Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informatiky a kvantitativních metod

Míry asociace mezi znaky a možnosti jejich využití

Diplomová práce

Autor: Bc. Marie Brixiová
Studijní obor: Informační management (2)

Vedoucí práce: prof. RNDr. Hana Skalská CSc.

Hradec Králové Srpen, 2021

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracovala samostatně a s použitím uvedené literatury.

V Hradci Králové dne 31. 7. 2021 ……

Marie Brixiová

Poděkování:

Děkuji vedoucí diplomové práce paní profesorce RNDr. Haně Skalské CSc. za metodické vedení, odborné rady a pomoc. Dále bych ráda poděkovala své rodině a příteli za podporu při studiu.

Anotace

Diplomová práce se zabývá statistickými metodami pro výpočet míry asociace, které jsou aplikovány na datech zvolených z veřejné databáze PISA 2018 (www.oecd.org), kde byly využity hodnoty datového setu uloženého ve formátu kompatibilním s IBM SPSS Statistics. Podle typů pozorovaných neznámých jsou rozlišeny jednotlivé statistické metody, tedy koeficienty měr asociací, testy nulové hypotézy o nezávislosti a další související koeficienty. Ty jsou aplikované za pomoci statistického programu IBM SPSS Statistics verze 26, kde většinu definovaných metod obsahuje procedura CROSSTABS. Použitá procedura napovídá tomu, že metodiky se opírají o základy kontingenční tabulky. Výsledky, které přímo souvisejí se sílou závislosti jsou vyhodnocovány za pomoci kombinace dvou metod z publikací Khamise (2008) a Mareše a spol. (2015). Výsledky jednotlivých koeficientů měřící míru asociace nabývají hodnot z intervalu <0;1> nebo také <-1;1>. Záporné hodnoty pak označují závislost nepřímou a kladné určují závislost přímou. Síla závislosti mezi proměnnými je určena dle výsledku koeficientu, především dle vzdálenosti od 0 (absolutní nezávislost) a 1 (absolutní závislost).

Annotation

Title: Measures of Association and Their Applications

The Master Thesis deals with statistical methods for measurement of association, which are applied on data chosen from the public database PISA 2018 (www.oecd.org), where the dataset and its values were already saved in a format compatible with IBM SPSS Statistics. The individual statistical methods are distinguished by the type of the observed variable. These methods are the association rate coefficients, tests of the null hypothesis of independence, and other related coefficients. These are applied using the statistical program IBM SPSS Statistics version 26, where most of the defined methods are contained within the CROSSTABS procedure. This procedure hints that the methodology is founded on the basics of the contingency table. The results, which directly relate to the strength of the association, are evaluated based on a combination of two methods from the publications of Kamis (2008) and Mareš et al. (2015). The results of individual coefficients measuring the strength of the association gain values from the interval <0;1> or <-1;1>. Negative values mean indirect dependence, while positive values mean a direct dependence. The strength of the association between the variables is defined by the result of the coefficient, mainly by the distance from 0 (absolute independence) and 1 (absolute dependence).

Obsah

1 Úvod

Ve výzkumných pracích různých oborů, se velice často klade otázka, zda existuje závislost mezi jednotlivými definovanými proměnnými. Tato otázka, nebo jinak také hypotéza o nezávislosti, se pokládá na začátku výzkumné práce a pomocí testů nezávislosti se hypotéza zamítá nebo přijímá. Pokud se zamítne hypotéza o nezávislosti, pak důležitou otázkou nastává, jak silná tato asociace je. Pro zjištění míry závislosti existují různé koeficienty, které často vycházejí z testů o nezávislosti. Jejich výsledky se pohybují v intervalu <0;1> nebo také <-1;1> podle čehož se dá určit síla dané závislosti. Z výsledku poté mohou výzkumníci zhodnotit, zda je důležité se na danou závislost zaměřit.

Pro zjištění, zda existuje závislost a určení její síly, je důležitá správná volba metody pro testování a měření. Při špatně zvolené metodě může nastat situace, kdy závěry ve výzkumné práci budou chybné a tedy nevěrohodné. Základem správného výběru testu a koeficientu je určit typ zkoumané proměnné. Proto se v prvních kapitolách práce zaměřuje právě na typy proměnných a práci s daty.

Jelikož může být výběr dat rozsáhlý a můžeme analyzovat proměnné které mohou mít i tisíce respondentů, je důležité znát nástroje které usnadní práci při manuálním počítání dle vzorců. Příkladem může být MS Excel, kde je potřeba znát vzorce pro výpočet. Dalším vhodným systémem je IBM SPSS Statistic, pomocí kterého jsou v této práci získávány výsledky testů a koeficientů definovaných v metodikách. MS Excel je využit v konkrétních výstupech k získání přesné kritické hodnoty, která se porovnává s výsledkem testu. Nulová hypotéza se dle psaných pravidel následně (ne)zamítá na dané hladině významnosti.

Cílem diplomové práce je tedy zpracování přehledu výpočtových možností měr asociací dle různých typů znaků, popsání jejich vlastností a možnosti využití. Součástí práce je také simulace definovaných metod nad vhodně zvolenými daty z veřejně dostupných databází, prostřednictvím již zmíněného softwaru IBM SPSS Statistic, který je rozebrán i z uživatelského pohledu.

2 Teoretická východiska

Asociace pochází z latinského slova „a socius“, v překladu, společník. Slovo asociace se v různých významech vyskytuje v běžném životě, v ekonomii a managementu, v psychologii, v přírodních vědách a v informatice. (Wikipedie, 2020)

Metody statistiky umožnují asociaci měřit. Při takovémto měření se zkoumá existence, druh a síla vazby mezi dvěma a více proměnnými. (Cvrček, 2019)

Míry asociace jsou popsané matematickými vzorci interpretující četnosti zjištěné v datech a odvozují se z nich konzistence dat. (Cvrček, 2019) Mnoho asociačních měr je založeno na statistických testech hypotéz, zatímco jiné jsou čistě heuristické a vycházejí z kombinace variant pozorovaných znaků. Tyto míry jsou měřitelné pouze v souborech objektů. (Rabušic, Soukup a Mareš, 2019)

Nesmí se však opomenout fakt, že nalezení vztahů mezi proměnnými, a změření míry prostřednictvím asociace a korelace (blíže popsáno v následujících kapitolách) je forma vyšší úrovně deskripce, která pro vysvětlení není postačující. Může sice posloužit k pochopení fenoménu, ale nepřinese odpověď na otázku „proč“. Daná otázka se ptá na příčinu a míra asociací přináší pouze stochastickou souvislost. (Rabušic, Soukup a Mareš, 2019)

2.1 Proměnné

Proměnné (jinak také například znak, měřená veličina, zkoumaná veličina, neznámá) představují logicky uspořádané charakteristiky, nebo vlastnosti zkoumaných jednotek. Díky nim lze dané jednotky uspořádat do kategorií podle jejich vlastností, dle nějaké míry, nebo určit číselně intenzitu. (Rabušic, Soukup a Mareš, 2019)

2.1.1 Typy proměnných

Pro zvolení vhodné míry asociace je nutné určit úroveň měření každé studované neznámé. (Khamis, 2008) Jsou-li výzkumné dotazy kódovány, používá se pojem kategorie a pracuje se tedy s kategoriálními proměnnými. Vhodné příklady kategoriálních proměnných uvádí Řezanková (2017):

- *„Národnost (česká, slovenská, …)*
- *Úroveň vzdělání (základní, střední, vysokoškolské)*
- *Počet dětí (0, 1, 2, 3, …)“*

„Uvedené příklady zároveň ilustrují úroveň vztahů mezi kategoriemi – v prvním případě kategorie nelze uspořádat, ve druhém je lze uspořádat, ve třetím navíc lze hovořit o škálách měření, jejichž základní dělení je následující:“ (Řezanková 2017)

- *Škála nominální*

  U hodnot této škály nelze stanovovat jejich pořadí, ale pouze jejich rozdílnost.
- *Škála ordinální*

  Hodnoty této škály můžeme seřadit, ale nemůžeme určit o kolik je jedna hodnota větší/menší než druhá.
- *Škála kardinální, která se ještě dělí na intervalovou a podílovou/poměrovou* (Rabušic, Soukup a Mareš, 2019)
  - *Škála intervalová*

    Hodnoty této škály jsou číselné a můžeme u nich stanovit o kolik je hodnota jedné proměnné větší/menší než druhá. (Řezanková 2017)
  - *Škála poměrová*

    Tato škála obsahuje pouze kladné hodnoty, u kterých můžeme určit o kolik i kolikrát je jedna hodnota větší/menší než druhá. (Řezanková 2017)

Uvedené typy škál jsou základem pro následující dělení proměnných:

- *Nominální proměnná*

  Hodnoty takového znaku jsou kategorie, které se dle daných pravidel označují číselnými kódy. Tyto kategorie nelze nijak uspořádávat, neboť dané číslo je pouze symbolem (označením) dané kategorie, a nikoliv množstvím měřené vlastnosti. (Rabušic, Soukup a Mareš, 2019)
- *Ordinální/Pořadové proměnné*

  Na rozdíl od kategorií nominálních proměnných, se mohou kategorie ordinálních proměnných uspořádat, a lze určit která z nich je v pořadí výše než jiná. Ordinální stupnice ale zobrazují pouze pořadí, nelze jimi proto určit stupeň odlišnosti. Stejně jako u nominálních neznámých nelze určit množství daného znaku. (Rabušic, Soukup a Mareš, 2019)
- *Kardinální/Kvantitativní proměnné*

  Na rozdíl od předchozích dvou typů zkoumaných veličin, číselné kódy kardinálních proměnných již nejsou arbitrární, tedy vyjadřují skutečné

množství sledované vlastnosti. Nejen, že takové proměnné se dají seřadit dle velikosti, ale také je možné určit o kolik se liší od ostatních. (Rabušic, Soukup a Mareš, 2019)

Kardinální znaky se obvykle rozlišují na:

- Intervalové – u těchto neznámých má smyl rozdíl hodnot
- Poměrové – mohou nabývat hodnot kladných, ale i nulových
- Diskrétní – nabývají izolovaných, často celočíselných hodnot
- Spojité – mohou nabývat libovolných hodnot z daného intervalu reálných čísel. (Řezanková, 2017)

Dle článku Measures of Association: How to Choose? (Khamis, 2008) se rozlišují především dvě úrovně zkoumaných veličin – kontinuální (spojité) a diskrétní.

Kontinuální proměnné mají hodnoty, které pocházejí ze spojité reálné číselné řady. Pro takové znaky teoreticky neexistují žádné omezení v možných hodnotách. Příkladem takovýchto proměnných je například gestační věk, krevní tlak nebo index tělesné hmotnosti. (Khamis, 2008)

Diskrétní proměnné nabývají hodnot, které jsou diskrétní neboli „oddělené“. Tyto hodnoty nepocházejí z kontinua skutečné číselné řady. Existují zde omezení mezi hodnotami měřených veličin. Tento typ proměnných obvykle má jen několik možných hodnot. Příkladem diskrétních proměnných jsou pohlaví, závažnost onemocnění, typ operace atp. Kvůli způsobu, jakým jsou tyto neznámé definovány, není možné pozorovat hodnotu mezi dvěma pozorovanými znaky. Například, každý pacient je buď muž nebo žena bez jakéhokoli jiného možného označení. (Khamis, 2008)

V případě diskrétních proměnných existují dvě podkategorie stupnice měření, které se označují jako ordinální a nominální. Pokud úrovně diskrétní proměnné nemají přiřazeny žádné pořadí, pak se neznámá nazývá diskrétní nominální proměnná. Například typ operace je nominální proměnná, protože různé typy operací nemají žádné přiřazené pořadí. Ale pokud je neznámá spojena například s provozními náklady, pak by se jednalo o ordinální proměnnou. (Khamis, 2008)

Další typem měřené veličiny, kterou označujeme jako **dichotomickou**, jinak také binární, je proměnná nabývající pouze dvou hodnot. Jako příklad lze uvést zaměstnaný x nezaměstnaný, živý x zemřelý apod. Tento typ znaku má poměrně zvláštní chování,

jelikož s ním lze operovat také jako se znakem kardinálním, i přestože z metodologického pohledu jde o znak nominální. Za kardinální jej lze považovat pouze za předpokladu, že využijeme zavedených kódovacích schémat (často 0 vs 1 nebo 0 vs 100). (Rabušic, Soukup a Mareš, 2019)

Hledáme-li souvislosti mezi jevy, obvykle se začíná úvahou, také nazývanou hypotézou, o vztahu dvou jevů (proměnných), přičemž jeden z jevů musí být příčinou a druhý následkem. Jev, který reprezentuje v naší hypotéze příčinu, je proměnná nezávislá. Závislá proměnná pak značí jev, který reprezentuje následek. Pokud toto určení není jasně dáno, je vhodné vzít v úvahu časový průběh. (Rabušic, Soukup a Mareš, 2019)

2.1.2 Počet proměnných

Pro statistiku, která popisuje, nebo vytváří závěry o jediném rozdělní stačí jednorozměrná statistika. Ta se využívá například pro výpočet souhrnné statistiky, získání odhadů intervalu pro parametry, nebo testování hypotéz týkajících se těchto parametrů. Jednorozměrná statistka sice tvoří základ jiných typů statistik, ale žádná se netýká vztahů mezi zkoumanými veličinami. Pro analýzu vztahů mezi proměnnými, je nutné přejít na úroveň dvourozměrné statistiky, která napomáhá analyzovat vztahy mezi dvěma znaky. Výzkumníci nebo analytikové ve firmách ale často chtějí zkoumat vztahy mezi více proměnnými. Pro tyto typy problémů využívají vícerozměrnou statistiku. (Gingrich, 1992)

Dvourozměrná a vícerozměrná statistika je významná nejen pro statistiku, ale také pro další vědy a obory. Pomocí těchto analýz mohou výzkumníci daného odvětví vysvětlit chování jednotlivých proměnných a vyřadit z výzkumu i takové neznámé, které mají malý vliv na chování jiných zkoumaných veličin. Takovýto výzkum se dá dobře představit v sociologii, kde jsou časté výzkumy zabývající se vysvětlením nějakého sociálního jevu, který má vždy nějakou příčinu. Vědec zde zkoumá povahu vztahů mezi proměnnými. Ty, které se zdají mít malý vztah ke zkoumané proměnné může ignorovat, a zaměřit pozornost na ty které daný jev ovlivňují více. Snadno tak vyvodí závěry ve svém výzkumu. (Gingrich, 1992)

2.2 Hromadná data

Před jakoukoliv analýzou problému je nutné nejdříve získat hromadná data. Tedy data, která jsou získána na základě výzkumu odvíjejícího se od zvolené výzkumné otázky. Ta vzniká na základě potřeby vysvětlení nějakého problému. „*Primární podmínkou je, aby byla získávána jako data standardizovaná*.“ (Rabušic, Soukup a Mareš, 2015) Příkladem sběru dat jsou dotazníky. V nich jsou splněny podmínky standardizace, tedy že všem respondentům byly kladeny stejné otázky ve stejném znění a pořadí. Dále musí být zajištěny také podmínky plynoucí ze statistických požadavků na proměnné. (Rabušic, Soukup a Mareš, 2015)

Dle Rabušice, Soukupa a Mareše (2015) ze statistických požadavků plynou následující podmínky:

1) **Rozlišitelnost** – *„proměnná musí variovat, musí tedy nabývat alespoň dvou hodnot“*
2) **Zařaditelnost** – *„ke každému stavu vlastnosti existuje příslušná hodnota znaku“*
3) **Jednoznačnost** – *„dvě různé hodnoty znaku nemohou odpovídat jednomu stavu vlastnosti“*
4) **Reprezentativnost** – *„naše data musí být reprezentativní, tak aby nám umožnila zobecnit výsledky našich výpočtů z výběrového souboru na soubor základní (inferenční statistika) a do výzkumu musí být zahrnut dostatečný počet výzkumných jednotek“*

Reprezentativnost však není vždy nutnou podmínkou. Je-li předmětem výzkumu malá populace, je potřeba vyčerpávajícím šetřením jednoduše zahrnout všechny členy výzkumu. *„V takovém to případě ztrácí smysl inferenční statistika, kterou lze aplikovat pouze v souborech, jehož jednotky byly vybrány náhodně.“* (Rabušic, Soukup a Mareš, 2015)

„Zdrojem hromadných dat pro statistickou analýzu může být především vlastní sběr dat, dále data posbíraná jinými výzkumníky nebo statistické výkaznictví a speciální šetření, jako je například sčítání lidu nebo mikrocensus.“ (Rabušic, Soukup a Mareš, 2015)

2.2.1 Práce s hromadnými daty

Ve většině statistických programových systémů je vstupem pro vícerozměrnou statistickou analýzu nejčastěji datová matice. Řádky této matice odpovídají statistickým jednotkám – například osoby, domácnosti, firmy atp. Tyto řádky se často označují jako případy (cases), také se ale vyskytuje termín pozorování (observation), nebo záznam. Sloupce dané matice odpovídají statistickým znakům. Existují také programovací systémy umožňující jako vstup využít četnosti kombinací výskytů kategorií u analyzovaných proměnných získaných v minulosti. V tomto případě každý řádek zaznamenává jednu kombinaci výskytů kategorií a jejich četnost, která je označována jako váha uvedené kombinace. Jiné systémy umí použít také oba způsoby, například IBM SPSS Statistics, o kterém je více zmíněno v kapitole 3.1. (Řezanková, 2017)

2.3 Kontingenční tabulka

Jako analýzu druhého stupně se v sociálních vědách označuje takové zobrazení rozdělení, které má 2 proměnné, jež jsou zobrazené za pomocí tabulky nebo grafu. Zjišťování četností proměnných kategoriálního typu probíhá pro takové dvojice kategorií, *„kdy jedna kategorie z dvojice přísluší první proměnné a druhá kategorie druhé proměnné“*. (Řezanková, 2017) Tímto se získá dvourozměrná tabulka četností neboli kontingenční tabulka. Na základě těchto četností pak analytik může usoudit, zda mezi dvěma kategoriálními proměnnými je nebo není závislost. (Řezanková, 2017)

Kontingenční tabulky nemusí být pouze dvourozměrné. *„Výsledkem třídění souboru dle tří kategoriálních proměnných je trojrozměrná kontingenční tabulka.“* (Hebák a kolektiv, 2013). Kategoriální proměnné trojrozměrné tabulky se značí písmeny X (R kategorií), Y (S kategorií) a Z (V kategorií). Tabulka je tedy tvořena z R řádků, S sloupců a V vrstev. *„Pro dvou rozměrné zobrazení v kontingenční tabulce je možné zvolit tzv. hierarchickou formu, kdy například tabulka o R řádcích má SV sloupců kombinujících kategorie veličin X a Z, má S sloupců. Dvourozměrná rozdělení veličin podmínění kategoriemi třetí veličiny (Z) pak představují soustavu dvourozměrných kontingenčních tabulek, jejichž sloučením vznikne marginální dvourozměrné rozdělení.“* (Hebák a kolektiv, 2013)

Ve skupině tří veličin může existovat následující situace vzájemných vztahů (Hebák a kolektiv, 2013):

a) *„všechny tři veličiny jsou vzájemně nezávislé, tedy každá dvojice z nich je nezávislá*
b) *jedna veličina neovlivňuje zbývající dvě ani jejich asociaci*
c) *jedna veličina ovlivňuje zbývající dvě, a tím způsobuje jejich zdánlivou asociaci, která se nazývá Simpsonův paradox*
d) *každá veličina ovlivňuje zbývající dvě, ale ne jejich asociaci, taková asociace se označuje jako homogenní*
e) *v případě nehomogenní asociace se jedná o takovou veličinu která nejen ovlivňuje zbývající dvě veličiny, ale také jejich asociaci.“*

Již výše zmíněným Simpsonovým paradoxem se tedy označuje situace, kdy změnou jedné veličiny dochází také ke změněně vztahu mezi zbývající dvojicí. To může vést mylným závěrům při vyhodnocení pouze dvourozměrné kontingenční tabulky. Veličiny, které jsou podmíněny nezávisle se pak mohou v marginálním rozdělení jevit jako závislé. (Hebák a kolektiv, 2013)

Zvyšuje-li se počet proměnných a jejich kategorií, pak se vztahy mohou více komplikovat. S rostoucím rozměrem kontingenční tabulky se zvyšují také nároky na rozsah souboru, pokud není potřeba mít výsledkem jeho třídění řídkou tabulku s neobsazenými nebo jen málo obsazenými políčky. (Hebák a kolektiv, 2013)

Četnosti na políčkách kontingenční tabulky mohou být buď absolutní nebo relativní. Udávají kolik hodnot daného vzorku se vyskytuje ve statickém souboru. (Matematika.cz, 2014) Hodnoty sloupcových a řádkových součtů se pak označují jako marginální četnosti.

