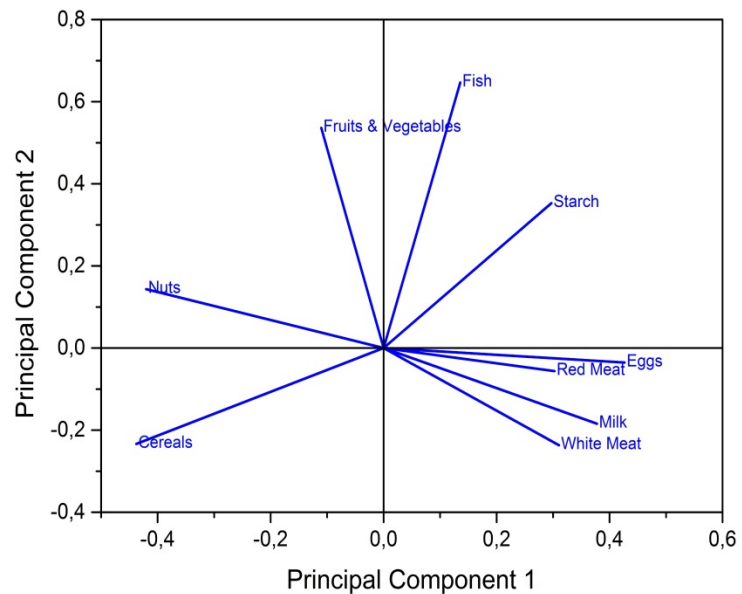
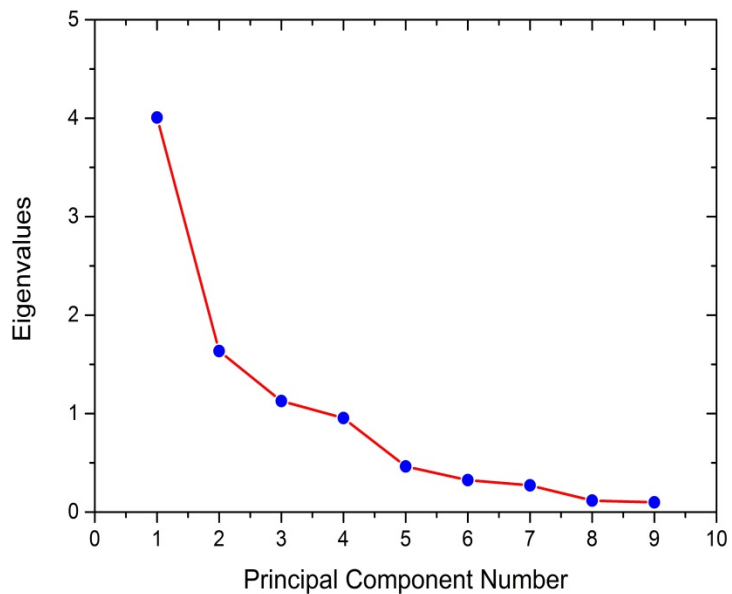
Graf komponentních vah (Plot of Components Weights)



Graf komponentního skóre (Plot of Components Score)







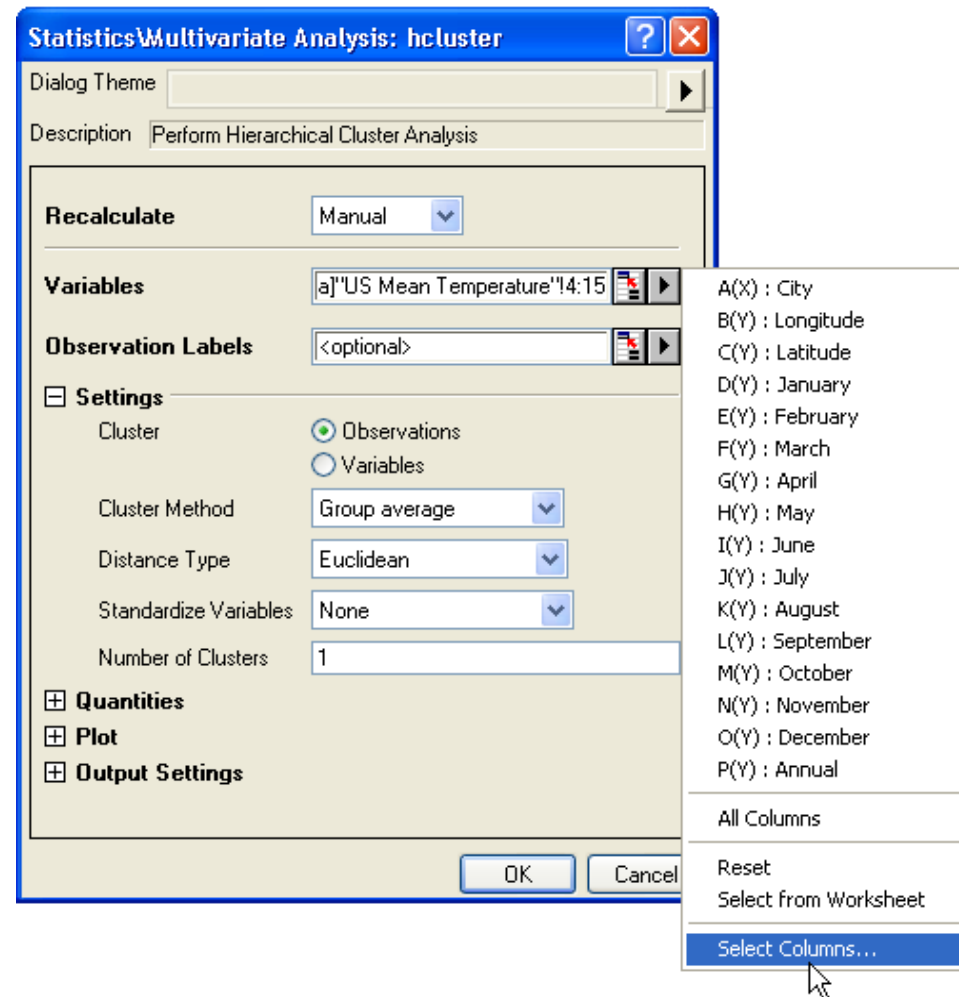
4.6.2 Shluková analýza (Cluster Analysis)

Na příkladu analýzy shluků průměrných teplot v amerických městech více než 3-leté periody se demonstruje analýza shluků. Výchozím bodem je hierarchická shluková analýza s náhodně vybranými daty s cílem nalézt nejlepší metodu pro shlukování. Analýza K-průměrů (K-Means) je rychlý způsob shlukování se provádí pro celé datové soubory.

