

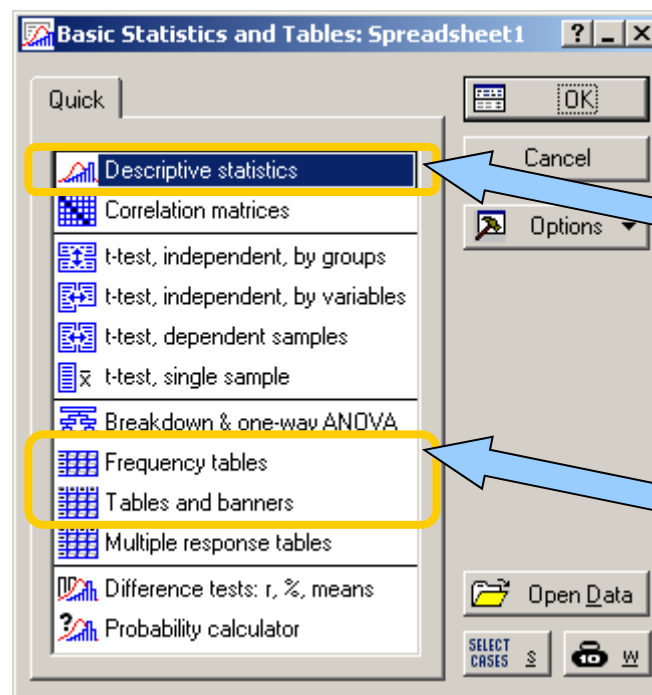
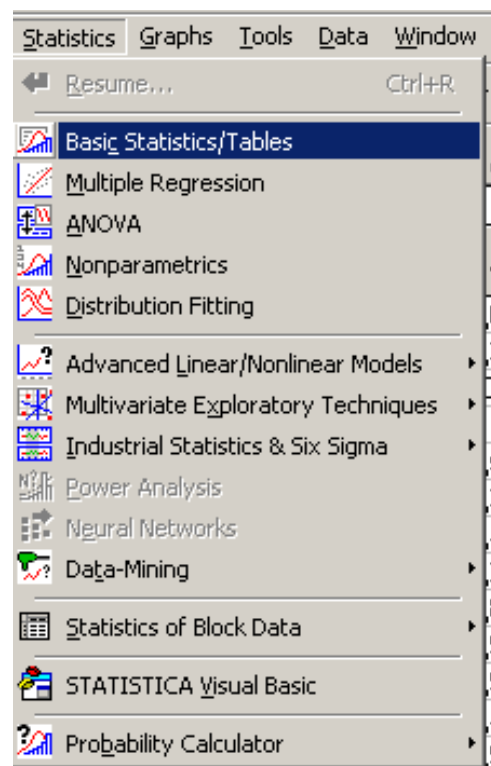
Analýza dat na PC I.

Popisná analýza v programu Statistica

IBA výuka

Analýza dat na PC I.

Základní popisná statistika



Popisná statistika

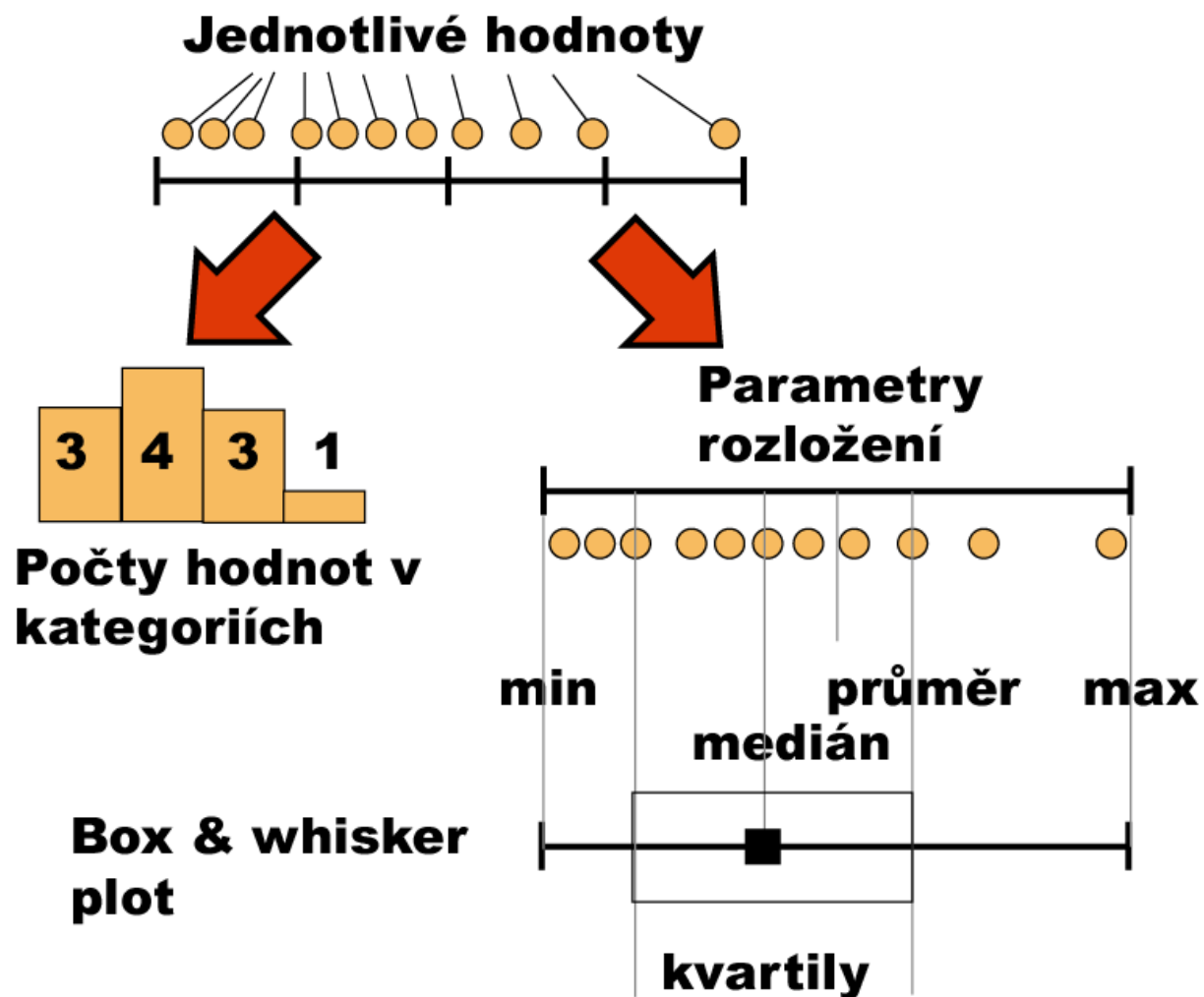
Frekvenční
tabulky, analýza
kontingenčních
tabulek

Typy proměnných

- ◆ **Kvalitativní/kategorická**
 - ◆ binární - ano/ne
 - ◆ nominální - A,B,C ... několik kategorií
 - ◆ ordinální - $1 < 2 < 3$...několik kategorií a můžeme se ptát, která je větší
- ◆ **Kvantitativní**
 - ◆ nespojitá – čísla, která však nemohou nabývat všech hodnot (např. počet porodů)
 - ◆ spojitá – teoreticky jsou možné všechny hodnoty (např. krevní tlak)

Analýza dat na PC I.

