Test χ^2 (chí kvadrát)

Test χ^2 (chí kvadrát)

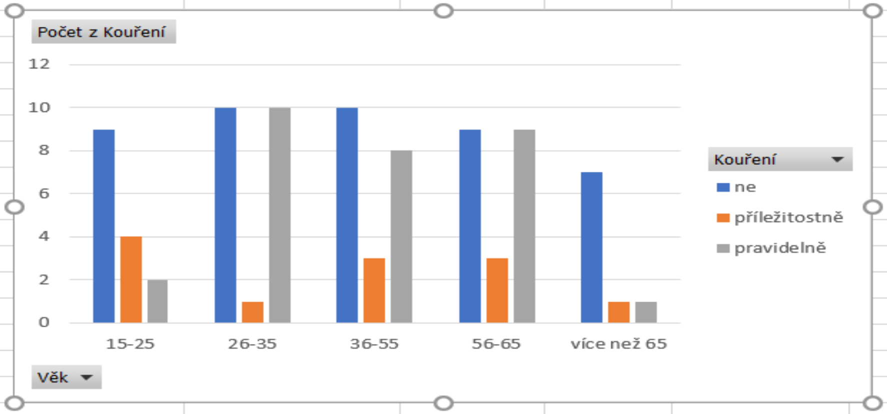
- Máme k dispozici náhodný výběr rozsahu *n* rozdělený do dvou znaků (znak 1, znak 2)
- Normální (Gaussovo rozdělení)
- Úkolem testu je rozhodnout, zda jsou znaky na sobě závislé nebo nezávislé (zda znak 1 má vliv na znak 2)
- Znak 1
- Znak 2

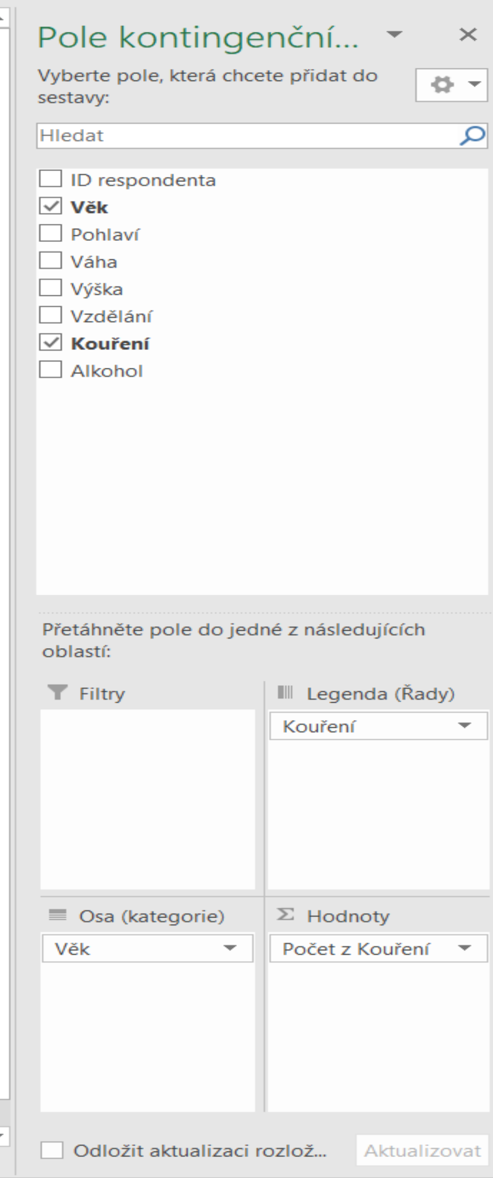
Surová data

ID respondenta	Věk	Pohlaví	Váha	Výška	Vzdělání	Kouření	Alkohol
1	36-55	žena	75	176	SŠ	ne	ne
2	26-35	žena	76	178	SŠ	pravidelně	ne
3	15-25	muž	90	183	SŠ	ne	příležitostně
4	36-55	žena	80	178	SŠ	ne	ne
5	15-25	žena	70	178	VŠ	ne	ne
6	15-25	muž	90	191	SŠ	ne	ne
7	26-35	žena	69	173	VŠ	ne	ne
8	56-65	žena	139	178	SŠ	ne	ne
9	36-55	žena	72	180	SŠ	ne	ne
10	26-35	žena	65	173	VŠ	ne	ne
11	56-65	muž	93	185	SŠ	ne	pravidelně
12	36-55	žena	70	175	ZŠ	pravidelně	pravidelně
13	56-65	žena	64	164	SŠ	pravidelně	ne
14	56-65	muž	70	171	ZŠ	ne	pravidelně
15	56-65	muž	67	162	SŠ	pravidelně	pravidelně
16	36-55	muž	89	178	SŠ	příležitostně	pravidelně
17	36-55	žena	63	165	SŠ	pravidelně	ne
18	56-65	muž	99	187	VŠ	ne	pravidelně
19	56-65	žena	85	155	SŠ	pravidelně	příležitostně
20	15-25	žena	53	170	SŠ	ne	ne
21	15-25	muž	82	190	SŠ	ne	ne
22	56-65	žena	69	152	SŠ	pravidelně	ne
23	36-55	muž	66	175	SŠ	pravidelně	příležitostně
24	56-65	muž	110	177	ZŠ	příležitostně	příležitostně
25	15-25	žena	65	175	VŠ	ne	příležitostně
26	15-25	žena	57	157	VŠ	příležitostně	příležitostně
27	56-65	muž	90	185	SŠ	příležitostně	ne
28	26-35	muž	70	190	VŠ	příležitostně	příležitostně
29	36-55	muž	94	180	SŠ	příležitostně	příležitostně
30	56-65	muž	75	170	SŠ	pravidelně	příležitostně
31	26-35	žena	54	169	VŠ	pravidelně	ne
32	26-35	žena	60	170	VŠ	ne	příležitostně
33	36-55	žena	62	171	VŠ	pravidelně	ne
34	56-65	žena	49	58	SŠ	ne	pravidelně
35	15-25	muž	60	183	SŠ	pravidelně	příležitostně
36	26-35	žena	50	160	SŠ	pravidelně	pravidelně
37	36-55	žena	68	153	SŠ	ne	ne
38	56-65	žena	65	167	ZŠ	pravidelně	příležitostně
39	56-65	žena	45	150	ZŠ	příležitostně	pravidelně
40	15-25	žena	54	155	SŠ	ne	příležitostně
41	36-55	žena	75	172	ZŠ	příležitostně	ne
42	26-35	žena	69	168	ZŠ	pravidelně	příležitostně

Kontingenční tabulka

Počet z Kouření	Popisky sloupců			
Popisky řádků	**ne**	**příležitostně**	**pravidelně**	**Celkový součet**
15-25	9	4	2	15
26-35	10	1	10	21
36-55	10	3	8	21
56-65	9	3	9	21
více než 65	7	1	1	9
Celkový součet	**45**	**12**	**30**	**87**





Hypotézy

- Nulová hypotéza: znaky 1 a 2 jsou nezávislé
- Alternativní hypotéza: mezi znaky 1 a2 existuje závislost

Chyby testu

- Chyba 1. druhu - hladina významnosti

  Chceme ji mít pod dostatečnou kontrolou. Požadujeme, aby pravděpodobnost chyby 1. druhu nepřekročila námi předem zvolenou mez α, tzv. hladinu testu, volíme zpravidla α = 0,05 nebo 0,01
- Chyba 2. druhu

