Biostatistika

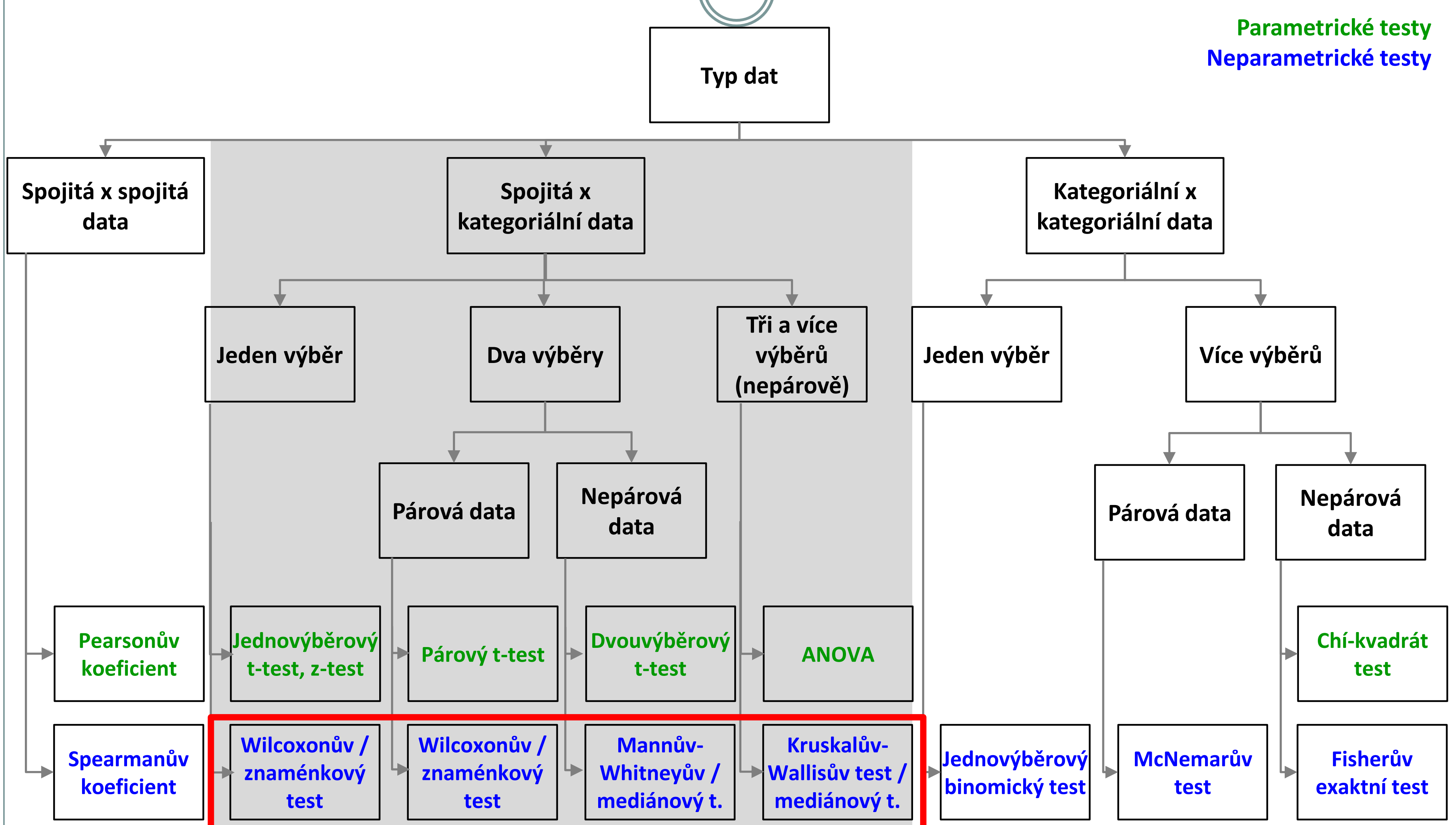
Opakování
Analýza kontingenčních tabulek
Základy korelační analýzy

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek

Co byste měli umět z minula:

1. Určit, kdy je vhodné použít pro testování hypotéz parametrické a neparametrické testy – ověřování předpokladů.
2. Vybrat typ neparametrického testu – jednovýběrový, párový nebo dvouvýběrový?
3. Provést testování v softwaru Statistica – Wilcoxonův test, znaménkový test, Mannův-Whitneyho test, Kruskalův-Wallisův test.
4. Interpretovat výsledky testování.

Základní rozhodování o výběru statistických testů - co budeme probírat dnes



Analýza kontingenčních tabulek

Kontingenční tabulky

Pearsonův chí-kvadrát test (test dobré shody)

Fisherův exaktní test

McNemarův test

Kontingenční tabulka - opakování

- Frekvenční sumarizace dvou kategoriálních proměnných (binárních, nominálních nebo ordinálních proměnných).
- Obecně: **R x C kontingenční tabulka** (R – počet kategorií jedné proměnné, C – počet kategorií druhé proměnné).
- Speciální případ: 2 x 2 tabulka = čtyřpolní tabulka.
- Kontingenční tabulky: **absolutních četností, celkových procent, řádkových/sloupcových četností**

- Př.: Sumarizace vyšetřených osob podle pohlaví a výsledku diagnostického testu.

	Výsledek vyšetření		
Pohlaví	Nemocný	Zdravý	**Celkem**
Muž	45	11	56
Žena	25	6	31
Celkem	70	17	87

Ukázka kontingenční tabulky

- **Vztah pohlaví a výskytu onemocnění** (pozor na hodnocení nesmyslného vztahu)

	Nemocný	Zdravý	Celkem
Muž	**a**	**b**	**a + b**
Žena	**c**	**d**	**c + d**
Celkem	**a + c**	**b + d**	**a + b + c + d = N**

Marginální absolutní četnost

Celkový počet hodnot

Simultánní absolutní četnost

	Nemocný	Zdravý	Celkem
Muž	45	11	56
Žena	25	6	31
Celkem	70	17	87

Jsou více nemocní muži nebo ženy?

Kontingenční tabulky – procenta

Kontingenční tabulka absolutních četností

Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	1	7	176	46	230
MCI	13	85	201	107	406
AD	9	34	90	64	197
Celkem	23	126	467	217	833

Kontingenční tabulka řádkových procent

Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	0,4	3,0	76,5	20,0	100,0
MCI	3,2	20,9	49,5	26,4	100,0
AD	4,6	17,3	45,7	32,5	100,0
Celkem	2,8	15,1	56,1	26,1	100,0

Kontingenční tabulka sloupcových procent

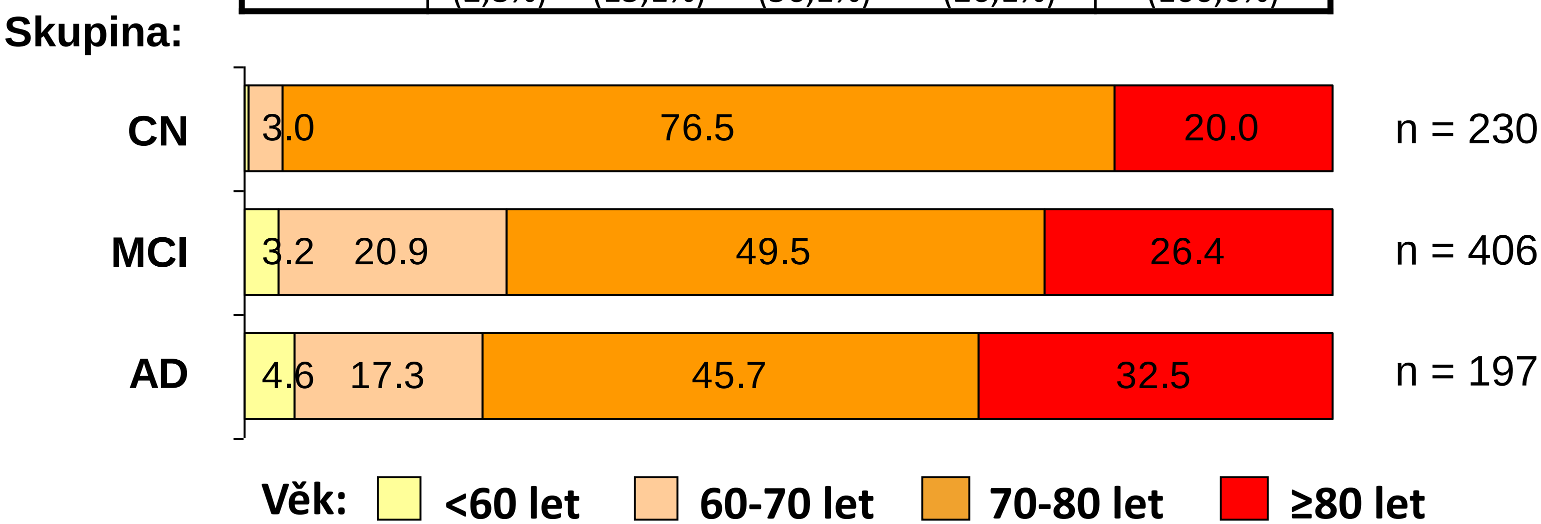
Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	4,3	5,6	37,7	21,2	27,6
MCI	56,5	67,5	43,0	49,3	48,7
AD	39,1	27,0	19,3	29,5	23,6
Celkem	100,0	100,0	100,0	100,0	100,0

Kontingenční tabulka celkových procent

Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	0,1	0,8	21,1	5,5	27,6
MCI	1,6	10,2	24,1	12,8	48,7
AD	1,1	4,1	10,8	7,7	23,6
Celkem	2,8	15,1	56,1	26,1	100,0

Kontingenční tabulky – popis a vizualizace

Skupina	Věk <60 let	Věk 60-70 let	Věk 70-80 let	Věk ≥80 let	Celkem
CN	1 (0,4%)	7 (3,0%)	176 (76,5%)	46 (20,0%)	230 (100,0%)
MCI	13 (3,2%)	85 (20,9%)	201 (49,5%)	107 (26,4%)	406 (100,0%)
AD	9 (4,6%)	34 (17,3%)	90 (45,7%)	64 (32,5%)	197 (100,0%)
Celkem	23 (2,8%)	126 (15,1%)	467 (56,1%)	217 (26,1%)	833 (100,0%)



Co analyzujeme u kontingenčních tabulek?

- Analýza kontingenčních tabulek umožňuje analyzovat **vazbu mezi dvěma kategoriálními proměnnými**. Základním způsobem testování je tzv. chí-kvadrát test, který **srovnává pozorované četnosti kombinací kategorií oproti očekávaným četnostem**, které vychází z teoretické situace, kdy je vztah mezi proměnnými náhodný.

- Test dobré shody je využíván také pro **srovnání pozorovaných četností proti očekávaným četnostem daných určitým pravidlem** (typickým příkladem je Hardy-Weinbergova rovnováha v genetice).

- Specifickým typem výstupů odvozených z kontingenčních tabulek jsou tzv. **poměry šancí** a **relativní rizika**, využívaná často v medicíně pro identifikaci a popis rizikových skupin pacientů.

Test dobré shody - základní teorie

Testová statistika:

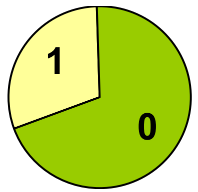
$$\chi^2 = \sum \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$

$$\chi^2 = \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{1. jev}} + \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{2. jev}} + \dots$$

$$\boxed{\chi^2 > \chi^2_{(1-\alpha)}(s.v.) \quad \text{... zamítáme } H_0}$$

1 - hladina významnosti *stupně volnosti*

Test dobré shody: příklad I



Binomické jevy (1/0)

$$\chi^2_{(1)} = \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{I. jev 1}} + \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{II. jev 2}}$$

Příklad

✓ **10 000 lidí hází mincí** → **rub: 4 000 případů (R)**
→ **líc: 6 000 případů (L)**

? **Lze výsledek považovat za statisticky významně odlišný (nebo neodlišný) od očekávaného poměru R : L = 1 : 1 (tzn. že je výsledek hodu mincí náhodný)?**

$$\chi^2 = \frac{(4000-5000)^2}{5000} + \frac{(6000-5000)^2}{5000} = 400$$

Tabulková hodnota: $\chi^2_{(0,95)}(\nu = k-1 = 1) = \underline{3{,}84} \quad (0{,}95 = 1-\alpha)$

Rozdíl je vysoce statisticky významný (p < 0,001)

Test dobré shody: příklad II

Celkem bylo zkoumáno 250 semen určitého druhu rostliny a roztříděno do následujících kategorií: žluté/hladké; žluté/vrásčité; zelené/hladké; zelené/vrásčité. Předpokládaný poměr výskytu těchto kategorií v populaci je **9 : 3 : 3 : 1**. Následující tabulka obsahuje původní data z pozorování a dále postup při testování H_0.

