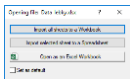
Statistica

Mgr. Mikoláš Jurda, Ph.D.

Ovládání programu

Otevření/import dat

Data je možné importovat ze souborů různých typů – excel, csv, txt a také vložením ze schránky – je možné je také editovat



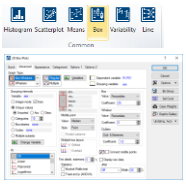
Práci ušetří, pokud jsou buňky v původním excelovském souboru ve správném formátu. Pokud po načtení formát proměnných neodpovídá našim požadavkům, jde formát upravit nastavením jednotlivých proměnných (dvojité poklikání na buňku s názvem proměnné) – typicky nastavení grupovacích proměnných na *Type > Double*

Datový soubor a výstupy analýz je možné uchovávat v různých formátech. Možnost uložit vše v přehledném stromu nabízí *Workbook*. Výstupy je možné také ukládat do wordu

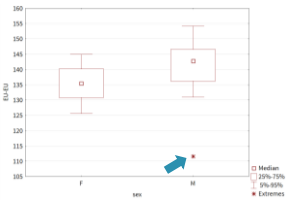
Popisná statistika

Vizuální posouzení dat

Krabicový graf



Pokud nezadáte grupovací proměnnou, zobrazí se graf pro celý soubor, pokud ano, pak

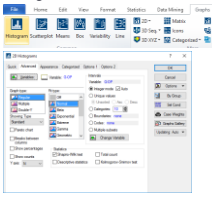


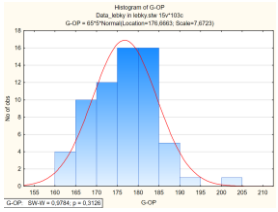
Krabicový graf pro dvě skupiny – m a f – dobrý pro vyhledávání **extrémních případů (například chyb v datech)** – při podržení myši nad odlehlou hodnotou se zobrazí její ID

Popisná statistika

Vizuální posouzení dat

Histogram





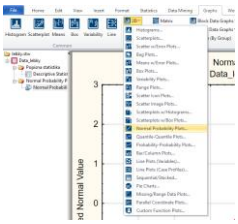
Umožňuje posoudit rozložení hodnot a srovnat je s předpokládaným rozložením (linie). Nastavuje se jako *Fit type* v předchozím grafu.

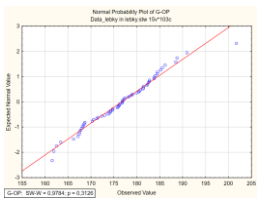
Popisná statistika

Vizuální posouzení dat

Normální pravděpodobnostní graf

porovnává kvantily spočtené z dat s kvantily standardizovaného normálního rozdělení



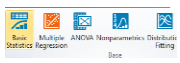


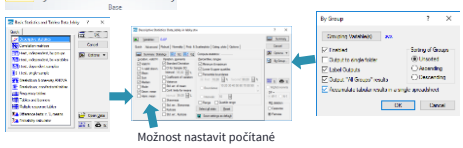
Dovoluje rovněž identifikovat extrémní hodnoty

Popisná statistika

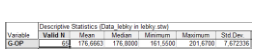
Číselná popisná statistika

Číselná forma



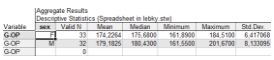


Možnost nastavit počítané parametry



Descriptive Statistics (Data_lebky in lebky.ste)						
Variable	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
G-OP	65	176.6663	176.8000	161.5500	201.6700	7.672336

Pro všechna data zároveň



Aggregate Results Descriptive Statistics (Spreadsheet in lebky.ste)						
Variable	sex	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
G-OP	F	33	174.2264	175.6000	161.8900	184.5100	6.417968
G-OP	M	32	179.1825	180.4300	161.5500	201.6700	8.133095
G-OP		0					