  Snažíme se ji minimalizovat
- Obě chyby jsou vzájemně nepřímo úměrné. Jestliže H_0 platí (tedy), pravděpodobnost zamítnutí H_0 má být menší než α

1. Hypotéza H_0 platí, hypotézu H_0 zamítneme (chyba 1. druhu),
2. Hypotéza H_0 platí, hypotézu H_0 nezamítneme,
3. Hypotéza H_0 neplatí, hypotézu H_0 zamítneme,
4. Hypotéza H_0 neplatí, hypotézu H_0 nezamítneme (chyba 2. druhu)

Chyba testu

- Podobná situace nastává u soudu, kde roli nulové hypotézy hraje presumpce neviny obžalovaného. Soudce na základě předložených důkazů zamítne jeho nevinu a odsoudí ho k trestu nebo naopak nezamítne jeho nevinu a neodsoudí ho, čímž však nijak netvrdí, že obžalovaný je skutečně nevinen. Buď je nevinen, nebo k prokázaní jeho viny nemá soudce dostatek důkazů.
- Stejně ve statistice, jestliže nulovou hypotézu nezamítneme, neznamená to ještě, že H_0 skutečně platí. Buď je pravdivá, nebo pro její zamítnutí nemáme dostatek potřebných měření, dostatek informací.

Chyba testu

1. Nevinen, odsouzen - H_0 platí, H_0 zamítneme (chyba 1. druhu)
2. Nevinen, neodsouzen - H_0 platí, H_0 nezamítneme
3. Vinen, odsouzen - H_0 neplatí, H_0 zamítneme
4. Vinen, neodsouzen - H_0 neplatí, H_0 nezamítneme (chyba 2. druhu)

	Vinen	Nevinen
Odsouzen	Pravda	Nepravda - chyba 1. druhu
Neodsouzen	Nepravda – chyba 2. druhu	Pravda

V každém soudním procesu se musí hledat jistá rovnováha mezi tvrdostí a mírností

Extrém 1 - mírný soudce, který k usvědčení obžalovaného vyžaduje velké množství důkazů
zřídka odsoudí nevinného (zřídka se dopustí chyby prvního druhu)
často osvobodí viníka (často se dopustí chyby druhého druhu)

Extrém 2 - přísný soudce, kterému k usvědčení stačí jen několik důkazů, posílá do vězení i jen při stínu podezření
častěji odsoudí nevinného (často se dopustí chyby prvního druhu)
zřídka osvobodí darebáka (zřídka se dopustí chyby druhého druhu)

Otázka: Která z chyb je závažnější: chyba prvního druhu, nebo chyba druhého druhu?
Má se za to, že závažnější je uvěznit nevinného, než osvobodit darebáka. A proto se chybě odsouzení nevinného přisuzuje číslo 1 a věnuje se jí větší pozornost.

Ale někde musí být stanovena jistá hranice, po jejímž překročení už soud přistoupí k rozhodnutí „vinen“ a člověka potrestá.

Pokud se soudce snaží být mírný a odsoudí člověka až po nahromadění velkého množství důkazů (snižuje tím možnost výskytu chyby prvního druhu) a současně narůstá nebezpečí, že i když je obžalovaný vinen, potřebné množství důkazů se nenajde a soud jej osvobodí (roste možnost výskytu chyby druhého druhu).

Tj. snižováním možnosti výskytu chyby prvního druhu roste možnost výskytu chyby druhého druhu – a naopak: pokud zvyšujeme možnost výskytu chyby prvního druhu, snižuje se možnost výskytu chyby druhého druhu.

Je vidět, že žádnou z chyb není možné naprosto vyrušit: pokud totiž snižujeme možnost výskytu chyby prvního druhu až téměř na nulu, roste tím možnost výskytu chyby druhého druhu do obludných rozměrů a rozhodnutí učiněná tímto stylem jsou nerozumná, až nemoudrá.

Strategií v rozhodovacích procesech tohoto typu je tedy zvolit pravděpodobnost výskytu chyby prvního druhu malou, ale ne příliš malou.

Postup výpočtu

1. Sestaví se tabulka skutečných (naměřených) četností
2. Vypočítají se očekávané četnosti
3. Zkontrolují se podmínky pro použití testu
4. Vypočte se testové kritérium
5. Testové kritérium se srovná s kritickou hodnotou
6. Vysloví se rozhodnutí

1. Sestaví se tabulka skutečných (naměřených) četností

Tabulka: Skutečné četnosti

	znak1 - 1.skupina	**znak1 - 2.skupina**	**...**	**celkem**
znak2 - 1.skupina	skutečné četnosti	skutečné četnosti	...	$n_{1\bullet}$
znak2 - 2.skupina	skutečné četnosti	skutečné četnosti	...	$n_{2\bullet}$
znak2 - 3.skupina	skutečné četnosti	skutečné četnosti	...	$n_{3\bullet}$
...	...	...	...	$n_{i\bullet}$
celkem	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet j}$	n

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij} \quad n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$$

2. Vypočítají se očekávané četnosti

Tabulka: Očekávané četnosti

$$n'_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

	znak1 - 1.skupina	**znak1 - 2.skupina**	**...**	**celkem**
znak2 - 1.skupina	očekávané četnosti	očekávané četnosti	...	$n_{1\bullet}$
znak2 - 2.skupina	očekávané četnosti	očekávané četnosti	...	$n_{2\bullet}$
znak2 - 3.skupina	očekávané četnosti	očekávané četnosti	...	$n_{3\bullet}$
...	...	...	...	$n_{i\bullet}$
celkem	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet j}$	n

3. Zkontrolují se podmínky pro použití testu

Podmínky pro použití testu nezávislosti v kontingenční tabulce:

- nejvíce 20 % teoretických četností může být menších než 5
- žádná teoretická četnost nesmí být menší než 1

Pro tabulku 2x2:

- $n > 40$
- pokud $20 < n < 40$, pak je nutná úprava testového kritéria pomocí Yatesovy korekce
- pokud $n < 20$, pak použijeme Fisherův test

Možnost při nesplnění podmínek

Skutečné četnosti

	18-20 let	21-25 let	26-30 let	31-35 let
Znak 2 – 1. skupina	2	5	26	45
Znak 2 – 2. skupina	2	7	28	30

Očekávané četnosti

	18-20 let	21-25 let	26-30 let	31-35 let
Znak 2 – 1. skupina	2,15	6,46	29,05	40,34
Znak 2 – 2. skupina	1,85	5,54	24,95	34,66

Test nelze použít, mohou se sloučit kategorie. Např. po sloučení

Skutečné četnosti

	18-25 let	26-30 let	31-35 let
Znak 2 – 1. skupina	7	26	45
Znak 2 – 2. skupina	9	28	30

Očekávané četnosti

	18-25 let	26-30 let	31-35 let
Znak 2 – 1. skupina	8,61	29,05	40,34
Znak 2 – 2. skupina	7,39	24,95	34,66

4. Vypočte se testové kritérium (dosazení do vzorce – výsledek hodnota)

Testové kritérium

$$G = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$

r	počet řádků
s	počet sloupců
n_{ij}	skutečné četnosti
n'_{ij}	očekávané četnosti

5. Testové kritérium se srovná s kritickou hodnotou (tabulková hodnota, je potřeba zohlednit počet stupňů volnosti)

Kritické hodnoty testového kritéria chí-kvadrát

Stupně volnosti	Hladina významnosti	
	0,05	0,01
1	3,841	6,635
2	5,991	9,21
3	7,815	11,341
4	9,483	13,277
5	11,070	15,086
6	12,592	16,812
7	14,067	18,475
8	15,507	20,09
9	16,919	21,666
10	18,307	23,209
11	19,675	24,725
12	21,026	26,217
13	22,362	27,688
14	23,685	29,141
15	24,996	30,578
16	26,296	32
17	27,587	33,409
18	28,868	34,805
19	30,144	36,191
20	31,410	37,566

6. Vysloví se rozhodnutí

- Je-li testové kritérium < kritická hodnota, potom nezamítáme nulovou hypotézu o nezávislosti a nezávislost lze předpokládat.
- Je-li testové kritérium > kritická hodnota, potom zamítáme nulovou hypotézu o nezávislosti a přijímáme alternativní hypotézu, která nám říká, že určitá závislost existuje.
- Buď je pravdivá, nebo pro její zamítnutí nemáme dostatek potřebných měření, dostatek informací.

