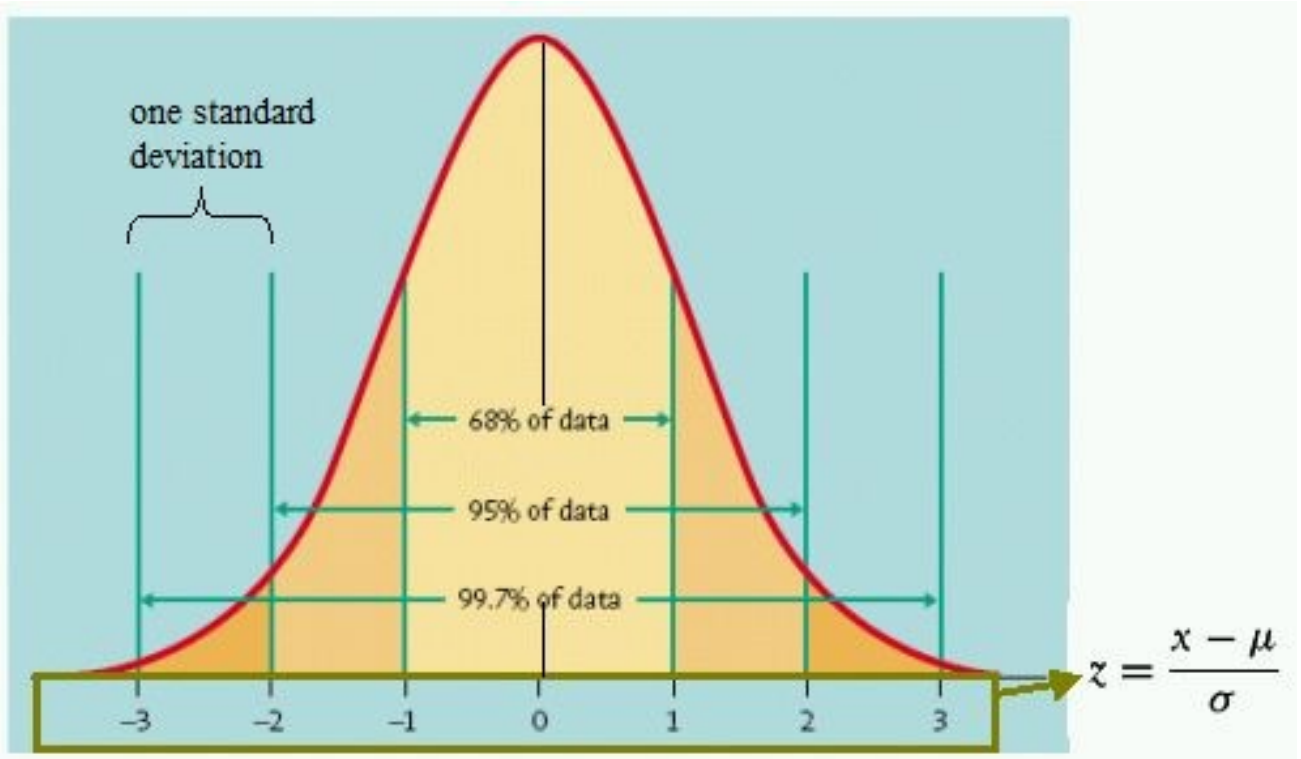
FSS MU, Katedra SPSP
Kvantitativní výzkum x118

Téma 12:
Chí-kvadrát distribuce a její využití pro test souvislosti mezi dvěma kategorickými proměnnými

Autor: Miroslav Suchanec

Logika vzniku chí-kvadrát distribuce:

- 1) Mějme normálně rozloženou distribuci v populaci, kterou standardizujeme

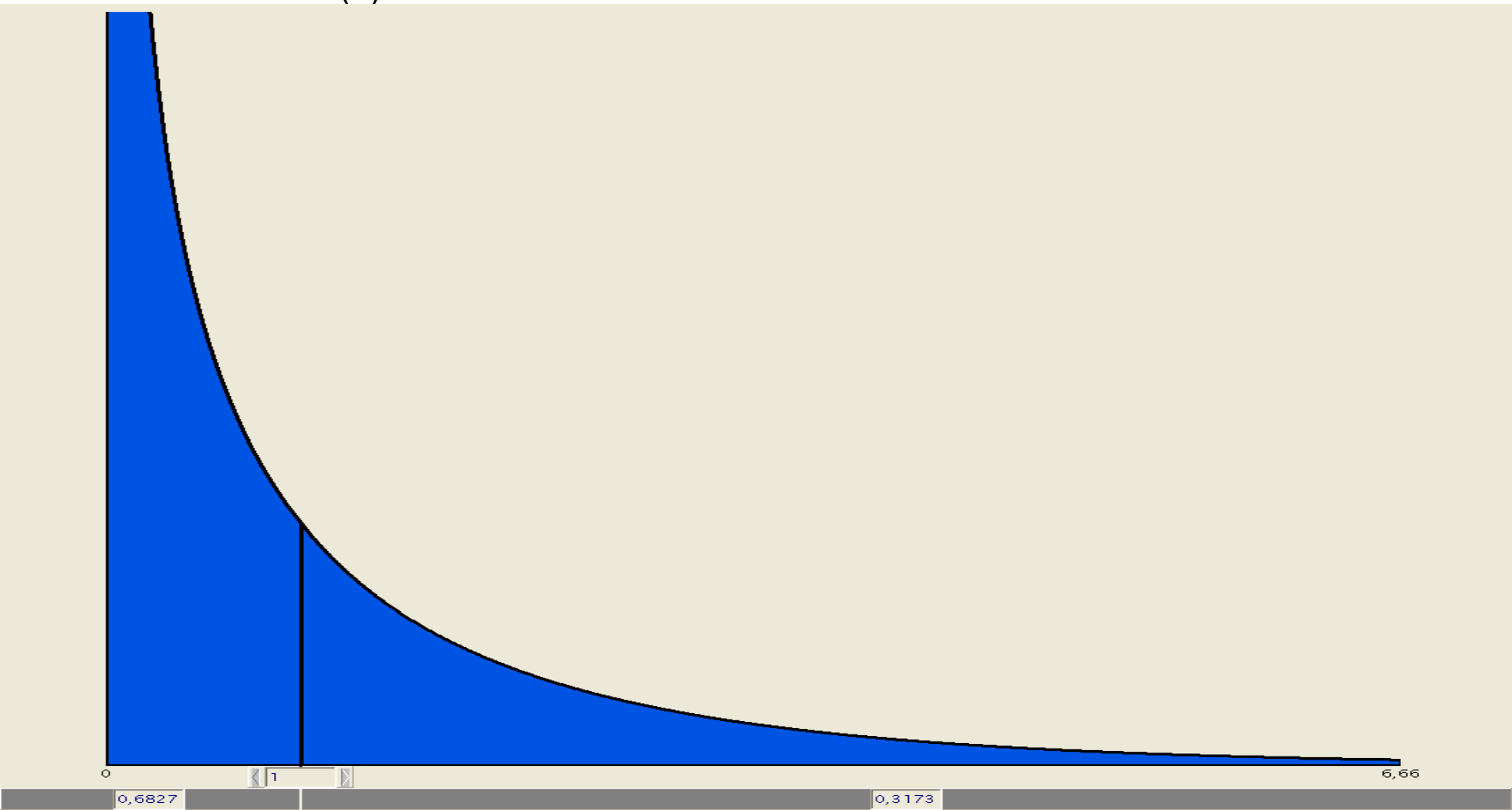


- 2) Z této distribuce vybereme náhodně jeden případ (N=1) a umocníme na druhou

- Př. z = 0,54

  $z^2 = 0,54*0,54 = 0,29$

- 3) Opakovali-li bychom tento postup vznikne distribuce výběrových „odchylek“ neboli distribuce náhodné proměnné $X^2_{(1)}$ neboli chí-kvadrát distribuce s jedním stupněm volnosti
- Tedy: $X^2_{(1)} = z^2$



- Výsledné vlastnosti distribuce:
- 1) sahá od 0 do ∞ protože cokoli na druhou je pozitivní číslo
- 2) zešikmená s vysokou pravděpodobností hodnot 0 až 1

  (protože pravděpodobnost z = -1 až +1 = 68% a tedy pravděpodobnost z^2 = 0 až 1 = 68%)

