Bi8352: Metody antropologie II
jaro 2023
Mgr. Mikoláš Jurda, Ph.D.

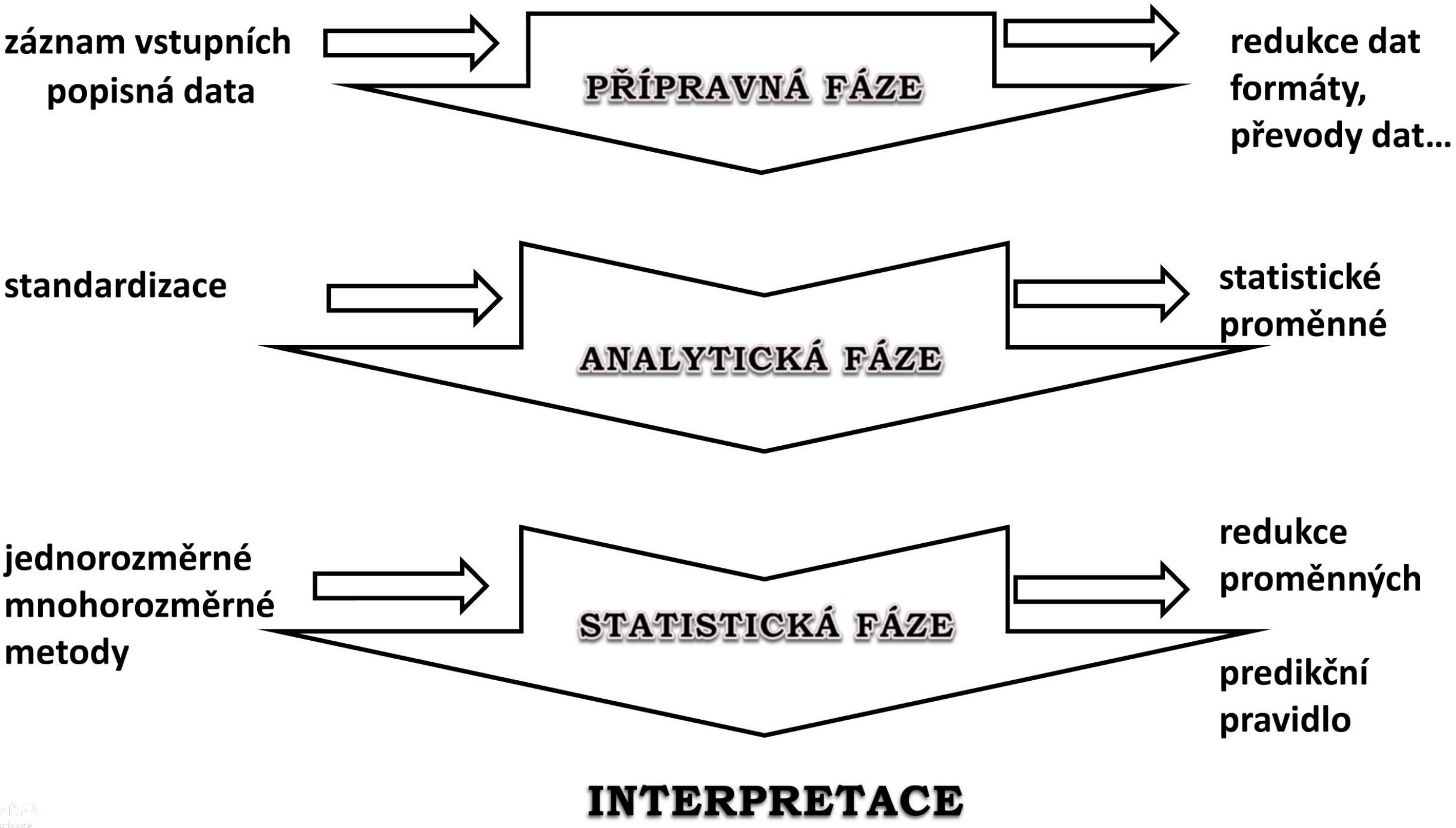
MUNI SCI

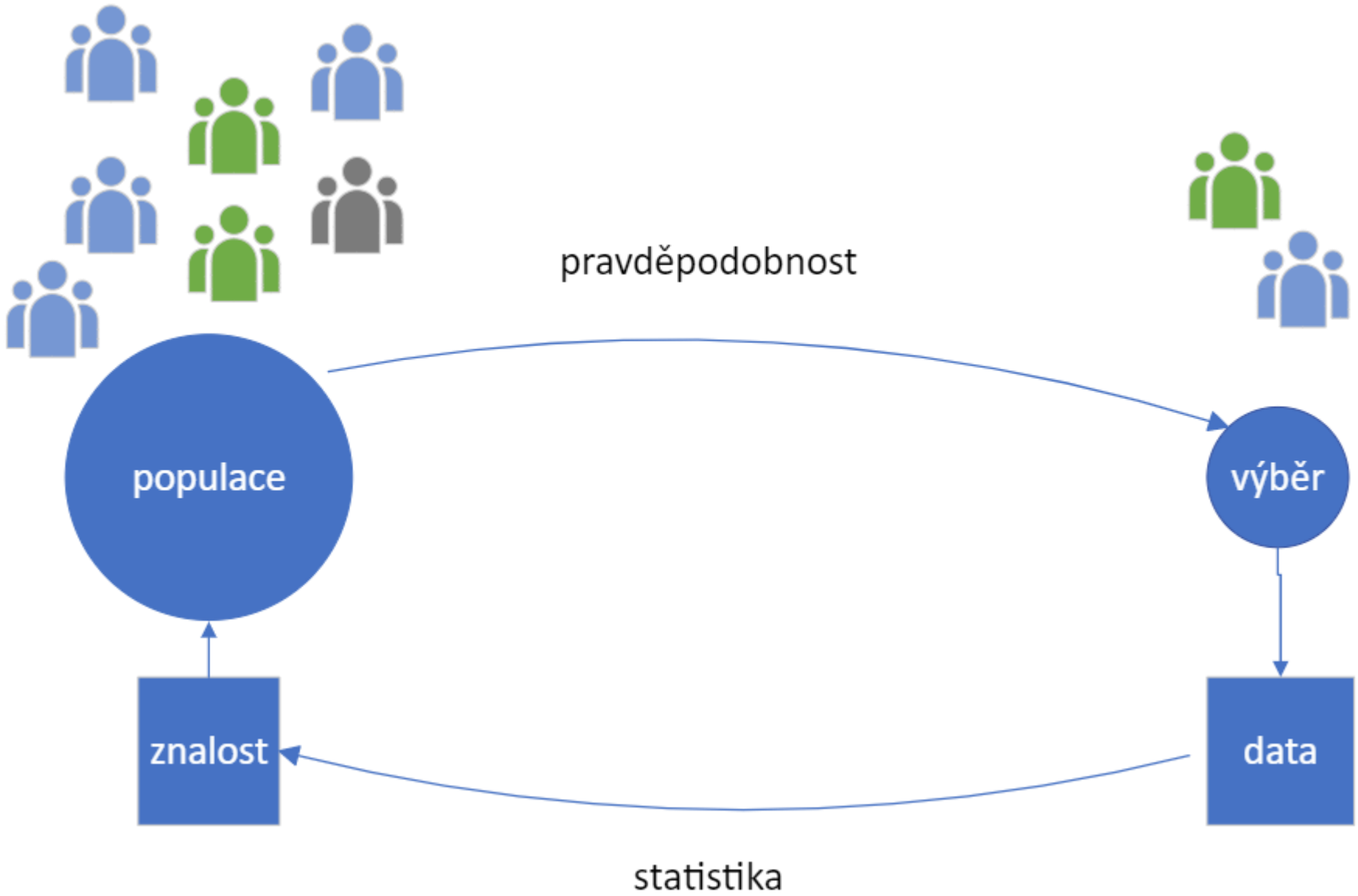
Statistické základy antropologických analýz

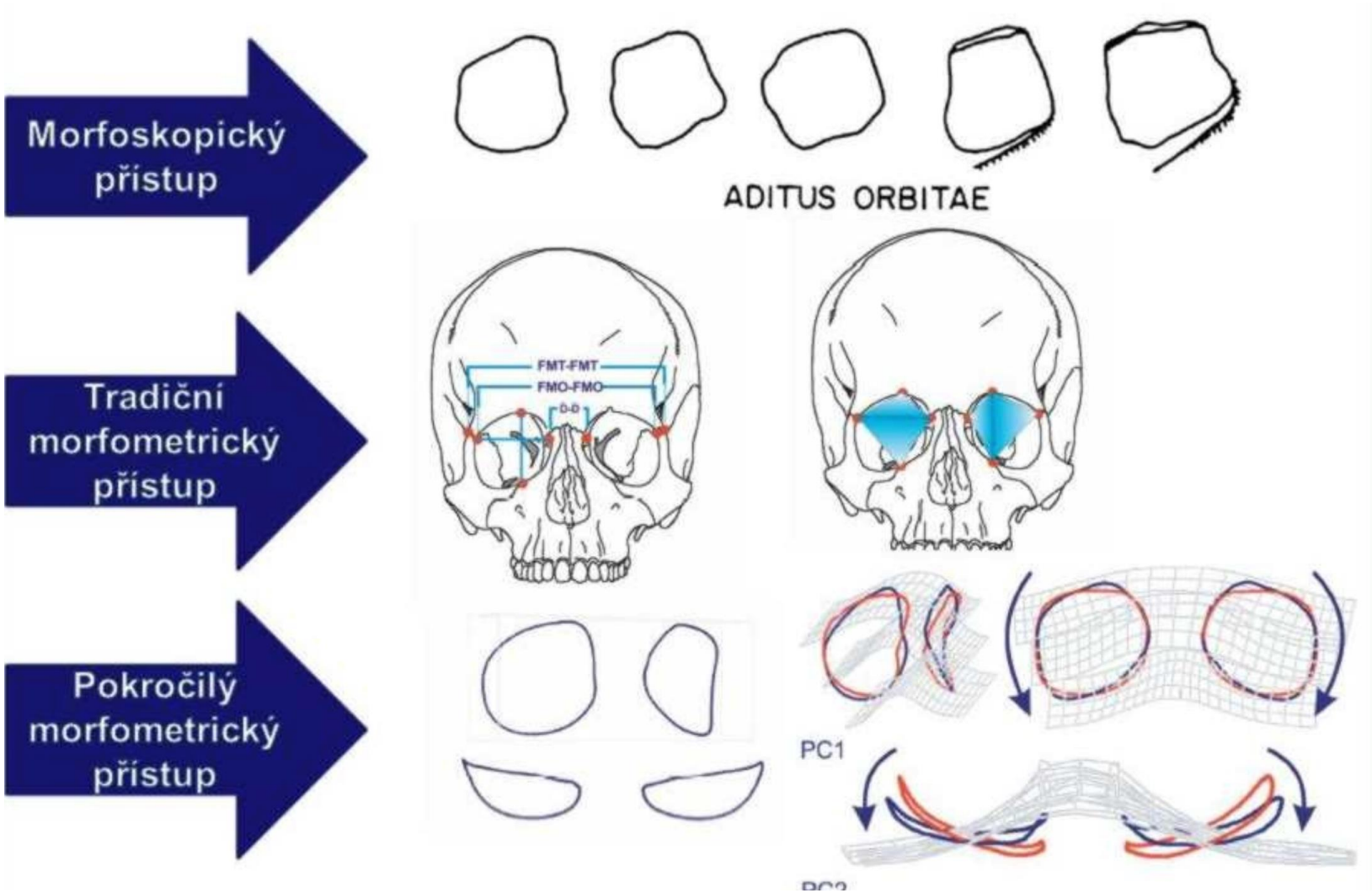
Proč aplikujeme statistické postupy???

Poskytují vědecký základ

- **objektivní** sumarizace výsledků!!!
- vytváření **predikčního pravidla** – odhad neznámých vlastností na základě známých vlastností
- zjišťování chyby odhadu/určení
- kombinace odlišných biologických vlastností do jednotného metodického postupu


záznam vstupních
popisná data
PŘÍPRAVNÁ FÁZE
redukce dat
formáty,
převody dat…
standardizace
ANALYTICKÁ FÁZE
statistické
proměnné
jednorozměrné
mnohorozměrné
metody
STATISTICKÁ FÁZE
redukce
proměnných
predikční
pravidlo
INTERPRETACE


pravděpodobnost
populace
výběr
znalost
data
statistika



rozdílné v **rozšíření**, **využití**, **zpracování** a **možnostech**

Popisná (deskriptivní) statistika

- **základní informace** o vlastnostech studovaného souboru a vztazích různých souborů a dat

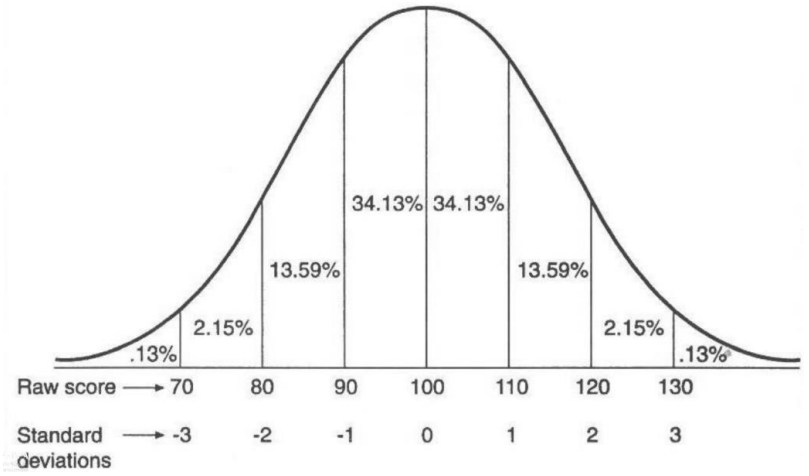
- kontrola splnění předpokladů statistických testů

průměr
rozptyl
medián
modus
SD
směrodatná/standardní chyba (S.E.)
variační koeficient
normalita rozložení

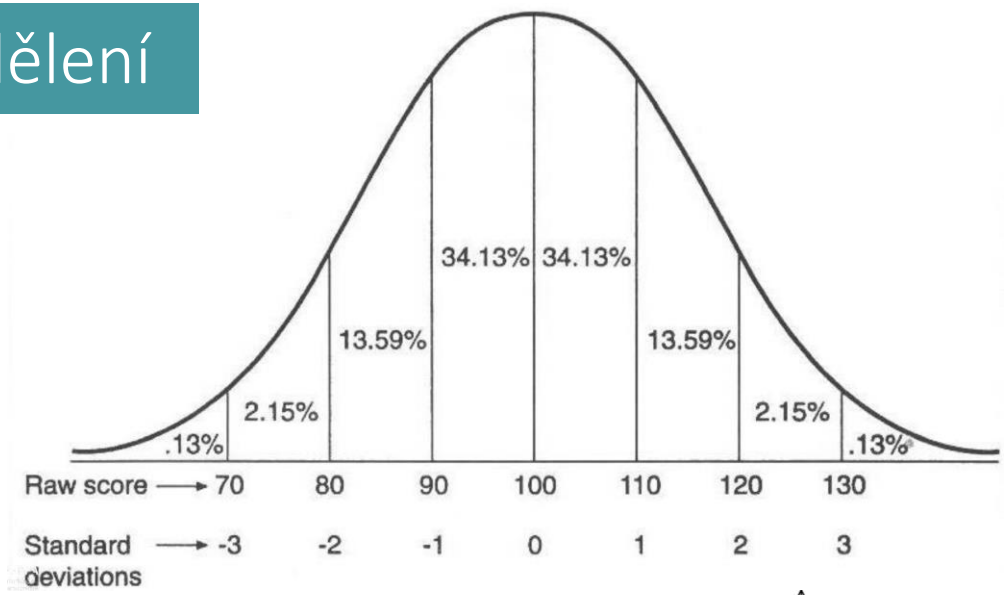
histogram
frekvenční tabulka

kontingenční tabulka

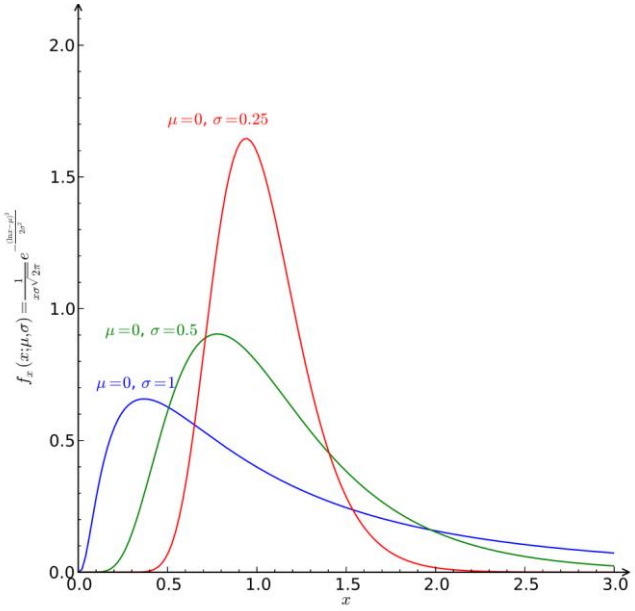
korelace proměnných



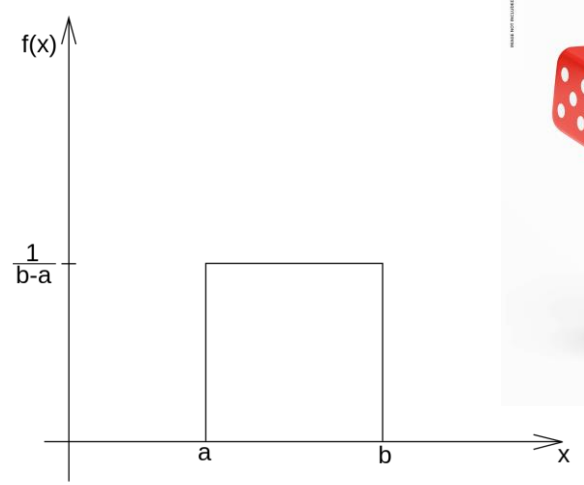
Rozdělení



normální rozdělení



logaritmicko normální rozdělení



rovnoměrné rozdělení (wikipedia.org)

Parametrické a neparametrické testy

PARAMETRICKE TESTY

- při výpočtech používají parametry (průměr, rozptyl)
- a pracují s předpoklady o populaci
- pokud tyto předpoklady nejsou splněny, nemusí být výsledky přesné

NEPARAMETRICKE TESTY

- nepracují s předpoklady o populaci
- nevyžadují např. normální rozložení ani nepřítomnost odlehlých hodnot

- ale mají menší statistickou sílu
- a ne pro všechny analýzy jsou k dispozici

v PASTU také permutační testy

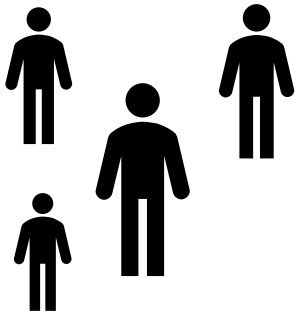
Parametrické a neparametrické testy – příklad rozdílů mezi skupinami

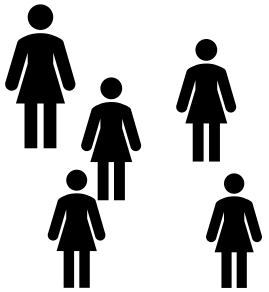
Dvou výběrový t-test

Jsou muži statisticky významně vyšší než ženy?