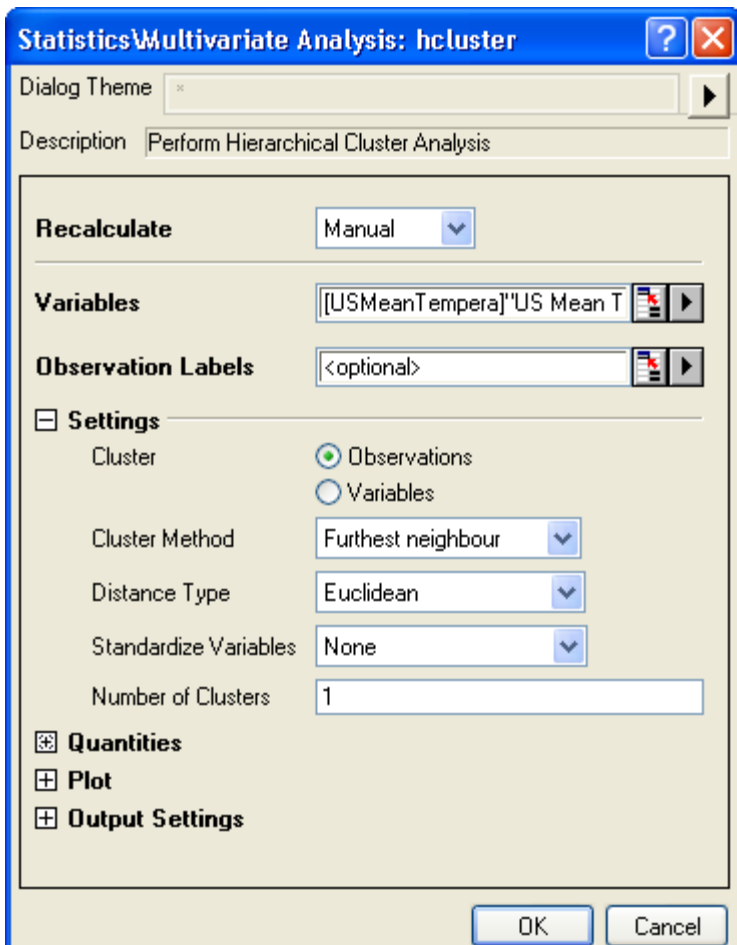
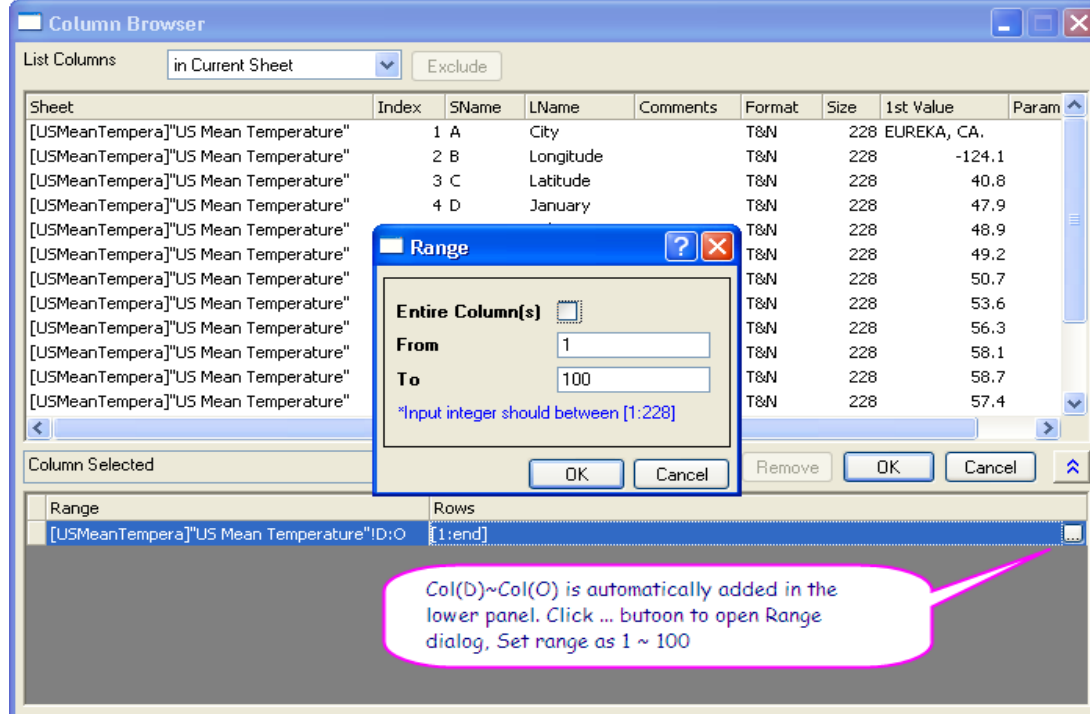
Hierarchická analýza shluků

Kroky:

1. Začněte s novým projektem (sešitem).
Nainportujte data souboru **File, Import, Single ASCII, \Samples\Graphing\US Mean Temperature.dat, Open, OK.**
2. Zvýrazněte sloupce **D** až **O**.
3. Vyberte **Statistics, Multivariate Analysis, Hierarchical Cluster Analysis** a pokračujte....
4. Klepnutím na trojúhelníkové tlačítko vedle proměnných **Variables** a pak klepněte na **Select Columns...**

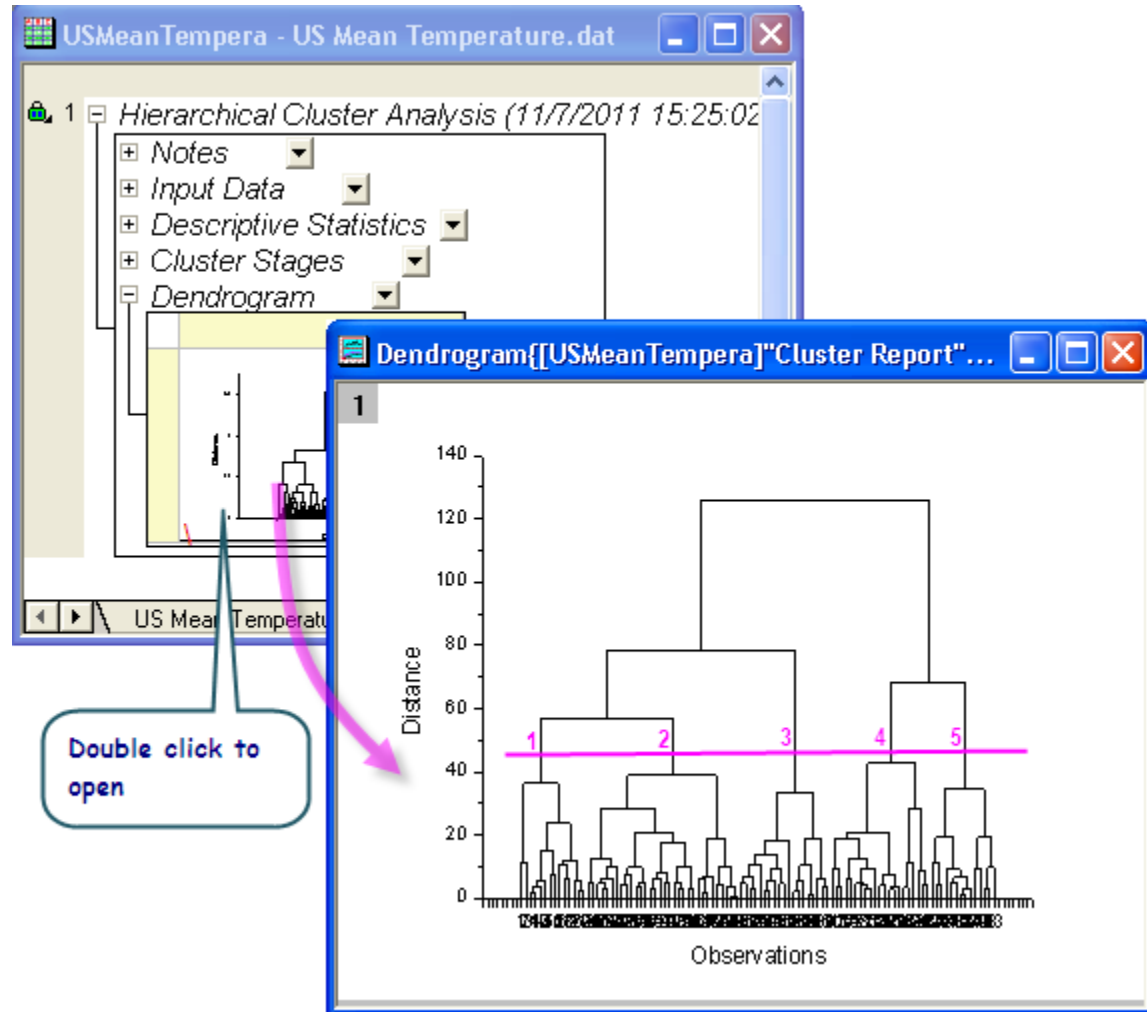


5. V dolním panelu dialogu **Column Browser** klikněte na tlačítko **...** tlačítko. Nastavte rozsah dat od 1 do 100. Klikněte na **OK** a **OK**.



6. V dialogovém okně **Statistics\Multivariate Analysis: hcluster** zaškrtněte v uzlu **Settings** v řádku **Cluster** na **Observations** a v řádku **Number of Clusters** na **1**. V řádku **Cluster Method** vyberte **Furthest Neighbour** dle obrázku vlevo a pak klikněte dole na **OK**.

