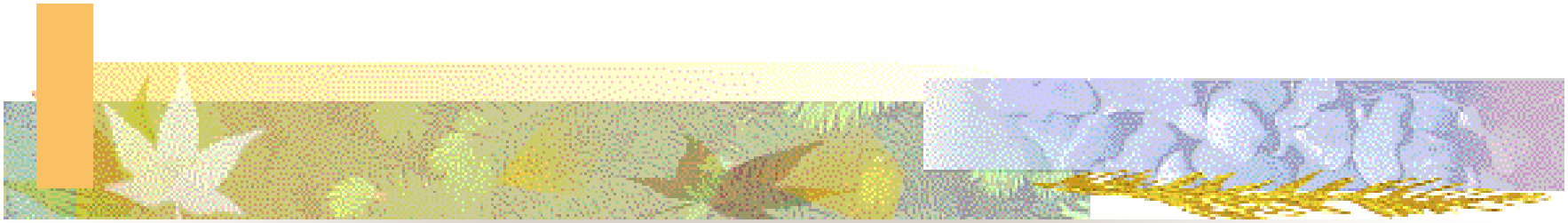
11_Chí-kvadrát (χ^2)



Test Chí-kvadrát: použití

- chí-kvadrát může být použit
  - pro testování rozdělení jedné nominální proměnné (test dobré shody)
  - testování nezávislosti zejména dvou nominálních nebo i ordinálních proměnných s málo kategoriemi

Chí-kvadrát pro testování nezávislosti mezi dvěma proměnnými

- Ze znalosti o průniku (kapitola PRAVDĚPODOBNOST) již víme, že jevy jsou také statisticky nezávislé pokud

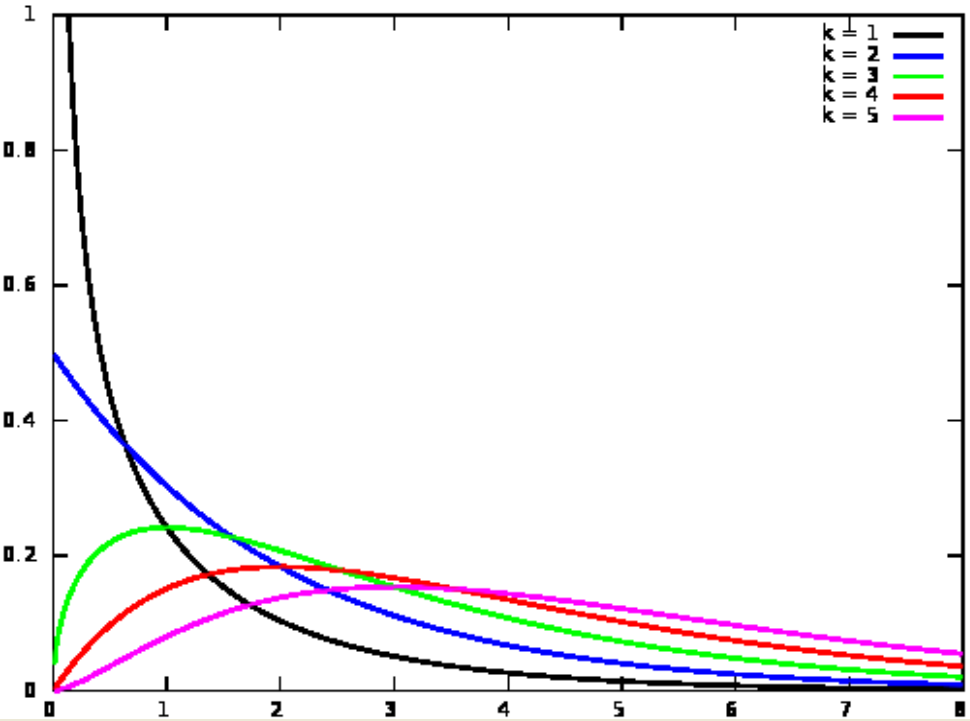
  p(A a B) = p(A)*p(B)
- Tedy pokud se četnost kombinace v buňce kontingenční tabulky rovná násobku marginálních celkových četností v příslušném řádku a sloupci vydělená celkovou velikostí vzorku
- Toto očekávámé pokud platí nulová hypotéza, že jevy jsou nezávislé
- Očekávané četnosti pod nulovou hypotézou
  - = $O_{ij} = (\check{r}_i * s_j)/ N$, tj. pro každou buňku tabulky se vynásobí celkové marginální četnosti z příslušného řádku se sloupcovými četnostmi a vydělí celkovým počtem osob