Předpoklady: normální rozdělení u mužů i u žen, shoda rozptylů

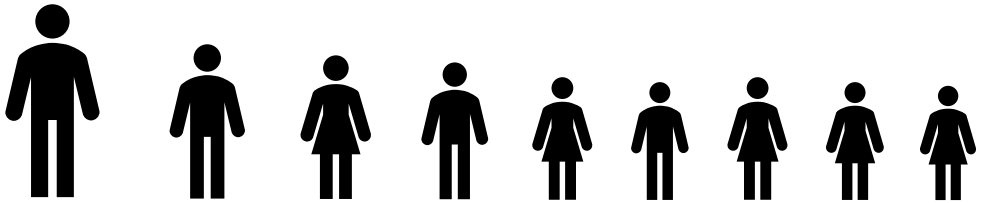
$$t\ hodnota = \frac{rozdíl\ průměrů}{variabilita} = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$





Mann-Whitney U-test

- každému jedinci přiřadí pořadí



Monte-Carlo permutace

- náhodně je promíchá a zjišťuje, v kolika „namícháních“ byl rozdíl mezi skupinami menší

Statistické testy – hodnota p

testovaná hypotéza – H0
Muži a ženy se neliší ve své výšce
Naměřená data mají normální rozložení
Šířka lebky a její délka vzájemně nekorelují

testovaná hypotéza – H1
Muži a ženy se liší ve své výšce
Naměřená data nemají normální rozdělení
Šířka lebky a její délka vzájemně korelují

Tyto hypotézy můžeme jen zamítnout

a v tom případě přijímáme platnost těchto

hodnota p – udává, jak pravděpodobná jsou naměřená data v případě, že platí testovaná hypotéza
neboli – jaká je pravděpodobnost, že bychom při platnosti Ho získali extrémnější hodnoty testové statistiky

pokud je p < hladina významnosti (z konvence 0,05, 0,02 nebo 0,01), pak H0 zamítáme a platí H1
pokud je p > hladina významnosti pak ji nezamítáme
POZOR – nezamítnutí není potvrzení – ve statistice jen zamítáme, nic nepotvrzujeme

jednorozměrné metody

vícerozměrné metody

lineární regrese

jednorozměrná
vícenásobná
vícerozměrná
vícenásobná vícerozměrná

diskriminační analýza

KLASIFIKACE DO JEDNÉ ZE DVOU SKUPIN

KLASIFIKACE DO JEDNÉ Z VÍCE SKUPIN

kanonická analýza

KLASIFIKACE DO JEDNÉ Z VÍCE SKUPIN

JAKÉKOLIV ZAŘAZENÍ DO TYPŮ

ODHAD TĚLESNÝCH PROPORCÍ

ODHAD VĚKU JEDINCE

DALŠÍ KVANTITATIVNÍ ODHADY

PAST

[Past 4 - the Past of the Future - Natural History Museum (uio.no)]

PAST – import souborů a práce s nimi

- importovat ze souborů různých typů – xls, txt, ale nejlépe vložením ze schránky

Zvykněte si mít data ve správném formátu!!!

proměnné (každá jedinečné pojmenování, bez mezer a bez čísla na začátku)

případy

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	sex	age	origin	YOD	YOB	G-OP	EU-EU	BA-B	ZYG-ZYG	D-D	RH-NS	ZM-ZM
2	184	F	25	admixed	1947	1922	184	147	131	126	21	26	90
3	208	F	47	admixed	1947	1900	177	130	129	118	23	36	86
4	213	F	24	admixed	1947	1923	177	135	132	131	22	34	91
5	237	F	30	admixed	1947	1710	180	130	132	120	18	31	91
6	377	F	49	admixed			174	133	130	117	21	29	89
7	389	F	100	admixed	1950	1905	173	140	128	126	22	28	93
8	475	F	48	admixed	1966	1905	168	130	123	114	20	34	80
9	491	F	38	admixed			169	124	118	110	16	33	81
10	5	M	40	European	1939	1899	183	136	134	123	20	39	93
11	16	M	23	European	1940	1917	175	140	131	128	25	41	90
12	18	F	34	European	1937	1903	166	133	127	111	20	31	88
13	19	F	19	European	1937	1918	172	140	130	122	17	36	85
14	26	M	40	European	1938	1898	182	146	141	128	25	36	86
15	28	M	47	European	1938	1891	174	140	143	131	20	37	93
16	34	M	33	European	1940	1907	183	146	126	121	16	39	88
17	37	F	30	European			163	138	131	118	17	34	85
18	48	M	75	European	1941	1866	170	147	131	130	25	40	89
19	49	M	22	European	1941	1919	168	144	130	132	20	37	90
20	50	M	39	European	1941	1902	177	134	134	126	18	38	94
21	57	M	44	European	1968	1924	184	154	138	136	21	31	102
22	59	M	66	European	1945	1879	178	147	132	138	23	38	89
23	62	M	66	European	1963	1897	184	154	137	135	22	38	102
24	64	M	40	European	1945	1905	183	136	141	129	24	33	91
25	72	M	60	European	1945	1885	181	145	123	132	27	32	95
26	74	F	50	European	1968	1918	162	144	130	119	20	35	86
27	78	F	40	European	1945	1905	172	133	124	116	19	30	116
28	81	F	68	European	1945	1877	174	141	133	126	26	37	92
29	96	F	30	European	1945	1915	178	143	140	131	23	36	97
30	117	M	58	European	1946	1888	177	146	132	127	20	40	97

PAST – import souborů a práce s nimi

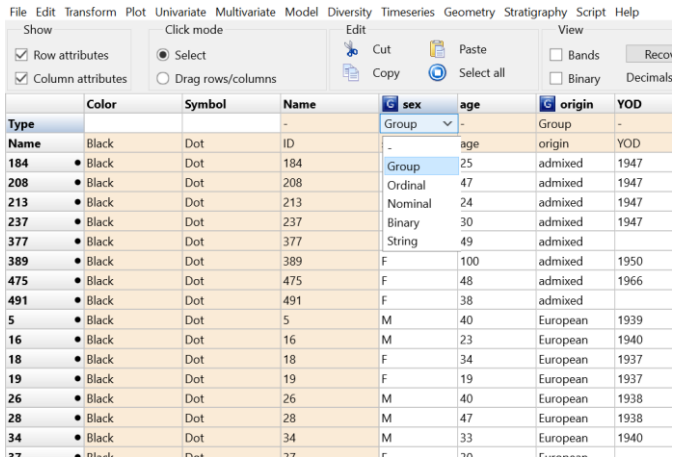
- importovat ze souborů různých typů – xls, txt, ale nejlépe vložením ze schránky

Pokud aktivujete „Row attributes“ a „Column attributes“, pak můžete překopírovat i záhlaví s názvy proměnných i ID případů.

Show: Row attributes, Column attributes | Click mode: Select, Drag rows/columns | Edit: Cut, Copy, Paste, Select all | View: Bands, Binary, Recover windows, Decimals:

	Color	Symbol	Name	sex	age	origin	YOD	YOB	G-OP	EU-EU	BA-B	ZYG-ZYG
Type			-	Group	-	Group	-	-	-	-	-	-
Name	Black	Dot	ID	sex	age	origin	YOD	YOB	G-OP	EU-EU	BA-B	ZYG-ZYG
184	Black	Dot	184	F	25	admixed	1947	1922	184	147	131	126
208	Black	Dot	208	F	47	admixed	1947	1900	177	130	129	118
213	Black	Dot	213	F	24	admixed	1947	1923	177	135	132	131
237	Black	Dot	237	F	30	admixed	1947	1710	180	130	132	120
377	Black	Dot	377	F	49	admixed			174	133	130	117
389	Black	Dot	389	F	100	admixed	1950	1905	173	140	128	126
475	Black	Dot	475	F	48	admixed	1966	1905	168	130	123	114
491	Black	Dot	491	F	38	admixed			169	124	118	110
5	Black	Dot	5	M	40	European	1939	1899	183	136	134	123
16	Black	Dot	16	M	23	European	1940	1917	175	140	131	128
18	Black	Dot	18	F	34	European	1937	1903	166	133	127	111
19	Black	Dot	19	F	19	European	1937	1918	172	140	130	122
26	Black	Dot	26	M	40	European	1938	1898	182	146	141	128
28	Black	Dot	28	M	47	European	1938	1891	174	140	143	131
34	Black	Dot	34	M	33	European	1940	1907	183	146	126	121
37	Black	Dot	37	F	30	European			163	138	131	118
48	Black	Dot	48	M	75	European	1941	1866	170	147	131	130
49	Black	Dot	49	M	22	European	1941	1919	168	144	130	132
50	Black	Dot	50	M	39	European	1941	1902	177	134	134	126
57	Black	Dot	57	M	44	European	1968	1924	184	154	138	136
59	Black	Dot	59	M	66	European	1945	1879	178	147	132	138
62	Black	Dot	62	M	66	European	1963	1897	184	154	137	135
64	Black	Dot	64	M	40	European	1945	1905	183	136	141	129
72	Black	Dot	72	M	60	European	1945	1885	181	145	123	132
74	Black	Dot	74	F	50	European	1968	1918	162	144	130	119
78	Black	Dot	78	F	40	European	1945	1905	172	133	124	116
81	Black	Dot	81	F	68	European	1945	1877	174	141	133	126
96	Black	Dot	96	F	30	European	1945	1915	178	143	140	131
117	Black	Dot	117	M	58	European	1946	1888	177	146	132	127

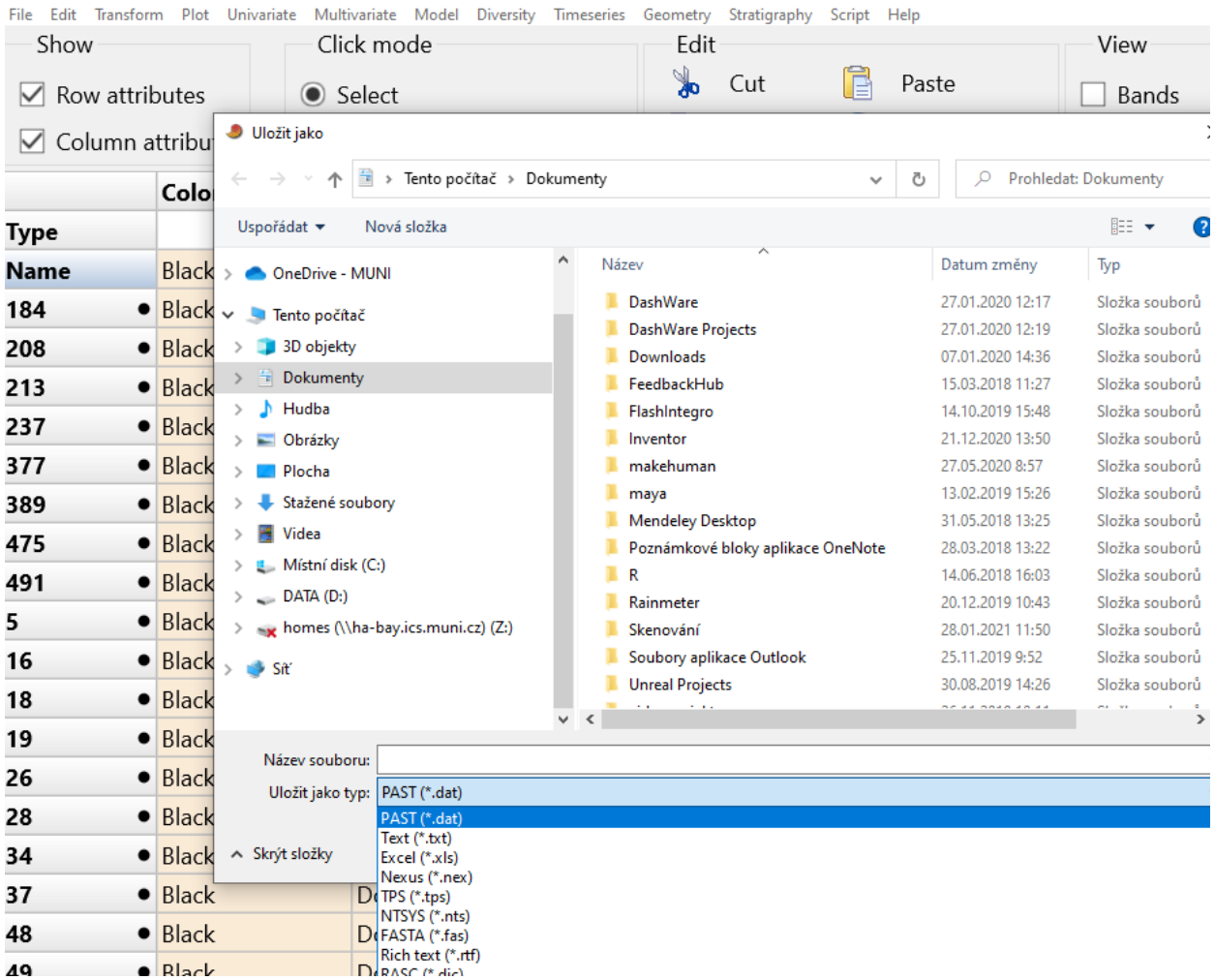
u grupovacích proměnných je potřeba nastavit jejich „Type“ na Group



až na výjimky se proměnné pro analýzy zadávají výběrem daného sloupce

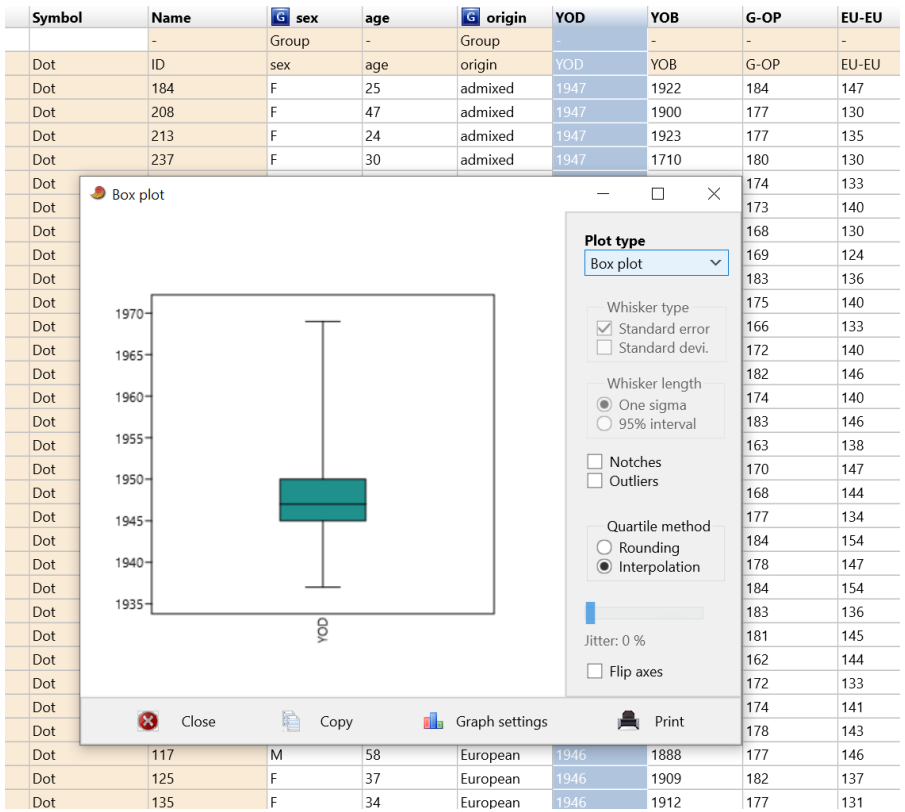
PAST – uložení

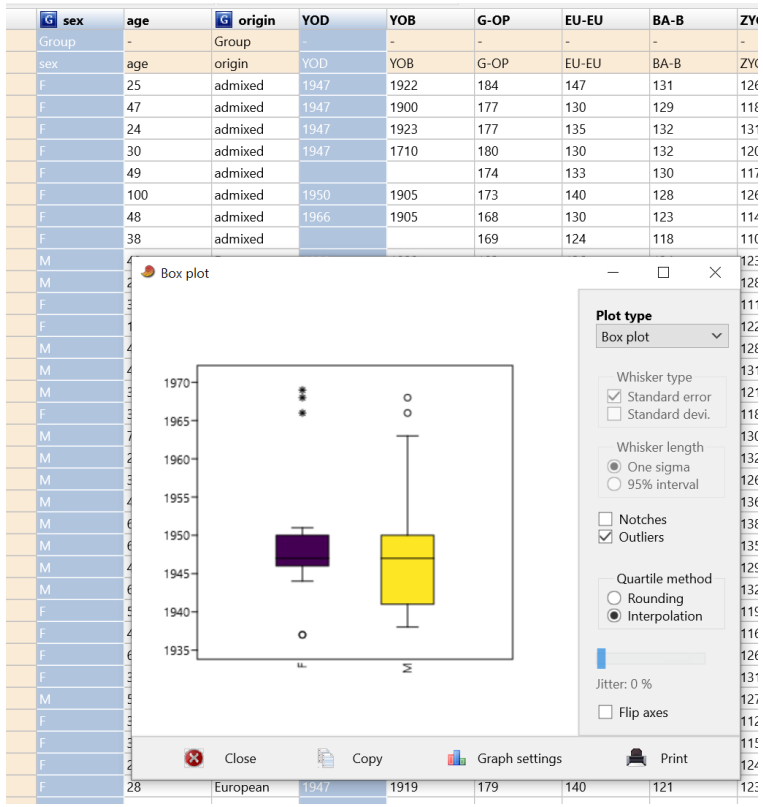
- tabulku je možné uložit jako nativní soubor programu PAST (formát .dat) i ve formě spousty dalších formátů
- *File > Save as...*



Popisná statistika – vizuální hodnocení – krabicový graf

Pokud nevyberete grupovací proměnnou, zobrazí se graf pro celý soubor, pokud vyberete grupovací proměnnou, pak odděleně pro definované skupiny *Plot* > *Barchart/Boxplot* (je potřeba přepnout *Plot type* na *Box plot* a případně zaškrtnout *Outliners*)

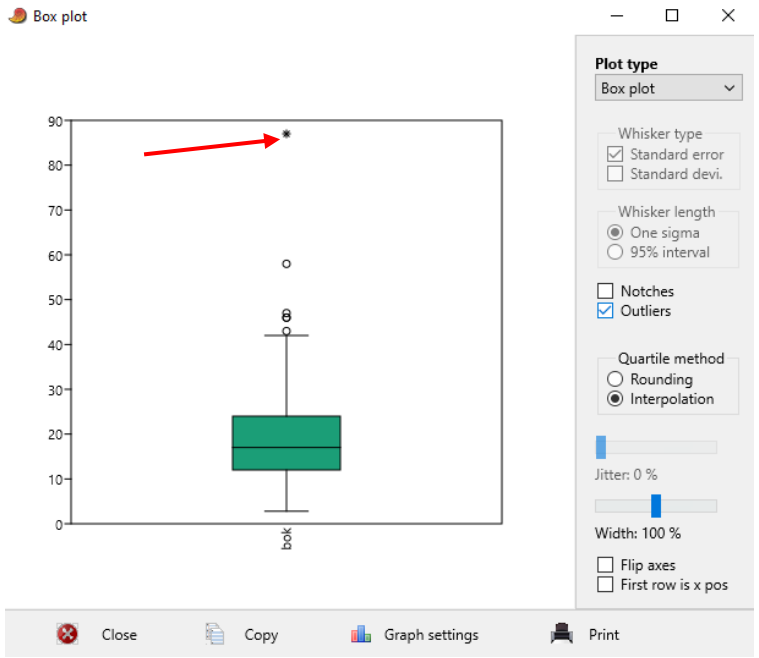




Popisná statistika – hledání odlehlých hodnot

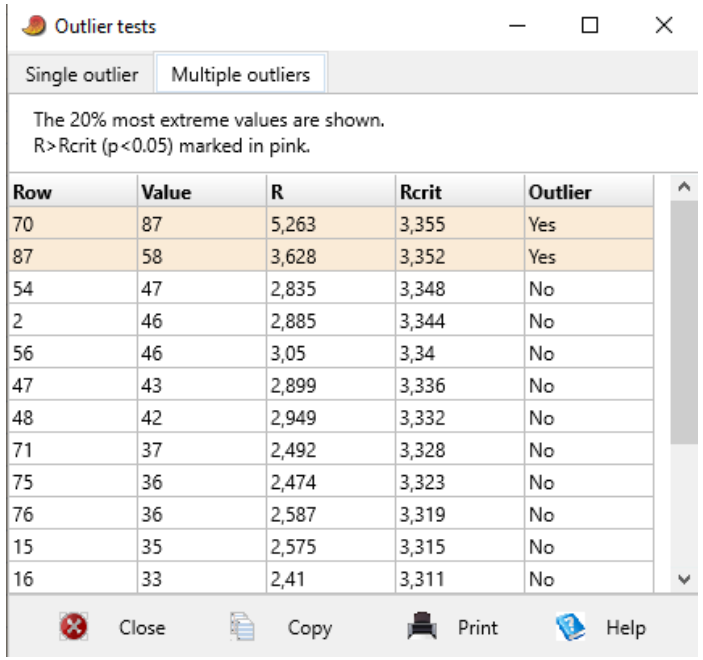
- mohou být extrémními jedinci
- mohou být patologičtí jedinci (často je přítomnost patologie vylučovacím kritériem
- může jít o chybu měření – někdy odvoditelné z ostatních hodnot

Vizuální hodnocení – např. Box-plot



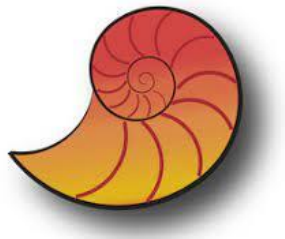
Nejde ale zatím zobrazit číslo jedince.