	žluté/hladké	žluté/vrásčité	zelené/hladké	zelené/vrásčité	n
$f_{poz.}$	152	39	53	6	250
$f_{oček.}$	140,6250	46,8750	46,8750	15,6250	

ν = k - 1 = 3

$$\chi^2 = \frac{11{,}3750^2}{140{,}6250} + \frac{7{,}8750^2}{46{,}8750} + \frac{6{,}1250^2}{46{,}8750} + \frac{9{,}6250^2}{15{,}6250} = 8{,}972$$

Tabulková hodnota: $\chi^2_{(0,95)}(\nu = k - 1 = 3) = \underline{7{,}81} \quad (0{,}95 = 1 - \alpha)$

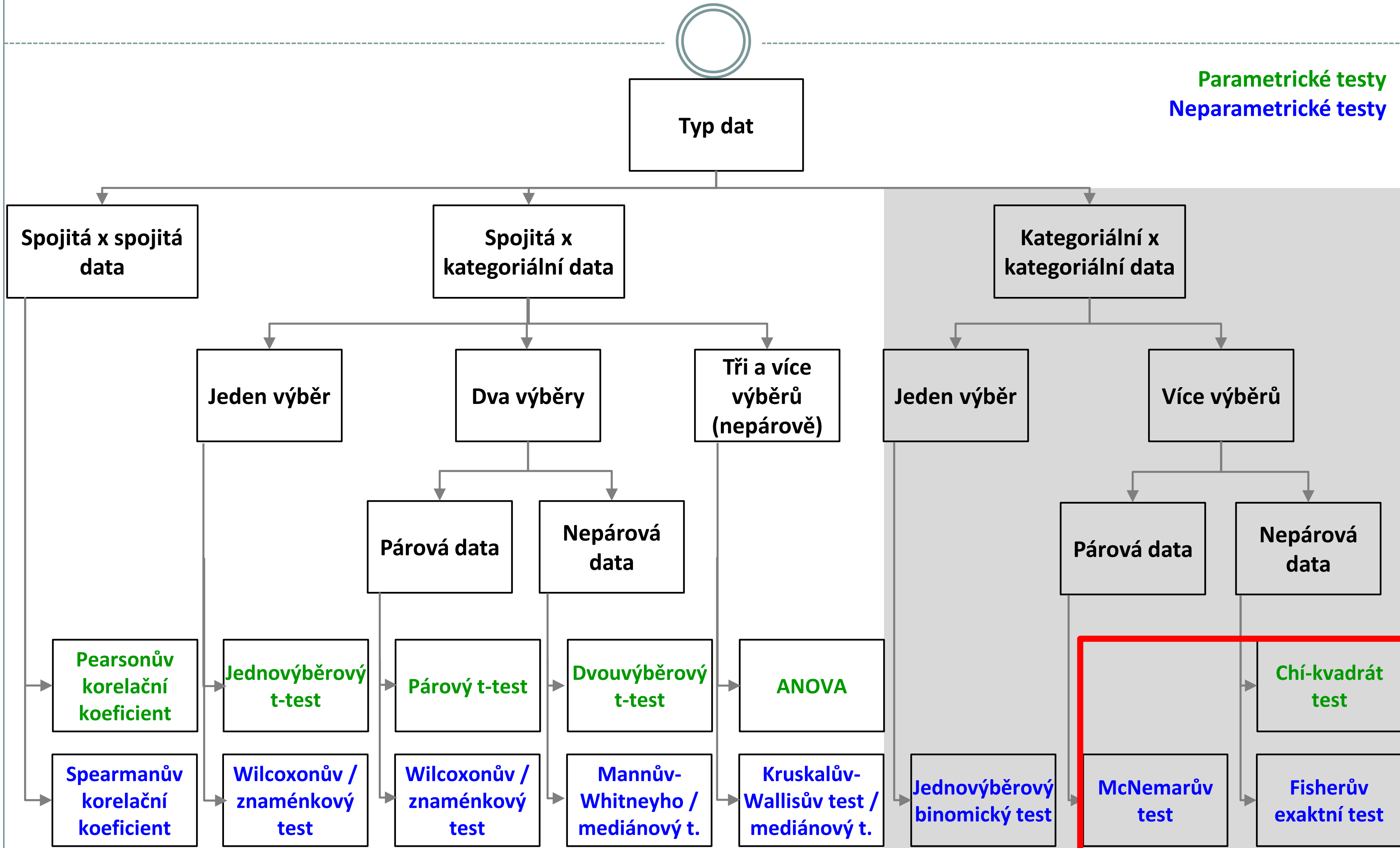
Zamítáme hypotézu shody pozorovaných četností s očekávanými

Kontingenční tabulka - hypotézy

- **NEZÁVISLOST** (Pearsonův chí-kvadrát test, Fisherův exaktní test)
  - Jeden výběr, 2 charakteristiky – obdoba nepárového uspořádání
  - Např.: existence vztahu mezi barvou očí a známkou z biostatistiky u studentů
- **SHODA STRUKTURY** (Pearsonův chí-kvadrát test, Fisherův exaktní test)
  - Tzv. test homogenity
  - Více výběrů, jedna charakteristika – obdoba nepárového uspořádání
  - Např.: pohlaví pacientů s diabetem v *K* nemocnicích (tj. *K* výběrů)
- **SYMETRIE** (McNemarův test)
  - Jeden výběr, opakovaně jedna charakteristika – obdoba párového uspořádání
  - Např.: posouzení stavu stromů ve dvou sezónách

Základní rozhodování o výběru statistických testů - analýza kontingenčních tabulek

Parametrické testy
Neparametrické testy



Kontingenční tabulka - obecně

- Máme dvě nominální veličiny, X (má r variant) a Y (má s variant)
- Kontingenční tabulka typu **r** x **s**

$x_{[j]}$ \ $y_{[k]}$	$y_{[1]}$	.....	.....	$y_{[s]}$	$n_{j.}$
$x_{[1]}$	n_{11}	.....	.....	n_{1s}	$\boxed{n_{1.}}$
.	.	.....	.....	.	.
.	.	.....	.....	.	.
$x_{[r]}$	n_{r1}	.....	.....	$\boxed{n_{rs}}$	$n_{r.}$
$n_{.k}$	$\boxed{n_{.1}}$	.	.	$n_{.s}$	n

Marginální absolutní četnost

Marginální absolutní četnost

Simultánní absolutní četnost

- Označení:

n_{jk}- simultánní absolutní četnost,

$n_{j.}$- marginální absolutní četnost

Testování nezávislosti – Pearsonův chí-kvadrát test

- Souvisí spolu výskyt dvou nominálních znaků měřených na jediném výběru?
- Příklad: Souvisí barva očí (modrá, zelená, hnědá) a barva vlasů (hnědá, černá, blond) u vybraných 30 studentů?
- **Nulová hypotéza:** Znaky X a Y jsou nezávislé náhodné veličiny.
- **Alternativní hypotéza:** Znaky X a Y jsou závislé náhodné veličiny.
- Test: **Pearsonův chí-kvadrát**

H_0 **platí**

$$K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(n_{jk} - e_{jk})^2}{e_{jk}} \approx \chi^2\left((r-1)(s-1)\right)$$

Očekávané (teoretické) četnosti e_{jk}: $e_{jk} = \frac{n_{j.} n_{.k}}{n}$

- H_0 zamítáme na hladině významnosti α, pokud $K \geq \chi^2_{1-\alpha}\left((r-1)(s-1)\right)$

- **Předpoklady testu ?**

Testování nezávislosti – Pearsonův chí-kvadrát test

- **Předpoklady Pearsonova chí-kvadrát testu**:

1. **Jednotlivá pozorování** shrnutá v kontingenční tabulce **jsou nezávislá**, tj. každý prvek patří jen do jedné buňky kont. tabulky, nemůže zároveň patřit do dvou.
2. **Podmínky dobré aproximace**: Očekávané (teoretické) četnosti jsou aspoň v 80 % případů větší nebo rovné 5 a ve 100 % případů nesmí být pod 2 (pokud není tento předpoklad splněn, je vhodné sloučit kategorie s nízkými četnostmi, ale tyto kategorie musí být slučitelné!).

- **Měření síly závislosti:**

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}, kde \quad m = \min\{r,s\}, V \;\; je \;\; z \; interval u \; (0,1)$

Význam hodnot: 0-0,1….zanedbatelná závislost

0,1-0,3…slabá závislost

0,3-0,7…střední závislost

0,7-1 silná závislost

Testování nezávislosti – Pearsonův chí-kvadrát test

- **Podmínka dobré aproximace – alespoň 80% buněk musí mít očekávanou (teoretickou) četnost větší než 5 – pro různě velké tabulky**:

tabulka 2x3 či 3x2 – alespoň 5 z 6 buněk
tabulka 2x4 či 4x2 – alespoň 7 z 8 buněk
tabulka 2x5 či 5x2 – alespoň 8 z 10 buněk
tabulka 3x3 – alespoň 8 z 9 buněk
tabulka 3x4 či 4x3 – alespoň 10 z 12 buněk
tabulka 3x5 či 5x3 – alespoň 12 z 15 buněk
tabulka 4x4 – alespoň 13 z 16 buněk
tabulka 4x5 či 5x4 – alespoň 16 z 20 buněk
tabulka 5x5 – alespoň 20 z 25 buněk

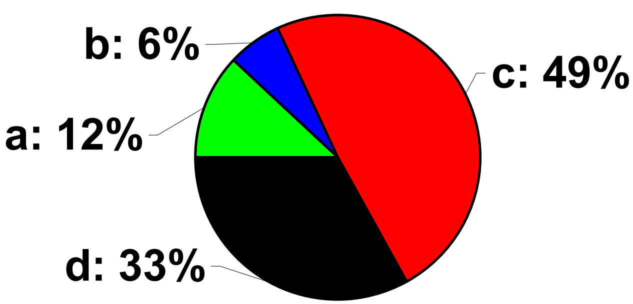
Kontingenční tabulky: příklad

gen \ †	Ano	Ne	Σ
Ano	20	82	102
Ne	10	54	64
Σ	30	136	166

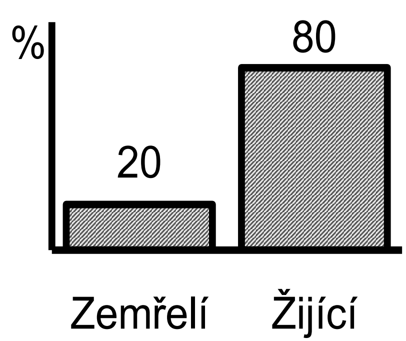
$F_A = 102 * 30 / 166 = 18{,}43$
$F_B = 102 * 136 / 166 = 83{,}57$
$F_C = 11{,}57$
$F_D = 52{,}43$

$$\chi^2_{(1)} = \frac{(20-18{,}43)^2}{18{,}43} + \frac{(82-83{,}57)^2}{83{,}57} + \frac{(10-11{,}57)^2}{11{,}57} + \frac{(54-52{,}43)^2}{52{,}43} = 0{,}423 \qquad 0{,}423 < \chi^{2\,(1)}_{0{,}95} = 3{,}84$$

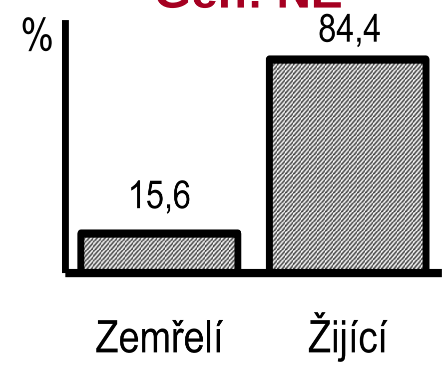
Kontingenční tabulka v obrázku



Gen: ANO



Gen: NE

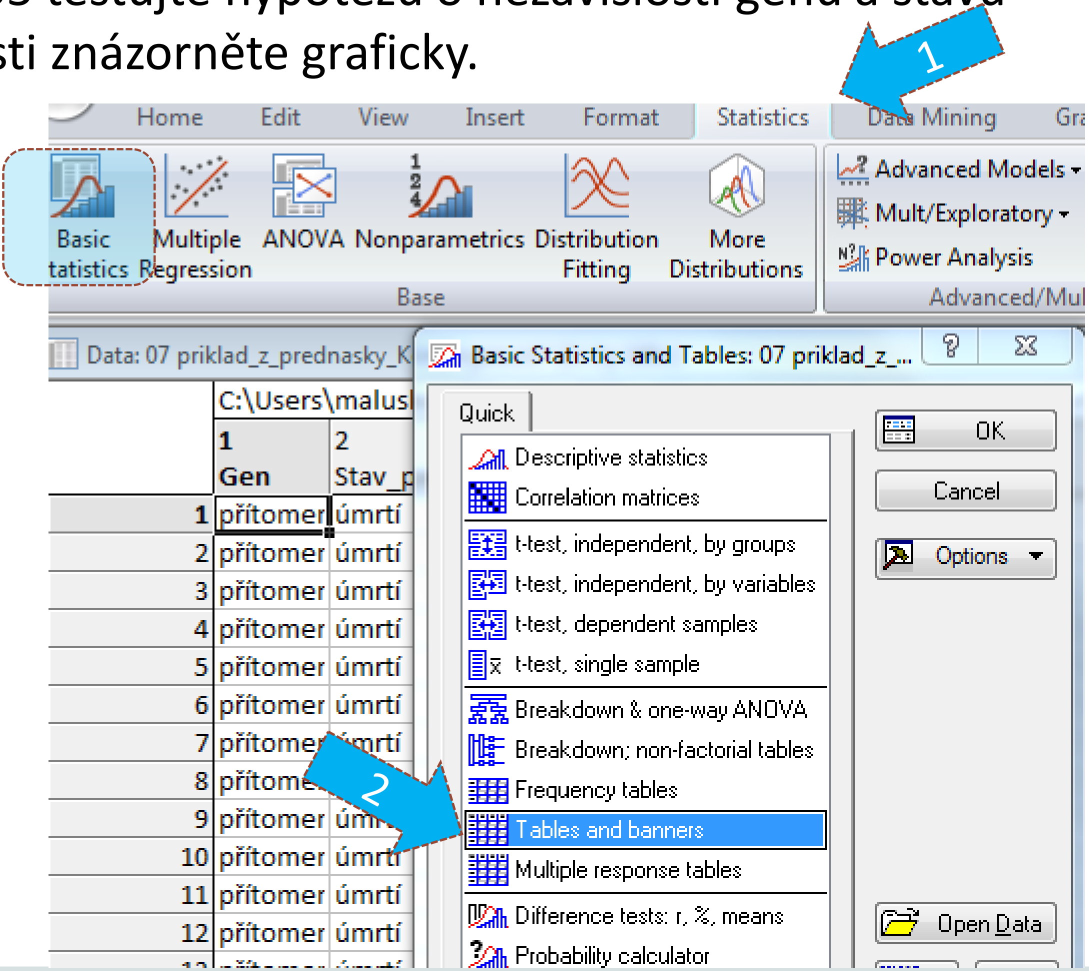


Řešení v softwaru Statistica

- Datový soubor může být zadán 2 způsoby:
    - **Původní data** (co řádek, to subjekt charakterizovaný danými kategoriálními proměnnými),
    - **Agregovaná data** (kontingenční tabulka, četnosti všech kombinací kategorií 2 kategoriálních proměnných) – analýza agregovaných dat možná i pomocí webových kalkulátorů.