7. Přejděte na list **Cluster 1**. Na výsledném dendrogramu vidíme data shlukovaná do 5 shluků.

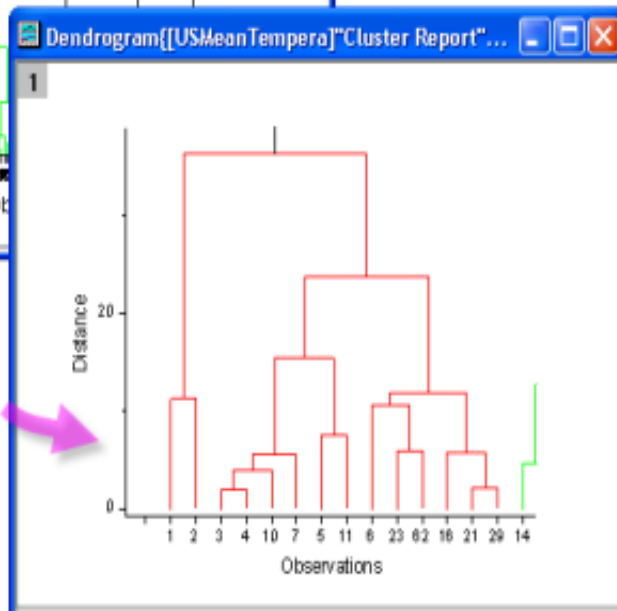
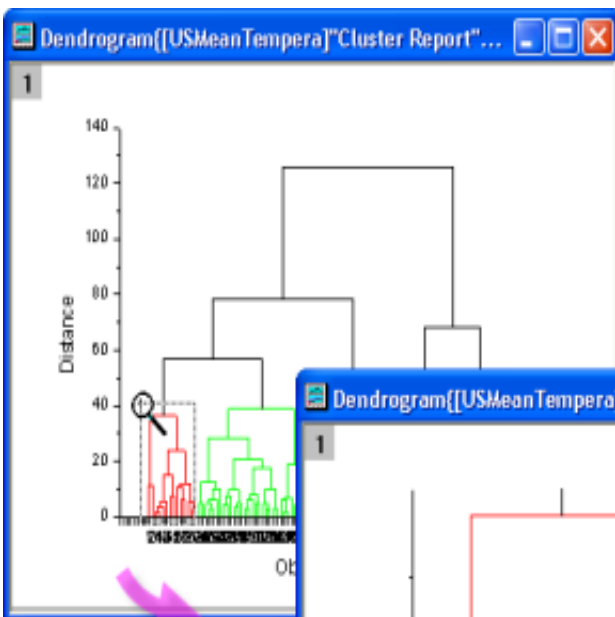
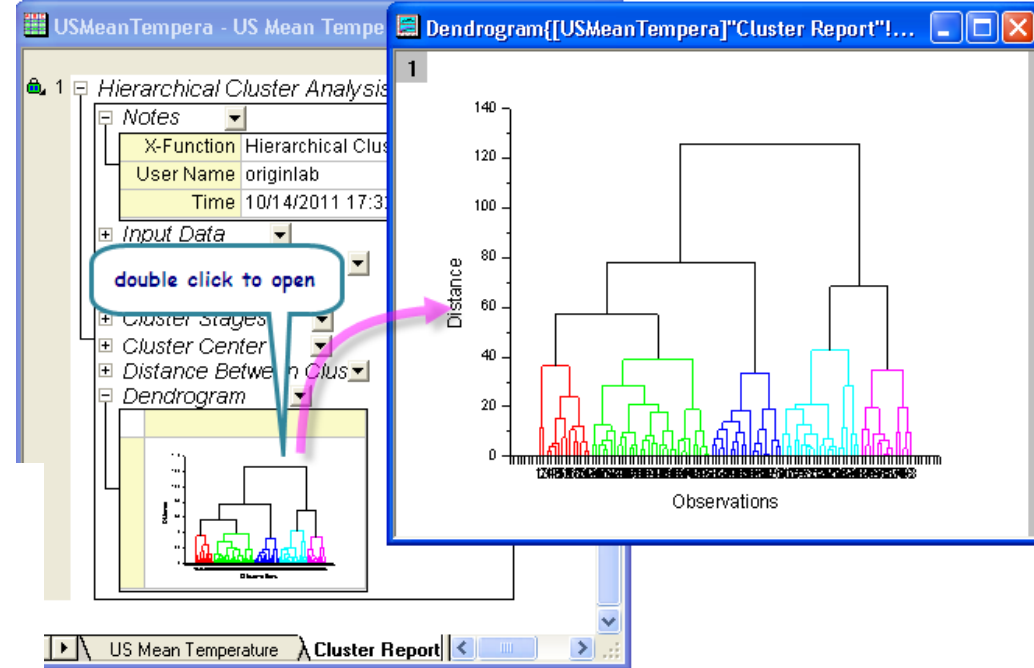


The figure shows the "Statistics/Multivariate Analysis: hcluster" dialog box. The "Recalculate" dropdown is set to "Manual". The "Variables" field contains "January" (1) 0 "December" (100). The "Observation Labels" field contains "<optional>". Under the "Settings" section, "Cluster" is set to "Observations", "Cluster Method" is "Furthest neighbour", "Distance Type" is "Euclidean", and "Standardize Variables" is "None". The "Number of Clusters" is set to 5. Under the "Quantities" section, "Cluster Stages" and "Cluster Center" are checked. The "Plot" and "Output Settings" sections are collapsed. The "OK" and "Cancel" buttons are at the bottom.

8. Klikněte na ikonu zámku v dendrogramu nebo výsledný strom a potom klikněte na **Change Parameters**.

9. Nastavte **Number of Clusters** na 5 a pak zaškrtněte políčko **Cluster Center** v uzlu **Quantities**. Klik na **OK**.

10. Ve výsledném dendrogramu lze jasně vidět, jak se pozorování seskupila do shluků. (Poznámka: dvoj-klikem lze otevřít a upravit dendrogram.)



11. Vzhledem k velkému počtu pozorování, se popisky osy překrývají v tomto dendrogramu. Použijte proto **Scale In** (lupu) v Tools-nástroji a vyberte si oblast, kterou chcete zvětšit.

Analýza metodou shlukování K-průměrů

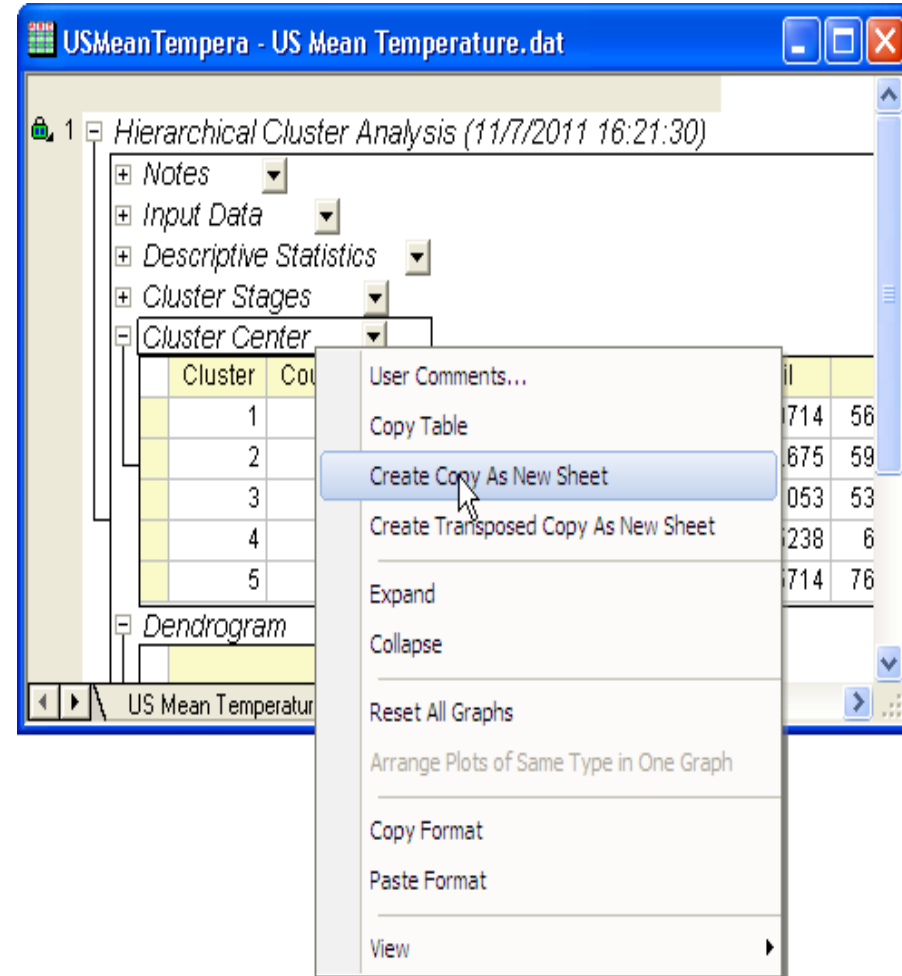
1. Klikněte pravou myší na **Cluster Center** a vyberte v roletce **Create Copy as New Sheet**. Budete používat nově vytvořený **Sheet2** jako **Initial Cluster Center**.

2. Vraťte se na list se zdrojovými daty (**US Mean Temperature**) a označte **col(D)** až **col(O)**. Vyberte **Statistics, Multivariate Analysis, K-Means Cluster Analysis, Open dialog** a pokračujte v okně **kmeans**.