Test *Univariate* > *Outlier*



Row	Value	R	Rcrit	Outlier
70	87	5,263	3,355	Yes
87	58	3,628	3,352	Yes
54	47	2,835	3,348	No
2	46	2,885	3,344	No
56	46	3,05	3,34	No
47	43	2,899	3,336	No
48	42	2,949	3,332	No
71	37	2,492	3,328	No
75	36	2,474	3,323	No
76	36	2,587	3,319	No
15	35	2,575	3,315	No
16	33	2,41	3,311	No

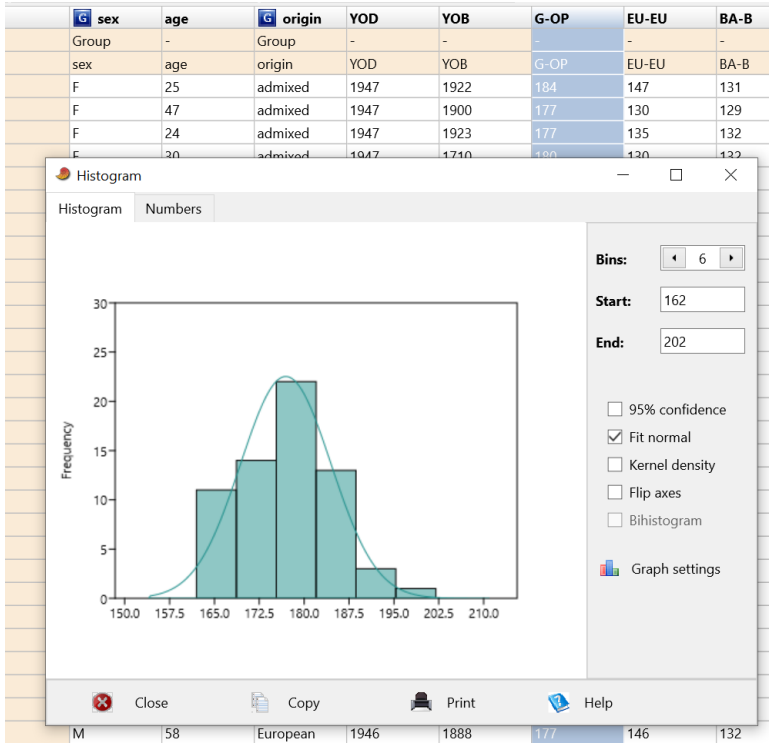
Test v první záložce najde jen jednoho!

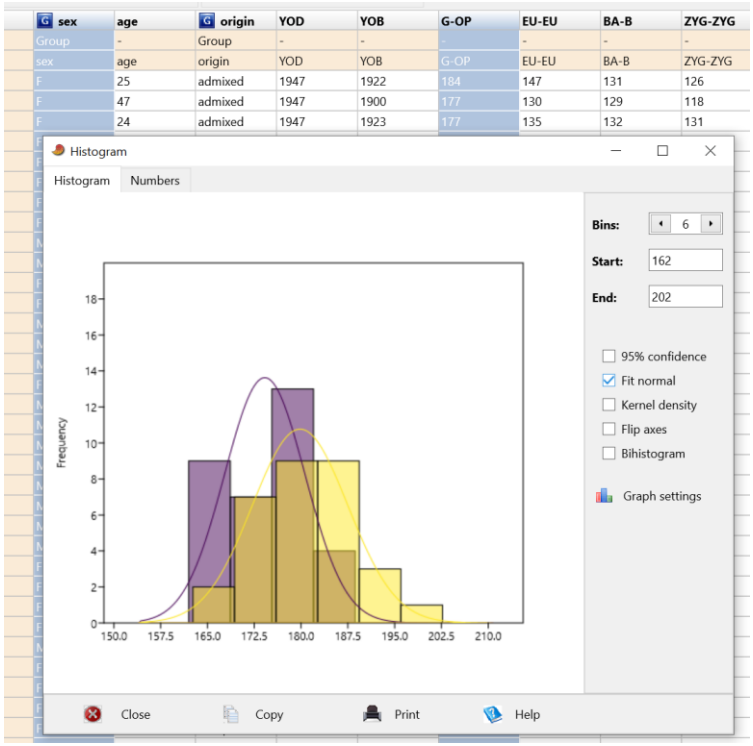


Popisná statistika – vizuální hodnocení – histogram

Umožňuje posoudit rozdělení hodnot a srovnat je s předpokládaným rozložením (křivka). Při současném výběru grupovací proměnné podle ní soubor rozdělí.

Plot > Histogram



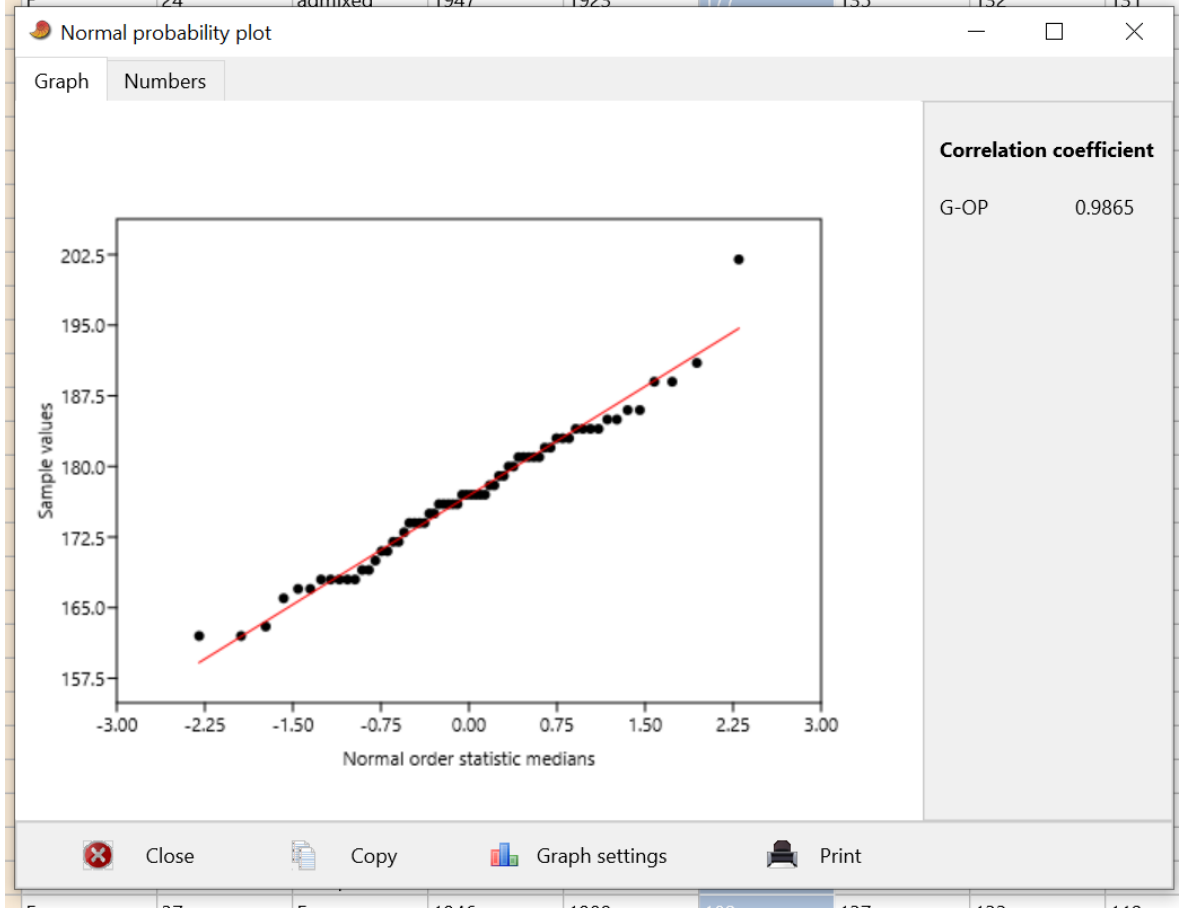


Popisná statistika – vizuální hodnocení – Normal probability plot

Alternativní způsob porovnání pozorovaných hodnot s normálním rozložením (pozorovaný kvantil vs. teoretický kvantil).

Plot > Normal probability plot

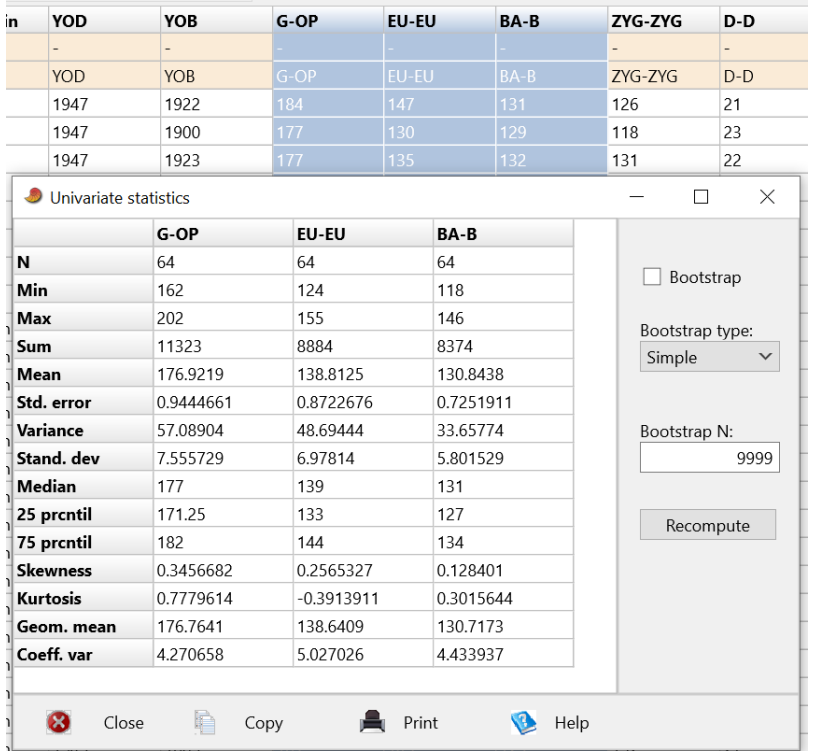
sex	age	origin	YOD	YOB	G-OP	EU-EU	BA-B	ZYG-ZY
Group	-	Group	-	-	-	-	-	-
sex	age	origin	YOD	YOB	G-OP	EU-EU	BA-B	ZYG-ZY
F	25	admixed	1947	1922	184	147	131	126
F	47	admixed	1947	1900	177	130	129	118
F	24	admixed	1947	1923	177	135	132	131

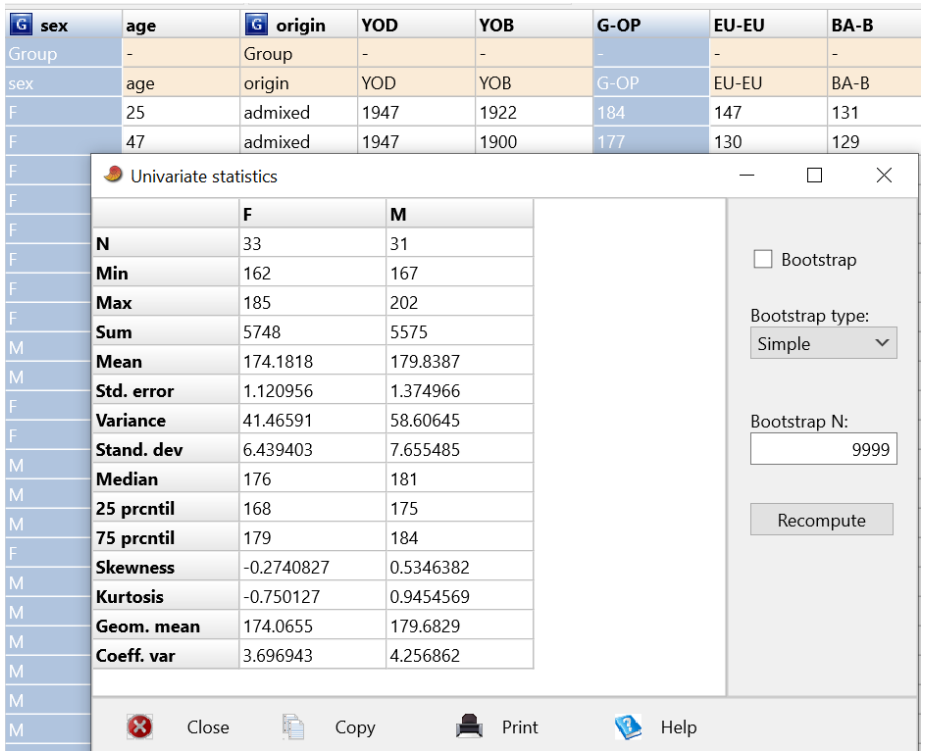


Popisná statistika – číselná popisná statistika

Pokud vyberete víc kontinuálních proměnných, pak se statistika zobrazí pro všechny. Pokud vyberete jednu kontinuální a jednu grupovací proměnnou, pak se zobrazí zvlášť pro každou skupinu.

Univariate > Summary statistics





Popisná statistika – normalita dat

Grafické posouzení

Srovnání s normálním rozložením – viz předchozí grafy

Testování

Statistické testy

Univariate > Normality tests

sex	age	origin	YOD	YOB	G-OP	EU-EU	BA-B
Group	-	Group	-	-	-	-	-
sex	age	origin	YOD	YOB	G-OP	EU-EU	BA-B
F	25	admixed	1947	1922	184	147	131
F	47	admixed	1947	1900	177	130	129
F	24	admixed	1947	1923	177	135	132
F	30	admixed	1947	1710	180	130	132
F							130
F							128
F							123
F							118
M							134
M							131
F							127
F							130
M							141
M							143
M							126
F							131
M							131
M							130
M							134
M							138
M							132
M							137
M							141
M							123
F	30	European	1968	1918	162	144	130

Tests for normal distribution

	F	M
N	33	31
Shapiro-Wilk W	0.9617	0.9641
p(normal)	0.288	0.3727
Anderson-Darling A	0.393	0.2685
p(normal)	0.3571	0.659
p(Monte Carlo)	0.365	0.6845
Lilliefors L	0.1263	0.08143
p(normal)	0.1931	0.8658
p(Monte Carlo)	0.1899	0.8593
Jarque-Bera JB	1.295	1.82
p(normal)	0.5233	0.4026
p(Monte Carlo)	0.3692	0.2083

Copy Print Monte Carlo N: 9999

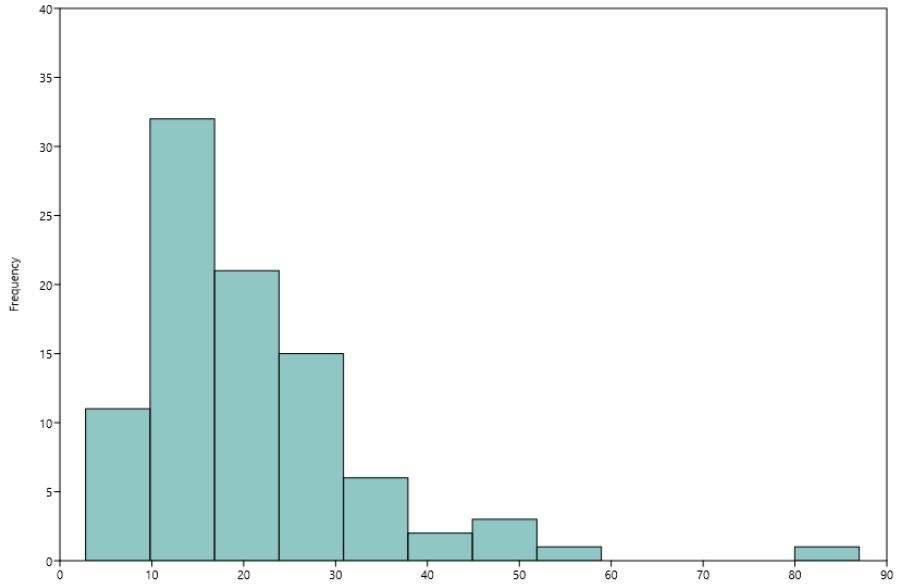
Close Help Recompute

Transformace dat (nelineární)

Mění rozložení dat, proto jsou použitelné k přiblížení se normálnímu rozdělení.

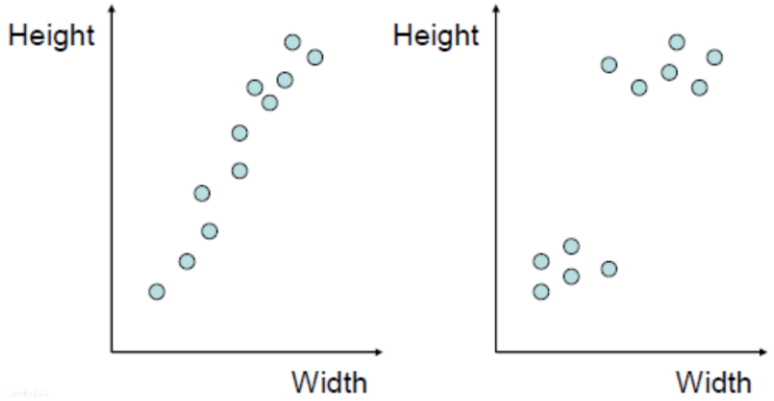
Logaritmická transformace (y = logx) se používá pro výrazně levostranné rozdělení.