Yatesova korekce (Yatesův chí-kvadrát test)

- upravuje vzorec pro Pearsonův chí-kvadrát test odečtením 0,5 od rozdílu mezi každou sledovanou hodnotou a její očekávanou hodnotou v kontingenční tabulce 2 × 2
- $G = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(n_{ij} - n'_{ij} \mathbf{-0{,}5}\right)^2}{n'_{ij}}$
- snižuje to získanou chí-kvadrát hodnotu a tím zvyšuje její *p*-hodnotu
- zabraňuje nadhodnocení statistické významnosti u malých dat

https://dbterapie.cz/encyklopedie/yatesova-korekce-kontinuity/
Yates, F (1934). Kontingenční tabulka s malými čísly a χ2 test. Journal of the Royal Statistical Society (Supplement) 1: 217-235.

Příklad 1

- 2 vybrané otázky z dotazníku (2 znaky)

1. pohlaví

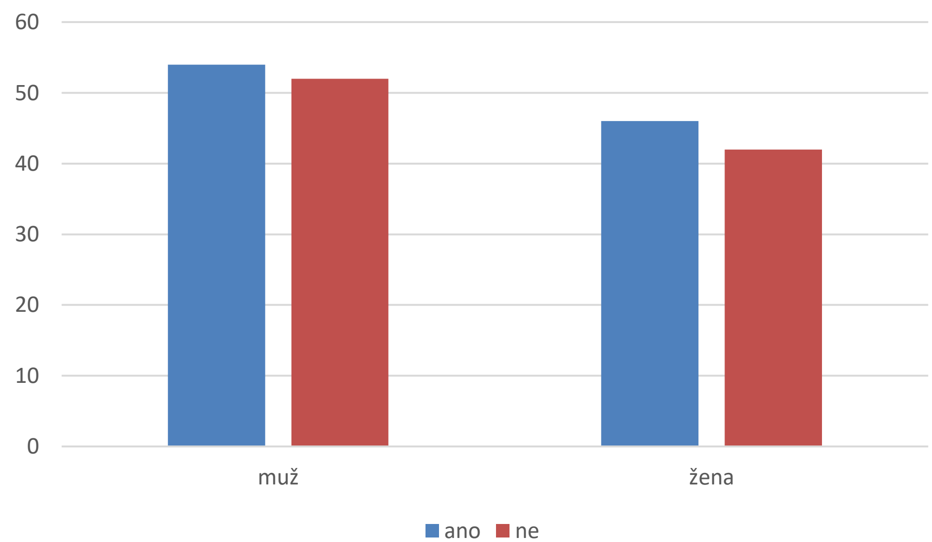
a) muž
b) žena

2. otázka

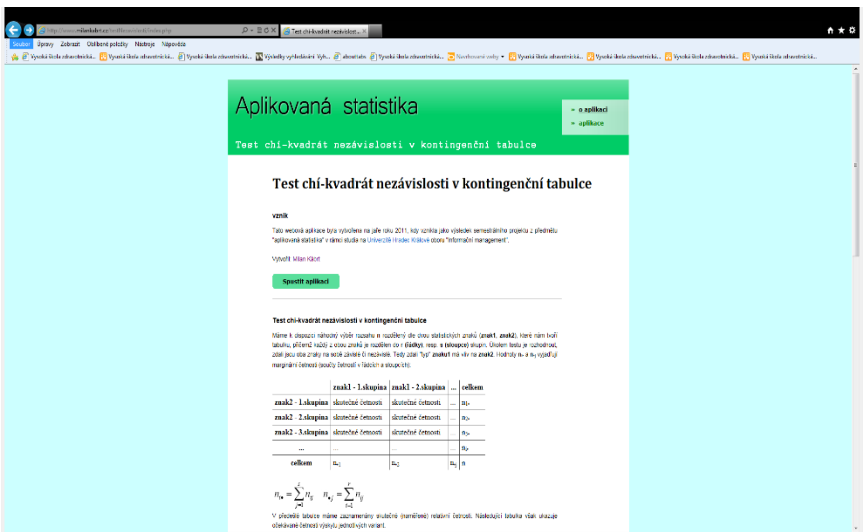
a) ano
b) ne

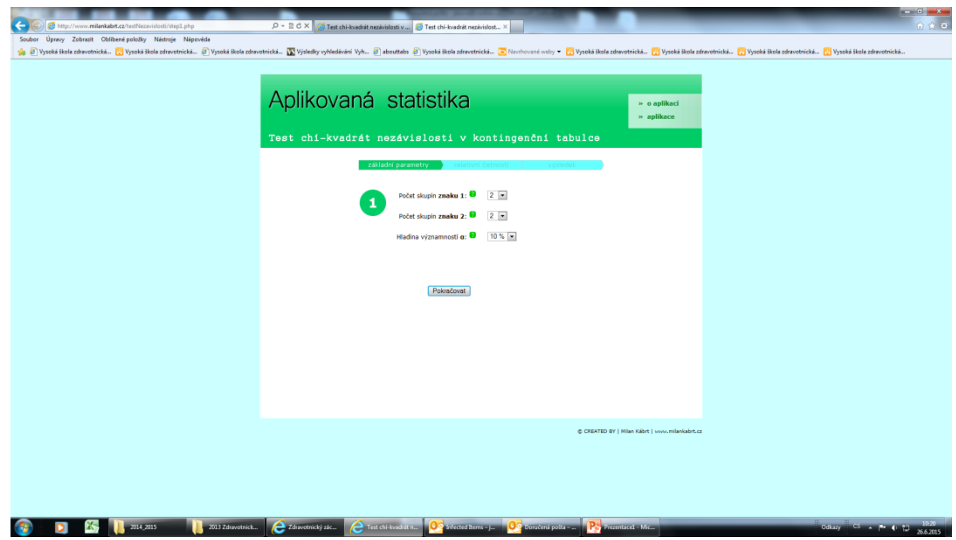
Sestaví se tabulka skutečných četností

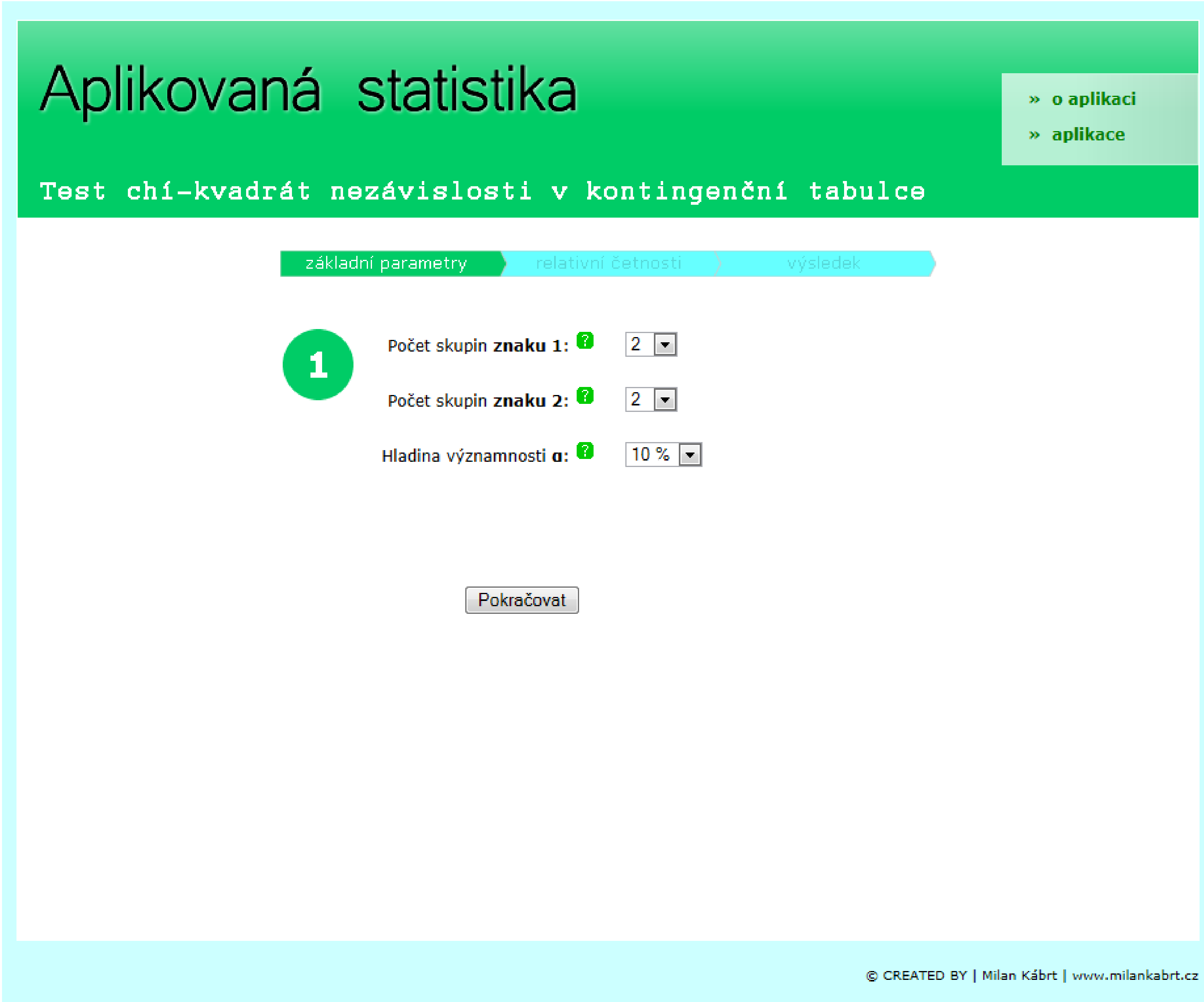
	ano	ne	celkem
muž	54	52	106
žena	46	42	88
celkem	100	94	194



- http://www.milankabrt.cz/testNezavislosti/index.php






Aplikovaná statistika
» o aplikaci
» aplikace
Test chí-kvadrát nezávislosti v kontingenční tabulce
základní parametry
relativní četnosti
výsledek
1
Počet skupin znaku 1:
2
Počet skupin znaku 2:
2
Hladina významnosti α:
10 %
Pokračovat
