

1. cvičení

30.9.2014

Obsah

- Typy dat, SW statistika
- Popisná statistika
- Normální rozložení a důsledky nenormality pro parametrické statistické metody + simulovaný příkladový soubor s různými daty
 - T-test, ANOVA
- Korelace (s důrazem na vizualizaci klidně více proměnných pomocí maticových grafů a vliv přítomnosti odlehlé hodnoty) + simulovaný příkladový soubor s různými typy závislosti (nebo klasické kosatce)
- Kontingenční tabulky – princip vypočtu, pozorované vs. Očekávané
- Práce s binárními a kategoriálními daty – dummies, srovnání s bazální kategorií
- Grafy – základní principy grafů ve statistice

Data

- Diskrétní
 - Kategoriální (několik kategorií: anamnéza, půdní typy,...)
 - Binární
 - speciální případ dat kategoriálních, kde máme jen dvě kategorie: muž vs. žena, výskyt vs. nevýskyt,...
- Spojitá
 - Data mohou nabývat jakýchkoliv hodnot: váha, výška, krevní tlak, nadmořská výška...
- Ordinální
 - Speciální typ dat, která jsou mezi daty spojitými a kategoriálními
 - Jedná se o kategorie, které je možno seřadit: kategorie BMI, věku, nadmořské výšky,...

Popisná statistika (1)

- Spojitá data – průměr, rozptyl, směrodatná odchylka, kvantily
- Ordinální data – kvantily, modus

Statistics → Basic Statistics/Tables → Descriptive statistics

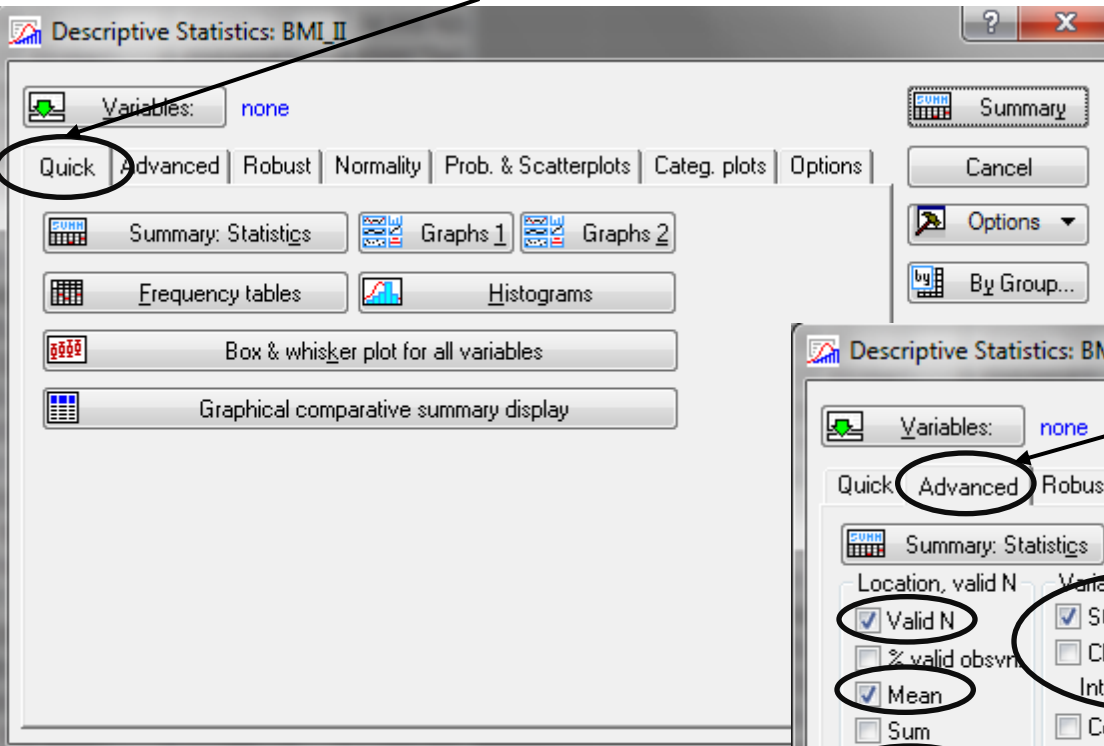
The image shows a screenshot of the STATISTICA software interface. The 'Statistics' menu is open, and the 'Basic Statistics/Tables' option is highlighted with a black oval. A large black arrow points from this menu item to a dialog box titled 'Basic Statistics and Tables: BMI_II'. In this dialog box, the 'Quick' tab is active, and the 'Descriptive statistics' option is highlighted with a black oval. The dialog box also shows other options like 'Correlation matrices', 't-test, independent, by groups', etc.

	1	prv
	vek	
78999.000000	90	1821
67805.000000	44	5531
21072.000000	85	9327
21080.000000	85	6161
22431.000000	66	8528
97366.000000	64	6590
75751.000000	12	6703
12590.000000	51	8487
109814.000000	11	2320
80222.000000	68	2057
71053.000000	13	7028
33525.000000	71	9760
33172.000000	69	4732
73793.000000	12	0714
47539.000000	84	8534
61320.000000	74	7328

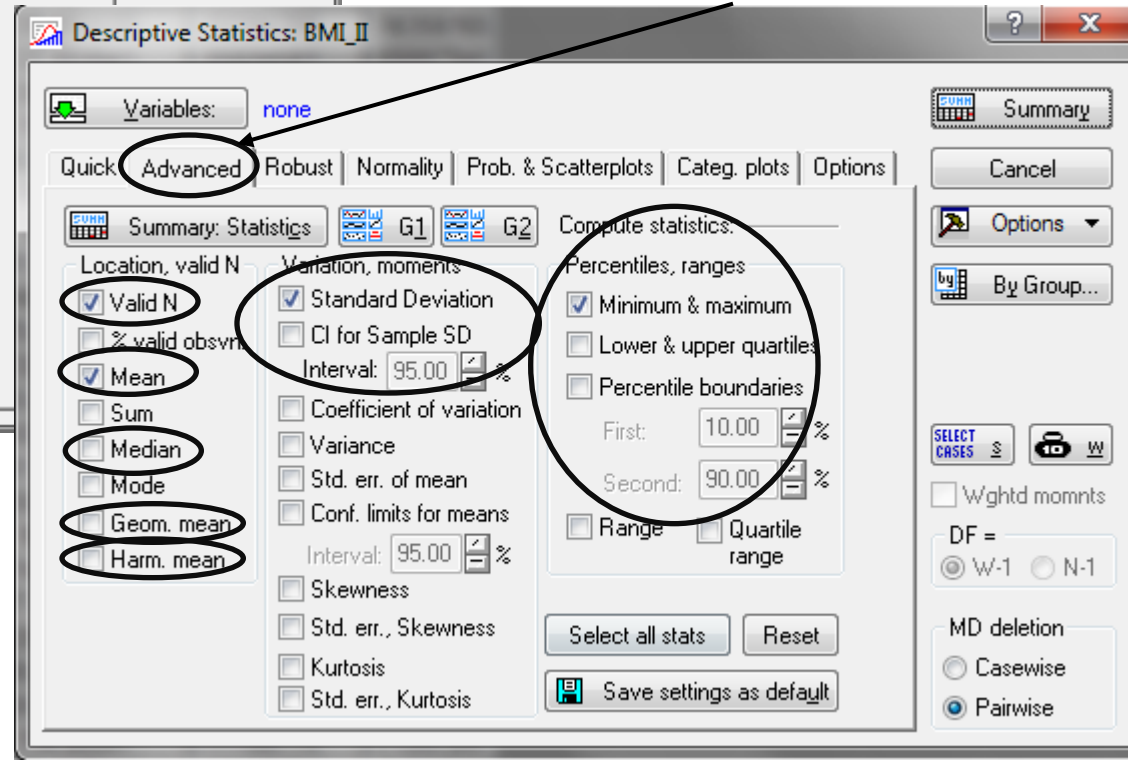
Popisná statistika (2)