Řada dat a její vlastnosti



Frekvenční rozložení

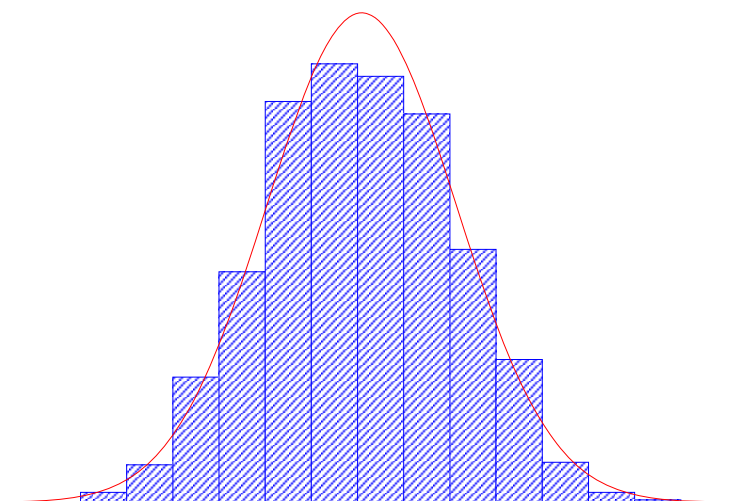
Kategorie	Četnost
B	5
C	8
D	1

Kvalitativní data

Tabulka s četností jednotlivých kategorií.

Kvantitativní data

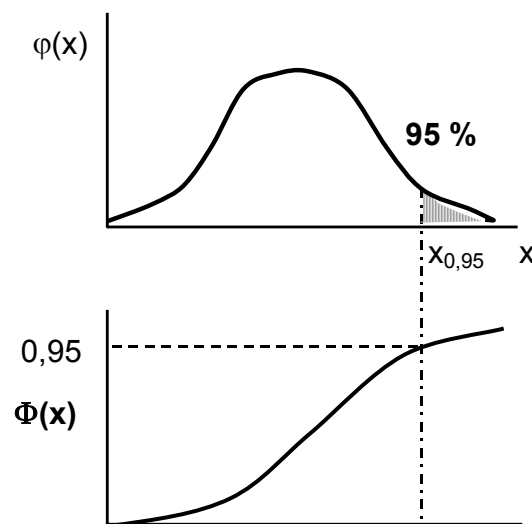
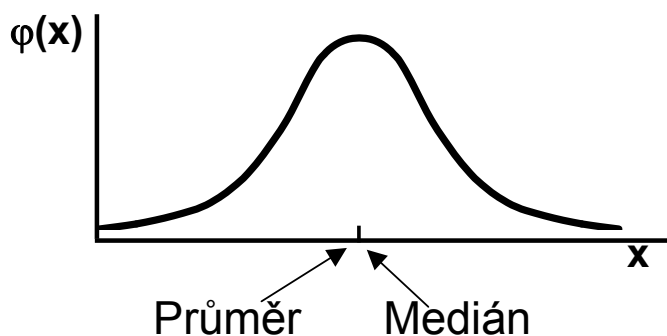
Četnost hodnot rozložení v jednotlivých intervalech.



Analýza dat na PC I.

Parametry rozložení

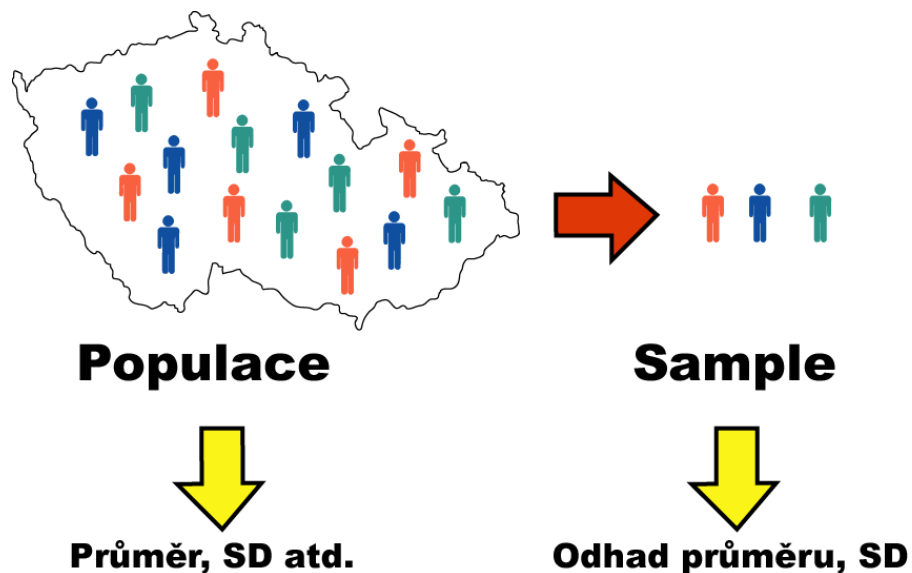
- ◆ Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- ◆ Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
 - ◆ Středu (medián, průměr, geometrický průměr)
 - ◆ Šířky rozložení (rozsah hodnot, rozptyl, směrodatná odchylka)
 - ◆ Tvaru rozložení (skewness, kurtosis)
 - ◆ Kvantily rozložení – kolik % řady dat leží nad a pod kvantilem



Analýza dat na PC I.

Populace a vzorek

- ◆ Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozložení
- ◆ Z populace je prováděno vzorkování za účelem získání reprezentativního vzorku (**sample**) populace, toto vzorkování by mělo být náhodné, důležitá je také velikost vzorku, ze vzorku získáme odhady parametrů rozložení

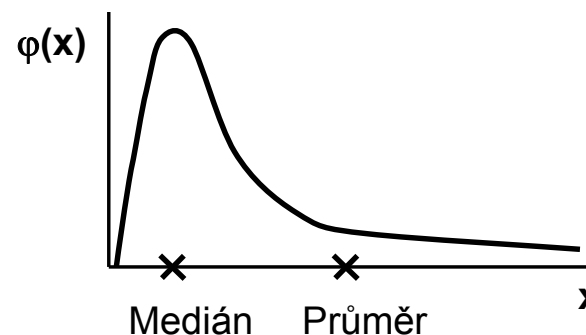
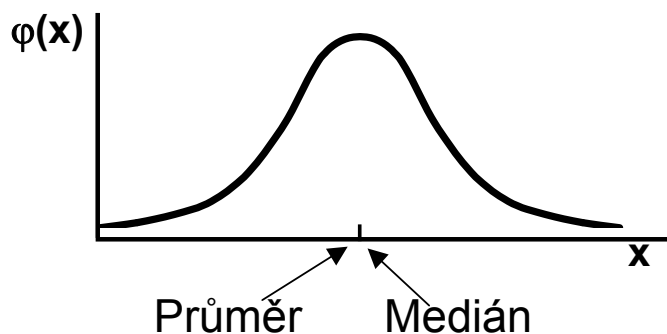


Ukazatele středu rozložení I

- ◆ **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde x_i jsou jednotlivé hodnoty a n jejich počet