© CREATED BY | Milan Kábrt | www.milankabrt.cz

Aplikovaná statistika

» o aplikaci
» aplikace

Test chí-kvadrát nezávislosti v kontingenční tabulce

základní parametry | relativní četnosti | výsledek

1

Počet skupin **znaku 1**: 2

Počet skupin **znaku 2**: 2

Hladina významnosti **α**: 0.05

2

Zadejte do tabulky naměřené relativní četnosti: ?

	znak1 - 1. sk.	znak1 - 2. sk.
znak2 - 1. sk.	54	52
znak2 - 2. sk.	46	42

zpět **pokračovat**

Aplikovaná statistika

» o aplikaci
» aplikace

Test chí-kvadrát nezávislosti v kontingenční tabulce

základní parametry | relativní četnosti | výsledek

Výsledek testu

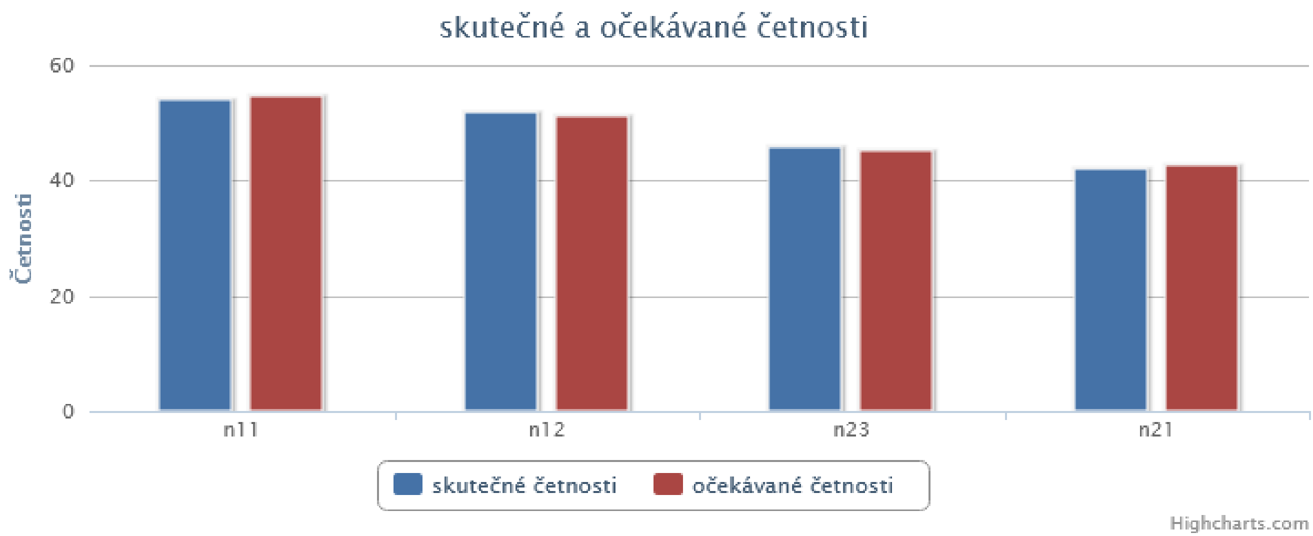
Počet skupin **znaku 1**:	2
Počet skupin **znaku 2**:	2
Hladina významnosti α:	5 %

Skutečné četnosti

	znak1 - 1. sk.	znak1 - 2. sk.	$n_{\cdot j}$
znak2 - 1. sk.	54	52	106
znak2 - 2. sk.	46	42	88
$n_{i\cdot}$	100	94	194

Očekávané četnosti

	znak1 - 1. sk.	znak1 - 2. sk.	$n_{\cdot j}$
znak2 - 1. sk.	54.64	51.36	106
znak2 - 2. sk.	45.36	42.64	88
$n_{i\cdot}$	100	94	194



testové kritérium:

$$G = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

Po dosazení do vzorce vychází testové kritérium:
G = 0.034

Kritická hodnota:

Kritická hodnota:
$\chi_{(1-\alpha);\, df}$ = **3.841**

Rozhodnutí:

Na hladině významnosti 5 % nulovou hypotézu (H_0) o nezávislosti jednotlivých znaků nezamítáme.

