12. Kontingenční tabulky a χ^2 test

χ^2 test

Shrnutí statistických testů

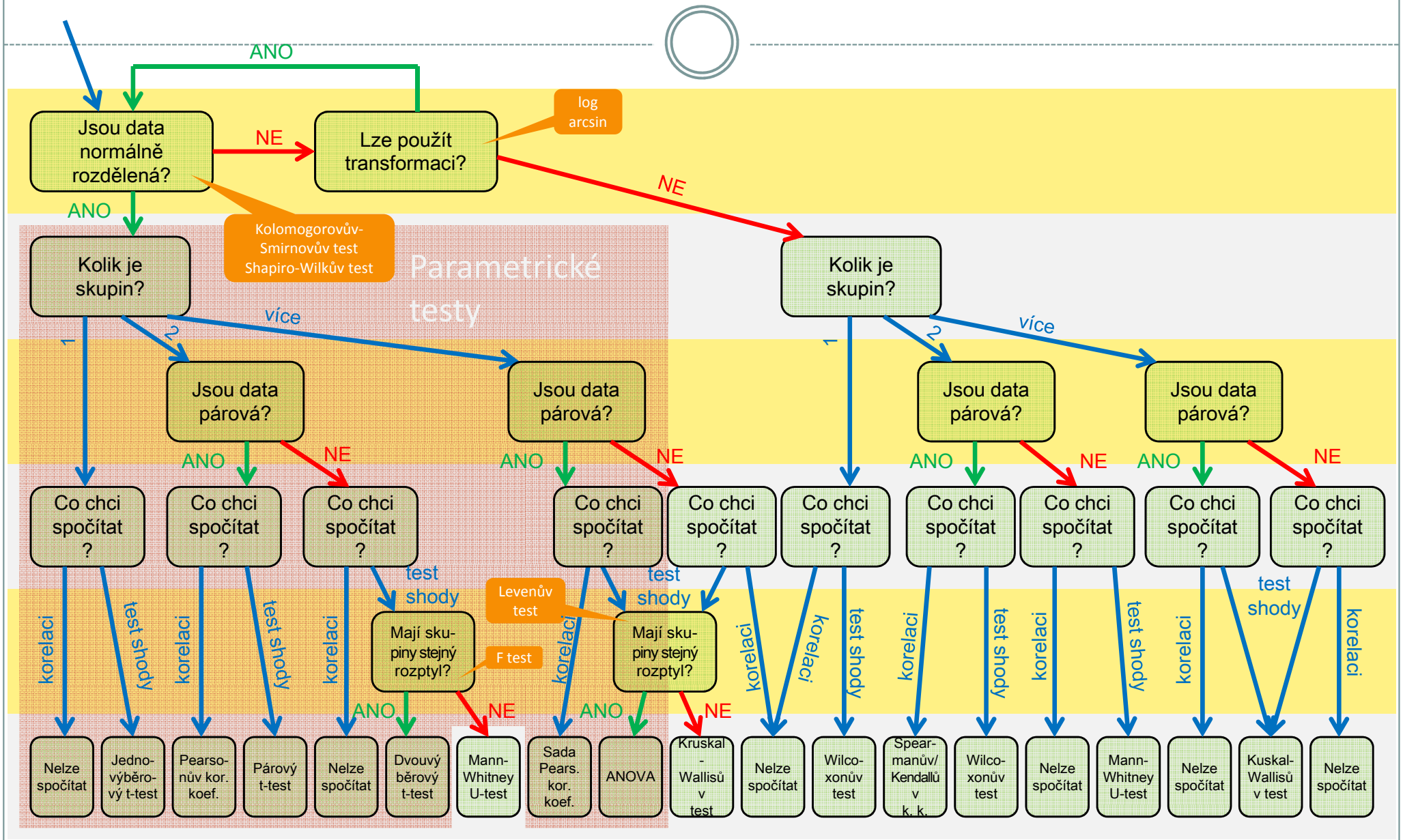
Kontingenční tabulky

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

Shrnutí statistických testů

Typ srovnání	Nulová hypotéza	Parametrický test	Neparametrický test
1 skupina dat vs. etalon	**Střední hodnota je rovna hodnotě etalonu.**	jednovýběrový t-test	Wilcoxonův test; znaménkový test
2 skupiny dat nepárově	**Obě skupiny hodnot pochází ze stejného rozdělení.**	nepárový t-test	Mann-Whitneyův test
2 skupiny dat párově	**Zkoumaný efekt mezi páry hodnot je nulový.**	Párový t-test	Wilcoxonův test; znaménkový test
shoda rozdělení	**rozdělení dat ve skupině odpovídá teoretickému (vybranému) rozdělení.**	Shapiro-Wilkův test; Kolmogorovův-Smirnovův test; Lilieforsův test	$\chi 2$ test, test dobré shody
homoskedasticita (shoda rozptylů)	**rozptyl obou (všech) skupin je shodný.**	Levenův test	
více skupin nepárově	**Zkoumaný efekt mezi skupinami hodnot je nulový.**	ANOVA	Kruskal- Wallisův test
korelace	**Neexistuje (příčinná, důsledková) vazba mezi skupinami hodnot.**	Pearsonův koeficient	Spearmanův koeficient; Kendallův koeficient

Shrnutí statistických testů

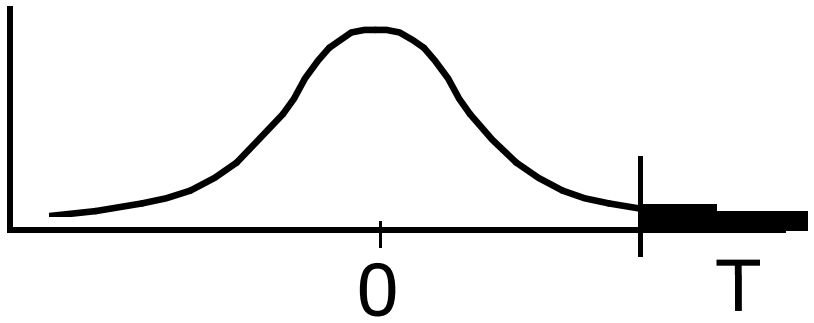