- Záložka Quick a Advanced

Základní varianta – pro rychlý přehled

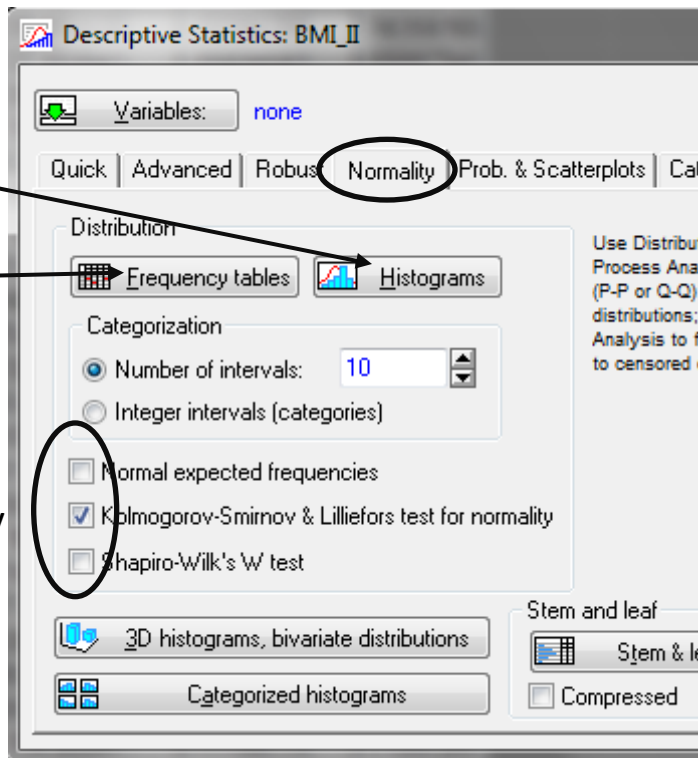


Detailnější varianta – pro pokročilé



Ověření předpokladu normality dat

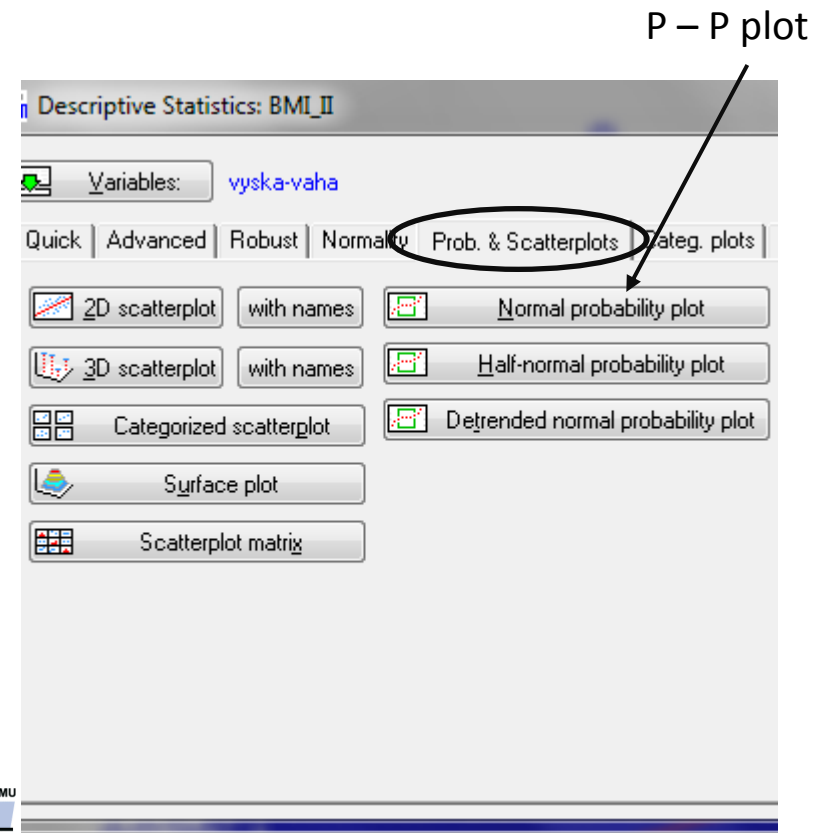
- Většina statistických metod předpokládá normalitu dat → parametrické metody: t-test, ANOVA
- Záložka Normality → Testy normality, histogram, frekvenční tabulka
- Záložka Prob. & Scatterplots → P-P plot



Histogram

Frekvenční tabulka

Testy normality



Transformace dat

- Nelineární, odmocňová, Boxova-Coxova transformace

Nelineární, odmocňová transformace

Postup: Data → Variable specs...

Variable 7

Name: Type: OK

Measurement Type: Length: Cancel

Excluded Label Case State MD code:

Display format

- General
- Number
- Date
- Time
- Scientific
- Currency
- Percentage
- Fraction
- Custom

Long name (label or formula with): Function guide

Labels: use any text. Formulas: use variable names or v1, v2, ..., v0 is case #.
Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after:)

Boxova-Coxova Transformace

Postup: Data → Box-Cox transformation

Box-Cox Transformation: BMI_II

Box-Cox

Variables:

Box-Cox transformation

Max. iterations:

Min. lambda:

Max. lambda:

Epsilon (convergence):

Shift variables with minimum ≤ 0 to:

OK

Cancel

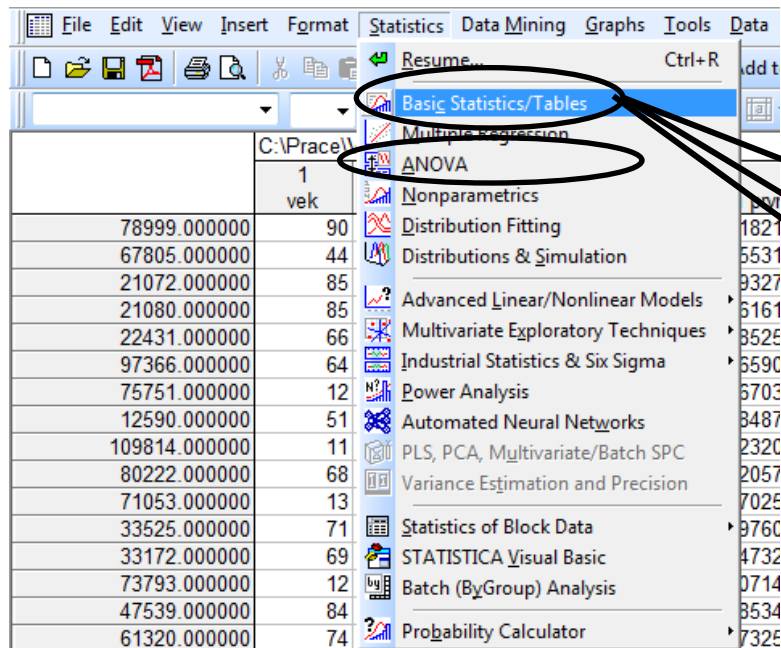
Options

SELECT CASES

Open Data

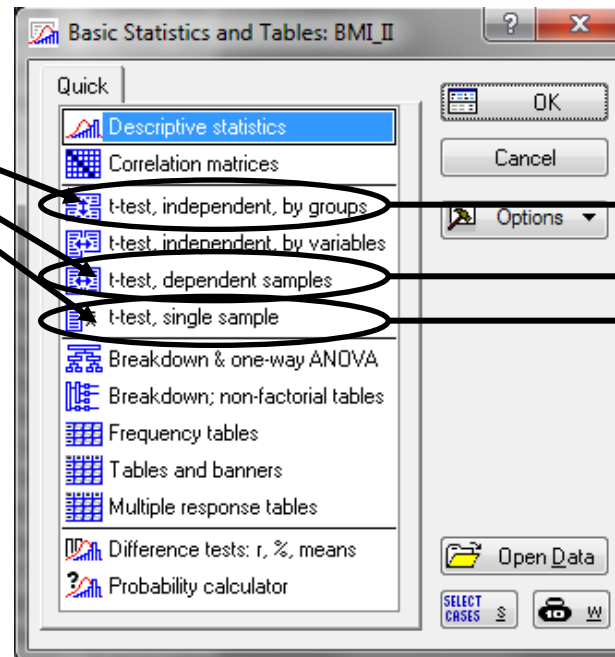
T-test, ANOVA

- Parametrické metody: obě předpoklad normality, ANOVA předpoklad homogenity rozptylu ve skupinách a nezávislost skupin



The screenshot shows the SPSS Statistics menu with 'Basic Statistics/Tables' and 'ANOVA' circled. The 'ANOVA' option is further expanded to show 't-test, independent, by groups', 't-test, independent, by variables', 't-test, dependent samples', and 't-test, single sample', all of which are also circled.