Spustit znovu

Bakalářská práce

- Provedli jsme test nezávislosti chí-kvadrát. Zkoumali jsme, zda existuje vztah mezi pohlavím a odpovědí na otázku … Hladinu významnosti jsme zvolili 5 %. Vytvořili jsme kontingenční tabulku, tabulku skutečných četností, dále jsme vypočítali očekávané četnosti. V tabulce očekávaných četností jsme zkontrolovali podmínky pro použití testu. Podmínky byly splněny a test jsme mohli použít. Hodnota vypočteného testového kritéria je … Protože kritická hodnota je větší než vypočtená hodnota, z provedeného testu vyplývá, že nezamítáme nulovou hypotézu o nezávislosti.

Příklad 2

- 2 vybrané otázky z dotazníku (2 znaky)

1. pohlaví

a) muž

b) žena

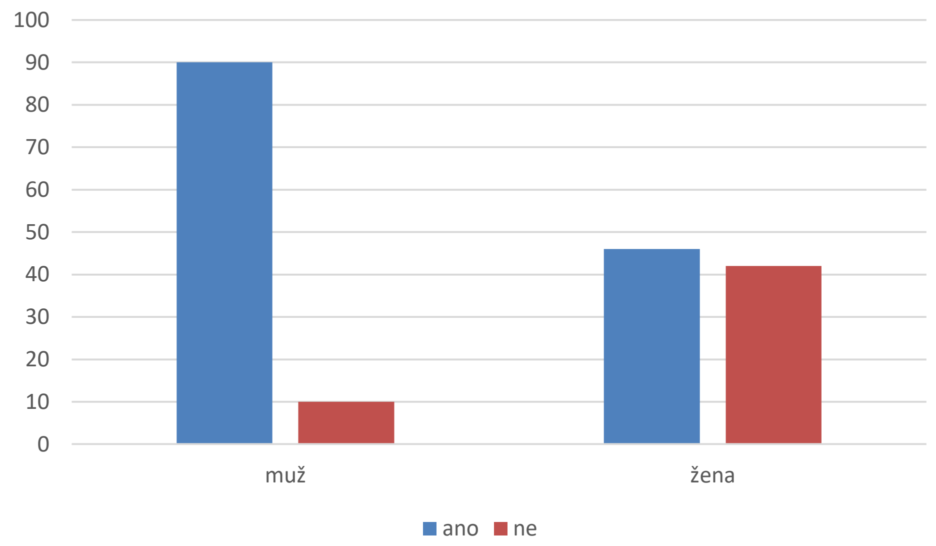
2. otázka

a) ano

b) ne

Sestaví se tabulka skutečných četností

	ano	ne	celkem
muž	90	10	100
žena	46	42	88
celkem	136	52	188



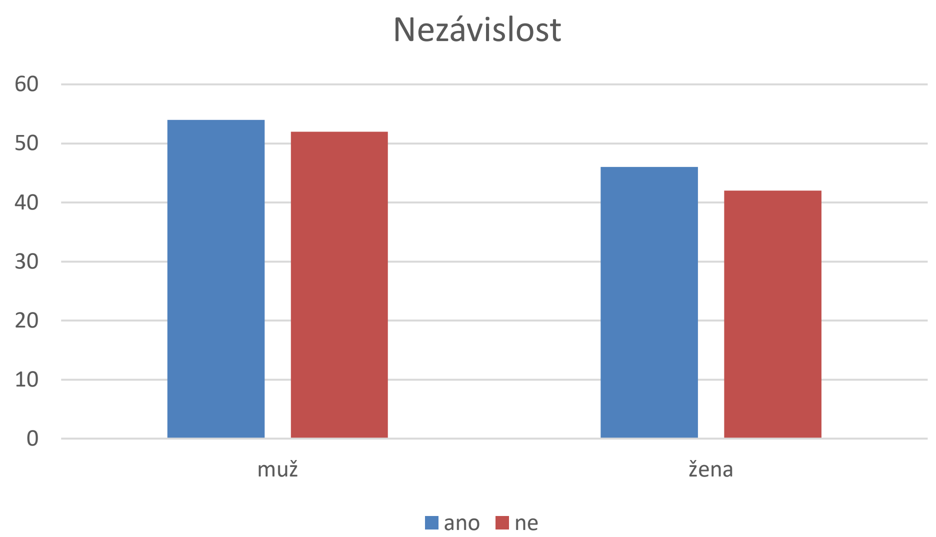
Vypočítají se očekávané četnosti

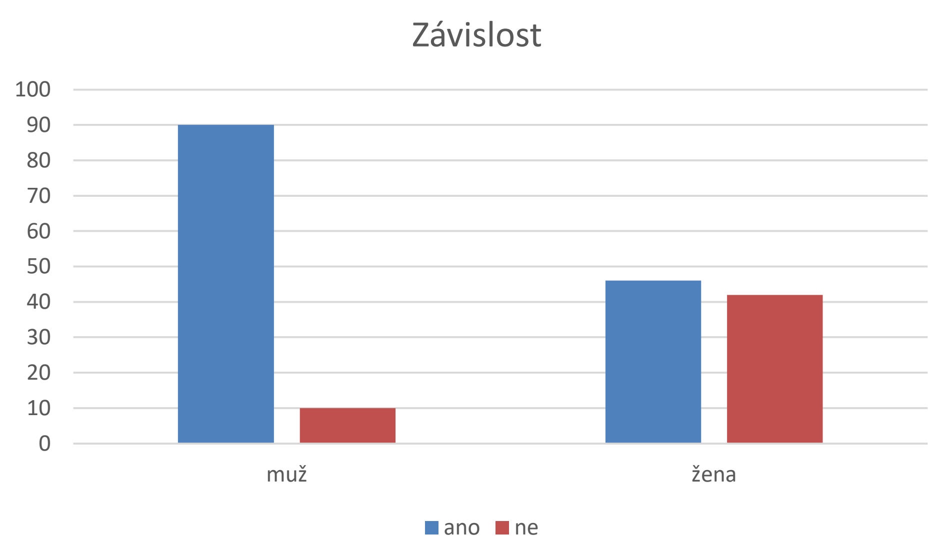
	ano	ne	celkem
muž	72,34	27,66	100
žena	63,66	24,34	88
celkem	136	52	188

- Testové kritérium 32,299
- Kritická hodnota pro 1 stupeň volnosti a hladinu významnosti 5 % je 3,841
- 32,299 > 3,841
- Rozhodnutí: Na hladině významnosti 5 % nulovou hypotézu (H_0) o nezávislosti jednotlivých znaků zamítáme a přijímáme hypotézu H_1, která nám říká, že zde určitá závislost existuje.

Bakalářská práce

- Provedli jsme test nezávislosti chí-kvadrát. Zkoumali jsme, zda existuje vztah mezi pohlavím a odpovědí na otázku … Hladinu významnosti jsme zvolili 5 %. Vytvořili jsme kontingenční tabulku, tabulku skutečných četností, dále jsme vypočítali očekávané četnosti. V tabulce očekávaných četností jsme zkontrolovali podmínky pro použití testu. Podmínky byly splněny a test jsme mohli použít. Hodnota vypočteného testového kritéria je … Protože kritická hodnota je menší než vypočtená hodnota, z provedeného testu vyplývá, že zamítáme nulovou hypotézu o nezávislosti a přijímáme alternativní hypotézu, která nám říká, že určitá závislost zde existuje.