Hodnoty z kontingenční tabulky lze zobrazit jako sloupcový graf, pro dvojice kategorií pak může mít 2 podoby, a to graf shlukový nebo graf kumulativní. (Řezanková, 2017)

Postup pro analýzu kontingenční tabulky shrnuje Rabušic, Soukup a Mareš (2015) do následujících kroků:

1. *„Vytvoříme četnostní tabulky pro obě proměnné a zvážíme případné sloučení či vynechání kategorií v těch případech, v nich jsou malé četnosti.*
2. *Zobrazíme si kontingenční tabulku pro první vhled, zpravidla ve formě sloupcových, či řádkových procent.*

3. *Vypočítáme chí-kvadrát test (více v kapitolách 2.8 a 3.3) a rozhodneme o případné závislosti.*
4. *V případě, že chí-kvadrát test bude statisticky významný, vypočteme adjustovaná rezidua a znaménkové schéma k posouzení struktury závislosti.*
5. *Interpretujeme nalezenou závislost (díky výstupům z kroků 2 a 4).*
6. *Popíšeme sílu souvislosti koeficientem asociace a tím posoudíme věcnou významnost nalezené souvislosti.“*

2.4 Volba správné míry asociace

Dle práce, Lexical association measures, P. Peciny (2009), poslední krok extrakce dat zahrnuje použití vybraného měřítka asociace. Žádoucím výsledkem celého procesu je seznam kandidátů seřazených podle jejich asociačního skóre, korespondujícího ke třem základním přístupům, kterým je určeno měřením:

1) statistické asociace mezi komponenty jednotlivých proměnných
2) souvislostí v kontextu proměnných
3) rozdílnost v kontextu proměnných a jejich komponent

Díky existenci mnoha měr nebo koeficientů, které ukazují vztah mezi dvěma proměnnými, není neobvyklé najít výzkumného pracovníka, který vybere nesprávný koeficient pro míru asociace. To může vést k nesprávnému či zavádějícímu závěru. (Khamis, 2008)

Výzkumníci v rámci obecné analýzy souboru dat často chtějí určit sílu vztahu mezi dvěma proměnnými pomocí jednoho koeficientu, nebo míry asociace. Konkrétně to bývají čísla mezi -1 a +1 nebo mezi 0 a 1. Tato čísla se pak využívají jako měřítko síly příbuznosti těchto dvou proměnných. Pro rozhodnutí o vhodném měřítku asociace je nutné určit úroveň měření každé studované proměnné. Podle těchto informací se určuje příslušná míra asociace. (Khamis, 2008)

Pro každou situaci může existovat několik různých asociačních měr, které jsou platné. Následující kapitola se zaměřuje na výběr jednoho až dvou nejvhodnějších způsobů pro každou situaci, s níž se můžeme v praxi setkat. (Khamis, 2008)

2.4.1 Výběr míry asociace

Při studování vztahu mezi dvěma znaky pomocí jediné míry nebo koeficientu, je nutné zvolit vhodný typ. Pro zjednodušení se výběr patřičné míry zakládá výhradně na typu měřené veličiny, tedy úrovně měr. (Khamis, 2008)

Z praktického hlediska je šest možných kombinací neznámých, které se nejčastěji vyskytují: (Khamis, 2008)

Dvě spojité proměnně

Vhodným koeficientem pro výpočet síly lineárního vztahu mezi dvěma spojitými proměnnými je ve většině případů Pearsonův korelační koeficient, r. Tato hodnota leží mezi -1 a +1. Hodnoty, které jsou blížící se -1 znamenají silný negativní lineární vztah. Pokud se tedy hodnota jedné proměnné zvyšuje, hodnota druhé klesá. Například s přibývajícím věkem se snižuje výkonost pracovníka. Hodnoty blízké +1 pak znamenají pozitivní lineární vztah, tedy jak se hodnota jednoho znaku zvyšuje, zvyšuje se i hodnota druhého. Například roční příjem se zvyšuje se zvyšováním zkušeností pracovníka. (Khamis, 2008)

Pokud jsou hodnoty r blízké nule tak naznačují, že lineární vztah mezi těmito dvěma proměnnými není. Tato skutečnost může nastat, pokud jsou dvě proměnné nezávislé, tj. znalost hodnoty jedné proměnné nijak nepomáhá předpovídat hodnotu druhé proměnné. Nebo v případě že jsou proměnné vysoce odlišné, či mají nelineární vztah. (Khamis, 2008)

Pro interpretaci hodnoty r neexistuje univerzální pravidlo, ale jsou k nim poskytnuty neformální pokyny. Většina studií zahrnujících lékařské, biomedicínské, biologické, zdravotnické, sociologické, vzdělávací a psychologické údaje, se řídí dle pokynů zahrnuty do následující tabulky:

r	**Interpretace měr lineárního vztahu**
0,8	Silně pozitivní
0,5	Středně pozitivní
0,2	Slabě pozitivní
0,0	Žádný vztah
-0,2	Slabě negativní
-0,5	Středně negativní
-0,8	Silně negativní

Tabulka 1: Interpretace lineárního vztahu *Autor: Khamis, 2008*

V případě extrémních hodnot jedné nebo obou zkoumaných veličin, je vhodnějším měřítkem lineárního vztahu Spearmanův korelační koeficient hodnosti. To platí zejména v případě, kdy je požadován test statistické významnosti vztahu. Hodnoty tohoto koeficientu se také pohybují od -1 do +1 a jsou interpretovány stejným způsobem jako r. (Khamis, 2008)

Spojitá – ordinální

Je-li jedna neznámá spojitá a druhá ordinální, pak je vhodné použít metodu výpočtu míry zvanou Kendallův koeficient hodnostní korelace, τ_b. Při označení proměnné X jako spojité, a Y jako ordinální, potom jsou úrovně Y numericky kódované podle pořadí úrovní. Následně Kendallův koeficient, τ_b, použije numerické hodnoty X a kódované numerické hodnoty Y k vykreslení čísla (koeficientu) mezi -1 a +1, které měří sílu vztahu mezi X a Y. (Khamis, 2008)

Pokud má ordinální proměnná Y velké množství rovní, tedy 5 nebo 6 a více, pak lze použít Spearmanův koeficient hodnosti k výpočtu sily míry asociace mezi X a Y. Při tomto použití je nutná opatrnost při numerickém kódování úrovní Y, aby bylo co nejvíce smysluplné. Článek Measures of Association, How to Choose? (Khamis, 2008) uvádí dva typické příklady takového kódování. Prvním příkladem je, pokud Y představuje stupeň shody, pak úrovně kódujeme následujícím způsobem: 1 = velmi silně souhlasím, 2 = silně souhlasím, 3 = souhlasím, 4 = neutrální, 5 = nesouhlasím, 6 = silně nesouhlasím a 7 = velmi silně nesouhlasím. Další příklad zahrnuje rozsahy příjmů, kde každá úroveň je kódována středem mezi nejnižším a nejvyšším příjmem v rozsahu. (Khamis, 2008)

Spojitá – nominální

Koeficient bodové biseriální korelace je vhodné požít, pokud je jedna proměnná spojitá a druhá nominální pouze se dvěma kategoriemi. Tento koeficient se opět pohybuje mezi -1 a +1. Jestliže nominální proměnná má více než dvě úrovně, pak lze vypočítat bodovou biseriální korelaci mezi spojitou proměnnou a všemi možnými páry úrovní nominální proměnné. Výsledkem takového počítání by vzniklo $\frac{k-1}{2}$ korelačních koeficientů, kde *k* představuje počet úrovní nominální proměnné. (Khamis, 2008)

Při výpočtu tohoto koeficientu se provádí kódování dvou úrovní binární proměnné a to, „0“ a „1“a získání Pearsonova korelačního koeficientu mezi spojitou proměnnou a těmito zakódovanými binárními proměnnými. (Khamis, 2008)

Ordinální – ordinální

Pro výpočet míry asociace mezi dvěma ordinálními znaky je vhodné použít Kendallův koeficient τ_b. Pokud obě ordinální neznámé mají velký počet úrovní, je vhodné použít schéma numerického kódování a vypočítat Spearmanův korelační koeficient hodnosti. Stejně jak bylo řečeno dříve v případě, kdy jsme řešili situaci míry asociace mezi spojitou a ordinální proměnnou. (Khamis, 2008)

Ordinální – nominální

V případě, že předmětem analýzy je jedna proměnná nominální a druhá ordinální, platí zde stejný princip počítání jako v případě, kdy by se počítalo s jednou proměnnou spojitou a druhou nominální. Proto i zde se použije biseriální koeficient korelace. (Khamis, 2008)

Nominální – nominální

Zvažujeme-li dvě diskrétní nominální neznámé a předpokládáme-li, že obě proměnné mají pouze dvě úrovně, pak je výsledné zobrazení dat ve formě kontingenční tabulky 2x2. Jedním z běžných způsobů výpočtu míry asociace v takové tabulce je použití koeficientu phi, ϕ. Hodnoty tohoto koeficientu leží mezi 0 a 1. Pokud jsou tyto hodnoty blízké 0, tak značí velmi malou asociaci, naopak hodnoty blížící se k hodnotě 1 značí téměř dokonalou spojitost. (Khamis, 2008)

Pro dvě nominální proměnné, kdy alespoň jedna z nich má více než dvě úrovně, je užitečným měřítkem asociace Goodmanova-Kruskalova lambda, λ. Tato hodnota je relativní pokles pravděpodobnosti chyby při odhadování úrovně jednoho ze znaků mezi úrovní jednoho známého a jednoho neznámého znaku. Hodnoty λ stejně jako ϕ opět leží mezi 0 a 1 se stejnou logikou míry asociace zmiňovanou výše. (Khamis, 2008)

Seznam možných statistických metod pro výpočet měr asociací v daných kombinacích neobsahuje všechny možné metody, ale nejčastěji používaných a prakticky užitečných koeficientů. Většina těchto možností je k dispozici ve většině statistických softwarových balíčcích. (Khamis, 2008)

Koeficient korelace nebo asociace počítaný ze vzorku, je měřen s určitým rozpětím chyby. Obecně platí, že čím menší je velikost vzorku, tím je větší míra chyby. Statistické softwary mohou poskytnou standardní chybu (SE) spolu s odhadem koeficientu. Míru chyby lze nejčastěji vypočítat jako přibližný dvojnásobek standardní chyby. Přibližně s 95 % spolehlivostí lze poté dospět k závěru, že interval, odhadový koeficient ± 2 * SE, obsahuje „skutečnou" nebo „populační" hodnotu. (Khamis, 2008)

Za skutečnou nebo populační hodnotou se považuje následující: vezme-li se v úvahu populaci subjektů, které jsou předmětem studia, jako jsou všichni 50 až 60 letí běloušští Američané. V praxi se získá z této populace náhodný vzorek (např. Náhodně vybere 100 subjektů z populace), získají se hodnoty dvou sledovaných proměnných (například věk a hladinu cholesterolu), a poté je vypočítán požadovaný koeficient (například Pearsonův korelační koeficient). Tato hodnota je odhadovaný koeficient. Pokud by byla hodnota koeficientu vypočtena z celé populace namísto pouhých 100 náhodně vybraných subjektů, byla by získána skutečná nebo populační hodnota koeficientu. (Khamis, 2008)

V následující tabulce jsou shrnuty míry asociací dle kombinací typů neznámých, které jsou podrobněji popsány výše v této kapitole. Hodnoty všech koeficientů měřící sílu asociace leží mezi -1 a +1 se stejnou logikou síly vztahu.

Míry asociace	Kombinace proměnných	Vysvětlení
Pearsonův korelační koeficient, r	Dvě spojité proměnně	Hodnota leží mezi -1 a +1 $0 < r \leq +1$ pozitivní lin. vztah $-1 \leq r < 0$ negativní lin. vztah Hodnota blízko 0 značí možnost neexistence vztahu mezi proměnnými
Spearmanův korelační koeficient hodnosti	Dvě spojité proměnně	Hodnota leží mezi -1 a +1. Použije se, pokud je požadován test statistické významnosti extrémních hodnot.
	Spojitá – ordinální	Použije se, pokud ordinální proměnná Y má 5 nebo 6 a více numericky kódovaných možností
	Ordinální – ordinální	Stejné jako v případě kombinace proměnných Spojitá – Ordinální
Kendallův koeficient hodnostní korelace, τ_b	Spojitá – ordinální	Koeficient využívá hodnoty proměnných X a Y k vykreslení čísla mezi -1 a +1. Kde X je spojitá proměnná s numerickými hodnotami a Y je

		ordinální proměnná s numericky zakódovanými úrovněmi.
	Ordinální – ordinální	Stejné jako v případě kombinace proměnných Spojitá – Ordinální
Koeficient bodové biseriální korelace	Spojitá – nominální	1) Používá se, pokud máme 1 proměnnou spojitou a druhou nominální pouze se 2 kategoriemi 2) Pokud má nominální prom. více než 2 kategorie, provádí se výpočet tohoto koeficientu mezi spojitou proměnnou a všemi možnými páry úrovní nominální proměnné. ➔ Vznik $\frac{k-1}{2}$ korelačních koeficientů (k = počet úrovní nominální proměnné)
	Ordinální – nominální	Stejný princip jako v případě kombinace proměnných spojitá – nominální
Koeficient phi, ϕ	Nominální – nominální	V případě, kdy obě nominální proměnné mají pouze 2 úrovně.
Goodmanova a Kruskalova lambda, λ	Nominální – nominální	λ je užitečným měřítkem pro 2 nominální proměnné, když alespoň 1 z nich má více než 2 proměnné.

Tabulka 2: Přehled základních měr asociací *Autor: Autor práce*

2.5 Hypotézy

Hypotéza je určité očekávání o povaze věcí odvezené většinou z teorie. Je to tedy tvrzení o tom, jaká má tato povaha být, má-li teorie ze které je odvozena být pravdivá. Výzkumná hypotéza je předběžný předpoklad, domněnka o existenci a příčinách jevů, vztahu mezi jevy, průběhu nějakého procesu, změn apod. (Rabušic, Soukup a Mareš 2019)

Rozlišují se následující druhy hypotéz (Rabušic, Soukup a Mareš, 2019):

- *Hypotézy teoretické a empirické*
- *Hypotézy výchozí a pracovní*

  Statistická analýza slouží k ověřování pracovních hypotéz. Tuto hypotézu definují ve své publikaci Rabušic, Soukup a Mareš (2019) způsobem, který převzali z publikace pana Dismana (1993). Definici rozdělili do následujících tří důležitých bodů:

1) *„Pracovní hypotéza je tvrzení předpovídající souvislosti mezi dvěma nebo více proměnnými*
2) *Všechny proměnné zmíněné v hypotéze musejí mít validní operační definici*
3) *Soubor pracovních hypotéz musí zahrnovat nejen proměnné reprezentující zkoumané koncepty ale i ty proměnné, které mohou významně zkreslit interpretaci testovaných vztahů"*

- *Hypotézy kauzální a vztahové*

  Ve velkém množství případů se ve statistické analýze vyskytuje testování hypotéz o vztahu dvou, nebo více neznámých. Takové hypotézy se také označují jako vysvětlující (explanační) hypotézy. V případě těchto hypotéz je důležité především zda „mezi sledovanými proměnnými existují nějaké vztahy a povaha těchto vztahů". Tyto vztahy mohou být identifikované jako kauzální nebo statistické. Způsob, jakým určit, jestli se jedná o kauzální vztah mezi dvěma znaky uvádí Rabušic, Soukup a Mareš ve své publikaci, dle návodu od Lazarsefelda:

  1) *„Proměnné musí být empiricky asociovány nebo korelovány, musejí existovat jejich souběžné změny.*
  2) *Kauzální proměnná, nebo také příčina, musí v čase předcházet proměnné, kterou ovlivňuje, tedy důsledku.*
  3) *Pozorovaný důsledek nemůže být vysvětlen působením jiných proměnných."*

  Nalezená asociace nebo korelace tedy nepředstavuje kauzalitu. Z pohledu empirické sociologie se ve většině případů pracuje se statistickými vztahy, které nesplňují třetí podmínku identifikace kauzálního vztahu. Závěry tohoto pohledu pak mají vždy pravděpodobnostní charakter. Tedy s určitou pravděpodobností proměnná X má vliv na proměnou Y.

- *Hypotézy věcné a statistické*

  „Věcnou hypotézou se rozumí domněnka o existenci vztahu mezi dvěma nebo více proměnnými." Hypotetickým tvrzením o relacích vyvozených ve vztahu ve věcné hypotéze se vyjadřuje statistická hypotéza o existenci vztahu mezi dvěma nebo více neznámýma.

Za věcné hypotézy lze považovat:

1) Existence, výskyty a stavy předmětů, jevů, událostí lidí, skupin – týkající se toho, zda existují či proběhly a jak byly početné.
2) Vlastnosti zmíněných případů – početnost jejich určité kategorie a jejich intenzita.
3) Vztahy v rámci těchto subjektů – *„resp. zda mezi jejich vlastnostmi existovaly, jaké byly povahy a kdy nastaly“.*
4) Vývoj subjektů, jeho etap a stadií a jejich charakteristik.
5) Procesy, jichž se dané subjekty účastní nebo v nich probíhají; *„resp. mohou se týkat vztahů mezi těmito skutečnostmi*“.

2.5.1 Testování hypotéz

Každá vědecká studie by měla začít jasně definovanou výzkumnou otázkou. Tato kapitola se bude zabývat základními předpoklady statistické analýzy, které jsou dodržovány při testech bivaričních asociací, které zahrnují jeden vstup a jeden výstup. (Gonzalez-Chica DA a další, 2015)

Hypotézu lze tedy také definovat jako „*domněnku či tvrzení, které zadavatel výzkumu formuluje“*. Stemmark (2015) Zformulované tvrzení lze zpochybnit nebo otestovat a ve vědeckých studiích může být vyvráceno. Zamítne-li se zformulovaná hypotéza, činí se tak ve prospěch takzvané nulové hypotézy, značené H0. Ta zpravidla tvrdí, že mezi dvěma zkoumanými vzorky či parametry není žádný rozdíl nebo žádná asociace. Tento předpoklad je přijat jako pravdivý. Pro danou situaci existuje takzvaná alternativní hypotéza značená jako HA. Tato hypotéza je dalším vysvětlením pro stejnou situaci, která může nahradit H0 a musí být testována. (Gonzalez-Chica DA a další, 2015)

Pokud je potřeba rozhodnout o zamítnutí H0, musí se definovat maximální přijatelná pravděpodobnost chyby typu I, také značená jako „alfa“. Obvykle je tento typ chyby pro dvoustranné testy nastaven na 5 %. Pro jednostranné testy je nastaven na 2,5 %. Hodnota, která se požívá ke stanovení. zda je výsledek „statisticky významný“ nebo není, se běžně označuje jako p-hodnota. Takovéto statistické testy pomáhají především vědcům hledajícím p-hodnotu na základě statistického vzorce a referenční tabulky pravděpodobností, které odpovídají chybě typu I. (Gonzalez-Chica DA a další, 2015)

Pro odhad p-hodnoty je třeba kombinovat výsledky statistického testu s počtem stupňů volnosti. Výsledkem je kombinace mezi počtem hodnocených jednotlivců a/nebo počtem srovnávaných skupin či kategorií. Stupně volnosti úzce souvisí s chybou typu I, protože čím menší je počet jednotlivců a/nebo čím větší počet srovnávaných skupin, tím menší je pravděpodobnost, že výsledek bude „statisticky významný". Tedy že p-hodnota bude menší než 0,05, nebo menší než 5 %. (Gonzalez-Chica DA a další, 2015)

2.5.2 Výběr testu

Při výběru statistického testu je potřeba znát několik aspektů. První se týká typu dostupných údajů, kdy tyto údaje mohou být „závislé" nebo „nezávislé". Údaje se považují za nezávislé, pokud parametry zjištěné u jednotlivce nezávisí na hodnotách pozorovaných u jiných respondentů ve vzorku. (Gonzalez-Chica DA a další, 2015)

Hlavním kritériem pro určení typu testu, který by měl být vybrán pro analýzu numerických výsledků je právě symetrie proměnné. (Gonzalez-Chica DA a další, 2015)

Proměnná je považována za symetrickou („normální"), pokud je její průměr a medián podobný a rozptyl hodnot je stejný vlevo i vpravo od centrální tendence. Pro takové znaky jsou nejvhodnější „parametrické" testy. (Gonzalez-Chica DA a další, 2015)

„Neparametrické" testy je pak vhodné použít v případě asymetrických hodnot, nebo se je pokusit transformovat na symetrické – například použitím přirozeného logaritmu. (Gonzalez-Chica DA a další, 2015)

T-test a ANOVA, oba testy mají další předpoklady, a to že rozptyl nebo směrodatná odchylka by měla být mezi srovnávanými skupinami homogenní. K ověření tohoto předpokladu lze použít specifické testy, jako je Bartlettův test. Pokud p-hodnota v tomto testu bude menší než 0,05, tak existuje heterogenita rozptylů mezi skupinami. Když odchylky nejsou homogenní, doporučuje se použít neparametrický test, přestože jsou hodnoty symetrické. Jinak by mohla být zkreslena p-hodnota vyplývající s t-testu nebo ANOVA testu. (Gonzalez-Chica DA a další, 2015)

T-test je však považován za „robustní", protože pokud je počet subjektů ve vzorku větší než 100 rovnoměrně rozdělen do exponovaných a neexponovaných skupin, a výstup je symetrický, bude p-hodnota spolehlivá. To i případě kdy je rozptyl mezi skupinami heterogenní. (Gonzalez-Chica DA a další, 2015)

Statistická konfirmace pomocí **Post-hoc testů,** jako jsou Bonferroniho, Scheffé, Newman-Keuls nebo Duncanův test, se doporučuje při kontrole průměrných výsledků v každé kategorii. Tedy ve všech možných kombinacích, tak aby bylo možné určit, které skupiny byly od sebe statisticky odlišné. Vzhledem k tomu, že post-hoc testy zahrnují opravu počtu provedených srovnání, kromě zohlednění počtu jednotlivců v každé kategorii, jsou konzervativnější. Je tak méně pravděpodobné, že vykazují významnou p-hodnotu, i když by výsledek testu ANOVA byl menší než 0,05. (Gonzalez-Chica DA a další, 2015)

Pokud je nezávislá proměnná ordinální a HA poukazuje na existenci určitého trendu tykající se výsledku, (zvýšení nebo snížení průměrného výsledku s kategoriemi vstupující proměnné), lze se rozhodnout pro využití testu lineárního trendu, namísto testu heterogenity. (Gonzalez-Chica DA a další, 2015)

Jsou-li vstupní a výstupní proměnné číselné, může výzkumný pracovník použít Pearsonův korelační test, nebo jednoduché lineární regrese. Také v těchto případech musí být vstupní proměnná symetrická. Spermanova korelace by pak měla být použita, pokud vstupní a výstupní proměnné nejsou symetrické, nebo pokud symetrické jsou, ale mezi zkoumanými veličinami není lineární vztah. (Gonzalez-Chica DA a další, 2015)

Každý subjekt má vstupní hodnotu (osa x) a výstupní hodnotu (osa y), kterou lze vykreslit jako podmnožinu bodů pozorování. Takové zobrazení se označuje jako rozptylový graf. Každý bod v tomto grafu se nazývá „pozorovaná" hodnota subjektu. Na základě sady pozorovaných hodnot je možné odhadnout „predikční linii" očekávané výstupní hodnoty pro každou hodnotu vstupní proměnné. Rozdíl mezi pozorovanými hodnotami a předpovězenými hodnotami se nazývá zbytková hodnota. Vzhledem k tomu, že predikční čára prochází sadou bodů v rozptylovém grafu, mohou být zbytky kladné a záporné hodnoty. Tyto hodnoty by měly být rozloženy nad, pod a podél této přímky, což se nazývá homoskedasticita. Test Pearsonovy korelace a jednoduchého lineárního testu vychází z odhadů těchto informací. Předpoklady pro oba testy zahrnují: (Gonzalez-Chica DA a další, 2015)

1) Přibližně lineární vztah mezi vstupem a výstupem, tedy predikční linie by měla být přímá a měla by mít nenulový sklon
2) Symetrické rozdělení ostatních hodnot

3) Symetrické rozdělení výsledků a homogenní rozdělení zbytků podél vstupních hodnot.

V rámci testu Pearsonovy korelace se vyhodnocuje takzvaný koeficient korelace (r), který měří lineární vtah mezi dvěma numerickými znaky, které mohou být kladné nebo záporné. Hodnota r je tedy míra, do jaké se „pozorované" hodnoty přibližují predikční linii a poskytuje informaci o směru asociace mezi proměnnými. Koeficient korelace může nabývat hodnot od -1 (perfektní negativní korelace) do +1 (perfektní pozitivní korelace), kde r = 0 znamená, že mezi oběma proměnnými není lineární vztah. Mezilehlé hodnoty lze hodnotit jako silné (r = 0,7-0,9), střední (r = 0,4-0,6) nebo slabé korelace (r = 0,1-0,3). (Gonzalez-Chica DA a další, 2015)

Jedním z parametrů lineární regrese je její koeficient (β), který určuje sklon predikční linie ve vztahu k vodorovné linii vytvořené na základě α. Poskytuje informace o směru asociace mezi proměnnými a také o síle (intenzitě) tohoto vztahu. Možné hodnoty tohoto koeficientu záleží na parametrech proměnných a může se pohybovat od -∞ do + ∞. (Gonzalez-Chica DA a další, 2015)

Zmíněná α označuje intercept nebo jinak výchozí hodnotu, která způsobuje vertikální posun přímky při nulové hodnotě vstupního parametru. (Bořil, 2015) Na základě parametrů proměnných se hodnota může pohybovat od -∞ do + ∞. (Gonzalez-Chica DA a další, 2015)

Při testování asociace mezi dvěma numerickými proměnnými je HA takové, že oba koeficienty „r" a „β" se liší od nuly. (Gonzalez-Chica DA a další, 2015)

V případě párových dat je HA spárovaného t-testu taková, že je rozdíl v prostředcích před (základní linie nebo „T0") a po zásahu (konec studie nebo „T1") odlišný od nuly. V případě více než jednoho hodnocení po zásahu (T1, T2, T3 atd.) je HA taková, že existuje rozdíl mezi alespoň dvěma momenty (ne nutně ve vztahu k T0). Stejné požadavky na symetrii výsledku a homogenitu rozptylů platí pro párovaná data, aby bylo možné použít parametrické testy. (Gonzalez-Chica DA a další, 2015)

Pearsonův chí-kvadrát test se používá v případě kategoriálních výstupů, bez ohledu na počet kategorií nebo neznámých. Příkladem, který uvádí ve svém článku Gonzalez-Chica DA a další (2015), je hodnocení vztahu mezi pohlavím a použitím opalovacího

krému. Chí-kvadrát test zde porovnává „pozorované“ číselné hodnoty (absolutní frekvence), které byly uspořádány v kontingenční tabulce, a „očekávané“ hodnoty.