Statistické testování – základní pojmy

- **Nulová hypotéza H_0** — H_0: sledovaný efekt je nulový
- **Alternativní hypotéza H_A** — H_A: sledovaný efekt je různý mezi skupinami
- **Testová statistika**

$$\textbf{Testová statistika} = \frac{\textbf{Pozorovaná hodnota} - \textbf{Očekávaná hodnota}}{\textbf{Variabilita dat}} * \sqrt{\textbf{Velikost vzorku}}$$

- **Kritický obor testové statistiky**



Statistické testování odpovídá na otázku zda je pozorovaný rozdíl náhodný či nikoliv. K odpovědi na otázku je využit statistický model – testová statistika.

P-hodnota

Významnost hypotézy hodnotíme dle získané tzv. p-hodnoty, která vyjadřuje pravděpodobnost, s jakou číselné realizace výběru podporují H_0, je-li pravdivá.

P-hodnotu porovnáme s α (hladina významnosti, stanovujeme ji na **0,05**, tzn., že připouštíme 5 % chybu testu, tedy, že zamítneme H_0, ačkoliv ve skutečnosti platí).

P-hodnotu získáme při testování hypotéz ve statistickém softwaru.

- Je-li **p-hodnota ≤ α**, pak **H_0 zamítáme** na hladině významnosti α a **přijímáme H_A**
- Je-li **p-hodnota > α**, pak **H_0 nezamítáme** na hladině významnosti α

P-hodnota vyjadřuje pravděpodobnost za platnosti H_0, s níž bychom získali stejnou nebo extrémnější hodnotu testové statistiky.