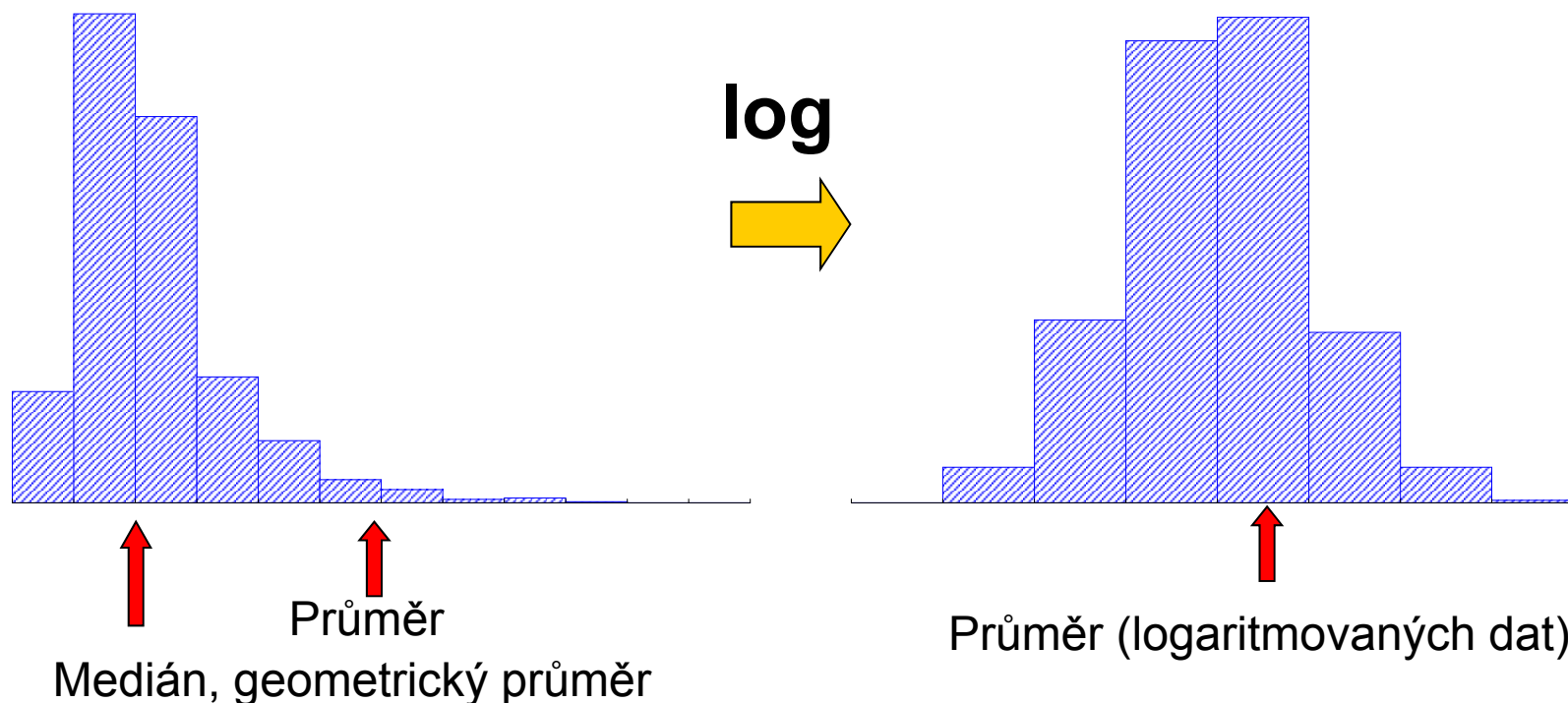
$$E(x) = \bar{x} = \sum_{i=1}^n \frac{x_i}{n}$$

- ◆ **Medián** – jde vlastně o 50% kvantil, tj. polovina hodnot leží nad a polovina pod mediánem
- ◆ V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné



Ukazatele středu rozložení II.

- ◆ Geometrický průměr – antilogaritmus průměru logaritmovaných dat, je vhodný pro doleva asymetrická data (lognormální rozložení), která jsou v biologii velmi častá, jeho hodnota v podstatě odpovídá mediánu
- ◆ Takto asymetrická data je možné převést logaritmickou transformací na

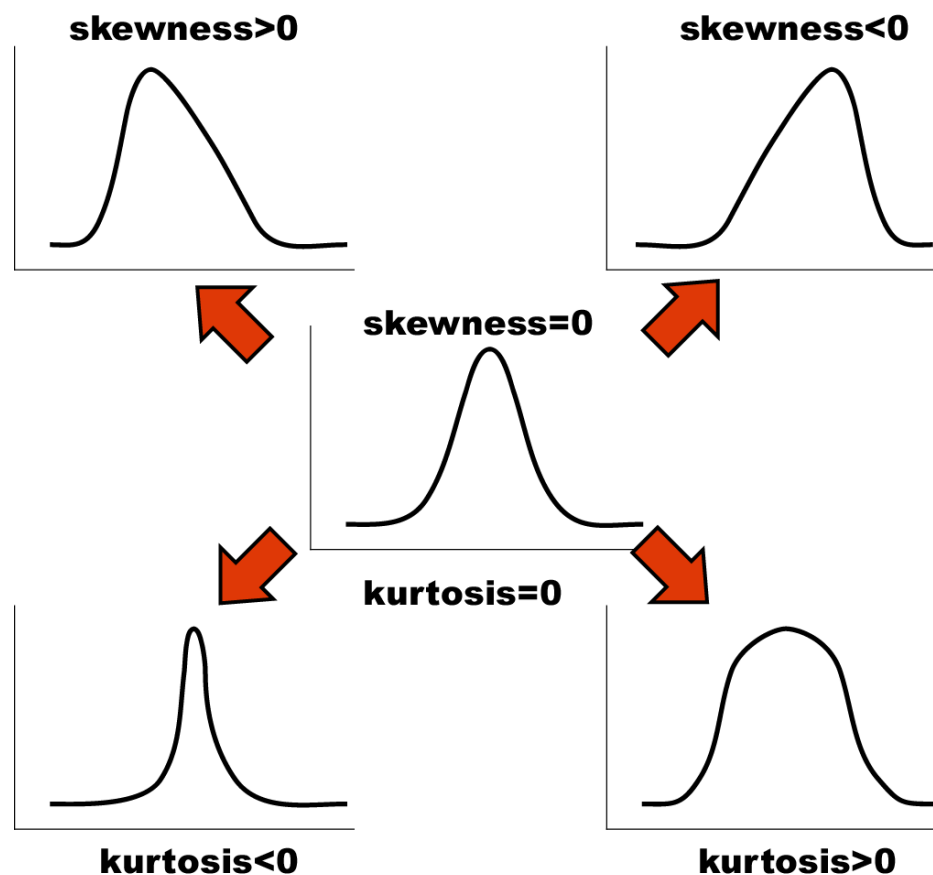


Ukazatele šířky rozložení

- ◆ **Rozptyl** je ukazatelem šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru.
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$
- ◆ Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení
- ◆ **Směrodatná odchylka** je druhá odmocnina z rozptylu
- ◆ **Koeficient variance** - podíl SD ku průměru (u normálního rozložení by se 95% hodnot mělo vejít do průměr ± 3 SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozložení – ukazatel problémů s normalitou dat

Ukazatele tvaru rozložení

- ◆ **Skewness** – ukazatel „šikmosti“ rozložení, asymetrie rozložení
- ◆ **Kurtosis** – ukazatel „špičatosti/plochosti“ rozložení

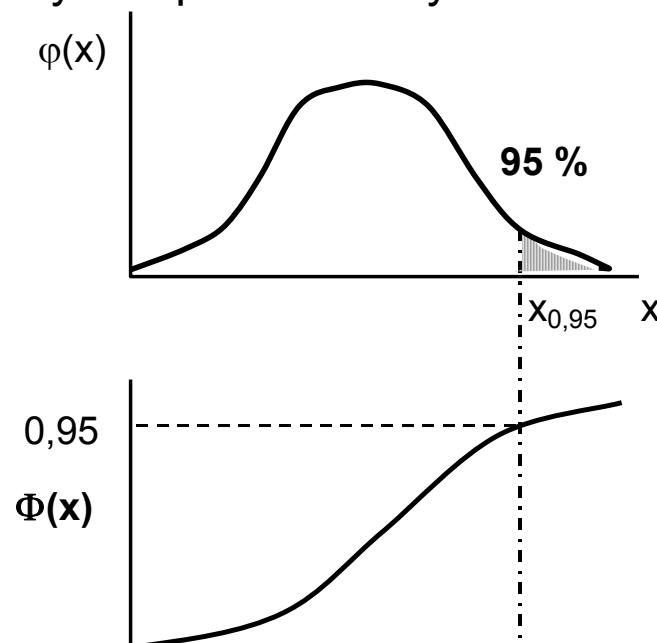


Další parametry rozložení

- ◆ **Počet hodnot** – důležitý ukazatel, znamená jak moc lze na data spoléhat
- ◆ **Střední chyba odhadu průměru** - je založena na směrodatné odchylce rozložení a **počtu hodnot**, vlastně jde o směrodatnou odchylku rozložení průměru. Říká jak přesný je náš výpočet průměru. Čím větší počet hodnot rozložení, tím je náš odhad skutečného průměru přesnější.
- ◆ **Suma hodnot**
- ◆ **Modus** – nejčastější hodnota, vhodný např. při kategoriálních datech
- ◆ **Minimum, maximum**
- ◆ **Rozsah hodnot**
- ◆ **Harmonický průměr** - převrácená hodnota průměru převrácených hodnot (vždy platí harmonický průměr < geometrický průměr < aritmetický průměr)

Distribuční funkce

- ◆ Definice kvantilu dle distribuční funkce - Kvantil rozložení ($X_{0,95}$) je číslo, jehož hodnota distribuční funkce je rovna pravděpodobnosti, pro kterou je kvantil definován ($\Phi(x)$... distribuční funkce), tj. pokud vezmeme nějaký bod rozložení a porovnáme jej s tímto bodem (kvantilem), máme 95% pravděpodobnost, že bude menší než hodnota kvantilu ($X_{0,95}$).
- ◆ Pomocí distribuční funkce můžeme určit jaký podíl hodnot rozložení je menší než daná hodnota – využití při statistických testech



Analýza dat na PC I.