Způsob 1: Řešení v softwaru Statistica I

- Na hladině významnosti 0,05 testujte hypotézu o nezávislosti genu a stavu pacienta. Simultánní četnosti znázorněte graficky.

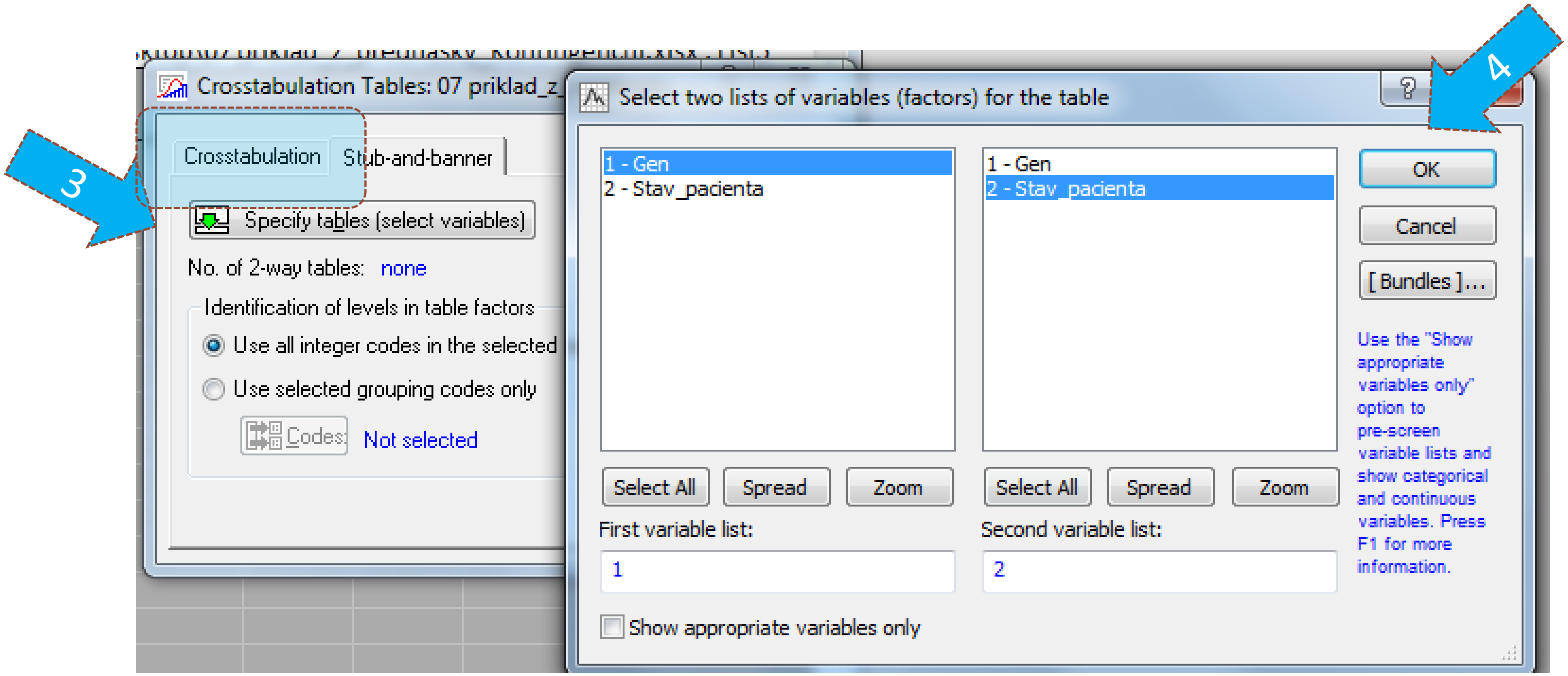


- **Původní datový soubor** (co řádek, to subjekt)

- V menu ***Statistics*** zvolíme ***Basic statistics,*** Vybereme ***Tables and banners*** (v češtině ***Kontingenční tabulky***)

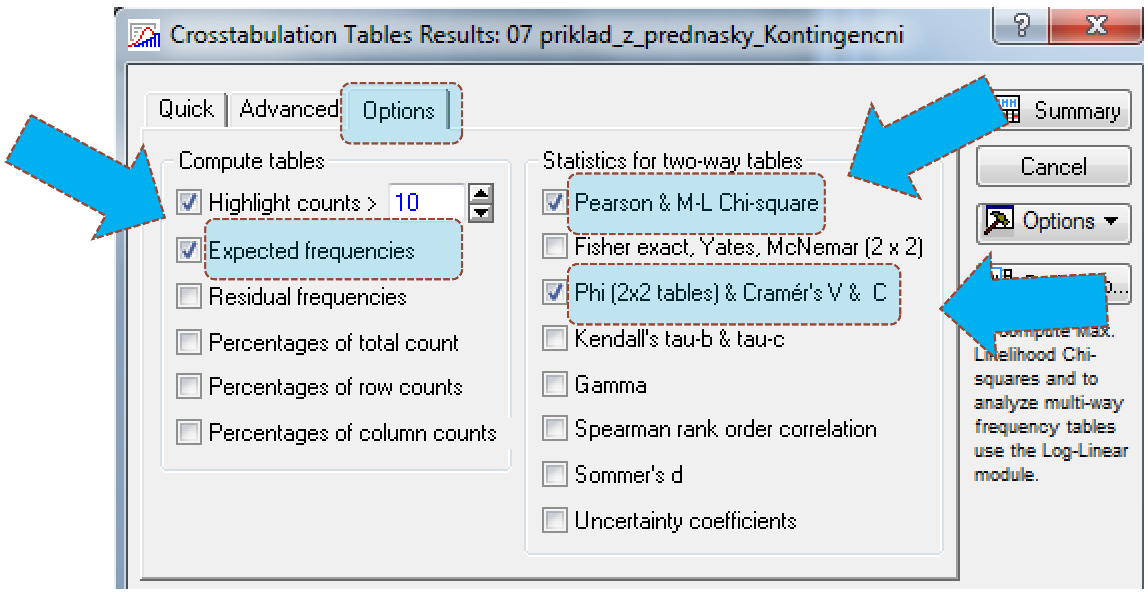
Způsob 1: Řešení v softwaru Statistica II

- Vybereme proměnné, které chceme testovat



Způsob 1: Řešení v softwaru Statistica III

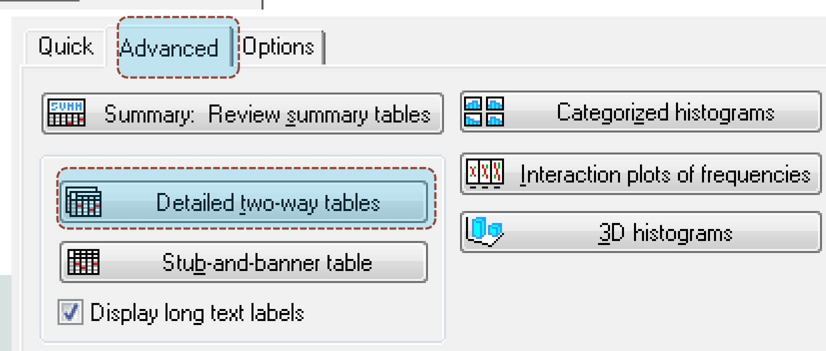
• Na záložce **Options** zaškrtneme ***Expected frequencies*** (***Očekávané četnosti***) (k ověření podmínek dobré aproximace)



• Zaškrtneme Pearsonův chí-kvadrát

• Pokud chceme vypočítat i Cramérův koeficient zaškrtneme Phi & Cramer‘s V

• Poté se vrátíme na záložku **Advanced,** kde a zvolíme ***Detailed two-way tables***



Způsob 1: Řešení v softwaru Statistica IV

Tab.1: Pozorované četnosti

Summary Frequency Table (07 priklad_z_prednasky_K
Marked cells have counts > 10
(Marginal summaries are not marked)

Gen	Stav_pacienta úmrtí	Stav_pacienta žijící	Row Totals
přítomen	20	82	102
nepřítomen	10	54	64
All Grps	30	136	166

Tab. 2: Očekávané četnosti

Summary Table: Expected Frequencies (07 priklad_z_pre
Marked cells have counts > 10
Pearson Chi-square: ,421322, df=1, p=,516278

Gen	Stav_pacienta úmrtí	Stav_pacienta žijící	Row Totals
přítomen	18,43373	83,5663	102,0000
nepřítomen	11,56627	52,4337	64,0000
All Grps	30,00000	136,0000	166,0000

Jsou splněny podmínky dobré aproximace?

Tab. 3: Pearsonův chí-kvadrát

Hodnota testové statistiky

Počet stupňů volnosti

p- hodnota

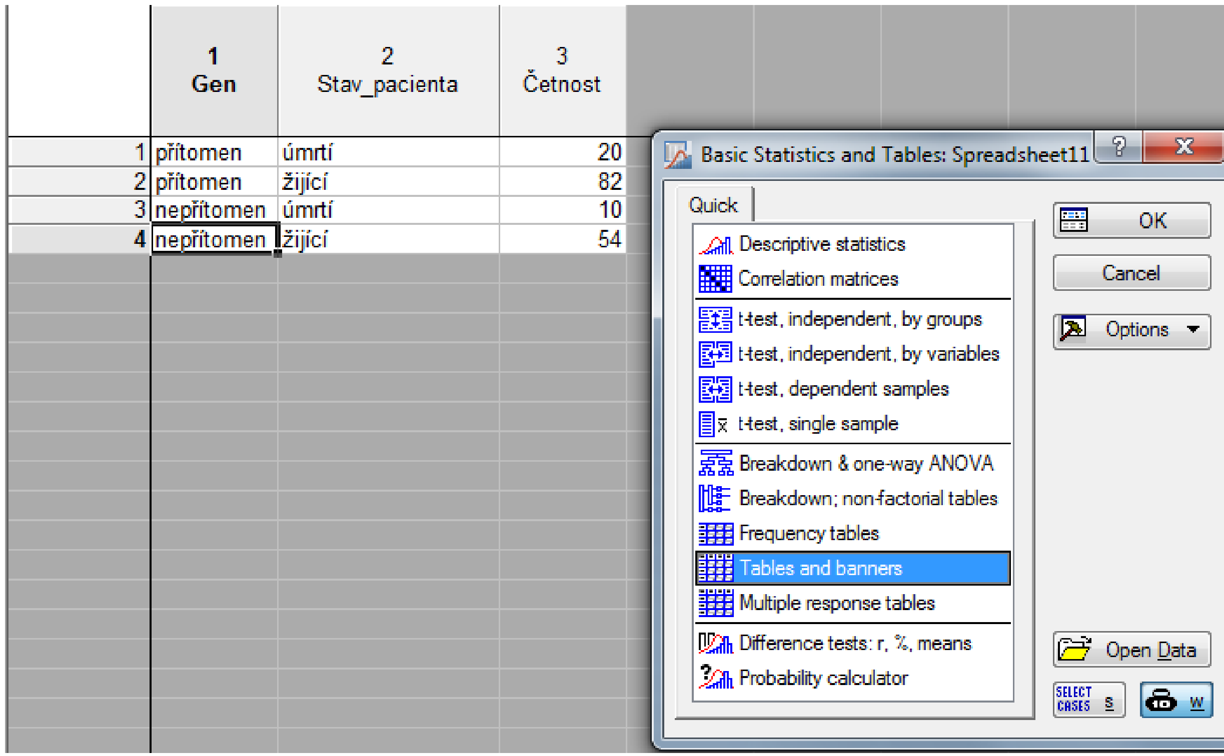
Statistic	Statistics: Gen(2) x Stav_pacienta(2) Chi-square	df	p
Pearson Chi-square	,4213223	df=1	p=,51628
M-L Chi-square	,4277117	df=1	p=,51311
Phi for 2 x 2 tables	,0503794		
Tetrachoric correlation	,0949754		
Contingency coefficient	,0503156		

Způsob 2: Řešení v softwaru Statistica I

- Na hladině významnosti 0,05 testujte hypotézu o nezávislosti genu a stavu pacienta. Simultánní četnosti znázorněte graficky.

