# Základní postup vyhodnocení závislosti dvou proměnných v TIBCO STATISTICE

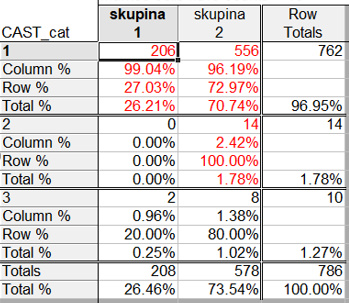
Tento dokument je určen pouze pro potřeby studentů výuky statistiky na KAD 1. LF UK. Jeho cílem je poskytnout návod k základnímu vyhodnocení závislosti dvou proměnných. Text neprošel jazykovou kontrolou.

První část textu uvádí konceptu kontingenční tabulky. Další část již vysvětluje na příkladu krok za krokem, jak vytvořit a interpretovat kontingenční tabulku.

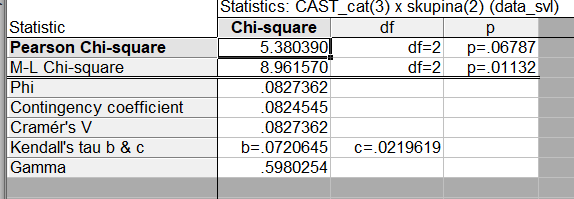
# Kontingenční tabulka

Základním nástrojem sloužícím k vyhodnocení závislosti dvou proměnných jsou kontingenční tabulky. Kontingenční tabulky se obvykle skládají z jedné tabulky, která zobrazuje hodnoty dvou proměnných ve vzájemném vztahu (obrázek 1) a z další tabulky, která zobrazuje základní koeficienty, které můžeme využít k analýze vztahu dvou proměnných (obrázek 2).

Obrázek



Obrázek



# Vyhodnocení kontingenční tabulky

Nyní se podívejme na vyhodnocení kontingenční tabulky. Prvním krokem je kontrola předpokladů pro použití Pearsonova chí kvadrát testu. Použití tohoto testu má předpoklady:

- 20 % buněk má očekávanou hodnotu menší než 5 (software nás upozorní)

- zároveň žádná buňka by neměla mít očekávanou hodnotu menší než 1 (2)

‚‘‘‘

Pokud nejsou naplněny předpoklady, tak případně můžeme tabulku také upravit, a to tak, že vybrané kategorie sloučíme pomocí překódování jednotlivých proměnných stejným postupem, jako je uveden níže při úpravě proměnných.

Pokud jsou naplněny předpoklady, tak přikročíme k vyhodnocení Pearsonova chí-kvadrát testu. Přitom sledujeme p-hodnotu. Záleží, jak je stanoven práh p-hodnoty, obvykle je stanovena hodnota alfa na 0,05. Pokud nám p-hodnota umožňuje zamítnutí nulové hypotézy o ne-existenci vztahu (obvykle P-hodnota menší než 0,05), tak můžeme usuzovat na existenci vztahu mezi proměnnými a dále pokračovat v analýze.

V tomto případě využijeme některý z koeficientů pro další vyhodnocení. Zde je třeba na základě charakteristik proměnných vybrat vhodný test pro vyhodnocení síly a případně směru vztahu (tabulka přehled základních testů a prezentace). V případě koeficientů pro nominální proměnné se hodnoty obvykle pohybují mezi 0 a 1, přičemž orientační hranice hodnot koeficientů je následující:

* Hodnota 0 – žádný vztah
* Hodnoty kolem 0,2 slabý vztah
* Hodnoty kolem 0,5 – středně silný vztah
* Hodnoty kolem 0,7 a výše – silný vztah

Pro ordinální a kardinální proměnné se hodnoty koeficientu pohybují mezi -1 a 1. Podobně jako u ostatních koeficientů, čím blíže 1 nebo -1, tím silnější vztah.

* pozitivní koeficient - s rostoucí mírou jedné proměnné roste i druhá (např. s rostoucí úrovní vzdělání roste zájem o politiku)
* negativní koeficient - s rostoucí mírou jedné proměnné druhá klesá (např. s rostoucím věkem klesá spokojenost se životem).



## Postup k vytvoření základní kontingenční tabulky

Obvykle ještě před ověřováním vztahu dvou proměnných je třeba se zamyslet nad tím, jaké vztahy mezi proměnnými nás zajímají a na základě toho si z dotazníku vybrat odpovídající proměnné. V případě výuky nemáme vlastní data, čili si musíme zvolit na základě dostupných proměnných, jaké vztahy nás zajímají.