Základní popisná statistika

Výběr proměnných

Tabulka četností hodnot

Box and whisker plot (následuje nastavení zobrazených parametrů)

Histogram

Základní výstup

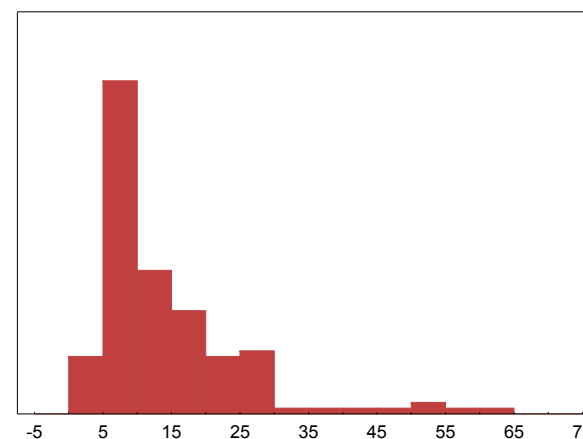
Výběr dat

Zpracování chybějících hodnot

The image shows a screenshot of the SPSS 'Descriptive Statistics: Spreadsheet 1' dialog box. The 'Variables:' field contains 'ALL'. The 'Quick' tab is selected, showing options for 'Summary: Descriptive statistics', 'Frequency tables', 'Histograms', and 'Box & whisker plot for all variables'. The 'Options' button is visible. The 'SELECT CASES' section shows 'W' (Weighted) selected. The 'Wghtd momnts' checkbox is unchecked. The 'DF' section shows 'W-1' selected. The 'MD deletion' section shows 'Pairwise' selected. Blue arrows point from text labels to specific elements in the dialog box: 'Výběr proměnných' points to the 'Variables:' field; 'Tabulka četností hodnot' points to the 'Frequency tables' button; 'Box and whisker plot (následuje nastavení zobrazených parametrů)' points to the 'Box & whisker plot for all variables' button; 'Histogram' points to the 'Histograms' button; 'Základní výstup' points to the 'Summary' button; 'Výběr dat' points to the 'SELECT CASES' section; and 'Zpracování chybějících hodnot' points to the 'MD deletion' section.

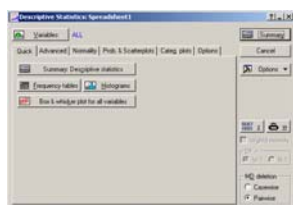
Popisné grafy I. Histogram a graf četnosti

- ◆ Tyto grafy se používají k zobrazení podílu výskytu hodnot v určitém intervalu proměnné. Oba grafy se liší způsobem zobrazení poměrů, zatímco sloupcový graf četností vynáší jako výšku sloupce přímo počet hodnot, u histogramu je důležitá plocha sloupce (počet hodnot zde odpovídá ploše a ne výšce sloupce), která vyjadřuje podíl objektů v daném intervalu, výška sloupce histogramu se získá jako podíl plochy (tj. počtu objektů) a šířky intervalu. V případě stejných šířek intervalů vypadají oba typy grafů stejně, liší se v případě nestejných intervalů (sloupce histogramu jsou u širších intervalů nižší – plocha sloupce odpovídá počtu objektů).
- ◆ Sloupce tedy odráží četnost objektů v daném intervalu, kterou vyjadřují buď svou výškou nebo plochou. Histogramy mohou existovat v několika formách 1) histogram relativních a absolutních četností a 2) histogram normální a kumulativní.

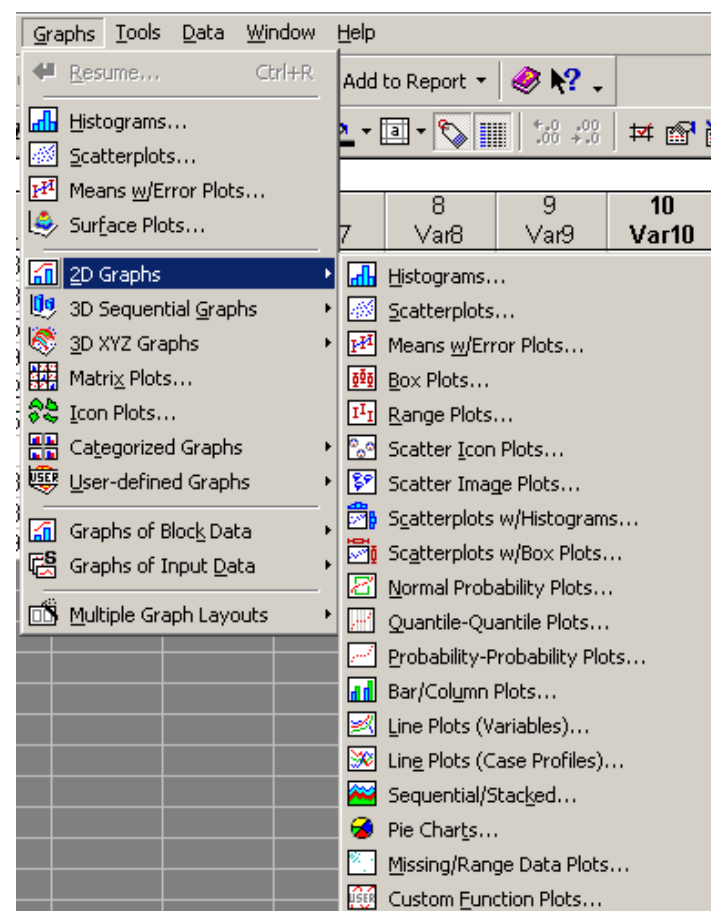
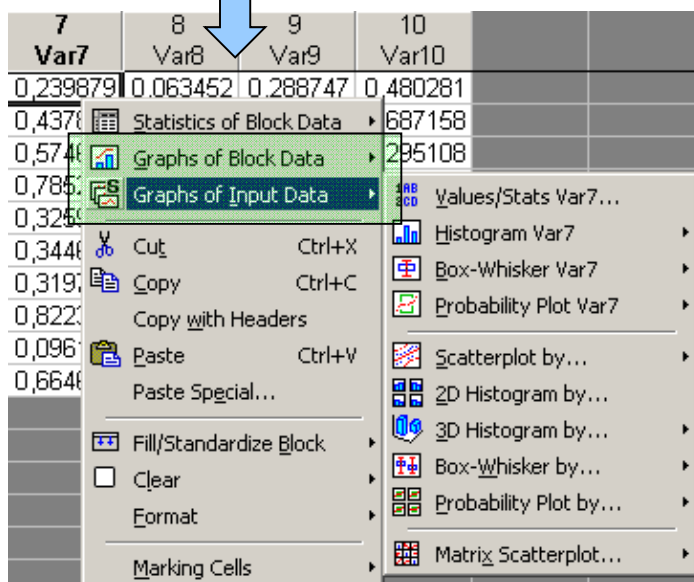


Analýza dat na PC I.