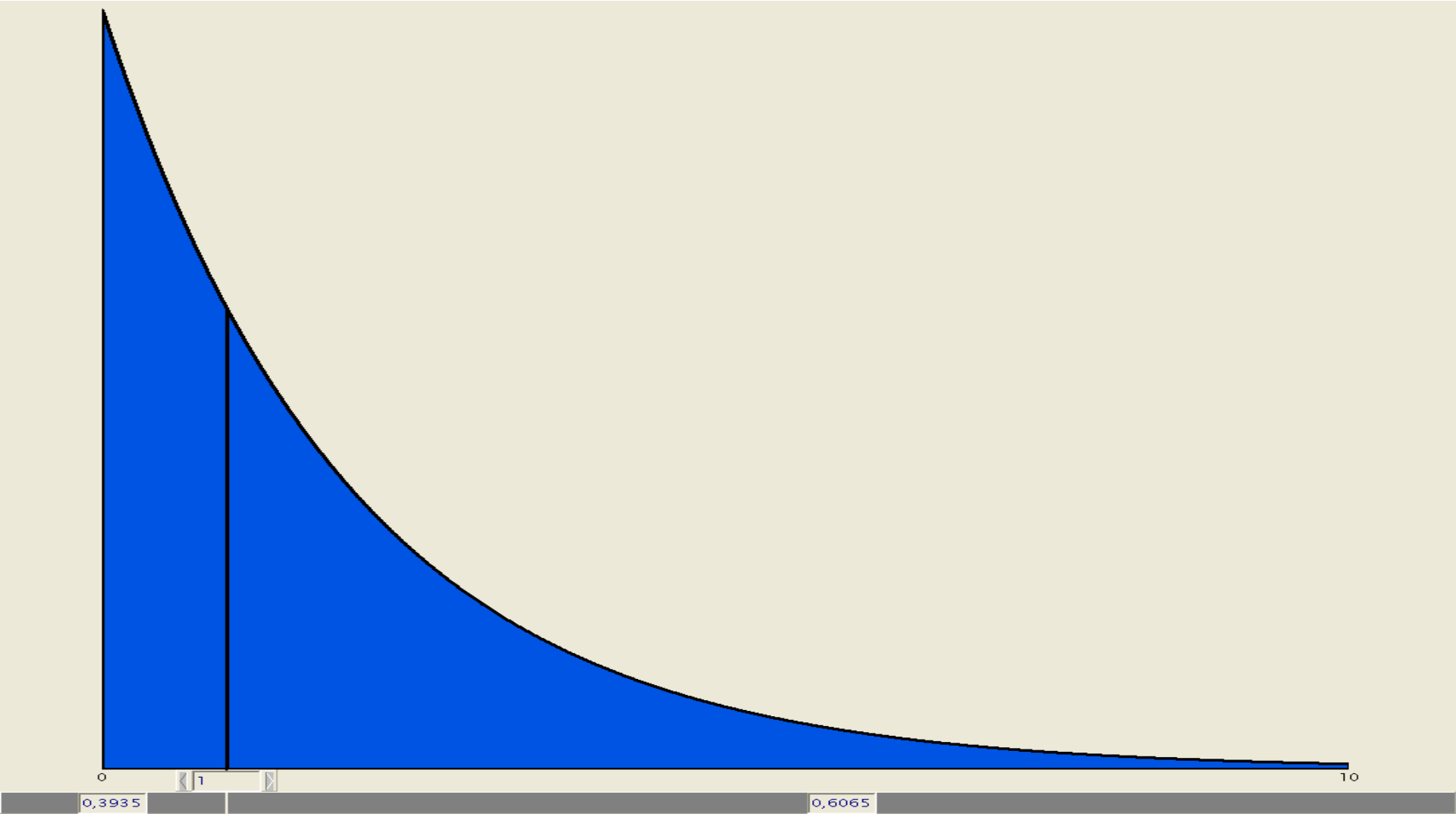
- Představme-si že vytáhneme náhodně a nezávisle vzorek o velikosti N=2 a každou z hodnot umocníme na druhou
- Př. $z_1 = 0,54$