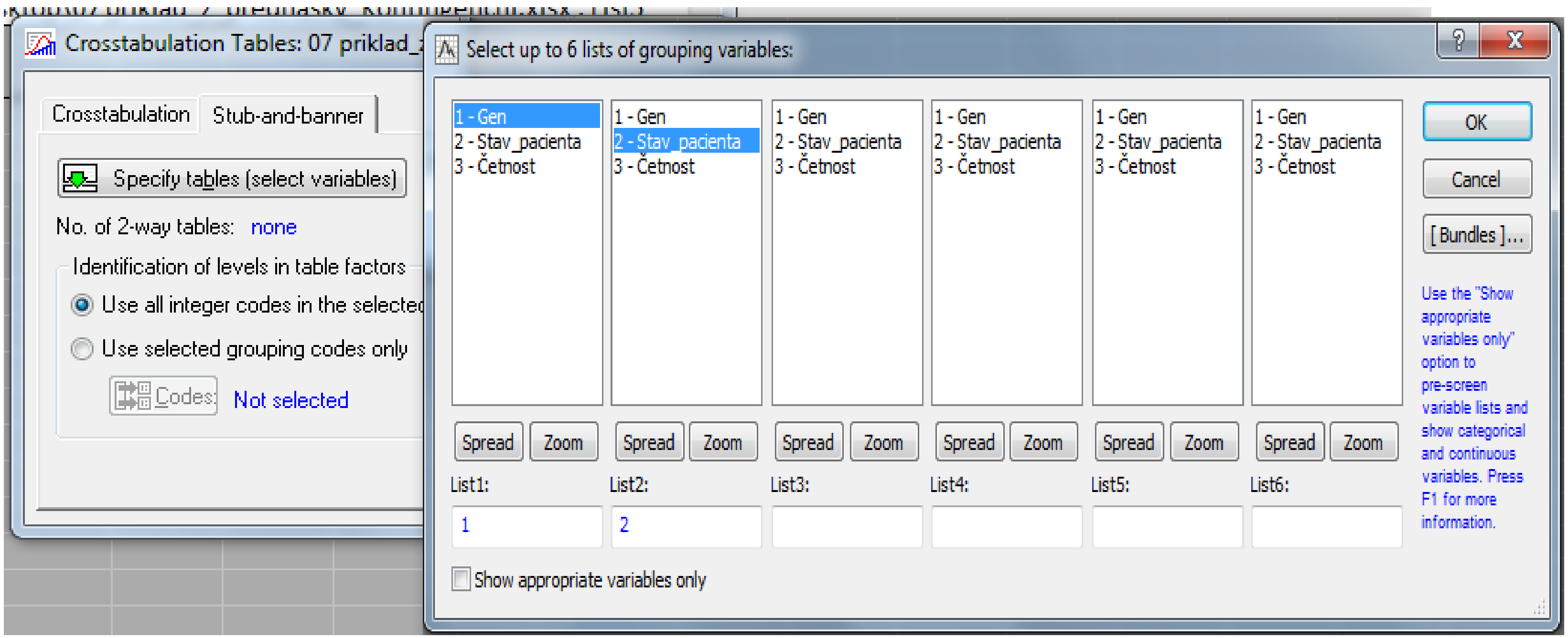
• **Agregovaný datový soubor**

• V menu ***Statistics*** zvolíme ***Basic statistics,*** vybereme ***Tables and banners*** (v češtině ***Kontingenční tabulky***)



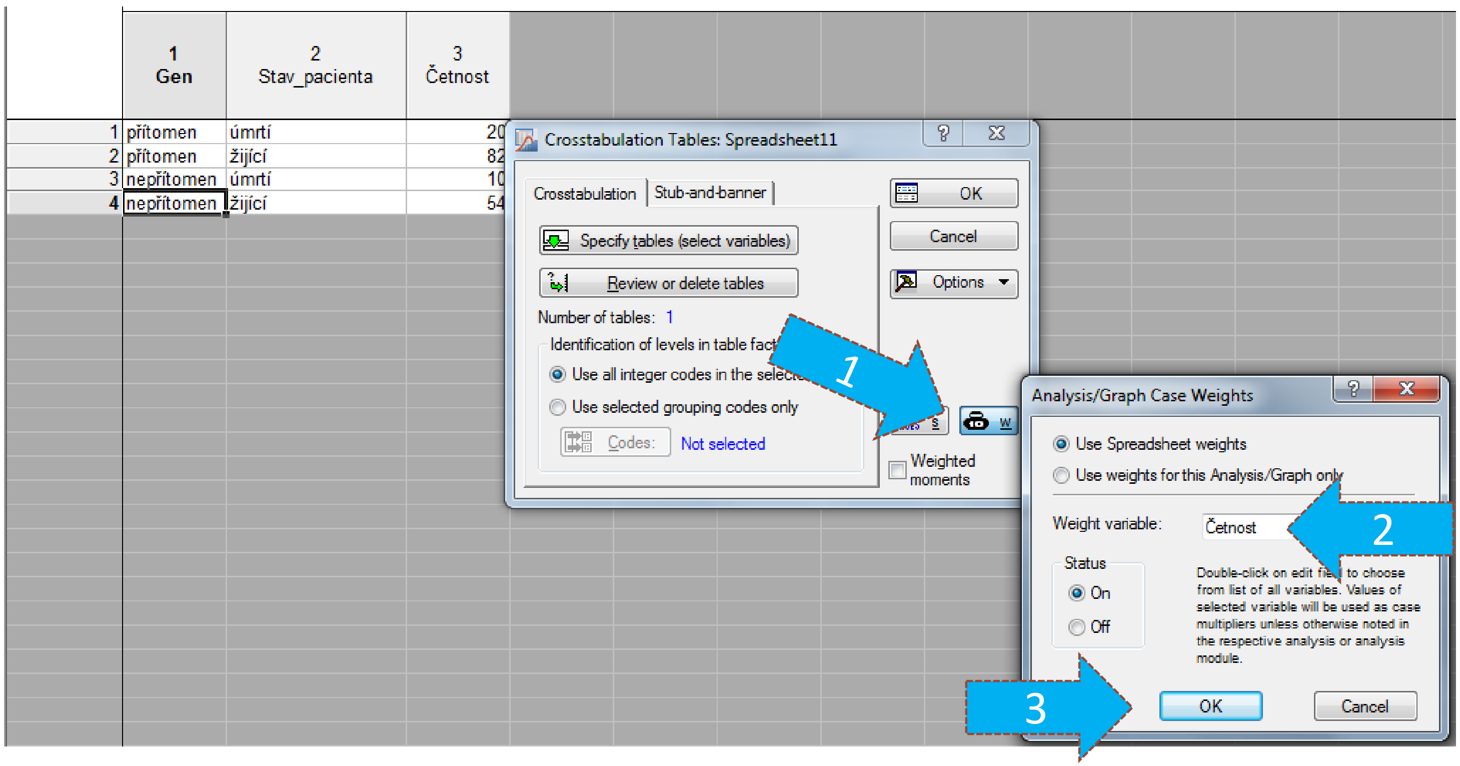
Způsob 2: Řešení v softwaru Statistica II

- Vybereme proměnné, které chceme testovat



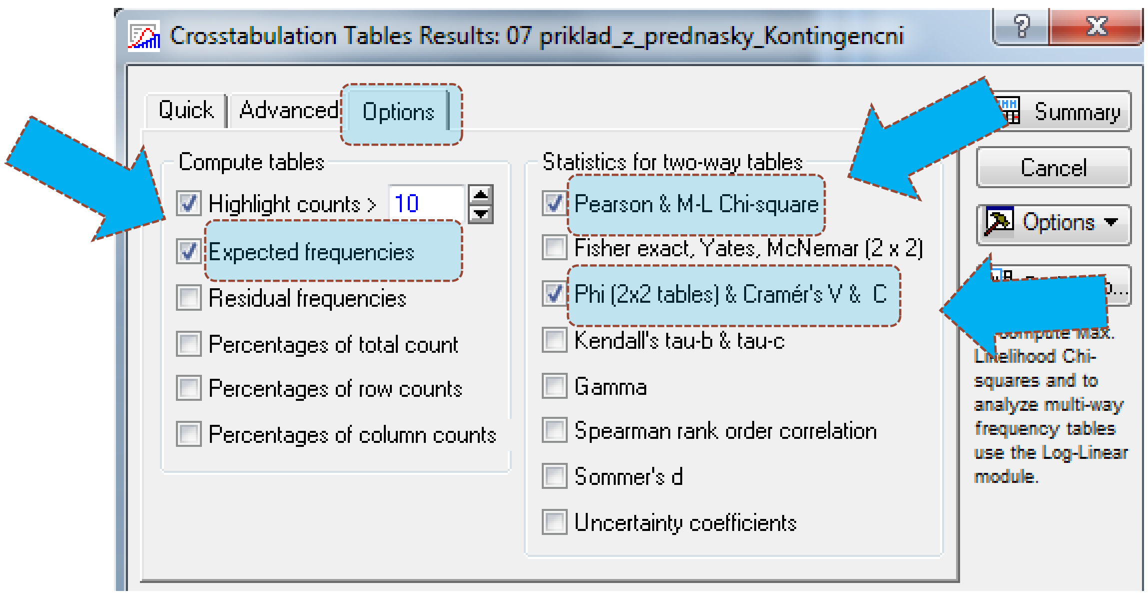
Způsob 2: Řešení v softwaru Statistica III

• Zapneme **váhy** (vpravo ikonka černých vah ***w***), jako váhy vybereme proměnnou **četnost** (tj. proměnnou, ve které jsou uvedeny počty případů jednotlivých kombinací kategorií)



Způsob 2: Řešení v softwaru Statistica IV

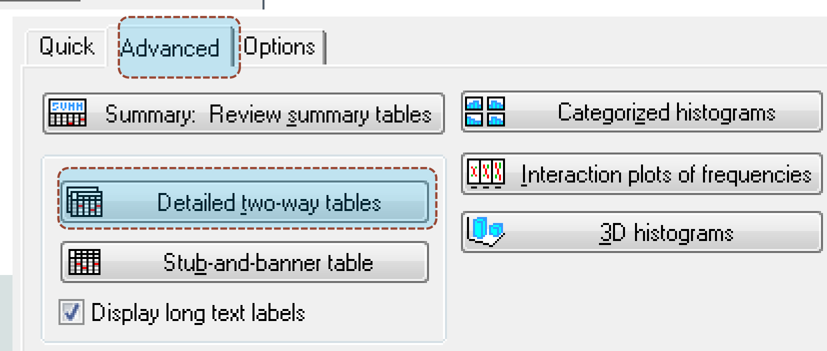
• Na záložce **Options** zaškrtneme ***Expected frequencies*** (***Očekávané četnosti***) (k ověření podmínek dobré aproximace)



• Zaškrtneme Pearsonův chí-kvadrát

• Pokud chceme vypočítat i Cramérův koeficient zaškrtneme Phi & Cramer‘s V

• Poté se vrátíme na záložku **Advanced,** kde a zvolíme ***Detailed two-way tables***



Testování homogenity (shody struktury)

- Motivace: Zajímá nás výskyt nominálního znaku u *r* nezávislých výběrů z *r* různých populací.
- Příklad: Je zájem o sport stejný u děvčat jako u chlapců?
- Nulová hypotéza: pravděpodobnostní rozdělení kategoriální proměnné je stejné v různých populací
- Test: **Pearsonův chí-kvadrát**

		Dívky	**Chlapci**	
Zájem o sport	**Ano**	*a*	b	a+b
	Ne	*c*	d	c+d
		a+c	b+d	n

Některé marginální četnosti (buď sloupcové, nebo řádkové) jsou předem pevně stanoveny

Fisherův exaktní test

- Využití ve čtyřpolní tabulce s nízkými četnostmi, které znemožňují použití Pearsonova chí-kvadrát testu.
- Patří mezi **neparametrické testy** pracující s daty na nominální škále, v nejjednodušší podobě ve dvou třídách: pozitivní/negativní, úspěch/neúspěch apod.
- Nulová hypotéza předpokládá rovnoměrné zastoupení sledovaného znaku u dvou nezávislých souborů.
- Slovo exaktní (přímý) znamená, že se přímo vypočítává pravděpodobnost odmítnutí, resp. platnosti nulové hypotézy.

Fisherův exaktní test

- Výpočet „přesné“ p-hodnoty, která zde hraje roli testové statistiky:
  - spočítá se parciální pravděpodobnost čtyřpolní tabulky p_1:

Sledovaný jev	Skupina		
	Experimentální	Kontrolní	Celkem
Ano	a	b	$a + b$
Ne	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

$$p_1 = \frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{N! * a! * b! * c! * d!}$$

  - Spočítá se p_a všech možných tabulek při zachování marginálních četností (řádkové a sloupcové součty) a výsledná p-hodnota je součtem p_a menších nebo stejných jako p_1, která přísluší pozorované tabulce.

Fisherův exaktní test – podrobněji

- Založen na výpočtu „přesné“ p-hodnoty (pravděpodobnosti, s jakou bychom dostali stejný nebo ještě extrémnější výsledek při zachování součtu řádků i sloupců v tabulce).
- **Příklad**: Chceme ověřit vztah dvou typů nežádoucích účinků, které jsou sumarizovány následující tabulkou:

		NÚ II ano	NÚ II ne
NÚ I	ano	2	3
	ne	6	4

- **Postup:** Všechny varianty tabulky při zachování součtu řádků a sloupců:

0	5	1	4	2	3	3	2	4	1	5	0
8	2	7	3	6	4	5	5	4	6	3	7

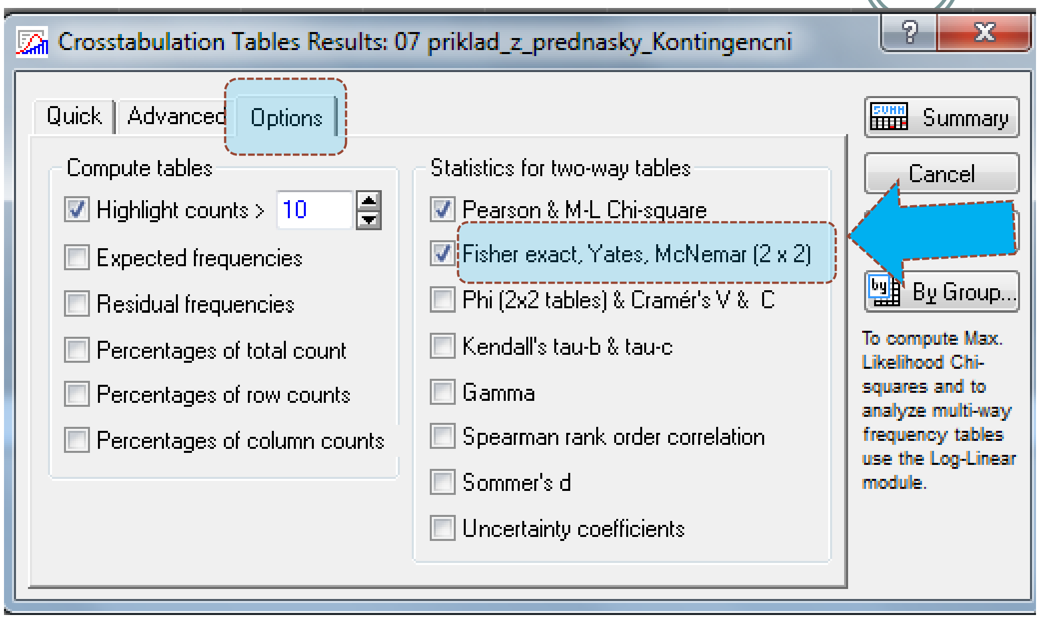
Pravděpodobnosti výskytu jednotlivých tabulek:

0,007 0,093 0,326 0,392 0,163 0,019

Oboustranná p-hodnota (sečtení pravděpodobností stejných nebo menších než je pravděpodobnost pozorované varianty):

p = 0,326 + 0,093 + 0,007 + 0,163 + 0,019 = **0,608**

Řešení v softwaru Statistica: Fisherův exaktní test



- Na záložce **Options** zaškrtneme ***Fisher exact***

- Výstupní tabulka

Statistic	Statistics: Gen(2) x Stav_paci Chi-square	df	p
Pearson Chi-square	,4213223	df=1	p=,51628
M-L Chi-square	,4277117	df=1	p=,51311
Yates Chi-square	,1952605	df=1	p=,65857
Fisher exact, one-tailed			p=,33259
two-tailed			p=,54314
McNemar Chi-square (A/D)	14,71622	df=1	p=,00012
(B/C)	54,79348	df=1	p=,00000