Test Chí-kvadrát

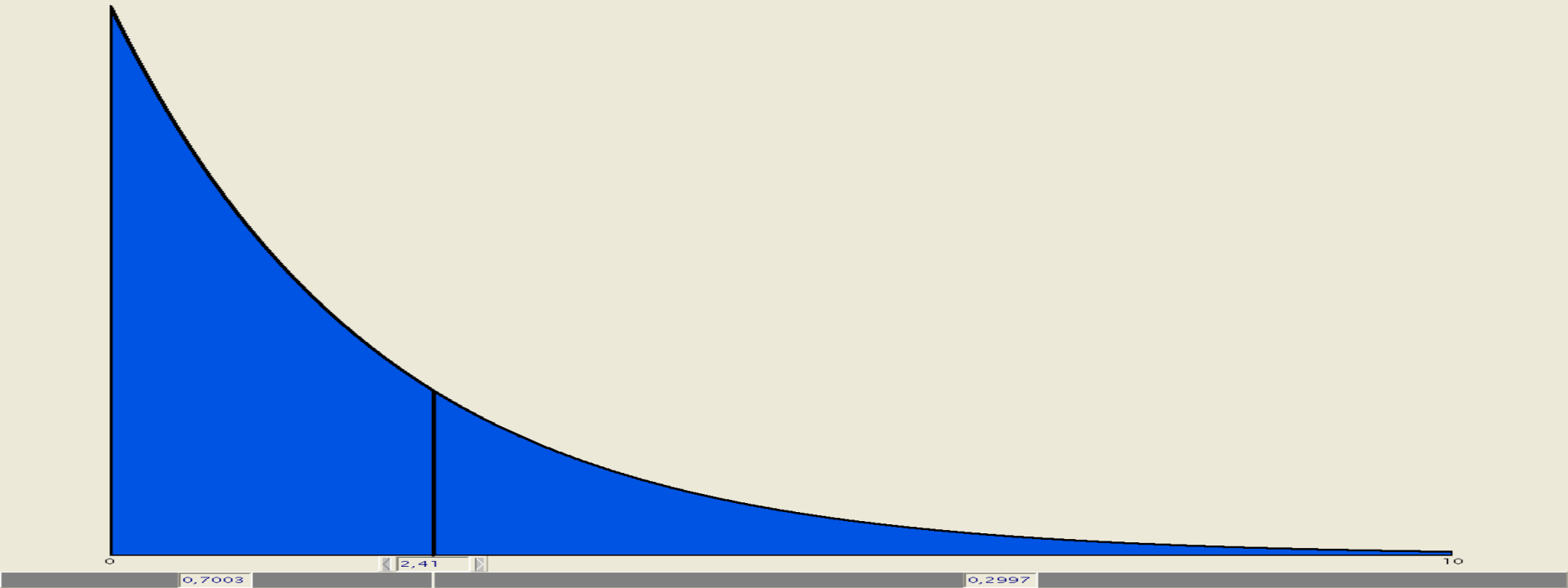
- chí-kvadrát porovná očekávané četnosti s pozorovanými ve všech buňkách
- $\chi^2 = \Sigma$ [(pozor. četnosti – oček.)^2/oček.]
  - čím více se očekávané odchylují od pozorovaných, tím vyšší je statistika X^2, a tím vyšší je evidence proti H_0 o statistické nezávislost (za předpokladu konstantního počtu stupňů volnosti)
  - Pokud mezi pozorovanými a očekávanými četnostmi není rozdíl, pak $\chi^2 = 0$