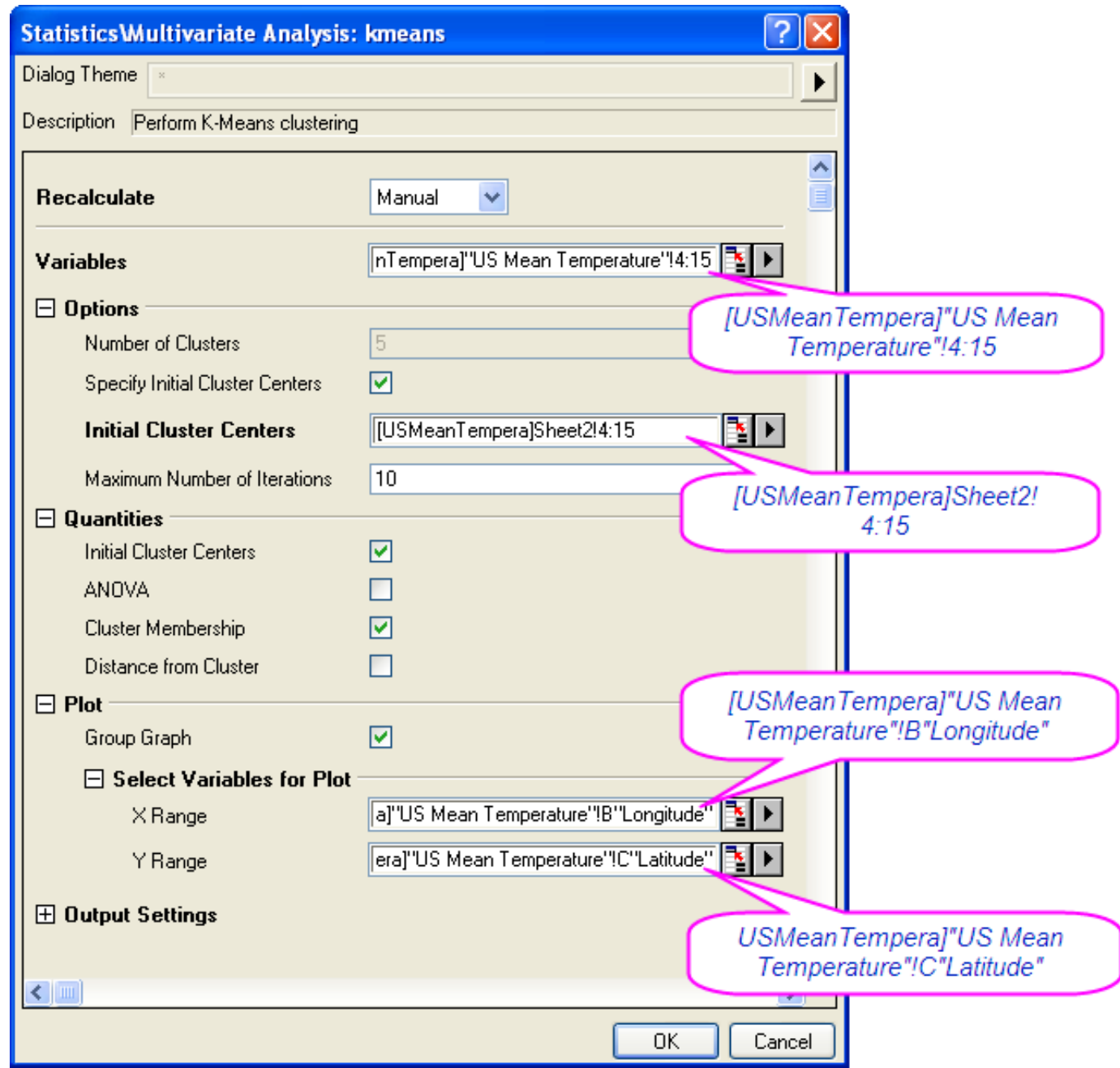
3. V uzlu **Options** zaškrtněte políčko **Specify Initial Cluster Centers**. Klikněte na interaktivní červeno-černé tlačítko vedle **Initial Cluster Center**. Dialogové okno naroluje.

4. Přejděte na záložku listu **Sheet2** přejděte na řádky s **Sheet2** a označte řádky od **Col(D)** až **Col(O)**. Klikněte na interaktivní červeno-černé tlačítko vedle **Initial Cluster Center** k obnovení dialogu.

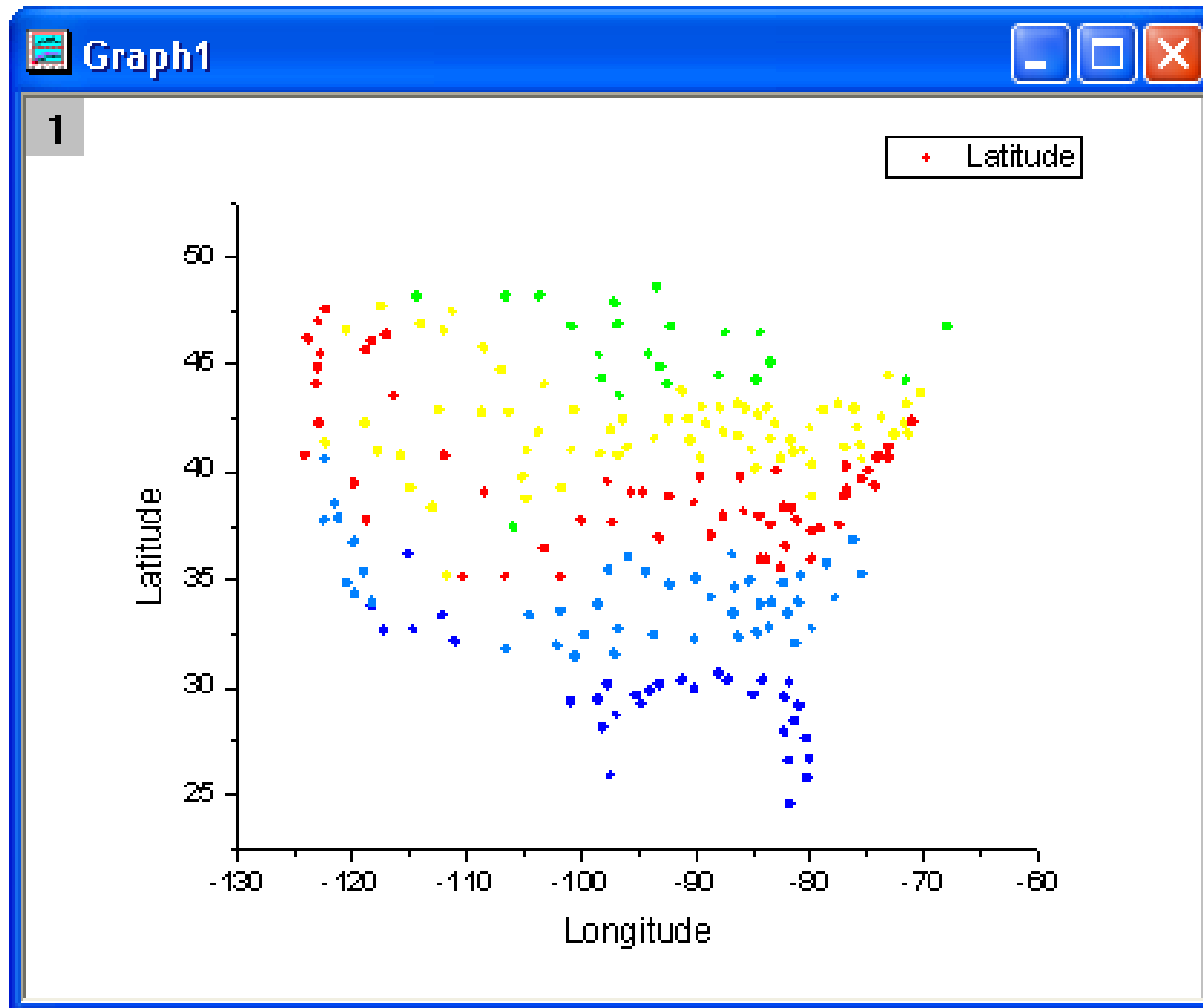
5. Otevřete uzel **Plot** a **Group Graph**. Klikněte na interaktivní tlačítko vedle **X Range**. Dialogové okno se naroluje. Vraťte se zpět do zdrojového listu **US Mean Temperature** a zvýrazněte **Col(B):longitude**. Klikněte na tlačítko v dialogovém okně roll up až do obnovení.



6. Kliknutím na trojúhelníkové tlačítko vedle **Y Range**, a pak vyberte **C(Y)**, **Latitude**. Klikněte na **OK**.



7. Aktivujte list **K-Means1**. Všimněte si, že data byla seskupena do 5 skupin podle zeměpisných šířkách měst.



4.6.3 Diskriminační analýza (Discriminant Analysis)

Sada 150 Fisherových kosatců Iris dataset představuje vícerozměrný výběr dat Sira Ronalda Aylmera Fishera z roku 1936. Tento výběr dat je často používán k ilustračním účelům v mnoha klasifikačních algoritmech. Skládá se z 50 kyticek z každého ze tří druhů kosatců (Iris setosa, Iris virginica a Iris versicolor). Čtyři naměřené hodnoty, a to délka a šířka kališního lístku a délka a šířka okvětního lístku v centimetrech tvoří zdrojovou matici dat pro každý vzorek kosatce. Lze použít diskriminační analýzu k identifikaci botanického druhu kosatce na základě těchto čtyř měr.

Cíl úlohy: Užije se náhodný vzorek 120 řádků dat k vytvoření modelu diskriminační analýzy (**trénovací** čili **analyzovaná data**), a poté zbývajících 30 řádků za účelem ověření přesnosti modelu (**testovací data**).