By group – pro skupiny zvlášť

Popisná statistika

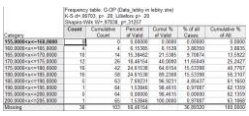
Souhrnné výsledky – histogram, krabicový graf, zvolené parametry a P-P plot

V základní, přednastavené podobě

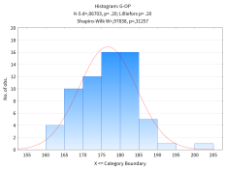
Normalita dat

Grafické posouzení
srovnání s normálním rozložením (viz předchozí grafy)

Testování
Shapiro-Wilksův test
Statistics > Basic statistics > Descriptive statistics > Normality



Záhlaví *Frequency table*



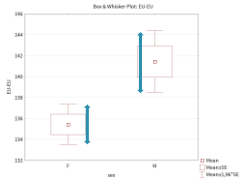
Záhlaví *Histograms*

T-test

Nepárový dvouvýběrový t-test

Předpoklady:
Normální rozložení v rámci porovnávaných skupin
- již představenými postupy

Shoda rozptylu
- testování je přímo součástí výsledků



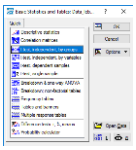
Pokud data nesplňují

Neparametrické alternativy
např. Mann-whitney U-test

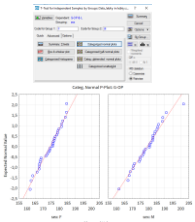
V případě různých rozptylů lze použít t-test se samostatnými odhady rozptylů

T-test

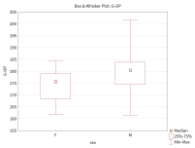
Ověření předpokladů testu přímo v jeho dialogovém okně



Normalita rozložení ve skupinách
Advanced > Categorized normal plots



Shoda rozptylů - graficky
Advanced > Box & Whiskers plot



T-test

Samotné výstupy testu – lze provést hromadně pro všechny zároveň

Diskriminační analýza

Závislost jedné kvalitativní proměnné na několika kvantitativních proměnných

Pro
- určení proměnných, které diskriminují mezi dvěma nebo více skupinami
- ke klasifikaci objektů do různých skupin

Předpoklady
- mnohorozměrné normální rozdělení (**především citlivé na odlehlé hodnoty**)
- shoda skupinových kovariančních matic
- **proměnné nejsou redundantní**

Diskriminační analýza

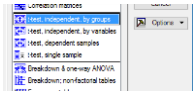
Jaké použít proměnné?
Význam mají **pouze ty proměnné**, které mají souvislost s kategoriální proměnnou

Redundantní proměnné snižují stabilitu modelu a mohou vést k nesmyslným výsledkům

Hodnocení vztahu nezávislých proměnných a kategoriální proměnné
Korelační analýza a XY grafy
Hlavní komponenty a faktorová analýza
Diskriminační analýza
„Expertní" znalost proměnných – pokud jsou redundantní, můžeme vyřadit ty proměnné zatížené chybami nebo vysokým počtem chybějících hodnot

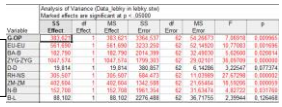
Diskriminační analýza

Hledání proměnných
Vztah ke kategoriální proměnné
- samostatný t-test pro jednotlivé proměnné – pro dvě skupiny vždy (*Basic statistics > t-test, independent, by groups*)



- ANOVA – pro dvě a více skupin (*Basic statistics > Breakdown & One-way ANOVA; Analysis of Variance*) – výsledky jsou ekvivalentní

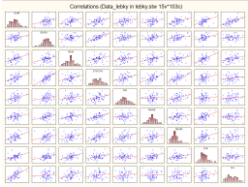
Může napovědět, ale diskriminace může být dána pouze kombinací proměnných



Diskriminační analýza

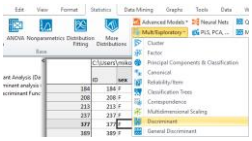
Hledání proměnných
Korelační analýza – vztah proměnných mezi sebou
(*Statistics > Basic statistics and tables > Correlation matrices*; *Summary: Correlations*)