Vlastnosti chi^2 distribuce výběrových odchylek

- 1) vždy pozitivní - sahá od 0 do ∞ neboť chi^2 statistika sčítá čtverce rozdílů mezi pozorovanými a očekávanými četnosti dělenými očekávanými četnostmi
- 2) tvar závisí na df (stupních volnosti)
- 3) zešikmená, s rostoucími stupni volnosti se normalizuje
- 4) za konstantních df vyšší chi^2 = vyšší evidence proti H_0

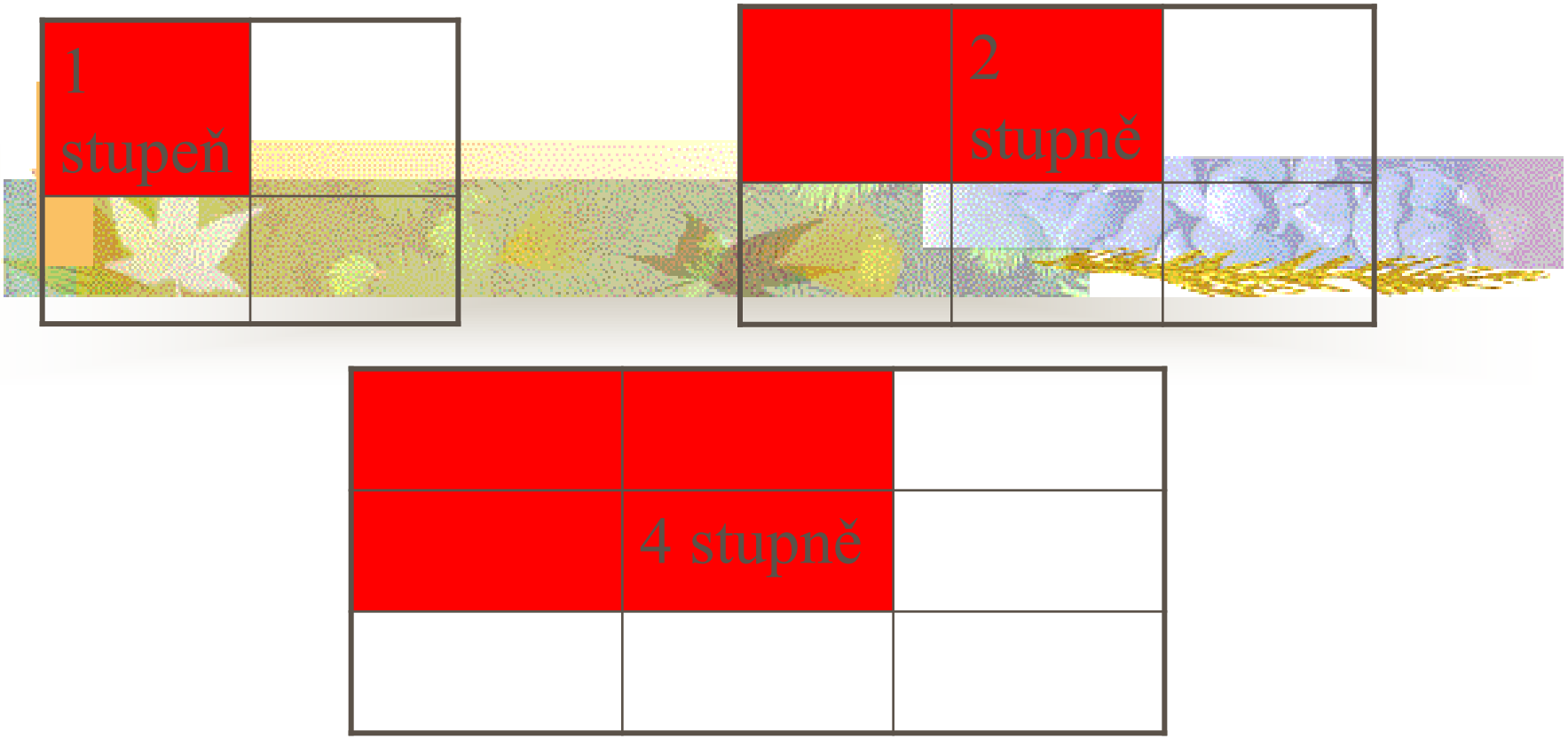


Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		



Stupně volnosti

- = počet hodnot používaných pro výpočet statistiky (např. chí-kvadrát statistiky z tabulky) které nejsou fixní – které se mohou pohybovat (nabývat různých hodnot)