Základním požadavkem Pearsonova chí-kvadrát testu je, že žádná očekávaná hodnota není rovna nule. Pokud toto není splněno, může být nezbytné znovu definovat výzkumnou otázku. Dalším požadavkem tohoto testu je, že frekvence v kontingenční tabulce nesmí být nižší než 5 ve více než 20 % případů. Pokud k tomu dojde, měl by se použít **Chí-kvadrát test** s Yatesovou korekcí kontinuity, za předpokladu, že celková velikost vzorku je větší než 20. Pro malé výběry se používá **Fisherův exaktní test**. (Gonzalez-Chica DA a další, 2015) (Rabušic, Soukup a Mareš, 2015)

HA ve výše uvedených případech je, že pozorovaná frekvence výsledku je mezi nejméně dvěma kategoriemi vstupní měřené veličiny (test chí-kvadrát heterogenity) odlišná. Pokud je výstup dichotomický a vstup je ordinální znak, lze použít trendový chí-kvadrát test. (Gonzalez-Chica DA a další, 2015)

U párových dat, kde byla expozice i výsledek dichotomický, by se měl použít McNemarův chí-kvadrát test. V případě polytomických proměnných jsou doporučeny testy pro posouzení mezní homogenity (Stuart Maxwell nebo Bhapkar). (Gonzalez-Chica DA a další, 2015)

V rámci jednoho výzkumného projektu může být často nutné použít více než jeden statistický test, a to z důvodu různých testovaných hypotéz. (Gonzalez-Chica DA a další, 2015)

Kromě výběru vhodného statistického testu je třeba zkontrolovat různé předpoklady, aby se zabránilo odhadu zkresleného typu I. chybové hodnoty, které ovlivňují nejen interní platnost studie, ale také extrapolaci výsledků na referenční populaci. (Gonzalez-Chica DA a další, 2015)

2.6 Testy o (ne)závislosti

„Závislost sledovaná u dvou proměnných může být buď symetrická (vzájemná) nebo asymetrická (jednostranná).“ (Řezanková, 2007) Například lze zkoumat symetrickou závislost ve výměně názorů mezi partnery nebo asymetrickou závislost názoru uchazeče na jeho vzdělání. Jednotlivé metody se používají v závislosti na typech proměnných. (Řezanková, 2007)

2.6.1 Testy pro dvě nominální proměnné

Základním testem pro zjišťování závislosti dvou nominálních proměnných je Chí-kvadrát test. (Řezanková, 2007) Tato metoda *"je založena na srovnávání napozorovaných (empirických) a očekávaných četností. Vychází z jednoduché myšlenky, že existuje model rozložení dat, které by vzniklo tak, že mezi sledovanými hodnotami dvou proměnných není žádná asociace".* (Rabušic, Soukup a Mareš, 2015) Rozložení dat by tedy vzniklo působením náhody. Četnostem vzniklých z náhodných dat říkáme očekávané, je to tedy taková četnost, která by vznikla, pokud by platila nulová hypotéza nezávislosti. *„Srovnáváme-li tento model se skutečným, empirickým rozložením reálných dat, zjistíme, zdali model rozložené odpovídá empirickým datům nebo ne."* (Rabušic, Soukup a Mareš, 2015)

Z numerického rozdílu mezi empirickou a očekávanou četností lze vyčíst, zda je vypozorovaná četnost vyšší nebo nižší než očekávaná, kdyby platila nulová hypotéza nezávislosti. Tato hodnota se označuje jako reziduum. (Rabušic, Soukup a Mareš, 2015)

Pro zjištění statistické významnosti, je nutné porovnat získanou hodnotu testové statistiky chí-kvadrát s matematickým modelem rozložení, v tomto případě s modelem chí-kvadrát. Zde je mimo jiné potřeba brát v úvahu také prvek zvaný počet stupňů volnosti, které odpovídají součinu $(k-1)*(l-1)$, kde k a l jsou počty řádkových, respektive sloupcových kategorií. (Rabušic, Soukup a Mareš, 2015)

Předpoklady pro použití tohoto testu byly zmíněny v předchozí kapitole 2.7.2. Výběr testu, dle Řezankové (2007) *„předpokladem pro použití tohoto testu je, aby očekávané četnosti v jednotlivých políčkách neklesly pod hodnotu 5 aspoň v 80% políček a ve zbylých políčkách se vyskytovaly alespoň hodnoty 1. Není-li předpoklad splněn, používají se exaktní testy"*. Exaktní test je více popsán v kapitole 2.6.9 Fisherův exaktní test. (Řezanková, 2007)

K testování nezávislostí dvou nominálních proměnných, kromě statistiky chí-kvadrát, se používá věrohodností poměr s totožnými stupni volnosti jako v případě chí-kvadrát testu. (Řezanková, 2007)

2.6.1.1 Koeficienty symetrických měr (ne)závislosti

Statistika Chí-kvadrát je základem pro koeficienty, které v případě nezávislosti nabývají hodnoty 0. Řezanková (2007) uvádí následující příklady koeficientů určující míru symetrické (ne)závislosti:

a) Pearsonův kontingenční koeficient
b) Koeficient φ
c) Cramérovo V
d) Čuprovův kontingenční koeficient

2.6.1.2 Koeficienty asymetrických měr (ne)závislosti

Mimo symetrických měr byly navrženy také míry asymetrické, které hodnotí intenzitu jednostranné závislosti vysvětlované proměnné a proměnné vysvětlující. „*Můžeme sledovat jednak závislost proměnné Y na proměnné X a závislost proměnné Y na proměnné Y. Ke většině takovýchto dvojic ještě existuje míra symetrická.*“ (Řezanková, 2007) Výpočty asymetrické závislosti jsou založeny na principu analýzy rozptylu, „*při níž se testuje shoda středních hodnot ve skupinách vytvořených na základě kategorií vysvětlující proměnné. Nemůžeme-li předpokládat, že proměnná Y je z normálního rozdělení, pak místo středních hodnot zkoumáme jiné míry polohy, tj u nominální proměnné modus, medián pak u ordinální proměnné.*“ (Řezanková, 2007)

Následující koeficienty pro výpočet intenzity míry asymetrické asociace:

a) **Goodmanova-Kruskalova λ** (lambda), kterou v roce 1854 navrhli Goodman a Kruskal, vychází ze vtahu, který definuje Řezanková (2007) jako proporcionální redukci chyby PRE (*proporcional reduction in error*), který se počítá dle schématu:

$$\mathrm{PRE} = \frac{P\{1\} - P\{2\}}{P\{1\}}.$$

Toto schéma nám říká že, „*pokud sledujeme závislost sloupcové proměnné Y na řádkové proměnné X, pak mohou nastat dvě následující situace:*

{1} sloupce jsou statisticky nezávislé na řádcích,
{2} sloupce jsou funkce řádků“. (Řezanková, 2007)

Je-li nový objekt u kterého je známa hodnota znaku X, ale neznámé hodnoty znaku Y a předpokládá-li se situace {1}, pak bychom pro nový objekt odhadovali hodnotu znaku Y podle modální kategorie, pro kterou platí, že $p_{+M} = max_j(p_{+j})$. Pak tedy P{1} je vyjádření pravděpodobnosti chyby vztahem P{1} = $(1 - p_{+Mo})$, což je variační poměr v znaku Y, kde variační poměr se počítá vzorcem $v = 1 - p_{Mo} = 1 - n_{Mo}/n$. Předpokládá-li se situace {2}, pak se hodnota znaku Y odhaduje vyhledáním maximum řádku, který odpovídá známé hodnotě znaku Y. Toto maximum Řezanková (2007) označuje jako $p_{iMo} = max_j(p_{ij})$. Pak se tedy pravděpodobnost chyby P{2} vyjádří vztahem P{2} = $(1 - \sum p_{iMo})$. (Řezanková, 2007)

Podmínkou pro výpočet tohoto koeficientu lambda je, aby se nenulové četnosti vyskytovaly ve více než jednom sloupci. K vlastnostem tohoto koeficientu patří, že nabývá hodnot z intervalu ⟨0; 1⟩. (Řezanková, 2007)

b) **Goodmanovo-Kruskalovo τ (tau)** je založeno na principu analýzy rozptylu s použitím míry *nomvar,* nebo také Giniho koeficientem, který vyjadřuje relativní počet dvojic, které nejsou ve stejné kategorii, vyjádřený vzorcem:

$$\text{nomvar} = 1 - \sum_{i=1}^{K} p_i^2 = \sum_{i=1}^{K} \left(p_i(1 - p_i)\right),$$

nebo

$$\text{nomvar} = 1 - \sum_{i=1}^{K} \left(\frac{n_i}{n}\right)^2 = \frac{n^2 - \sum_{i=1}^{K} n_i^2}{n^2}.$$

i=1,2…, K, kde K je počet kategorií. (Řezanková, 2007)

Podmínkou pro výpočet tohoto koeficientu je, aby se nenulové četnosti vyskytovaly ve více než jednom sloupci. (Řezanková, 2007)

c) **Informační koeficient** nebo jinak také **koeficient nejistoty** lze získat, pokud jako charakteristiku variability se uvažuje entropie *H*, která je dána vzorcem

$$\text{H} = -\sum_{i=1}^{K} p_{i\ln}\, p_i$$

a jeho dosazením do vzorce

$$S_{Y|X} = \frac{var(Y) - \sum_{i=1}^{R} p_{i+}\, var(Y|x_i)}{var(Y)}.$$

Tento vzorec je odvozen pro výpočet na základě kontingenční tabulky. Obecně se tato míra zapisuje

$$S_{Y|X} = \frac{var(Y,X)}{var(Y)} = \frac{var(Y) - var(Y|X)}{var(Y)}.$$

Míra slouží při analýze rozptylu, kdy se měří intenzita závislosti pomocí poměru determinace, který se počítá jako podíl mezi skupinové variability na celkové variabilitě. (Řezanková, 2007)

Míra souhlasu lze počítat v případě, kdy má tabulka stejný počet sloupců a řádků, jinak také jmenovanou jako čtvercovou. Jedná se o případ, kdy dva sledované znaky, například žena a muž, nabývají stejných kategorií, například rodinný stav. Pro výpočet této míry se používá koeficient, který se nazývá **Cohenovo κ** (kappa). (Řezanková, 2007)

Pokud se nenulové četnosti vyskytují pouze na diagonále, pak koeficient Cohenovo κ nabývá svého maxima, tedy hodnoty 1. Dále pokud jsou hodnoty míry souhlasu větší než 0,75 pak lze předpokládat výborný souhlas a u hodnot menší než 0,4 se souhlas předpokládat nedá. (Řezanková, 2007)

2.6.2 Testy pro dvě ordinální proměnné

Statistická závislost pro dvě ordinální proměnné, také označovaná jako korelace, se rozlišuje na dva typy – pozitivní a negativní. Pozitivní korelace se vyznačuje tím, že *„nízkým hodnotám jedné proměnné odpovídají hodnoty proměnné druhé“* a negativní pak tím, že *„nízkým hodnotám jedné proměnné odpovídají vysoké hodnoty druhé proměnné“*. (Řezanková, 2007)

Důležité je odlišit případy, kdy je pouze jedna proměnná ordinální a kdy obě. V případě, kdy jsou obě sledované proměnné ordinální, může se použít testování, které je založené na pořadí. (Řezanková, 2007)

2.6.2.1 Symetrické koeficienty dvou ordinálních proměnných

Spearmanův koeficient pořadové korelace, se považuje za základní míru mezi dvěma ordinálními znaky, vycházející z toho, *„že každé hodnotě proměnné X je*

přiřazeno pořadí a_l *tak, že* $\sum_{l=1}^{n} a_l = n\frac{n+1}{2}$*, a každé hodnotě proměnné Y je přiřazeno pořadí* b_l *tak, že* $\sum_{l=1}^{n} b_l = n\frac{n+1}{2}$*“.* (Řezanková, 2007)

Koeficient pořadové korelace nabývá hodnot z intervalu $\langle -1; 1 \rangle$. V případě že jsou u každého respondenta obou proměnných stejné pořadí, pak koeficient nabývá hodnoty 1. To značí pozitivní korelaci což znamená přímou závislost. Jestliže vzestupné seřazení hodnot proměnné X způsobilo sestupné pořadí proměnné Y, pak hodnota koeficientu je -1, což je výsledek pro negativní korelaci, takzvanou nepřímou závislost. Výsledek pro lineární nezávislost je roven 0. Test nulové hypotézy H_0: $\rho_s = 0$ se provádí pomocí statistiky

$$\mathrm{t} = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

„která má za předpokladu nulové hypotézy Studentovo t rozdělení $\mathrm{s}(n-2)$ *stupni volnosti“.* (Řezanková, 2007)

Mimo Spearmanova koeficientu existují míry zkoumající dvojice objektů. Sledované dvojice se mohou označovat jako konkordantní, diskordantní nebo vázané. Za konkordantní pár se považují takové dvojice jednoho objektu, jejichž hodnoty u obou proměnných jsou menší (resp. větší) než u druhého objektu. V případě že je u jednoho znaku hodnota menší a u druhé větší, pak jde o pár diskordantní. V ostatních případech, tedy pokud je hodnota jedné proměnné nebo hodnoty u obou proměnných shodné, hovoříme o párech vázaných. (Řezanková, 2007)

Řezanková (2007) uvádí pro zjednodušení zápisů vzorců následující symboly:

„C – počet konkordantních párů,

D – počet diskordantních párů,

T_X *– počet párů, které obsahují stejnou hodnotu proměnné X, ale různou hodnotu Y,*

T_Y *– počet párů, které obsahují stejnou hodnotu proměnné Y, ale různou hodnotu X“.*

Matematicky se tyto symboly vyjadřují dle následujících vzorců:

$$C = \sum_{i=2}^{R}\sum_{j=2}^{S}\left(n_{ij}\sum_{h<i}\sum_{k<j}n_{hk}\right) \qquad D = \sum_{i=2}^{R}\sum_{j=1}^{S-1}\left(n_{ij}\sum_{h<i}\sum_{k>j}n_{hk}\right)$$

$$T_X = \sum_{i=1}^{R}\sum_{j=2}^{S}\left(n_{ij}\sum_{h=i}\sum_{k<j}n_{hk}\right) \qquad T_x = \sum_{i=2}^{R}\sum_{j=1}^{S}\left(n_{ij}\sum_{h<i}\sum_{k=j}n_{hk}\right)$$

Na základě výše zmíněné logiky dvojic existují následující symetrické míry dvou ordinálních znaků:

a) **Goodmanova-Kruskalova γ** (gama), pro který platí že $|\gamma| \in \langle 0; 1\rangle$, přičemž pro tabulky 2 x 2 nemusí nabýt hodnoty 0. Zde opět platí obvyklá interpretace, že v případě, kdy se koeficient rovná 0 to značí nezávislost a hodnota 1 pak znamená plnou závislost. Takový případ nastává, pokud jsou nenulová políčka pouze na diagonále.
b) **Kendallovo τ_b** (tau-b), jinak také Kendallův koeficient pořadové korelace, pro který platí, že $\tau_b \in \langle -1; 1\rangle$. Pouze v tabulce R x R může tento koeficient dosáhnout mezních hodnot, jestliže se žádná marginální četnost nerovná nule.
c) **Kendallovo τ_c** (tau-c). Pro typ tabulky R x S platí $\tau_c \in \langle -1; 1\rangle$.

2.6.2.2 Asymetrické koeficienty dvou ordinálních proměnných

Jediným uvedeným koeficientem pro měření míry mezi dvěma pořadovými znaky Řezanková (2007) uvádí Somersovo d. Jeho symetrická podoba se získá jako harmonický průměr obou asymetrických variant. K vlastnostem tohoto koeficientu dále patří, že geometrickým průměrem jeho dvou asymetrických variant je koeficient τ_b.

2.6.3 Testy pro ordinální vysvětlovanou proměnnou

Pokud hodnoty vysvětlované veličiny Y rozdělíme do skupiny podle vysvětlující proměnné X, pak pro porovnání skupin použijeme **Kruskalův-Wallisův test.** „*Nulová hypotéza předpokládá, že všechny skupiny lze charakterizovat stejnou hodnotou mediánu vysvětlované proměnné, obsahem alternativní je, že alespoň jeden z mediánů je různý od ostatních.*“ (Řezanková, 2007)

„Výpočet testového kritéria je založen na pořadových číslech, která jsou přiřazena hodnotám v souboru, vzniklým spojením všech výběrů.“ (Řezanková, 2007)

Hodnota Kuskalovy-Wallisovy statistiky má při platnosti nulové hypotézy o nezávislosti přibližně rozdělení Chí-kvadrát s (R-1) stupni volnosti. (Řezanková, 2007)

2.6.4 Testy pro dvě kvantitativní proměnné

Pearsonův korelační koeficient se používá v případě kdy znaky mezi kterými je zkoumaná vzájemná závislost jsou kvantitativní a nabývá hodnot z intervalu $\langle -1; 1 \rangle$. Stejně jak je již zmíněno pro dvě ordinální proměnné, *„hodnota 0 znamená lineární nezávislost, hodnoty 1 pozitivní korelaci (přímou závislost) a hodnota -1 negativní korelaci (nepřímou závislost)“.* (Řezanková, 2007)

Statistika pro test o nezávislosti dvou proměnných založený na tomto koeficientu se počítá stejným způsobem jako v případě Spearmanova korelačního koeficientů. Předpokladem tohoto testu v tomto případě je, že se jedná o výběr z dvourozměrného normálního rozdělení. Jestliže tato podmínka není splněna, obvykle se postupuje přiřazením pořadí hodnotám a použitím výběrového Spearmanového korelačního koeficientu. (Řezanková, 2007)

2.6.5 Testy pro kvantitativní vysvětlovanou proměnnou

Pro zjištění závislosti vysvětlované kvantitativní neznámé na nominálním znaku se používá analýza rozptylu. Je pro ni nutné, aby vysvětlovaná proměnná splňovala podmínky normality. *„Intenzitu statistické závislosti můžeme posoudit buď podle poměru determinace nebo podle jeho odmocniny, která bývá označována jako koeficient η (éta). Tento koeficient nabývá hodnot z intervalu* $\langle 0; 1 \rangle$*.“* (Řezanková, 2007)

2.6.6 Testy pro dvě dichotomické proměnné

„Kontingenční tabulka pro dvě dichotomické proměnné má čtyři políčka a nazývá se čtyřpolní.“ (Řezanková 2007) V případě čtyřpolní tabulky Řezanková (2007) uvádí následující možné statistiky pro výpočet:

a) **Chí-kvadrát statisktika**, kdy náhodná veličina má za předpokladu platnosti nulové hypotézy asymptoticky chí-kvadrát rozdělení pouze s jedním stupněm volnosti.
b) **Yatesova korekce**, nebo také statistika chí-kvadrát korigované na spojitost.

c) **Věrohodnostní poměr** – statistika pro tento výpočet má *„asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti“*. Stejně jako u Chí-kvadrát statistiky by měl být splněn požadavek na velikost očekávaných četností.
d) **Mantelova-Heanstelova statistika chí-kvadrát** je stejně jako Pearsonova statistika chí-kvadrát funkcí korelačního koeficientu. Stejně jak bude zmíněno u Fisherova exaktního testu, tato statistika vychází z předpokladu, že četnosti jsou výběrem hypergeometrického rozdělení. Statistika má opět chí-kvadrát rozdělení s jedním stupněm volnosti.

Pro zjišťování míry asociace dále Řezanková (2007) uvádí následující možnosti koeficientů:

a) **Pearsonův korelační koeficient**, jinak také označován jako **koeficient asociace**, pro který Řezanková (2007) uvádí odvození zjednodušeného tvaru. Pro výpočet je potřeba znát kovarianci a směrodatné odchylky, jejichž výpočet je založen na odchylkách od aritmetického průměru. Korelační koeficient je pak podílem kovariance a součinu směrodatných odchylek. V tomto případě se jedná o míru symetrickou, která hodnotí intenzitu vzájemné závislosti a též její typ, tedy zda je přímá nebo nepřímá. Závislosti vycházející z chí-kvadrát statistiky jsou v případě čtyřpolní tabulky funkcí korelačního koeficientu. Dále platí, že vzorec pro Cramerovo V je v tomto případě shodný se vzorcem pro koeficient φ.
b) **Yuleovo Q** je další symetrickou mírou. Zde se předpokládá že oba analyzované dichotomické znaky jsou stejně kódované, například pomocí hodnot 0 a 1. Součin četností na hlavní diagonále pak vyjadřuje počet párů objektů, z nichž první obsahuje u jedné proměnné nulu a u druhé jedničky, zatímco u druhého objektu je to naopak. Tento koeficient lze mimo jiné použít i u ordinálních proměnných, kde zajišťuje relativní přebytek konkordantních párů nad páry diskordantními. Jde tedy o speciální název pro **koeficienty γ** (více o tomto koeficientu a párů je zmíněno v kapitole 2.6.2.1)**.**
c) **Hamannův koeficient** označuje míru souhlasu pro čtyřpolní tabulku a nabývá hodnot z intervalu $\langle -1; 1 \rangle$. Nacházejí-li se nenulové četnosti pouze na hlavní diagonále, což znamená, že byly zaznamenány pouze dvojice shodných kategorií, pak koeficient nabývá hodnoty 1. Naopak v případě, kdy nebyla

zjištěna žádná dvojice shodných kategorií, koeficient nabývá hodnoty -1, a hodnoty 0 v případě stejného počtu shod jako případů neshod.

d) **Poměr šancí** je případ míry, která je určena pouze pro čtyřpolní tabulky. *„Používá se zejména v případech, kdy se máme rozhodnout pro jednu ze dvou možností, například zda se léčit nebo neléčit, zda se zúčastnit školení atp.*
Pokud jsou kategorie ve čtyřpolní tabulce vhodně uspořádány, pak poměr součinů četností na hlavní diagonále je poměrem šancí, který indikuje, zda je lepší podniknout určitou aktivitu, či nikoliv.“ Pokud je hodnota menší než 1 je výhodnější aktivitu nepodniknout.

e) **Procentní rozdíl** je jedna z nejjednodušších vyjádření míry závislosti založených na chí-kvadrát pro čtyřpolní tabulku. Jedná se o míru asymetrickou, která hodnotí intenzitu závislosti neznámé vysvětlované na proměnné vysvětlující.

f) **Yuleův koeficient vazby**

g) **Kendallovo τ_b** (tau-b)

h) **Sommersovo d**

Řezanková (2007) dále uvádí speciální test pro případ čtyřpolní tabulky, **McNemarův test**. Jedná se o párový test, který lze využít *„například při zjišťování, zda se shodují názory týchž respondentů ve dvou různých obdobích“.* Řeší tedy vzájemný vztah mezi dvojicemi hodnot, kde se testuje nulová hypotézu o shodě četností v políčkách na vedlejší diagonále.

Polovina hodnoty minimální hladiny významnosti pro exaktní test vyjadřuje pravděpodobnost, že náhodná veličina s binomickým rozdělením nabude minimální hodnoty z vedlejší diagonály nebo menší.

Je-li součet hodnot z vedlejší diagonály vetší než 25, lze použít aproximaci binomického rozdělení rozdělením chí-kvadrát.