File	Edit	View	Insert	Format	Statistics	Data Mining	Graphs	Tools	Data
78999.000000	90	1821							
67805.000000	44	5531							
21072.000000	85	9327							
21080.000000	85	6161							
22431.000000	66	8525							
97366.000000	64	6590							
75751.000000	12	6703							
12590.000000	51	8487							
109814.000000	11	2320							
80222.000000	68	2057							
71053.000000	13	7025							
33525.000000	71	9760							
33172.000000	69	4732							
73793.000000	12	0714							
47539.000000	84	8534							
61320.000000	74	7325							

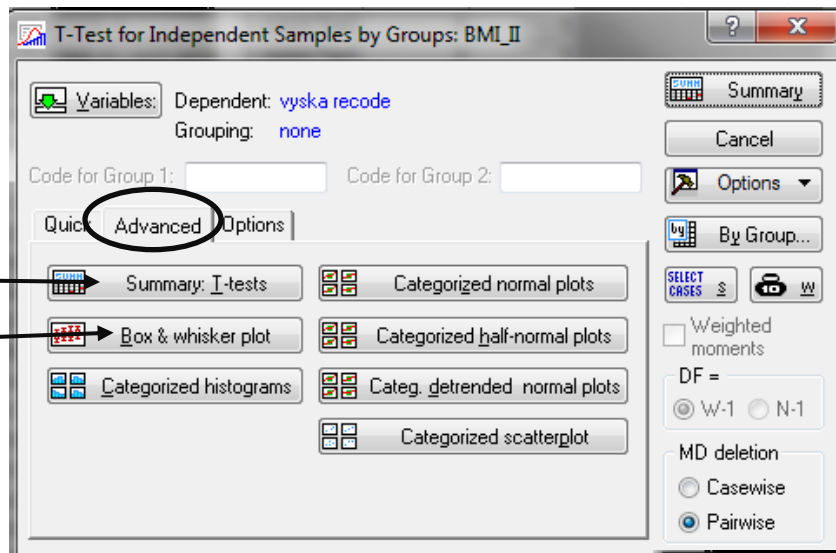


Dvouvýběrový t-test

Párový t-test

Jednovýběrový t-test

Dvouvýběrový T-test

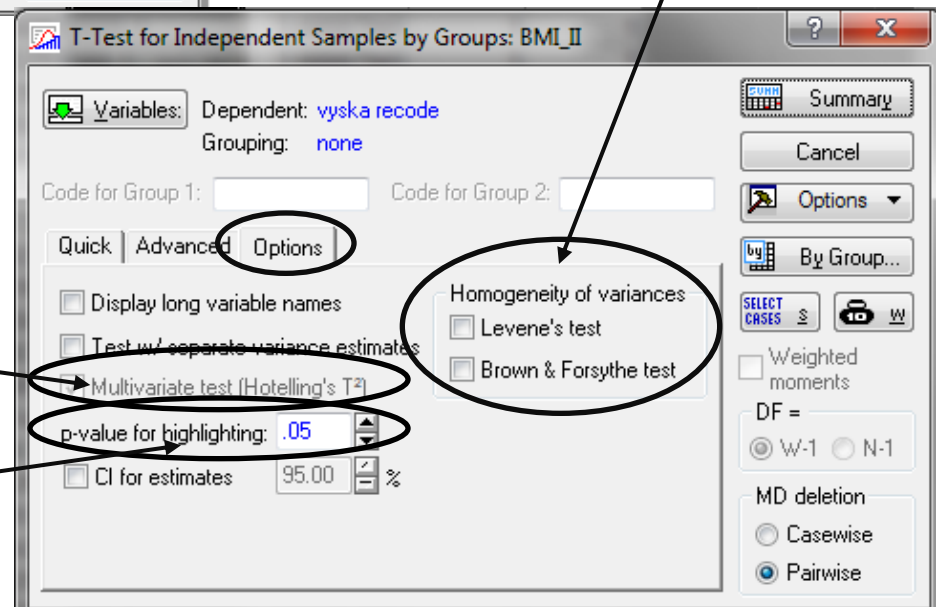


Výsledek
Boxplot

Testy homogeneity

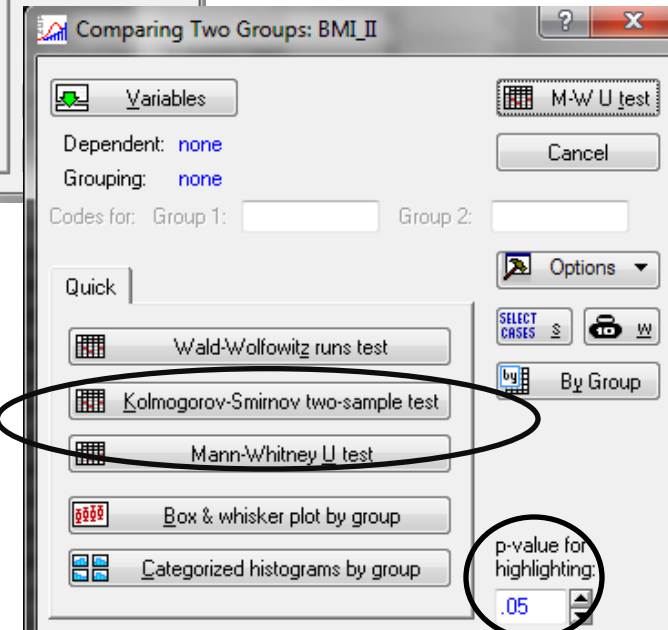
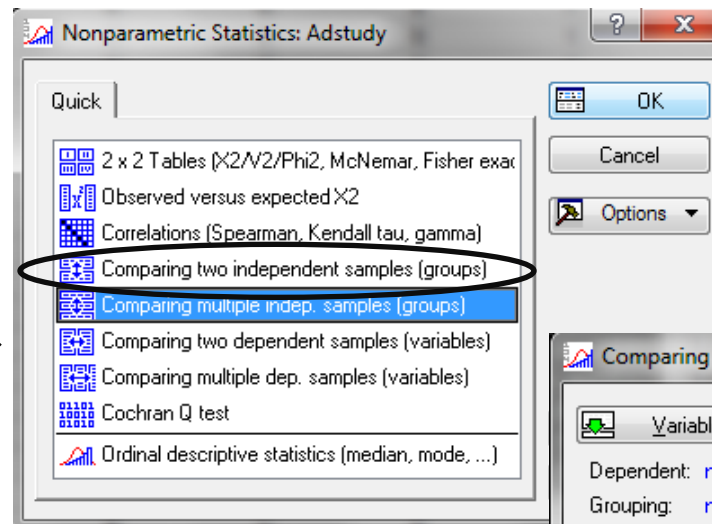
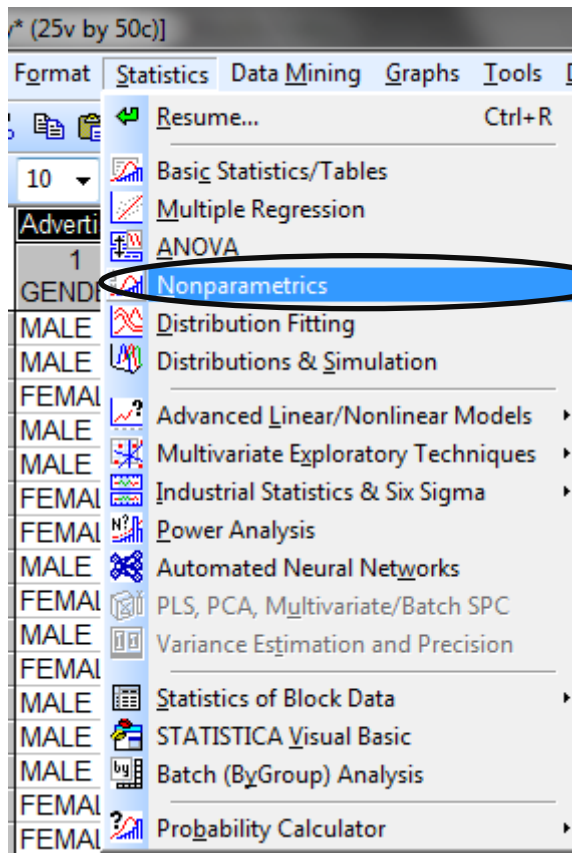
T-test pro více rozměrů