Při samotném ověření vztahu dvou proměnných je třeba učinit následující kroky, které nemusí jít nutně za sebou, ale obvykle jdou po sobě. Mezi tyto základní kroky patří:

1) Příprava analýzy.

2) Hledání vztahu mezi proměnnými a sepsání interpretace.

V tomto případě bude postup demonstrován na příkladu vyhodnocení závislosti mezi výsledky v testu CAST a pobytem v sociálně vyloučených lokalitách.

### Příprava analýzy

Před samotným zahájením analýzy si musíme správně určit proměnné, které vstupují do analýzy, zkontrolovat jejich charakteristiky a formulovat výchozí hypotézy.

#### Stanovení hypotéz

Nejprve se podívejme na stanovení výzkumné hypotézy. V našem případě nás zajímá následující výzkumná otázka: Souvisí rizikové chování v souvislosti s užíváním konopných drog na základních školách s pobytem v sociálně vyloučených lokalitách (SVL)? Na základě této výzkumné otázky si můžeme stanovit následující hypotézu H1 (existuje vztah) a H0 (neexistuje vztah). V rámci vyhodnocení kontingenční tabulky pak ověřujeme H0. Statistickým postupem je totiž očekávat a ověřovat žádný vztah. V tomto případně si je stanovíme takto:H1: Rizikové chování v souvislosti s užíváním konopných drog souvisí s pobytem SVL.H0: (Nulová „statistická“ hypotéza): Rizikové chování v souvislosti s užíváním konopných drog nesouvisí s pobytem v SVL.

#### Kontrola a případná úprava proměnných

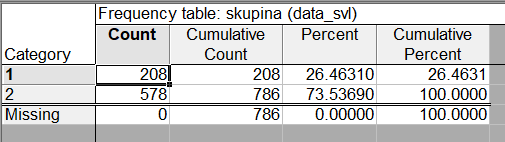
Prvně musíme určit proměnné, jejichž vztah budeme ověřovat. V tomto případě nás zajímá vztah mezi těmito proměnnými: a) nezávislá proměnná: příslušnost k populaci SVL a b) závislá proměnná: rizikovost užívání konopí – výsledky testu CAST.

Prvním krokem kontroly a případné úpravy dat je kontrola hodnot, chybějících hodnot a případně jejich úprava.

##### Nezávislá proměnná

V případě proměnné příslušnosti k populaci SVL je situace specifická, protože proměnná nebyla v rámci dotazníku a byla přiřazena na základě toho, v jaké lokalitě proběhl sběr dat. Není tedy uvedena v dotazníku, a proto ji tam na rozdíl od ostatních proměnných nebudeme hledat. Nicméně i tak si zobrazíme hodnoty této proměnné, abychom zjistili, zda se neobjevují neočekávané hodnoty. Víme, že proměnná se jmenuje skupina a měla by nabývat hodnot 1 (SVL) a 2 (neSVL).

Tabulku s hodnotami proměnných získáme, když půjdeme na kartu Statistics, vybereme v levém horním rohu Basic Statistics a dále vybereme možnost Frequency tables (viz prezentace 1). V novém dialogovém okně klikneme na Variables a vybereme proměnnou skupina a potvrdíme OK. Následně celý výběr potvrdíme kliknutím na tlačítko Summary. Tím získáme následující tabulku:



Z té je patrné, že proměnná nabývá hodnot 1 a 2, což odpovídá předpokladům. Nevidíme žádné chybějící ani divné hodnoty, a tak tuto proměnnou nemusíme upravovat. Hodnota reprezentuje SVL.

##### Závislá proměnná

V případě závislé proměnné, jelikož nás zajímají výsledky testu CAST, tak je naše zadání složitější. Test CAST musíme vyhodnotit.

Test CAST

Cílem testu CAST je zmapovat rizikové chování v souvislosti s užíváním konopných drog v posledních 12 měsících. Test CAST se skládá ze šesti otázek, na které respondent odpovídá ANO/NE. Každá odpověď ANO se hodnotí jedním bodem. Výsledkem součtu jednotlivých kladných odpovědí na otázky dotazníku má respondent možnost získat 0–6 bodů

* 0–1 bod představuje malé riziko rizikového chování,
* 2–3 body představují střední riziko rizikového chování,
* 4–6 bodů představují vysoké riziko rizikového chování užívání konopných drog.