2.6.9 Fisherův exaktní test

Fisherův exaktní test se provádí v případě, kdy je potřeba použít test o nezávislosti dvou znaků, ale nebyl by splněn požadavek týkající se očekávaných četností. *„Fisherův exaktní test vychází z předpokladu, že marginální četnosti jsou považovány za neměnné a že data jsou tudíž výběrem z hypergeometrického rozdělení.“* (Řezanková 2007) Jelikož

jsou četnosti všech políček, kromě levého políčka, které odpovídá prvnímu řádku a prvnímu sloupci, určeny marginálními četnostmi, nelze je testovat. Lze tedy testovat pouze zmíněné levé políčko, které je jako jediné určené relativní četností. (Řezanková, 2007)

„Počítají se pravděpodobnosti výskytu všech možných variant četností v kontingenční tabulce, které dávají stejné marginální četnosti jako tabulka četností za zjištěných dat.“ (Řezanková 2007) Takovéto pravděpodobnosti se získávají vyhledáním hodnot pravděpodobnostní funkce hypergeometrického rozdělení. (Řezanková, 2007)

2.6.10 Testy pro tři proměnné (2 dichotomické a 1 více kategoriální)

„Postupy používané při analýze dvourozměrných kontingenčních tabulek lze rozšířit na analýzu tří proměnných.“ (Řezanková 2007) V takovém to případě se jedná o trojrozměrnou kontingenční tabulku 2 x 2 x L nebo o soustavu L čtyřpolních tabulek, což je výsledek rozdělení čtyřpolní tabulky podle třetí proměnné o L kategorií. (Řezanková 2007)

Řezanková (2007) uvádí pro nezávislost dvou dichotomických proměnných podmíněnou další kategoriální proměnnou následující statistiky:

a) **Cochranova statistika**
b) **Mantelova-Haenszelova statistika**

V obou případech se používá korekce na spojitost a úprava pří výpočtu rozptylu a obě mají asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti. (Řezanková 2007)

Koeficient pro stanovení míry asociace pro L čtyřpolních tabulek Řezanková (2007) uvádí **Mantelův-Haenszelův odhad společného poměru šancí**. Míra tohoto koeficientu nabývá v případě nezávislosti hodnoty 1, a proto *„pro testování o nezávislosti se vychází z přirozeného logaritmu společné hodnoty. Test vyžaduje, aby se charakter závislosti v jednotlivých tabulkách příliš nelišil.“* Homogenitu v jednotlivých tabulkách lze testovat pomocí *Breylovy-Dayovy statistiky,* případně pomocí *Taroneovy statistiky*. (Řezanková 2007)

Na základě zmíněných metod a vlastností zmíněných v publikaci Mareše, Rabušice a Soukupa (2015), je sepsán následující přehled koeficientů pro výpočet míry asociace a jejich základní vlastnosti.

Koeficienty	**Rozsah hodnot**	**Symetrický**	**Typy proměnných**
Pearsonův kontingenční koeficient	$\langle 0;\ \sqrt{(q-1)/q}\rangle$	Ano	Dvě nominální proměnné
Koeficient φ	$\langle 0;\ 1\rangle$	Ano	Dvě dichotomické proměnné
Cramérovo V	$\langle 0;\ 1\rangle$	Ano	Dvě nominální proměnné
Čuprovův kontingenční koeficient	$\langle 0;\ 1\rangle$	Ano	Dvě nominální proměnné
Goodmanova-Kruskalova λ	$\langle 0;\ 1\rangle$	Obě verze	Dvě nominální proměnné
Goodmanovo-Kruskalovo τ	$\langle 0;\ 1\rangle$	Obě verze	Dvě nominální proměnné
Somersovo d	$\langle -1; 1\rangle$	Obě verze	Dvě ordinální proměnné Dvě dichotomické proměnné
Spearmanův koeficient pořadové korelace	$\langle -1; 1\rangle$	Ano	Dvě ordinální proměnné
Goodmanova-Kruskalova γ	$\langle -1; 1\rangle$	Obě verze	Dvě ordinální proměnné
Kendallovo τ_b	$\langle -1; 1\rangle$	Ano	Dvě ordinální proměnné Dvě dichotomické proměnné
Kendallovo τ_c	$\langle -1; 1\rangle$	Ano	Dvě ordinální proměnné
Pearsonův korelační koeficient	$\langle -1; 1\rangle$	Ano	Dvě kvantitativní proměnné Dvě dichotomické proměnné
Spearmanův korelační koeficient	$\langle -1; 1\rangle$	Ano	Dvě kvantitativní proměnné
Koeficient η (éta)	$\langle 0; 1\rangle$	Ne	Kvantitativní vysvětlovaná proměnná
Yuleovo Q	$\langle -1; 1\rangle$	Ano	Dvě dichotomické proměnné
Yuleův koeficient vazby	$\langle -1; 1\rangle$	Ano	Dvě dichotomické proměnné
Hamannův koeficient	$\langle -1; 1\rangle$	Ano	Dvě dichotomické proměnné
Procentní rozdíl	$\langle -1; 1\rangle$	Ne	Dvě dichotomické proměnné

Tabulka 3: Souhrn koeficientů měr asociací

3 Metodika

V této kapitole jsou představeny jednotlivé statistické metody související s výpočtem měr asociací, které jsou uvedeny na základě publikace Hany Řezankové (2007). Dále je zde popsán systém pro analýzu dat. Jedná se především o stručný popis základních obrazovek a prvků, které jsou důležité pro ovládání této aplikace.

3.1 Systém pro analýzu dat

Veškeré statistické postupy a aplikaci následujících principů zjišťování závislostí proměnných budou prováděny prostřednictvím statistického programu IBM SPSS Statistics verze 26.

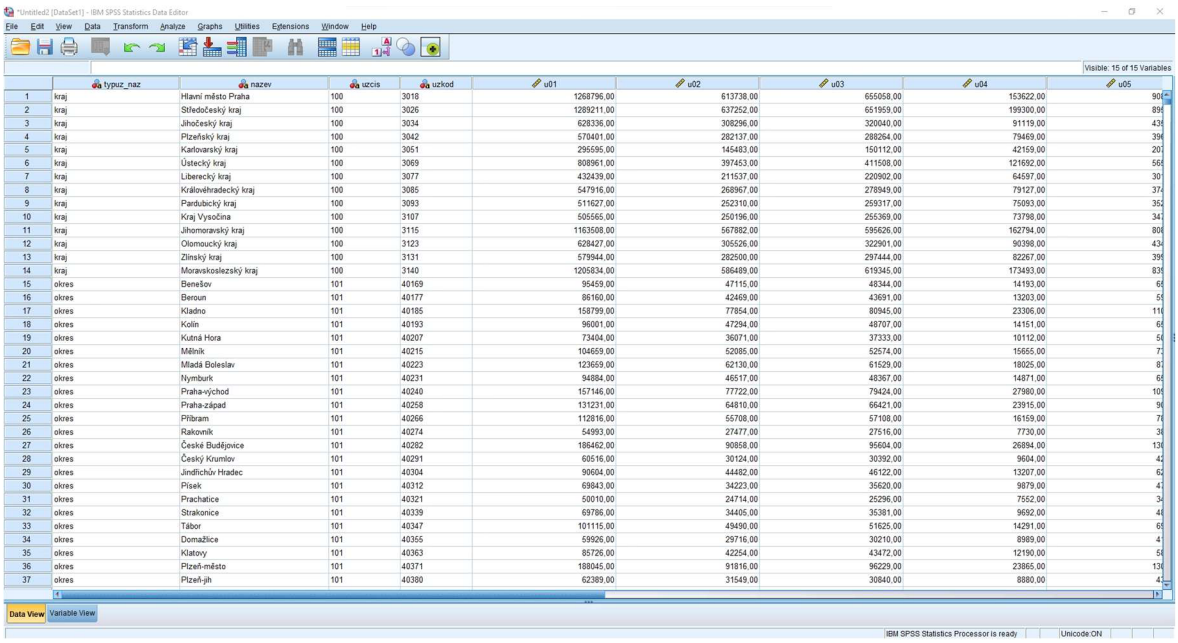
Software SPSS byl vyvinut v době před existencí osobních počítačů, tedy vzniknul v éře velkých sálových počítačů a jeho původní název zněl Statistical Package for Social Sciences neboli SPSS (dále jen SPSS). (Rabušic, Soukup a Mareš, 2019)

3.1.1 Popis základních obrazovek

V této kapitole se jsou uvedeny základní prvky po přípravu dat a následnou analýzu.

Před začátkem analýzy je potřeba data nahrát. Pro správné nahrání dat je nutné nejdříve definovat jednotlivé proměnné a těm pak přiřadit výzkumem zjištěné konkrétní hodnoty. K tomu nám slouží okna *Data View* a *Variable View.* (Rabušic, Soukup a Mareš, 2019)

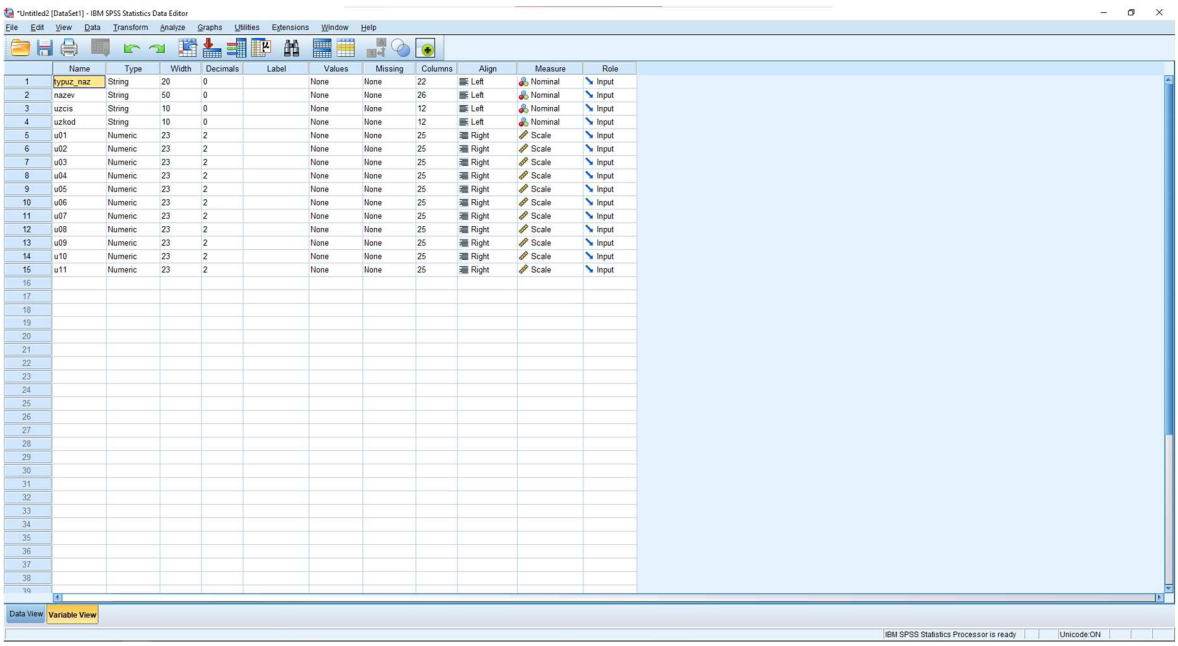
Data View (viz. obrázek 1) je okno, které zobrazuje datovou matici, ve které řádky značí případy (cases) výzkumné jednotky. Jedná většinou o osoby, ale mohou to být také skupiny osob, územní celky, nebo předměty jako texty a podobně. Sloupce matice pak charakterizují dané jednotky v řádcích. Každá zkoumaná jednotka tedy představuje vektor a číslice v něm přestavují kódy hodnot proměnných, tj. u nominálních a ordinálních proměnných, nebo čísla, tj u kardinální nebo také spojité proměnné. Tyto hodnoty popisují vlastnosti/charakteristiky jednotky. (Rabušic, Soukup a Mareš, 2019)



	typuz_naz	nazev	uzcis	uzkod	u01	u02	u03	u04	u05
1	kraj	Hlavní město Praha	100	3018	1268796,00	613738,00	655058,00	153622,00	90
2	kraj	Středočeský kraj	100	3026	1289211,00	637252,00	651959,00	199300,00	89
3	kraj	Jihočeský kraj	100	3034	628336,00	308296,00	320040,00	91119,00	43
4	kraj	Plzeňský kraj	100	3042	570401,00	282137,00	288264,00	79469,00	39
5	kraj	Karlovarský kraj	100	3051	295595,00	145483,00	150112,00	42159,00	20
6	kraj	Ústecký kraj	100	3069	808961,00	397453,00	411508,00	121692,00	56
7	kraj	Liberecký kraj	100	3077	432439,00	211537,00	220902,00	64597,00	30
8	kraj	Královéhradecký kraj	100	3085	547916,00	268967,00	278949,00	79127,00	37
9	kraj	Pardubický kraj	100	3093	511627,00	252310,00	259317,00	75093,00	35
10	kraj	Kraj Vysočina	100	3107	505565,00	250196,00	255369,00	73798,00	34
11	kraj	Jihomoravský kraj	100	3115	1163508,00	567882,00	595626,00	162794,00	80
12	kraj	Olomoucký kraj	100	3123	628427,00	305526,00	322901,00	90398,00	43
13	kraj	Zlínský kraj	100	3131	579944,00	282500,00	297444,00	82267,00	39
14	kraj	Moravskoslezský kraj	100	3140	1205834,00	586489,00	619345,00	173493,00	83
15	okres	Benešov	101	40169	95459,00	47115,00	48344,00	14193,00	6
16	okres	Beroun	101	40177	86160,00	42469,00	43691,00	13203,00	5
17	okres	Kladno	101	40185	158799,00	77854,00	80945,00	23306,00	11
18	okres	Kolín	101	40193	96001,00	47294,00	48707,00	14151,00	6
19	okres	Kutná Hora	101	40207	73404,00	36071,00	37333,00	10112,00	5
20	okres	Mělník	101	40215	104659,00	52085,00	52574,00	15655,00	7
21	okres	Mladá Boleslav	101	40223	123659,00	62130,00	61529,00	18025,00	8
22	okres	Nymburk	101	40231	94884,00	46517,00	48367,00	14871,00	6
23	okres	Praha-východ	101	40240	157146,00	77722,00	79424,00	27980,00	10
24	okres	Praha-západ	101	40258	131231,00	64810,00	66421,00	23915,00	9
25	okres	Příbram	101	40266	112816,00	55708,00	57108,00	16159,00	7
26	okres	Rakovník	101	40274	54993,00	27477,00	27516,00	7730,00	3
27	okres	České Budějovice	101	40282	186462,00	90858,00	95604,00	26894,00	13
28	okres	Český Krumlov	101	40291	60516,00	30124,00	30392,00	9604,00	4
29	okres	Jindřichův Hradec	101	40304	90604,00	44482,00	46122,00	13207,00	6
30	okres	Písek	101	40312	69843,00	34223,00	35620,00	9879,00	4
31	okres	Prachatice	101	40321	50010,00	24714,00	25296,00	7552,00	3
32	okres	Strakonice	101	40339	69786,00	34405,00	35381,00	9692,00	4
33	okres	Tábor	101	40347	101115,00	49490,00	51625,00	14291,00	6
34	okres	Domažlice	101	40355	59926,00	29716,00	30210,00	8989,00	4
35	okres	Klatovy	101	40363	85726,00	42254,00	43472,00	12190,00	5
36	okres	Plzeň-město	101	40371	188045,00	91816,00	96229,00	23865,00	13
37	okres	Plzeň-jih	101	40380	62389,00	31549,00	30840,00	8880,00	4

Obrázek 1: Data view – datová matice

Variable view (viz. obrázek 2) je okno ve kterém jsou jednotlivé proměnné popsány. V SPSS to je zabudovaný speciální tabulkový procesor, který tento popis umožňuje. Před samotným nahráváním dat je potřeba proměnné v tomto okně popsat, bez toho by se konkrétní hodnoty nemohly nahrát. Při popisu proměnných se převádí dotazník do formalizované podoby, kterou SPSS vyžaduje. (Rabušic, Soukup a Mareš, 2019)



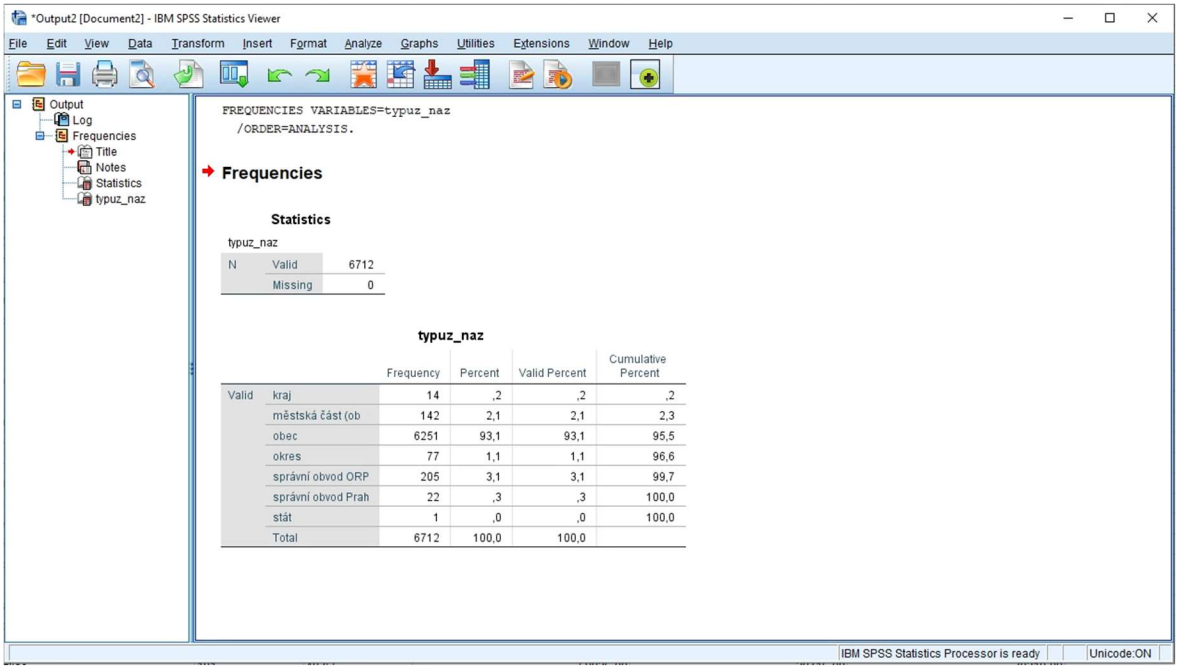
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	typuz_naz	String	20	0		None	None	22	Left	Nominal	Input
2	nazev	String	50	0		None	None	26	Left	Nominal	Input
3	uzcis	String	10	0		None	None	12	Left	Nominal	Input
4	uzkod	String	10	0		None	None	12	Left	Nominal	Input
5	u01	Numeric	23	2		None	None	25	Right	Scale	Input
6	u02	Numeric	23	2		None	None	25	Right	Scale	Input
7	u03	Numeric	23	2		None	None	25	Right	Scale	Input
8	u04	Numeric	23	2		None	None	25	Right	Scale	Input
9	u05	Numeric	23	2		None	None	25	Right	Scale	Input
10	u06	Numeric	23	2		None	None	25	Right	Scale	Input
11	u07	Numeric	23	2		None	None	25	Right	Scale	Input
12	u08	Numeric	23	2		None	None	25	Right	Scale	Input
13	u09	Numeric	23	2		None	None	25	Right	Scale	Input
14	u10	Numeric	23	2		None	None	25	Right	Scale	Input
15	u11	Numeric	23	2		None	None	25	Right	Scale	Input

Obrázek 2: Variable view – popis proměnných

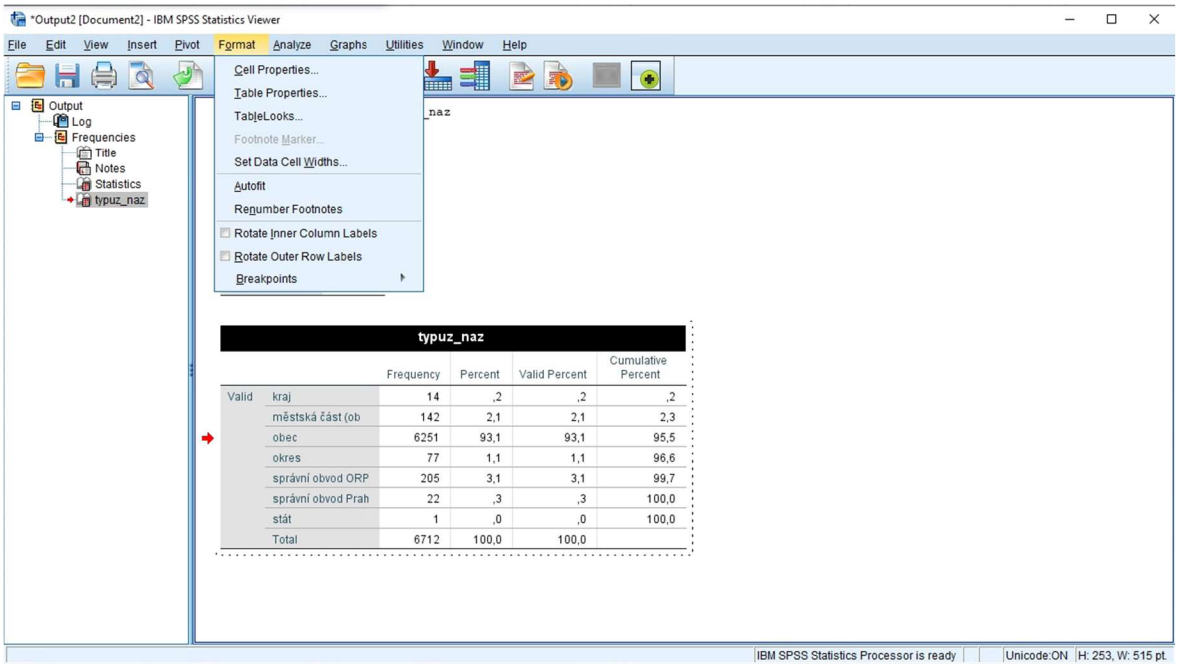
V okně pro popis proměnných jsou jednotlivé znaky (charakteristiky) z datové matice umístěny v řádcích. Sloupce těchto řádků pak vyjadřují jejich základní charakteristiky,

tedy technické jméno proměnné *(Name)*, její popisek *(Label),* popisky jednotlivých hodnot *(Values)*, chybějící hodnoty *(Missing)* atd. (Rabušic, Soukup a Mareš, 2019)

Třetím základním oknem je okno výstupů (viz obrázek číslo 3), které se otevře vždy při spuštěné nějaké procedury. V tomto okně lze vidět výsledky požadovaných výpočtů, tabulky, grafy atp. Tyto výsledky mohou být v okně editovány. Po dvojkliku na výstup, například na tabulku, nebo graf, se objeví jiná nabídka (viz obrázek číslo 4) a je možné měnit jeho grafickou podobu, texty popisků apod. Edituje se především prostřednictvím menu *Edit, Format* nebo také *Pivot*. (Rabušic, Soukup a Mareš, 2019)



Obrázek 3: Output



Obrázek 4: Editace výstupu

Výsledky zobrazeny v okně Output lze uložit příkazem *Save as*. Výsledek se pak uloží v novém souboru s příponou *.spo* nebo *.spv*. Takovéto typy souboru lze pak otevřít jen v IBM SPSS Statistics. Rabušic, Soukup a Mareš (2019) dále upozorňují že různé verze SPSS mají různý formát tohoto souboru, proto zmíněné dva typy souboru lze spolehlivě otevřít jen ve verzi SPSS ve které byl vytvořen. Pro datové soubory (*.sav*) tato nepřenositelnost neplatí. Rabušic, Soukup a Mareš (2019) však doporučují je ukládat ve formátu *.porm* který umí spolehlivě číst všechny verze SPSS a nadto i mnohé jiné softwary jako je například SAS, STATA Transfer a podobně.