  $z_1^2 = 0,54*0,54 = 0,29$

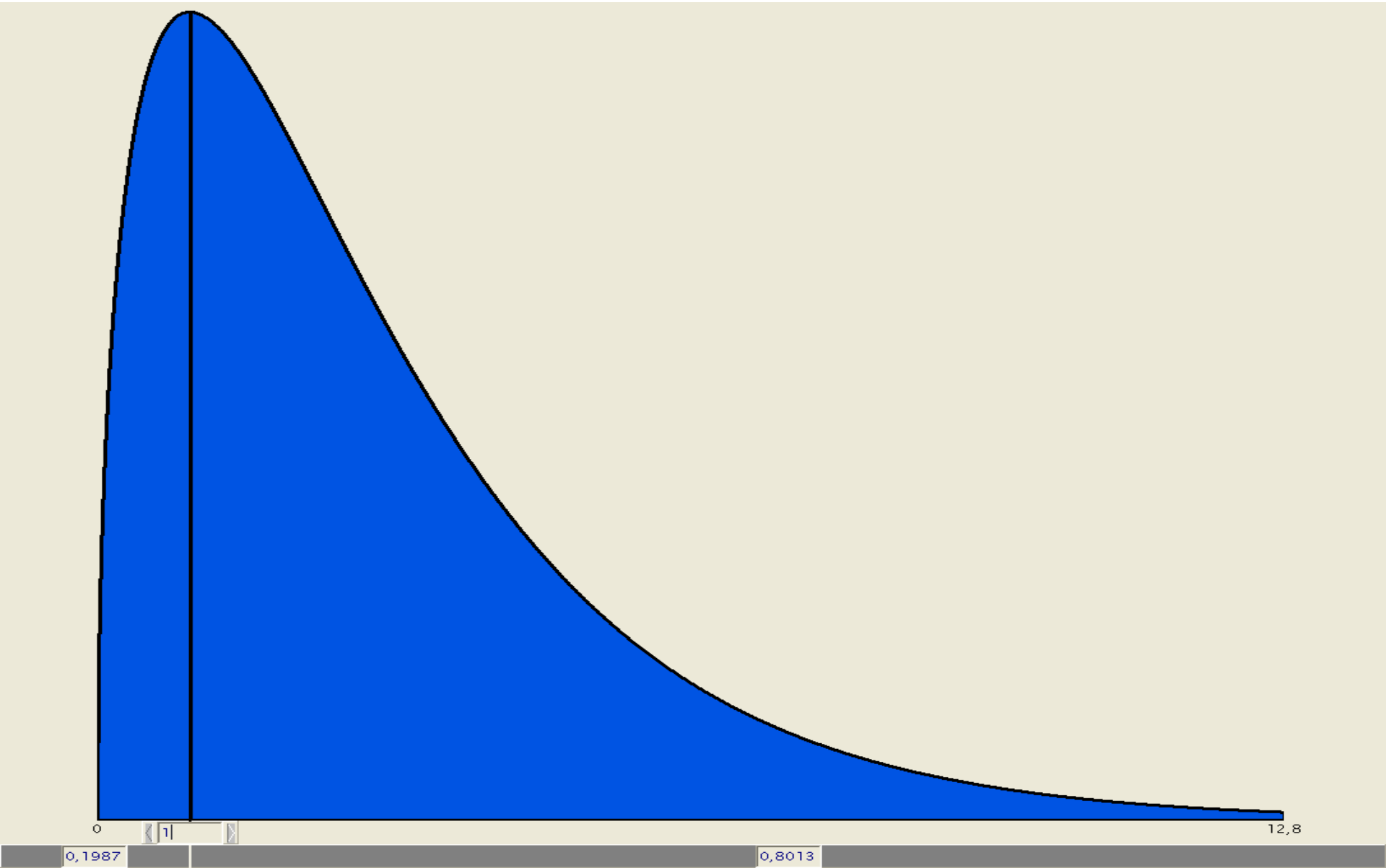
  $z_2 = -0,78$

  $z_2^2 = -0,78 * -0,78 = 0,61$

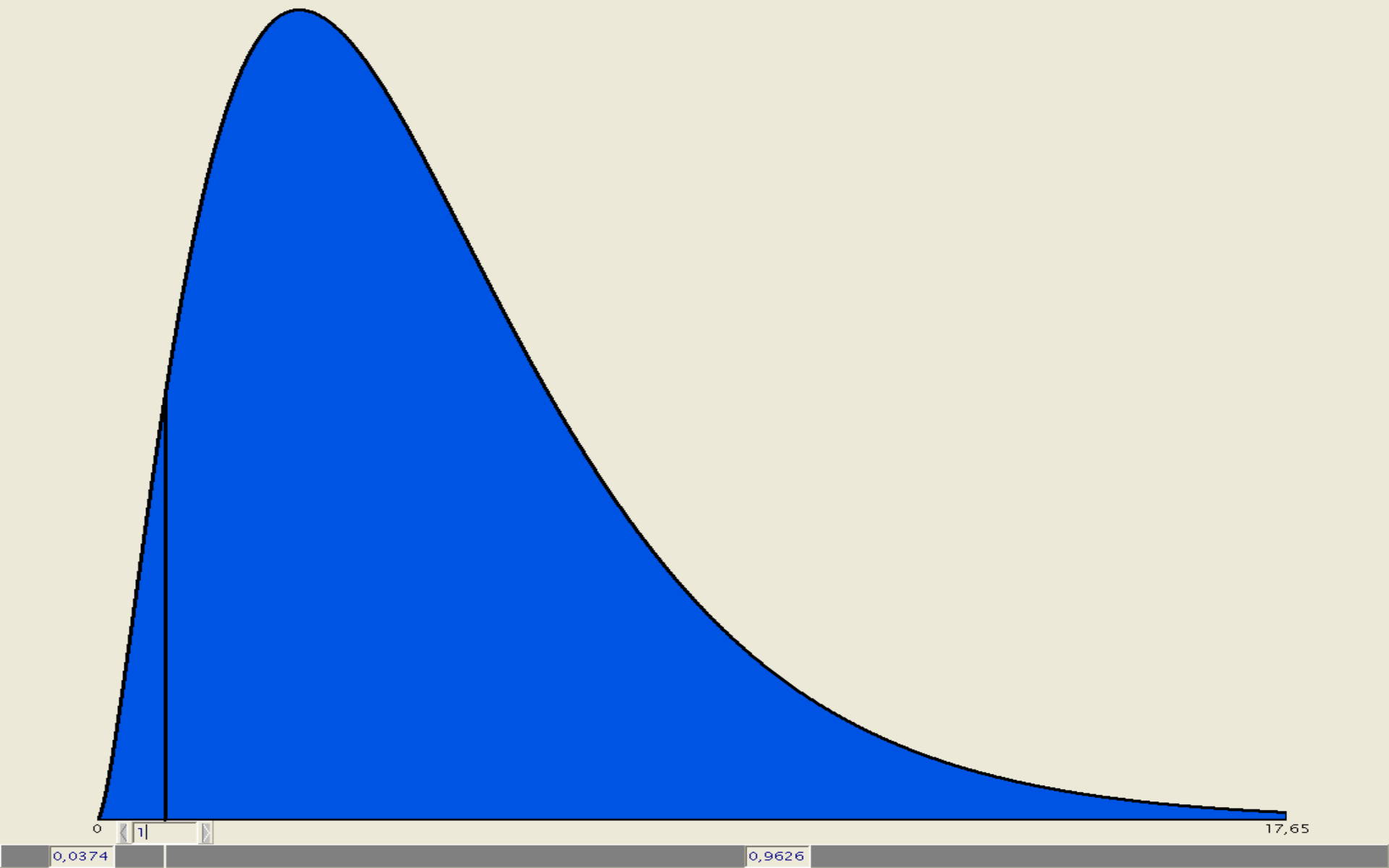
- V tomto případě distribuce náhodné proměnné $X^2_{(2)}$ neboli chí-kvadrát distribuce s dvěma stupněmi volnosti bude dána součtem z^2
- Tedy: $X^2_{(2)} = z_1^2 + z_2^2$