- *Transform > Log*



Vztahy dvou proměnných – kovariance a korelace

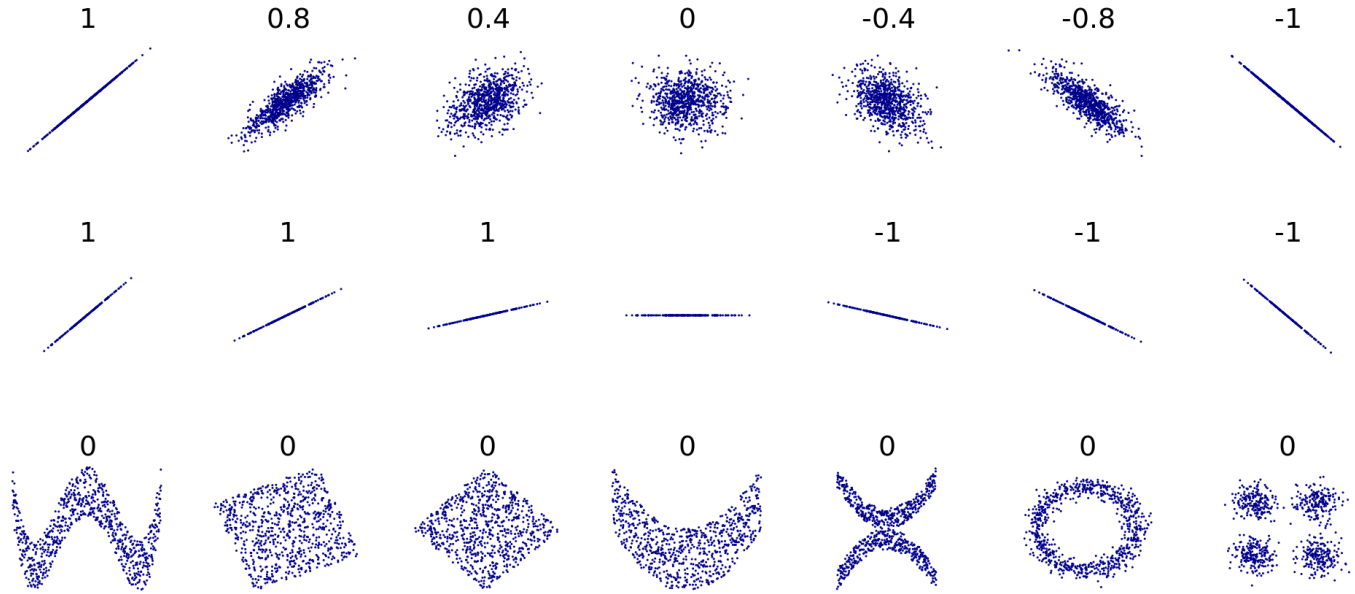
- kovariance – vztah proměnných, důležité je znaménko, **hodnota čísla se neřeší** (ta totiž odvisí i od absolutní hodnoty analyzované proměnné)



pozitivní – negativní – bez trendu

$$\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

- korelace vyjadřuje také sílu vzájemného vztahu a je nezávislá na jednotkách. Síla korelace neznamená statistickou významnost, ta totiž závisí na velikosti souboru.



Korelační analýza – vizuální posouzení

Univariate > Correlation

	YOB	G-OP	EU-EU	BA-B	ZYG-ZYG	D-D	RH-NS	ZM-ZM	M	N	O
	-	-	-	-	-	-	-	-	-	-	-
	YOB	G-OP	EU-EU	BA-B	ZYG-ZYG	D-D	RH-NS	ZM-ZM	M	N	O
	1922	184	147	131	126	21	26	90			
	1900	177	130	129	118	23	36	86			
	1923	177	135	132	131	22	34	91			

Correlation

Table | Plot

	G-OP	EU-EU	BA-B	ZYG-ZYG	D-D	RH-NS	ZM-ZM
G-OP		0.1094	0.00086259	0.0002867	0.0056623	0.030566	0.0011248
EU-EU	0.20203		0.038304	4.8862E-12	0.026958	0.0087577	0.021134
BA-B	0.40637	0.25961		0.0047586	0.03681	0.0050267	0.025465
ZYG-ZYG	0.4388	0.73459	0.3486		0.00084281	4.9356E-05	0.00049608
D-D	0.34208	0.27655	0.26158	0.40708		0.42881	0.0081012
RH-NS	0.2706	0.32513	0.34656	0.4847	0.10064		0.31293
ZM-ZM	0.39804	0.28775	0.27921	0.42306	0.32822	0.12814	

Correlation statistic
- (•) Linear r (Pearson)
- ( ) Spearman's D
- ( ) Spearman's rs
- ( ) Kendall's tau
- ( ) Polyserial rho
- ( ) Partial linear
- ( ) Phi (compositional)

Table format
- (•) Statistic \ p(uncorr)
- ( ) Statistic
- ( ) p(uncorr)
- ( ) Permutation p

[ ] Bonferroni correction

Close | Copy | Print | Help

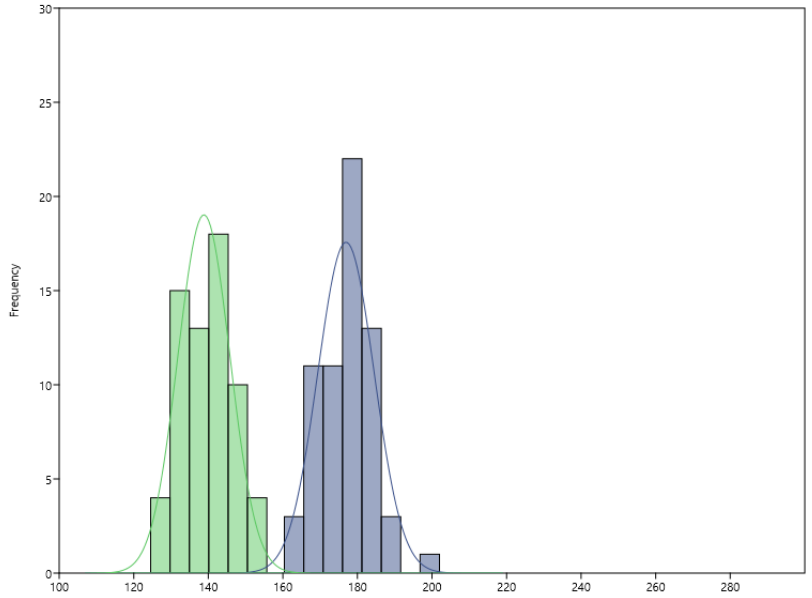
1888 177 146 132 127 20 40 97

velikost korelace

význam korelace (p)

parametrická (Pearson) nebo neparametrická (Spearman)

Korelační analýza



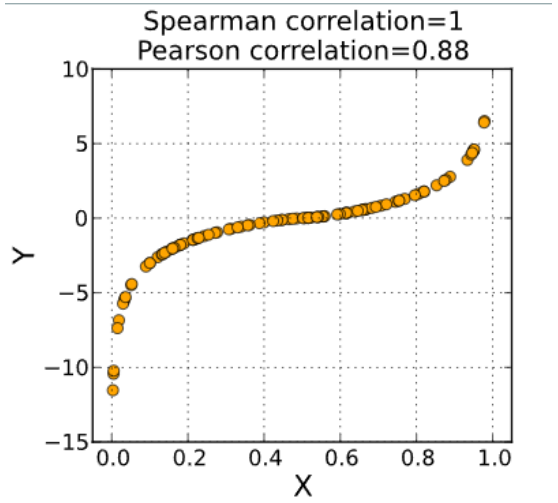
Mají obě proměnné normální rozložení?

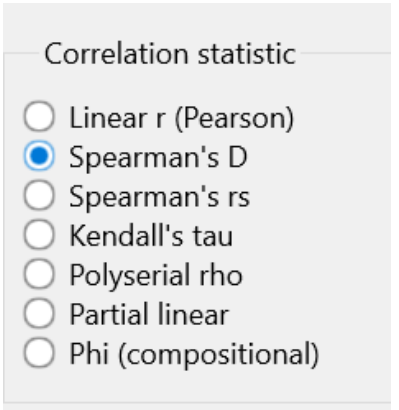
ANO?

Pearsonův korelační koeficient

NE?

Spearmanův korelační koeficient





Jednorozměrná lineární regrese – teorie

Regrese vyjadřuje, jak lze z **nezávislé** proměnné odhadnout **závislou proměnnou**

Jakou výšku postavu měl pravděpodobně člověk s femurem dlouhým 48 cm?

jedna vstupní proměnná

nezávislá = predikující = známá

délka dlouhé kosti

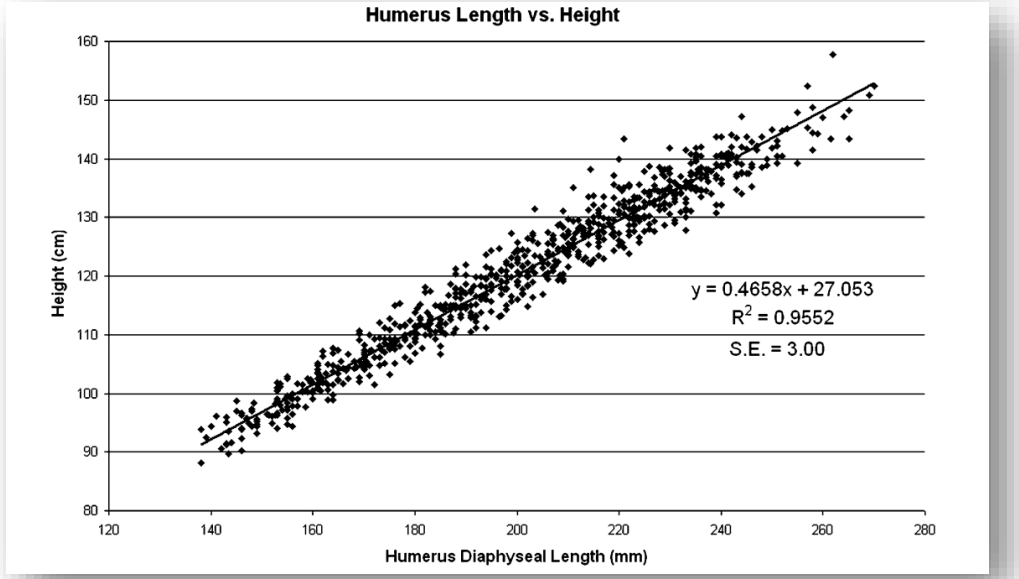
obsah dřeňové dutiny

jedna výstupní proměnná (numerická vlastnost)

závislá = predikovaná = neznámá

výška postavy

dožitý věk



(Smith 2007)

Jednorozměrná lineární regrese – teorie

- analýza hledá přímku, která nejlépe vyjadřuje závislost jedné proměnné na druhé

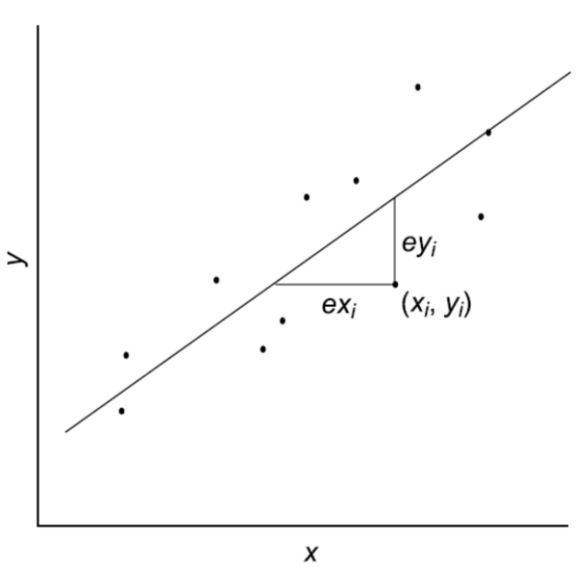
$$y = ax + b\ (+E)$$

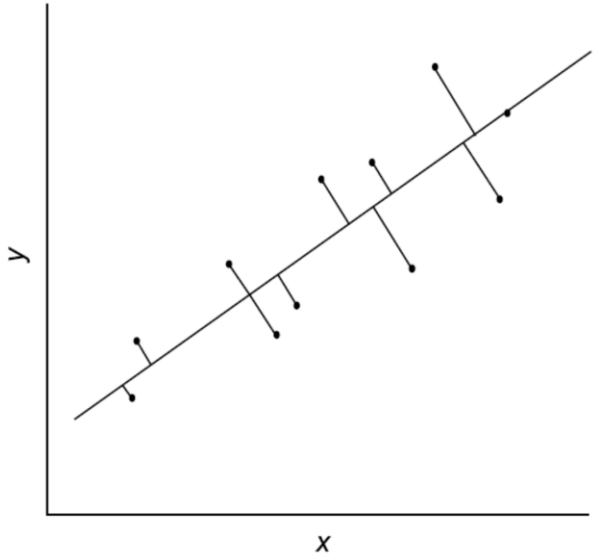
různé metody vyjádření chyby
(jak proložit body přímku?)

metoda nejmenších čtverců

- **pouze y-proměnné**
- x i y (RMA)

euklidovská vzdálenost





(RMA – reduced major axis)

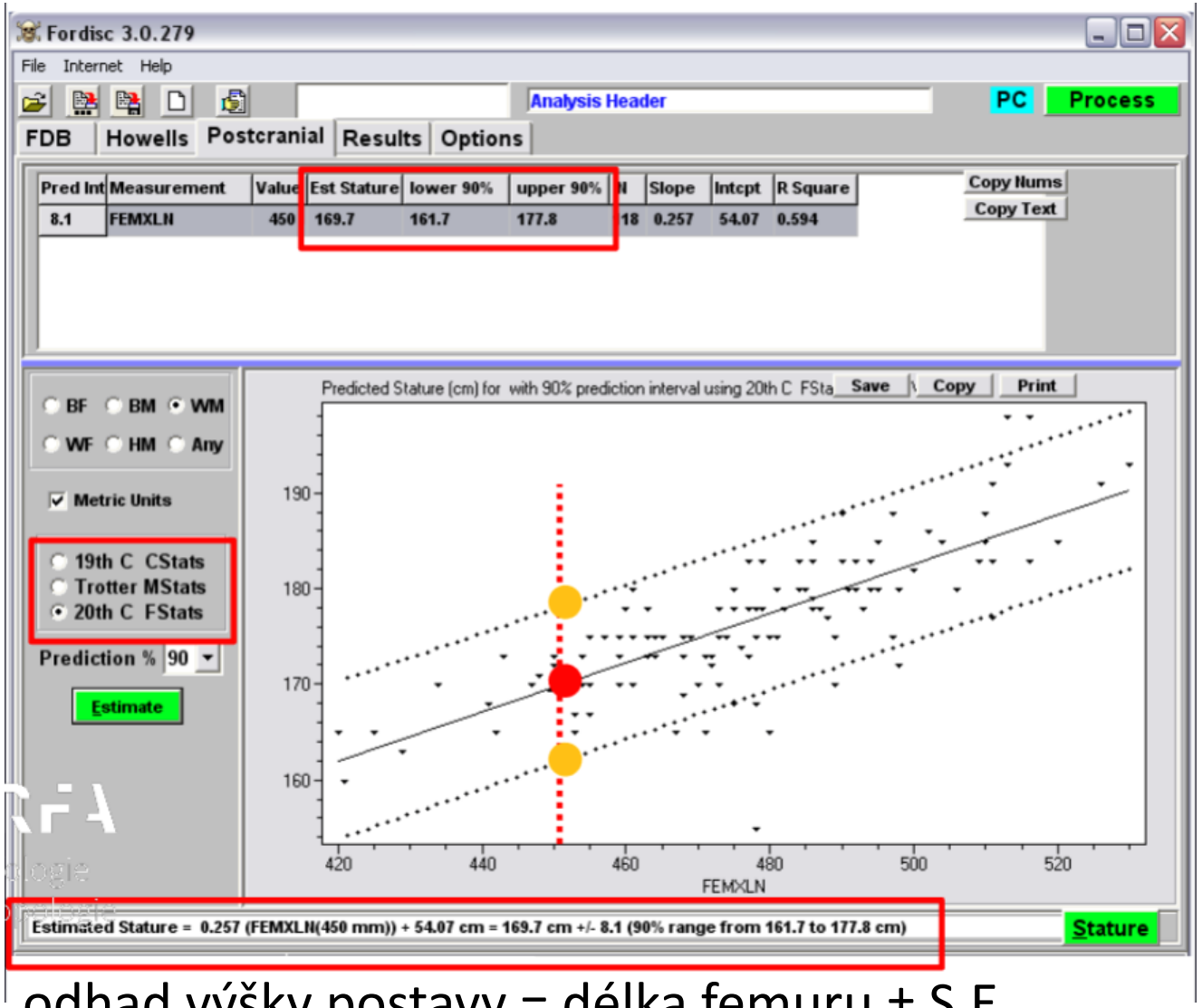
Jednorozměrná lineární regrese

Výstupy $y = ax + b (+E)$

predikční pravidlo = lineární rovnice
- hodnota závislé proměnné (y)
- koeficient závislosti (a)
- položení v prostoru (b)

interval spolehlivosti = konfidenční interval – jak spolehlivě odhadujeme regresní přímku na základě populačního výběru

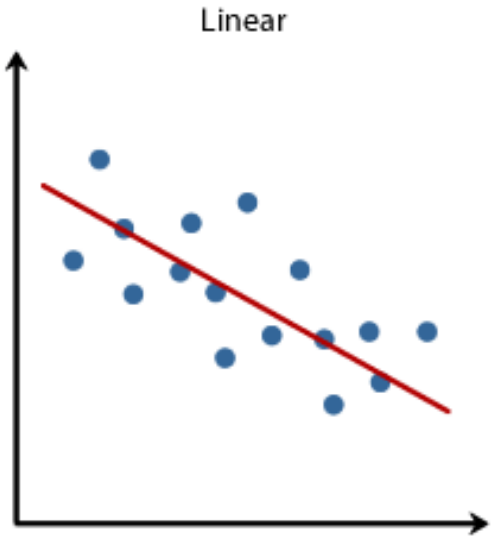
standardní chyba odhadu = jakou chybu odhadu můžeme při dané přímce očekávat

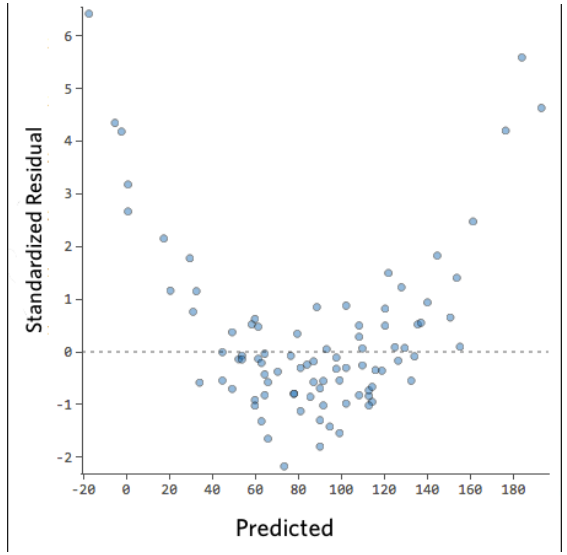


odhad výšky postavy = délka femuru ± S.E.