Příklad

- zajímá nás, jak souvisí model manželství s jeho vydařeností
  - model manželství má kategorie:
    dominance žena, dominance muž, kooperace
  - vydařenost má 3 kategorie – vydařené, průměrné, nevydařené
  - pozn.: jde o manželství rodičů respondentů, tak jak je posuzují oni (zdroj: Plaňava)
- otázka zní: liší se podíl (podmíněné proporce) vydařených, průměrných a nevydařených manželství u rodin, kde dominovala matka, rodin, kde dominoval otec a u rodin, kde nedominoval ani jeden z nich?

Absolutní pozorované četnosti

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

Count

		hodnoceni manzelstvi rodicu - muz			
		vydarene	prumerne	nevydarene	Total
model rodic. rodiny - muz	matka dominance	22	29	18	69
	otec dominance	14	19	11	44
	kooperativnost	29	8	4	41
Total		65	56	33	154

Podmíněné proporce

- Vydařená manželství jsou relativně více zastoupena v kooperujících svazcích (70,7%) než v ostatních svazcích (31,9 %, resp. 31,8%)

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

			hodnoceni manzelstvi rodicu - muz			
			vydarene	prumerne	nevydarene	Total
model rodic. rodiny - muz	matka dominance	Count	22	29	18	69
		% within model rodic. rodiny - muz	31,9%	42,0%	26,1%	100,0%
	otec dominance	Count	14	19	11	44
		% within model rodic. rodiny - muz	31,8%	43,2%	25,0%	100,0%
	kooperativnost	Count	29	8	4	41
		% within model rodic. rodiny - muz	70,7%	19,5%	9,8%	100,0%
Total		Count	65	56	33	154
		% within model rodic. rodiny - muz	42,2%	36,4%	21,4%	100,0%

Test Chí-kvadrát

- chí-kvadrát porovnává očekávané a pozorované četnosti v každé buňce
- očekávané četnosti jsou četnosti za předpokladu, že proměnné jsou nezávislé tj, jaké bychom očekávali četnosti v každé buňce, pokud by mezi proměnnými nebyla souvislost?

Očekávané četnosti výpočet

- očekávané četnosti (expected count)
- = (celkový počet v příslušném řádku * celkový počet v příslušném sloupci) / celková velikost vzorku
- Tj. $O_{ij} = (\check{r}_i * s_j) / N$

  tj. pro každou buňku tabulky se vynásobí celkové marginální četnosti z příslušného řádku se sloupcovými četnostmi a vydělí celkovým počtem osob)

Příklad výpočtu očekávané četnosti

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

			hodnoceni manzelstvi rodicu - muz			
			vydarene	prumerne	nevydarene	Total
model rodic. rodiny - muz	matka dominance	Count	22	29	18	69
		Expected Count	29,1	25,1	14,8	69,0
	otec dominance	Count	14	19	11	44
		Expected Count	18,6	16,0	9,4	44,0
	kooperativnost	Count	29	8	4	41
		Expected Count	17,3	14,9	8,8	41,0
Total		Count	65	56	33	154
		Expected Count	65,0	56,0	33,0	154,0

- pro první políčko tabulky (vydařená manželství s dominantní matkou) je očekávaná četnost

$O_{11} = (69*65)/ 154$

$\underline{O_{11} = 29,12}$

Chí-kvadrát statistika: výpočet

- chí-kvadrát porovná očekávané četnosti s pozorovanými ve všech buňkách tabulky
- $\chi^2 = \Sigma$ [(pozor. četnosti – oček.)^2/oček.]
- Př.