Nastavení hladiny
statistické významnosti

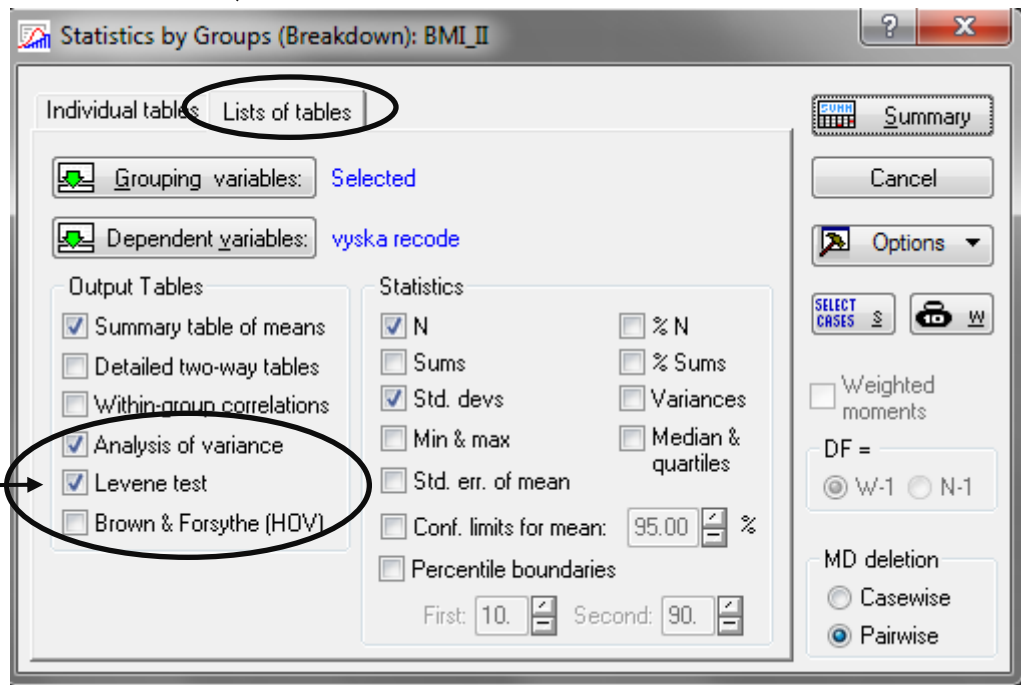
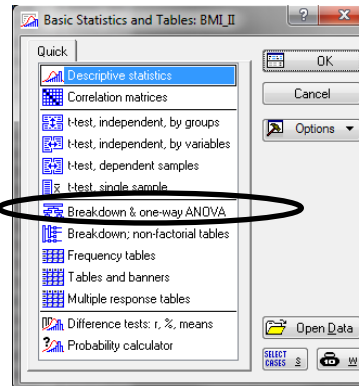
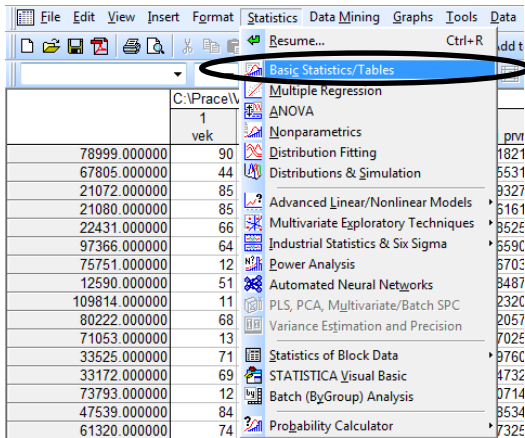


Nesplnění předpokladů pro T-test

- Neparametrické testy
 - Kolmogorův-Smirnovův test
 - Mannův-Whitneyův test

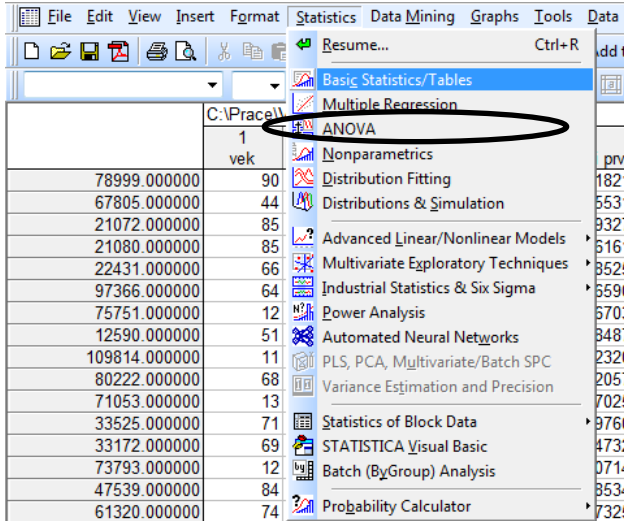


Ověření předpokladu homogenity rozptylu pro ANOVu

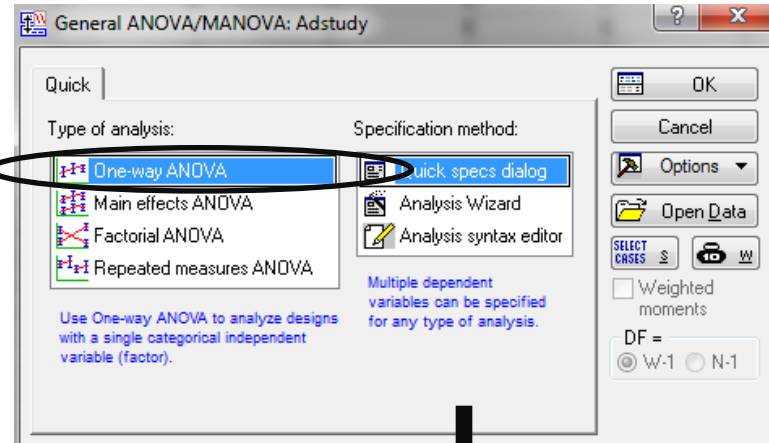


ANOVA – analýza rozptylu

- Spojité proměnné dle kategorií
- Zda na hodnotu spojité proměnné má vliv hodnota kategoriální proměnné

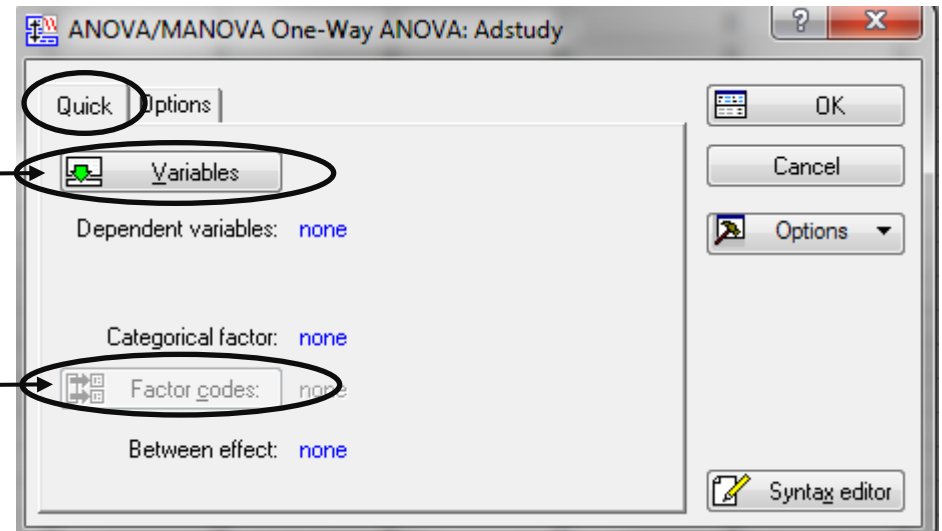


The screenshot shows the SPSS menu structure. The 'ANOVA' option is circled in red. The menu items include: Basic Statistics/Tables, Multiple Regression, ANOVA, Nonparametrics, Distribution Fitting, Distributions & Simulation, Advanced Linear/Nonlinear Models, Multivariate Exploratory Techniques, Industrial Statistics & Six Sigma, Power Analysis, Automated Neural Networks, PLS, PCA, Multivariate/Batch SPC, Variance Estimation and Precision, Statistics of Block Data, STATISTICA Visual Basic, Batch (ByGroup) Analysis, and Probability Calculator.



