# Analýza dat v softwaru STATISTICA

Software STATISTICA je produkt StatSoft, Inc. (www.statsoft.com, www.statsoft.cz). STATISTICA je dostupná v rámci MU z inet.muni.cz (login stejný jako do www.is.muni.cz, seznam dostupných softwarů lze najít v oddílu Provozní služby).

### Načtení datového souboru

File -> Open -> *vybrat datový soubor* -> Open -> Import selected sheet to a Spreadsheet -> *vybrat list Excelovského souboru* -> OK -> *nechat zatržené* Get variable names from first row -> OK -> Import as Text Labels

### Uložení datového souboru

File -> Save -> *zadáme název souboru* -> Save

### Zapnutí automatického filtru

*Označit všechny sloupce (např. pomocí CTRL+A nebo kliknutím do levého horního rohu tabulky)* -> Data -> Auto Filter -> Auto Filter

## 1. Vizualizace dat

Vytváření grafů pomocí záložky Graphs.

### Koláčový graf

Graphs -> 2D Graphs -> Pie Charts -> *zvolit proměnnou (např.* Gender*) (v záložce* Advanced *je možné zvolit, jakou legendu, typ a tvar grafu chceme* (Pie Legend, Type, Shape) -> OK
*Po dvojím kliknutí na graf se nám ukáže okno* Graph Options, *kde lze libovolně měnit barvu grafu i typ a tvar grafu a další parametry*

### Sloupcový graf (na ose y počty lidí)

Graphs -> Histograms -> Variables -> *zvolit proměnnou (např.* Group) -> OK -> *zrušit zatržení* Fit type: Normal -> *na záložce* Advanced *zatrhnout* Breaks between columns -> OK

### Sloupcový graf (na ose y procenta)

Graphs -> Histograms -> Variables -> *zvolit proměnnou (např.* Group) -> OK -> *zrušit zatržení* Fit type: Normal -> *na záložce* Advanced *zatrhnout* Breaks between columns -> *na záložce* Advanced *změnit u* Y axis *hodnotu* N *na* % -> OK

### Histogram (na ose y procenta)

Graphs -> Histograms -> Variables -> *zvolit proměnnou (např.* Age) -> OK -> *na záložce* Advanced *změnit u* Y axis *hodnotu* N *na* % (*lze např. si vypsat i základní popisnou statistiku zatrhnutím* Descriptive statistics) -> OK

### Krabicový graf (s vykreslením odlehlých hodnot)

Graphs -> 2D Graphs -> Box Plots... -> Variables -> *zvolit proměnnou (např.* Age) *jako* Dependent variable -> OK -> OK

### Krabicový graf (s minimem a maximem)

Graphs -> 2D Graphs -> Box Plots... -> Variables -> *zvolit proměnnou (např.* Age) *jako* Dependent variable -> OK -> *na záložce* Advanced -> *u* Whisker *zvolit* Min-Max -> *u* Outliers *zvolit* Off -> OK

# 2. Příprava dat pro analýzu

### Nastavení formátu u MMSE na double

*Dvakrát kliknout na šedé políčko s názvem proměnné* -> *nastavit* Type *na* Double -> *nastavit* Display format *na* Number -> OK

### Nastavení formátu u scan_date na datum

*Dvakrát kliknout na šedé políčko s názvem proměnné* -> *nastavit* Display format *na* Date -> *vybrat formát* 17/03/10 -> OK

### Identifikace a odstranění duplikací

Data -> Data Filtering/Recoding -> Filter Duplicate Cases -> Input: Variables -> ID -> OK -> *u* Output *zatrhnout* Create duplicates spreadsheet -> OK
*Je patrné, že se vždy zachová první záznam a druhý záznam je vyřazen bez ohledu na datum pořízení skenu. Pokud chceme, aby byl vždy odstraněn záznam se starším datem, je nejprve nutné data seřadit podle data pořízení skenu (sestupně) pomocí:* Data -> Sort -> *označit* 1-ID -> Add Var(s) -> *označit* 30-scan_date -> Add Vars(s) -> Descending -> OK -> Include Formatting
*Nový datový soubor bez duplikací uložit.*