XY grafy
(*Statistics > Basic statistics and tables > Correlation matrices*; *Scatterplot matrix for selected correlations*)

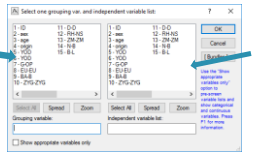


Diskriminační analýza

Diskriminanční analýza
(*Statistics > Mult/Exploratory > Discriminant*)



Variables



Grupovací proměnná – kategorie, do kterých bychom případně chtěli klasifikovat

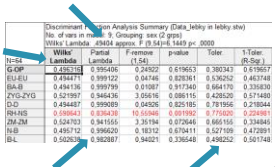
Nezávislé proměnné

Diskriminační analýza

Číselný výstup analýzy

Celkové Wilks lambda – popisuje celkovou kvalitu modelu všech proměnných (0 = nejlepší diskriminace)

Wilks lambda celého modelu při vyřazení dané proměnné



Unikátní příspěvek proměnné k diskriminaci

Variabilita proměnné nevysvětlená ostatními proměnnými

Var. proměnné vysvětlená kombinací ostatních proměnných v modelu

Diskriminační analýza

Číselný výstup analýzy

Klasifikační funkce
Sada rovnic – objekt je zařazen do té skupiny, jejíž klasifikační funkce nabývá nejvyšší hodnoty

F = 0,666 . G-OP + 1,128 . EU-EU + 1,715 . BA-B + ... + (-573,608)

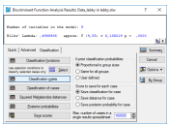
M = 0,621 . G-OP + 1,124 . EU-EU + 1,715 . BA-B + ... + (-623,991)

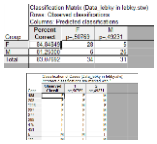
Jindy jako jedna rovnice, jejíž výsledek se porovnává se dělicím bodem

Diskriminační analýza

Hodnocení úspěšnosti klasifikačního kritéria
Klasifikační tabulka – procentuální vyjádření úspěšnosti zařazení objektů do skupin

Resubstituce – klasifikační rovnici testujeme na stejném souboru, na kterém byla vytvořena





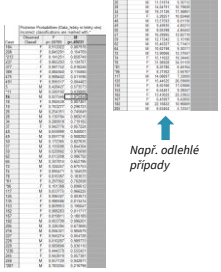
Křížové ověření (*leave-one-out cross-validation*) – vybereme n-1 objektů, z nich vytvoříme kritérium a to pak aplikujeme na vypuštěný případ. Postup opakujeme se všemi dalšími případy

Aplikace na nezávislý vzorek – kritérium vytvoříme například pouze na části případů a ověříme na tom zbytku

Diskriminační analýza

Co může dále napovědět?
Mahalanobisova vzdálenost
- popisuje vzdálenost centroidů skupin (bere v úvahu korelaci mezi parametry a je nezávislá na rozsahu parametrů)

Posterior probability – pravděpodobnost zařazení objektu do skupiny (p toho, že objekt patří do té které skupiny) - vychází z Mahalanobisových vzdáleností ke skupinám a *a priori* pravděpodobnosti

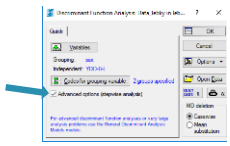


Např. odlehlé případy

Diskriminační analýza

„Step-wise" analýza (dopředná a zpětná eliminace) – výběr proměnných samotnou analýzou

- proměnné jsou přidávány/ubírány, podle jejich významu v modelu
- zpravidla je vybrán pouze zlomek původních proměnných



Discriminant Function Analysis Summary (Data_lebky in lebky.stw)
(Step 3, N of vars in model: 3; Grouping: sex (2 grps)
Wilks' Lambda: .51168 approx. F (3,61)=19.405 p<.0000