Spojité proměnná

Kategoriální proměnná



ANOVA – výsledky 1

Boxploty

Výsledný vliv kategoriální proměné

Celkový výsledek

ANOVA Results 1: BMI_II

Profiler | Resids | Matrix | Report

Quick | Summary | Means | Comps

All effects/Graphs | All effects

Univariate results | Cell statistics

Between effects

Design terms | Whole model R

Coefficients | Estimate

Alpha values

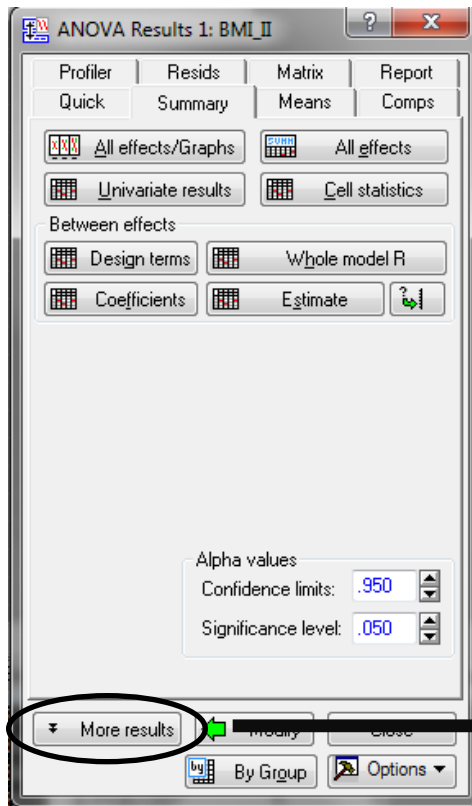
Confidence limits: .950

Significance level: .050

More results | Modify | Close

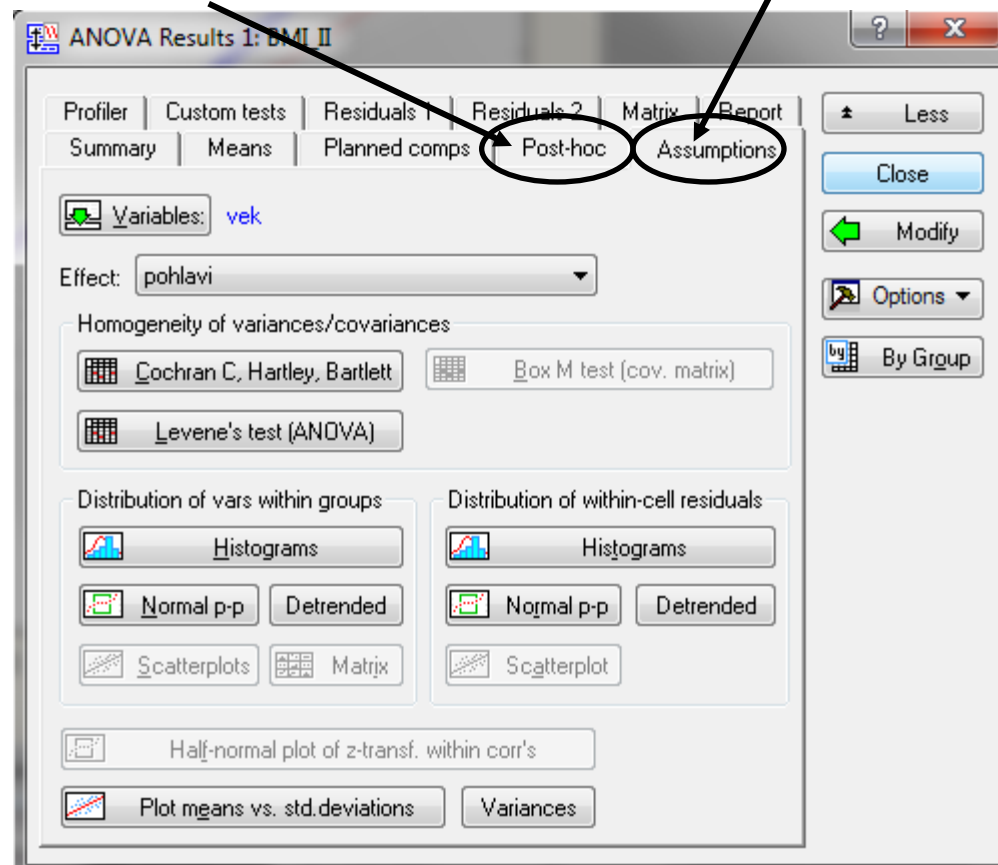
By Group | Options

ANOVA – výsledky 2



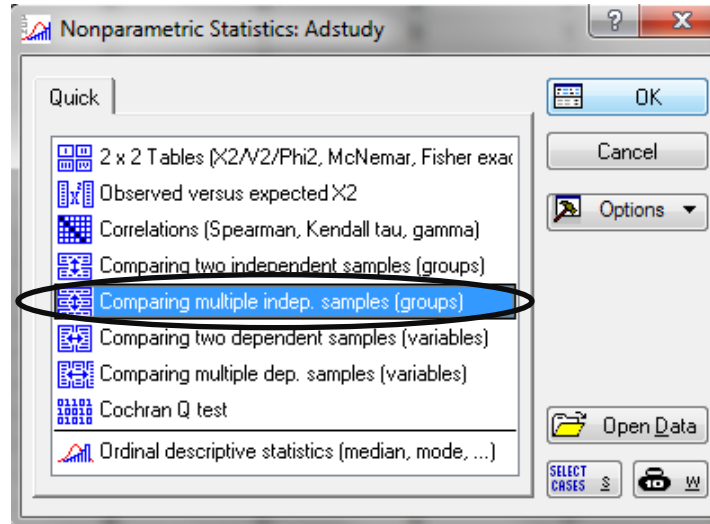
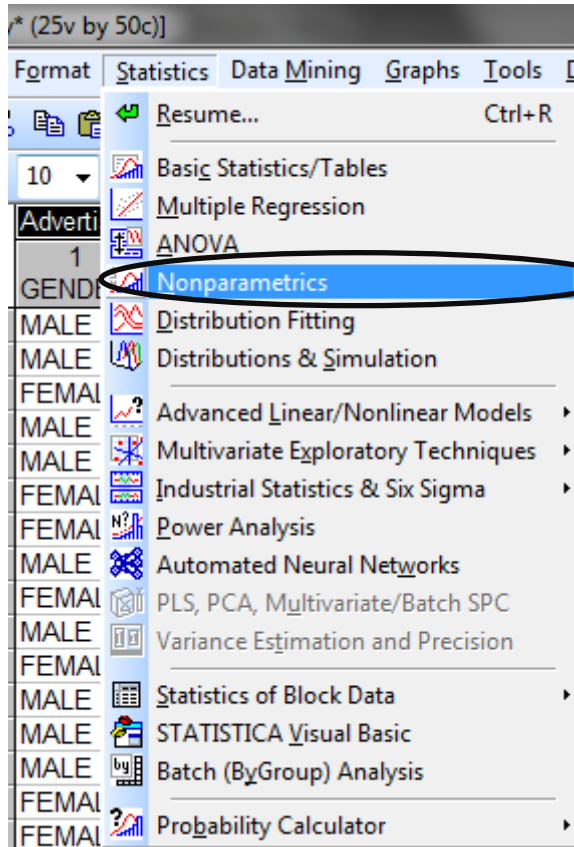
Post-Hoc testy
→ stanovení homogenních skupin

Ověření předpokladů

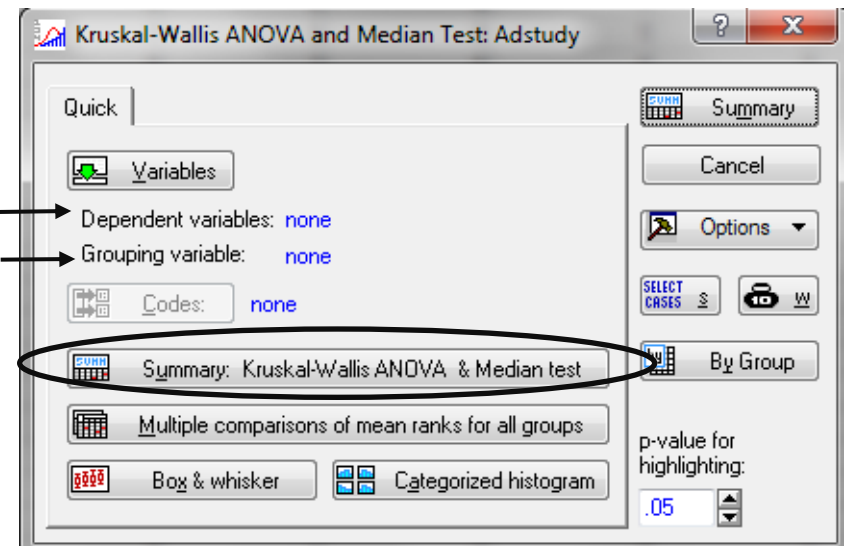


ANOVA – porušení homogenity rozptylu

- Neparametrická metoda Kruskalova – Willisova analýza rozptylu



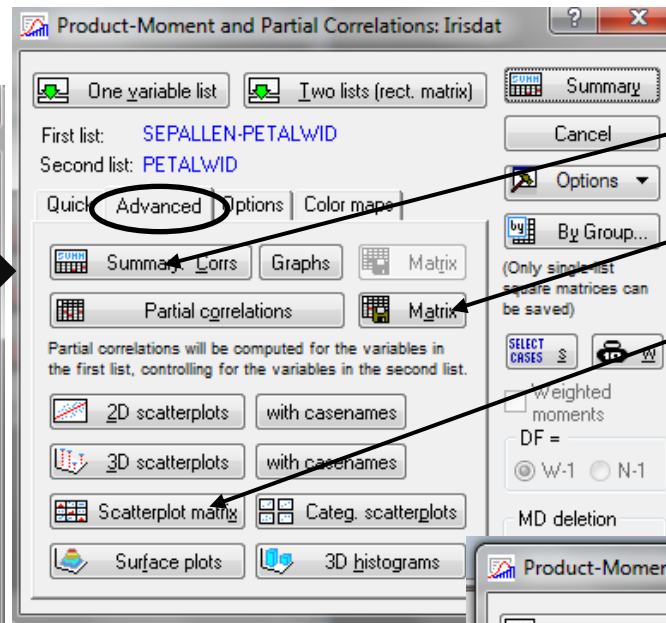
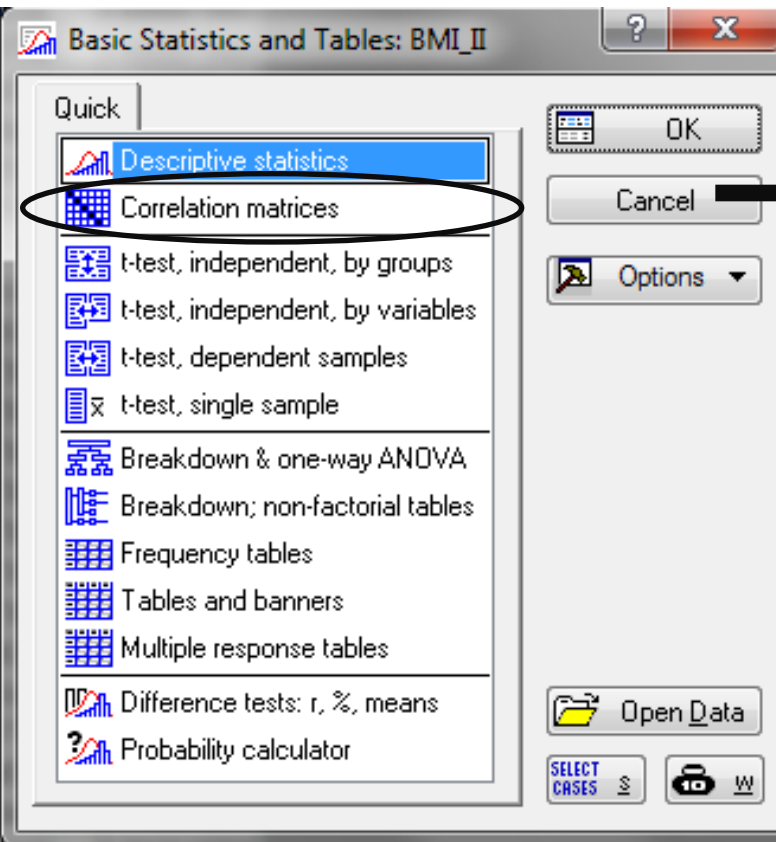
Spojité proměnná →
Kategoriální proměnná →