Tvorba grafů

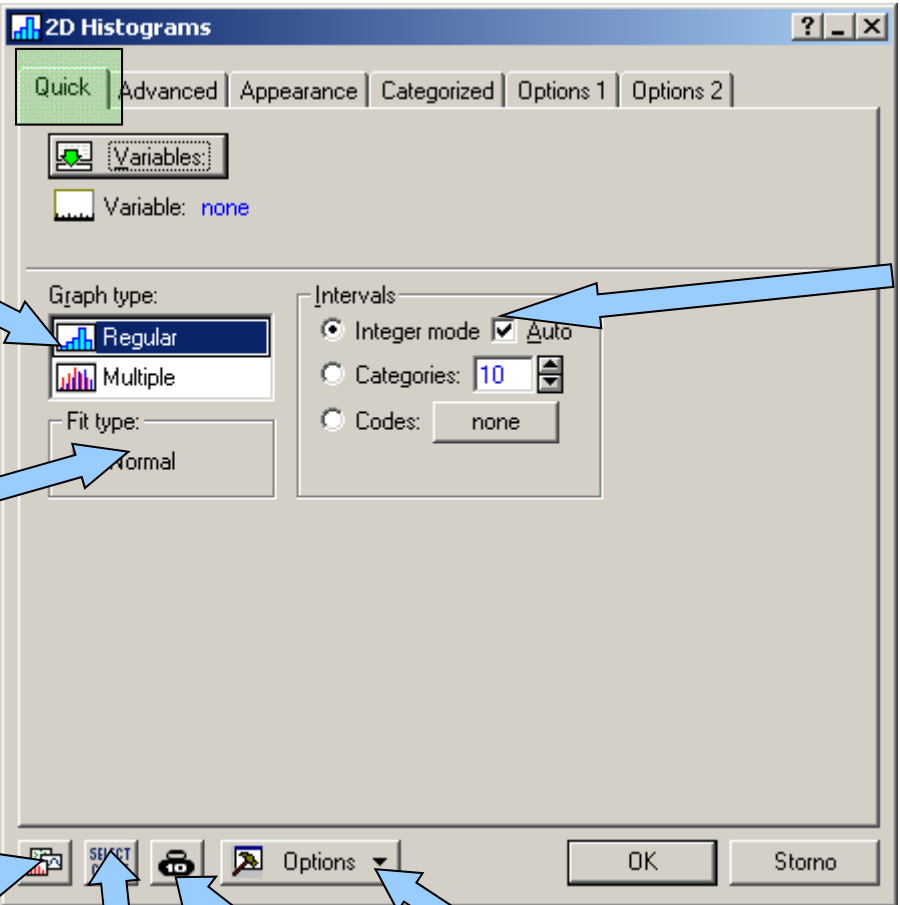


- ◆ Jako součást analýzy
- ◆ Lišta grafů
- ◆ Samostatné menu grafů
- ◆ Graphs of block and input data



Analýza dat na PC I.

Tvorba histogramu/grafu četnosti



The screenshot shows the '2D Histograms' dialog box in Minitab. The 'Quick' tab is selected. The 'Variables' section shows 'Variable: none'. The 'Graph type' section has 'Regular' selected. The 'Fit type' section has 'Normal' selected. The 'Intervals' section has 'Integer mode' selected with 'Auto' checked, 'Categories' set to 10, and 'Codes' set to none. The bottom toolbar contains icons for 'Galerie všech grafů', 'Výběr dat', 'Vážení dat', and 'Možnosti nastavení'. The 'OK' and 'Storno' buttons are at the bottom right.

Způsob zobrazení více proměnných

Proložení normálního rozložení

Nastavení intervalů grafu:

- Na základě celých čísel v datech
- Počet intervalů
- Podle kódů

Galerie všech grafů Výběr dat Vážení dat Možnosti nastavení

Analýza dat na PC I.