Jednotlivé výstupy i celek lze také exportovat do Wodru, což se dle Rabušice, Soukupa a Mareše (2019) nedoporučuje, protože se tabulky často rozpadnou. Doporučují zde výsledek raději v menu *Edit* zablokovat pomocí *Copy Object* nebo lze také provést příkazem *Ctrl+C* a následně ji do wordovského dokumentu vložit pomocí příkazu *Ctrl+V*.

3.1.2 Definice jednotlivých proměnných

Proměnné datového souboru je potřeba po nahrání do SPSS nadefinovat. V záložce *Variables View,* která je již představena výše (viz obrázek 2). V tomto okně se provádí následující úkony, které jmenují ve své publikaci Rabušic, Soukup a Mareš (2019):

a) *„Připsání technického jména proměnné, určení jeho místa v matici (sloupce/sloupců).*
b) *Definice charakteru proměnné jako číselné (numeric) či textové (string); textovou proměnnou počítač chápe jako označení a neprovádí s ní početní operace jako je sčítání či násobení.*
c) *Připsání širšího/podrobnějšího označení proměnné (variables labels)*
d) *Připsání verbálního označení jednotlivým hodnotám (kategorizované) proměnné (values label). Ty zpřehledňují tištěné výstupy, neboť přiřazují ke jménům proměnných i vysvětlující popis.*
e) *Určení počtu desetinných míst (v případě spojitých proměnných)*
f) *Definování uživatelsky chybějící hodnot, takzvané user missing values. Někdy do missing values přeřazujeme některé hodnoty proměnných při jejich transformaci. Týká se to například varianty „nevím“, která sice někdy může být součástí ordinální proměnné jakožto její středová hodnota, častěji jsou ale případy, kdy ji používáme jen proto, abychom nenutili respondenta zaujímat postoj, který nemá.*

V další analýze se pak často soustřeďujeme jen na ty, kdo postoj zaujali a v modulu Transform – Recode přiřadíme odpověď "nevím" číslici označující chybějící hodnotu (missing values). SPSS také zná systémové chybějící hodnoty. Takové hodnoty se zobrazují jako tečky a znamenají, že pro danou proměnnou a daný případ není k dispozici žádná hodnota.

g) *Rozumné je určit u každé proměnné v kolonce Measure úroveň měření. Tato informace totiž u některých statistických operací rozhoduje o volbě počítaných statistik."*

Pole *Name* a *Label* lze vyplňovat ručně přímo vepsáním daného textu. U dalších se zobrazí nabídka možností po kliknutí na dané políčko. V případě možnosti *Value* se otevře nové okno, kde se vyplňují kolonky *Value* a *Label*. *Value* definuje hodnotu v podobně čísla a *Label* její popisek. Po doplnění těchto kolonek se definice hodnoty uloží pomocí tlačítka *Add*. Poté je možné vytvořenou hodnotu editovat nebo ji smazat. V případě editace je potřeba vybrat příslušnou definici a změnit jí atribut *Value* nebo *Label*, poté kliknout na tlačítko *Change*. Pokud je potřeba smazání definice, provede se její volba a kliknutím na tlačítko *Remove* se odstraní.

Poslední možností je definování hodnot, které nejsou potřeba zahrnout do analýzy. Jedná se o políčka ve sloupci Missing. Po kliknutí na jedno z políček se zobrazí dialogové okno, kde lze zvolit možnosti *No missing values* (žádné hodnoty nejsou definovány), *Discrete missing values* (kde lze vyplnit tři možnosti, které nebudou zahrnuty do analýzy), *nebo Range plus one optional discrete missing value* (kde lze určit číselný rozsah možností hodnot, které nejsou relevantní pro analýzu).

3.1.3 Plnění matice daty

Existují různé způsoby, jak do matice dostat data. Rabušic, Soukup a Mareš (2019) ve své publikaci uvádí dva hlavní způsoby:

a) Výzkumníkem definovanou matici lze naplnit vlastními daty
b) Data do matice se mohou naimportovat ze souboru jiného typu, například z textového editoru, databáze, tabulkového procesoru, nebo z programu jako je například Excel.

   Lze však také použít i dříve nebo někým jiným vytvořenou matici dat, takzvaný systémový soubor.

3.2 Data zvolena pro prezentaci zvolených statistických metod

Pro simulaci měr asociací byla zvolena data z PISA 2018 Database. Tato databáze obsahuje set odpovědí jednotlivých studentů, učitelů a rodičů. Web www.oecd.org ke stažení nabízí dotazníky, soubory s daty ve formátu pro SAS nebo SPSS a také soubory, ve kterých lze nalézt vysvětlení jednotlivých proměnných. Pro zjednodušení zápisu do datové matice byly zakódovány.

Jednotlivé asociační míry budou aplikovány na datovém souboru určeném pro otevření v SPSS softwaru, který obsahuje odpovědi jednotlivců na dotazník určený pro studenty. Datová matice a jednotlivé proměnné jsou již vyplněny, včetně definovaných proměnných ve sloupci Values. Pro některé proměnné jsou zde také definovány Missing hodnoty.

3.3 Principy zjišťování závislostí proměnných

V této kapitole jsou uvedeny principy zjišťování závislostí proměnných, které budou využity pro simulaci na zvolených datech. U těchto možností je dále uvedeno, jak lze snadno získat hodnoty daných koeficientů pomocí procedur softwaru SPSS.

3.3.1 Kontingenční tabulka

„Kontingenční tabulka je základem pro testování nezávislosti a pro výpočet měr intenzity závislosti.“ (Řezanková, 2017)

Jak již bylo zmíněno v kapitole 2.3, kontingenční tabulka může pomocí četností pomoci posoudit závislost či nezávislost mezi kategoriálními proměnnými a s jejich rozlišením dle typu četností na tabulky s absolutní četností, nebo s relativní četností. Relativní četnosti, které jsou obvykle udávány v procentech, lze počítat třemi odlišnými způsoby. Paní Řezanková je uvádí ve své knize Analýza dat z dotazníkových šetření (2017):

1) Podíly počítané na základě celého rozsahu
2) Řádkové podíly
3) Sloupcové podíly

Ve všech třech případech se součet rovná hodnotě 1, respektive 100 pokud se hodnota vyjádří v procentech.

V IBM SPSS Statistics *„můžeme buď nechat zobrazit několik tabulek s různými typy četností, nebo zapsat několik hodnot do jednoho políčka“.* (Řezanková, 2017)

Rozsah souboru bude označen symbolem n, počet kategorií proměnné X jako R a počet kategorií proměnné Y jako S. Četnosti, které jsou zaznamenány v tabulce, jsou označeny jako n_{ij}, kde i = 1, 2, …, R a j = 1, 2, …, S. Pomocí tohoto značení lze vytvořit následující kontingenční tabulku absolutních četností. (Řezanková, 2017)

		Znak Y					Celkem
		1. kategorie	…	j-tá kategorie	…	S-tá kategorie	
Znak X	1. kategorie	n_{11}	…	n_{1j}	…	n_{1S}	n_{1+}
	…	…	…	…	…	…	…
	i-tá kategorie	n_{i1}	…	n_{ij}	…	n_{iS}	n_{i+}
	…	…	…	…	…	…	…
	R-tá kategorie	n_{R1}	…	n_{Rj}	…	n_{RS}	n_{R+}
Celkem		n_{+1}	…	n_{+j}	…	n_{+S}	n

Schéma 1: Značení pro kontingenční tabulku absolutních četností Autor: Řezanková (2017)

Mimo sdružených absolutních četností jsou v schématu číslo 1 uvedené také marginální četnosti značené n_{i+} a n_{+j}. Pro tyto četnosti platí $n_{i+} = \sum_{j=1}^{S} n_{ij}$ a $n_{+j} = \sum_{i=1}^{R} n_{ij}$. (Řezanková, 2017)

Podobnou kontingenční tabulku lze vytvořit pro relativní četnosti, které jsou označeny jako p_{ij}. (Řezanková, 2017)

		Znak Y					Celkem
		1. kategorie	…	j-tá kategorie	…	S-tá kategorie	
Znak X	1. kategorie	p_{11}	…	p_{1j}	…	p_{1S}	p_{1+}
	…	…	…	…	…	…	…
	i-tá kategorie	p_{i1}	…	p_{ij}	…	p_{iS}	p_{i+}
	…	…	…	…	…	…	…
	R-tá kategorie	p_{R1}	…	p_{Rj}	…	p_{RS}	p_{R+}
Celkem		p_{+1}	…	p_{+j}	…	p_{+S}	p

Schéma 2: Značení pro kontingenční tabulku relativních četností Autor: Řezanková (2017)

Pokud se jedná o podíly počítané na základě celého souboru, pak pro i = 1,2, …, R a j = 1,2, …, S platí, že

$$p_{ij} = \frac{n_{ij}}{n},\ p_{i+} = \sum_{j=1}^{S} p_{ij}\ ,\ p_{+j} = \sum_{i=1}^{R} p_{ij} \text{ a } \sum_{i=1}^{R} p_{i+} = \sum_{j=1}^{S} p_{+j} = 1\ .$$

Hodnoty kontingenční tabulky absolutních nebo relativních četností, pak lze graficky zobrazit jako *sloupcový graf*, a to buď jako shlukový graf, nebo jako graf kumulativní. (Řezanková, 2017)

Spuštěním procedury CROSSTABS v systému IBM SPSS Statistice, lze a charakterizovat vztah dvou proměnných (označeny A, B) pomocí kontingenční tabulky. Před spuštěním

této procedury je potřeba do sekce *Row(s)* zvolit proměnnou A, a do části *Column(s)* proměnnou B. (Řezanková, 2017)

Graficky lze hodnoty z kontingenční tabulky zobrazit jako sloupcový graf. Dvojice kategorií pak mohou mít 2 podoby, a to graf shlukový nebo graf kumulativní. Pro zobrazení sloupcového shlukového grafu s absolutními četnostmi, v již zmiňovaném systému, v proceduře CROSSTABS se zvolí možnost *Display clustered bar charts*. (Řezanková, 2017)

3.3.2 Chí-kvadrát test o nezávislosti

Při Chí-kvadrát testu se zvažuje, jestliže jsou dva znaky nezávislé, pak je rozdělení četností v kontingenční tabulce úměrné řádkovým a sloupcovým marginálním četnostem. Takové četnosti se nazývají očekávanými, obecně je budeme označovat m_{ij}. (Řezanková, 2017) Vzorec pro očekávané četnosti je:

$$m_{ij} = \frac{n_{i+} n_{+j}}{n}$$

Kde n_{i+} a n_{j+} jsou řádkové, respektive sloupcové marginální četnosti s n označuje celkový počet.

Po výpočtu všech očekávaných četností pro všechna pole kontingenční tabulky se musí provést další výpočetní operace. Následuje tedy výpočet rozdílu mezi empirickou četností, která je označena jako n_{ij}, a očekávanou četností pro všechna pole tabulky. Tento rozdíl se pak umocní na druhou, podělí hodnotou očekávané četnosti a jednotlivé výsledky se následně sečtou. Tím lze získat hodnotu testové statistiky chí-kvadrát. Zápis této statistiky pak je:

$$\chi^2_P = \sum_{i=1}^{R} \sum_{j=1}^{S} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}.$$

Výše zmíněné operace provede SPSS. V proceduře CROSSTABS je potřeba zvolit všechny požadované možnosti tak, že v dialogovém okně Cells se v boxu Counts také zaškrtne políčko Expected a v boxu Residuals políčko Unstandardized. Pro vyhodnocení testu chí-kvadrát je potřeba v proceduře v sekci *Statistics* vybrat Chi-square. (Rabušic, Soukup a Mareš, 2015)

Řádek Residuals udává numerický rozdíl mezi empirickou (Count) a očekávanou (Expected Count) četností. Z této hodnoty lze vyčíst především to, zda je napozorovaná četnost vyšší nebo nižší, než by bylo očekáváno, kdyby platila nulová hypotéza nezávislosti. (Rabušic, Soukup a Mareš, 2015)

Jak bylo zmíněno v kapitole 2.6.1 pro testování dvou nominálních neznámých se mimo jiné také využívá věrohodností poměr, který se v případě multinomického rozdělení spočte jako

$$G^2 = 2\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij}\ln\frac{n_{ij}}{m_{ij}}.$$

Obě statistiky mají asymptotické stupně volnosti (R – 1)(S – 1). (Řezanková 2007)

Pro zamítnutí nulové hypotézy o nezávislosti je potřeba znát kritickou hodnotu. Ta se určuje stupni volnosti a hladinou významnosti α (například 0,05). Hodnota testového kritéria χ_P^2 se pak porovnává s kvartilem

$$\chi_{1-\alpha}^2[(R-1)(S-1)].$$

Je-li hodnota testového kritéria vetší nebo rovna kritické hodnotě, pak se nulová hypotéza o nezávislosti zamítá. V opačném případě nulovou hypotézu zamítnout nelze.

V této kapitole jsou dále uvedeny vzorce koeficientů symetrických měr, které vycházejí ze statistiky chí-kvadrát: (Řezanková, 2007)

a) **Pearsonův kontingenční koeficient**, které bude označeno písmenem C_p, se počítá dle vztahu

$$C_p = \sqrt{\frac{\chi_P^2}{\chi_P^2+n}}.$$

Koeficient nabývá hodnot z intervalu $\langle 0; \sqrt{(q-1)/q}\rangle$, kde q = min{R, S}. V případě nezávislosti nabývá hodnoty 0 a čím více se hodnota (pozitivně) vzdaluje od této hodnoty se stejným n, R a S, tím je závislost silnější.

b) **Koeficient φ (fí)** je další mírou intenzity závislosti pro dvě nominální proměnné, kde platí vztah

$$\varphi = \sqrt{\frac{\chi_P^2}{n}}.$$

c) **Cramérovo V** se počítá pomocí vzorce

$$\mathrm{V} = \sqrt{\frac{\chi_P^2}{n(q-1)}},$$

„kde q = min{R, S}. Ve jmenovateli je tedy maximální hodnota, které může dosáhnout Pearsonova statistika chí-kvadrát. To znamená, že tento koeficient nabývá hodnot z intervalu od 0 do 1. Pro tabulku, kdy je alespoň jedna proměnná dichotomická, tedy počet odpovídajících řádků a/nebo sloupců je 2, je získán koeficient φ.

d) **Čuprovův kontingenční koeficient** se uvádí ve tvaru

$$C_T = \sqrt{\frac{\chi_P^2/n}{\sqrt{(R-1)(S-1)}}}.$$

Pokud se jedná o čtvercovou tabulku, tedy počet jejích řádků je stejný jako počet sloupců, pak platí, že

$$\mathrm{q} - 1 = \sqrt{(R-1)(S-1)}$$

a hodnoty Cramérova V a Čuprovova kontingenčního koeficientu jsou shodné.

V části *Statistics* procedury *CROSSTABS* v systému SPSS je potřeba zaškrtnout možnosti *Contingency coefficient* a *Phi and Cramér´s V.* Součástí výstupu pak budou minimální hladiny významnosti, od kterých se zamítá nulová hypotéza o nezávislosti dvou sledovaných veličin. (Řezanková, 2007)

Pro asymetrické míry uvádí Řezanková (2007) následující koeficienty:

a) **Goodmanova-Kruskalova λ**, pro kterou lze odvodit vzorec pro asymetrickou variantu

$$\lambda_{Y|X} = \frac{1 - p_{+M} - (1 - \sum_{i=1}^{R}\{p_{iMo})\}}{1 - p_{+Mo}} = \frac{\sum_{i=1}^{R} p_{iMo} - p_{+Mo}}{1 - p_{+Mo}} = \frac{\sum_{i=1}^{R} n_{iMo} - n_{+Mo}}{n - n_{+M}},$$

kde $n_{+Mo} = max_j(n_{+j})$ a $n_{iMo} = max_j(n_{ij})$.

b) **Goodmanova-Kruskalova τ (tau)**, jehož obecný vzorec je psán ve tvaru

$$\tau_{Y|X} = \frac{nomvar(Y) - \sum_{i=1}^{R} p_{i+} nomvar(Y|x_i)}{nomvar(Y)}.$$

Po rozepsání a úpravách obecného tvaru je získán vzorec

$$\tau_{Y|X} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{S}\frac{\left(p_{ij} - p_{i+}p_{+j}\right)^2}{p_{i+}}}{1 - \sum_{j=1}^{S} p_{+j}^2}.$$

Pomocí absolutních četností jej lze také psát

$$\tau_{Y|X} = \frac{n\sum_{i=1}^{R}\sum_{j=1}^{S}\frac{\left(n_{ij} - m_{ij}\right)^2}{n_{i+}}}{n^2 - \sum_{j=1}^{S} n_{+j}^2}$$

c) **Koeficient nejistoty,** pro který se uvádí vzorec v podobě:

$$U_{Y|X} = \frac{-\sum_{i=1}^{R} p_{i+}\ln p_{i+} - \sum_{j=1}^{S} p_{+j}\ln p_{+j} + \sum_{i=1}^{R}\sum_{j=1}^{S} p_{ij}\ln p_{ij}}{-\sum_{j=1}^{S} p_{+j}\ln p_{+j}} =$$

$$= \frac{H(X) + H(Y) - H(XY)}{H(Y)}.$$

V případě výpočtu asymetrických koeficientů v systému SPSS, se opět vybere procedura CROSSTABS a v části *Statistics* je potřeba zaškrtnout možnost *Lambda*, *Uncertainty coefficient*. *„Součástí výstupu jsou opět minimální hladiny významností, od kterých zamítáme nulovou hypotézu o nezávislosti dvou sledovaných proměnných. (sloupec Approx. Sig.)“* (Řezanková, 2007)

Pokud se jedná o čtvercovou tabulku, kde sledované znaky nabývají stejných kategorií lze využít koeficient Cohenovo **κ** (kappa), který se počítá podle vzorce

$$\kappa = \frac{\sum_{i=1}^{R} n_{ii} - \sum_{i=1}^{R} m_{ii}}{n - \sum_{i=1}^{R} m_{ii}}.$$

Pro tento koeficient v proceduře CROSSTABS v části *Statistics* je potřeba zvolit možnost *Kappa*. (Řezanková 2007)

3.3.3 Korelace

O korelaci se jedná v případě, kdy je počítána závislost mezi dvěma ordinálními neznámými. V předešlých vzorcích se tedy jednalo o statickou závislost, která je označována jako kontingence. (Řezanková, 2007)

Následující koeficienty s používají pro výpočet míry asociace mezi dvěma ordinálními znaky, které definuje Řezanková (2007) ve své publikaci:

a) **Spearmanův koeficient pořadové korelace** vychází z toho, *„že každé hodnotě proměnné X je přiřazeno pořadí a_l tak, že $\sum_{l=1}^{n} a_l = n\frac{n+1}{2}$, a každé hodnotě proměnné Y je přiřazeno pořadí b_l tak, že $\sum_{l=1}^{n} b_l = n\frac{n+1}{2}$“.* (Řezanková 2007)

„Při výpočtu z četností kontingenční tabulky je postupováno následujícím způsobem

a) Kategoriím proměnné X se přiřadí postupně pomocné skóry a_i:

$$a_1 = \frac{n_{1+}+1}{2},\ a_i = \sum_{l=1}^{i-1} n_{l+} + \frac{n_{i+}+1}{2} \text{ pro } \leq i \leq R,$$

kategoriím proměnné Y se přiřadí pomocné skóry b_j:

$$b_1 = \frac{n_{+1}+1}{2},\ b_j = \sum_{l=1}^{j-1} n_{+l} + \frac{n_{+j}+1}{2} \text{ pro } \leq j \leq S,$$

b) spočítají se hodnoty

$$D^2 = \sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij}\left(a_i - b_j\right)^2,$$

$$T_X = \frac{1}{12}\left(n^3 - \sum_{i=1}^{R} n_{i+}^3\right),$$

$$T_Y = \frac{1}{12}\left(n^3 - \sum_{j=1}^{R} n_{+j}^3\right),$$

které jsou dosazeny do vzorce

$$r_s = \frac{T_X + T_Y - D^2}{2\sqrt{T_X T_Y}}.$$

Jestliže $\sum_{i=1}^{R} n_{i+}^3 = \sum_{j=1}^{R} n_{+j}^3$, pak $T_X = T_Y$. Je tak získán jednodušší vzorec, a to

$$r_s = \frac{2T_X - D^2}{2\sqrt{T_X^2}} = 1 - \frac{D^2}{2T_X}.$$

Platí-li navíc $\sum_{i=1}^{R} n_{i+}^3 = \sum_{j=1}^{R} n_{+j}^3 =$ n, lze napsat známější vzorec tohoto koeficientu, a to“

$$r_s = 1 - \frac{D^2}{2\frac{1}{12}(n^3 - n)} = 1 - \frac{6D^2}{n(n^2-1)}.$$

V případě, kdy jde o prosté ordinální proměnné X a Y, které vyjadřují jednoznačné pořadí (žádná hodnota se neopakuje), pak není potřeba skóry a_i a b_j počítat a pro D^2 se použije následující vzorec (Řezanková, 2007):

$$D^2 = \sum_{l=1}^{n} (x_l - y_l)^2.$$

Koeficient může nabývat hodnot z intervalu $\langle -1; 1 \rangle$. Zde 1 znamená pozitivní korelaci, takzvanou přímou závislost, a –1 pak negativní korelaci, nebo jinak také nepřímou závislost. V případě že hodnota koeficientu vyjde 0, jedná se o lineární nezávislost. Nulovost tohoto koeficientu se provádí za pomocí již zmíněného testu statistiky

$$\mathrm{t} = r_s \sqrt{\frac{n-2}{1-r_s^2}},$$

„*která má za předpokladu platnosti nulové hypotézy Studentovo t rozdělení s(n-2) stupni volnosti*“. (Řezanková 2007)

c) **Goodmanova-Kruskalova γ** (gama), která je počítána pomocí vzorce

$$\gamma = \frac{C - D}{C + D},$$

a hodnota tohoto koeficientu nabývá z intervalu $\langle 0; 1 \rangle$.

d) **Kendallovo τ_b** (tau-b) dáno vztahem

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X)(C + D + T_Y)}},$$

může nabývat hodnot z intervalu $\langle -1; 1 \rangle$.

e) **Kendallovo τ_c** (tau-c), jehož výpočet je založen na C, D a dále na $q = \min\{r, s\}$, se vyjadřuje pomocí vztahu

$$\tau_c = \frac{2q(C - D)}{n^2(q - 1)}$$

Pokud je tabulka typu *R x S, pak platí* $\tau_c \in \langle -1; 1 \rangle$.

f) **Somersovo d** je koeficient pro kterého jsou vyjádřeny následující varianty výpočtu:

a. Asymetrická míra, vyjádřena vztahem

$$d_{Y|X} = \frac{C - D}{C + D + T_Y}$$

b. Symetrická míra je vyjádřena jako harmonický průměr obou asymetrických variant

$$d_{sym} = \frac{2}{\frac{C + D + T_Y}{C - D} - \frac{C + D + T_X}{C - D}} = \frac{2(C - D)}{2(C + D) + T_X + T_Y}.$$

Matematické vyjádření pro C, D, T_Y a T_X je zmíněno v kapitole 2.6.2.1.