Nezávislost
60
50
40
30
20
10
0
muž
žena
ano
ne


Závislost
100
90
80
70
60
50
40
30
20
10
0
muž
žena
ano
ne

	ano	ne	celkem
muži	50	50	=SUMA(B2:C2)
ženy	50	50	=SUMA(B3:C3)
celkem	=SUMA(B2:B3)	=SUMA(C2:C3)	=SUMA(B4:C4)
	ano	ne	celkem
muži	=$D2*B$4/$D$4	=$D2*C$4/$D$4	=SUMA(B7:C7)
ženy	=$D3*B$4/$D$4	=$D3*C$4/$D$4	=SUMA(B8:C8)
celkem	=SUMA(B7:B8)	=SUMA(C7:C8)	=SUMA(B9:C9)
testová statistika		=POWER(B2-B7;2)/B7	=POWER(C2-C7;2)/C7
		=POWER(B3-B8;2)/B8	=POWER(C3-C8;2)/C8
		=C11+D11+C12+D12	
kritická hodnota		3,841	
Cramerův koeficient		=ODMOCNINA(C14/D9*(2-1))	
Pearsonův koeficient		=ODMOCNINA((C14/(C14+D4))/((2-1)/2))	

Korigovaný koeficient kontingence pomocí Pearsona

$$C_{kor} = \frac{\sqrt{\frac{\chi^2}{\chi^2 + n}}}{\sqrt{\frac{m-1}{m}}}$$

kde χ^2 je hodnota testového kritéria

n je rozsah souboru

m je počet řádků nebo počet sloupců v kontingenční tabulce (je-li větší počet řádků, je *m* počet řádků; je-li větší počet sloupců, je m počet sloupců)

Cramerův koeficient

$$V = \sqrt{\frac{\chi^2}{n(m-1)}}$$

χ^2 je hodnota testového kritéria

n je rozsah souboru

m je počet řádků nebo počet sloupců v kontingenční tabulce (je-li větší počet řádků, je *m* počet řádků; je-li větší počet sloupců, je m počet sloupců)

p hodnota - Excel

CHISQ.TEST (funkce)

Vrátí test nezávislosti. Funkce CHISQ.TEST vrátí hodnotu rozdělení chí-kvadrát (χ2) pro dané testové kritérium a příslušné stupně volnosti. Pomocí testů χ2 můžete určit, zda experiment potvrzuje předpokládané výsledky.

Syntaxe

CHISQ.TEST(aktuální, očekávané)

$p < 0{,}05 \rightarrow$ nezávislost zamítáme, určitá závislost existuje

Skutečné četnosti

	ano	ne	celkem
muž	54	52	106
žena	46	42	88
celkem	100	94	194

$$n'_{11} = \frac{106*100}{194} = 54{,}6$$

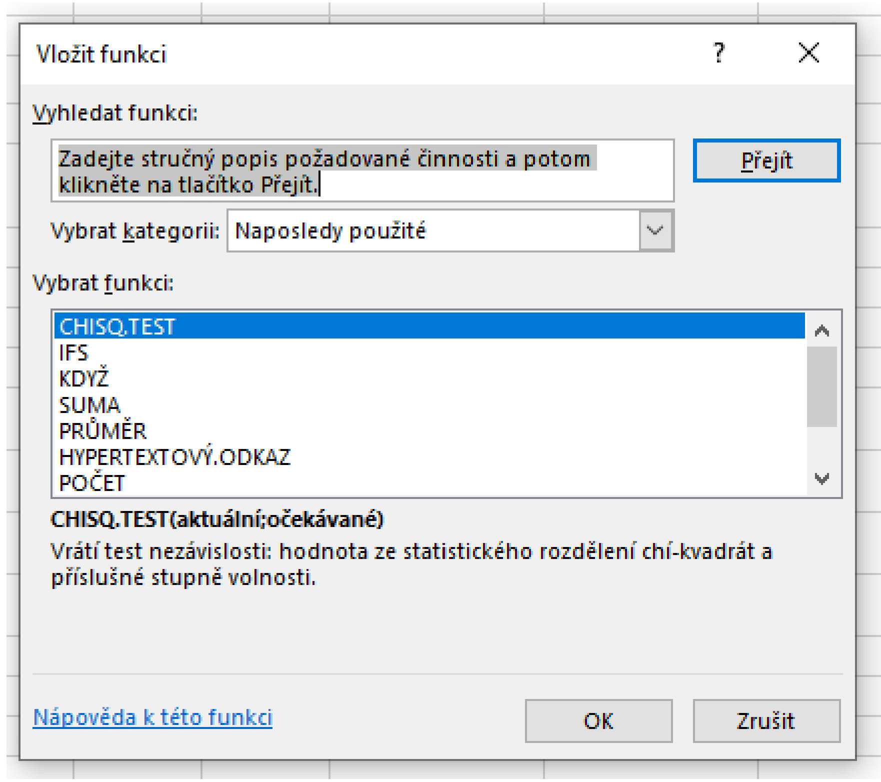
$$n'_{12} = \frac{106*94}{194} = 51{,}4$$

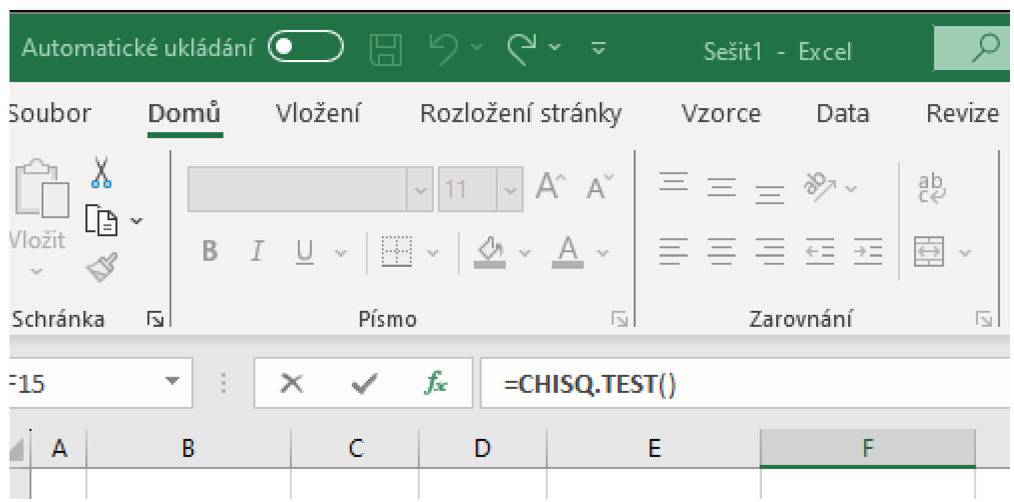
$$n'_{21} = \frac{88*100}{194} = 45{,}4$$