Nicméně postupujeme stejně jako u kontroly jakýchkoliv jiných proměnných. První musíme zjistit, kde proměnnou najdeme v dotazníku, jaký je její stav a zda ji případně musíme upravit. Část dotazníku, kterou využijeme k vyhodnocení testu CAST je baterie otázek pod číslem 39 (viz obrázek níže). Jak vidíme, tak baterie se skládá ze šesti otázek. V případě, že je chceme identifikovat v rámci našeho datasetu v SASu, tak použijeme následující postup. Jednotlivé otázky se identifkují pomocí kombinace čísel a případně písmen, pokud se jedná o baterii více otázek. V případě, že chceme určit označení první otázky z baterie tak spojíme písmeno Q (jako question), číslo baterie (39) a písmeno otázky (a). Proměnnou v data setu najdeme pod Q39a.

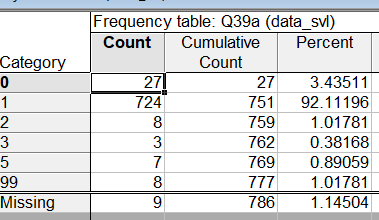
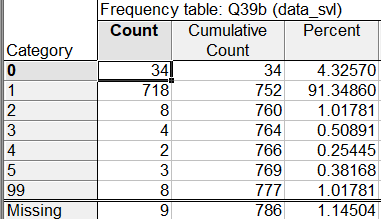
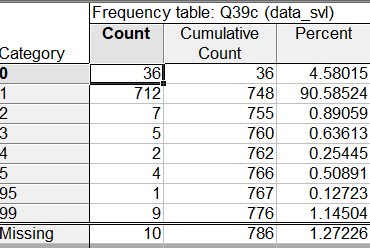
Obsah obrázku snímek obrazovky

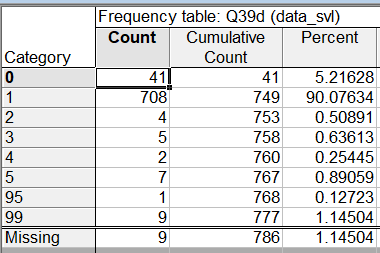
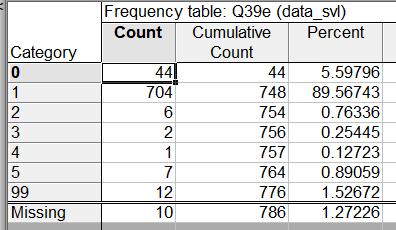
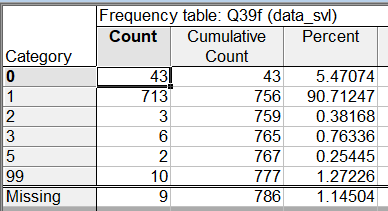
Popis byl vytvořen automaticky

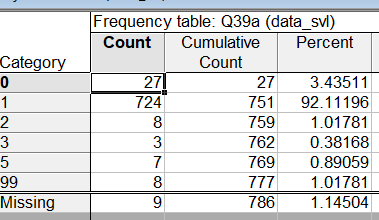
Stejným postupem pak získáme označení všech otázek:

Q39a Q39b Q39c Q39d Q39e Q39f

Nyní si musíme zkontrolovat proměnné v rámci data setu pomocí následujícího, přičemž využijeme uvedeného postupu pro zobrazení frekvenční tabulky. Frekvenční tabulky proměnných:





Ilustrujme si celý popis na základě proměnné Q39a.

Jako první se snažíme porozumět hodnotám, které nám tabulka s proměnnou Q39a zobrazuje. Při pohledu na otázku v rámci dotazníku vidíme (dva obrázky nahoře), že je pět možností odpovědí (nahoře nad otázkami), které reprezentuje 5 hodnot (dole pod otázkami): nikdy = 1, zřídka = 2, občas = 3, celkem často = 4 a velmi často = 5. Tyto hodnoty by se nám také měli zobrazit ve frekvenční tabulce níže. Avšak při pohledu na tabulku zjistíme, že místo očekávaných pěti hodnot se nám ukazují i další hodnoty: 0, missing a 99. V tomto případě se jedná u hodnoty 0 o špatné zakódováné odpovědi. 99 a missing jsou chybějící hodnoty. Nicméně, aby software rozeznal 99 jako chybějící hodnotu, tak je třeba to programu určit. Stejně tak 0 potřebujeme převést na chybějící hodnoty. Proto je tyto hodnoty třeba překódovat.