### Odstranění chybějících a chybných hodnot

Data -> Subset -> Cases -> *zatrhnout* Enable Selection Condition-> *do* By expression *napsat* v4="" OR v4>110 OR v7="" -> OK -> OK
*Nový datový soubor bez chybějících a chybných hodnot uložit.*

### Rekódování proměnné Gender, aby obsahovala pouze hodnoty F a M

1. způsob – ručně: *Vyfiltrovat si řádek s hodnotou* FF *a hodnotu* FF *přepsat na* F
2. způsob – vytvořením nové proměnné: *Označit proměnnou za proměnnou* Gender -> Insert -> Add Variables -> Name -> *zadat název nové proměnné (např.* Gender_rek*)* -> *do* Long name *napsat* =iif(v3="FF","F",v3) -> OK

### Rekódování proměnné Group, aby obsahovala pouze hodnoty 1 (CN), 2 (MCI) a 3 (AD)

*Označit proměnnou za proměnnou* Group -> Insert -> Add Variables -> Name -> *zadat název nové proměnné (např.* Group_3kat*)* -> *do* Long name *napsat* =iif(v2=3;2;iif(v2=4;3;v2)) -> OK
*Jiný způsob pomocí* Data -> Recode...

### Vytvoření textových popisků u kvalitativní proměnné

*Dvakrát kliknout na šedé políčko s názvem proměnné* -> Text Labels...-> *zadat textové popisky a jejich příslušné číselné hodnoty* -> OK -> OK

# 3. Popisná sumarizace dat

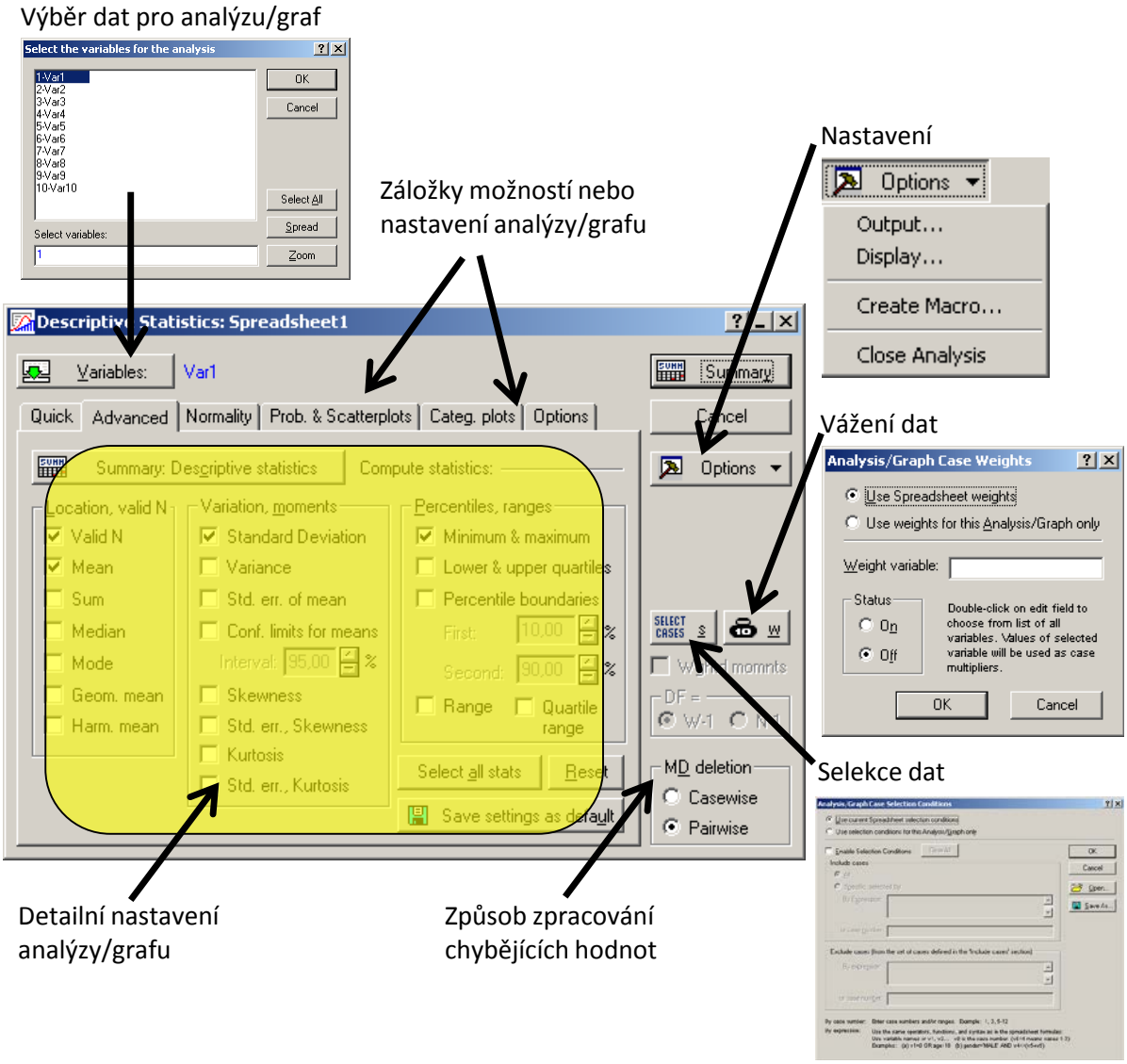
Popisná sumarizace dat pomocí Statistics -> Basic Statistics/Tables. Obecný popis dialogového okna pro sumarizaci dat, vizualizace a další analýzy je uveden na Obr. 1.