Jednorozměrná lineární regrese – předpoklady vstupů

Mezi proměnnými musí existovat lineární vztah





Data byla získána nezávisle na sobě

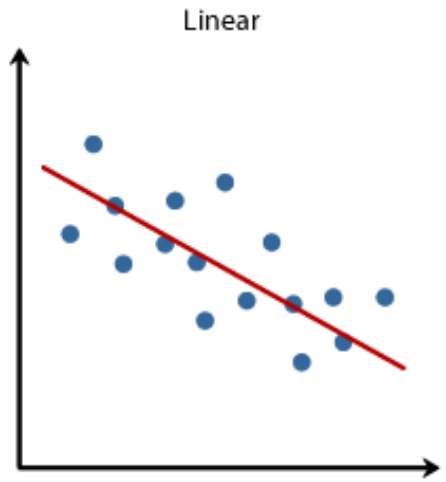
- neodvodili jsme jednu míru z druhé

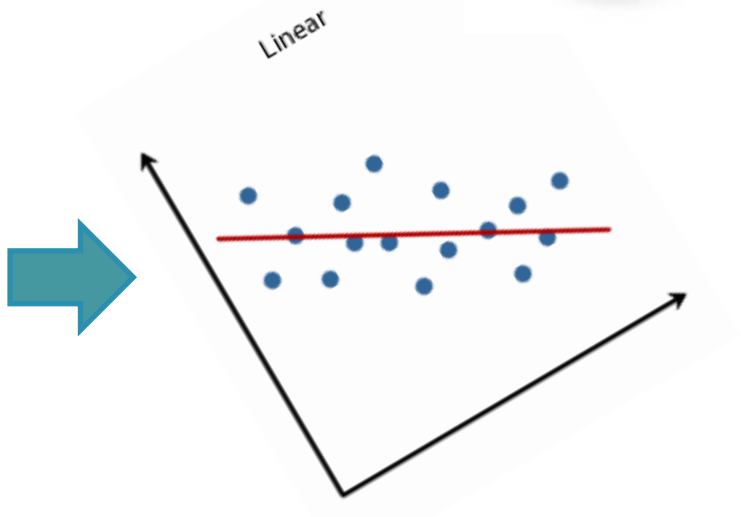
Jednorozměrná lineární regrese – kdy věřit výstupům

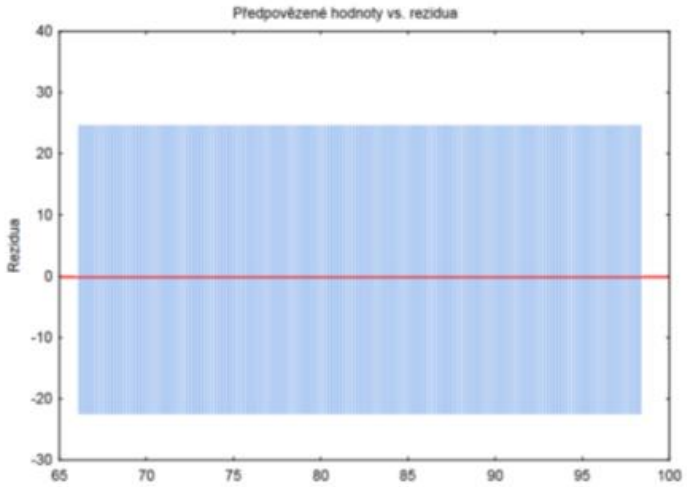
- střední hodnota chybové složky je 0
- chybová složka má konstantní rozptyl
- jednotlivé složky chybového vektoru jsou nekorelované
- reziduální složka má normální rozdělení

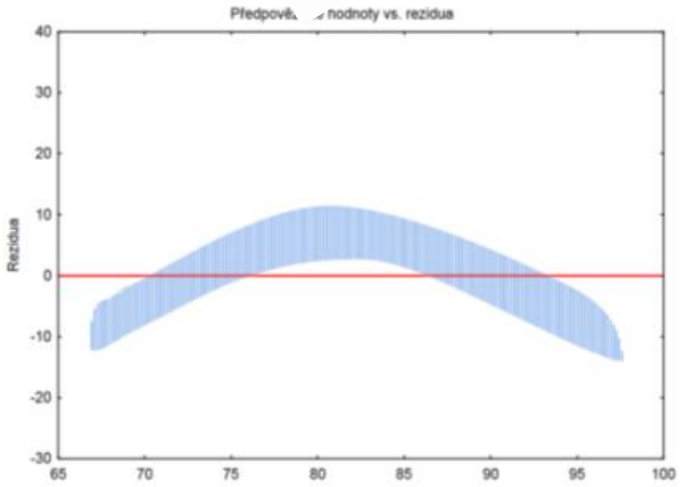
Pokud ne?

- vztah nemusí být lineární





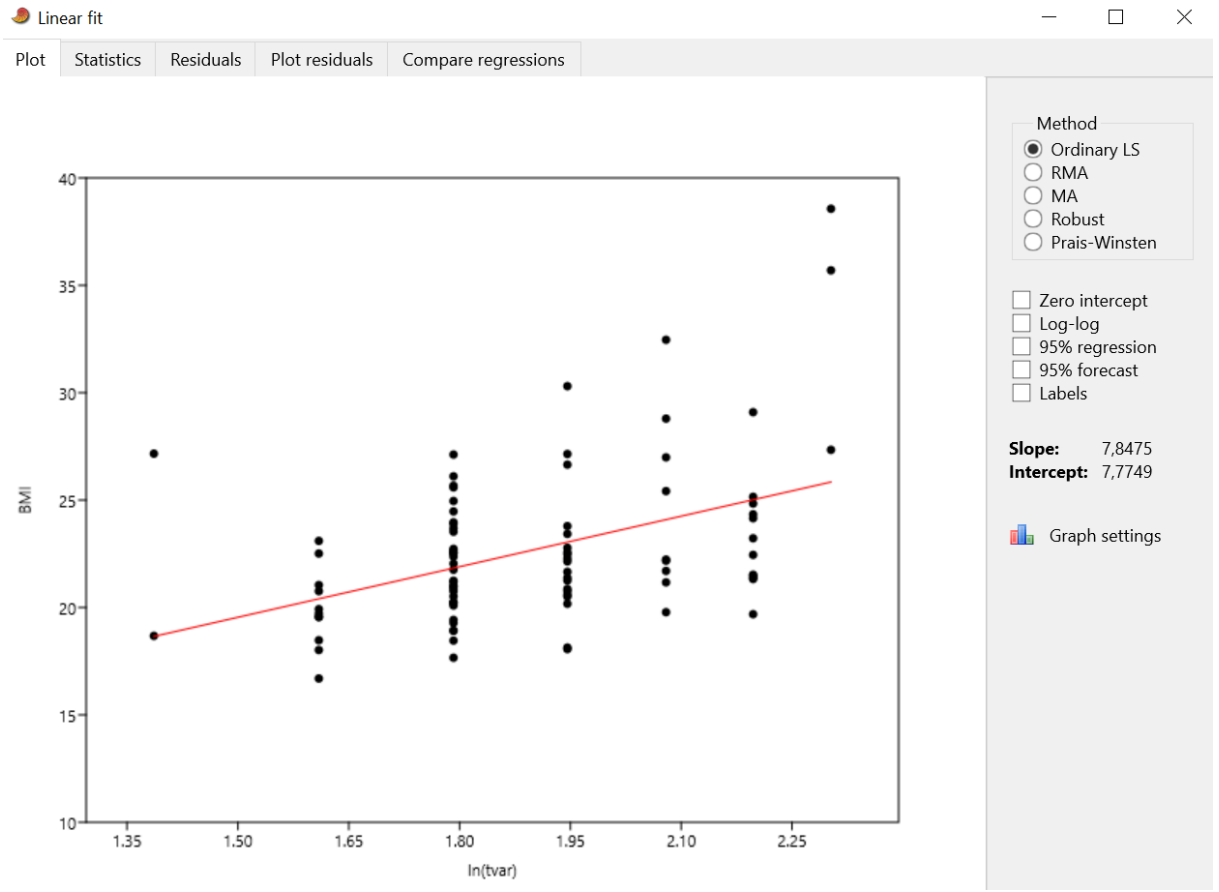




Regresní analýza

Model > *Linear* > *Bivariate* (vysvětlovaná proměnná musí být napravo od vysvětlující)

vysvětlovaná/závislá proměnná



	ln(tvar)	BMI	ln(krk
	2.079441542	32.46489733	2.3025
	1.791759469	22.51938193	2.0794
	2.079441542	25.41927804	2.5649
	1.945910149	20.59728187	2.3025
	1.791759469	22.04466846	1.9459
	2.197224577	21.5253565	2.0794
	1.945910149	22.14773406	2.1972

vysvětlující/nezávislá proměnná

Regresní analýza

Záložka *„Statistics“*

slope a intercept dovolují sestavit regresní rovnici

$$y = ax + b$$

$$BMI = 7{,}8475\ln \times (tvar) + 7{,}7749$$

korelace mezi proměnnými – pokud neexistuje, nemá cenu provádět regresní analýzu!

Ordinary Least Squares Regression: ln(tvar)-BMI

Slope *a*:	7,8475	**Std. error *a*:**	1,7054
***t*:**	4,6015	***p* (slope):**	1,3685E-05
Intercept *b*:	7,7749	**Std. error *b*:**	3,2455

95% bootstrapped confidence intervals (*N*=1999):

Slope *a*:	(2,9276, 12,636)
Intercept *b*:	(-1,1593, 16,849)

Correlation:

***r*:**	0,43642
r^2:	0,19046
***t*:**	4,6015
***p* (uncorr.):**	1,3685E-05
Permutation *p*:	0,0001

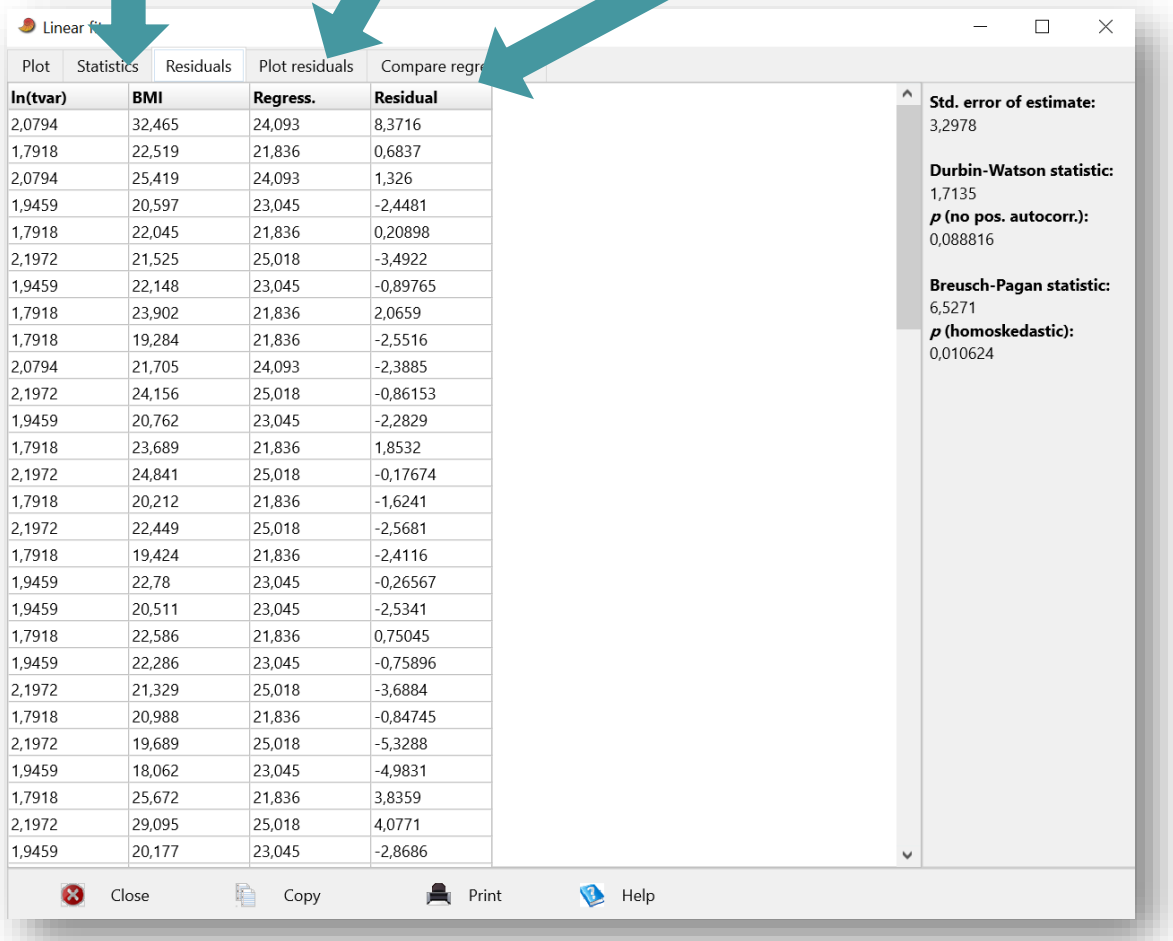
Regresní analýza

Záložka *„Residuals“*

původní hodnoty ln(tvar) a BMI u jednotlivých jedinců

hodnoty BMI, které by byly vypočítané regresí (leží na regresní přímce)

rezidua – rozdíly mezi predikovanou a skutečnou hodnotou



ln(tvar)	BMI	Regress.	Residual
2,0794	32,465	24,093	8,3716
1,7918	22,519	21,836	0,6837
2,0794	25,419	24,093	1,326
1,9459	20,597	23,045	-2,4481
1,7918	22,045	21,836	0,20898
2,1972	21,525	25,018	-3,4922
1,9459	22,148	23,045	-0,89765
1,7918	23,902	21,836	2,0659
1,7918	19,284	21,836	-2,5516
2,0794	21,705	24,093	-2,3885
2,1972	24,156	25,018	-0,86153
1,9459	20,762	23,045	-2,2829
1,7918	23,689	21,836	1,8532
2,1972	24,841	25,018	-0,17674
1,7918	20,212	21,836	-1,6241
2,1972	22,449	25,018	-2,5681
1,7918	19,424	21,836	-2,4116
1,9459	22,78	23,045	-0,26567
1,9459	20,511	23,045	-2,5341
1,7918	22,586	21,836	0,75045
1,9459	22,286	23,045	-0,75896
2,1972	21,329	25,018	-3,6884
1,7918	20,988	21,836	-0,84745
2,1972	19,689	25,018	-5,3288
1,9459	18,062	23,045	-4,9831
1,7918	25,672	21,836	3,8359
2,1972	29,095	25,018	4,0771
1,9459	20,177	23,045	-2,8686

Std. error of estimate:
3,2978

Durbin-Watson statistic:
1,7135
***p* (no pos. autocorr.):**
0,088816

Breusch-Pagan statistic:
6,5271
***p* (homoskedastic):**
0,010624

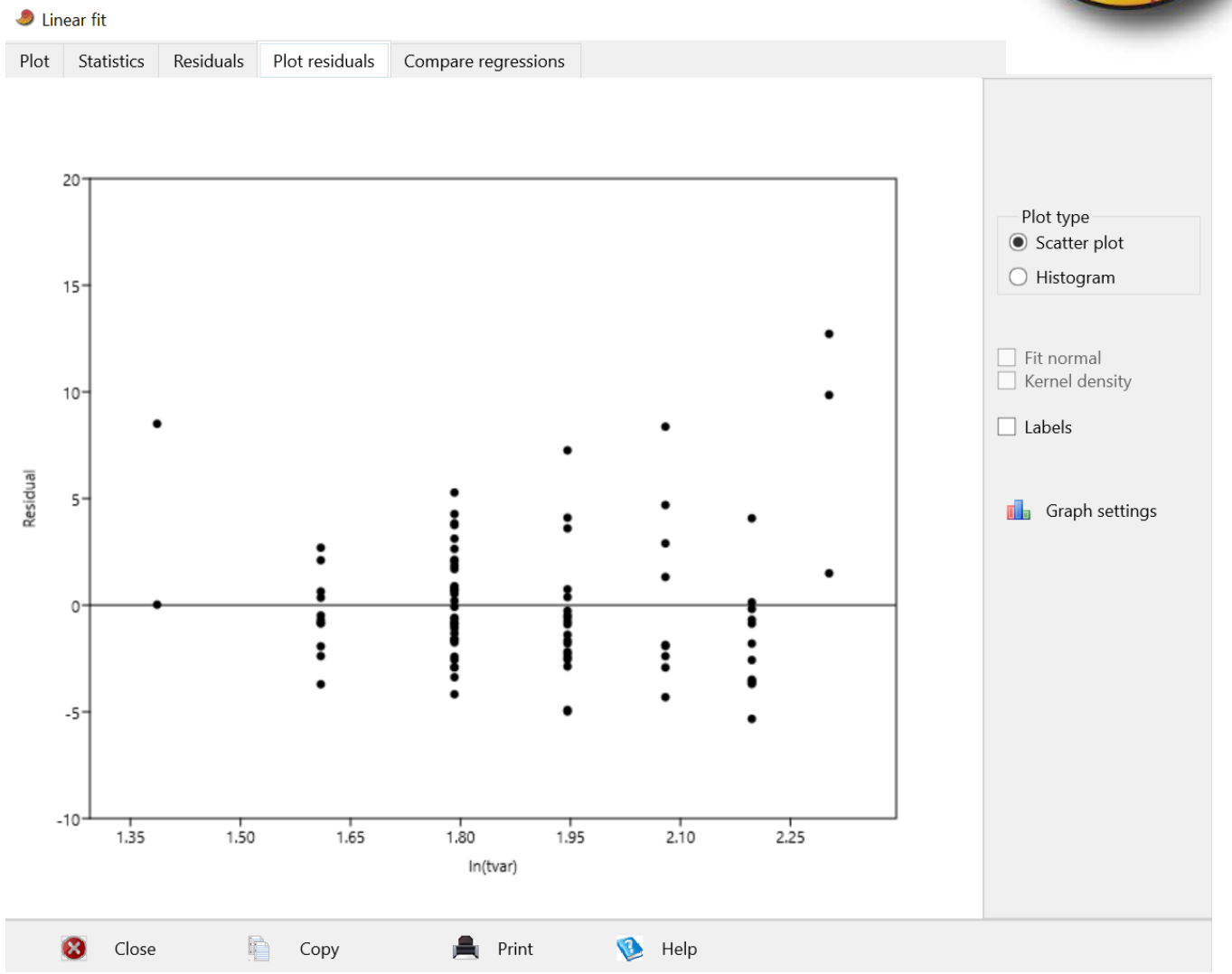
Regresní analýza

Záložka *Plot residuals*

Dovoluje posoudit vhodnost modelu.