Před překódováním se ale ještě zamysleme, zda hodnoty této proměnné můžeme využít pro vyhodnocení testu CAST (viz popis testu výše). Přitom zjistíme, že potřebujeme získat u každé hodnoty místo pěti hodnot dvě hodnoty, které budou reprezentovat ANO a NE. Tuto úvahu také budeme muset zahrnout do překódování proměnné, které provedeme následujícím způsobem. Nejprve je vhodné si zapsat, jakým způsobem chceme, aby se změnily hodnoty. Může to být následujícím způsobem:

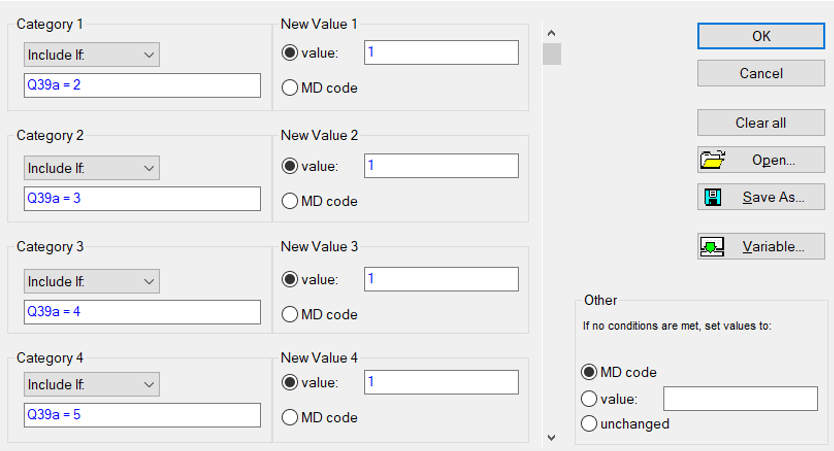
Staré hodnoty → nové hodnoty

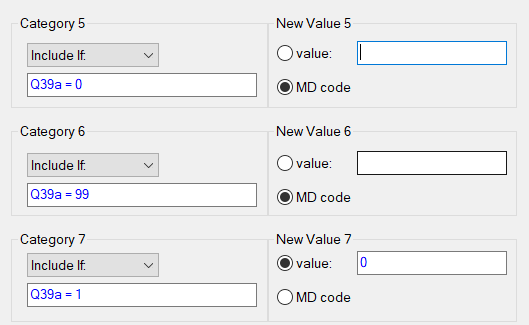
missing, 99, 0 → missing

1 → 0 (NE v testu CAST)

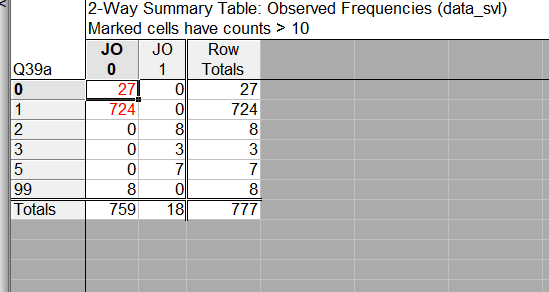
2,3,4,5 → 1 (ANO v testu CAST)

Jakmile provedeme tuto úvahu, tak ji aplikujeme pomocí rekódování. Proměnnou Q39a rekodujeme do nove promenne, kterou musíme vytvořit a kterou nazveme kupříkladu CAST\_1. Kliknete na novou proměnnou a vyberete kartu Data a posléze RECODE. Zde můžete vyplnit kupříkladu tímto způsobem:





Nyní je třeba zobrazit si nově vytvořenou proměnné, zda jsme ji správně zakódovali. To učiníme tak, že srovnáme staré a původní proměnné, a to tak že překřížíme původní a novou proměnnou a vytvoříme si kontingenční tabulku těchto dvou proměnných. Opět využijeme karty basic statistics, kde ovšem zvolíme možnost (tables and banners, v češtině kontingenční tabulky) a specifikujeme proměnné q39a a CAST, potvrdíme dvakrát ok. V dalším menu vybereme možnost advanced a detailed-two way tables. Graficky je to zobrazeno v prezentaci 2. Následující tabulka nám zobrazuje staré a nové hodnoty, takže můžeme vidět, že hodnota 1 proměnná CAST1 obsahuje hodnoty 2,3 a 5 původní proměnné Q39a. Vidíme tedy, že jsme proměnnou překódovali správně. Hodnota 4 se totiž neobjevila vůbec jako odpověď v dotazníku.