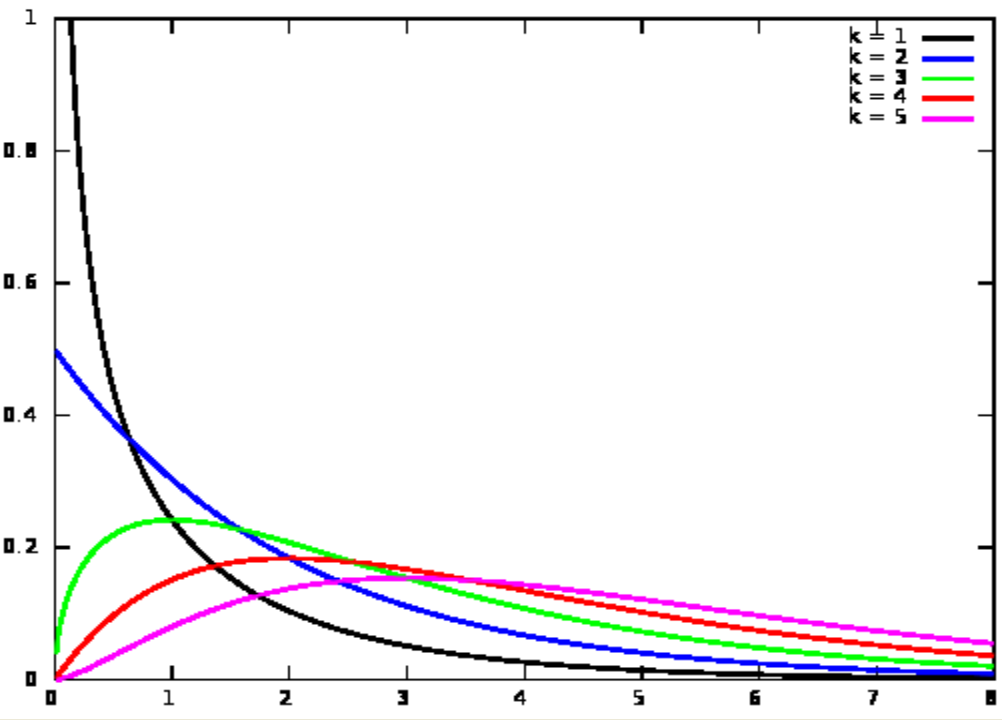


- Obdobně pro N=3
- $X^2_{(3)} = z_1^2 + z_2^2 + z_3^2$

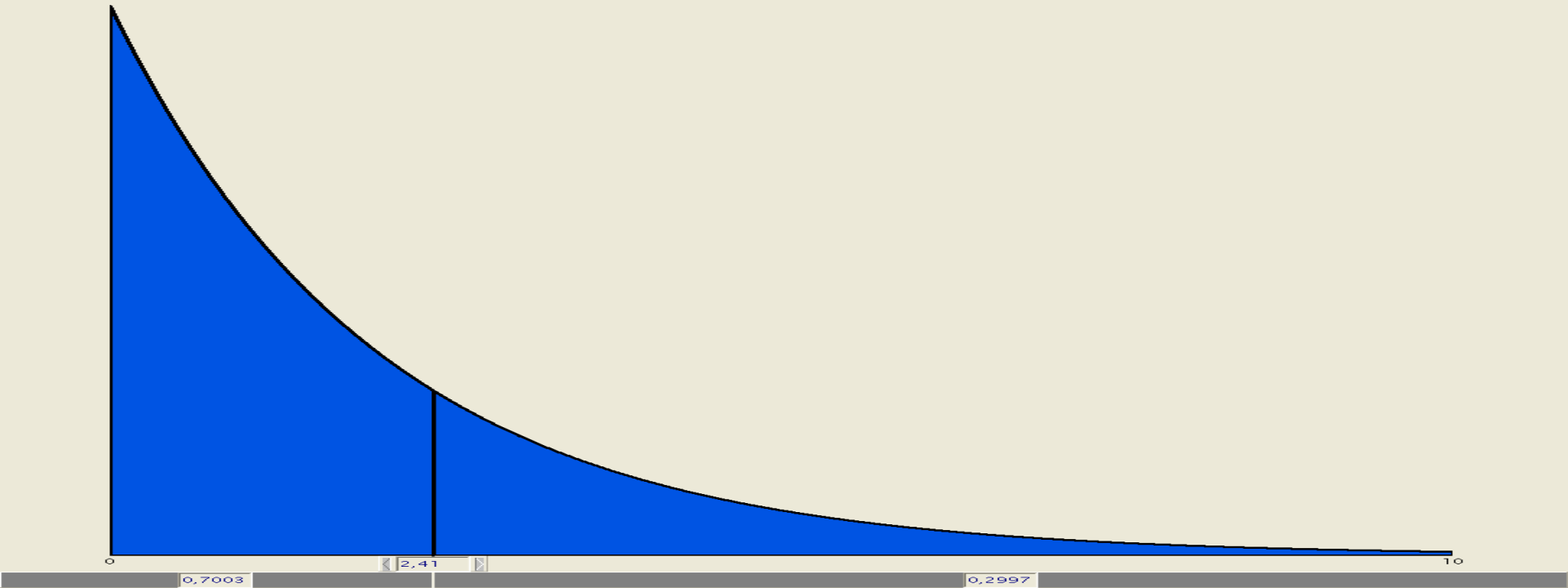


- Pro N=5
- $X^2_{(5)} = z_1^2 + z_2^2 + z_3^2 + z_4^2 + z_5^2$





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		



Využití pro testování souvislosti mezi dvěma kategorickými proměnnými

Základní postup:

V tzv. kontingenční tabulce porovnáváme pozorované a očekávané hodnoty každé kombinace, čím více se očekávané odchylují od pozorovaných, tím vyšší je statistika X^2, a tím vyšší je statistická závislost (za předpokladu konstantního počtu stupňů volnosti)

- Vzorec: $X^2 = \Sigma ((O - E) / E))$

P (chlapec)= .55	(chlapec a „ano“)	(chlapec a „ne“)
P(Dívka) = .45	(dívka a „ano“)	(dívka a „ne“)
	P(Ano) = .40	P(Ne) = .60

- Jevy/třídy jevů vyskytující se podél okraje jsou vzájemně neslučitelné a vyčerpávající (sada těchto jevů tedy formuje S) = dimenze
  - Př. 2 dimenze: jevy „ano“ a „ne“ a jevy „chlapec“ a „dívka“
- Statistická nezávislost : každá kategorie nebo jev podél jednoho okraje musí být nezávislý na každém jevu podél druhého okraje = pravděpodobnost každého spojeného jevu se musí rovnat součinu pravděpodobností korespondujících (v řádku a sloupci) marginálních jevů - p (A ∩ *B*) = p (A) * p (B)
  - Př. Pokud dimenze „pohlaví“ a „odpověď ano/ne“ jsou nezávislé,
    pak p(chlapec a „ano“) = p(chlapec) * p(„ano“) = .55 * .40 = .22
    Stejně postupujeme v ostatních případech a vznikne tabulka:

P (chlapec)= .55	p(chlapec)p („ano“) =.55 * .40 = .22	p(chlapec)p(„ne“) = .55 * .60 = .33
P(Dívka) = .45	p(dívka)p(„ano“) =.45 * .40 = .18	p(dívka)p(„ne“) = .45 * .60 = .27
	P(Ano) = .40	P(Ne) = .60