### Popisná sumarizace kvalitativních dat – frekvenční tabulka

Statistics -> Basic Statistics/Tables -> Frequency tables -> Variables -> *zvolit proměnnou (např.* Group) -> Summary

### Popisná sumarizace kvantitativních dat

Statistics -> Basic Statistics/Tables -> Descriptive statistics -> Variables -> *zvolit proměnnou (např.* Age) -> OK -> *na záložce* Advanced *zatrhnout* Median, Coefficient of variation, Lower & upper quartiles -> Summary




Obr. 1. Popis dialogového okna sloužícího pro sumarizaci, vizualizaci a další analýzy dat.

**Popisná sumarizace kvantitativních dat – zapnutí filtru (vyfiltrování posledních 20 pacientů)**

Statistics -> Basic Statistics/Tables -> Descriptive statistics -> Variables -> *zvolit proměnnou (např.* Height *a* Height_cor) -> OK -> Select Cases -> Enable Selection Conditions -> Specific, selected by: -> *do* "or case number" *zadat* 814-833 -> OK -> *na záložce* Advanced *zatrhnout* Median -> Summary

# 4. Transformace dat

### Logaritmická transformace

*Označit proměnnou za proměnnou, kterou chceme logaritmovat* -> Insert -> Add Variables -> Name -> *zadat název nové proměnné* (*např.* Weight_log) -> *do* Long name *napsat* =Log(v9) (*Pozor, v softwaru STATISTICA je přirozený logaritmus označen jako Log*(*x*) *místo Ln*(*x*)*!*) -> OK