Pro získání výstupu v SPSS pro předchozí koeficienty je nutné vybrat opět proceduru CROSSTABS a v části *Statistics* zaškrtnout možnosti *Correlations, Gama, Somers' d, Kendall's tau-b* a *Kendall's tau-c.* Výsledek zahrnuje pouze symetrické míry a asymetrické Somersovo *d*. Výstup opět obsahuje minimální hladiny významnosti, *„od nichž zamítáme nulové hypotézy o nulovosti koeficientů".* (Řezanková 2007)

Dalším korelačním koeficientem je **Pearsonův korelační koeficient**, který lze použít, pokud je potřeba znát vzájemná závislost mezi dvěma kvantitativními proměnnými. Koeficient je vyjádřen ve tvaru

$$\mathrm{r} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{S} x_i x_j p_{ij} - \sum_{i=1}^{R} x_i p_{i+} \times \sum_{j=1}^{S} x_j p_{+j}}{\sqrt{\left(\sum_{i=1}^{R} x_i^2 p_{i+} - \left(\sum_{i=1}^{R} x_i p_{i+}\right)^2\right) \cdot \left(\sum_{j=1}^{S} x_j^2 p_{+j} - \left(\sum_{j=1}^{S} x_j p_{+j}\right)^2\right)}},$$

lze také vyjádřit pomocí absolutních četností ve tvaru

$$\mathrm{r} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij} x_i y_j - m \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^{R} n_{i+} x_i^2 - n\bar{x}^2\right) \cdot \left(\sum_{j=1}^{S} n_{+j} x_j^2 - n\bar{y}^2\right)}},$$

kde $\bar{x} = \frac{1}{n}\sum_{i=1}^{R} n_{i+} x_i$ a $\bar{y} = \frac{1}{n}\sum_{j=1}^{R} n_{+j} y_j$.

Výsledky koeficientu mohou nabývat hodnot z intervalu $\langle -1; 1 \rangle$. (Řezanková 2007)

V SPSS lze korelační koeficient získat jednak v rámci kontingenčních tabulek, a jednak ve speciální proceduře pro výpočet korelačních koeficientů. Pro případ výpočtu v rámci kontingenční tabulky se opět vybere procedura CROSSTABS a v části *Statistics* se zaškrtne možnost *Correlation*. V případě výpočtu za použití speciální procedury je nutné se přesměrovat přes *Analyze* a v části *Correlate* vybrat proceduru *Bivariate*. Zde

je potřeba zaškrtnout položku *Pearson*. Výsledkem je korelační matice, kde lze vyčíst hodnota korelačního koeficientu. Ta se nachází vždy v políčku matice jako první. (Řezanková 2007)

3.3.4 Kruskalův-Wallisův test

Výpočet testového kritéria je založen na pořadových číslech, která jsou přiřazena hodnotám v souboru, vzniklým spojením všech výběrů. Při výpočtu z četností kontingenční tabulky se postupuje tak, že kategoriím proměnné Y se postupně přiřadí pomocné skóry w_j dle vztahů

$$w_1 = \frac{n_{+1}+1}{2}, w_j = \sum_{l=1}^{j-1} n_{+l} + \frac{n_{+j}+1}{2}, \text{pro } 2 \le j \le S,$$

pro každou kategorii proměnné X se vypočítá průměrné pořadí

$$\bar{R}_i = \sum_{j=1}^{S} w_j p_{ij}, i = 1,2,\dots,R$$

a součet pořadí

$$R_i = \sum_{j=1}^{S} w_j n_{ij} = \bar{R}_i n_{i+}, i = 1,2,\dots,R,$$

opravu na spojitost

$$H_S = 1 - \sum_{j=1}^{S} \frac{n_{+j}^3 - n_{+j}}{n^3 - n} = 1 - \frac{\sum_{j=1}^{S} n_{+j}(n_{+j}^2 - 1)}{n \cdot (n^2 - 1)} = \frac{n^3 - \sum_{j=1}^{S} n_{+j}^3}{n \cdot (n^2 - 1)}$$

a nakonec se spočítá hodnota Kruskalovy-Wallisovy statistiky

$$KW = \frac{\frac{12}{n \cdot (n+1)} \sum_{i=1}^{R} \frac{R_i^2}{n_{i+}} - 3 \cdot (n+1)}{H_S}.$$

Tato veličina má při platnosti nulové hypotézy o nezávislosti přibližně rozdělené chí-kvadrát s (R-1) stupni volnosti. (Řezanková 2007)

V SPSS je potřeba si pro analýzu pro ordinální vysvětlovanou proměnnou v proceduře *K Independent Samples* zaškrtnout položku *Kuskal-Wallis H*. Tato procedura je pod

výběrem *Analyze -> Nonparamterics Tests*. Výstupem jsou dvě tabulky. V první jsou údaje o pořadí pro jednotlivé kategorie vysvětlované proměnné a ve druhé je hodnota testového kritéria, počet stupňů volnosti a minimální hladina významnosti. Tato hladina významnosti je v tabulce na řádku pojmenovaném *Asymp. Sig*. Výstup však neobsahuje hodnotu H_S, který je potřeba pro použití vzorce. (Řezanková 2007)

3.3.5 Tabulka s kvantitativní vysvětlovanou proměnnou

Intenzitu statistické závislosti vysvětlované kvantitativní proměnné na proměnné nominální se posuzuje buď poměrem determinace, nebo podle jeho odmocniny. (Řezanková 2007)

Poměr determinace se počítá podle vzorce

$$I^2_{Y|X} = \frac{s^2(Y) - \sum_{i=1}^{R} p_{i+} s^2(Y|x_i)}{s^2(Y)},$$

kdy po vyjádření a úpravách ve vzorci je získán tvar (celý postup pro vyjádření koeficientu lze dohledat v publikaci paní Řezankové 2007, str. 107)

$$I^2_{Y|X} = \frac{\sum_{i=1}^{R} \frac{1}{p_{i+}} \left(\sum_{j=1}^{S} p_{ij} y_j\right)^2 - \left(\sum_{j=1}^{S} p_{+j} y_j\right)^2}{\sum_{j=1}^{S} p_{+j} y_j^2 - \left(\sum_{j=1}^{S} p_{+j} y_j\right)^2}.$$

Odmocnina vyjádřené determinace se označuje také jako koeficient η (éta). Vzorec pro daný koeficient lze vyjádřit také pomocí absolutních četností následujícím způsobem

$$\eta_{Y|X} = \sqrt{I^2_{Y|X}} = \sqrt{\frac{\sum_{i=1}^{R} \frac{1}{n_{i+}} \left(\sum_{j=1}^{S} n_{ij} y_j\right)^2 - \frac{1}{n} \left(\sum_{j=1}^{S} n_{+j} y_j\right)^2}{\sum_{j=1}^{S} n_{+j} y_j^2 - \frac{1}{n} \left(\sum_{j=1}^{S} n_{+j} y_j\right)^2}}.$$

Pro analýzu v SPSS je potřeba spustit proceduru CROSSTABS v jejíž části *Statistics* je zaškrtnuta možnost *Eta* a typ měr *Nominal by Interval*. (Řezanková 2007)

3.3.6 Kontingenční tabulka pro dvě dichotomické proměnné

Kontingenční tabulka pro dvě dichotomické proměnné se také jinak označuje jako čtyřpolní, díky její podobě čtyř políček, kterou vystihuje *Schéma 3,* tedy forma zjednodušené kontingenční tabulky, jejíž podoba je již uvedena v kapitole 3.8.9 (*Schéma 1 a Schéma 2*).

Znaky X	Znaky Y		Celkem
	0	1	
0	p_{11}	p_{12}	p_{1+}
1	p_{21}	p_{22}	p_{2+}
Celkem	p_{+1}	p_{+2}	1

Schéma 3: Čtyřpolní tabulka relativních četností *Autor: Řezanková (2007)*

Pro chí-kvadrát statistiku lze upravit vzorec do následující podoby:

$$\chi_P^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1+}p_{2+}p_{+1}p_{+2}}.$$

Pro případ absolutních četností lze psát

$$\chi_P^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}.$$

Pro testování nezávislosti dvou dichotomických proměnných lze také využít statistiky chí-kvadrát korigované na spojitost, známou jako **Yatesova korekce,** a využít lze dvou následujících vzorců

$$\chi_C^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left[\max\left(0, \left|n_{ij} - m_{ij}\right| - 0{,}5\right]\right]^2}{m_{ij}},$$

nebo

$$\chi_C^2 = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - 0{,}5n)^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \text{ pro } |n_{11}n_{22} - n_{12}n_{21}| > 0{,}5n,$$

neplatí-li tato podmínka pak $\chi_C^2 = 0$. (Řezanková 2007)

Dále Řezanková (2007) pro testování nezávislosti uvádí **věrohodnostní poměr**, který lze počítat pomocí vzorce:

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \ln\left(\frac{n_{ij}}{m_{ij}}\right).$$

Výstup výše uvedených statistiky pro čtyřpolní kontingenční tabulku lze v SPSS získat spuštěním procedury CROSSTABS, která má v části *Statistics* je zvolenu možnost C*hi-square*. Ve výstupu lze zjistit výsledek Pearsonovy statistiky chí-kvadrát (*Pearson Chi-Square)*, statistiku korigovanou na spojitost (*Continuity Correction)* a věrohodnostní poměr *(Likelihood Ratio)*.

Pro dvě dichotomické proměnné lze také odvodit zjednodušený tvar pro **Pearsonův korelační koeficient** a *Schéma 3* lze přepsat jako „jednorozměrnou“ tabulku následujícím způsobem:

Znak X	Znak Y	Četnosti
0	0	p_{11}
0	1	p_{12}
1	0	p_{21}
1	1	p_{22}

Schéma 4: Přepis Schéma 3 na vstupní matici vážených kombinací hodnot *Autor: Řezanková (2007)*

Ve schématu jsou uvedeny jednotlivé kombinace kategorií a k nim dané četnosti.

Pro výpočet Pearsonova korelačního koeficientu dvou dichotomických proměnných je potřeba znát kovarianci a směrodatné odchylky, *„jejichž výpočet je založen na odchylkách aritmetického průměru*“. (Řezanková 2007)

Kovarianci lze počítat dle vztahu $s_{XY} = \overline{xy} = \bar{x} \cdot \bar{y}$, kde $\bar{x}$ je aritmetický průměr proměnné X a $\bar{y}$ je pak aritmetický průměr proměnné Y. Vztahy pro tyto aritmetické průměry se počítají dle vztahu $\bar{x} = \sum x_i \cdot p_i$, pak lze psát

$$\bar{x} = 0 \cdot p_{21} + 0 \cdot p_{12} + 1 \cdot p_{21} + 1 \cdot p_{22} = p_{21} + p_{22} = p_{2+},$$
$$\bar{y} = 0 \cdot p_{21} + 1 \cdot p_{12} + 0 \cdot p_{21} + 1 \cdot p_{22} = p_{12} + p_{22} = p_{+2}.$$

Po dosazení do vztahu kovariance a úpravách je získán vzorec

$$s_{XY} = p_{11} \cdot p_{22} - p_{12} \cdot p_{21}.$$

Rozptyl může vycházet ze vztahu $s^2 = \sum x_i^2 p_i - \bar{x}^2$. Jelikož se počítá s dichotomickými znaky, pro které platí, že $\sum x_i^2 p_i = \sum x_i p_i$, pak lze rozptyl zapsat ve tvaru $s_X^2 = \bar{x} - \bar{x}^2$, nebo též jako $s^2 = \bar{x}(1 - \bar{x})$. Pro neznámou X je pak výsledkem

$$s_X^2 = p_{1+} p_{2+},$$

a pro proměnnou Y

$$s_Y^2 = p_{+1} p_{+2}.$$

Korelační koeficient je pak podílem kovariance a součinu směrodatných odchylek. Vzorec této míry je tedy tvaru

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

„Pro čtyřpolní tabulku platí, že $\chi_P^2 = nr^2$. To znamená, že míry asociací jsou v tomto případě funkcí korelačního koeficientu". (Řezanková 2007) Jak již bylo zmíněno v kapitole 2.6.6 vzorec pro *Cramerovo V* je v tomto speciálním případě shodný se vzorcem koeficientu φ, proto lze také odvodit, že

$$|r| = \sqrt{\frac{\chi_P^2}{n}} = \varphi = V.$$

Pro výpočet míry asociace se také může využít Pearsonův kontingenční koeficient (vzorec pro tento koeficient je zmíněn v kapitole 3.8.10).

K získání výstupu v SPSS je pro tyto koeficienty potřeba opět spustit proceduru CROSSTABS, kdy jsou v části *Statistics* zvoleny možnosti *Contingency coeficient, Phi and Cramér´s V* a *Correlations*. (Řezanková 2007)

Pro analýzu dvou dichotomických proměnných lze dále požít následující koeficienty (Řezanková, 2007):

a) **Yuleovo Q**, dáno vzorcem

$$Q = \gamma = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$

b) **Yuleův koeficient vazby**, vyjádřen

$$Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}}.$$

c) **Kendallovo τ_b** (viz kapitola 3.8.11), stejně jako Pearsonův korelační koeficient, lze tuto míru zapsat jednodušeji. Po úpravách (viz Řezanková 2007, str. 125) je získán vztah

$$\tau_b = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{11}n_{22}n_{12}n_{21}}} = r.$$

d) **Somersovo d** (viz kapitola 3.8.11), upravený vzorec pro dvě dichotomické neznámé lze po úpravách (viz Řezanková 2007, str. 125), vyjádřit jako

$$d = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{1+}n_{2+} + n_{+1}n_{+2}}.$$

e) **Hamanův koeficient,** je mírou souhlasu, pro kterou platí

$$\frac{(n_{11}+n_{22}) - (n_{12}+n_{21})}{n}.$$

f) **Poměr šancí**, matematicky se píše jako

$$\Psi = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{RR_1}{RR_2},$$

kde RR je označení koeficientu relativního rizika, které lze vypočítat dle následujícího vzorce

$$RR_1 = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}},$$

$$RR_2 = \frac{n_{12}/n_{1+}}{n_{22}/n_{2+}}.$$

g) **Procentní rozdíl** se na základě absolutních četností vypočítá podle vztahu

$$PR_{Y|X} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}.$$

3.3.7 Fisherův exaktní test

V případě Fisherova exaktního testu se testuje hypotéza

$$H_0: \pi_{11} = p_{1+}p_{+1}.$$

Řezanková (2007) uvádí, že pravděpodobnost četnosti buňky n_{11}, která nabude hodnoty *t*, je dána vztahem

$$P(n_{11} = t) = \frac{\binom{n_{+1}}{t}\binom{n_{+2}}{n_{1+}-t}}{\binom{n}{n_{1+}}}.$$

Pravděpodobnost pro každou variantu četností lze tedy spočítat pomocí následujícího vztahu

$$p = \frac{n_{1+}!\, n_{2+}!\, n_{+1}!\, n_{+2}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}.$$

Dle vztahu

$$\alpha' = \sum_A p,$$

se určí minimální hladina významnosti, od které se zamítá nulová hypotéza (H_0), pro pravostrannou a oboustrannou alternativní hypotézu. (Řezanková, 2007)

A, uvedené ve výpočtu minimální hladiny významností je

a) *„pro oboustranný test, množina tabulek, kde p je menší nebo rovno pravděpodobnosti zjištěných četností*
b) *pro jednostranný test, množina tabulek se stejnými hodnotami p, jako pro oboustranný test a zároveň n_{11} je buď rovno zjištěné četnosti, nebo ve stejné relaci jako zjištěná četnost k teoretické ".* (Řezanková, 2007)

Pro levostrannou alternativní hypotézu je hladina významnosti počítána dle vztahu

$$\alpha' = 1 - \sum_A p,$$

„kde A je množina tabulek shodná s množinou pro výše uvedený jednostranný test s výjimkou tabulky zjištěných četností". (Řezanková, 2007)

Pro získání výsledku exaktního testu se v části *Statistics* CROSSTABS procedury zvolí možnost *Chi-Square*. Ve výstupu jsou získány pouze výsledky dvou variant testů, a to pro oboustrannou alternativní hypotézu a pravostrannou alternativní hypotézu. (Řezanková, 2007)

3.3.8 McNemarův test

McNemarovým testem se testuje nulová hypotéza o shodě četností v políčkách na vedlejší diagonále čtyřpolní tabulky dvou dichotomických proměnných. Hypotéza se uvádí ve tvaru

$$H_0: \pi_{12} = \pi_{21}, H_1: \pi_{12} \neq \pi_{21}.$$

Hypotéze je odvozena ze vztahu

$$\frac{n_{1+}}{n} = \frac{n_{+1}}{n},$$

neboli

$$\frac{n_{11} + n_{12}}{n} = \frac{n_{11} + n_{21}}{n},$$

čili

$$\frac{n_{12}}{n} = \frac{n_{21}}{n}.$$

Hladina významnosti pro exaktní test je dána vztahem

$$\alpha' = 2 \sum_{i=0}^{\min\{n_{12},n_{21}\}} \binom{n_{12}+n_{21}}{i} (0{,}5)^{n_{12}+n_{21}}.$$

Polovina hodnoty α' vyjadřuje pravděpodobnost, že náhodná veličina s binomickým rozdělením nabude hodnoty min $\{n_{12}, n_{21}\}$ nebo menší. (Řezanková, 2007)

Pokud je $n_{12} + n_{21} > 25$, pak „*lze použít aproximaci binomického rozdělení rozdělením chí-kvadrát*". (Řezanková, 2007) Testové kritérium s opravou na spojitost je dáno vztahem

$$Q_M = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}.$$

Hodnota Q_M se porovnává s kvantilem z chí-kvadrát rozdělení s jedním stupněm volnosti. (Řezanková, 2007)

Pro získání výsledku v SPSS v proceduře CROSSTABS v části *Statistics* je potřeba zvolit možnost *McNemar*. SPSS dále nabízí možnost použít proceduru v rámci neparametrických testů. Provedením voleb *Analyze*, *Nonparametric Test*, *2 Related Samples* a v části *Test Type* zvolením možnosti *McNemar*. (Řezanková, 2007)

3.3.9 Trojrozměrná kontingenční tabulka

Pro výpočet nezávislosti dvou dichotomických znaků podmíněných dalším kategoriální znakem jsou uvedeny dvě statistiky, kde L je označení pro soustavu čtyřpolních tabulek, a n_l je celkový počet objektů zařazených do l-té tabulky (l=1,2,…L)

a) Cochranova statistika je dána vztahem

$$Q_c = \frac{\left(\sum_{l=1}^{L} n_{l11} - \sum_{l=1}^{L} m_{l11}\right)^2}{\sum_{l=1}^{L} v_{l11}}$$

b) Matelova-Haenszelova statistika, pro kterou se používá vzorec

$$Q_{MH} = \frac{\left(\sum_{l=1}^{L} n_{l1} \ - \sum_{l=1}^{L} m_{l11} + 0{,}5\right)^2}{\sum_{l=1}^{L} v'_{l1}}$$

kde n_{lij} označuje sdružené četnosti, n_{li+} řádkové marginální četnosti a n_{l+j} sloupcové marginální četnosti. „*Při platnosti nulové hypotézy o nezávislosti dichotomických*

proměnných je očekávaní četnosti v poličku v l-té tabulce, i-tém řádku a j-tém sloupci dána vztahem

$$m_{lij} = \frac{n_{li}\ n_{l+j}}{n_l}$$

a rozptyl této četnosti vztahem

$$v_{lij} = \frac{n_{l1+}n_{l2+}n_{l+1}n_{l+2}}{n_l^3},$$

resp. při použití analogie pro výpočet výběrového rozptylu jedné proměnné vztahem

$$v_{lij} = \frac{n_{l1+}n_{l2+}n_{l+1}n_{l+2}}{n_l^2(n_l - 1)}.$$ "(Řezanková, 2007)

Obě statistiky mají asymptoticky chí-kvadrát rozdělení s jedním stupněm volnost. Rovnají-li se součty zjištěných a teoretických četností v prvním poličku tabulky, pak se používá Cochranova statistika.

Mantelův-Haenszelův odhad společného poměru šancí pro L čtyřpolních tabulek pro stanovení míry asociace je dán vztahem

$$\Psi_{MH} = \frac{\sum_{l=1}^{L} \frac{n_{l1}\ n_{l11}}{n_l}}{\sum_{l=1}^{L} \frac{n_{l12}n_{l21}}{n_l}}.$$

V SPSS spuštěním procedury CROSSTABS, která má v části *Statistics* nastavenou možnost *Cochran's and Matel-Haenszel statistics* jsou ve výstupu získány 3 tabulky. Ty popisuje Řezanková (2007)

1) ***Test of Conditional Independence*** obsahuje výsledky testů podmíněné nezávislosti, tedy statistiky *Cochrans* a *Matel-Haenszel*. V obou případech se vypisuje hodnota statistiky, počet stupňů volnosti a minimální hladina významnosti, od které se zamítá nulová hypotéza.
2) ***Test of Homogenity of the Odds Ratio***, který se vztahuje k testům homogenity poměru šancí. Obsahuje výsledky pro Breslowův-Dayovův test a Taroneův test.
3) ***Mantel-Haenszel Common Odds Ratio Estimate*** je tabulka jejíž výsledky se vztahují k odhadu společného poměru šancí. Jsou zde uvedeny výsledky bodového odhadu poměru šancí, jeho logaritmus a směrodatnou chybu odhadu zlogaritmované hodnoty. Dále lze z tabulky vyčíst hladina významnosti pro

zamítnutí hypotézy o nezávislosti a výsledky pro oboustranný interval spolehlivosti, a to jak pro společný poměr šancí, tak pro logaritmus této hodnoty pomocí dolních a horních mezí.

3.4 Interpretace výsledků měr asociací

Pro interpretaci výsledků měr asociací, tedy síly závislosti, je využita tabulka 1 z kapitoly 2.4.1, kterou uvádí ve svém článku Khamis (2008) a podle interpretace souvislostí hodnot korelace, kterou uvádí ve své publikaci Mareš, Rabušic a Soukup (2015) (viz tabulka 4).

Tabulka 3 je upravenou verzí tabulky od Khamis (2008). Lze zde pozorovat jednotlivé interpretace měr lineárního vztahu na daných intervalech. Uvedeny jsou pouze pozitivní vztahy, tedy přímé závislosti. Stejná interpretace se dá uvést pro hodnoty záporné, pro které lze konstatovat vztah negativní a hovoří se o nepřímé závislosti.

r	**Interpretace měr lineárního vztahu**
0,80 – 0,99	Silně pozitivní
0,50 – 0,79	Mírně pozitivní
0,20 – 0,49	Slabě pozitivní
0,00 – 0,19	Žádný vztah

Tabulka 4: Zjednodušená interpretace měr lineárního vztahu od Khamis (2008)

Hodnota korelace	**Interpretace souvislosti**
0,01 – 0,09	Triviální, žádná
0,10 – 0,29	Nízká až střední
0,30 – 0,49	Střední až podstatná
0,50 – 0,69	Podstatná až velmi silná
0,70 – 0,89	Velmi silná
0,90 – 0,99	Téměř perfektní

Tabulka 5: Interpretace souvislosti Mareš, Rabušic a Soukup (2015)

4 Vlastní popis a výsledky

Tato kapitola se věnuje aplikaci výše zmíněných metod měr asociací a testování nulové hypotézy o nezávislosti za použití SPSS softwaru.