Příklad:
Testování efektivity dvou léků oproti placebu při prevenci chřipky

	Lék 1	Lék 2	Placebo (cukr)	Celkem
nemocný	20	30	30	80
zdravý	100	110	90	300
Celkem	120	140	120	380

H0: Léky nemají žádný efekt

H1: Léky mají nějaký (pozitivní nebo negativní) efekt

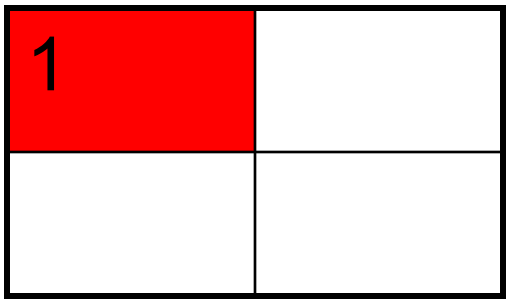
alfa = 0,1

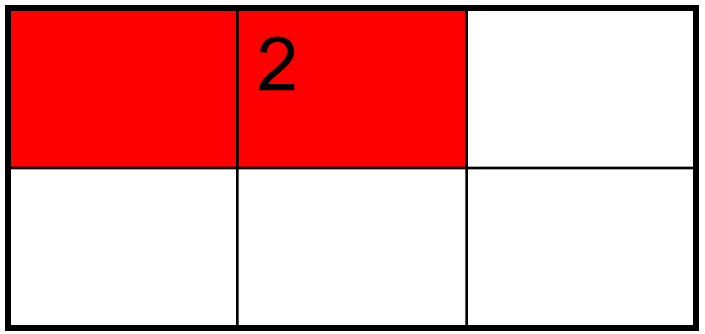
	Lék 1	Lék 2	Placebo (cukr)	Celkem
nemocný	20 25	30 29	30 25	80 (21%)
zdravý	100 95	110 111	90 95	300 (79%)
Celkem	120 (32%)	140 (37%)	120 (32%)	380 (100 %)

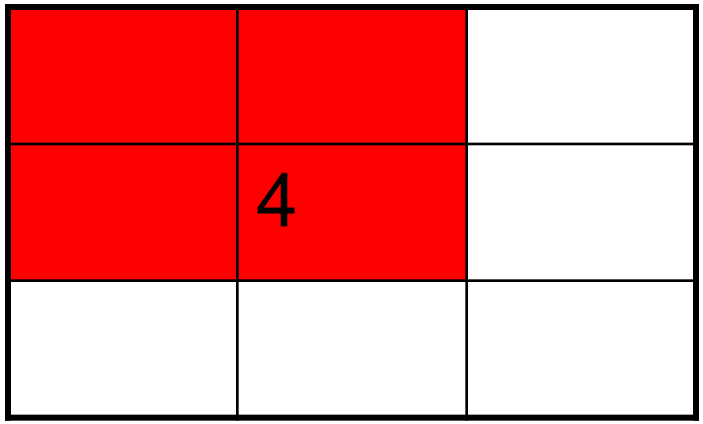
- Černá = pozorované (observed = „O“) hodnoty
- Červená = očekávané (expected = „E“) hodnoty
- např. E (lék1, nemocný)=p(lék1)*p(nemocný)*celkový počet = 0.21*0.32*380= 25
- Alternativní postup zjištění očekávané hodnoty: E (lék1, nemocný)= 0,21 * 120 = 25

Stupně volnosti

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







	Lék 1	Lék 2	Placebo (cukr)	Celkem
nemocný	20 25	30 29	30 25	80 (21%)
zdravý	100 95	110 111	90 95	300 (79%)
Celkem	120 (32%)	140 (37%)	120 (32%)	380 (100 %)