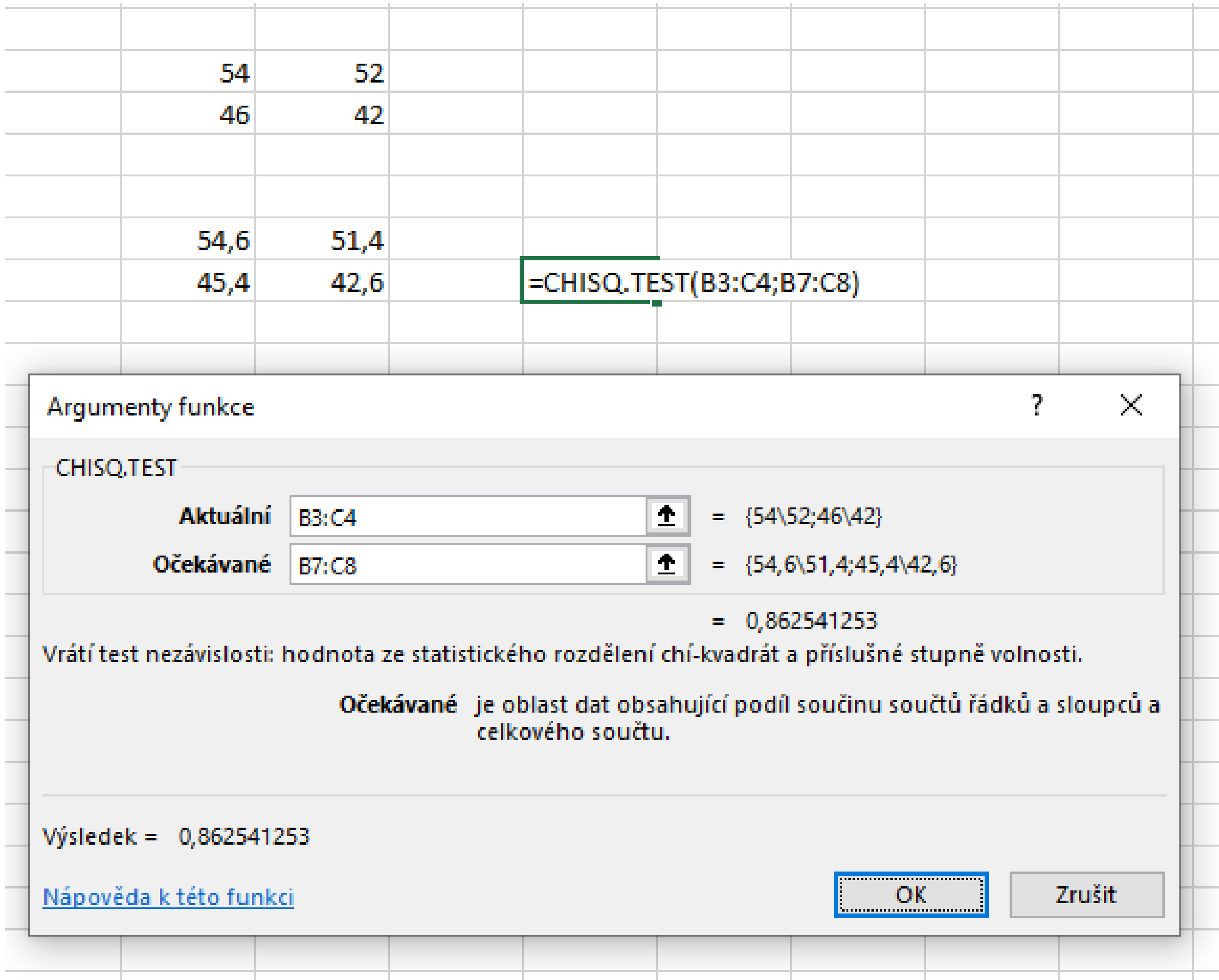
$$n'_{22} = \frac{88*94}{194} = 42{,}6$$

Očekávané četnosti

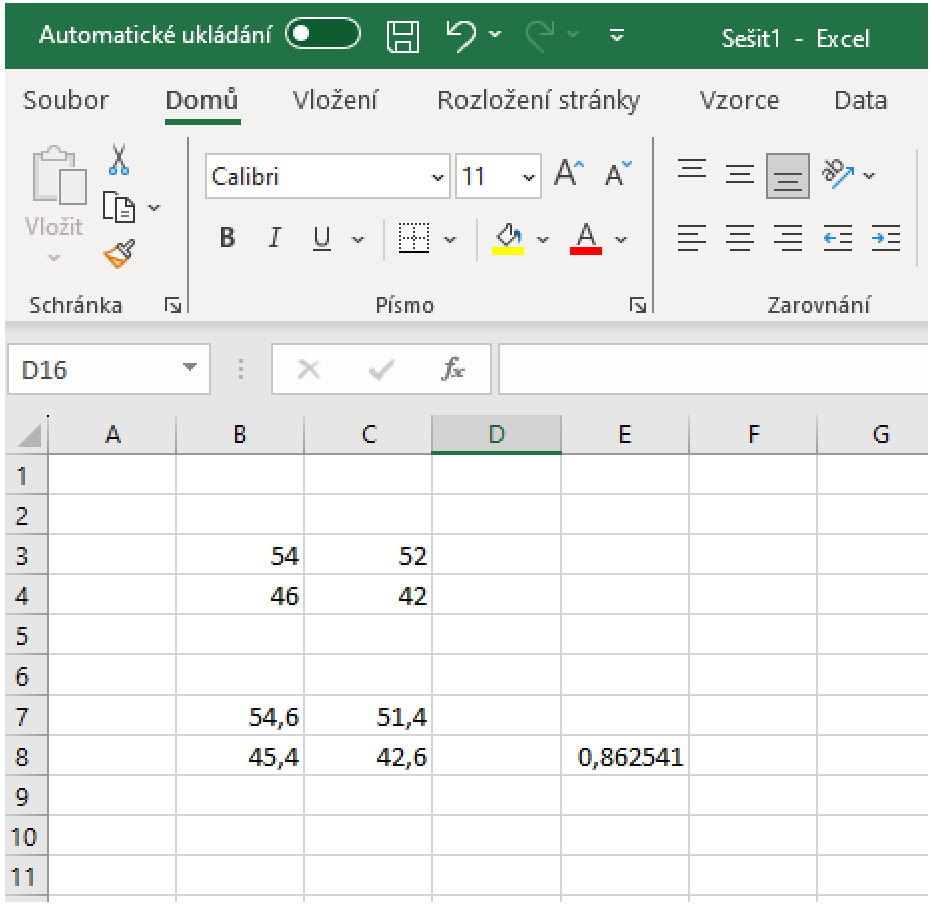
	ano	ne	celkem
muž	54,6	51,4	106
žena	45,4	42,6	88
celkem	100	94	194







54 52
46 42
54,6 51,4
45,4 42,6
=CHISQ.TEST(B3:C4;B7:C8)
Argumenty funkce
CHISQ.TEST
Aktuální B3:C4 = {54\52;46\42}
Očekávané B7:C8 = {54,6\51,4;45,4\42,6}
= 0,862541253
Vrátí test nezávislosti: hodnota ze statistického rozdělení chí-kvadrát a příslušné stupně volnosti.
Očekávané je oblast dat obsahující podíl součinu součtů řádků a sloupců a celkového součtu.
Výsledek = 0,862541253
Nápověda k této funkci
OK
Zrušit