N=65	Wilks' Lambda	Partial Lambda	F-remove (1,61)	p-value	Toler.	1-Toler. (R-Sqr.)
ZYG-ZYG	0.606466	0.843794	11.38024	0.001342	0.884970	0.115030
RH-NS	0.610534	0.838083	11.78516	0.001078	0.933334	0.066666
ZM-ZM	0.530334	0.964823	2.22404	0.141031	0.919489	0.080511

V našem případě (dopředná analýza) vybrány pouze tři proměnné

Kontingenční tabulky

Test dobré shody (Pearsonův chí-kvadrát test)

Testuje shodu reálné distribuce hodnot do n skupin s teoretickou distribucí.

V případě platnosti nulové hypotézy **je poměr mezi buňkami jednoho řádku v různých sloupcích nezávislý na výběru tohoto řádku**

je A nezávislé na B a naopak

↓ B \ A	+	-	Σ
+	a	b	
-	c	d	
Σ			**suma sum**

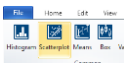
Statistics > Basic statistics > Tables and banners .... > Options > Expected frequencies

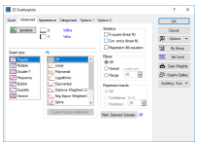
zobrazení výsledků
Advanced > Detailed Two-way Tables

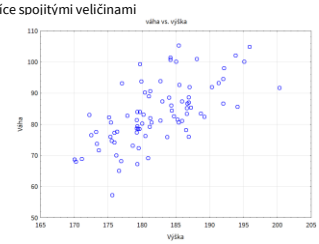
Korelační analýza

Hodnocení vztahu mezi dvěma a více spojitými veličinami

Bodový graf



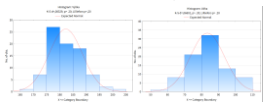




Vztah existuje, ale vztah není přesně lineární

Korelační analýza

Korelační koeficienty
Předpokladem parametrických je normalita rozložení



Vyhovuje?

Pearsonův korelační koeficient

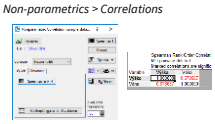
$$R(X,Y) = \frac{E((X - EX)(Y - EY))}{\sqrt{DX}\sqrt{DY}}$$

Basic statistics > Correlation matrices

Nevyhovuje?

Spearmanův korelační koeficient

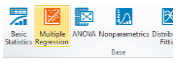
Non-parametrics > Correlations

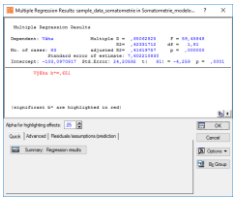


Regresní analýza

Vysvětluje vztah dvou a více proměnných. Jak **vysvětlovaná proměnná** závisí na jiných **proměnných (prediktorech). Model musí odpovídat typu vztahu** – pokud je přímkový, můžeme použít lineární model.

$$Y = b_0 + b_1X + E$$





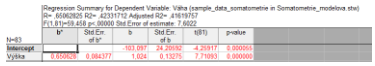
Dependent – závislá (vysvětlovaná proměnná
Multiple R – koeficient vícerozměrné korelace
R2 – koeficient determinace – podíl vysvětlované variability
Adjusted R2 – podobný, ale bere v úvahu počet regresorů
F, df a p – F test vztahů mezi závislou proměnnou a množinou neávislých proměnných – F=regresní průměr čtverců/reziduální průměr čtverců
Standard error of estimate – směrodatná chyba odhadu – rozptýlení hodnot kolem přímky

Intercept (Absolutní člen) – hodnota B0
Std. Error – směrodatná chyba absolutního členu (následují testy Ho – intercept je roven nule)
b* – standardizované koeficienty – umožňují porovnat vliv jednotlivých proměnných

Regresní analýza

Další výsledky
Summary: regression results
První tabulka – statistiky z předchozího souhrnného okna

Druhá tabulka – podrobnější výsledky regrese, včetně nestandardizovaného koeficientu (b) (ten standardizovaný ukazuje relativní přispění jednotlivých proměnných)



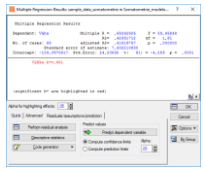
Pro každý koeficient jsou vypočítány hodnoty t-statistiky a p testující, zda je daný parametr významně odlišný od 0 (jestli má proměnná v modelu **své opodstatnění – součást verifikace modelu**).