Test dobré shody - základní teorie

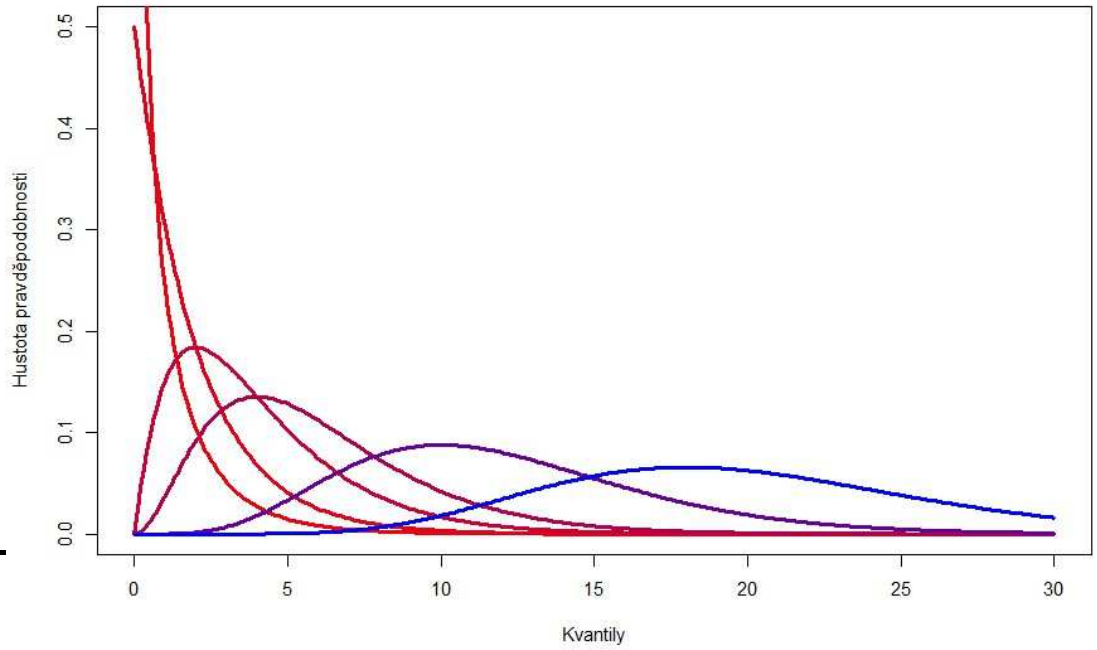
Testuje shodu reálné distribuce hodnot do n skupin s teoretickou distribucí.

Předpokladem je, že velikost rozdílu mezi očekávaným a skutečným počtem hodnot v každé skupině je náhodně rozdělená → multinomické rozdělení.

Součet druhých mocnin relativních rozdílů očekávaného a skutečného počtu hodnot má přibližně χ^2 rozdělení.

χ^2 rozdělení pro kladné hodnoty (suma čtverců) se liší podle počtu stupňů volnosti *k* (počtu skupin) - se zvyšujícím se *k* přechází v normální rozdělení.

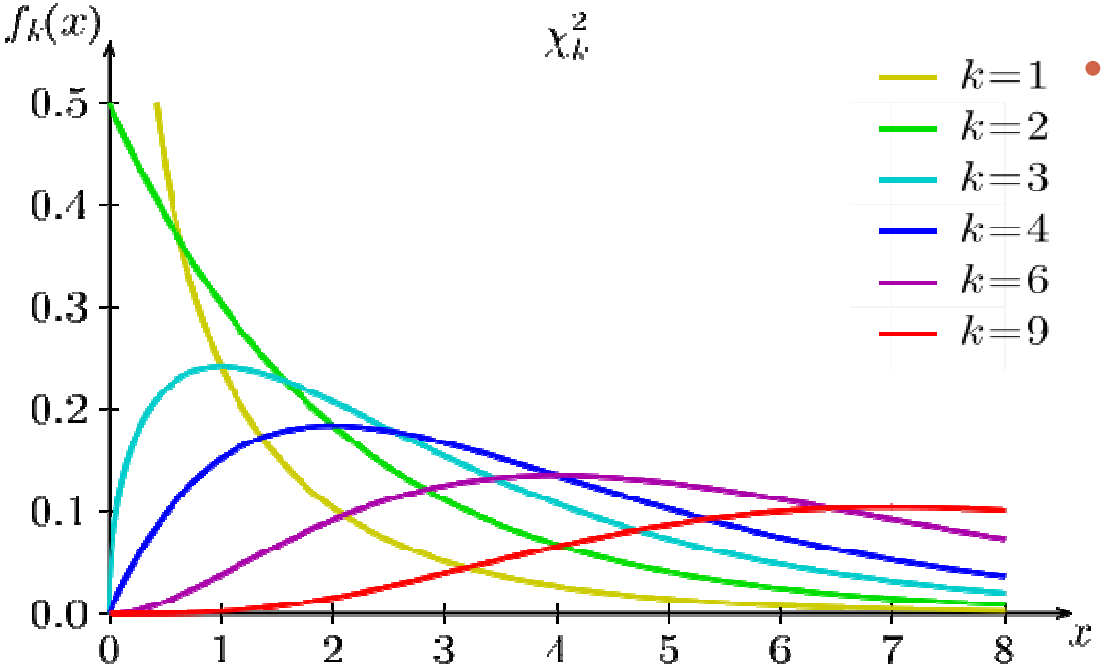
$$\chi^2_{(s.v.)} = \sum \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$