Pro jednostranný test

Pro oboustranný test

Fisherův exaktní test na webu

- 2 x 2 tabulky: http://graphpad.com/quickcalcs/contingency1/
- 2 x 3 tabulky: http://www.vassarstats.net/fisher2x3.html
- 2 x 5 (nebo menší) tabulky:
  http://www.quantitativeskills.com/sisa/statistics/fiveby2.htm
- 3 x 3 tabulky: http://vassarstats.net/fisher3x3.html

Test hypotézy o symetrii (McNemarův test pro čtyřpolní tabulku)

- Motivace: Na osobách sledujeme binární proměnnou před pokusem a po něm, cílem je zjistit, zda došlo ke změně v rozdělení této proměnné.
- **Analýza párových dichotomických proměnných**

Četnostní tabulka

		po		
		+	**-**	$n_{j.}$
před	**+**	*a*	*b*	*a+b*
	-	*c*	*d*	*c+d*
	$n_{.k}$	*a+c*	*b+d*	*n*

Tabulka teoretických pravděpodobností

		po		
		+	**-**	
před	**+**	p_{11}	p_{12}	$p_{1.}$
	-	p_{21}	p_{22}	$p_{2.}$
		$p_{.1}$	$p_{.2}$	

- Nulová hypotéza: $p_{ij} = p_{ji}$, pokus nemá vliv na výskyt daného znaku
- Testová statistika: $\chi^2 = \dfrac{(|b-c|-1)^2}{b+c}$ pokud je větší než kritická hodnota χ^2 rozdělení o jednom stupni volnosti (vhodné pro počty údajů b+c > 8), pak nulovou hypotézu zamítáme

McNemarův test: příklad I

Zjistěte, zda úspěch našich sportovců na olympiádě vede ke změně postojů žáků ke sportování.

Nulová hypotéza: Počet žáků, kteří změní svůj postoj pozitivním směrem, je pouze náhodně odlišný od počtu žáků, kteří změní svůj postoj negativním směrem.

		Postoj po olympiádě		
		+	-	
Postoj před olympiádou	+	*5*	3	8
	-	*16*	2	18
		21	5	26

$$\chi^2 = \frac{(|3-16|-1)^2}{3+16} = 7{,}58$$

Stupně volnosti

Tabulky: $\chi^2_{1-\alpha}(v = k(k-1)/2 = 1) = 3{,}84$

H_0 zamítnuta

Závěr: Úspěch našich sportovců má pozitivní vliv na postoj žáků vzhledem k provozování sportu.

Řešení v softwaru Statistica: McNemarův test

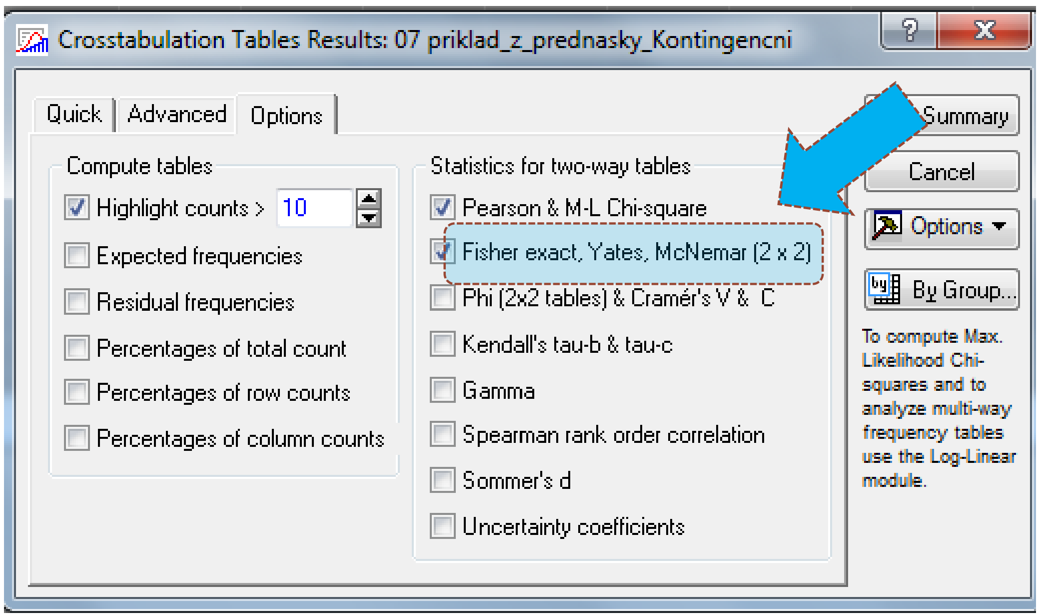
Datový soubor

	1 postoj_pred_vyukou	2 postoj_po_vyuce	3 cetnost
1	kladný	kladný	5
2	záporný	kladný	16
3	kladný	záporný	3
4	záporný	záporný	2

Výstupní kontingenční tabulka

2-rozměrná tabulka: Pozorované četnosti (priklad_pos
Četnost označených buněk > 10

postoj_pred_vyukou	postoj_po_vyuce kladný	postoj_po_vyuce záporný	Řádk. součty
kladný	5	3	8
záporný	16	2	18
Celk.	21	5	26



- Na záložce **Options** zaškrtneme ***McNemar (2x2)***

- Výstupní tabulka

Statist.	Statist. : postoj_pred_vyukou(2) x p Chí-kvadr.	sv	p
Yatesův chí-kv.	1.074735	df=1	p=.29988
Fisherův přesný, 1-str.			p=.15026
Fisherův přesný, 2-str.			p=.28051
McNemarův chí-kv. (A/D)	.5714286	df=1	p=.44969
McNemarův chí-kv. (B/C)	7.578948	df=1	p=.00591

2 hodnoty testových statistik a p-hodnoty, podle toho, kde jsou ve výstupní kontingenční tabulce uloženy četnosti, u kterých jsme při opakovaném měření zaznamenali rozdílné výsledky (A/D nebo B/C)

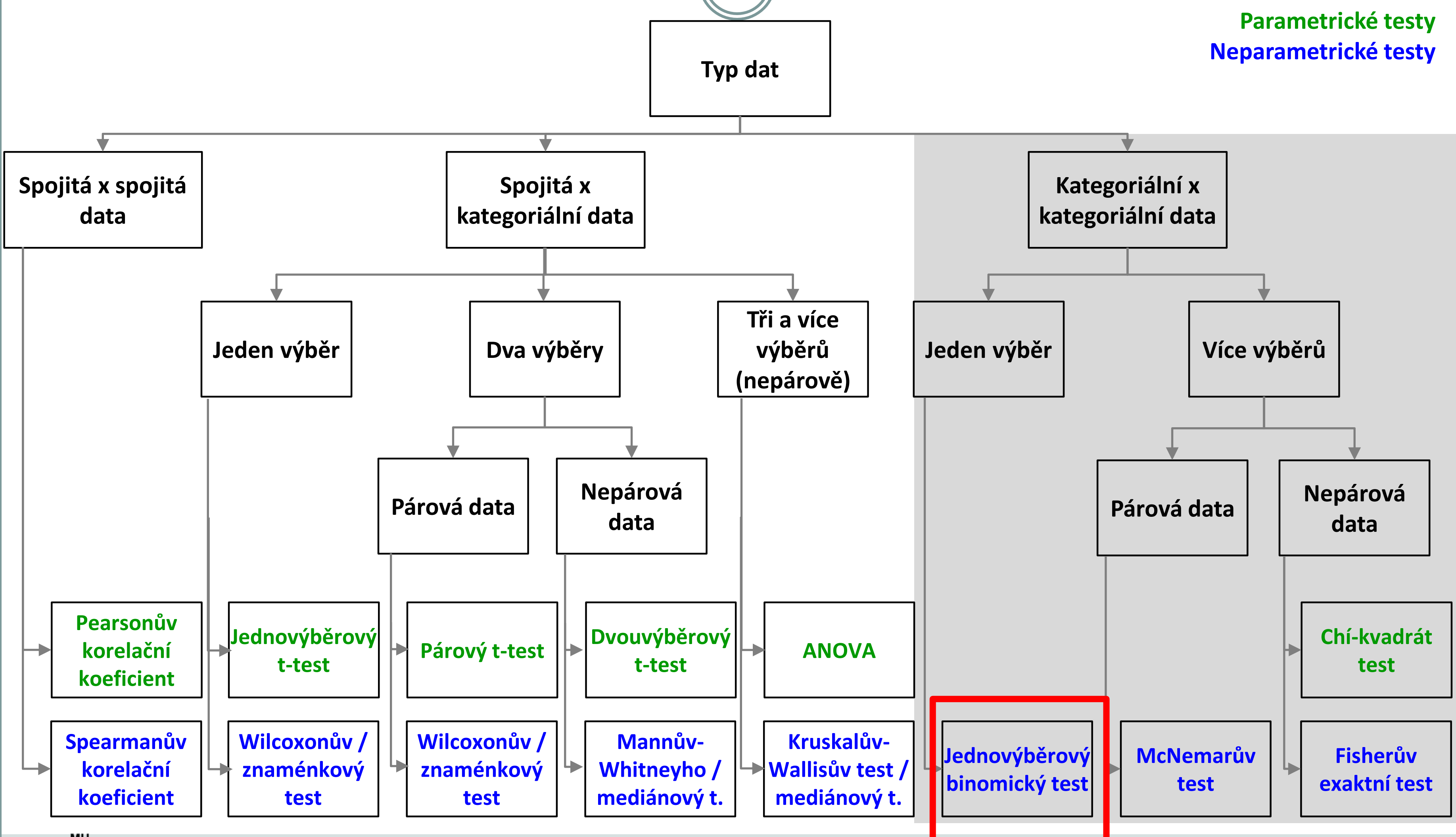
Společný příklad – testování homogenity

Očkování proti chřipce se zúčastnilo 460 dospělých, z nichž 240 dostalo očkovací látku proti chřipce a 220 dostalo placebo. Na konci experimentu onemocnělo 100 lidí chřipkou, 20 z nich bylo z očkované skupiny a 80 z kontrolní skupiny. Je to dostatečný důkaz, že očkovací látka byla účinná?

Nulová hypotéza: Procento výskytu chřipky je v očkované a kontrolní skupině stejné.

1. *Vytvořte si na základě zadání datový soubor v softwaru STATISTICA (agregovaná data ve formě kontingenční tabulky).*
2. *Testujte platnost nulové hypotézy pomocí Pearsonova chí-kvadrát testu.*
3. *Testujte platnost nulové hypotézy pomocí Fisherova exaktního testu.*
4. *Který z testů je vhodné použít a proč?*

Základní rozhodování o výběru statistických testů
- jednovýběrový binomický test



Jednovýběrový binomický test

- test pro podíl u jednoho výběru
- **Liší se podíl (*p*) pacientů s výskytem sledovaného jevu od předpokládané (referenční) hodnoty (π)?**
- např. liší se procento pacientů s nežádoucími účinky léčby od předpokládaného procenta?
- výpočet: https://www.medcalc.org/calc/test_one_proportion.php

Jednovýběrový binomický test – příklad

- **Příklad:** Z 50 studentů, kteří si zvolilo maturitu z matematiky, ji v letošním roce neudělalo 12. Ověřte, zda podíl neúspěšných studentů je stejný jako v předchozím roce, kdy byla neúspěšnost 5%.
- Tzn. hypotézy budou mít tvar:
  $H_0 : p = \pi$ a $H_1 : p \neq \pi$
- **Řešení:**
- π = 5%
- p = 12/50 = 0,24 => 24%
- **Závěr:**
  Podíl neúspěšných studentů je statisticky významně odlišný od podílu v předchozím roce.

Test for one proportion calculator

Observed proportion

Observed proportion (%): 24

Sample size: 50

Null hypothesis value

Null hypothesis value (%): 5

Test

Results

z-statistic	6.164
Significance level	P < 0.0001
95% CI of observed proportion	13.06 to 38.17

Základy korelační analýzy

Korelace a regrese
Pearsonův korelační koeficient
Spearmanův korelační koeficient

Proč hodnotit vztah dvou spojitých veličin?