$\chi^2 = (-7,1)^2/29,1 + 3,9^2/25,1 + 3,2^2/14,8 +$
$(-4.6)^2/18,6 + 3^2/16 + 1,6^2/9,4 + 11,7^2/17,3 +$
$(-6,9)^2/14,9 + (-4,8)^2/8,8 =$ **18, 71**

Test Chí-kvadrát: stupně volnosti

- pro vyhledání kritické hodnoty χ^2 v tabulce musíme vypočítat počet stupňů volnosti (df)
- **df = (ř-1) (s-1)**

(tj. počet řádků -1 krát počet sloupců -1)

df v našem případě = (3-1) * (3-1) = 4

- v tabulkách vyhledáme kritickou hodnotu χ^2 pro df = 4 a 5% hladinu významnosti
  - **$\chi^2_{krit} = 9{,}49$**

Závěr porovnání vypočtené a kritické hodnoty chi

- $\chi^2_{krit} = 9{,}49$
- $\chi^2 = 18{,}71$
- **závěr:** vypočítaná hodnota je větší než kritická hodnota - očekávané a pozorované četnosti se liší na 5% hladině významnosti (tj. je malá pravděpodobnost, že proměnné jsou nezávislé)

Výsledek v SPSS

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18,712[a]	4	,001
Likelihood Ratio	18,837	4	,001
Linear-by-Linear Association	11,482	1	,001
N of Valid Cases	154		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 8,79.

- Pokud platí nulová hypotéza o nezávislosti mezi proměnnými v populaci, pak pravděpodobnost, že dostanu hodnotu chi=18,712 nebo větší při 4 stupních volnosti je 0,001, proto zamítám nulovou hypotézu.

Chí-kvadrát pro 1 proměnnou

- tzv. test dobré shody (goodness-of-fit test)
- opět porovnává očekávané a pozorované četnosti
- předpokladem očekávaných četností není tentokrát nezávislost proměnných (máme jen 1)

Test dobré shody

- Jak dobře sedí nulovou hypotézou očekávané rozložení pozorovaným datům ve vzorku
- jak určíme očekávané četnosti?
- 2 způsoby:
  - předpoklad vyplývá z teorie nebo ze znalosti parametru v populaci v minulosti
  - nebo můžeme předpokládat náhodné rozdělení do kategorií

Příklad

- je počet sebevražd stejný každý den v týdnu? H_0= proporciální rozložení sebevražd v populaci do jednotlivých dnů je stejné
- zjistíme data pro rok 2000 (ČR)

Příklad

pondělí	255
úterý	247
středa	240
čtvrtek	206
pátek	236
sobota	192
neděle	226

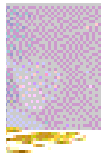
Příklad

- **očekávané četnosti**
  - stejný počet sebevražd pro každý den v týdnu
  - celkem 1602 sebevražd
  - očekávaná četnost pro každý den je 228,9

Příklad

sebevrazdy - den v tydnu

	Observed N	Expected N	Residual
pondeli	255	228,9	26,1
utery	247	228,9	18,1
streda	240	228,9	11,1
ctvrtek	206	228,9	-22,9
patek	236	228,9	7,1
sobota	192	228,9	-36,9
nedele	226	228,9	-2,9
Total	1602		



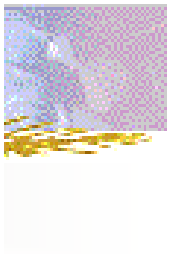
Příklad

- vzorec pro výpočet je stejný
- $\chi^2 = 13{,}44$
- df = k -1 (počet kategorií -1)
- df = 6
- pro df =6 a 5% hladinu významnosti je $\chi^2_{krit} = 12{,}59$
- **rozdíl je statisticky významný**

Výstup v SPSS

Test Statistics

	sebevrazdy - den v tydnu
Chi-Square[a]	13,444
df	6
Asymp. Sig.	,036



a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 228,9.