Kosatce (Iris)



Iris Setosa

Wild Iris - *Iris setosa*



Iris Versicolor

© 2005 Tajet Novak



Iris Virginica

Kosatce (Iris)



Iris Setosa



Iris Versicolor

© 2005 Tajet Novak



Iris Versicolor

© 2001 Elzevir P. Moly



Iris Virginica



Iris Virginica

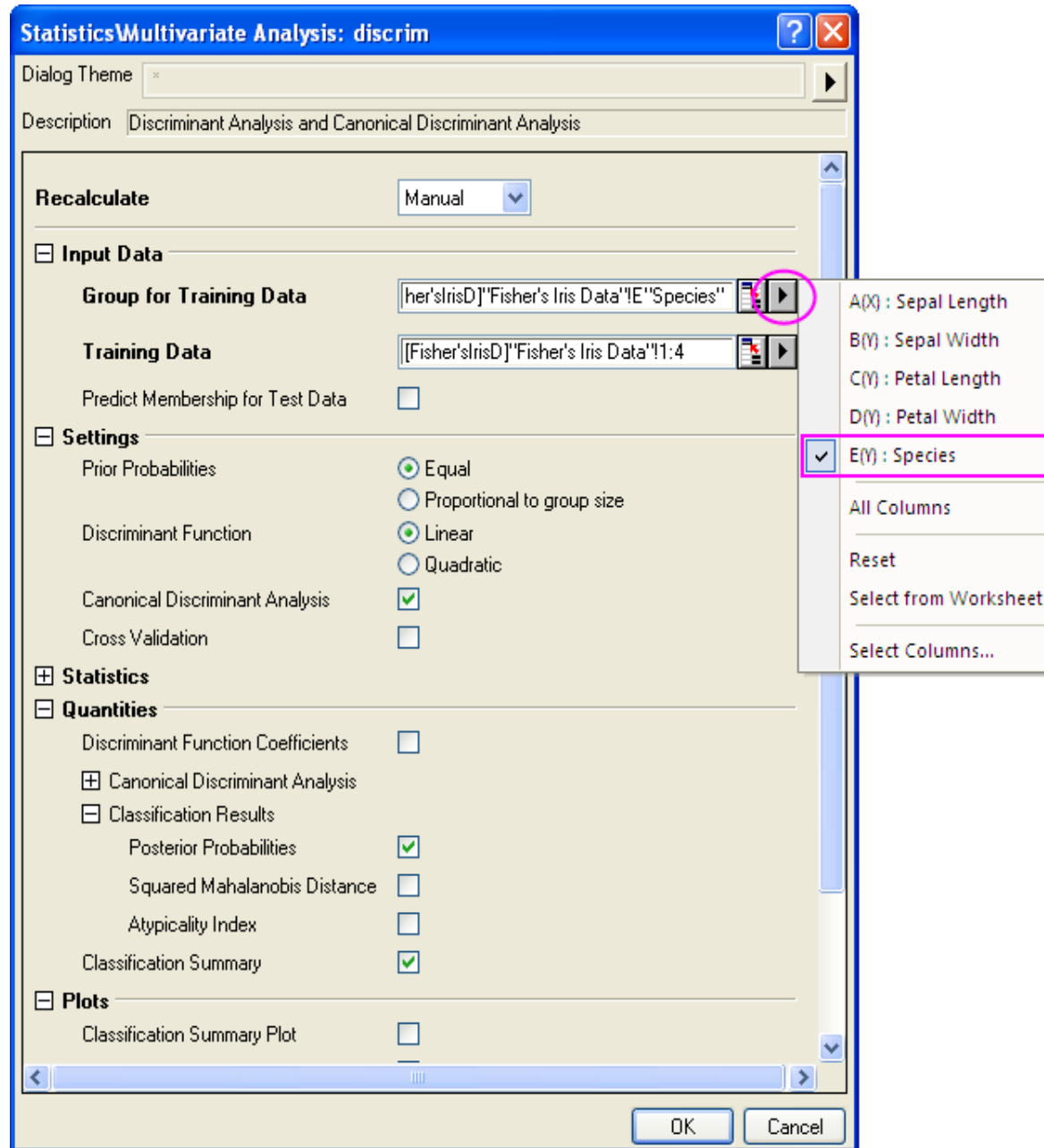


Iris Virginica

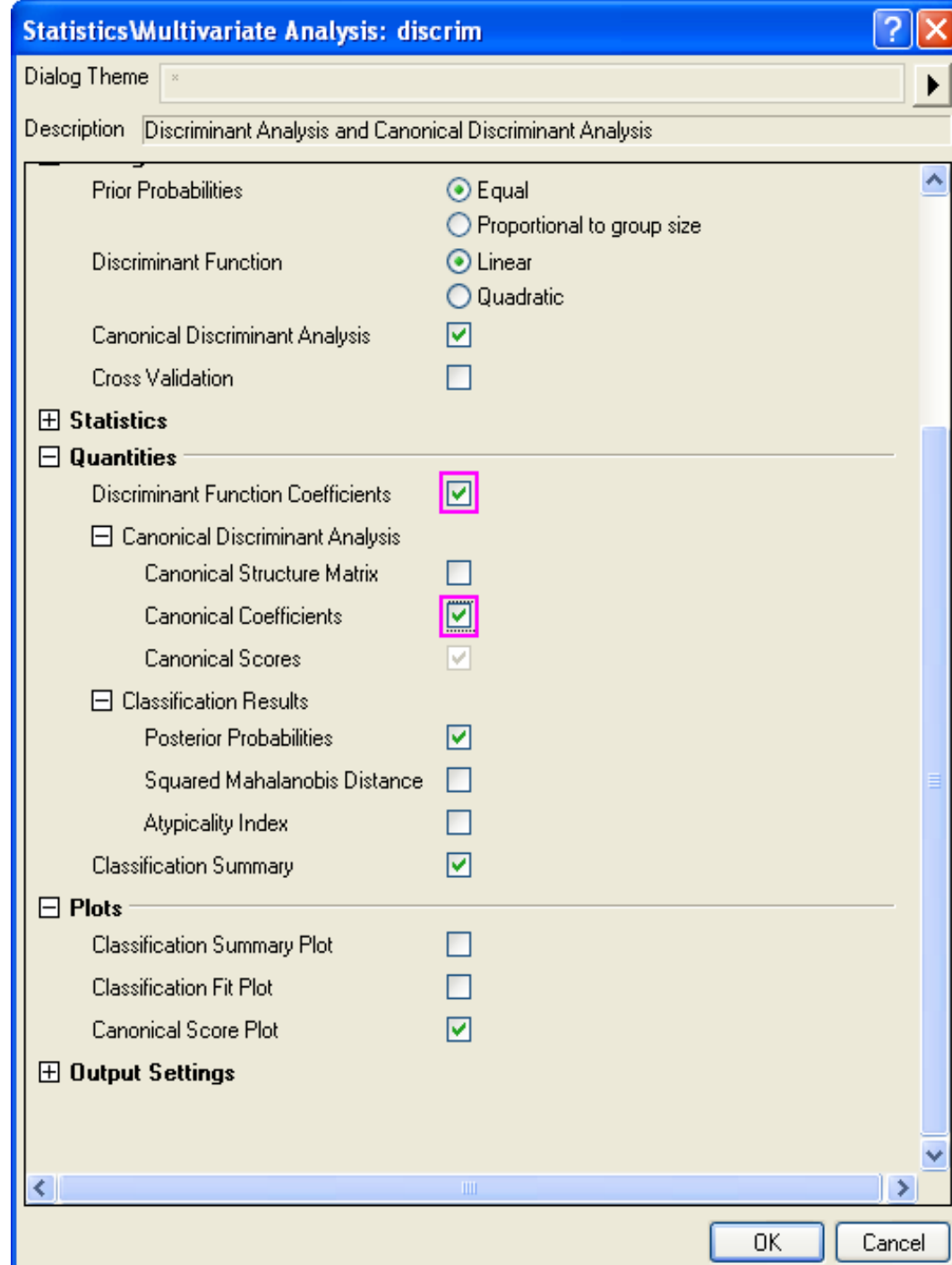
A. Načtení analyzovaných (trénovacích) dat:

1. Otevřete nový projekt a nainportujte data **File, Import, Import Wizard**, kliknutím na ... pak soubor `\Samples\Statistics\Fisher's Iris Data.dat`, **Add File(s), OK, Finish**.

2. Zvýrazněte sloupce **A až D** a zvolte **Statistics, Multivariate Analysis, Discriminant Analysis, Open Dialog** otevře dialog diskriminační analýzy. Sloupce **A až D** jsou automaticky přidány jako **Training Data (trénovací nebo analyzovaná data)**. Klikněte na **triangle button** vedle **Group for Training Data** a vyberte **E(Y): Species** v **otevřené roletce**.



3. Otevřete dále uzel **Quantities**, a pak zaškrtněte políčko **Discriminant Function Coefficients**. Zaškrtněte políčko **Canonical coefficients** v oddíle **Canonical Discriminant Analysis**. Přijměte všechna ostatní defaultní nastavení a klikněte na **OK**.