2 stupně volnosti

Výpočet chí-kvadrát statistiky a stupňů volnosti

$$X^2 = \Sigma [(O – E)^2 / E)]$$

$$= (20 – 25)^2 / 25$$

$$+ (30 - 29)^2 / 29$$

$$+ (30 - 25)^2 / 25$$

$$+ (100 – 95)^2 / 95$$

$$+ (110 - 111)^2 / 111$$

$$+ (90 - 95)^2 / 95$$

$$= 2,53$$

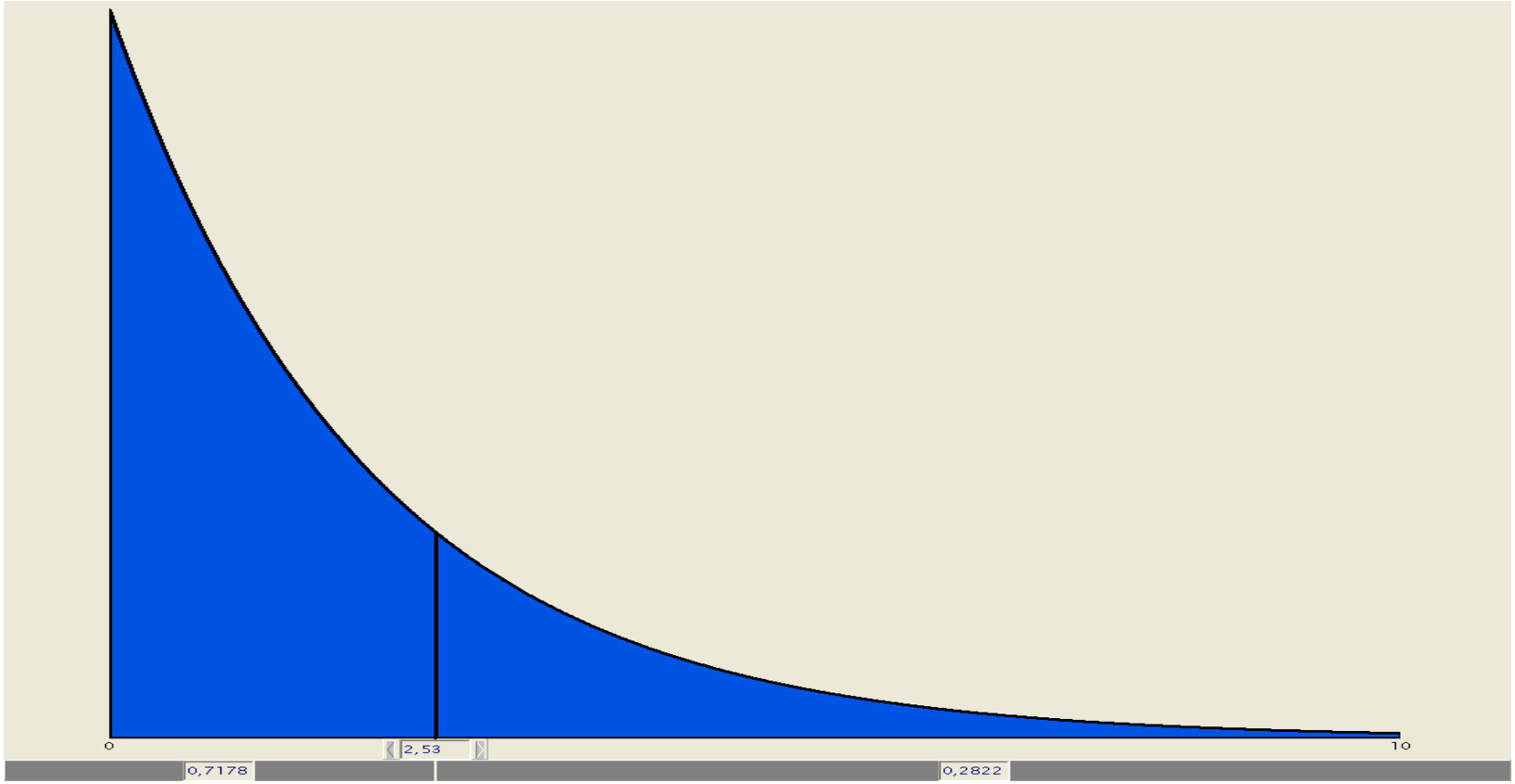
Stupně volnosti: (ř – 1) * (s – 1) = (2 – 1) * (3 – 1) = 1 * 2 = 2

$$X^2_{(2)} = 2,53$$

Závěr pomocí tabulky

- Kritická hodnota pro alfa=0,1 a 2 stupně volnosti je $x^2 = 4,6$ což je více než 2,53. Hodnota 2,53 nespadá do regionu zamítnutí proto H0 nezamítám

Závěr pomocí PQRS



- Pokud léky nemají efekt (tj. H0 je pravdivá) pak pravděpodobnost hodnoty 2,53 nebo extrémnější je 0,28 což je docela velká pravděpodobnost (více než 0,1) a proto H0 nezamítám

Závěr pomocí SPSS

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	**2,526[a]**	**2**	**,283**
Likelihood Ratio	2,559	2	,278
Linear-by-Linear Association	2,500	1	,114
N of Valid Cases	380		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 25,26.

- Pokud léky nemají efekt (tj. H0 je pravdivá) pak pravděpodobnost hodnoty 2,53 nebo extrémnější je 0,28 což je docela velká pravděpodobnost (více než 0,1) a proto H0 nezamítám