Omezení (předpoklady) Chí-kvadrátu

- 2 potenciální problémy:
  - malý počet osob – pokud má velké % políček tabulky očekávanou četnost menší než 5 (v ideálním případě by všechna měla mít oček. četnost nejméně 5 osob)
  - příliš velký počet osob – čím vyšší N, tím vyšší χ^2 (vyjdou významné i malé rozdíly)

Míry asociace

- míry asociace vyjadřují **těsnost vztahu proměnných** (a případně **směr** vztahu)
- z chí-kvadrátu se dozvíme pouze, **zda nějaký vztah mezi proměnnými existuje** (tj. zda se liší četnosti pozorované a četnosti očekávané za předpokladu, že proměnné jsou nezávislé)

Míry asociace

- **těsnost (síla) vztahu** – vyjádřena absolutní hodnotou koeficientu
- není shoda v tom, od jaké hodnoty je vztah považován za těsný (někdy uváděno >0.70, jindy >0.30), středně těsný či slabý

Míry asociace

- **směr vztahu** – pouze u ordinálních a kardinálních proměnných
- **pozitivní vztah** – čím vyšší hodnoty jedné proměnné, tím vyšší hodnoty druhé proměnné
- **negativní vztah** - čím vyšší hodnoty jedné proměnné, tím nižší hodnoty druhé proměnné

Míry asociace pro nominální data

- míry asociace pro nominální data ukazují pouze sílu vztahu dvou proměnných, nikoli směr či jiné informace o povaze vztahu

Míry založené na chí-kvadrátu

- velikost hodnoty chí-kvadrát je ovlivněna velikostí výběru a počtem kategorií tabulky
- účelem koeficientů založených na chí-kvadrátu je eliminovat tyto vlivy

Míry založené na chí-kvadrátu

- rozsah koeficientů je obvykle mezi 0 a 1
  - čím vyšší hodnota, tím těsnější vztah
  - 0 – žádný vztah
  - 1 – absolutní vztah (z hodnot jedné proměnné můžeme předpovědět hodnoty druhé proměnné)
- pro koeficienty je možno spočítat statistickou významnost

Míry založené na chí-kvadrátu

- mezi nejčastěji užívané míry asociace založené na chí-kvadrátu patří koeficienty
  - Fí (Phi)
  - Cramerovo V (Cramer's V)

Míry založené na chí-kvadrátu

- **Fí koeficient** - užívá se pro tabulky 2x2 (tj. pro dichotomické proměnné, např. pohlaví)
- vypočte se tak, že se hodnota chí-kvadrátu vydělí počtem osob a výsledek se odmocní

Míry založené na chí-kvadrátu

- Cramerovo V – podobný výpočet jako Fí; počet osob se navíc násobí počtem řádků - 1
  - (pokud je počet řádků menší než počet sloupců, jinak počtem sloupců – 1)
- používá se pro tabulky větší než 2x2

Příklad

- Př. Jak souvisí model manželství s jeho vydařeností?
- Chí-kvadrát = 18.71
- počet osob N = 154
- m = počet řádků - 1 = 3 – 1 = 2

model rodic. rodiny - muz * hodnoceni manzelstvi rodicu - muz Crosstabulation

Count

		hodnoceni manzelstvi rodicu - muz			
		vydarene	prumerne	nevydarene	Total
model rodic. rodiny - muz	matka dominance	22	29	18	69
	otec dominance	14	19	11	44
	kooperativnost	29	8	4	41
Total		65	56	33	154

Výpočet Cramerova V

- tabulka 3x3 – použijeme Cramerovo V
- $V = \sqrt{\chi^2/(N*m)}$
- $V = \sqrt{18.71/(154*2)}$
- **V = 0,246**

Interpretace

- Hodnota V = 0,246 je poměrně nízká – vztah mezi modelem manželství a jeho vydařeností není příliš těsný (i když statisticky významný – viz výstup v SPSS)

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,349	,001
	Cramer's V	,246	,001
N of Valid Cases		154	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.