Veškeré typy úloh jsou rozčleněny podle zkoumaných typů znaků, konkrétně podle:

a) Dvou nominálních proměnných, kde je řešena Chí-kvadrát statistika a všechny důležité koeficienty. Jednak takové, které existují na základě této statistiky, ale také Goodman-Kruskal λ, τ (tau), koeficient nejistoty, nebo Cohenovo κ. Cohenovo κ lze použít pouze v případě, že možnosti zkoumaných dvou nominálních znaků vytvoří kontingenční tabulku n x n.
b) Dvou ordinálních proměnných, jinak také pořadových, na kterých je aplikován Spearmanův koeficient pořadové korelace, Goodmanova-Kruskalova gama, Kendallovo tau-b a tau-c a Somersovo d.
c) Ordinální vysvětlované proměnné, kde je řešen Kruskalův-Wallisův test. Tato neznámá je vysvětlována za pomocí vysvětlujícího znaku, který je nominálního typu.
d) Dvou kvantitativních proměnných, kde je aplikován Pearsonův korelační koeficient.
e) Kvantitativní vysvětlované proměnné, která je vysvětlovaná pomocí nominálního znaku. Aplikuje se zde výpočet koeficientu éta.
f) Dvou dichotomických proměnných, pro které lze využít zjednodušenou verzi statisticky Chí-kvadrát, a některých z koeficientů vycházející z této statistiky. Dále jsou zde aplikovány speciální koeficienty, které mohou být využity pouze pro případ dichotomických proměnných včetně McNemarova testu, vyjadřující míru mezi dvěma stejnými znaky z různých období. Pro případ malých výběrů dvou dichotomických znaků je aplikován Fisherův exaktní test.
g) Dvou dichotomických a jedné vícekategoriální proměnné, kde je aplikována trojrozměrná kontingenční tabulka a další statistické metody, definovány v kapitole 3.8.17.

Jednotlivé statistické metody jsou aplikovány na datech veřejné databáze PISA 2018, která je k nalezení na webové stránce www.oecd.org. Jedná se o odpovědi respondentů na dotazník nesoucí název STUDENT QUESTIONNAIRE FOR PISA 2018. Tento dotazník byl vyplněn studenty z většiny části světa. Otázky jsou koncipovány tak, aby společnost

PISA získala přehled o podmínkách a úrovni vzdělávání v jednotlivých zemích. Jsou zde například otázky zaměřeny na vzdělání rodičů, kolik hodin se student věnuje vzdělávání, jestli se věnuje studiu před a po škole atp.

Ve většině případů je základem analýzy kontingenční tabulka, kterou lze zobrazit ve výstupu po spuštění procedury CROSSTABS po zvolení sloupcové a řádkové proměnné. Právě v této proceduře lze vybrat většinu z metod pro výpočet míry asociací a testovacích statistik.

Pro vyhodnocení výsledků síly závislostí je postupováno podle definovaných podmínek v kapitole 3.9 Intepretace výsledků.

4.1 Dvě nominální proměnné

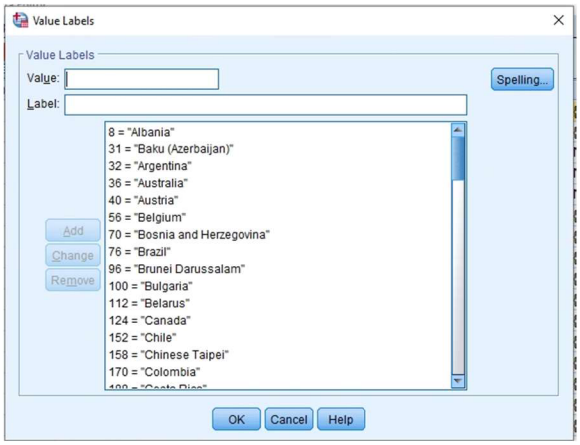
Pro zjištění (ne)závislosti dvou nominálních proměnných si z datového souboru zvolíme například znaky „*Country*“ a nejvyšší dosažené vzdělání jejich matky. Jména těchto znaků jsou vyznačena na obrázku 5.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	CNTRYID	Numeric	3	0	Country Identifier	{8, Albania}...	None	9	Right	Nominal	Input
2	CNT	String	3	0	Country code 3-character	{ALB, Albania}...	None	11	Left	Nominal	Input
3	CNTSCHID	Numeric	8	0	Intl. School ID	None	None	10	Right	Nominal	Input
4	CNTSTUID	Numeric	8	0	Intl. Student ID	None	None	10	Right	Nominal	Input
5	CYC	String	4	0	PISA Assessment Cycle (2 digits + 2 character Assessment type - MS/FT)	None	None	14	Left	Nominal	Input
6	NatCen	String	6	0	National Centre 6-digit Code	{000800, Albania}...	None	20	Left	Nominal	Input
7	STRATUM	String	7	0	Stratum ID 7-character (cnt + region ID + original stratum ID)	{ALB0101, ALB - stratum 01: Urban / North / Public}...	None	23	Left	Nominal	Input
8	SUBNATIO	String	7	0	Adjudicated sub-region code 7-digit code (3-digit country code + region ID + stratum ID)	{0080000, Albania}...	None	23	Left	Nominal	Input
9	OECD	Numeric	1	0	OECD country	{0, No}...	None	6	Right	Nominal	Input
10	ADMINMODE	Numeric	1	0	Mode of Respondent	{1, Paper}...	None	11	Right	Nominal	Input
11	LANGTEST...	Numeric	3	0	Language of Questionnaire	{113, Indonesian}...	None	14	Right	Nominal	Input
12	LANGTEST...	Numeric	3	0	Language of Assessment	{113, Indonesian}...	None	14	Right	Nominal	Input
13	LANGTEST...	Numeric	3	0	Language of Assessment (PAQ)	{113, Indonesian}...	None	14	Right	Nominal	Input
14	BOOKID	Numeric	3	0	Form Identifier	{1, Form 1}...	None	8	Right	Nominal	Input
15	ST001D01T	Numeric	2	0	Student International Grade (Derived)	{7, Grade 7}...	None	11	Right	Ordinal	Input
16	ST003D02T	Numeric	2	0	Student (Standardized) Birth - Month	{1, January}...	95 - 99	11	Right	Ordinal	Input
17	ST003D03T	Numeric	4	0	Student (Standardized) Birth -Year	{9995, Valid Skip}...	9995 - 9999	11	Right	Ordinal	Input
18	ST004D01T	Numeric	1	0	Student (Standardized) Gender	{1, Female}...	5 - 9	11	Right	Nominal	Input
19	ST005Q01TA	Numeric	2	0	What is the <highest level of schooling> completed by your mother?	{1, ISCED level 3A}...	95 - 99	12	Right	Nominal	Input

Obrázek 5: Vybrané proměnné

Proměnná určující nejvyšší vzdělání matka má mimo jiné vyplněn sloupec „*Missing*“, o kterém se zmiňuje kapitola 3.1.2 Definice jednotlivých proměnných. Hodnoty definované v „*Missing*“ tedy určují zakódované možnosti proměnné 95-99, které se do analýzy nezahrnou. Možnosti, které jsou zakódovány těmito čísly, jsou zobrazeny na obrázku 7. Vyznačují se tím, že odpověď na danou otázku byla validně přeskočena, bez odpovědi atp. I takovéto hodnoty je vhodné zapojit do analýzy, která by mohla dokázat příčinu těchto odchylek, ale to není cílem této práce.

Obrázek 7 mimo jiné zobrazuje také kódy možností, které v analýze zahrnuty budou. Následně na obrázku 6 lze vyčíst kódové definice zemí které budou zahrnuty v analýze a jelikož zde nejsou nevalidní odpovědi, nejsou žádné definované v „*Missing*“.

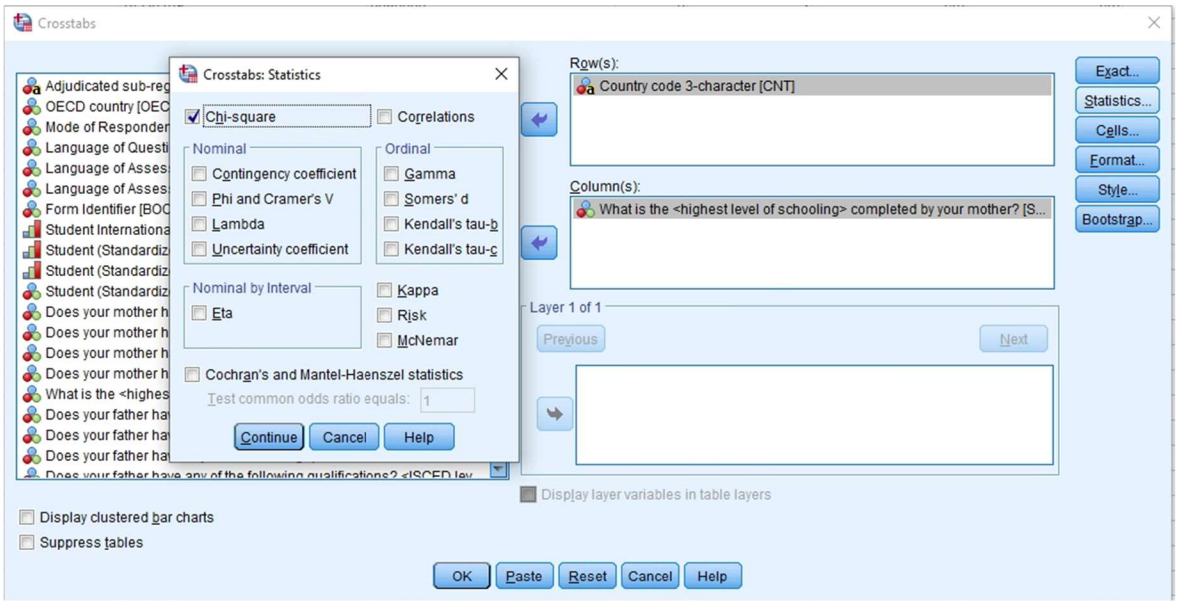


Obrázek 6: Země

Obrázek 7: Dosažené vzdělání matky studenta

4.1.1 Chí-kvadrát test o nezávislosti

Pro výpočet Pearsonovy statistky a věrohodnostního poměru jsou potřeba tabulky absolutních četností a teoretických četností[1] (viz. příloha). K získání výsledků statistiky je pak potřeba otevřít *Statistics* a zaškrtnout položku *Chi-Square* (viz Obrázek 8).



Obrázek 8: Výběr Chi-square statistiky v SPSS

Výsledek Pearsonova Chí-kvadrát testu a věrohodnostního poměru lze vyčíst z výstupu 1, který vrátil SPSS po spuštění dané procedury. Jedná se tedy o hodnoty ve sloupci *Value* pro první dva řádky, *Pearson Chi-Square* a *Likelihood Ratio*.

[1] Teoretické četnosti se ve výstup automaticky nezobrazují. Je-li potřeba je mít součástí výstupu, pak před spuštěním procedury CROSSTABS je potřeba nastavit v *Cells* a v sekci *Counts* možnosti *Observed* a *Expected*. Po spuštění procedury se tyto četnosti zobrazí v jedné kontingenční tabulce.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	224107,656[a]	316	,000
Likelihood Ratio	196704,858	316	,000
N of Valid Cases	586737		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 61,42.

Výstup 1: Chí-kvadrát test

Z Výstupu 1 lze dále vyčíst stupně volnosti, které jsou ve sloupci *df*, a ve sloupci *Asymptotic Significance (2-sided)* je hodnota minimální hladiny významnosti, od které se zamítá hypotéza H_0 o nezávislosti sledovaných proměnných.

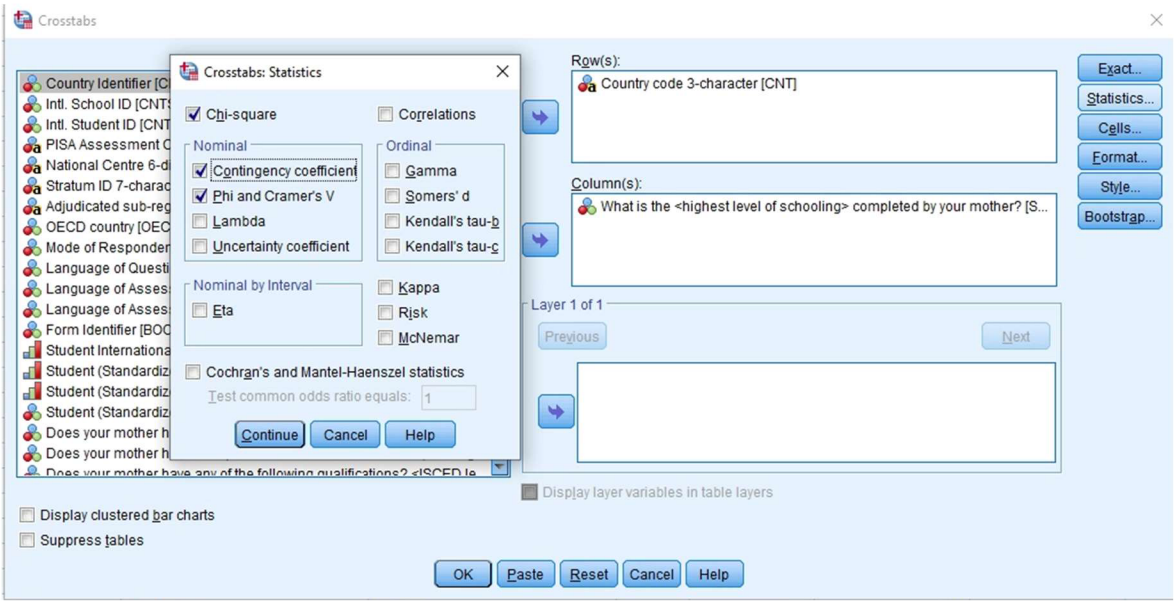
Pod tabulkou je uvedena také poznámka o tom, že žádná z očekávaných hodnot není menší než 5, čímž byla splněna podmínka pro použití Chí-kvadrát testu, a minimální hodnota očekávané četnosti je 61,42.

Pro zamítnutí nulové hypotézy o nezávislosti je potřeba znát kritickou hodnotu. Pokud je zvolena hladina významnosti 5%, pak kvartil $\chi^2_{0,95}[316] = 275{,}8169$. Nulová hypotéza o nezávislosti se zamítá, protože hodnota Pearsonova Chí-kvadrát testu je vyšší než kritická hodnota. Zamítnutí hypotézy o nezávislosti, lze také určit již podle minimální hladiny významnosti z výstupu 1. Jelikož je hodnota 0 lze zamítnout nulovou hypotézu na hladině významnosti 1% i 5%.

4.1.2 Symetrické koeficienty měr asociací

Hodnota Chí-kvadrát testu se následně používá pro výpočet koeficientů, které určí míru dané závislosti mezi sledovanými znaky. V této části jsou uvedeny koeficienty, které jsou symetrické. Jedná se o Pearsonův kontingenční koeficient, Cramerovo V, koeficient φ a Čuprovův kontingenční koeficient (vzorce těchto koeficientů jsou uvedeny v kapitole 3.8.10).

V SPSS je potřeba pro výpočet všech koeficientů v proceduře CROSSTABS otevřít modálním okně *Statistic* a v sekci *Nominal* zaškrtnout *Contingency coefficient* a *Phi and Caramer´s V* (viz obrázek 9).



Obrázek 9: Výběr symetrických měr asociací pro dvě nominální proměnné

Po spuštění procedury SPSS zobrazuje následující výstup vypočítaných symetrických měr asociací.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,618	,000
	Cramer's V	,309	,000
	Contingency Coefficient	,526	,000
N of Valid Cases		586737	

Výstup 2: Míry závislosti založené na Chí-kvadrátu

Řádek *Contingenci coefficient* je hodnota Pearsonova kontingenčního koeficientu, který v našem případě může nabývat hodnot z intervalu $\langle 0; 0{,}89\rangle$ (hodnota 0,89 byla vypočtena podle vzorce $\sqrt{(q-1)/q}$ kde $q = \min\{R, S\}$ tedy $q = 5$). Hodnota tohoto koeficientu je zde podstatně daleko od nulové hodnoty, proto pro lze konstatovat podstatnou až velmi silnou závislost. Stejná interpretace platí pro koeficient *Phi*.

Pro vyhodnocení závislosti pouze jednoho vztahu se dle Řezankové vybere nejspíše Cramerova V, které může nabývat hodnoty z intervalu $\langle 0; 1\rangle$ a proto jeho hodnotu můžeme nejlépe porovnat, zda se jeho výsledek blíží více 0 nebo 1. V tomto případě lze hovořit o mírné až střední silné síle závislosti. Tedy hodnota se vzdaluje od nulové hodnoty, která označuje plnou nezávislost.

Ve výstupu 2 není uvedena hodnota další symetrické míry Čuprovova kontingenčního koeficientu. Jelikož se nejedná o čtvercovou tabulku, hodnota tohoto koeficientu se nemusí shodovat s Cramerovým V, (jak je zmíněno v kapitole 3.8.10). Daný koeficient je proto potřeba spočítat dosazením do vzorce. Po dosazení již vypočtené testové hodnoty, N a hodnoty stupních volnosti získané z Výstupu 1 je získán výsledek koeficientu:

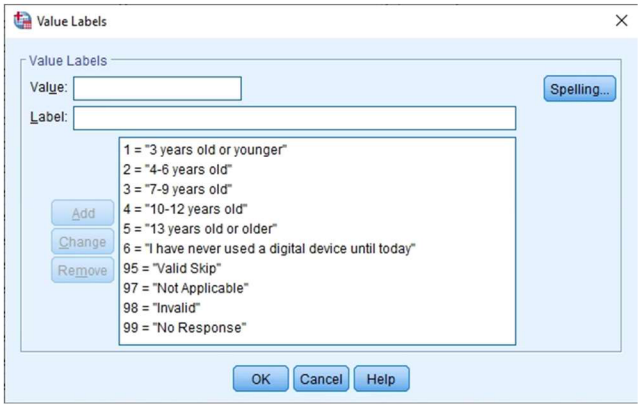
$$C_T = \sqrt{\frac{\chi_P^2/n}{\sqrt{(R-1)(S-1)}}} = \sqrt{\frac{224107{,}656/586737}{\sqrt{316}}} = 0{,}146583.$$

Výsledek Čuprovova kontingenčního koeficientu je poměrně nižší než v případě Cramerova V, což naznačuje na velmi nízkou sílu závislosti sledovaných znaků. Výsledek je tedy blíže k nulové hodnotě než v případě Cramerova V.

Součástí výstupu v SPSS je také sloupec minimálních hladin nezávislosti, od kterých se zamítá nulová hypotéza o nezávislosti. Podle zjištěných výsledků lze říct, že ve všech případech se může uvažovat závislost mezi sledovanými proměnnými. Nicméně podle výsledků jednotlivých měr nelze tuto závislost považovat za významnou.

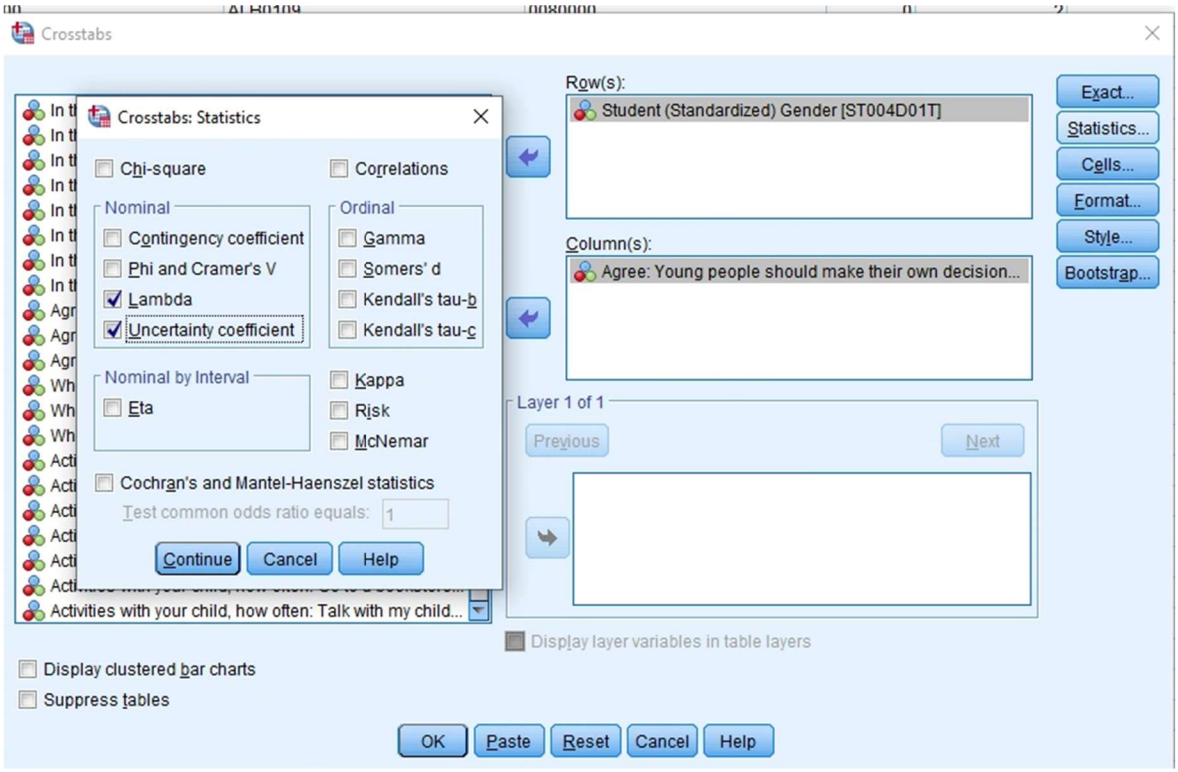
4.1.3 Asymetrické koeficienty měr asociací

Pro aplikaci asymetrických koeficientů jsou zvoleny jiné proměnné, než to bylo v předchozím případě. Jsou to konkrétně Pohlaví studenta a odpověď na otázku *„How old are you were you first used a digital device“* (ve volném překladu, jak staří byli respondenti, když použili první digitální zařízení). Na obrázku 10 jsou zobrazeny možnosti odpovědí, které mohli respondenti vyplnit. V *„Missing“* jsou definovány hodnoty 95-99. Stejné odpovědi 95-99 může nabývat tako proměnná týkající se pohlaví, a proto i u ní jsou tyto hodnoty definované jako *„Missing“*.



Obrázek 10: Možnosti proměnné „How old are you were you first used a digital device“

V proceduře CROSSTABS, je potřeba vybrat si dané proměnné a ve *Statistics* v sekci *Nominal* zaškrtnout *Lambda* a *Uncertainty coefficient* (viz obrázek 11)



Obrázek 11: Výběr asymetrických měr v SPSS

V tomto případě se sleduje jednostranná závislost, tedy zda a případně do jaké míry je odpověď závislá na pohlaví studenta.

Po spuštění procedury jsou získány výstupy 3 a 4. Přičemž ve výstupu 3 jsou uvedeny marginální a teoretické (nebo jinak také očekávané) četnosti a ve výstupu 4 lze vyčíst hodnoty koeficientů. Součástí výstupu 4 jsou i minimální hladiny významnosti od kterých se zamítá nulová hypotéza o nezávislosti. Sloupec, ve kterém se tato hladina nachází nese název *Approximate Significance*.

Podle hodnot koeficientů, které se čtou z prvního sloupce a z řádku *Dependent* daného koeficientu, si lze povšimnout, že výsledky jsou velice blízké nule, v případě lambdy lze pozorovat i nulovou hodnotu. Proto ve všech případě nelze nulovou hypotézu o nezávislosti zamítnout na hladině významnosti 5 %.

Student (Standardized) Gender * How old were you when you first used a digital device?