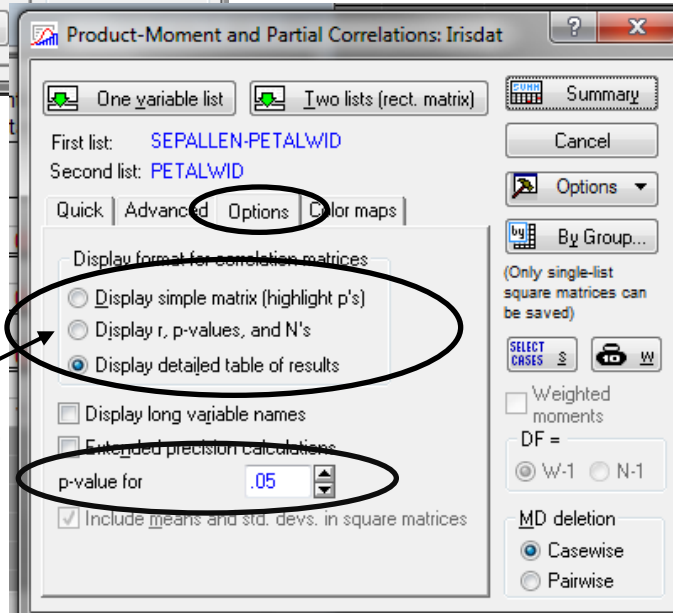
Korelace 1

- Spojité a normálně rozložené proměnné → Pearsonův korelační koeficient
- Pro ostatní proměnné → Spearmanův (pořadový) korelační koeficient

Pearsonův korelační koeficient



Celkový výsledek
Možnost uložení
korelační matice
Maticový graf

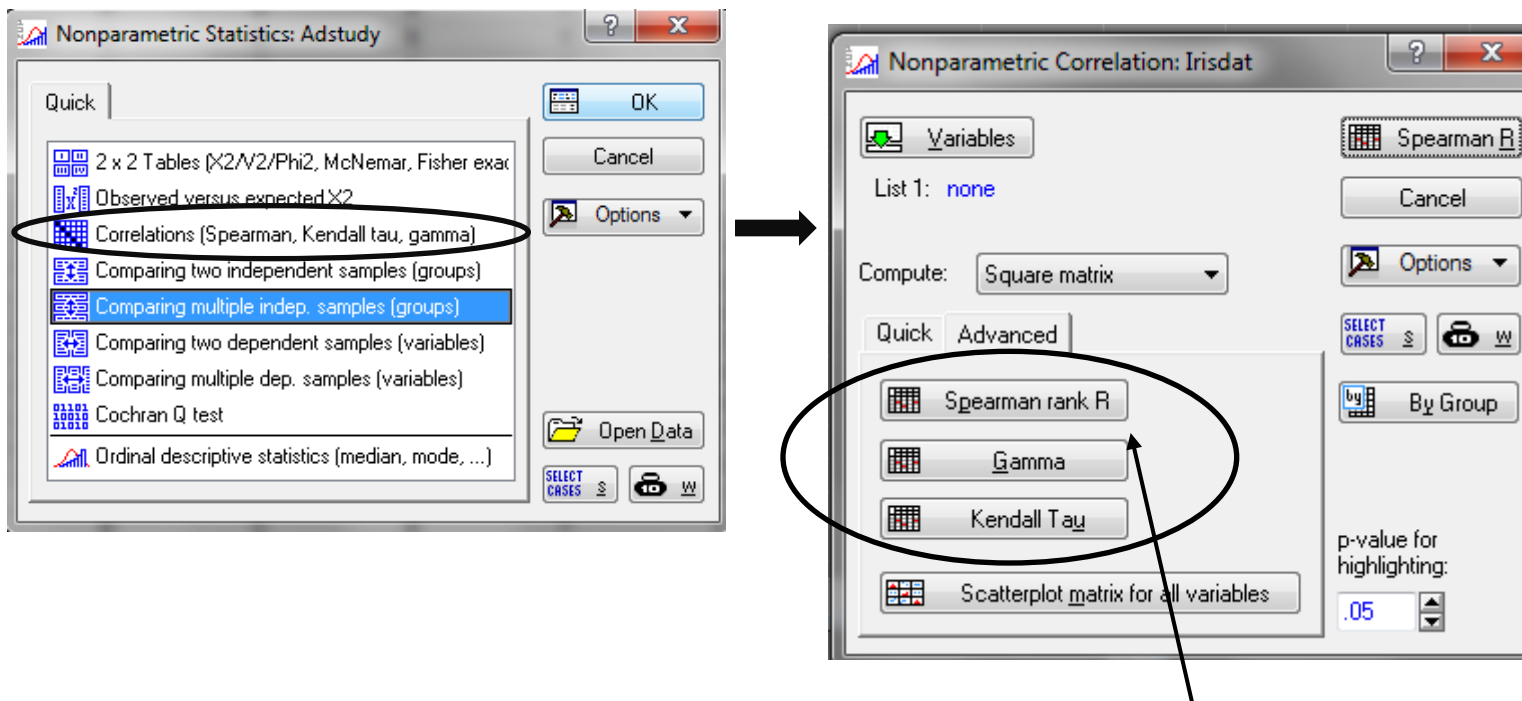


Nastavení výsledné matice

Korelace 2

- Spojité a normálně rozložené proměnné → Pearsonův korelační koeficient
- Pro ostatní proměnné → Spearmanův (pořadový) korelační koeficient