Pokročilá tvorba histogramu/grafu četnosti

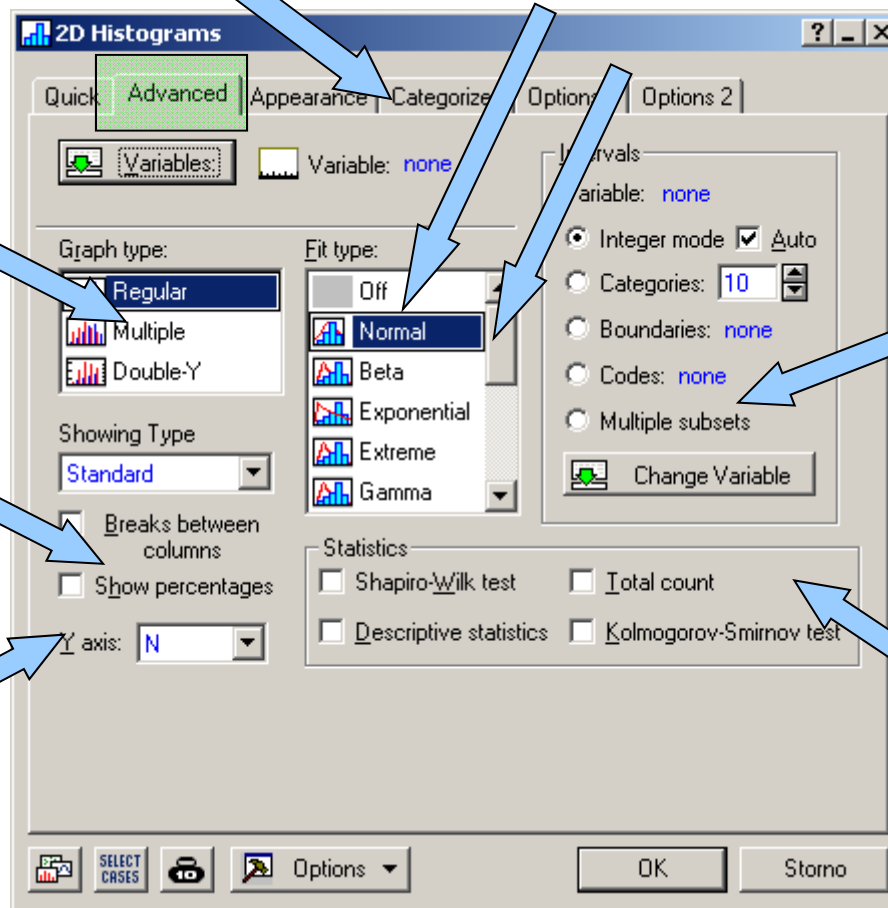
Kategorizace grafu

Proložení různých rozložení

Způsob zobrazení
více proměnných

Způsob zobrazení

Zobrazení hodnot na
ose Y



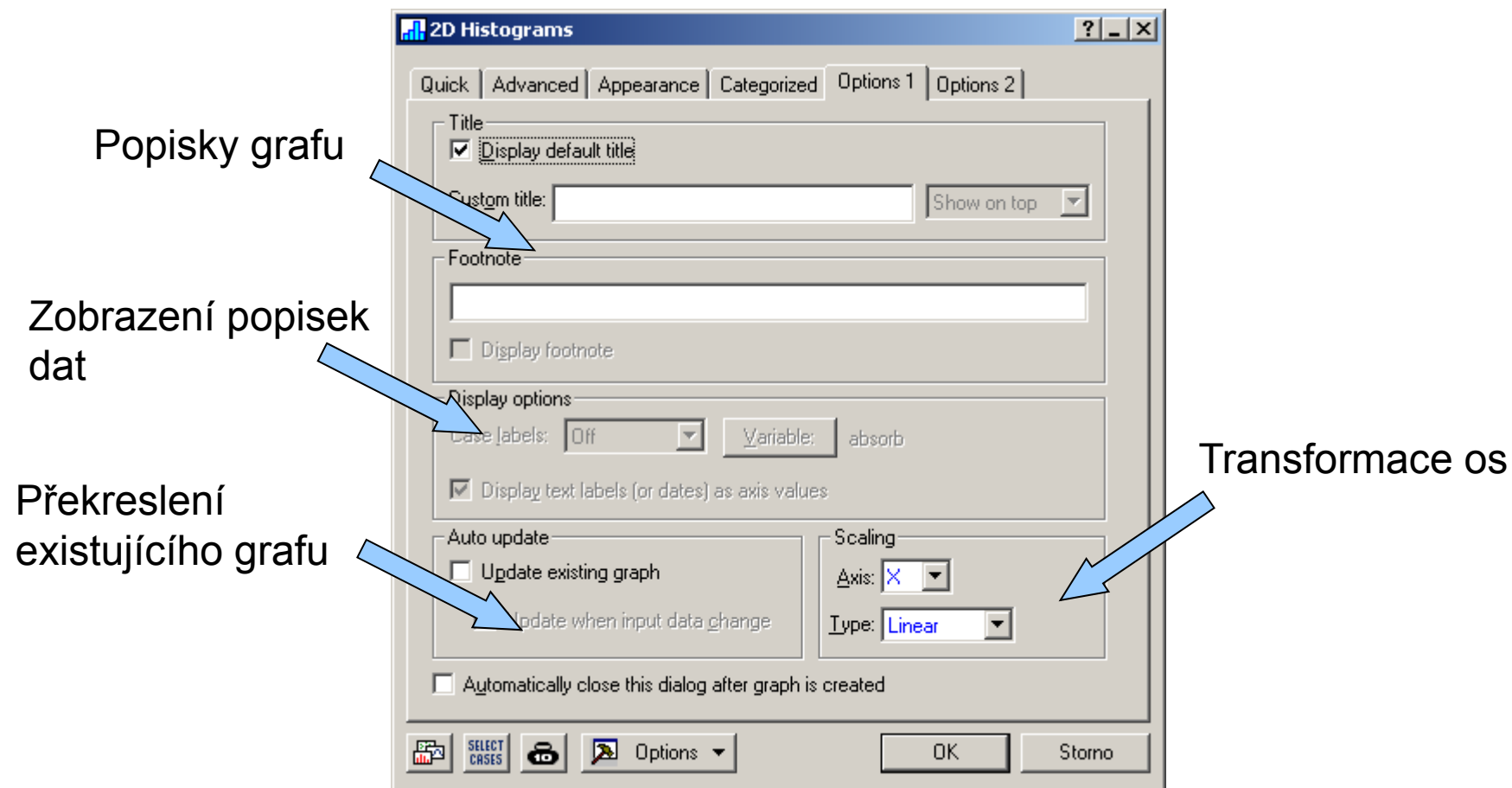
Kategorizace

- celá čísla v datech
- počet kategorií
- hranice
- kódy
- výběr dat

Testy normality a
popisná statistika

Analýza dat na PC I.

Nastavení společná různým typům grafů I



Analýza dat na PC I.

Nastavení společná různým typům grafů II

Normální (karteziánský)
nebo polární systém

Pozice os

Zobrazení regresní
funkce, nastavení
polynomu pro proložení,
základ logaritmu

Zobrazit postup výpočtu
Výběr z dat
Přidání upraveného grafu do
menu

2D Histograms

Quick Advanced Appearance Categorized Options 1 Options 2

Layout

Coordinate system: Standard (Cartesian)

X-Y Axis position: Standard

Fitted function

Display fit expression: In title

Polynomial order: Quadratic

Logarithmic base: 10

Display progress bar: Enable

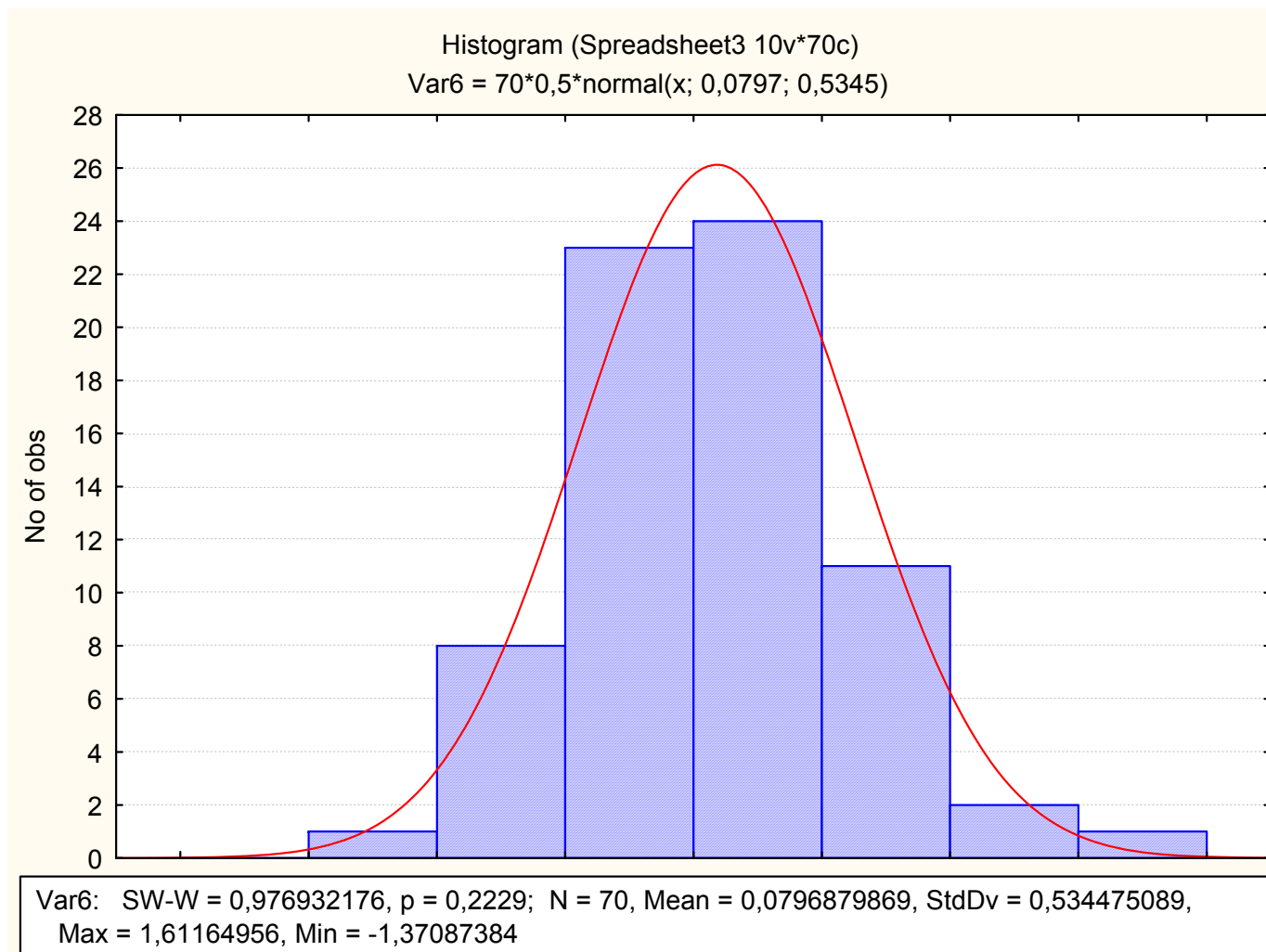
☐ Random subsampling; approximate subset: 1

Add As User-defined Graph to Menu...

OK Storno

Analýza dat na PC I.