Interpretace výsledků trénovacích dat:

1. Přejděte dole na list **Discrim1**. Oddíl výstupu zvaný **Canonical Discriminant Analysis** přináší odhady parametrů Fisherovy lineární diskriminační funkce pro data (**trénovací** neboli **analyzovaná data**).
2. Použitím **orámovaných hodnot v tabulce Unstandardized Canonical Coefficient** lze postavit kanonické diskriminační funkce **D1** a **D2**.

Unstandardized Canonical Coefficients

	Canonical Variable 1	Canonical Variable 2
Constant	-2.10511	-6.66147
Sepal Length	-0.82938	0.0241
Sepal Width	-1.53447	2.16452
Petal Length	2.20121	-0.93192
Petal Width	2.81046	2.83919

$$D1 = -2.10511 - 0.82938 * SL - 1.53447 * SW + 2.20121 * PL + 2.81046 * PW$$

$$D2 = -6.66147 + 0.0241 * SL + 2.16452 * SW - 0.93192 * PL + 2.83919 * PW$$

kde SL = Sepal Length, SW = Sepal Width, PL = Petal Length, PW = Petal Width

2. Kliknutím na uzel **Eigenvalues** výstupové tabulky se odkryjí vlastní čísla, která odhalí důležité kanonické diskriminační funkce. První funkce vysvětluje **99,12%** rozptylu a druhá vysvětluje zbývajících **0,88%**.

Eigenvalues

	Eigenvalue	Percentage of Variance	Cumulative	Canonical Correlation
1	32.19193	99.12%	99.12%	0.98482
2	0.28539	0.88%	100.00%	0.4712

3. Kliknutím na uzel **Wilk's Lambda Test** výstupové tabulky se otevrou hodnoty testu **Wilk's Lambda**, která ukazují, že obě diskriminační funkce výrazně vysvětlují účast v diskriminační třídě, protože obě hodnoty ve sloupci **Sig** jsou menší než **0.05**. Obě hodnoty by proto měly být zahrnuty do výsledků diskriminační analýzy.

Wilks' Lambda Test

	Wilks' Lambda	Chi-square	df	Sig.
1 to 2	0.02344	546.1153	8	8.87078E-113
2 to 2	0.77797	36.52966	3	5.78605E-8

At the 0.05 level, the dimensionality is significantly 2.

Klasifikace neznámých kosatců:

1. Aby bylo možné klasifikovat neznámé kosatce, vyčíslí se skóre čili souřadnice každého kosatce z odhadů parametrů Fisherovy lineární diskriminační funkce (**Coefficients of Linear Discriminant Function**), a poté každého kosatce zařadí do své třídy (Setosa, Versicolor, Virginica).

Coefficients of Linear Discriminant Function

	setosa	versicolor	virginica
Constant	-86.30847	-72.85261	-104.36832
Sepal Length	23.54417	15.69821	12.44585
Sepal Width	23.58787	7.07251	3.68528
Petal Length	-16.43064	5.21145	12.76654
Petal Width	-17.39841	6.43423	21.07911

2. Přepněte na list **Training Results**. Na příkladu sedmého kosatce je ukázáno, jak lze vypočítat souřadnicové skóre v každé ze tří tříd pomocí odhadů parametrů **Coefficient of Linear Discriminant Function** (výše).

Long Name	A(Y)	B(Y)	C(Y)	D(Y)	E(Y)	F(Y)
Units						
Comments	Source Data					
UserParam1						
1	5.1	3.5	1.4	0.2		
2	4.9	3	1.4	0.1		
3	4.7	3.2	1.3	0.1		
4	4.6	3.1	1.5	0.1		
5	5	3.6	1.4	0.2	setosa	setosa
6	5.4	3.9	1.7	0.4	setosa	setosa
7	4.6	3.4	1.4	0.3	setosa	setosa
8	5	3.4	1.5	0.2	setosa	setosa
9	4.4	2.9	1.4	0.2	setosa	setosa
10	4.9	3.1	1.5	0.1	setosa	setosa
11	5.4	3.7	1.5	0.2	setosa	setosa
12	4.8	3.4	1.6	0.2	setosa	setosa
13	4.8	3	1.4	0.1	setosa	setosa
14	4.3	3	1.1	0.1	setosa	setosa
15	5.8	4	1.2	0.2	setosa	setosa
16	5.7	4.4	1.5	0.4	setosa	setosa

$$\text{Score}(\text{setosa}) = - 86.30847 + 23.54417 * 4.6 + 23.58787 * 3.4 - 16.43064 * 1.4 - 17.39841 * 0.3 = \mathbf{73.971051}$$

$$\text{Score}(\text{versicolor}) = - 72.85261 + 15.69821 * 4.6 + 7.07251 * 3.4 + 5.21145 * 1.4 + 6.43423 * 0.3 = \mathbf{32.631989}$$

$$\text{Score}(\text{virginica}) = - 104.36832 + 12.44585 * 4.6 + 3.68528 * 3.4 + 12.76654 * 1.4 + 21.07911 * 0.3 = - \mathbf{10.390569}$$

3. Z ukázkových výpočtů pro sedmý kosatec je vidět, že skóre setosy **Score(setosa) = 73,971051** dosahuje největší hodnoty ze tří tříd (**73.971051 setosa**, **32.631989 versicolor**, **10.390569 virginica**) čili sedmý kosatec by měl být zařazen do skupiny **setosa**.

Kosatce (Iris)



Iris Setosa

Wild Iris - *Iris setosa*



Iris Versicolor

© 2005 Tajet Novak



Iris Virginica

Kosatce (Iris)



Iris Setosa



Iris Versicolor

© 2005 Tajet Novak



Iris Versicolor

© 2001 Elzevir P. Moly



Iris Virginica



Iris Virginica



Iris Virginica

4. Classification Summary for Training Data

v listu **Discrim1** ukazuje, že zařazení neznámých kosatců do skupiny **setosa** je 100%ně správné. Pro **versicolor** jsou pouze **2** kosatce chybně zařazené jako **virginica**. Pro **virginica** je pouze **1** kosatec chybně zařazen. Chybovost je pouze **2.0%**. Nalezený model je proto dobrý.

Classification Summary for Training Data

Classification Count

	Predicted Group			
	setosa	versicolor	virginica	Total
setosa	50	0	0	50
	100.00%	0.00%	0.00%	100.00%
versicolor	0	48	2	50
	0.00%	96.00%	4.00%	100.00%
virginica	0	1	49	50
	0.00%	2.00%	98.00%	100.00%
Total	50	49	51	150
	33.33%	32.67%	34.00%	100.00%

Error Rate

	setosa	versicolor	virginica	Total
Prior	0.33333	0.33333	0.33333	
Rate	0.00%	4.00%	2.00%	2.00%

Error rate for classification of training data is 2.00%.

Validate modelu:

Classification Summary of Training Data vyhodnocuje kosatce via Fisherovy diskriminační funkce, sestavenou z týchž dat. „Chybovost“ však bývá větší, když se klasifikují neznámá data, která nebyla užita k sestavení odhadu diskriminační funkce. Existují dva způsoby, jak to napravit:

- **Cross-validate:**

V křížové validaci je každý tréninkový údaj o kosatci považován za testovací data, zda má být vyloučen z tréninkových dat nebo posouzen, do které skupiny by měl být zařazen a tak se ověří, zda provedená klasifikace je správná nebo ne.

- **Podskupina validace:**

Obvykle se náhodně rozdělí množina kosatců do dvou podskupin, z nichž první se použije pro odhad diskriminačního modelu (**trénovací výběr**) a druhý je k testování spolehlivosti výsledků (**testovací výběr**).