- Vztah mezi dvěma spojitými veličinami v jedné skupině:

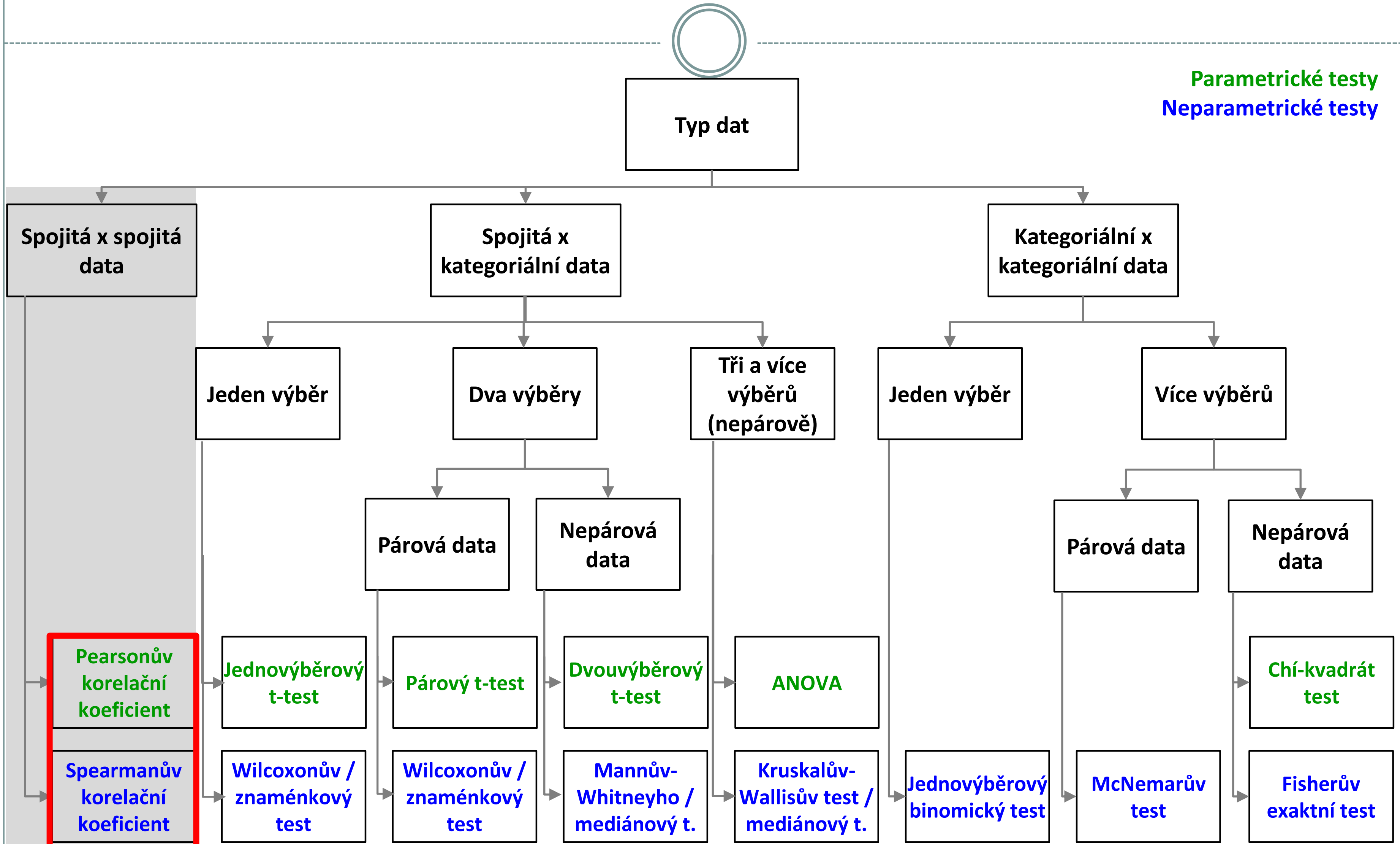
1. Chceme zjistit, jestli mezi nimi **existuje vztah** – např. jestli vyšší hodnoty jedné veličiny znamenají nižší hodnoty jiné veličiny;
2. Chceme **predikovat hodnoty** jedné veličiny na základě znalosti hodnot jiných veličin;
3. Chceme **kvantifikovat vztah** mezi dvěma spojitými veličinami – např. pro použití jedné veličiny namísto druhé veličiny.

Korelační a regresní analýza

- **Korelační analýza** je využívána pro vyhodnocení míry vztahu dvou spojitých proměnných. Obdobně jako jiné statistické metody, i korelace mohou být parametrické nebo neparametrické.

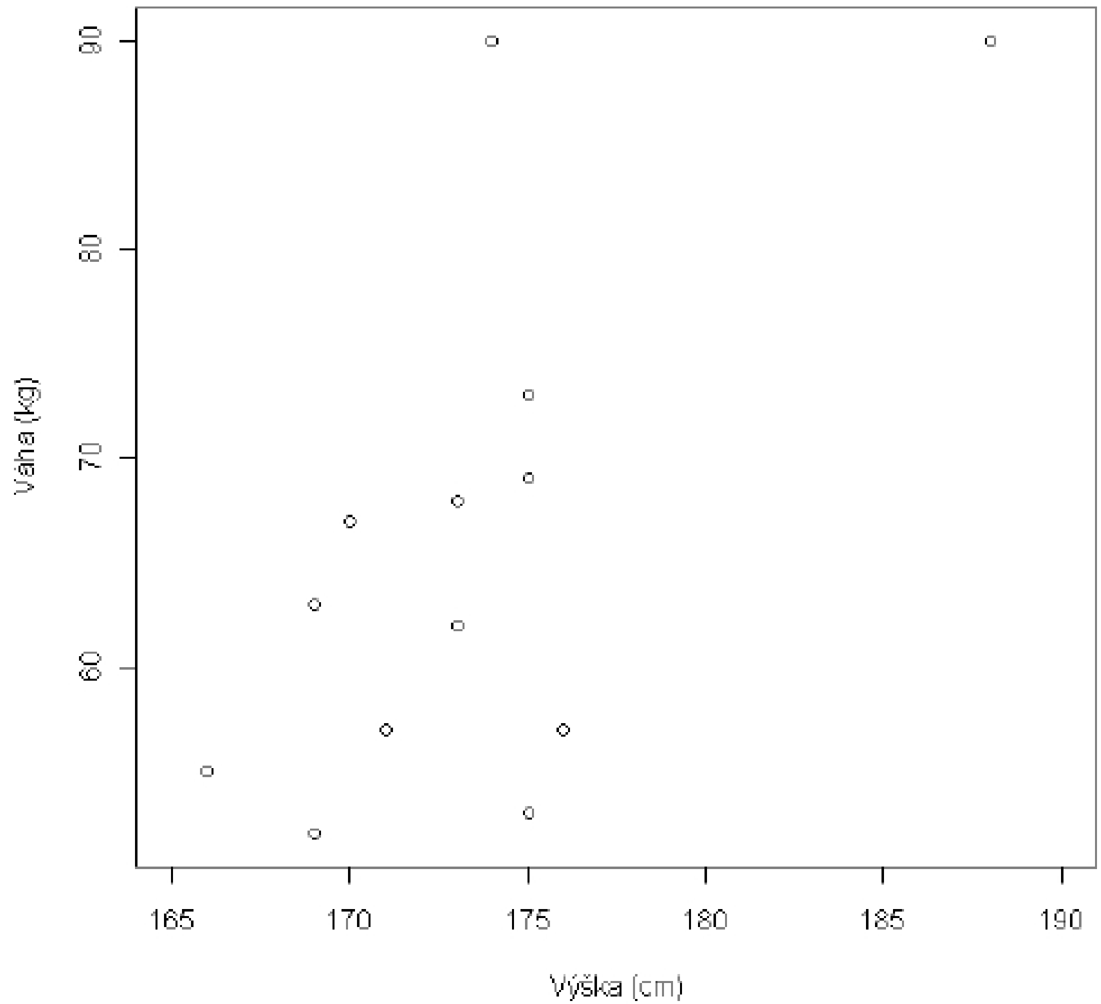
- **Regresní analýza** vytváří model vztahu dvou nebo více proměnných, tedy jakým způsobem jedna proměnná (vysvětlovaná) závisí na jiných proměnných (prediktorech). Regresní analýza je obdobně jako ANOVA nástrojem pro vysvětlení variability hodnocené proměnné.

Základní rozhodování o výběru statistických testů - korelační analýza



Vizuální hodnocení vztahu dvou proměnných

- Nejjednodušší formou je **bodový graf** (x-y graf), tzv. scatterplot.
- Vztah výšky a váhy studentů Biostatistiky pro matematické biology – jaro 2010:



Korelační koeficienty

- **Korelační koeficient** (*r*) – kvantifikuje míru vztahu mezi dvěma spojitými veličinami (*X* a *Y*).
  - **Pearsonův korelační koeficient –** parametrický, hodnotí míru **lineární** závislosti mezi 2 spojitými proměnnými,
  - **Spearmanův korelační koeficient –** neparametrický, hodnotí míru **lineární** pořadové závislosti mezi 2 spojitými proměnnými.
  - Hodnota *r* je kladná, když vyšší hodnoty *X* souvisí s vyššími hodnotami *Y*, naopak hodnota r je záporná, když nižší hodnoty *X* souvisí s vyššími hodnotami *Y*.
  - Nabývá hodnot od -1 do 1:

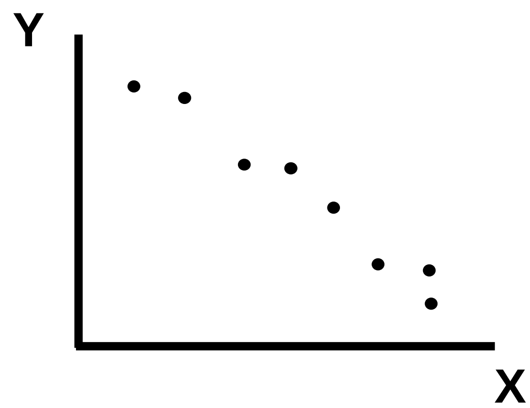
    $r = 0 \rightarrow$ nekorelované

    $r > 0 \rightarrow$ kladně korelované

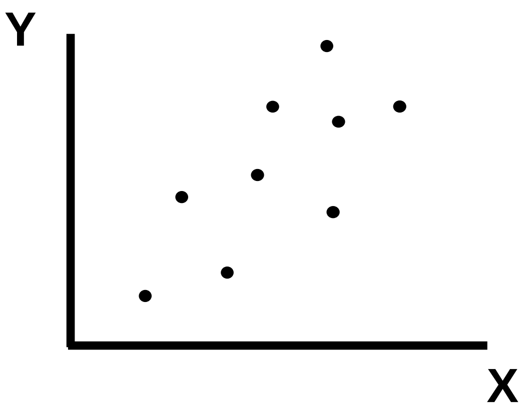
    $r < 0 \rightarrow$ záporně korelované

Korelační koeficienty - ukázky

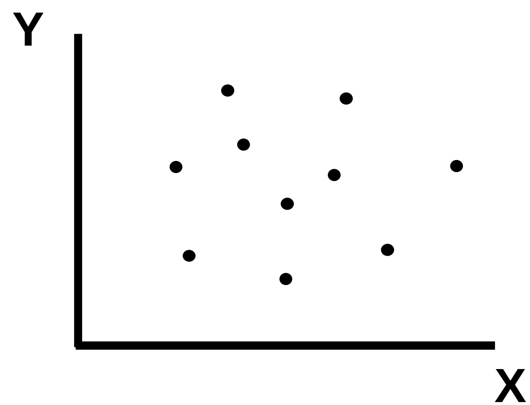
r = 1,0



r = -0,9



r = 0,4



r = 0,05

Test hypotézy H_0: $r = 0$

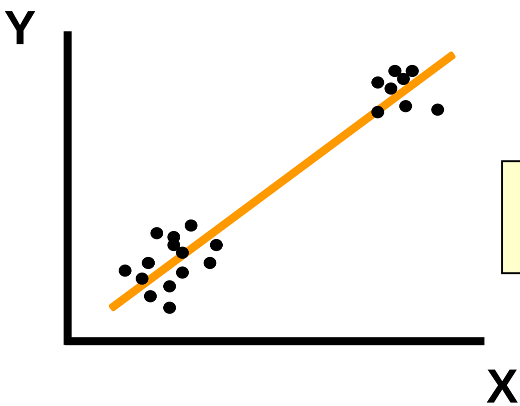
- K měření těsnosti lineárního vztahu 2 spojitých proměnných
  $r = 0$ → nekorelované
  $r > 0$ → kladně korelované
  $r < 0$ → záporně korelované

- H_0: proměnné X, Y jsou nezávislé náhodné veličiny
  ($r = 0$)
  H_A: proměnné X, Y nejsou nezávislé náhodné veličiny ($r \neq 0$)

- Testování pomocí intervalu spolehlivosti nebo výpočet testové statistiky (srovnání s kritickou hodnotou nebo výpočet p-hodnoty)

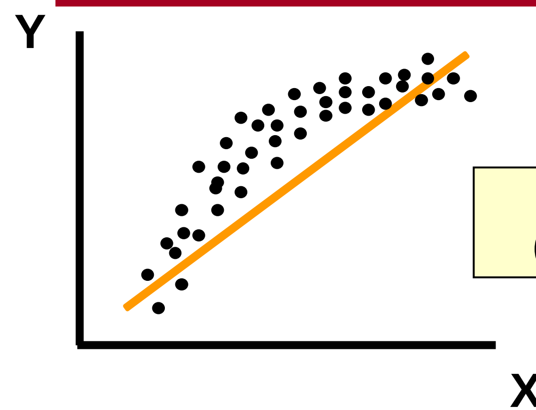
Problémy s výpočtem *r*

Problém více skupin



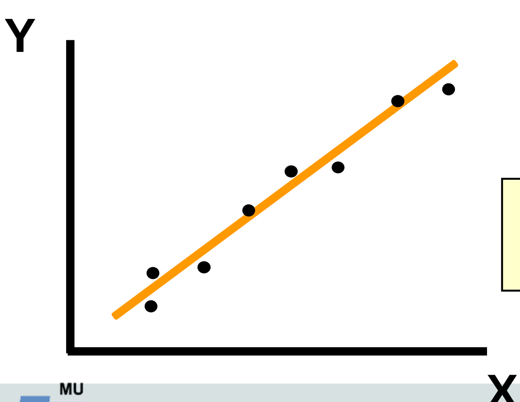
r = 0,981
(p < 0,001)

Nelineární vztah



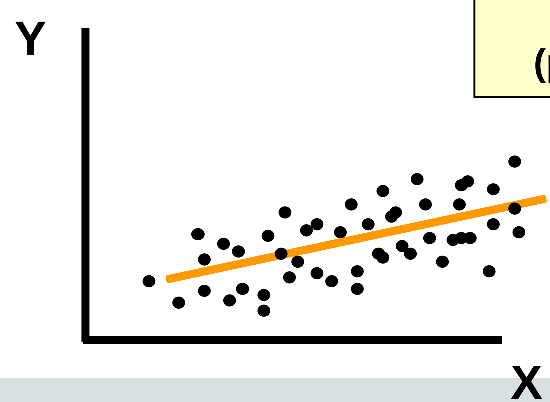
r = 0,761
(p = 0,032)

Problém velikosti výběru



r = 0,891
(p = 0,214)