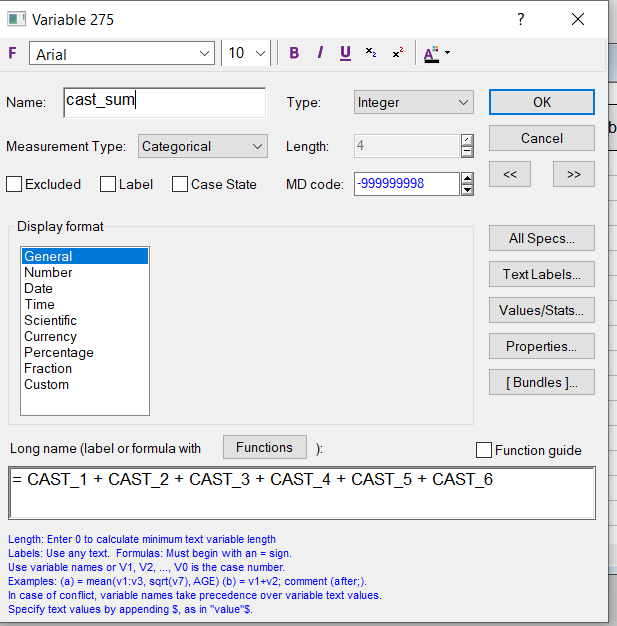
V případě testu CAST je třeba překódovat i ostatní proměnné analogickým způsobem. Po jejich překódování je třeba vyhodnotit test CAST. Nejprve musíme pro každého respondenta určit součet bodů, který získal v testu CAST (počet odpovědí ANO) a poté na základě toho určit do které kategorie patří. Kategorie jsou následující:

0–1 bod představuje malé riziko rizikového chování,

2–3 body představují střední riziko rizikového chování

4–6 bodů představují vysoké riziko rizikového chování užívání konopných drog.

Pro každého respondenta vytvoříme součet následujícím způsobem na základě proměnných, které jsme si vytvořili v předchozích krocích. Jedná se o tyto proměnné: CAST1, CAST2, CAST3, CAST4, CAST5, CAST6. Nejprve ovšem vytvoříme novou proměnnou CAST\_sum, kde v dolní části tabulky můžeme specifikovat součet.



Nyní je třeba překódovat proměnnou podobně, jako jsme to již dělali výše. Nejprve si zapíšeme, jak překódujeme hodnoty proměnných:

0–1 → 1 (představuje malé riziko rizikového chování)

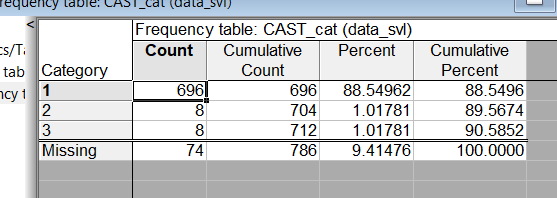
2–3 → 2 (představují střední riziko rizikového chování)

4–6 → 3 (představují vysoké riziko rizikového chování užívání konopných drog)

Posléze vytvoříme novou proměnnou CAST\_cat. Poté pomocí recode na karatě data překódujeme následujícím způsobem:



Výsledky rekódování:



#### Určení charakteristik proměnných

Dále si musíme určit základní charakteristiky proměnných. Ty jsou následující:

1. příslušnost k populaci SVL a n-SVL

* proměnná: skupina
* nominální proměnná

2. Rizikovost užívání konopí – výsledky CAST

- proměnná, kterou jsme upravili a vytvořili již dříve: CAST\_cat

- kvazi-kardinální, nicméně můžeme využít testy jako pro kardinální či ordinální proměnné