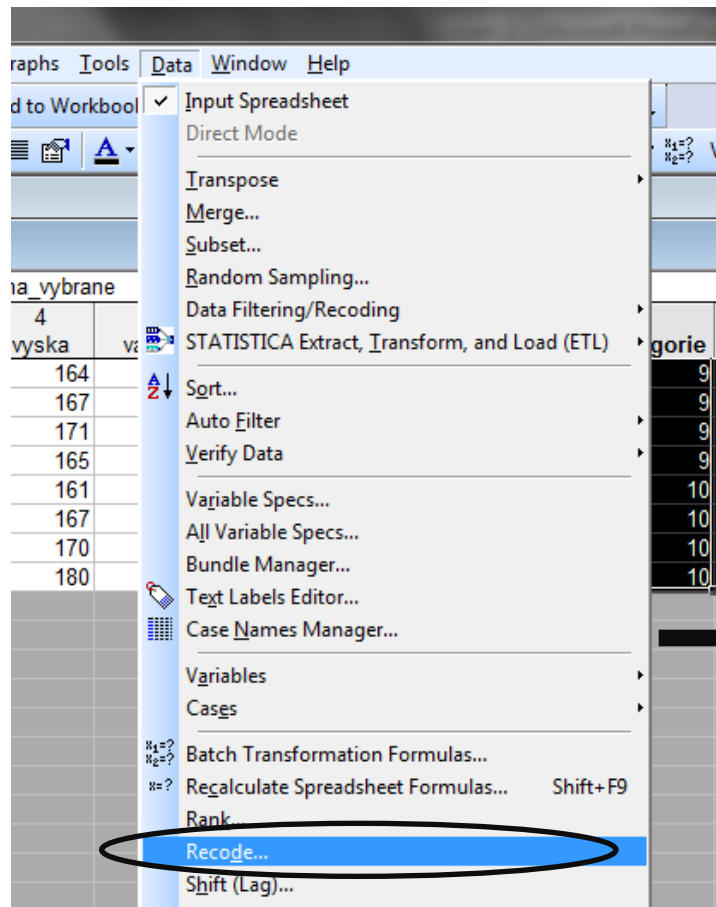
Spearmanův korelační koeficient



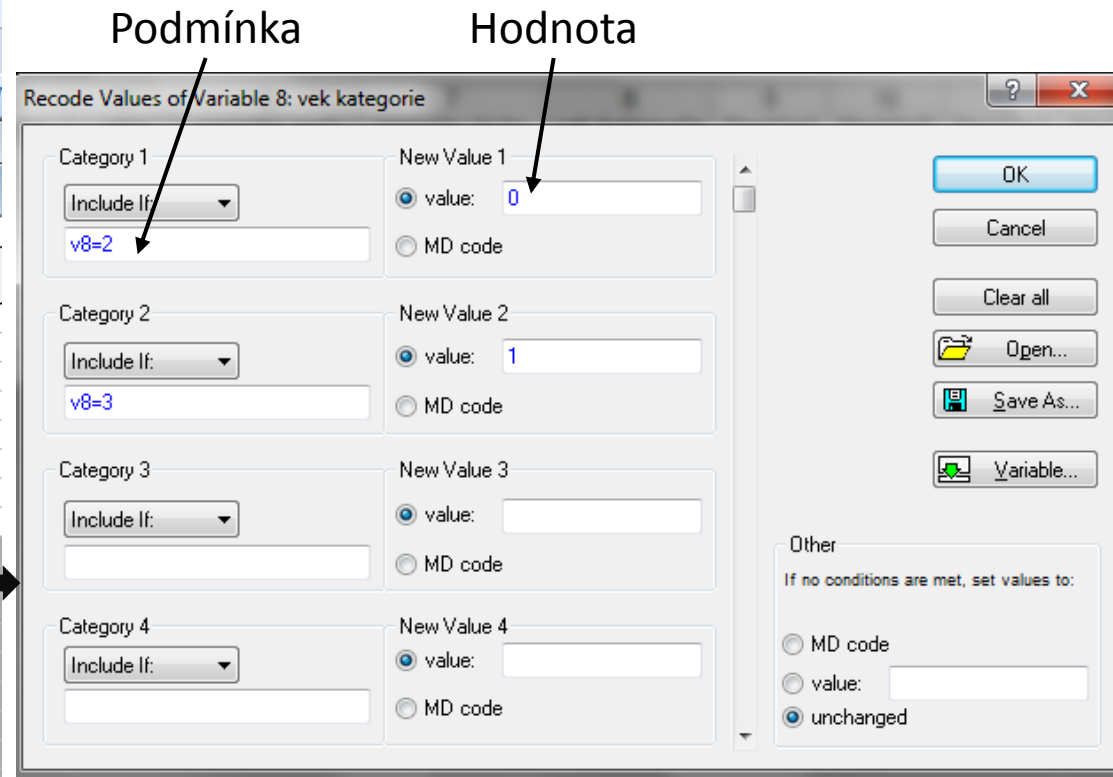
Spearmanův korelační koeficient
a jiné neparametrické korelační koeficienty

Binární a kategoriální data

- Binární → výskyt vs. nevýskyt, muž vs. žena, ano vs. ne, ...
- Kategoriální → kategorie, možno převést na binární
- Srovnání s bazální kategorií → použití při logistické regresi, jedna kategorie vybrána jako základní a k ní jsou vztahovány ostatní



The screenshot shows the SPSS Data menu with the 'Recode...' option highlighted in blue. The menu also includes options like 'Transpose', 'Merge...', 'Subset...', 'Random Sampling...', 'Data Filtering/Recoding', 'STATISTICA Extract, Transform, and Load (ETL)', 'Sort...', 'Auto Filter', 'Verify Data', 'Variable Specs...', 'All Variable Specs...', 'Bundle Manager...', 'Text Labels Editor...', 'Case Names Manager...', 'Variables', 'Cases', 'Batch Transformation Formulas...', 'Recalculate Spreadsheet Formulas...', 'Rank', and 'Shift (Lag)...'. The background shows a data grid with columns 'vyska' and 'vek' and rows of numerical data.

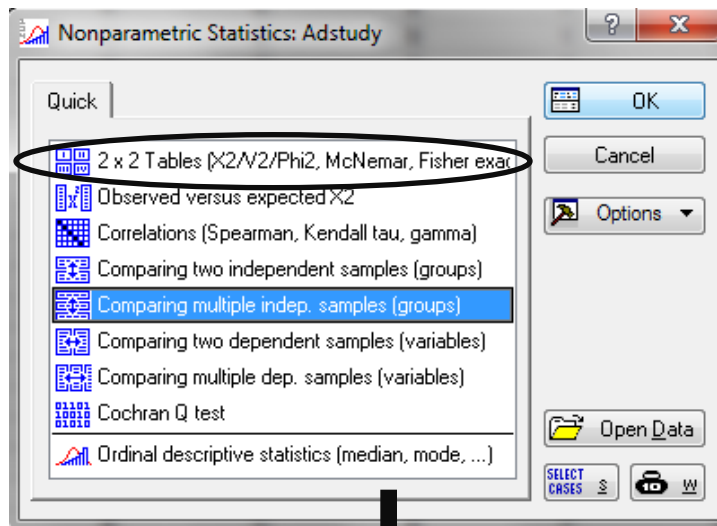


The screenshot shows the 'Recode Values of Variable 8: vek kategorie' dialog box. It has two main sections: 'Podmínka' (Condition) and 'Hodnota' (Value). The 'Podmínka' section contains four 'Include If:' dropdown menus, with the first two containing 'v8=2' and 'v8=3'. The 'Hodnota' section contains four 'New Value' sections, each with a radio button for 'value:' and 'MD code'. The first 'New Value 1' section has 'value:' set to '0'. The 'Other' section at the bottom has radio buttons for 'MD code', 'value:', and 'unchanged', with 'unchanged' selected. The dialog box has 'OK', 'Cancel', 'Clear all', 'Open...', 'Save As...', and 'Variable...' buttons.

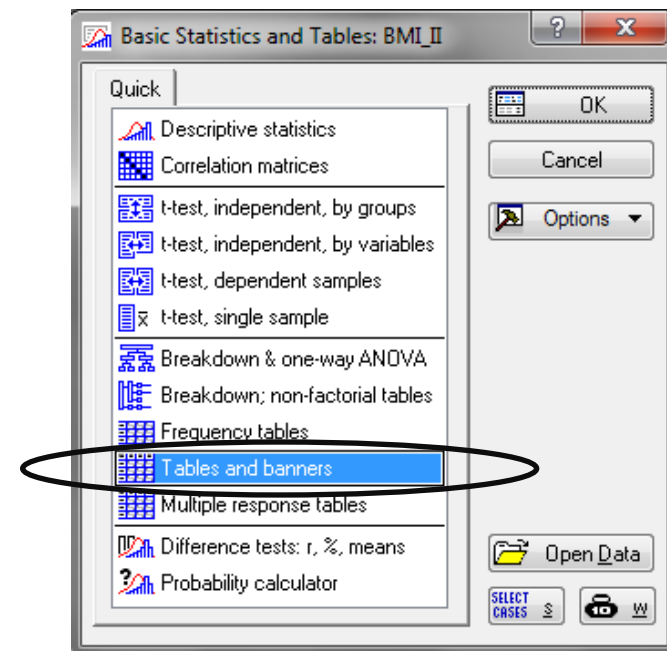
Kontingenční tabulky

- K přehledné vizualizaci vztahu dvou kategoriálních proměnných
- Excel nebo Statistika
- Chi-square test, Fisherův exaktní test,....

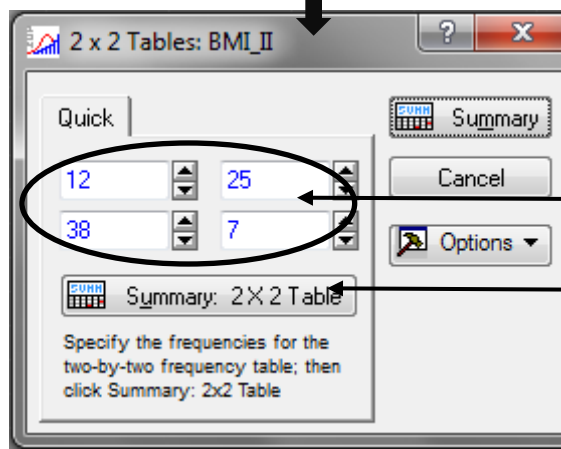
Varianta 2 x 2



Varianta n x m



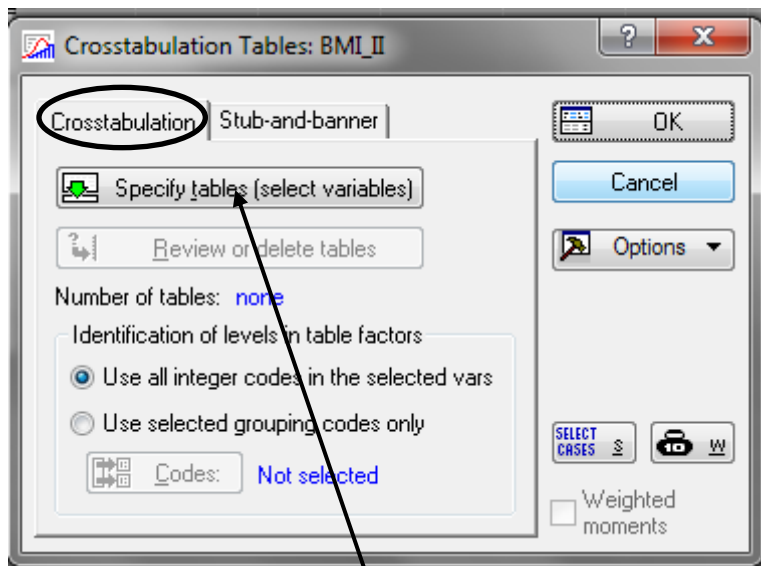
VS.