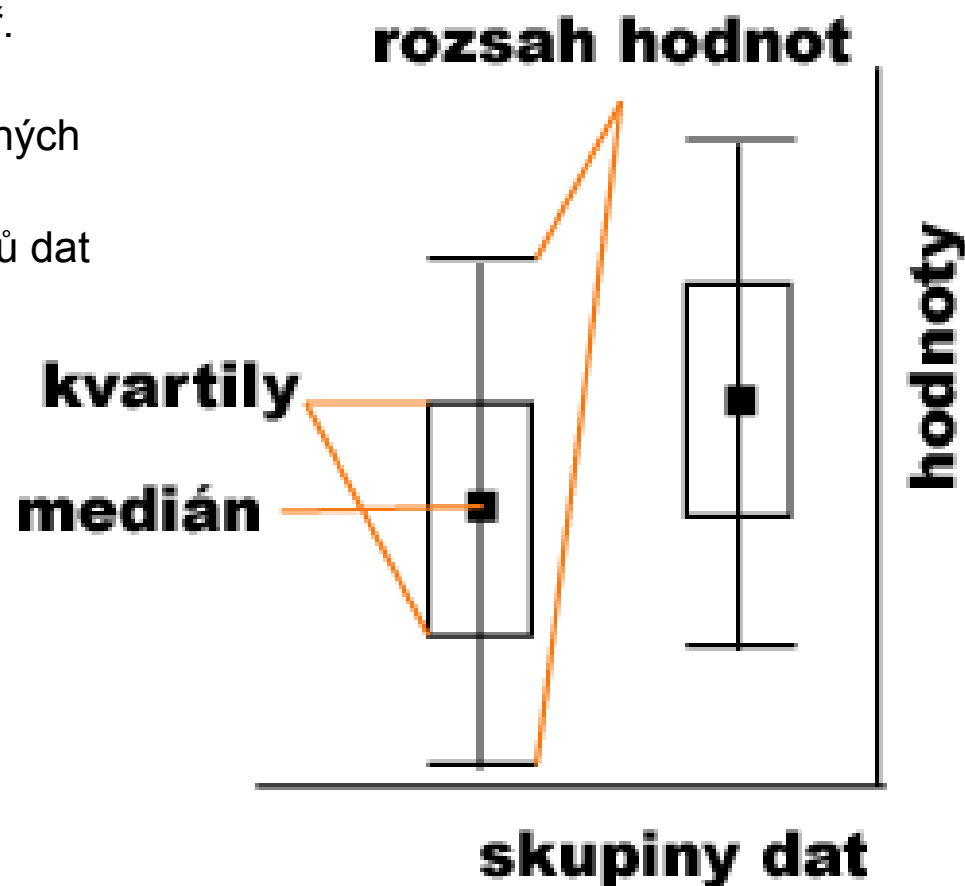
Ne - Histogram



Analýza dat na PC I.

Box & whisker plot

- ◆ Typ grafu vynášející několik významných bodů rozložení, např. medián, kvartily a rozsah hodnot
- ◆ Poskytuje grafický přehled popisných statistik
- ◆ Rychlé srovnání několika souborů dat
- ◆ Umožňuje orientačně posoudit normalitu dat



Analýza dat na PC I.

Box and whisker plot

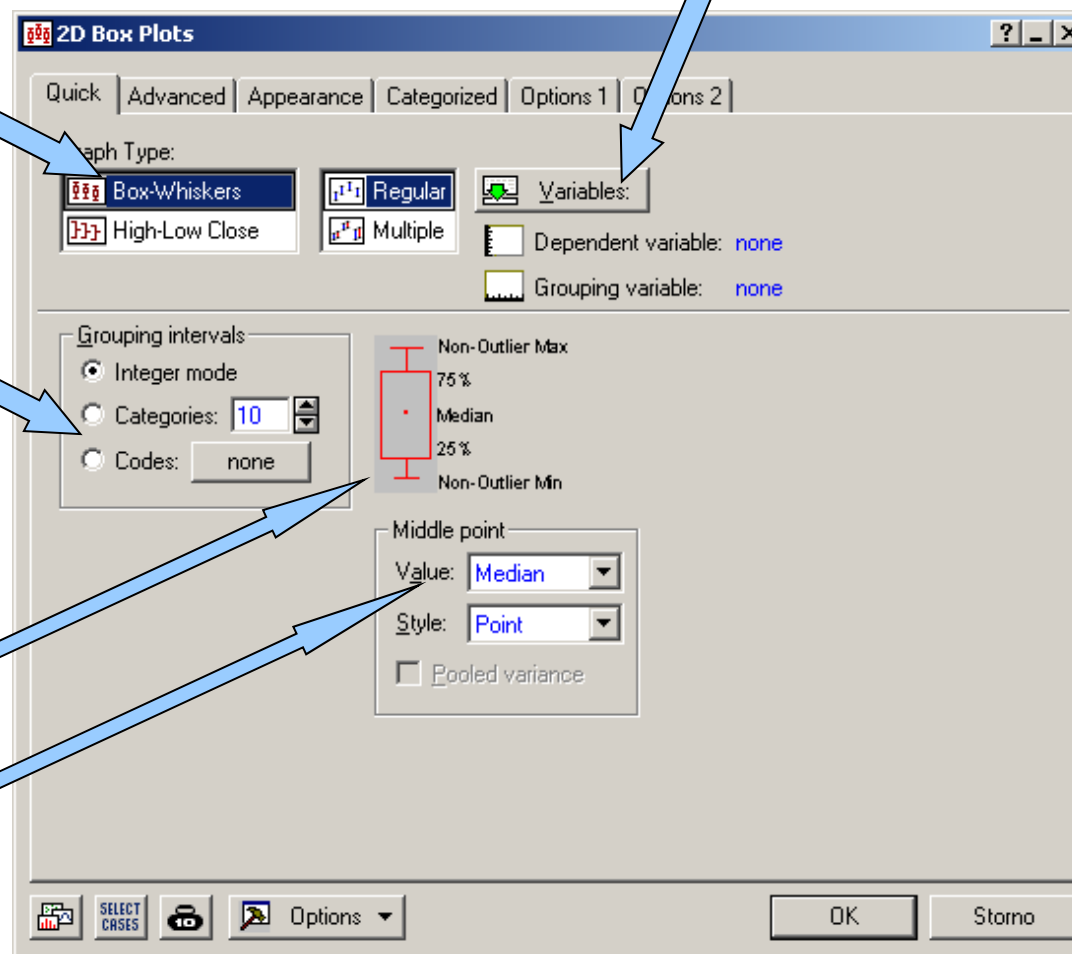
Datová a kategorizační proměnná

Způsob zobrazení box and whisker plotu

Kategorizace hodnot do jednotlivých grafů

Preview grafu

Ukazatel středu



Analýza dat na PC I.

Box & whisker plot II

Datová a kategorizační proměnná

Typ grafu

Kategorizace hodnot
do jednotlivých grafů

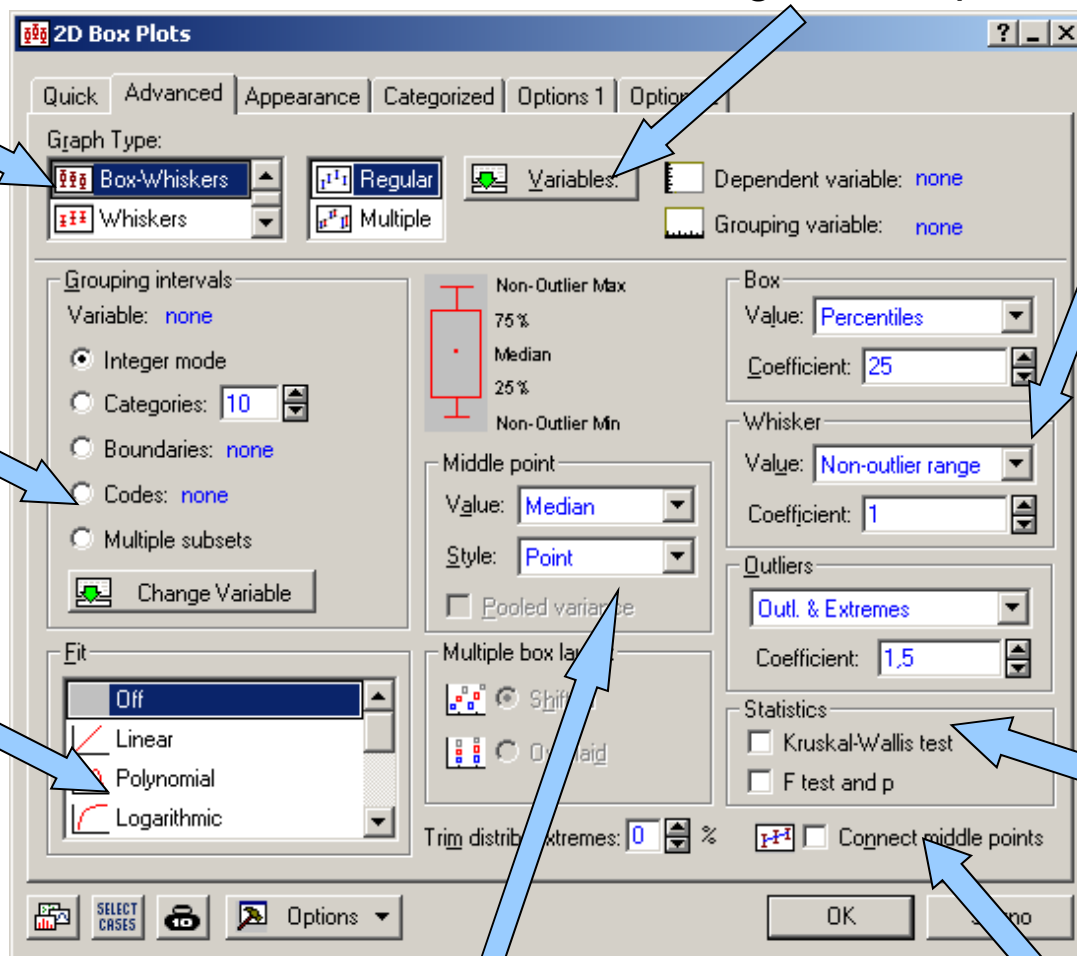
Proložení křivky

Které
statistiky
budou
zobrazeny

Statistické
testy

Středová hodnota

Spojení středů



Analýza dat na PC I.