Ověření předpokladů

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



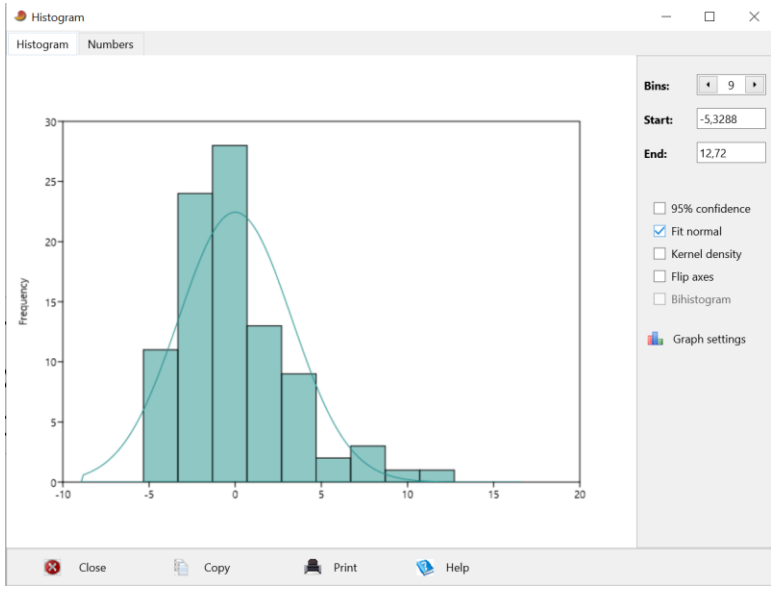
Regresní analýza

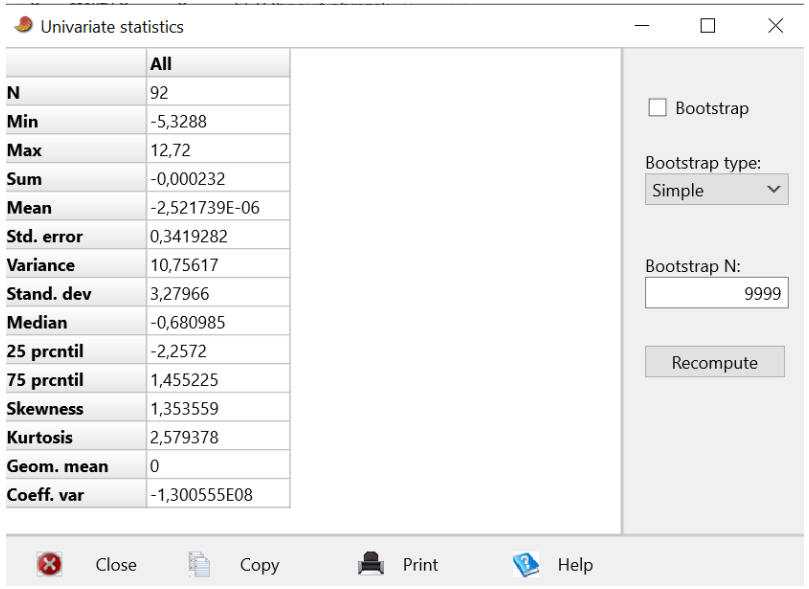
Ověření předpokladů

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

máme nezávislé jedince

Pro ověření bodů 2 a 5 je nutné kopírovat rezidua a zobrazit jejich popisnou statistiku a histogram.





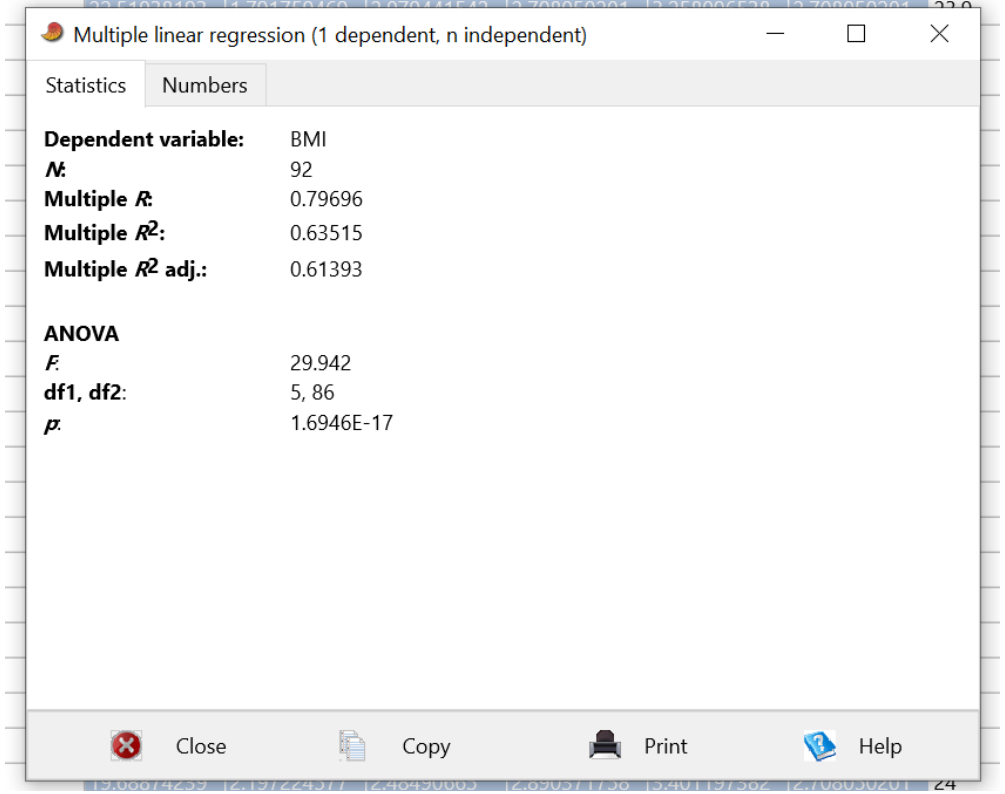
	All
N	92
Min	-5,3288
Max	12,72
Sum	-0,000232
Mean	-2,521739E-06
Std. error	0,3419282
Variance	10,75617
Stand. dev	3,27966
Median	-0,680985
25 prcntil	-2,2572
75 prcntil	1,455225
Skewness	1,353559
Kurtosis	2,579378
Geom. mean	0
Coeff. var	-1,300555E08

Vícenásobná regresní analýza - výstupy

$$y = a_1x_1 + a_2x_2 + \cdots + b$$

Data musejí být uspořádána tak, že závislá proměnná je ve sloupci nejvíce nalevo a nezávislé proměnné jsou napravo od ní.

	BMI	ln(tvar)	ln(krk)	ln(paze)	ln(bricho)	ln(bok)	Bioimp
	-	-	-	-	-	-	-
	BMI	ln(tvar)	ln(krk)	ln(paze)	ln(bricho)	ln(bok)	Bioimp
	32.46489733	2.079441542	2.302585093	3.091042453	3.988984047	3.828641396	42.5



Dependent – závislá (vysvětlovaná) proměnná
N – počet validních případů
Multiple R – koeficient determinace – podíl modelem vysvětlované variability

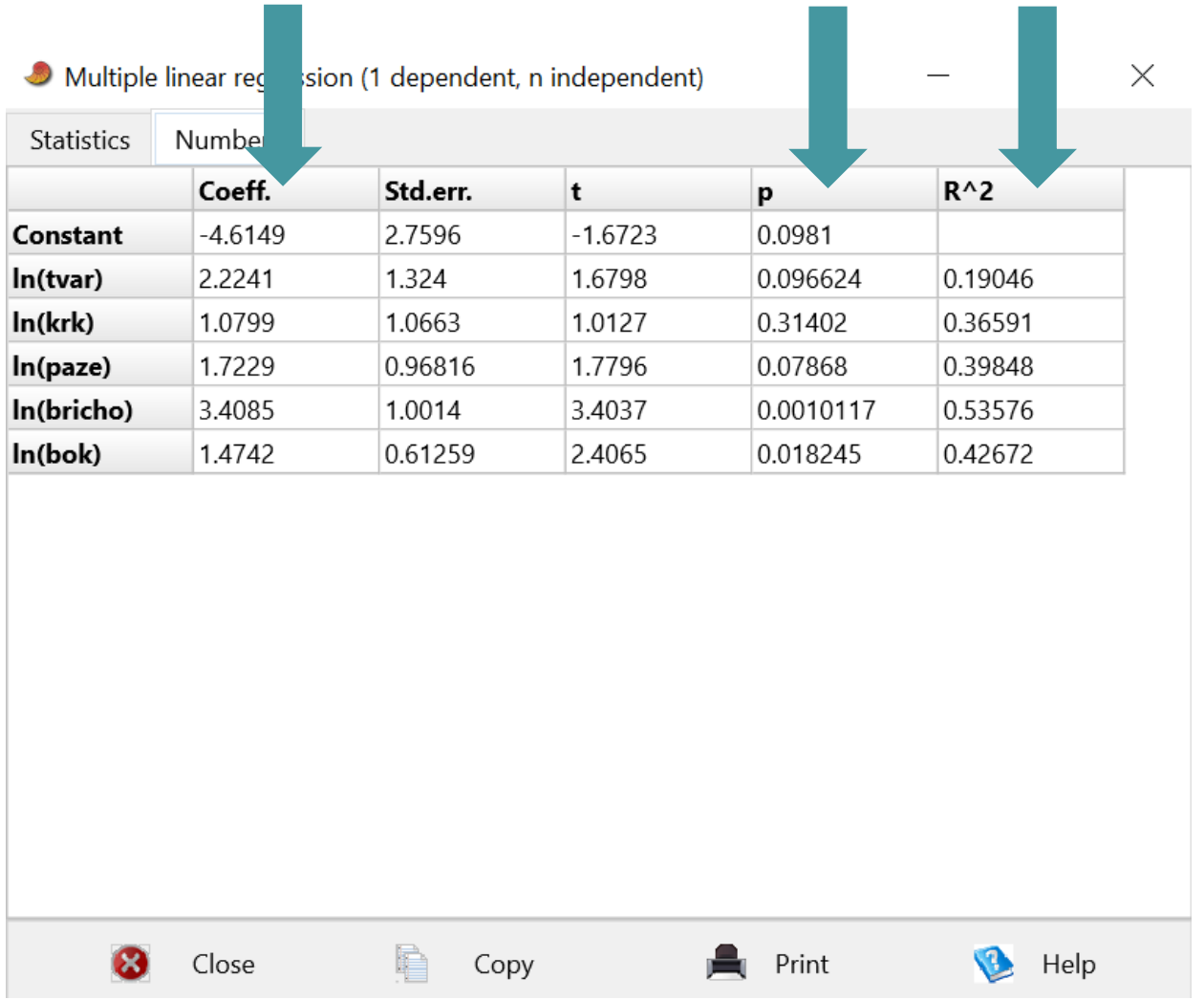
Anova (p) – test statistické významnosti regresního modelu

Regresní analýza

koeficienty/regresní rovnice

statistická významnost proměnné pro model

proměnnou vysvětlovaná variabilita



	Coeff.	Std.err.	t	p	R^2
Constant	-4.6149	2.7596	-1.6723	0.0981	
ln(tvar)	2.2241	1.324	1.6798	0.096624	0.19046
ln(krk)	1.0799	1.0663	1.0127	0.31402	0.36591
ln(paze)	1.7229	0.96816	1.7796	0.07868	0.39848
ln(bricho)	3.4085	1.0014	3.4037	0.0010117	0.53576
ln(bok)	1.4742	0.61259	2.4065	0.018245	0.42672

Regresní analýza – správná podoba výsledků

Predikce

pro odhad podle dané rovnice vypočítejte rovnici podle prvního sloupce

	Coeff.
Constant	-4.6149
ln(tvar)	2.2241
ln(krk)	1.0799
ln(paze)	1.7229
ln(bricho)	3.4085
ln(bok)	1.4742

Jednorozměrná lineární regrese

REGRESE není **KORELACE**

vyjadřuje, jak lze z **nezávislé** proměnné odhadnout **závislou proměnnou**

regrese vyjadřuje vliv změny hodnoty známé proměnné na hodnotu neznámé proměnné

proměnné nemůžeme zaměnit

vyjadřuje **vztah mezi dvěma rovnocennými proměnnými**

vypovídá o tom, do jaké míry se dvě proměnné mění společně

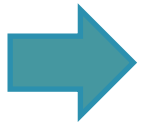
proměnné můžeme zaměnit

Vícenásobná lineární regrese

Řešená otázka
Jak odhadnout jednu spojitou proměnnou ze hodnoty druhé proměnné?

Jakou výšku postavy měl člověk, jehož femur měřil 48 cm a humerus 32?

více vstupních proměnných
nezávislá = predikující = známá



jedna výstupní proměnná (numerická vlastnost)
závislá = predikovaná = neznámá

$$y = ax + bx + c\ (+E)$$

Předpoklady

- vstupní proměnné by neměly korelovat

T-test

Nepárový dvouvýběrový **t-test**

Řešená otázka

Je mezi dvěma skupinami v konkrétní kvantitativní proměnné významný rozdíl?

Jsou muži statisticky významně vyšší než ženy?

Párový dvouvýběrový **t-test**

Řešená otázka

Je mezi stejnými jedinci v různé situaci rozdíl?

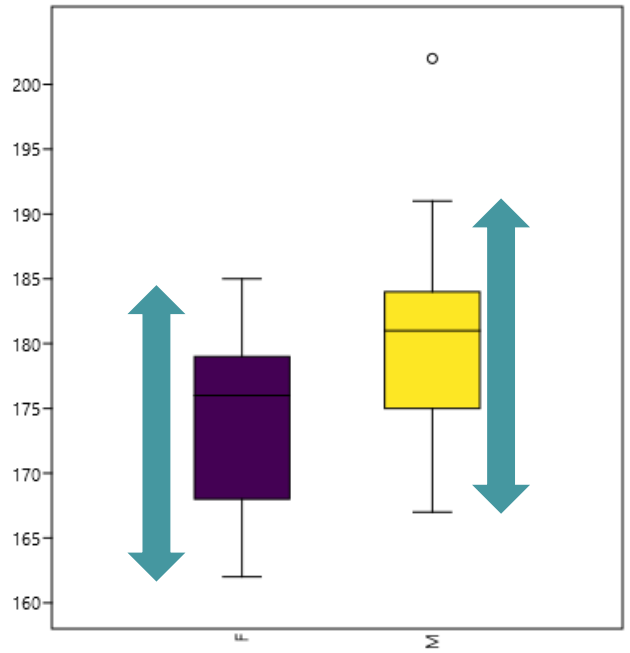
Mají titíž lidé po tréninku silnější stisk ruky než před ním?

T-test

Nepárový dvouvýběrový **t-test**

Předpoklady:
normální rozložení v rámci porovnávaných skupin
- **již představenými postupy**

shoda rozptylu těchto skupin
- F-statistika (je součástí testu)



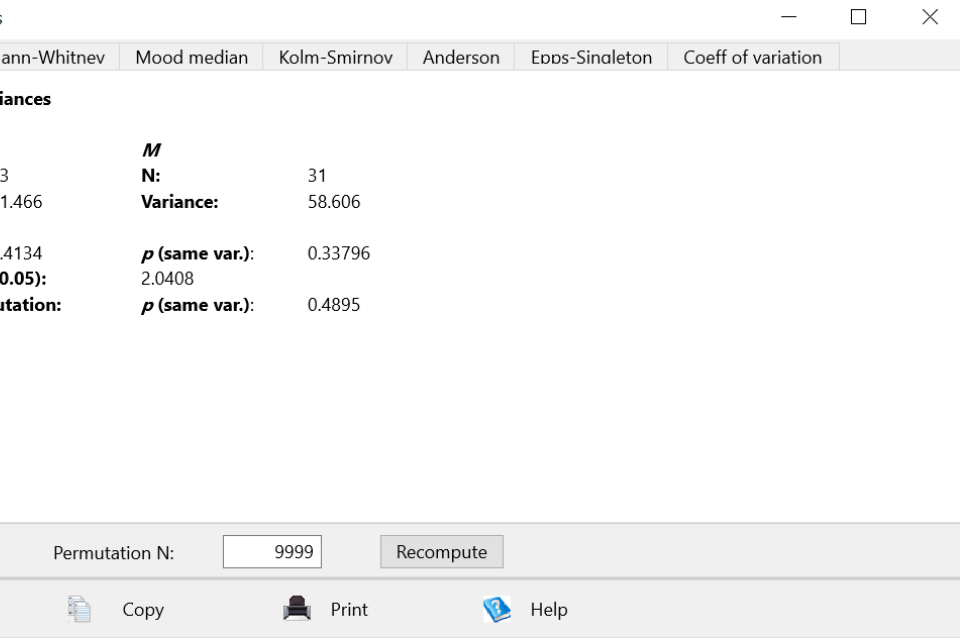
pokud data nesplňují

neparametrické alternativy
v tomto případě
Mann-whitney U-test

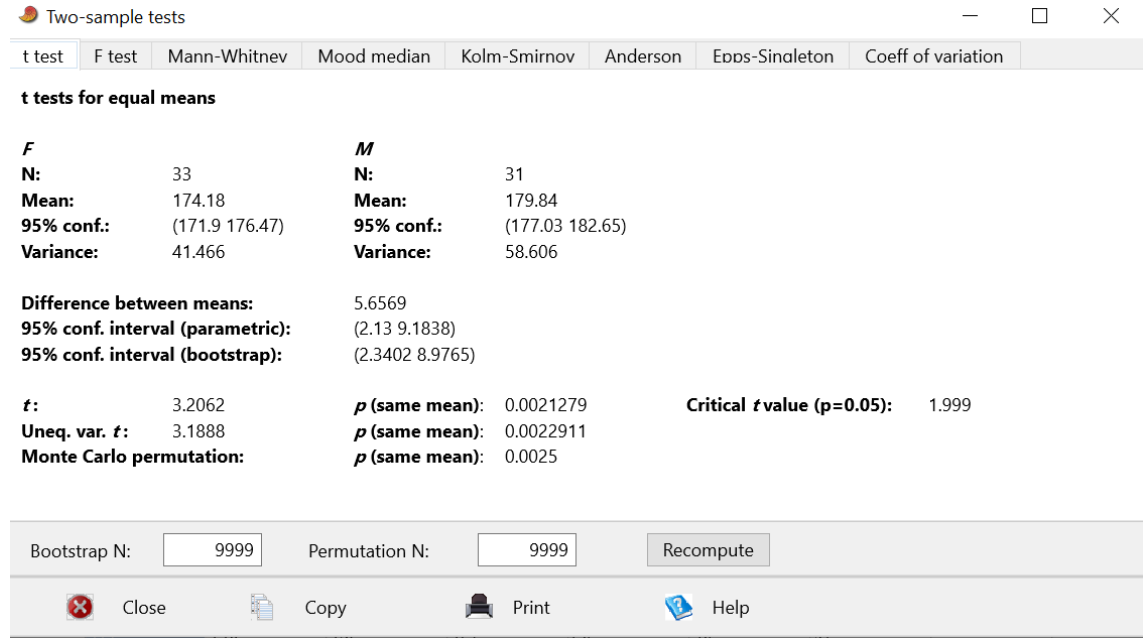
V případě různých rozptylů možno použít t-test se samostatnými odhady rozptylů

T-test

Univariate > Two sample tests > Two sample test...