Test dobré shody – stupně volnosti

Počet stupňů volnosti je roven počtu nezávislých skupin vstupujících do χ^2 testu:

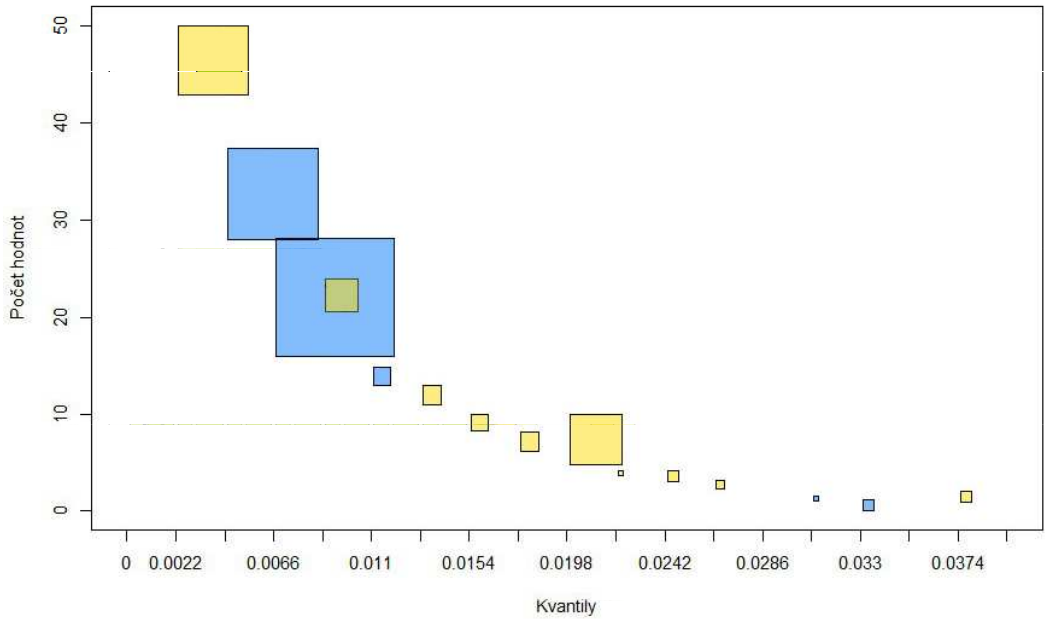
- V případě jednorozměrné distribuce (např. test shody rozdělení n hodnot náhodné veličiny) je roven n-1



- V případě vícerozměrné distribuce (např. test shody očekávaných a pozorovaných hodnot v kontingenční tabulce m × n) je roven (m-1)×(n-1).

Test dobré shody - základní teorie

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



Očekávané četnosti

V případě platnosti nulové hypotézy je poměr mezi buňkami jednoho sloupce v různých řádcích nezávislý na výběru tohoto sloupce.

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

Pokud tyto poměry normalizujeme, získáváme tabulku očekávaných četností.

Řádkové a sloupcové součty se touto operací nemění.

Pozorované četnosti

	Ano	Ne	Σ
Ano	20	82	**102**
Ne	10	54	**64**
Σ	**30**	**136**	**166**

102 × 30 / 166

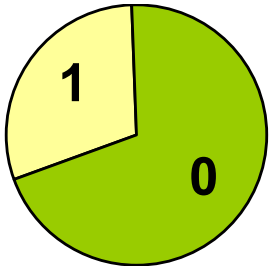
Očekávané četnosti

	Ano	Ne	Σ
Ano	18,4	83,6	**102**
Ne	11,6	52,4	**64**
Σ	**30**	**136**	**166**

Test dobré shody - základní teorie

Binomické jevy (1/0)

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



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 ?**

$$\chi^2_{(1)} = \frac{(4000 - 5000)^2}{5000} + \frac{(6000 - 5000)^2}{5000} = 400$$