Detailní popisná statistika

Percentily, rozsahy

Ukazatele
středu rozložení
a počet hodnot

Tvar rozložení (šířka,
asymetrie atd.)

Uložení nastavení

Zrušení výběru statistik

Vybrání všech statistik

Analýza dat na PC I.

Normalita dat

Histogram

Frekvenční tabulky

Nastavení histogramu

Testy normality

3D histogram

Kategorizovaný histogram

Srovnání rozložení

Steam and leaf plot

Descriptive Statistics: Spreadsheet1

Variables: none

Quick Advanced **Normality** Prob. & Plots Categ. plots Options

Distribution

☐ Frequency tables ☒ Histograms

Categorization

Number of intervals: 10

☐ Integer intervals (categories)

☐ Normal expected frequencies

☒ Kolmogorov-Smirnov & Lilliefors test for normality

☐ Shapiro-Wilk's W test

Use Nonparametrics, Process Analysis, or Graphs (P-P or Q-Q) to fit other distributions; use Survival Analysis to fit distributions to censored data.

Stem and leaf

☒ Stem & leaf plot

☐ Compressed

Options

SELECT CASES ☐ ☒ ☐ W

Wghtd momnts

DF = ☒ W-1 ☐ N-1

MD deletion

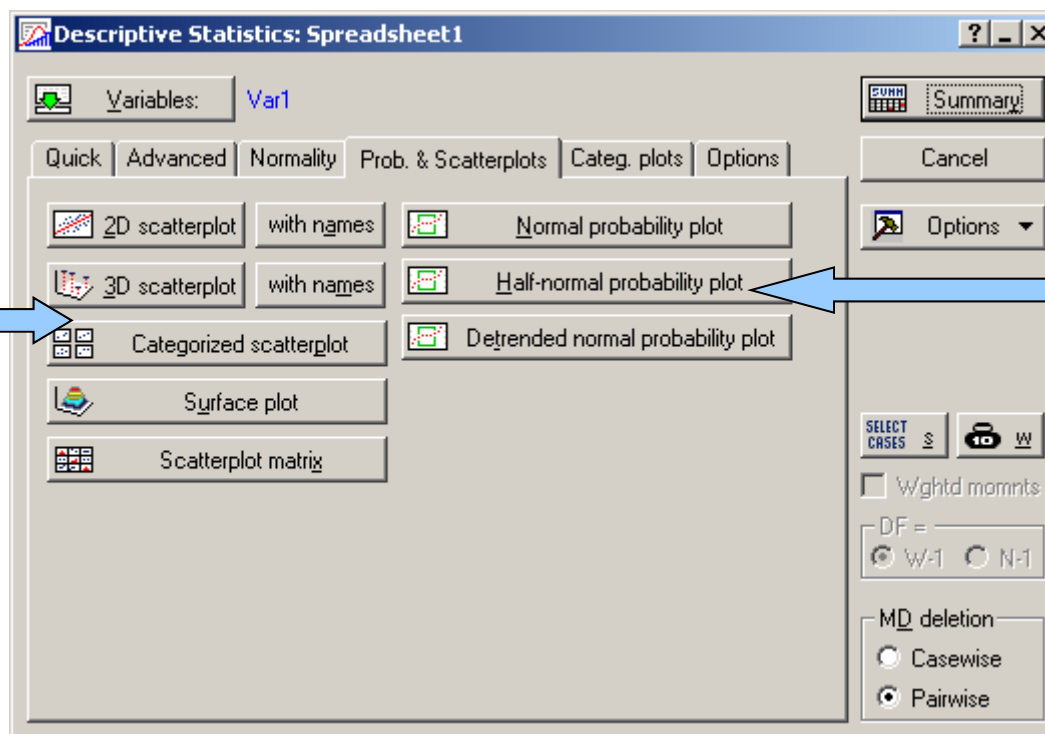
☐ Casewise

☒ Pairwise

Analýza dat na PC I.

Popisné grafy

Grafy
vynášející proti
sobě různým
způsobem
proměnné



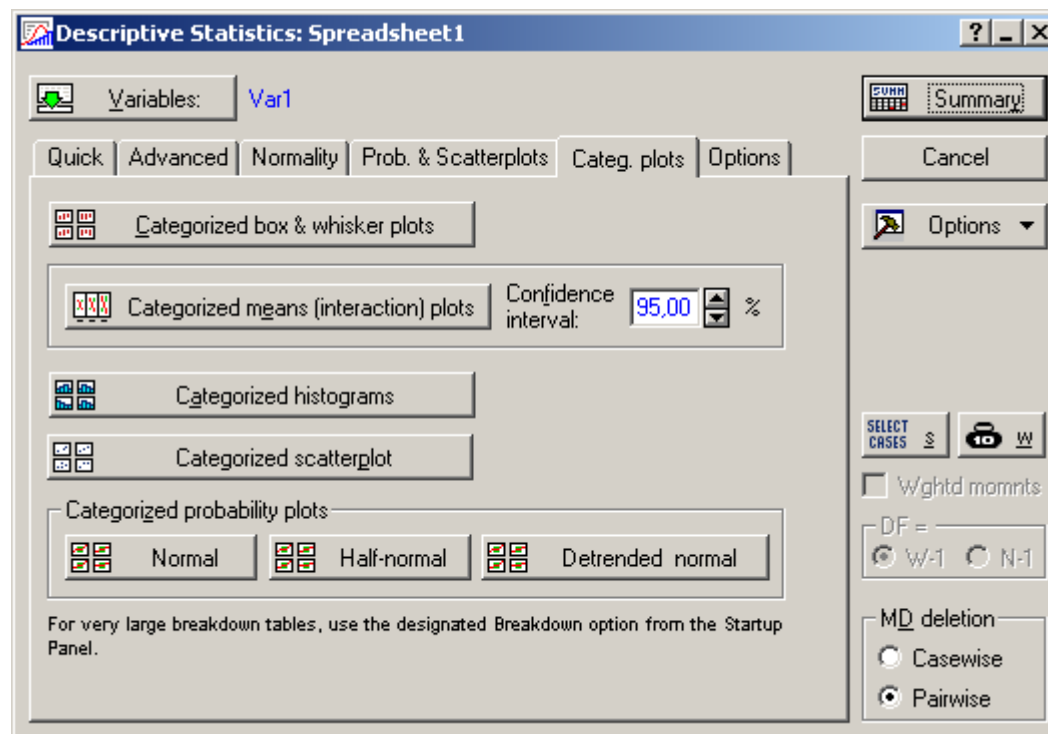
Grafy normality

Analýza dat na PC I.

Kategorizované grafy

Kategorizované grafy

proměnné jsou rozloženy na skupiny dané kategorizační proměnnou (např. proměnná obsahující výšku postavy může být rozdělena podle pohlaví jinou proměnnou obsahující informaci o pohlaví jednotlivých osob (řádků první proměnné))



Analýza dat na PC I.

Nastavení popisné statistiky