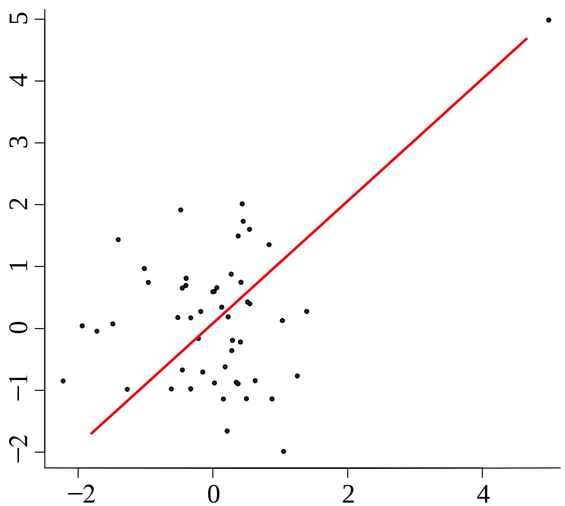
r = 0,212
(p = 0,008)



Problémy s výpočtem Pearsonova kor. koef. I

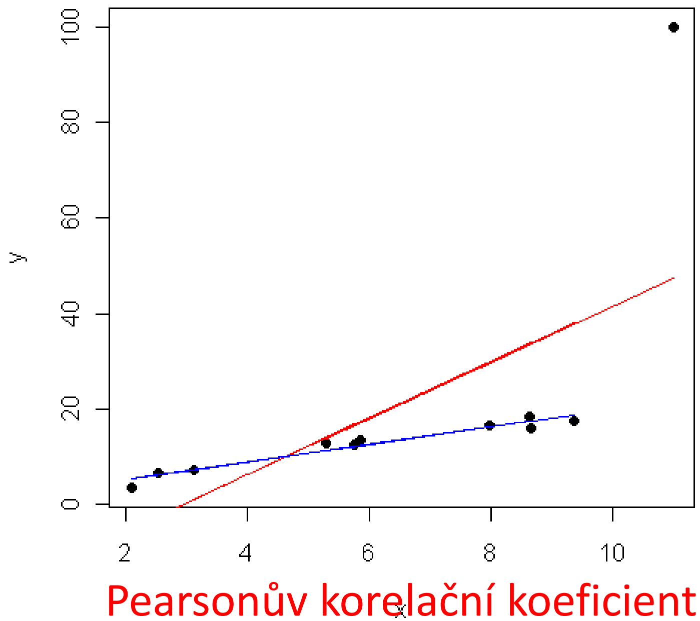
Odlehlá hodnota

Může způsobit, že korelace vyjde významně, i když ve skutečnosti tam žádný vztah není!



r = 0,36 (p = 0,009)

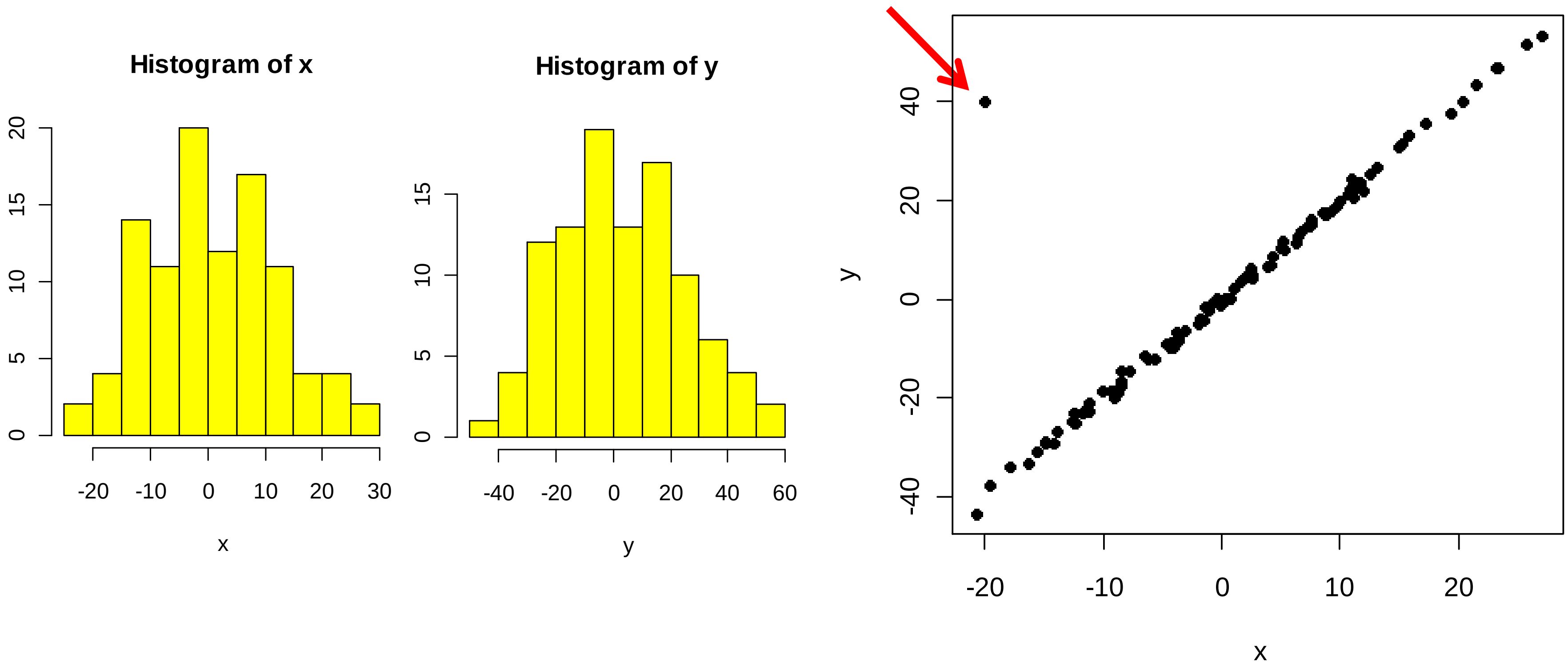
Může způsobit, že korelace bude méně významná (či dokonce nevýznamná), i když ve skutečnosti tam vztah je!



Pearsonův korelační koeficient:
r = 0,65 (p = 0,029)
Spearmanův korelační koeficient:
r_S = 0,95 (p < 0,001)

Problémy s výpočtem Pearsonova kor. koef. II

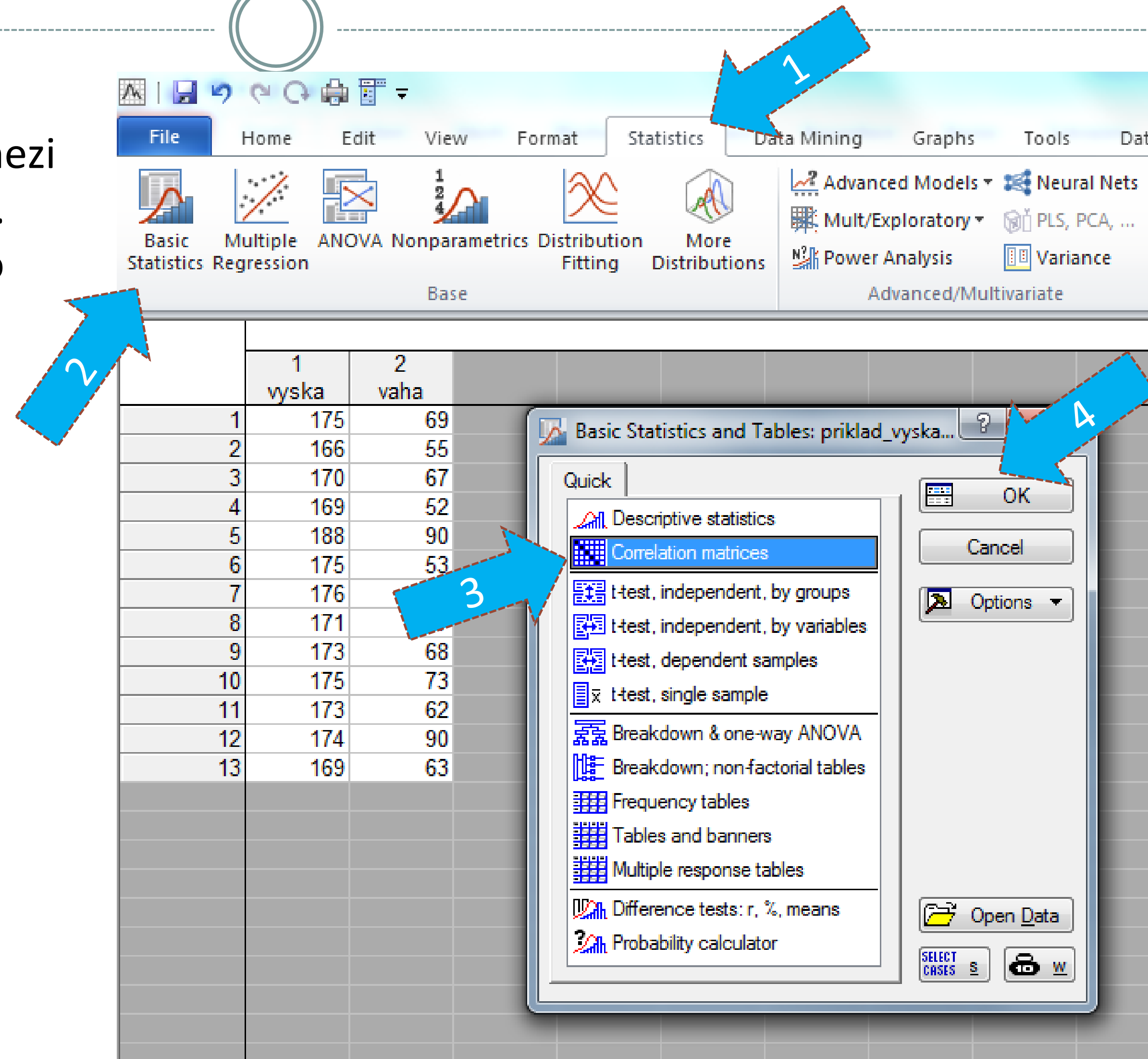
- Při srovnání dvou spojitých proměnných je nutné vykreslovat bodový graf, protože histogramy pro jednotlivé proměnné zvlášť nám nemusejí odhalit odlehlé hodnoty!



Řešení v softwaru Statistica: Pearsonův korelační koeficient I

Prozkoumejte lineární vztah mezi výškou a váhou u 13 studentů. Testujte hypotézu, že jsou tyto proměnné nezávislé.

1. Záložka **Statistics**
2. **Basic Statistics**
3. **Correlation matrices**
4. Potvrdíme: **OK**

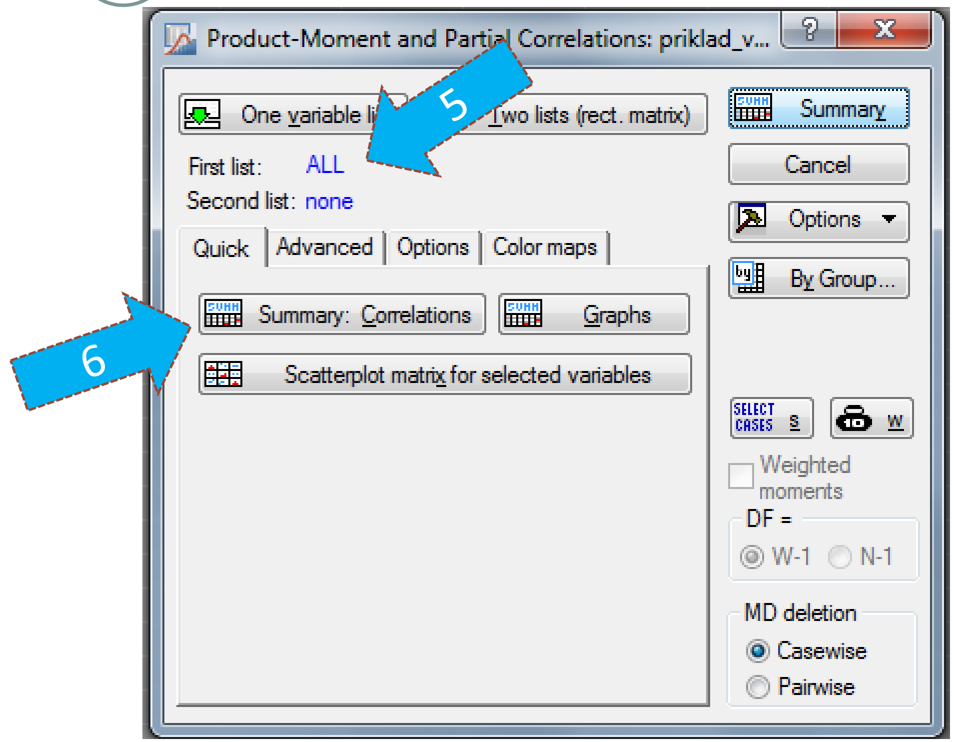


Řešení v softwaru Statistica: Pearsonův korelační koeficient II

5. Vybereme spojité proměnné pro hodnocení vztahu (váha a výška) ve ***One variable list***.

Na záložce ***Options*** můžeme vybrat formu výstupu (pouze p-hodnoty, matice korelačních koeficientů a p-hodnot ap.).