Záložka F-test pro testování shody rozptylů



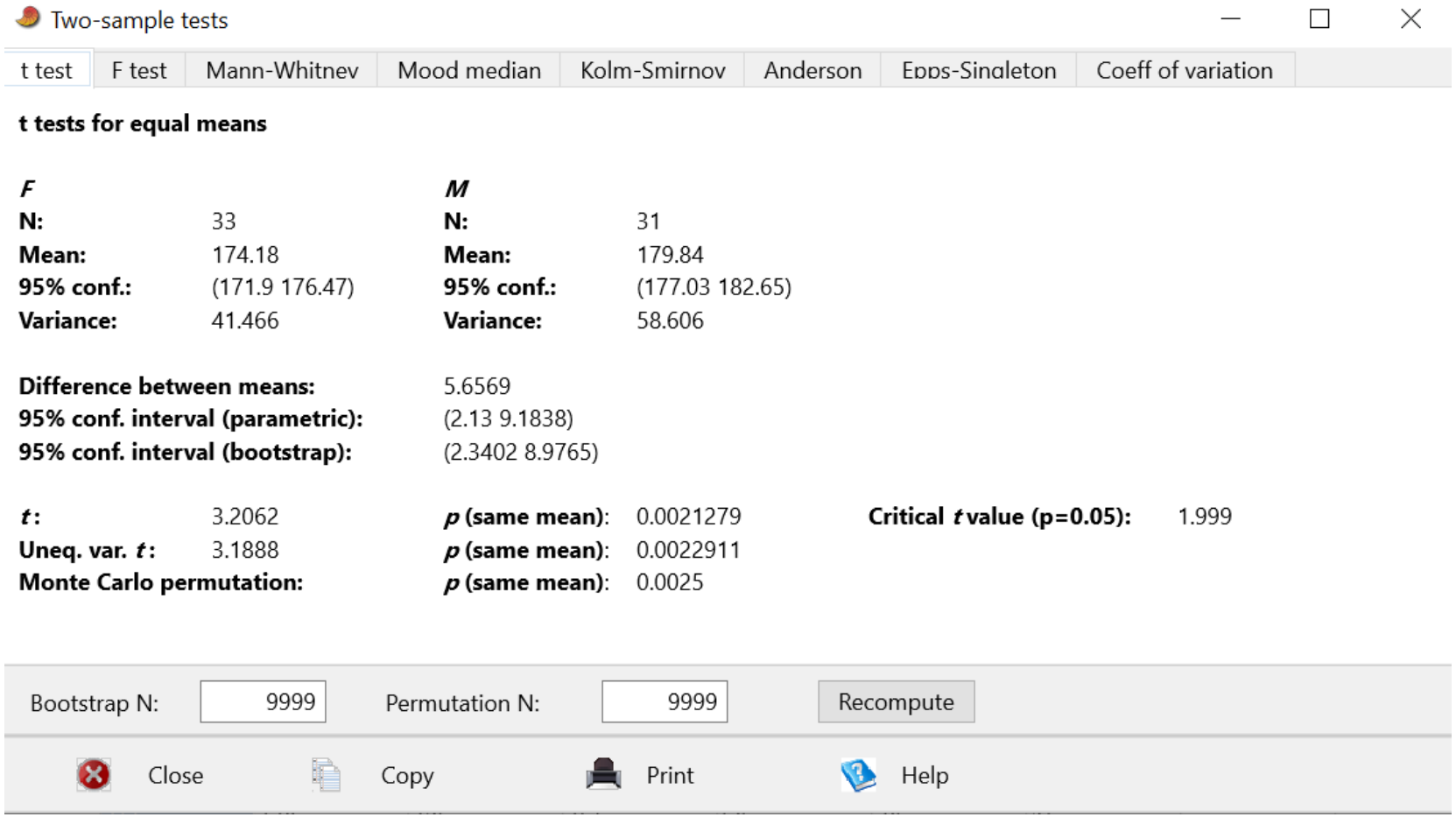
Záložka t-test se samotným testem (alternativně je zde i záložka s Mann-Whitney U testem, který je neparametrickou alternativou t-testu)

T-test – výstupy

Samotné výstupy testu – lze provést hromadně pro všechny zároveň

skupinové průměry a rozptyly

samotná statistika a permutační alternativy

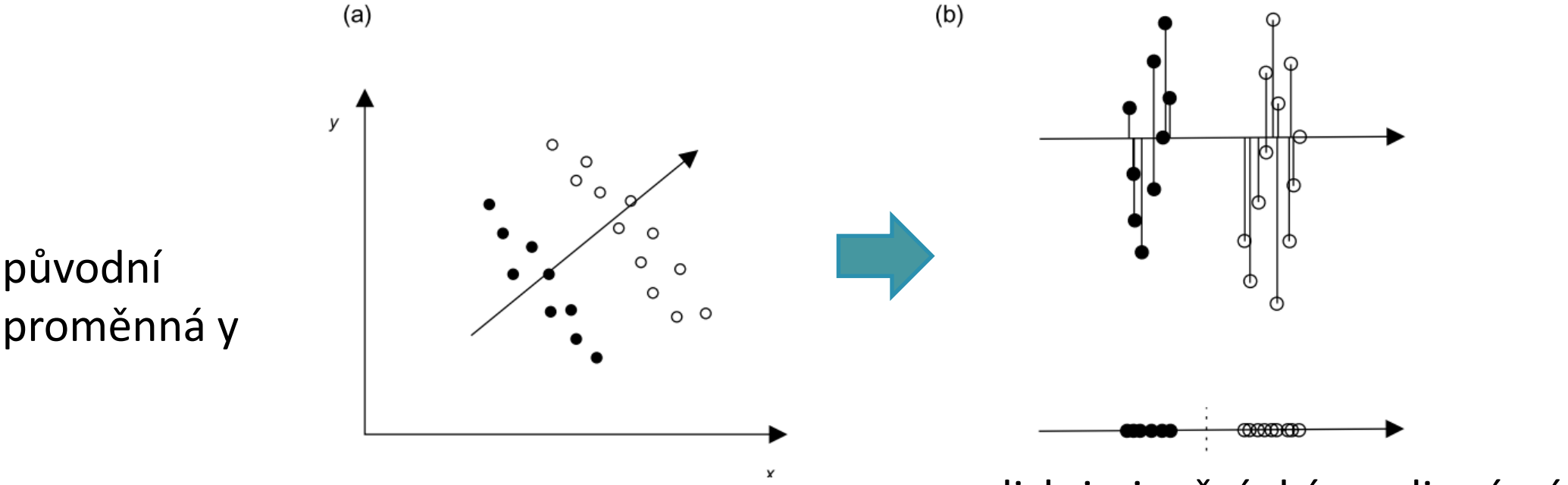


Diskriminační analýza

minimálně dvě nezávislé proměnné

predikční model pro odlišení mezi dvěma skupinami

Variabilita proměnných je zpracována s ohledem na předem dané (a priori známé) rozdělení do skupin.



původní proměnná y

původní proměnná x

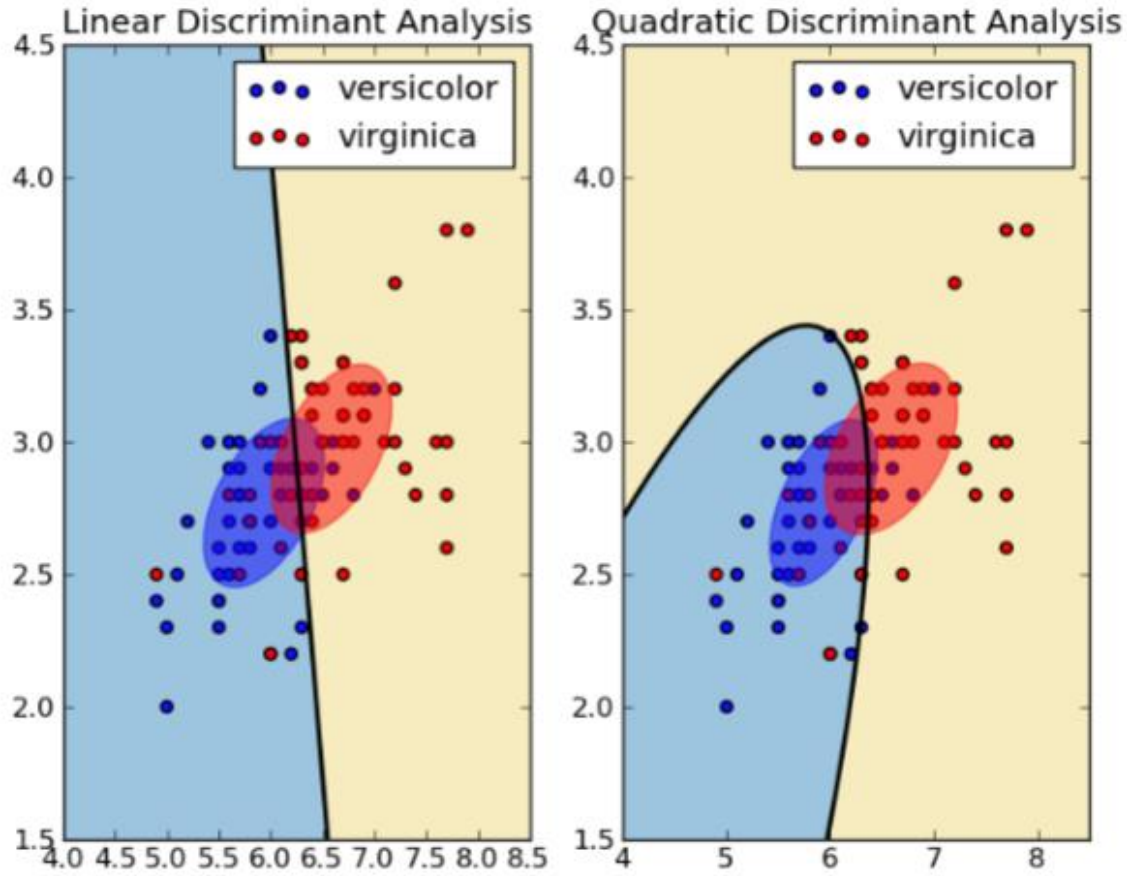
diskriminační skóre – lineární kombinace proměnných

Diskriminační analýza

minimálně dvě nezávislé proměnné

predikční model pro odlišení mezi dvěma skupinami

- založeno na **lineárním modelu**
- diskriminační skóre – lineární kombinace původních proměnných



Diskriminační analýza

Výstupy:

- diskriminační skóre pro každý případ
- **nestandardizované koeficienty pro každou proměnnou (použity v diskriminační rovnici)**

- **predikční pravidlo**

umožňuje přiřadit neznámému objektu regresní skóre a na základě jeho hodnoty jej zařadit do skupiny

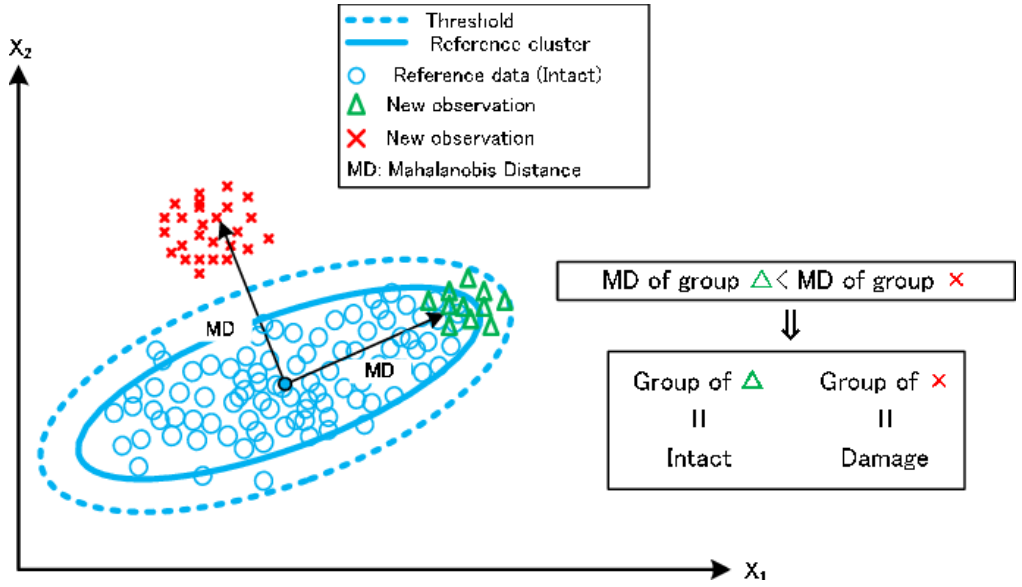
- standardizované koeficienty (vyjadřují podíl dané veličiny na diskriminačním skóre)
- spolehlivost pravidla
- Mahalanobisova vzdálenost
- aposteriorní pravděpodobnost
- spolehlivost klasifikace

White-Indian:

3.05(*Basion-prosthion*) — 1.04(*Glabello-occipital length*) — 5.41(*Maximum width*) + 4.29(*Basion-bregma height*) — 4.02(*Basion-nasion*) + 5.62(*Maximum diameter bi-zygomatic*) — 1.00(*Prosthion-nasion height*) — 2.19 (*Nasal breadth*).

Diskriminační analýza

Mahalanobisova vzdálenost
popisuje vzdálenost centroidů skupin (bere v úvahu korelaci mezi parametry a je nezávislá na jejich rozsahu)



apriorní pravděpodobnost –
pravděpodobnost, že objekt patří do dané skupiny daná například četností skupin

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

Diskriminační analýza

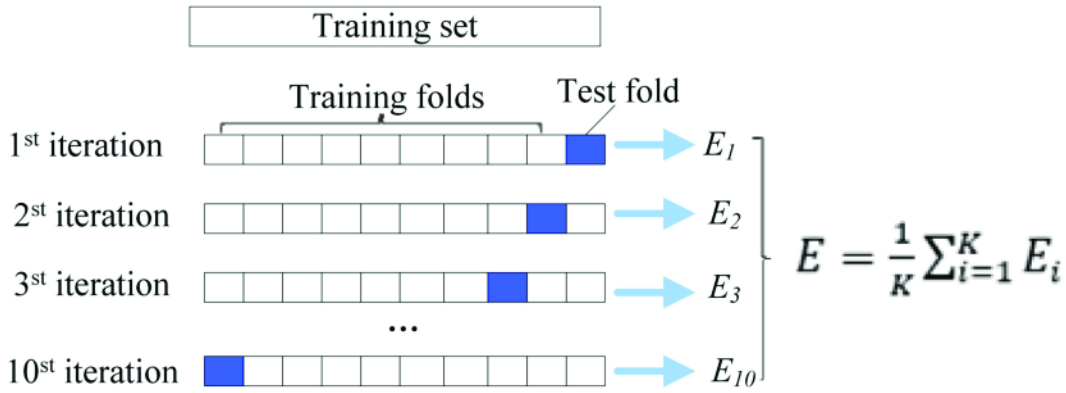
Spolehlivost zařazení případů do skupin na základě predikčního pravidla

each examining team.

			sex_real		Total female
			Female	Male	
Team B	Estimated sex	Female	20	1	21
		Male	1	28	29
	Total		21	29	50
Team A	Estimated sex	Female	19	0	19
		Male	2	29	31
	Total		21	29	50

Both teams achieved 96% accuracy (48 of 50) in determining the correct sex classification.

resubstituce



křížová validace (cross-validace)

NEJLÉPE

testování na nezávislém vzorku

Kanonická analýza

minimálně tři proměnné → predikční model pro odlišení mezi více **než dvěma skupinami**

Variabilita proměnných je zpracována s ohledem na předem dané (**a priori známé**) rozdělení do skupin – nové proměnné (kanonické osy), maximalizují rozdíly mezi skupinami.

Vlastnosti popsané původními proměnnými jsou převedeny na kanonické proměnné (k-1, kde k je počet skupin)

Pro každý prvek existuje hodnota kanonické proměnné – místo, kam dopadne na kanonické ose

Výstupy:
kanonické rovnice (k-1)
CS1 = a1x1 + b1x2 + c1x3 ... + C1
CS2 = a2x1 + b2x2 + c2x3 ... + C2

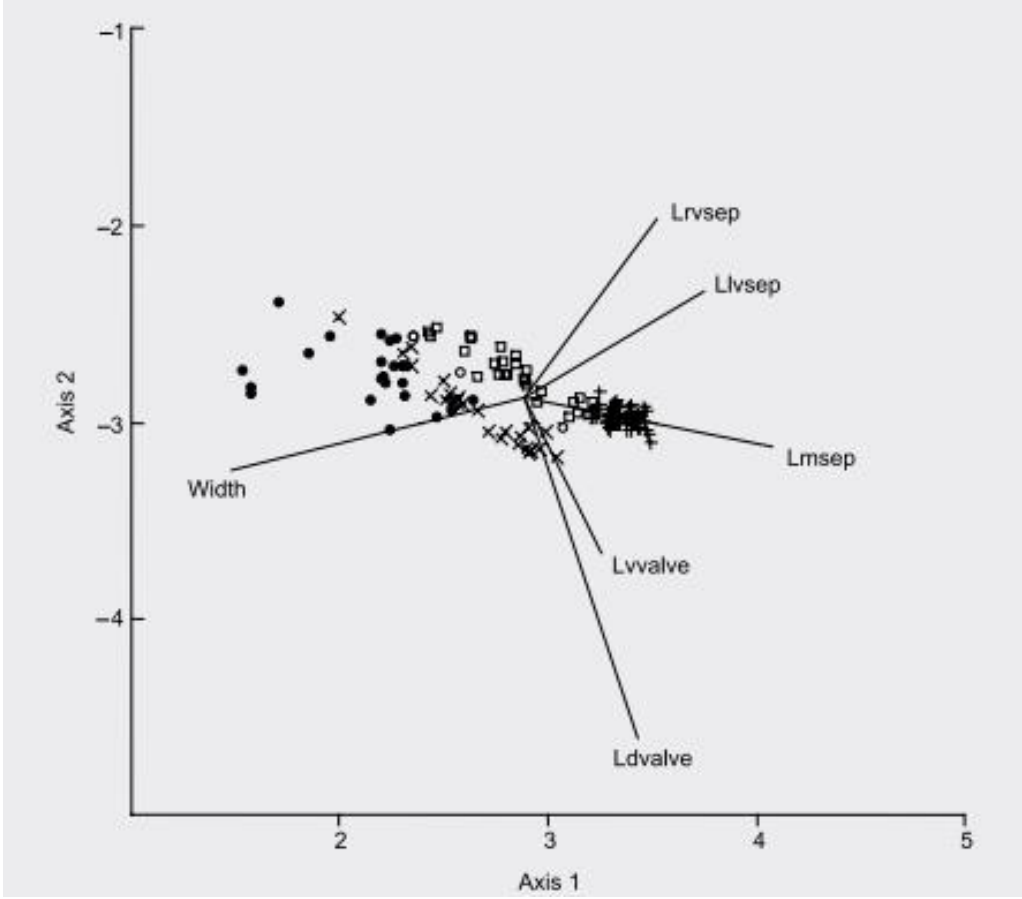
standardizované a nestandardizované koeficienty

Kanonická analýza

minimálně tři proměnné

predikční model pro odlišení mezi více **než dvěma skupinami**

Grafy – redukce proměnných na to „podstatné“



Diskriminační analýza

Jaké použít proměnné

význam mají pouze ty, které mají nějakou **souvislost s kategoriální proměnnou**

Hodnocení vztahu nezávislých proměnných a kategoriální proměnné

- t-test a ANOVA
- korelační analýza a XY grafy
- hlavní komponenty a faktorová analýza
- diskriminační analýza
- „expertní znalost proměnných“

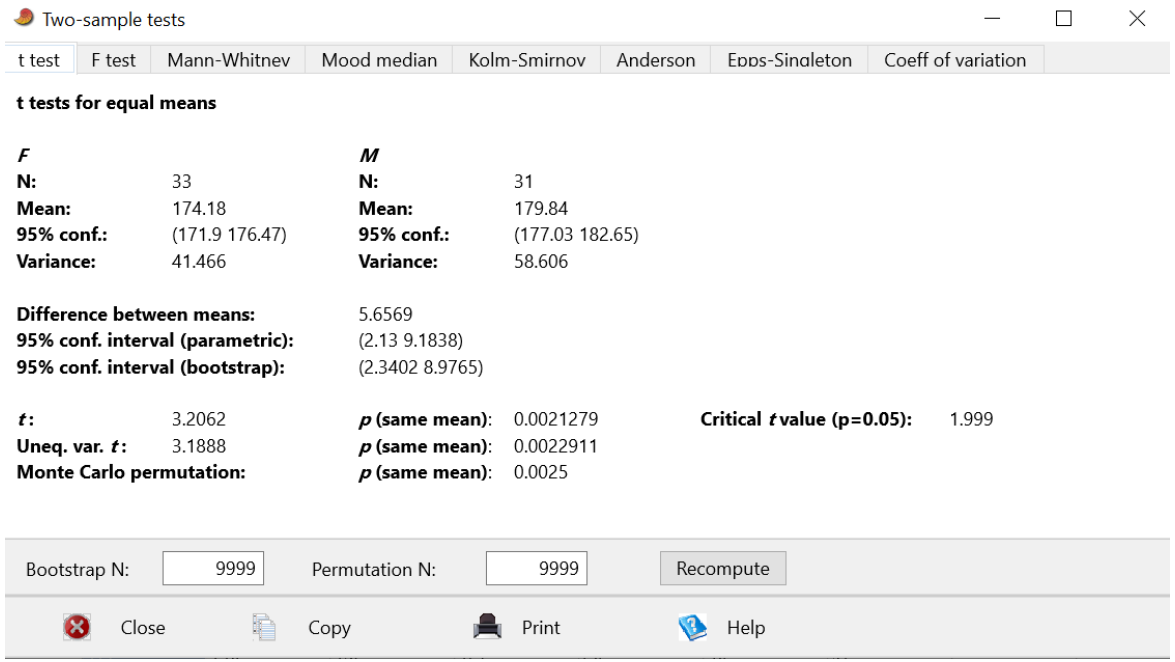
redundantní proměnné **snižují stabilitu modelu** a mohou vést k nesmyslným výsledkům

korelační analýza

Diskriminační analýza

Vztah ke kategoriální proměnné

Samostatný **t-test** pro jednotlivé proměnné – pro dvě skupiny!!