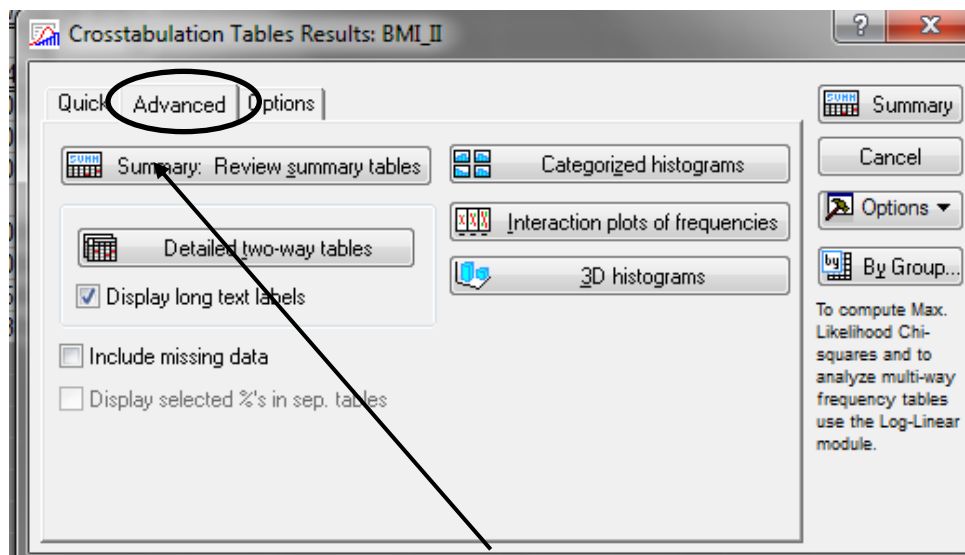
Hodnoty v kontingenční tabulce

Výpočet testů

Kontingenční tabulky rozměrů $n \times m$



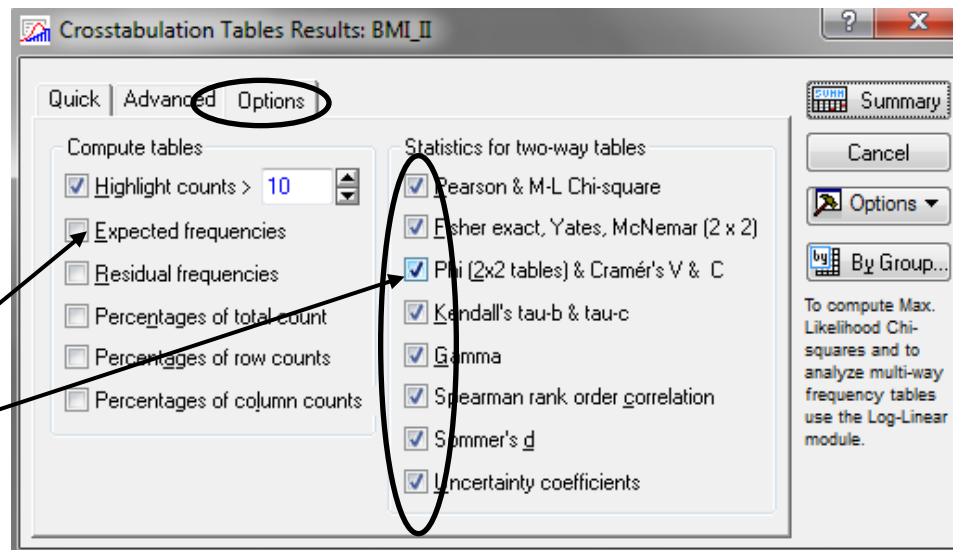
Proměnné



Výpočet kontingenční tabulky

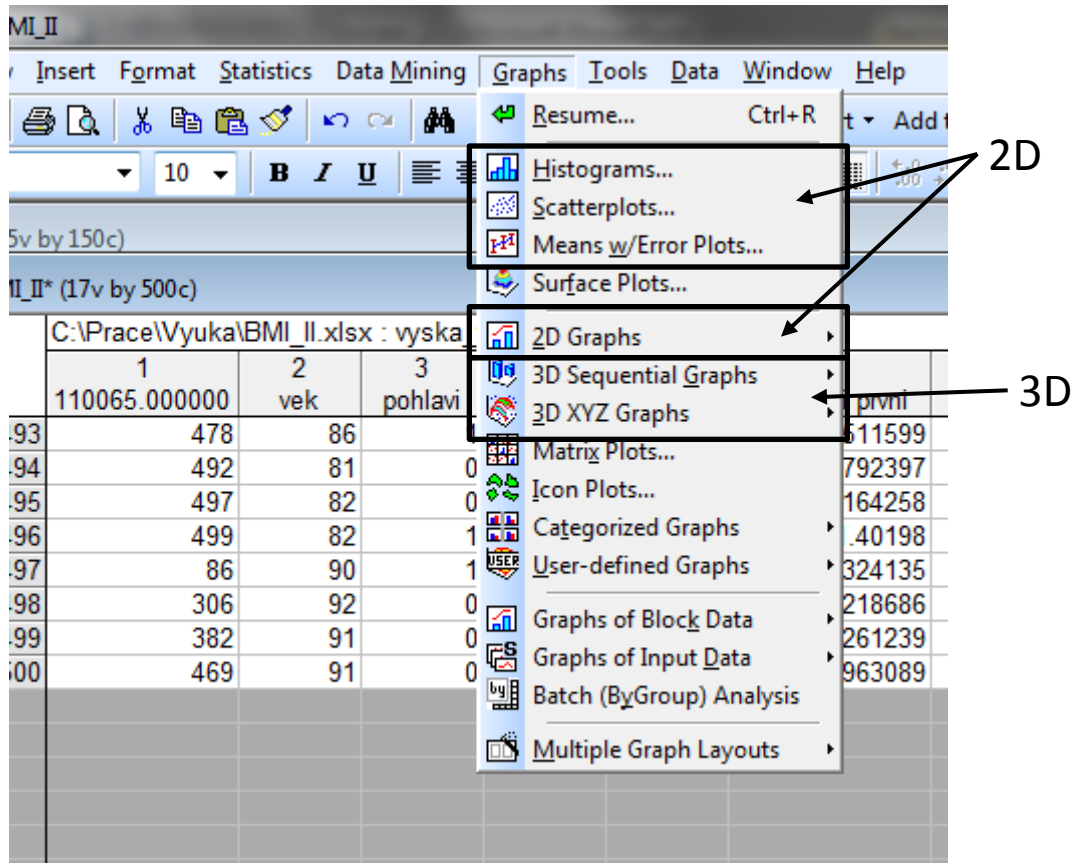
Očekávané četnosti

Testy



Grafy ve statistice

- Histogramy, boxploty, xy grafy



The screenshot shows the Minitab software interface with the 'Graphs' menu open. The menu items are as follows:

- Resume... (Ctrl+R)
- Histograms...
- Scatterplots...
- Means w/Error Plots...
- Surface Plots...
- 2D Graphs
- 3D Sequential Graphs
- 3D XYZ Graphs
- Matrix Plots...
- Icon Plots...
- Categorized Graphs
- User-defined Graphs
- Graphs of Block Data
- Graphs of Input Data
- Batch (ByGroup) Analysis
- Multiple Graph Layouts

Annotations in the image:

- A box labeled '2D' encompasses the 'Histograms...', 'Scatterplots...', and 'Means w/Error Plots...' options.
- A box labeled '3D' encompasses the '3D Sequential Graphs' and '3D XYZ Graphs' options.

The background shows a spreadsheet with the following data:

	1	2	3	
	110065.000000	vek	pohlavi	
.93	478	86		
.94	492	81	0	
.95	497	82	0	
.96	499	82	1	
.97	86	90	1	
.98	306	92	0	
.99	382	91	0	
.00	469	91	0	