### Hledání vztahu mezi proměnnými a interpretace

#### Úvod

Dalším krokem základního vyhodnocení dvou proměnných je zobrazení kontingenční tabulky a vyhodnocení základních koeficientů včetně jejich předpokladů. Dále si uvedeme dva příklady vyhodnocení kontingenční tabulky. Prvním bude analýza vztahu proměnné CAST a pobytu v sociálně vyloučené lokalitě. Zde ovšem nejsou naplněny předpoklady, a tak krátce okomentuji ještě příklad, kdy jsou předpoklady naplněny. Druhá bude analýza vztahu proměnné lie/bet a pobytu v sociálně vyloučené lokalitě

#### Praktické příklady

CAST

Kontingenční tabulku získáme stejným způsobem jako v předchozím příkladu. Opět využijeme karty basic statistics, kde ovšem zvolíme možnost (tables and banners, v češtině kontingenční tabulky) a specifikujeme proměnné CAST\_cat a skupina, potvrdíme dvakrát ok. V menu options vybereme mezi statistikami (Statistics for two way-tay tables) Pearsonův chí-kvadrát test (Pearson and M-L Chi square) v Compute tables můžeme kupříkladu zaškrnout všechny možnosti. Poté zvolíme v Advanced možnost detailed two-way tables. Dále je třeba zvolit zobrazení očekávaných hodnot.

V případě tabulky pod textem vidíme, že nejsou naplněny předpoklady (2 buňky ze 6 mají očekávanou četnost menší než 5, tj. 33 %). V tomto případě by nám nepomohlo ani sloučení jednotlivých kategorií pomocí překódování.

Obsah obrázku text, snímek obrazovky, Písmo, číslo

Popis byl vytvořen automaticky

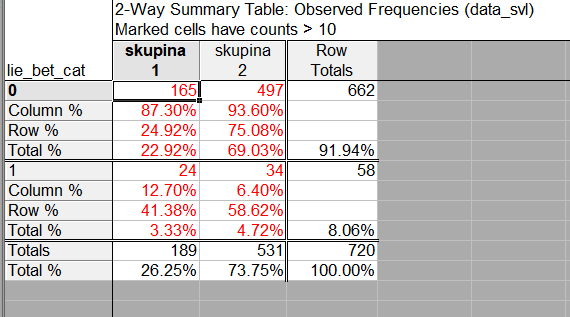
Interpretace je následující. Nejsou naplněny předpoklady pro využití Pearsonova chí-kvadrát testu, a proto nemůžeme postupovat dále s ověřením hypotézy.

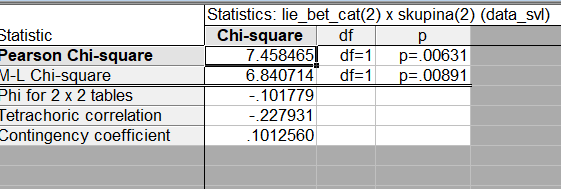
LIE/BET

V tomto případě využijeme proměnné lie\_bet\_cat, jejíž tvorba je popsána v prezentaci 2. Opět využijeme karty basic statistics, kde ovšem zvolíme možnost (tables and banners, v češtině kontingenční tabulky) a specifikujeme proměnné lie\_bet\_cat a skupina, potvrdíme dvakrát ok. V menu options vybereme mezi statistikami (Statistics for two way-tay tables) Pearsonův chí-kvadrát test (Pearson and M-L Chi square) v Compute tables můžeme kupříkladu zaškrnout všechny možnosti. Poté zvolíme v Advanced možnost detailed two-way tables. Dále je třeba zvolit zobrazení očekávaných hodnot.

Obsah obrázku text, snímek obrazovky, Písmo, číslo

Popis byl vytvořen automaticky





U této tabulky jsou předpoklady naplněny. Můžeme tedy přistoupit k dalším krokům. Hodnota chí-kvadrátu 0,0063, což je nižší než 0,05, což nám umožňuje vyvrátit hypotézu o ne-existenci vztahu mezi těmito dvěma proměnnými.

Můžeme tedy přistoupit k další analýze síly vztahu. V tomto případě, můžeme využít třech koeficientů dole v tabulce (Phi, Contigency a Cramers V). Všechny tyto koeficienty ukazují na to, že vztah je velmi slabý až žádný.

Interpretace tabulky je následující:

H0 o neexistenci vztahu mezi problémovým chováním a pobytem v SVL se podařilo vyvrátit. Můžeme tedy usuzovat na existenci vztahu mezi těmito dvěma jevy. Avšak síla vztahu je velmi nízká až zanedbatelná. Z tohoto důvodu můžeme říci, že mezi těmito populacemi není silný rozdíl.