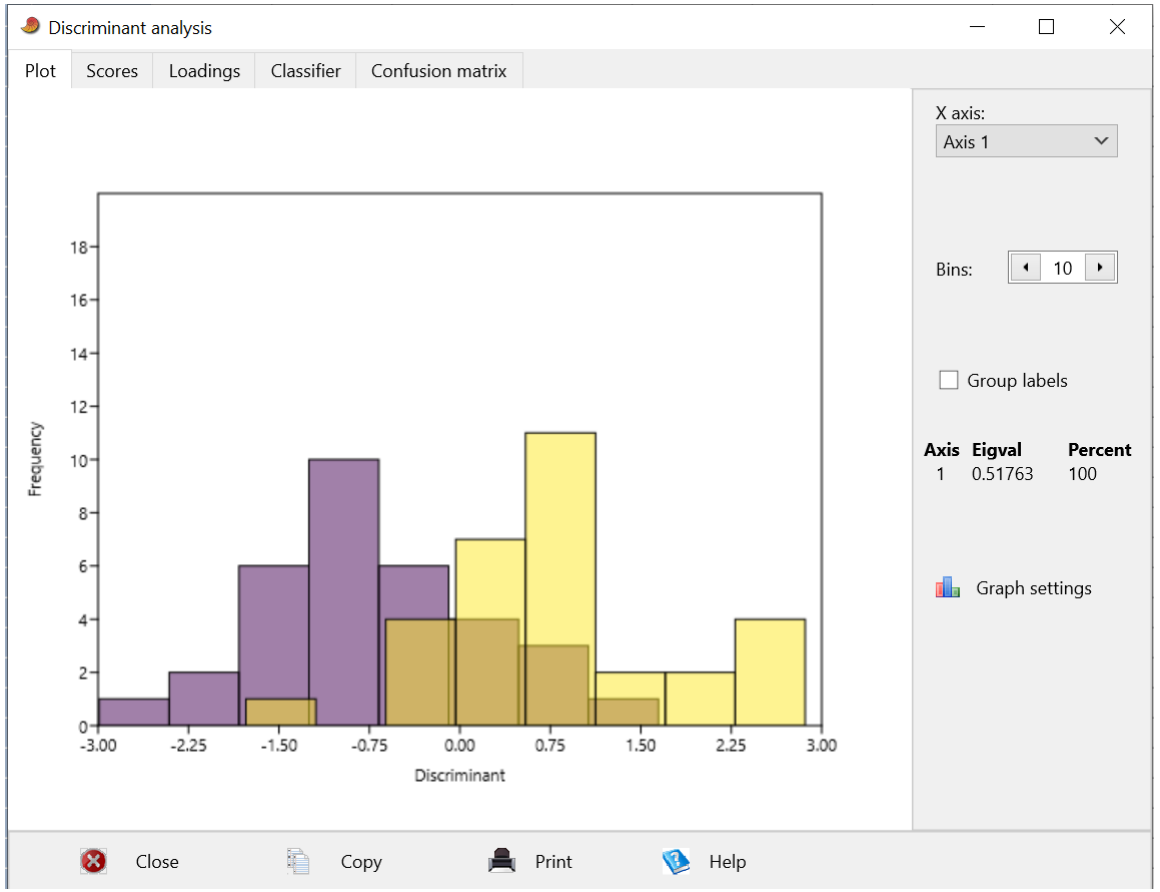


Obě analýzy mohou napovědět, ale diskriminace může uspět i díky kombinaci proměnných.

Diskriminační analýza

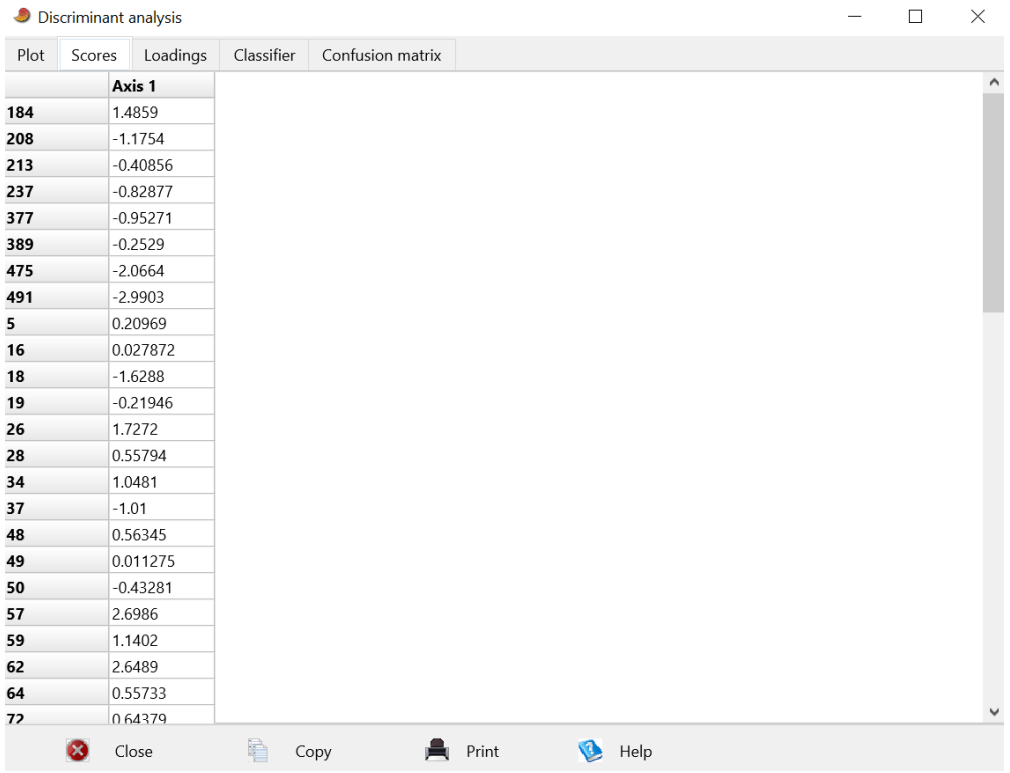
Multivariate > *Ordination* > *Discriminate* (při výběru jedné grupovací a jedné a více nezávislých proměnných) – při více kategoriích grupovací proměnné provede analýzu kanonických proměnných

histogram diskriminačního skóre pro obě skupiny



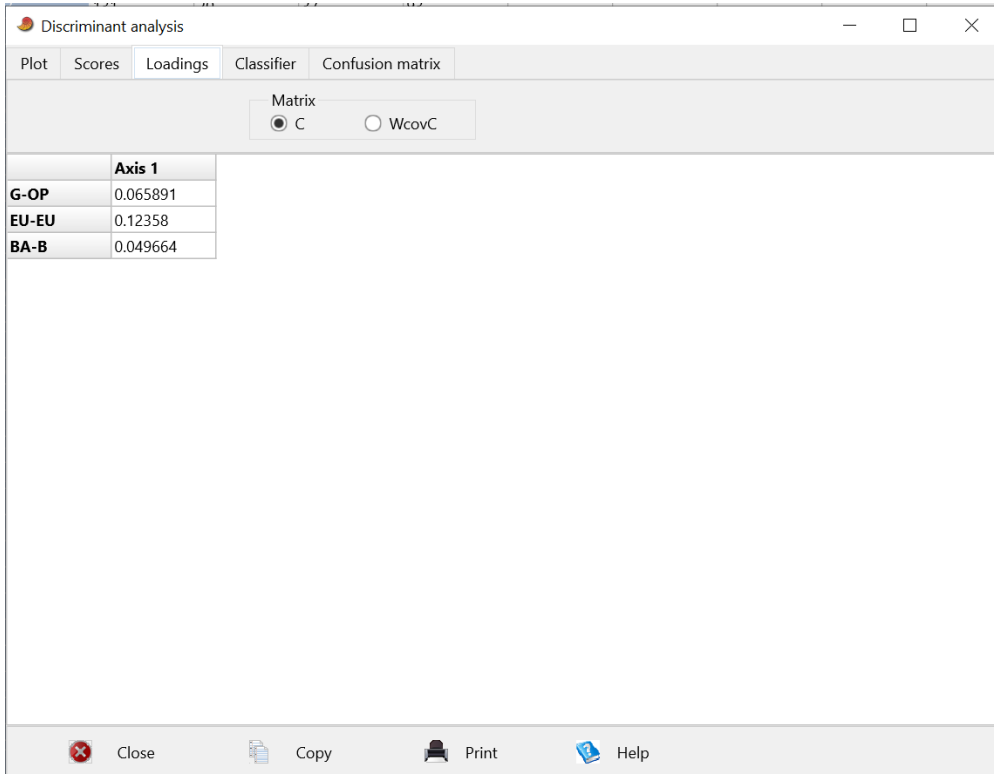
Eigenvalue – množství variability vyjádřené daným modelem (podíl SS v rámci skupiny a SS mezi skupinami)

Diskriminační analýza – číselný výstup analýzy



	Axis 1
184	1.4859
208	-1.1754
213	-0.40856
237	-0.82877
377	-0.95271
389	-0.2529
475	-2.0664
491	-2.9903
5	0.20969
16	0.027872
18	-1.6288
19	-0.21946
26	1.7272
28	0.55794
34	1.0481
37	-1.01
48	0.56345
49	0.011275
50	-0.43281
57	2.6986
59	1.1402
62	2.6489
64	0.55733
72	0.64379

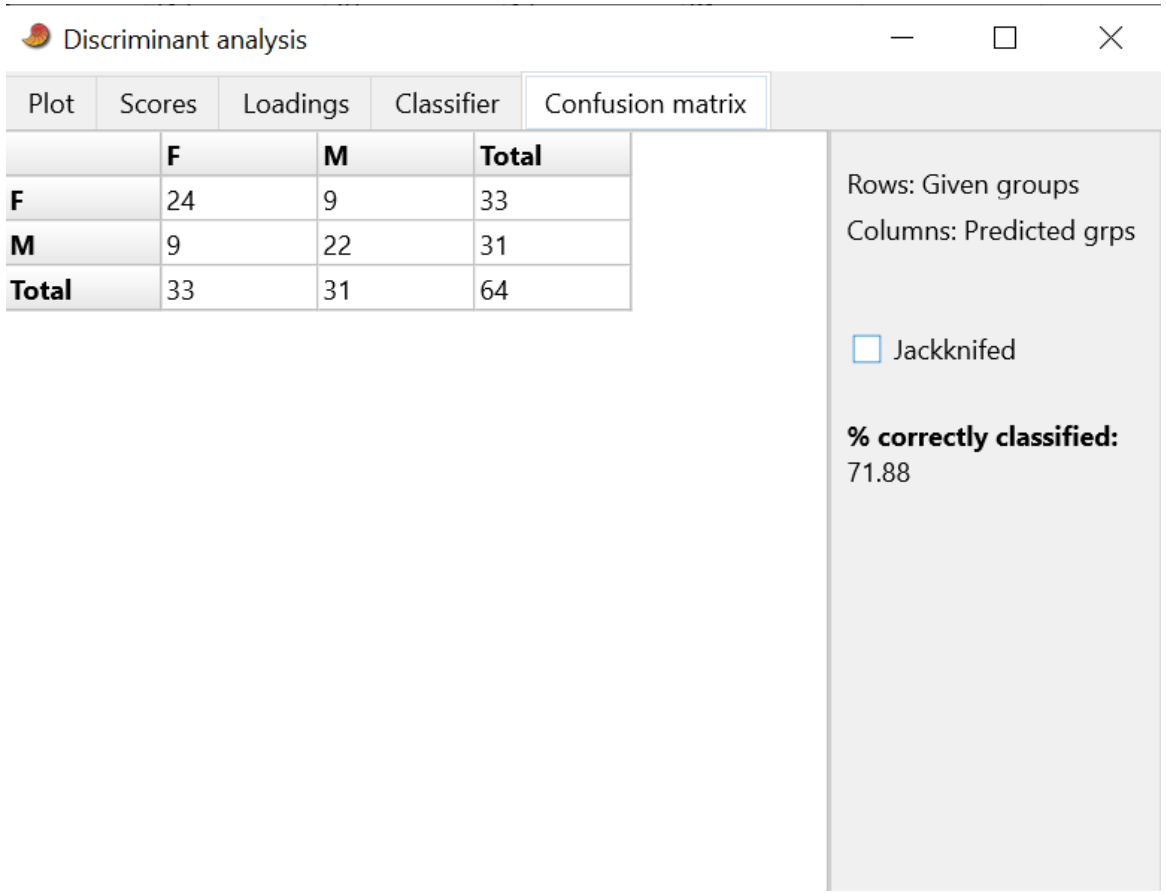
diskriminační skóre jednotlivých případů



	Axis 1
G-OP	0.065891
EU-EU	0.12358
BA-B	0.049664

Význam proměnných pro výpočet diskriminačního skóre

Diskriminační analýza – hodnocení klasifikačního kritéria



	F	M	Total
F	24	9	33
M	9	22	31
Total	33	31	64

klasifikační tabulka s procenty správně klasifikovaných případů (resubstituce)

Diskriminační analýza – výstupy v jiných programech – predikční pravidlo

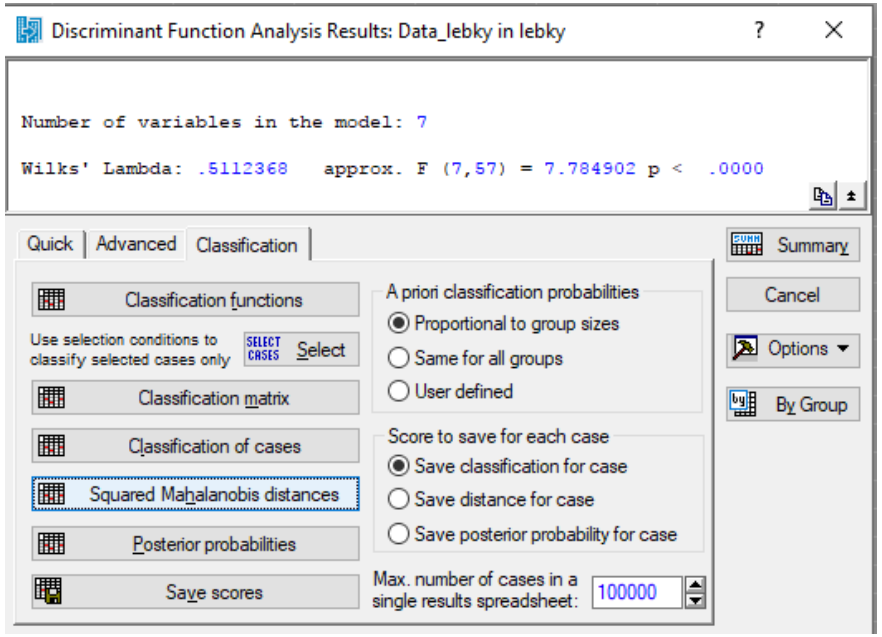
	Classification Functions; grouping: sex (list_1 in Data_lebky)	
Variable	F p=,50769	M p=,49231
ZYG-ZYG	3,362	3,577
RH-NS	2,157	2,503
ZM-ZM	1,972	2,065
Constant	-325,876	-373,301

rozepsané funkce
pro jednu a pro
druhou kategorii

V tomto případě se spočítají obě rovnice **a případ je klasifikován** do té skupiny, pro kterou je výsledek vyšší

Diskriminační analýza – další výstupy

Case	Observed Classif.	F p=.50769	M p=.49231
184	F	11.04751	17.18187
208	F	4.26211	7.58988
*213	F	6.96715	4.43218
237	F	3.90346	8.32681
377	F	1.85923	9.65731
389	F	4.18512	8.32306
475	F	4.78020	12.18686
491	F	8.18306	18.33574
5	M	4.54599	2.39293
*11	M	25.00021	26.27006
16	M	10.70175	5.22294
18	F	5.44417	15.17601
19	F	2.81365	4.45846
26	M	7.72133	6.64935
28	M	12.63816	7.46188



Case	Observed Classif.	F p=.50769	M p=.49231
	Posterior Probabilities (Data_lebky in lebk Incorrect classifications are marked with '		
184	F	0.956808	0.043192
208	F	0.844836	0.155164
*213	F	0.225009	0.774991
237	F	0.903997	0.096003
377	F	0.980731	0.019269
389	F	0.890878	0.109122
475	F	0.976663	0.023337
491	F	0.993983	0.006017
5	M	0.260040	0.739960
*11	M	0.660540	0.339460
16	M	0.062466	0.937534
18	F	0.992584	0.007416
19	F	0.701234	0.298766
26	M	0.376316	0.623684
28	M	0.071933	0.928067
34	M	0.454583	0.545417

Mahalanobisova vzdálenost
– vzdálenost od centroidů obou skupin

Aposteriorní pravděpodobnost
– pravděpodobnost, s jakou patří do obou skupin

Cohenova Kappa – nominální ale i ordinální data

Vyčísluje shodu mezi hodnotiteli.

Otázka: Do jaké míry přisuzují dva hodnotitelé dvě kategorie (nominální nebo ordinální).

$k = (p_o - p_e) / (1 - p_e)$

p0 je relativní shoda mezi řešiteli = (Both said Yes + Both said No) / (Total Ratings)
pe hypotetická šance náhodné shody

pe = p(„yes“) + p(„no“)
p („yes“) = (hodnocení yes hodnotitele 1/počet hodnocení) * (hodnocení yes hodnotitele 2/počet hodnocení)
p („no“) = (hodnocení no hodnotitele 1/počet hodnocení) * (hodnocení no hodnotitele 2/počet hodnocení)

Cohen's Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Fleiss Kappa – nominální ale i ordinální data

Vyčísluje shodu mezi hodnotiteli.

Otázka: Do jaké míry přisuzují dva hodnotitelé více kategorie (nominální nebo ordinální).

Korelace uvnitř třídy – intraclass correlation – ordinální data

Vyčísluje shodu mezi hodnotiteli.

Otázka: Do jaké míry přisuzují dva hodnotitelé dvě kategorie (nominální nebo ordinální).

ANOVA – analýza variance

Porovnává střední hodnoty pro více než dvě skupiny. pro dvě skupiny je výsledek stejný jako u t-testu.

Jednocestná analýza (One-way ANOVA): jeden **faktor** určující příslušnost ke skupině, například populační příslušnost, město narození

Otázka: Liší se nějak délka lebky u mužů, pocházejících z Evropské, Asijské a Americké populace

Předpoklady: uvnitř skupin je normální rozdělení (rozdělení reziduí) a skupiny mají stejný rozptyl (Bartlettův test)

Výstup: při zamítnutí víme, že se alespoň jedna skupina svými hodnotami liší. Která to je, je ale možná zjistit dalšími testy.