p-hodnota = 0,862541

$p > 0{,}05 \rightarrow$ přijímáme nulovou hypotézu o nezávislosti

Skutečné četnosti

	ano	ne	celkem
muž	90	10	100
žena	46	42	88
celkem	136	52	188

$$n'_{11} = \frac{100*136}{188} = 72{,}3$$

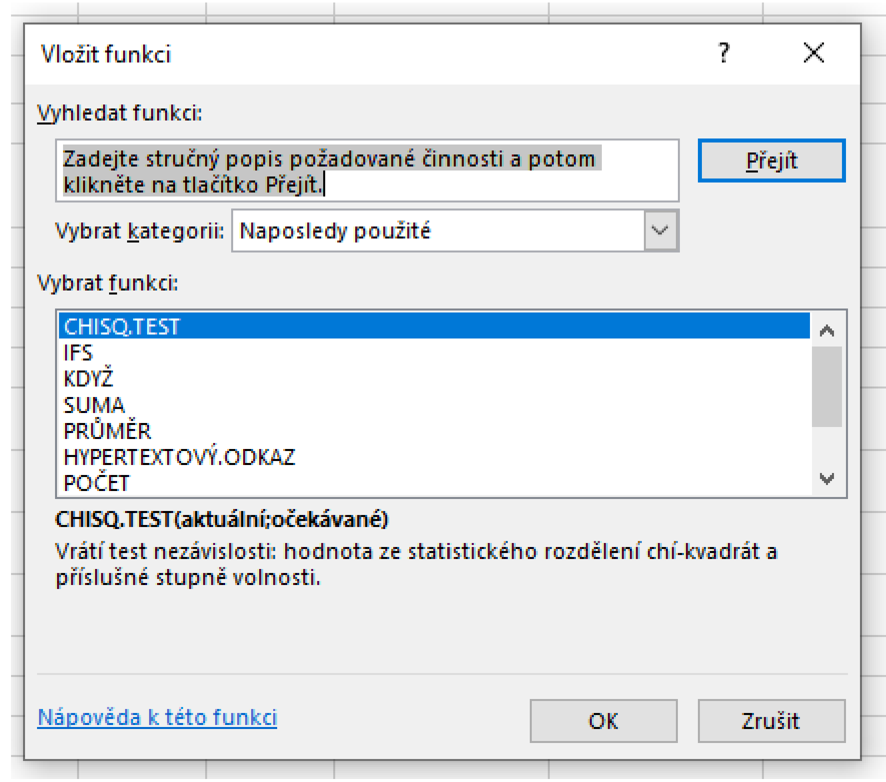
$$n'_{12} = \frac{100*52}{188} = 27{,}7$$

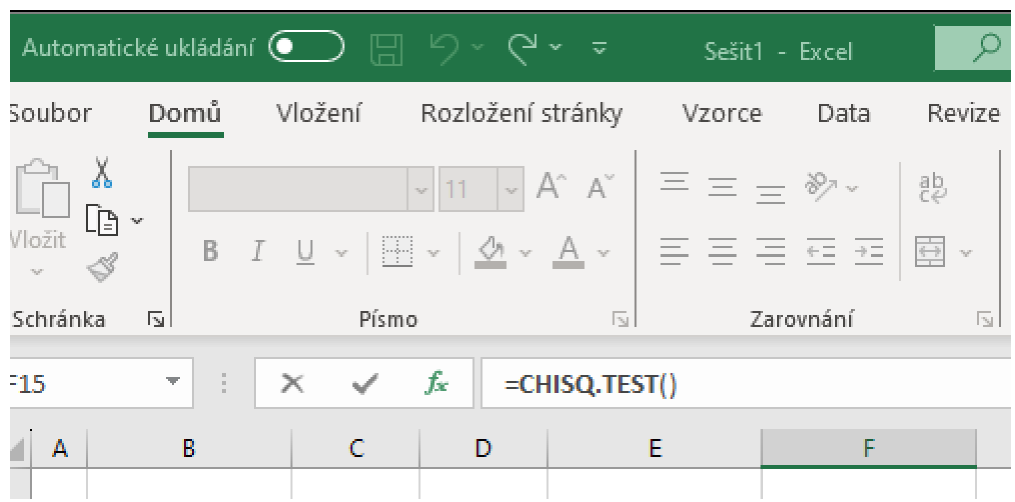
$$n'_{21} = \frac{88*136}{188} = 63{,}7$$

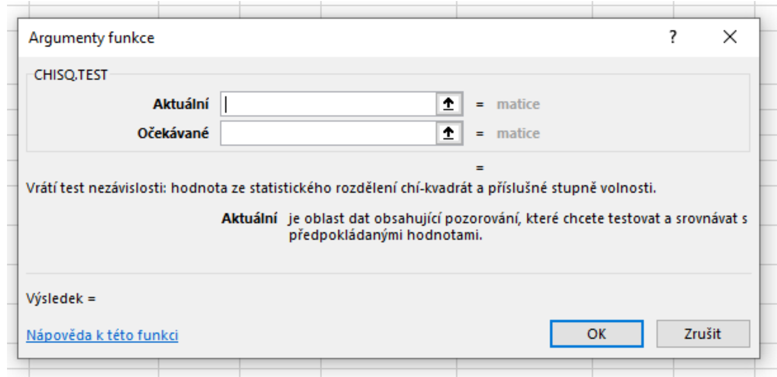
$$n'_{22} = \frac{88*52}{188} = 24{,}3$$

Očekávané četnosti

	ano	ne	celkem
muž	72,3	27,7	100
žena	63,7	24,3	88
celkem	136	52	188

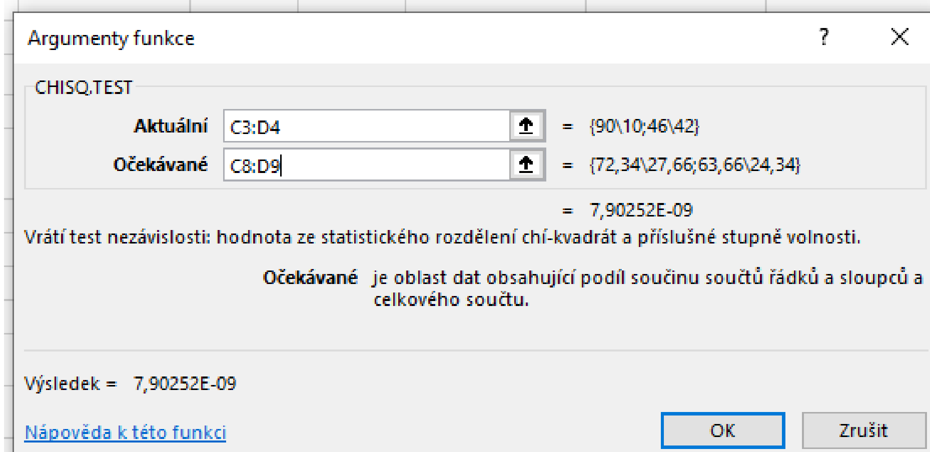






	ano	ne	celkem
muž	90	10	100
žena	46	42	88
celkem	136	52	188

	ano	ne	celkem
muž	72,3	27,7	100
žena	63,7	24,3	88
celkem	136	52	188



	ano	ne	celkem
muž	90	10	100
žena	46	42	88
celkem	136	52	188

	ano	ne	celkem
muž	72,3	27,7	100
žena	63,7	24,3	88
celkem	136	52	188

7,9E-09

p-hodnota = $7{,}9 \cdot 10^{-9}$

$p < 0{,}05 \rightarrow$ nezávislost zamítáme, určitá závislost existuje

p-hodnota

p-hodnota je nejmenší hladina významnosti, při které ještě zamítneme nulovou hypotézu

p-hodnota je pravděpodobnost, že při platnosti nulové hypotézy nabývá testová statistika své stávající hodnoty anebo hodnot ještě extrémnějších (nepříznivějších vůči nulové hypotéze)

p-hodnota je pravděpodobnost, s jakou bychom mohli obdržet pozorovaná data nebo data stejné, či ještě více odporující nulové hypotéze, za předpokladu, že je nulová hypotéza pravdivá. Čím menší je *p*, tím neudržitelnější čili méně důvěryhodná je nulová hypotéza