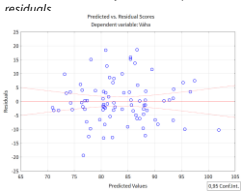
v našem případě – hmotnost = -103,097 + 1,024*výška+*E*

Regresní analýza

Ověření předpokladů:
1) **Správně specifikovaný model**
2) **Střední hodnota chybové složky je rovna 0**
3) **Chybová složka má konstantní rozptyl**
4) Jednotlivé složky chybového vektoru jsou nekorelované
5) Reziduální složka má normální rozdělení

Perform residual analysis > Scatterplots > Predicted vs. residuals



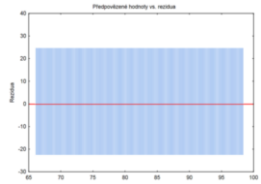


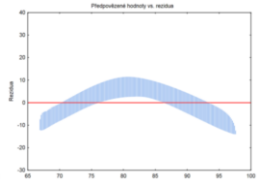
Rezidua konstantně rozptýlena kolem nulové střední hodnoty

Nepřímková závislost

Regresní analýza

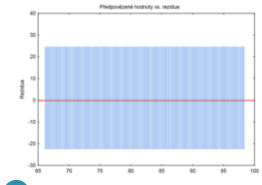
Ověření předpokladů:
1) **Správně specifikovaný model**
2) Střední hodnota chybové složky je rovna 0
3) Chybová složka má konstantní rozptyl
4) Jednotlivé složky chybového vektoru jsou nekorelované
5) Reziduální složka má normální rozdělení

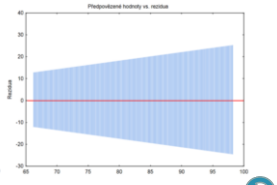




Regresní analýza

Ověření předpokladů:
1) Správně specifikovaný model
2) Střední hodnota chybové složky je rovna 0
3) **Chybová složka má konstantní rozptyl**
4) Jednotlivé složky chybového vektoru jsou nekorelované
5) Reziduální složka má normální rozdělení



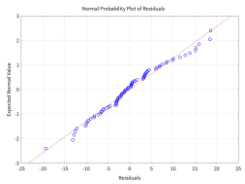


Regresní analýza

Ověření předpokladů:
1) Správně specifikovaný model
2) Střední hodnota chybové složky je rovna 0
3) Chybová složka má konstantní rozptyl
4) Jednotlivé složky chybového vektoru jsou nekorelované
5) **Reziduální složka má normální rozdělení**

Nemusíme ověřovat – jde o nezávislé jedince

Perform residual analysis > Basics > Normal plot of residuals (Kvantilový graf)



V případě normality musí body ležet na proložené přímce

Pokud neleží (dá se dále ověřit testem reziduí) – odhady parametrů modelu a regr. rovnice jsou v pořádku, ale **ne významnost regr. parametrů a konfidenční intervaly**

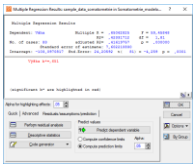
Regresní analýza

Predikce

Predict dependent variable

Compute confidence limits
interval spolehlivosti pro průměrnou hodnotu odezvy

udává rozmezí, ve kterém se s 95% spolehlivostí nachází „*true best fit*“ dané populace



Compute prediction limits
interval spolehlivosti pro individuální hodnotu odezvy

pokud použijete stejnou rovnici na další jedince z dané populace, bude se 95% z nich nacházet v daném rozmezí