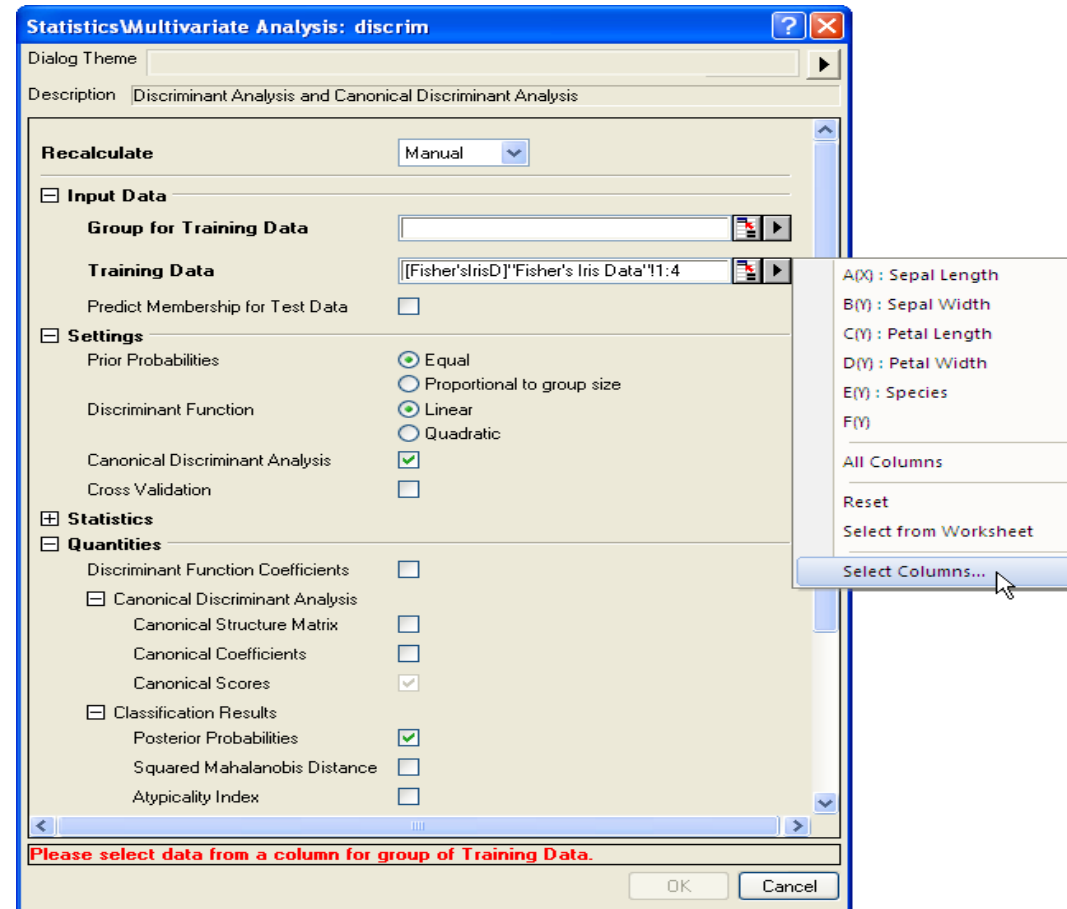
Příprava dat pro analýzu

Data lze třídit v náhodném pořadí. Použije se prvních 120 řádků kosatců jako **trénovací data** a posledních 30 kosatců jako **testovací data**.

1. Vraťte se zpět na záložku listu **Fisher's Iris Data**.
2. Přidejte nový sloupec **Column, Add New Column** a vyplňte ho normálně generovanými náhodnými čísly postupem **Column, Fill Column with, Normal Random Numbers**.
3. Označte nově přidaný a naplněný sloupec. Klikněte na něj pravou myší a vyberte **Sort Worksheet, Ascending** a hodnoty jsou seřazeny dle velikosti od záporných do kladných..

Průběh diskriminační analýzy

1. Vyberte a označte sloupce **A až D**.
2. Zvolte **Statistics, Multivariate Analysis, Discriminant Analysis, Open Dialog**.
3. Nastavte prvních 120 řádků sloupců **A až D** za trénovací data postupem: klikněte na trojúhelníkové tlačítko vedle **Training Data** a zvolte **Select Columns** v otevřené roletce a pokračujte v okně **Column Browser**.



4. V dialog **Column Browser** klikněte na tlačítko **...** umístěné vpravo dole na dolním panelu. Vypněte zaškrtnutí v řádku **Entire Column(s)** a zadejte **From** na **1** a **To** na **120**. Klikněte na **OK** a **OK**.

The image shows a screenshot of the **Column Browser** dialog box in a software application. The dialog box is titled "Column Browser" and has a "List Columns" section with a dropdown menu set to "in Current Sheet" and an "Exclude" button. Below this is a table of columns with the following data:

Sheet	Index	SName	LName	Comments	Format	Size	1st Value	Param
[Fisher'sIrisD]"Fisher's Iris Data"	1	A	Sepal Length		T&N	150	5.7	
[Fisher'sIrisD]"Fisher's Iris Data"	2	B	Sepal Width		T&N	150	4.4	
[Fisher'sIrisD]"Fisher's Iris Data"	3	C	Petal Length		T&N	150	1.5	
[Fisher'sIrisD]"Fisher's Iris Data"	4	D	Petal Width		T&N	150	0.4	
[Fisher'sIrisD]"Fisher's Iris Data"	5	E	Petal Width		T&N	150	setosa	
[Fisher'sIrisD]"Fisher's Iris Data"	6	F	Petal Width		T&N	150	-2.75751	

Overlaid on the bottom part of the Column Browser dialog is a smaller **Range** dialog box. It has the following fields and controls:

- Entire Column(s)**:
- From**:
- To**:
- Footnote: **Input integer should between [1:150]*
- Buttons: **OK** and **Cancel**

A callout bubble with a pink border and arrow points to the **...** button in the bottom right corner of the Column Browser dialog. The text inside the bubble reads: "Click the ... button next the data range to open Range dialog, set range as 1~120 and click OK."

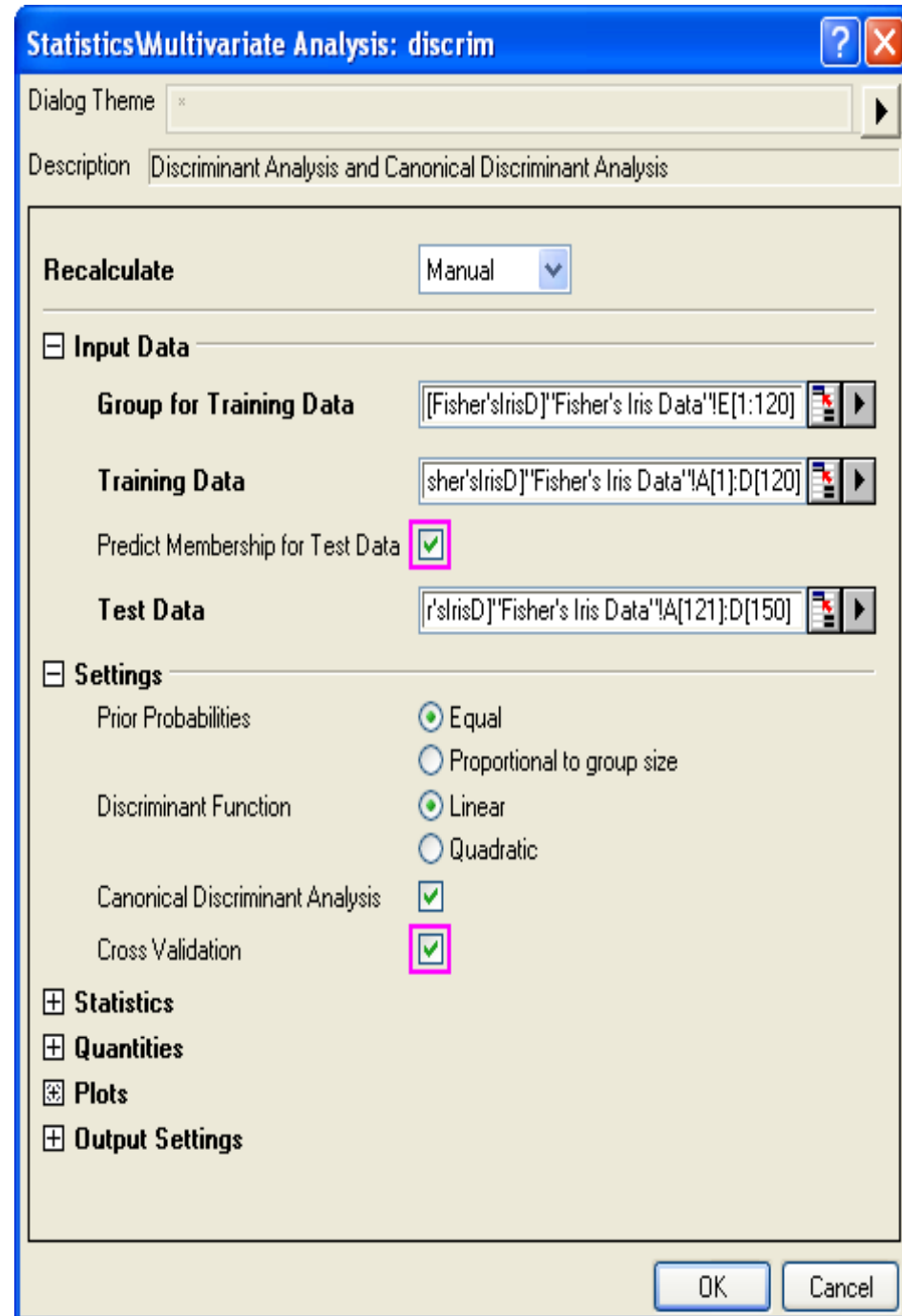
At the bottom of the Column Browser dialog, there is a section titled "Column Selected" with a table showing the selected range:

Range	Rows
[Fisher'sIrisD]"Fisher's Iris Data"!A:D	[1:end]

5. Chcete-li nastavit prvních 120 řádků **Col(E)** pro **Group for Training Data**, klikněte na tlačítko trojúhelníku vedle **Group for Training Data** a vyberte v roletce **E(Y): Species**. Poté klikněte na toto tlačítko trojúhelníku znovu, zvolte **Select Columns** a nastavte rozsah 1 až 120 ve sloupcovém prohlížeči. Klikněte na **OK** a **OK**.