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

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

Kontingenční tabulky
H0 :Nezávislost dvou jevů A a B

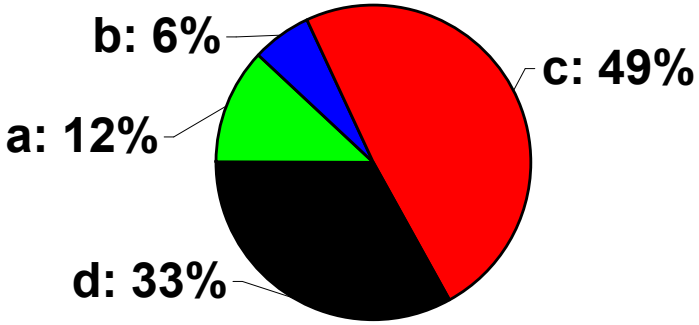
Kontingenční tabulka 2 x 2

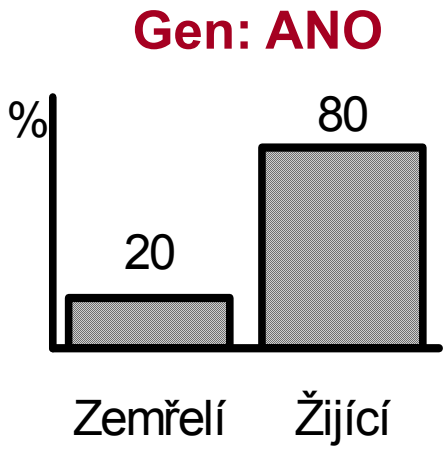
B \ A	+	-	Σ
+	a	b	
-	c	d	
Σ			**suma sum**

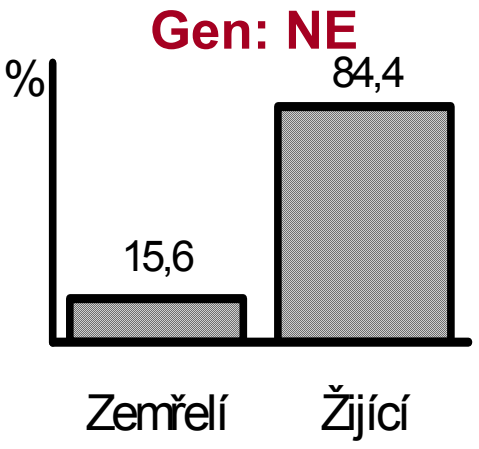
Kontingenční tabulky: příklad

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

Kontingenční tabulka v obrázku







Příklad – závislost pohlaví na onemocnění

	Zdraví	Nemocní	Celkem
Muži	50	50	100
Ženy	50	50	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	45	55	100
Ženy	55	45	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	35	65	100
Ženy	65	35	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	5	95	100
Ženy	95	5	100
Celkem	100	100	200

Příklad – závislost pohlaví na onemocnění

Pozorované hodnoty

	Zdraví	Nemocní	Celkem
Muži	50	50	100
Ženy	50	50	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	45	55	100
Ženy	55	45	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	35	65	100
Ženy	65	35	100
Celkem	100	100	200

	Zdraví	Nemocní	Celkem
Muži	5	95	100
Ženy	95	5	100
Celkem	100	100	200

Očekávané hodnoty pro <u>všechny</u> tabulky vlevo

	Zdraví	Nemocní	Celkem
Muži	50	50	100
Ženy	50	50	100
Celkem	100	100	200

$$\chi^2_{(s.v.)} = \sum \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$

Příklad – závislost pohlaví na onemocnění

	Zdraví	Nemocní	Celkem
Muži	50	50	100
Ženy	50	50	100
Celkem	100	100	200

$X^2 = 0,0$
$p = 1,000$

	Zdraví	Nemocní	Celkem
Muži	45	55	100
Ženy	55	45	100
Celkem	100	100	200

$X^2 = 2,0$
$p = 0,157$

	Zdraví	Nemocní	Celkem
Muži	35	65	100
Ženy	65	35	100
Celkem	100	100	200

$X^2 = 18,0$
$p < \mathbf{0,0001}$

	Zdraví	Nemocní	Celkem
Muži	5	95	100
Ženy	95	5	100
Celkem	100	100	200

$X^2 = 162,0$
$p < \mathbf{0,0001}$