### Standardizace dat

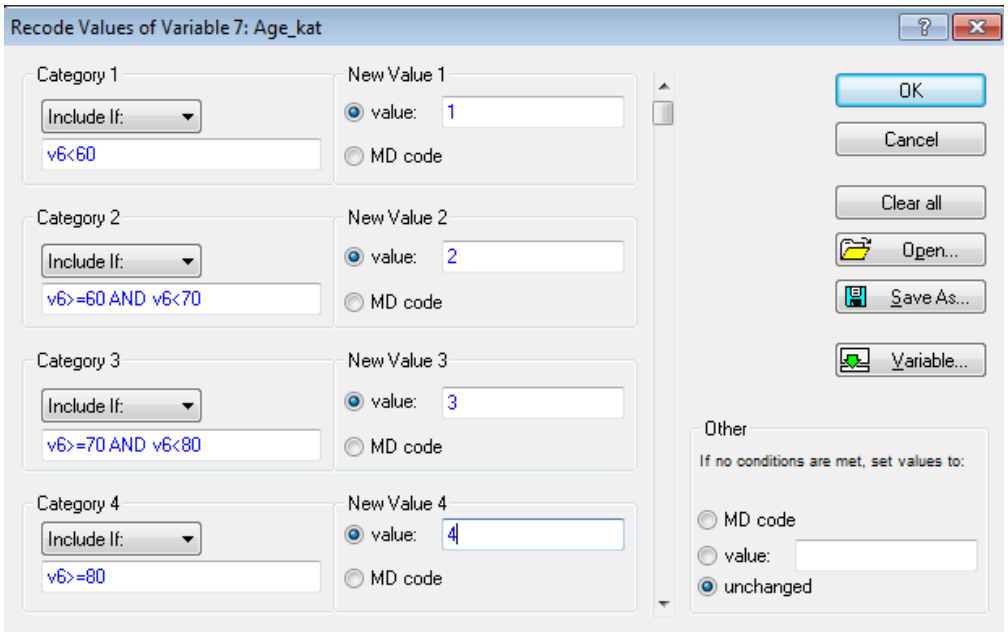
*Označit proměnnou za proměnnou, kterou chceme standardizovat* -> Insert -> Add Variables -> Name -> *zadat název nové proměnné* (*např.* Age_st) -> *do* Long name *napsat* =v6 -> OK -> Data -> Standardize... -> OK

### Centrování dat

*Označit proměnnou za proměnnou, kterou chceme centrovat*-> Insert -> Add Variables -> Name -> *zadat název nové proměnné* (*např.* Height_centr) -> *do* Long name *napsat* =v9-174.15 (*průměr vypočítaný pomocí* Descriptive statistics) -> OK

### Kategorizace

*Označit proměnnou za proměnnou, kterou chceme kategorizovat* -> Insert -> Add Variables -> Name -> *zadat název nové proměnné* (*např.* Age_kat) -> OK -> Data -> Recode... (*zkontrolovat si, že v záhlaví je správný název proměnné, jinak vybrat správnou proměnnou pomocí tlačítka* Variable...) -> *zadat podmínky a nové hodnoty* (*viz* Obr. 2) -> OK




Obr. 2. Ukázka kategorizace věku.

## 5. Intervaly spolehlivosti

### Výpočet intervalu spolehlivosti a střední chyby průměru (standard error)

Statistics -> Basic Statistics/Tables -> Descriptive statistics -> Variables -> *zvolit proměnnou (např.* Age) -> OK -> *na záložce* Advanced *zatrhnout* Conf. limits for means *a* Std. err. of mean -> Summary

### Výpočet kvantilů Studentova rozložení

Statistics -> Probability Calculator -> Distributions... -> t (Student) -> *zatrhnout* Inverse -> *jako p zadat* 0.975 -> *jako df (degrees of freedom – počet stupňů volnosti) zadat* 832 -> Compute *(vypočítá nám to hodnotu t)*

# 6. Další užitečná nastavení

### Vypnutí automatického překreslování grafů

File -> Output Manager -> Graphs -> Settings -> Data Update *přepnout na* Locked -> *zrušit zatržení u* Update spreadsheet case states -> OK