6. V bloku **Input Data** zaškrtněte políčko **Predict Membership of Test Data**. Klikněte na interaktivní mramorované tlačítko **Test Data**. Dialog se zbalí. Vyberte sloupce **A** až **D** v listu. Klikněte na tlačítko roletky až do obnovení dialogu. Poté klikněte na tlačítko trojúhelníku otevřít **Column Browser** a vyberte **Select Columns**. Klikněte na **...** tlačítko v dolním panelu, a nastavit v rozmezí od **121** do **150**. Klikněte na **OK** a **OK**.

7. Otevřete uzel **Settings**, a pak zaškrtněte políčko **Cross Validation** vyberte. Klikněte dole na **OK**.



Cross-validation:

Přejděte na list **Discrim2**. Tabulka **Cross-validation Summary for Training Data** poskytuje predikovanou chybu klasifikováním každého kosatce a zároveň jej vyloučí z dalšího modelového výpočtu. Přesto je tato metoda stále optimističtější než validace podskupiny.

Validace podskupiny:

1. Classification Summary for Test Data

poskytuje informace, jak jsou testovací data jsou klasifikována.

2. Na listu **Fisher's Iris Data** okopírujte posledních 30 řádků (121 až 150) jenom ze sloupce **Col(E) Species**.

3. Na listu **Test Result** přidejte jeden sloupec **Col(I)**. Vložte zkopírované hodnoty do nového sloupce.

Cross-validation Summary for Training Data

Classification Count

	Predicted Group			
	setosa	virginica	versicolor	Total
setosa	44	0	0	44
	100.00%	0.00%	0.00%	100.00%
virginica	0	36	2	38
	0.00%	94.74%	5.26%	100.00%
versicolor	0	2	36	38
	0.00%	5.26%	94.74%	100.00%
Total	44	38	38	120
	36.67%	31.67%	31.67%	100.00%

Error Rate

	setosa	virginica	versicolor	Total
Prior	0.33333	0.33333	0.33333	
Rate	0.00%	5.26%	5.26%	3.51%

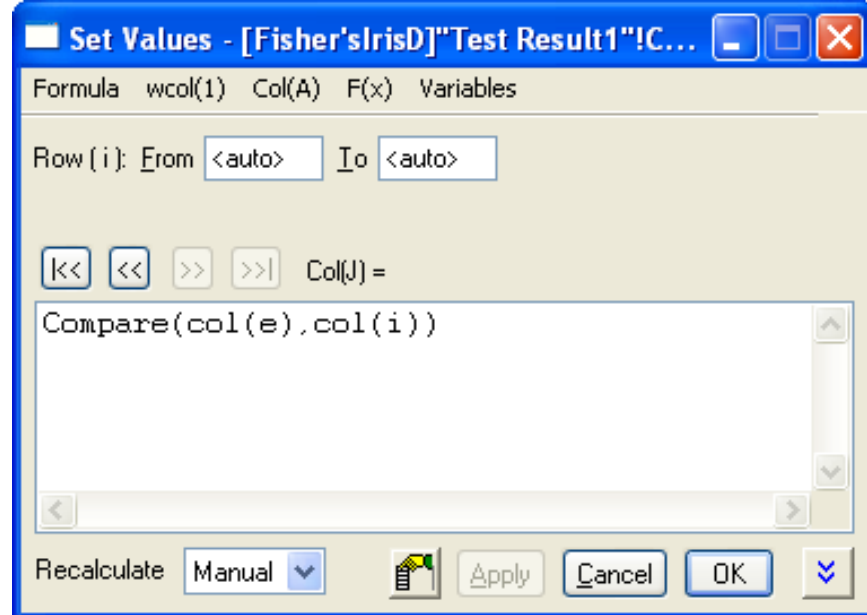
Error rate for Cross-validation of training data is 3.51%.

Classification Summary for Test Data

	setosa	virginica	versicolor	Total
Count	6	12	12	30
Percent	20.00%	40.00%	40.00%	100.00%

4. Přidejte nový sloupec **Col(J)** do listu, klikněte na něj pravou myší a zvolte nastavení **Set Column Values**. V otevřeném dialogu zadejte **Compare(col(e),col(i))** v dialogu a klikněte na **OK**.

5. Žádná z 30 hodnot není 0, což znamená, že chybovost testování dat je 0. Nalezený diskriminační model je dobrý.



Nastavení priorní pravděpodobnosti

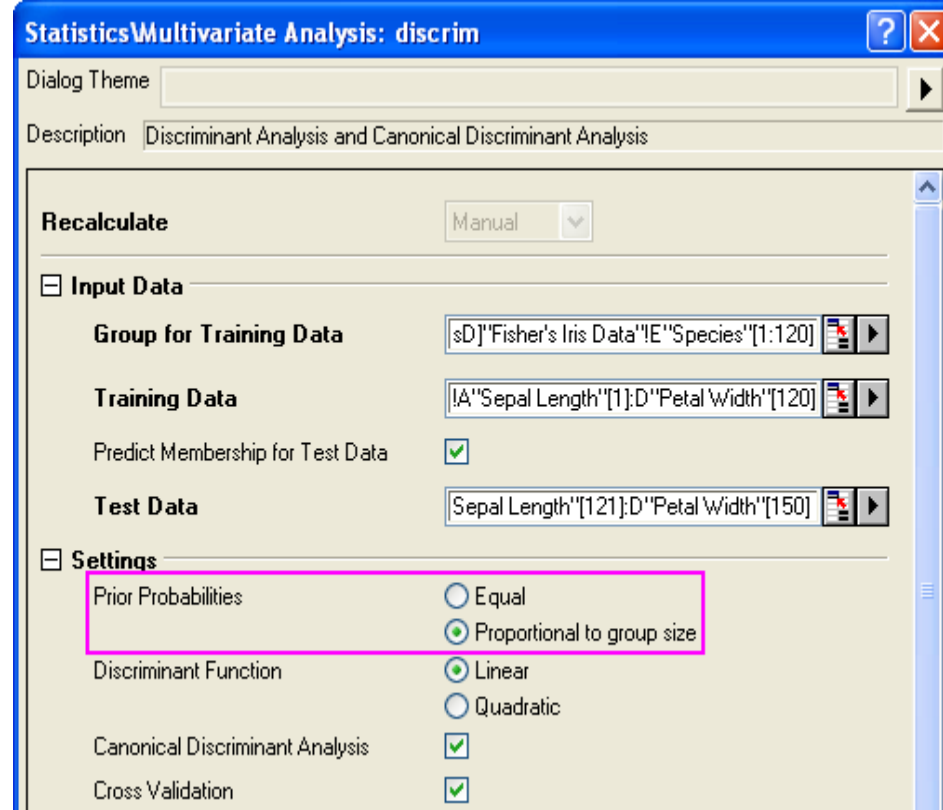
Diskriminační analýza předpokládá, že priorní pravděpodobnost příslušnosti ve skupině kosatců je identifikovatelná. Pokud se totiž velikosti skupin kosatců liší, priorní pravděpodobnosti se také liší. V tomto případě lze použít **Proportional to group size** pro priorní pravděpodobnost.

	setosa	virginica	versicolor	Total
Prior	0.33333	0.33333	0.33333	
Rate	0.00%	2.63%	5.26%	2.63%

Error rate for classification of training data is 2.63%.

1. Přejděte na list **Discrim2**, **Prior** řádek tabulky **Error Rate** v **Classification Summary for Training Data** indikuje priorní pravděpodobnost pro příslušnost ve skupině. Předpokládá se, že kosatec má stejnou pravděpodobnost, že bude v jedné ze tří skupin. Nastavení priorní pravděpodobnosti v závislosti na velikosti skupiny může zlepšit celkovou klasifikaci kosatců.

2. Klikněte na tlačítko zámku v grafu a klikněte na **Change Parameter**. Vyberte **Proportional to group size** pro políčko **Prior Probabilities**. Klikněte na **OK**.



3. Chyba klasifikace je 2,50%, což je lepší než 2,63% u míry chyb se stejnými priorními pravděpodobnostmi.

Prior Probabilities = Proportional to group

Error Rate

	setosa	virginica	versicolor	Total
Prior	0.36667	0.31667	0.31667	
Rate	0.00%	2.63%	5.26%	2.50%

Prior Probabilities = Equal

Error Rate

	setosa	virginica	versicolor	Total
Prior	0.33333	0.33333	0.33333	
Rate	0.00%	2.63%	5.26%	2.63%