6. ***Summary: Correlations***



Jedna z možných výstupních tabulek:

Variable	Correlations (priklad_vyska_vaha.sta) Marked correlations are significant at p < ,05000 N=13 (Casewise deletion of missing data)			
	Means	Std.Dev.	vyska	vaha
vyska	173,3846	5,31568	1,000000	0,639675
vaha	65,8462	12,54224	0,639675	1,000000

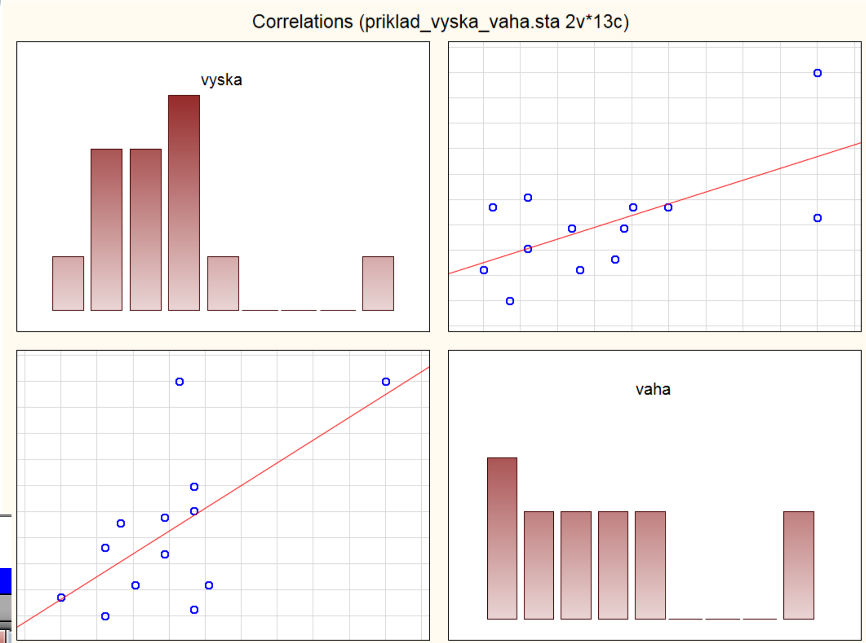
p-hodnota < 0,05 - test hypotézy H_0: $r = 0$, lze vypsat i konkrétní hodnotu (změna formy výstupu na záložce ***Options*** – skutečnou p-hodnotu zjistíme zatrhnutím ***Display r, p-values, and N's***

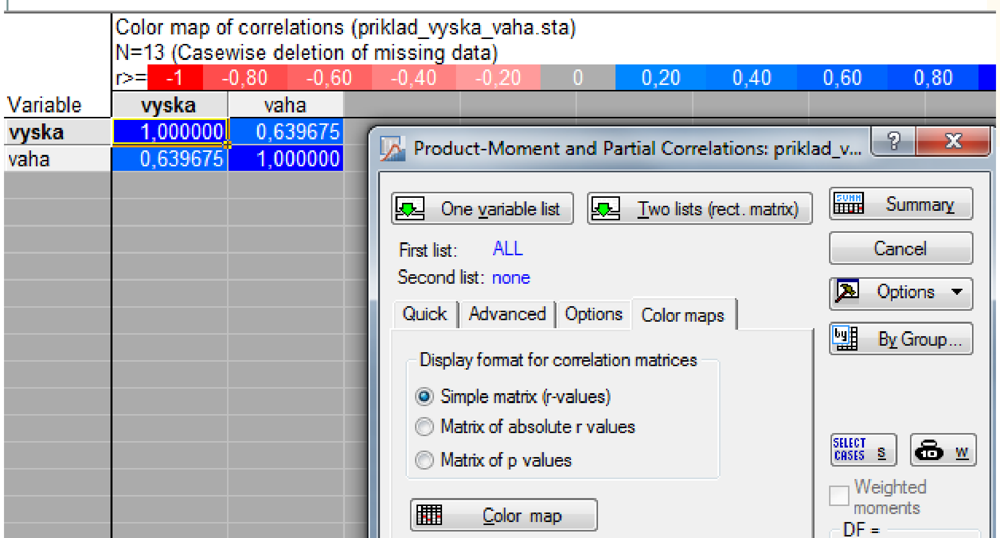
Pearsonovy korelační koeficienty

Řešení v softwaru Statistica: Pearsonův korelační koeficient III

Záložka ***Quick / Advanced*** umožňuje vykreslit různé druhy grafů (2D, 3D v případě více proměnných, matice bodových grafů s histogramy na diagonále ap.).

Jsou v daném případě splněny předpoklady (dvourozměrné normální rozdělení, absence odlehlých pozorování, lineární vztah)?



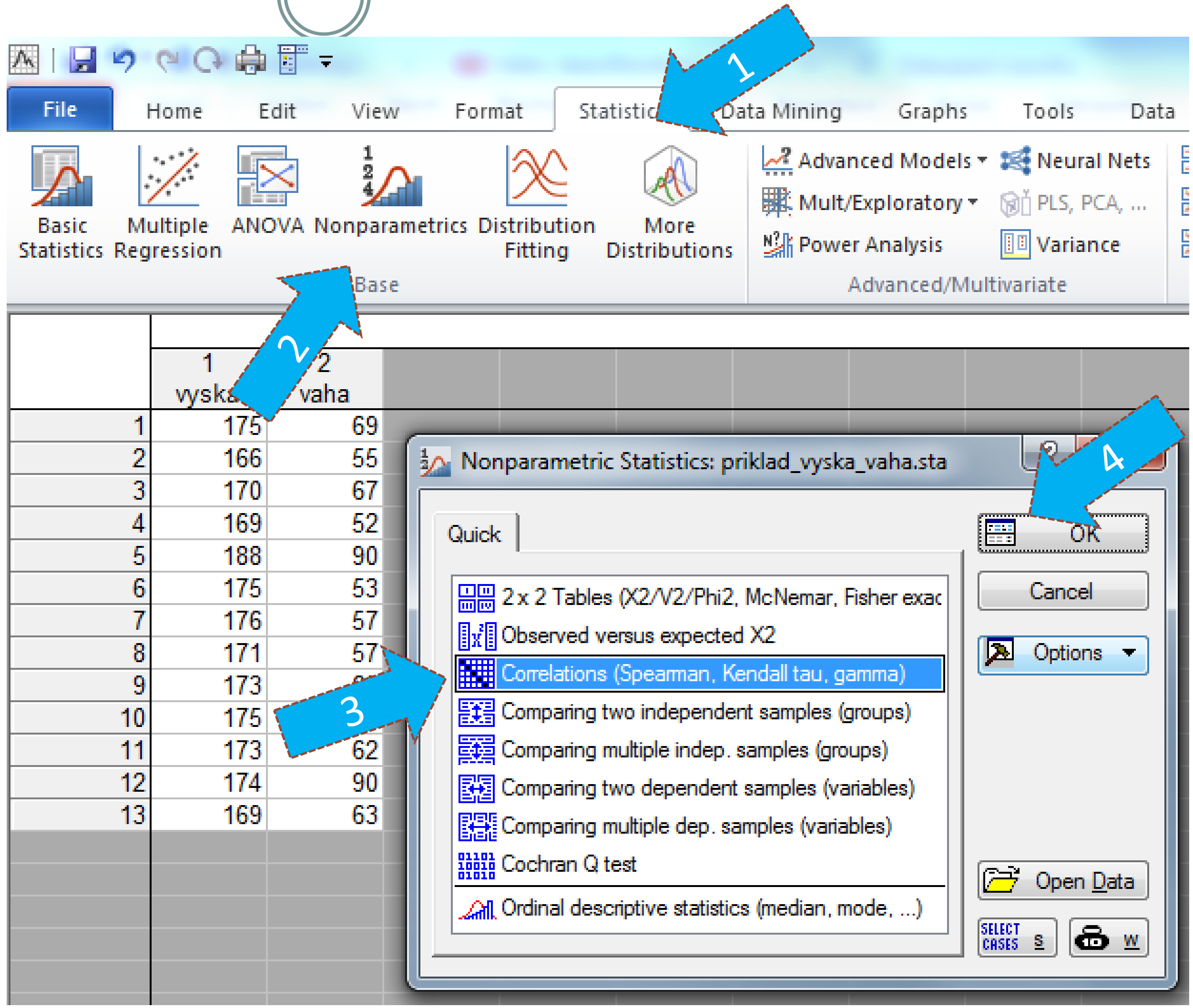


Na záložce ***Color maps*** můžeme získat matici korelačních koeficientů (nebo příslušných p-hodnot) obarvenou dle odpovídající barevné škály. Vhodné zejména při zkoumání vztahů mezi více spojitými proměnnými.

Řešení v softwaru Statistica: Spearmanův korelační koeficient I

Prozkoumejte pořadový vztah mezi výškou a váhou u 13 studentů. Testujte hypotézu, že jsou tyto proměnné nezávislé.

1. Záložka ***Statistics***
2. ***Nonparametrics***
3. ***Correlations***
4. Potvrdíme: ***OK***

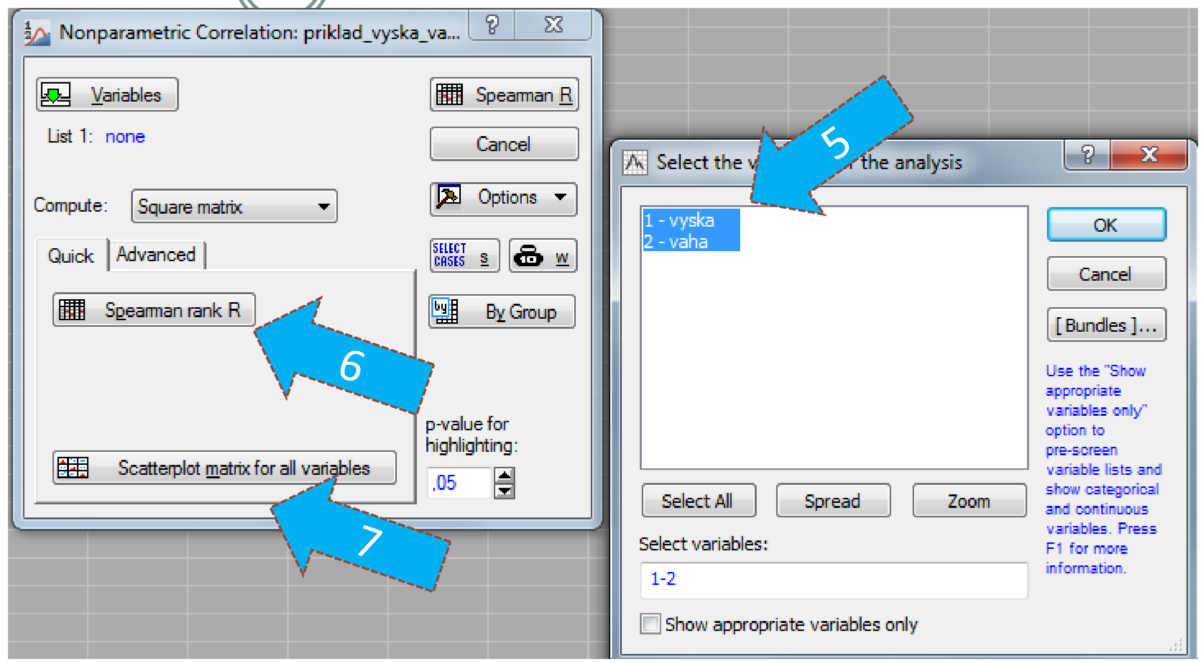


Řešení v softwaru Statistica: Spearmanův korelační koeficient II

5. Výběr proměnných – ***Variables*** – ***Select variables*** (vyska, vaha) – ***OK***

6. Pod možností Compute můžeme vybrat formu výstupu (čtvercová matice - ***Square matrix***, příp. detailní výsledky).

7. Lze vykreslit i matici bodových grafů s histogramy na diagonále (***Scatterplot matrix for all variables***).



p-hodnota <0,05 - test hypotézy H_0: $r = 0$,
lze vypsat i konkrétní hodnotu přepnutím ***Compute*** na ***Detailed report***

Jedna z forem výstupní tabulky:

Spearman Rank Order Correlations (priklad_vyska_vaha.sta)
MD pairwise deleted
Marked correlations are significant at p <,05000

Variable	vyska	vaha
vyska	1,000000	0,469452
vaha	0,469452	1,000000

Spearmanovy korelační koeicienty

Samostatný úkol

Testování nezávislosti
Testování homogenity

1. Příklad k procvičení

1. Testujte hypotézu, že **barva vlasů a barva očí spolu nesouvisí**. K dispozici jsou údaje od 6 800 mužů (*Yule, G. U., Kendall, M.G.: An Introduction to the Theory of Statistics, 14th ed. Griffin, London, 1950*).
2. Vypočítejte Cramérův koeficient a interpretujte jej.

	Světlá	Kaštanová	Černá	Zrzavá	Celkem
Modrá	1768	807	189	47	2811
Šedá nebo zelená	946	1387	746	53	3132
Tmavohnědá	115	438	288	16	857
Celkem	2829	2632	1223	116	6800

Nezapomeňte ověřit podmínky dobré aproximace!

2. Příklad k procvičení

1. Ve Skotsku byla provedena studie, která měla prokázat, **zda procentuální zastoupení krevních skupin na celém území je homogenní nebo není**. V oblasti Eskdale bylo náhodně vybráno 100 osob, v Annadale 125 osob a v Nithsdale 253 osob (*Osborn J. F., 1979, Statistical Exersice in Medical Research, Blackwell Scientific publications, Oxford*)

	A	B	0	AB	Celkem
Eskade	33	6	56	5	100
Annandale	54	14	52	5	125
Nithsdale	98	35	115	5	253
Celkem	185	55	223	15	478

Nezapomeňte ověřit podmínky dobré aproximace!

Výsledky k samostatnému úkolu

Testování nezávislosti

Testování homogenity

1. Příklad k procvičení

1. Testujte hypotézu, že **barva vlasů a barva očí spolu nesouvisí**. K dispozici jsou údaje od 6 800 mužů (*Yule, G. U., Kendall, M.G.: An Introduction to the Theory of Statistics, 14th ed. Griffin, London, 1950*).
2. Vypočítejte Cramérův koeficient a interpretujte jej.

Výsledky:

chí-kvadrát = 1073,51

p < 0,001 … na hladině významnosti zamítáme nulovou hypotézu o nezávislosti barvy očí a barvy vlasů (před provedením testu jsme zkontrolovali podmínky dobré aproximace),

Cramérův koeficient = 0,28 … mezi barvou očí a barvou vlasů je slabá závislost.

2. Příklad k procvičení

1. Ve Skotsku byla provedena studie, která měla prokázat, **zda procentuální zastoupení krevních skupin na celém území je homogenní nebo není**. V oblasti Eskdale bylo náhodně vybráno 100 osob, v Annadale 125 osob a v Nithsdale 253 osob (*Osborn J. F. , 1979, Statistical Exersice in Medical Research, Blackwell Scientific publications, Oxford*)

Výsledky:

chí-kvadrát = 10,454

P = 0,107 … nelze zamítnout nulovou hypotézu, že procentuální zastoupení krevních skupin na celém území je homogenní / stejné
(před provedením testu jsme zkontrolovali podmínky dobré aproximace).