Crosstabulation

			How old were you when you first used a digital device?						
			3 years old or younger	4-6 years old	7-9 years old	10-12 years old	13 years old or older	I have never used a digital device until today	Total
Student (Standardized) Gender	Female	Count	10011	45605	65282	40761	13043	1584	176286
		Expected Count	13280,5	49060,5	62679,5	37151,1	12075,9	2038,5	176286
	Male	Count	16393	51936	59336	33102	10966	2469	174202
		Expected Count	13123,5	48480,5	61938,5	36711,9	11933,1	2014,5	174202
Total		Count	26404	97541	124618	73863	24009	4053	350488
		Expected Count	26404	97541	124618	73863	24009	4053	350488

Výstup 3: Kontingenční tabulka marginálních a teoretických četností

Directional Measures

			Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,034	,001	38,086	,000
		Student (Standardized) Gender Dependent	,078	,002	38,086	,000
		How old were you when you first used a digital device? Dependent	,000	,000	.[c]	.[c]
	Goodman and Kruskal tau	Student (Standardized) Gender Dependent	,010	,000		,000[d]
		How old were you when you first used a digital device? Dependent	,002	,000		,000[d]
	Uncertainty Coefficient	Symmetric	,004	,000	29,364	,000[e]
		Student (Standardized) Gender Dependent	,007	,000	29,364	,000[e]
		How old were you when you first used a digital device? Dependent	,003	,000	29,364	,000[e]

a. Not assuming the null hypothesis.

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

c. Cannot be computed because the asymptotic standard error equals zero.

d. Based on chi-square approximation

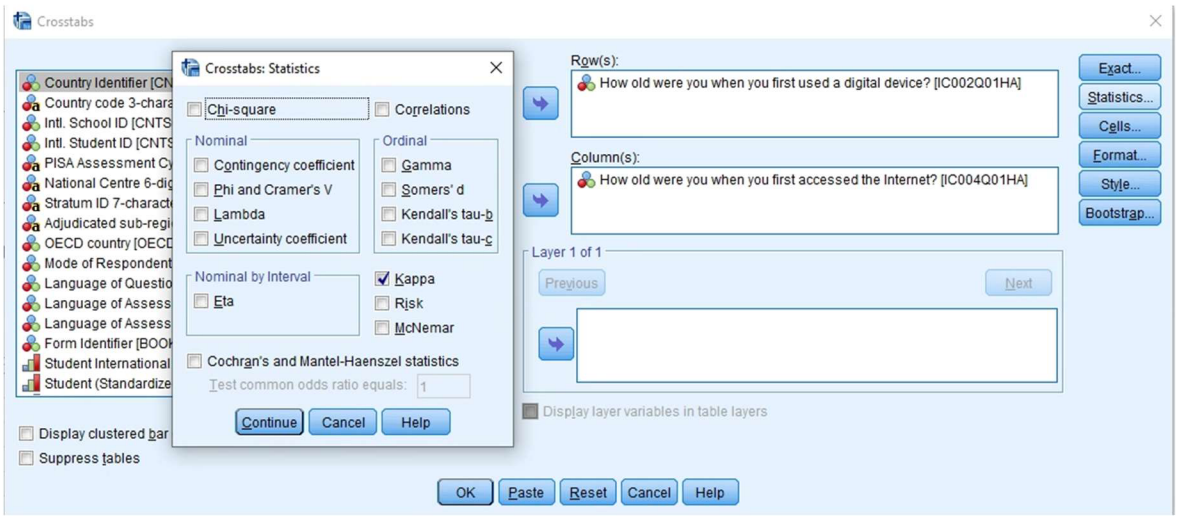
e. Likelihood ratio chi-square probability.

Výstup 4: Koeficienty lambda, tau a koeficient nejistoty

4.1.4 Cohenovo κ (kappa)

Koeficient Cohenovo kappa lze použít pouze v případě kdy má tabulka stejný počet řádků a sloupců. Proto je potřeba si pro tento případ zvolit takové nominální proměnné, které mohou nabývat hodnot (odpovědí) stejného počtu. Pro simulaci tohoto koeficientu jsou zvoleny odpovědi na otázky *„How old are you were you when you first accessed the Internet“* a *„How old are you were you first used a digital device“.* Tedy souvislost mezi tím v kolika letech student poprvé použil nějaké digitální zařízení a prvním připojením na internet. Možnosti odpovědí na tyto dvě otázky jsou stejné, tedy je stejný i jejich počet a lze simulovat koeficient.

Je potřeba spustit proceduru CROSSTABS, kde je ve *Statistics* zvolena možnost *Kappa*, viz obrázek 12.



Obrázek 12: Zvolení koeficientu kappa

Ve výstupu SPSS je poté získána kontingenční tabulka těchto dvou proměnných, tedy marginální četnosti[2] (viz výstup 3) a také výsledek koeficientu Kappa (viz. výstup 4), který je roven 0,353. Kvůli tomu se může konstatovat slabá až střední míra závislosti, podle již zmiňované kombinaci interpretací z kapitoly 3.9. Hodnota se nachází v intervalu (0,2; 0,5) a je tedy mírně vzdálená od 0.

Nulovou hypotézu o nezávislosti lze zamítnout na hladině významnosti 1% i 5%, vyhodnoceno na základě minimální hladiny významnosti uvedeném ve sloupci *Approximate Significance* výstupu 4.

[2] Pro výpočet koeficientu kappa je nutné mimo jiné znát i teoretické četnosti, tuto četnost lze získat způsobem, který byl již zmiňován v poznámce 1.

How old were you when you first accessed the Internet? * How old were you when you first used a digital device? Crosstabulation

Count

		How old were you when you first used a digital device?					
		3 years old or younger	4-6 years old	7-9 years old	10-12 years old	13 years old or older	Total
How old were you when you first accessed the Internet?	3 years old or younger	7450	1366	553	312	198	9879
	4-6 years old	10626	41557	7969	2152	500	62804
	7-9 years old	5841	41912	70636	13285	2023	133697
	10-12 years old	1585	10798	40574	45629	5678	104264
	13 years old or older	488	1202	4053	11728	14905	32376
Total		25990	96835	123785	73106	23304	343020

Výstup 5: Kontingenční tabulka absolutních četností pro výpočet koeficientu kappa

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	,352	,001	366,621	,000
N of Valid Cases		343020			

a. Not assuming the null hypothesis.

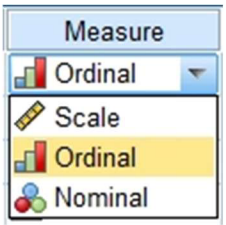
b. Using the asymptotic standard error assuming the null hypothesis.

Výstup 6: Míra shody věku prvního použití digitálního zařízení a Internetu

4.2 Dvě ordinální proměnné

Jak již bylo zmíněno v případě dvou ordinálních proměnných se hovoří o korelaci, kdy existují dva typy, pozitivní a negativní. V této kapitole jsou představeny Spearmanův koeficient pořadové korelace, Goodmanovu-Kruskalovu gamu, Kendallovo tau-b a tau-c a Somersovo d, kdy pouze v případě Somersova d se jedná o asymetrickou variantu.

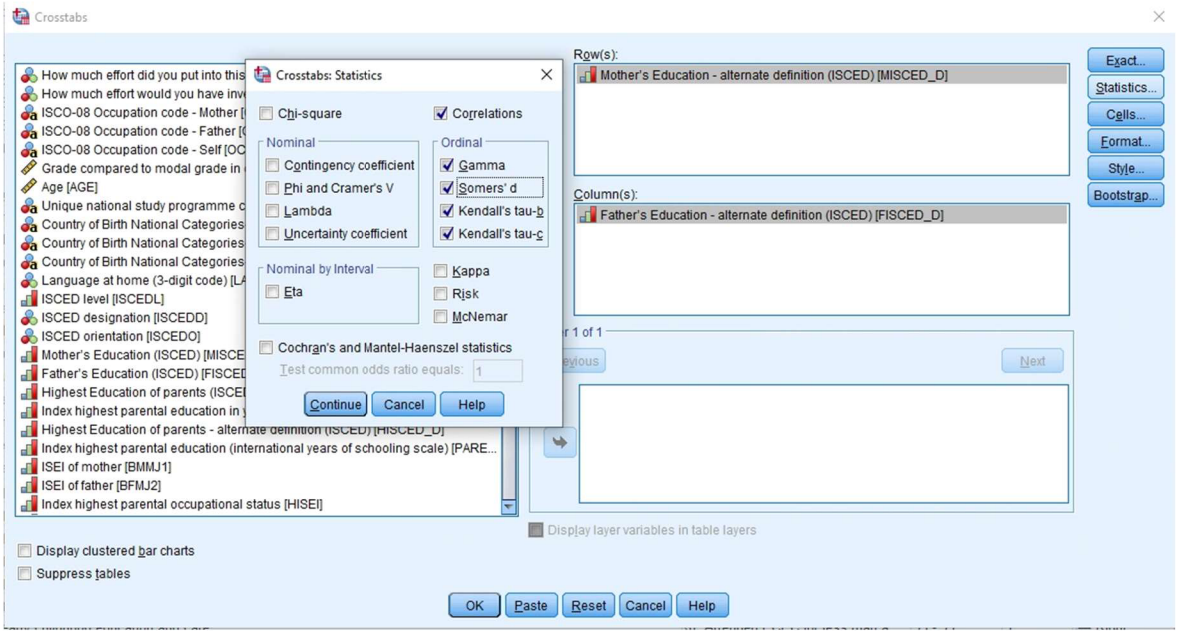
Pro simulaci jmenovaných koeficientů je potřeba si vybrat ze souboru dat dvě ordinální, nebo jinak také pořadové neznámé. Tedy takové neznámé, které lze logicky seřadit. V SPPS byly tyto znaky ve sloupci *Measure* přenastaveny z možnosti *Nominal* na možnost *Ordinal* (viz. obrázek 13).



Obrázek 13: Nastavení proměnné vhodný typ

Pro simulaci byly zvoleny proměnné: *„Mother´s Education – alternate definition (ISCED)“* a *„Father´s Education – alternate definition (ISCED)“*. Tyto úrovně vzdělání se dají snadno seřadit, proto jsou dané proměnné vhodné pro použití. Bude se tedy zkoumat síla závislosti mezi vzděláním matky a otce studenta.

V SPSS je nutné jít opět do procedury CROSSTABS, a ve *Statistics* a provést volbu *Correlations* a v sekci *Ordinal* zaškrtnout možnosti *Gamma*, *Somers´ d, Kendall´s tau -b* a *tau-c* (viz. obrázek 14).



Obrázek 14: Zvolení koeficientů pro dvě ordinální proměnné

Po spuštění procesu ve výstupu je zobrazena kontingenční tabulka s absolutními četnostmi (viz výstup 7), přičemž pro výpočet těchto koeficientů, již není potřeba znát četnosti teoretické. Dále ve výstupu lze najít dvě tabulky (výstup 8 a výstup 9), ve kterých jsou vypočítané zmiňované koeficienty asociačních měr.

Ve výstupech 8 a 9 lze vyčíst, že hodnoty koeficientů jsou mírně pozitivní, jedná se tedy o závislost přímou. Sílu závislosti sledovaných proměnných lze považovat za středně až podstatně silnou, protože se hodnota nachází v intervalu <0,5; 0,8). Z důvodu významného rozdílu od nulové hodnoty lze konstatovat zamítnutí hypotézy o nezávislosti na hladině významnosti 5%, tzn. že se úrovně dosaženého vzdělání matky a otce respondenta spíše shodují. Minimální hladina významnosti je opět v obou výstupech měr ve sloupci nesoucí název *Approximate Significance*.

Mother's Education - alternate definition (ISCED) * Father's Education - alternate definition (ISCED) Crosstabulation

Count

		Father's Education - alternate definition (ISCED)							
		None	ISCED 1	ISCED 2	ISCED 3B, C	ISCED 3A, ISCED 4	ISCED 5B	ISCED 5A, 6	Total
Mother's Education - alternate definition (ISCED)	None	6783	3671	2736	556	1836	713	1124	17419
	ISCED 1	3098	13133	6503	1383	4658	1884	2751	33410
	ISCED 2	2652	6339	28210	5312	13140	4891	5716	66260
	ISCED 3B, C	598	1282	4356	18211	6577	3233	4119	38376
	ISCED 3A, ISCED 4	2308	5129	13142	10100	77047	13841	23080	144647
	ISCED 5B	794	1945	5205	4025	14590	36841	20211	83611
	ISCED 5A, 6	1438	3094	7406	6095	25697	22963	120468	187161
Total		17671	34593	67558	45682	143545	84366	177469	570884

Výstup 7: Kontingenční tabulka absolutních četností dvou ordinálních proměnných

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	,513	,001	528,022	,000
		Mother's Education - alternate definition (ISCED) Dependent	,510	,001	528,022	,000
		Father's Education - alternate definition (ISCED) Dependent	,515	,001	528,022	,000

a. Not assuming the null hypothesis.

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

Výstup 8: Somersovo d dvou ordinálních proměnných

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-b	,513	,001	528,022	,000
	Kendall's tau-c	,472	,001	528,022	,000
	Gamma	,613	,001	528,022	,000
	Spearman Correlation	,589	,001	550,685	,000[c]
Interval by Interval	Pearson's R	,594	,001	558,547	,000[c]
N of Valid Cases		570884			

a. Not assuming the null hypothesis.

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

c. Based on normal approximation.

Výstup 9: Symetrické míry asociací pro dvě ordinální proměnné

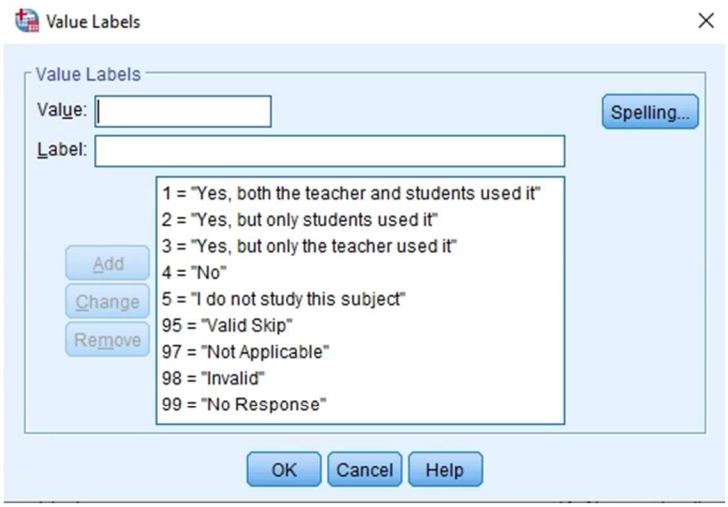
Ve výstupu symetrických měr je uveden mimo jiné výsledek Pearsonova korelačního koeficientu (*Pearson's R*) pro intervalové proměnné. Hodnotu SPSS zobrazí v tomto

výstupu vždy, když je vybrána možnost *Correlation* v proceduře CROSSTABS. Touto hodnotou se zabývá kapitola 4.4 Dvě kvantitativní proměnné, kde jsou zvoleny vhodné proměnné pro výpočet a vyhodnocení tohoto koeficientu.

4.3 Ordinální vysvětlovaná proměnná

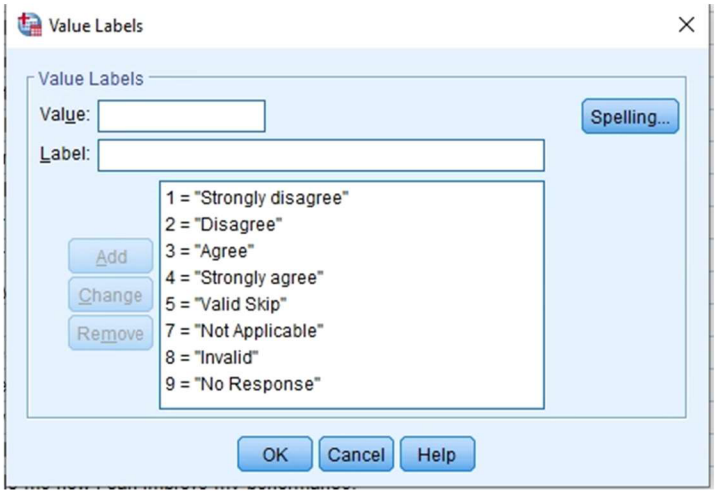
Pokud se se jedná o ordinální vysvětlovanou proměnnou, tak se pro zjištění nezávislosti používá takzvaný Kruskalův-Wallisův test. Přičemž pří platnosti nulové hypotézy o nezávislosti se počítá s rozdělením chí-kvadrát s (R-1) stupni volnosti.

Je třeba si zvolit jednu ordinální proměnnou, která bude proměnnou vysvětlovanou a druhou, podle které se bude vysvětlovaná proměnná dělit do skupin. Pro vysvětlovanou neznámou je zvolena otázka *„Thinking often past two <Test language lessons>" The Teacher made me feel confident in my ability to do well in the course"*, která je vyjádřena pomocí hodnotící škály. Jako vysvětlující znak je zvolen dotaz *„Digital device used for learning or teaching during lessons within the last month <Test language>"*, možnosti odpovědí na tuto otázku lze vidět na obrázku 15. Do analýzy jsou zahrnuty pouze možnosti 1-4, takže 5-99 je nutné definovat jako „Missing".



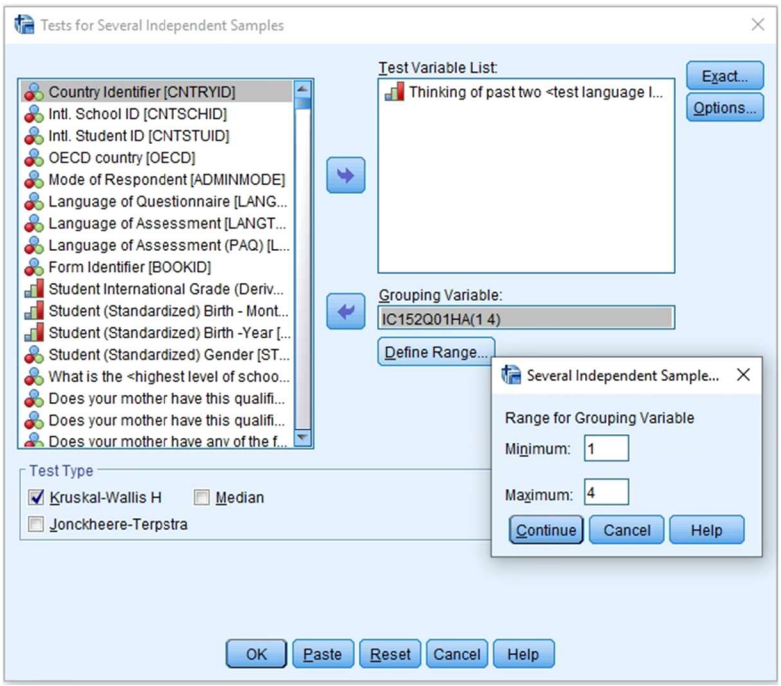
Obrázek 15: Možnosti odpovědí na otázku k proměnné o používání digitálních zařízení na předmětu matematika

Hodnotící škálu vysvětlované proměnné lze pozorovat na obrázku 16. Přičemž možnosti 5-9 jsou definovány ve jako „Missing", což znamená že do analýzy nejsou zahrnuty tyto dopovědi respondentů.



Obrázek 16: Možnosti proměnné

Pro zjištění nezávislosti takového případu je potřeba použít Kruskalův-Wallisův test. SPSS výsledek zobrazí ve výstupu po spuštění Neparametrického testu *K Independent Samples*. Před spuštěním je potřeba zaškrtnout možnost *Kruskal-Wallis H*, do *„Test Variable List"* vložit proměnnou vysvětlovanou a do *„Grouping Variable"* proměnnou vysvětlující, pro kterou je zvolena její maximální a minimální hodnota, v tomto případě případě 1-4 (viz. obrázek 17).



Obrázek 17: Kruskalův-Wallisův test v SPSS

Pro samotný výpočet testového kritéria je potřeba znát tabulku sdružených četností, která je ve výstupu 10 (tato tabulka byla získána pomocí spuštění jednoduché deskriptivní procedury CROSSTABS).

Digital device used for learning or teaching during lessons within the last month: <Test language lessons> * Thinking of past two <test language lessons>: The teacher made me feel confident in my ability to do well in the course. Crosstabulation

Count

		Thinking of past two <test language lessons>: The teacher made me feel confident in my ability to do well in the course.				
		Strongly disagree	Disagree	Agree	Strongly agree	Total
Digital device used for learning or teaching during lessons within the last month: <Test language lessons>	Yes, both the teacher and students used it	11469	16756	60867	25942	115034
	Yes, but only students used it	4440	6663	21629	7050	39782
	Yes, but only the teacher used it	6593	14064	43802	15115	79574
	No	10027	17861	49382	17979	95249
Total		32529	55344	175680	66086	329639

Výstup 10: Kontingenční tabulka absolutních četností pro KW test

V samotném výstupu SPSS pro Kruskalův-Wallisův test jsou dvě tabulky. Ve výstupu 11 jsou údaje o pořadí pro jednotlivé kategorie vysvětlované proměnné, tj. počet objektů (*N*) a průměrné pořadí (*Mean Rank*). V druhé tabulce, která je označena jako výstup 12 je uvedena hodnota testového kritéria, stupeň volnosti (*df*) a minimální hladina významnosti.

Ranks

	Digital device used for learning or teaching during lessons within the last month: <Test language lessons>	N	Mean Rank
Thinking of past two <test language lessons>: The teacher made me feel confident in my ability to do well in the course.	Yes, both the teacher and students used it	115034	170251,75
	Yes, but only students used it	39782	159993,26
	Yes, but only the teacher used it	79574	165048,48
	No	95249	160085,05
	Total	329639	

Výstup 11: Výsledek Kruskalova-Wallisova testu (Pořadí)

Test Statistics[a,b]

	Thinking of past two <test language lessons>: The teacher made me feel confident in my ability to do well in the course.
Kruskal-Wallis H	854,554
df	3
Asymp. Sig.	,000

a. Kruskal Wallis Test

b. Grouping Variable: Digital device used for learning or teaching during lessons within the last month: <Test language lessons>

Výstup 12: Výsledek Kruskalova-Wallisova testu (hodnota testového kritéria)

Součástí výstupu není hodnota opravy na spojitost. Tato hodnota je vypočtena dosazením do vzorce

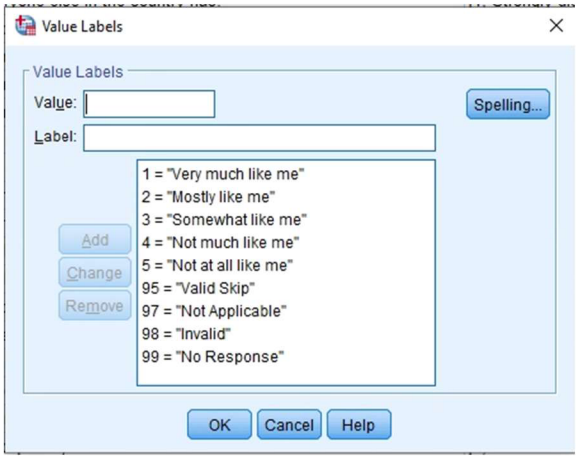
$$H_S = \frac{n^3 - \sum_{j=1}^{S} n_{+j}^3}{n \cdot (n^2 - 1)} = \frac{329639^3 - (32529^3 + 55344^3 + 175680^3 + 66086^3)}{329639 \cdot (329639^2 - 1)} = 0{,}695.$$

Nulová hypotéza o nezávislosti lze zamítnout na hladině významnosti 1% i 5%, protože hodnota minimální hladiny významnosti ve výstupu 12 je 0. Kritická hodnota pro hladinu významnosti 5% je kvantil $\chi^2_{0,95}[3] = 0{,}352$, od které se nulová hypotéza zamítá. Existuje tedy závislost mezi vysvětlovanou a vysvětlující proměnnou, což značí že hodnocení pocitu z jazykové lekce závisí na tom, zda se během této lekce používají nějaká digitální zařízení.

Vhodným koeficientem pro zjištění míry této závislosti by bylo Cramerovo V nebo Lamnda. Tyto dva koeficienty jsou již interpretované v případě dvou nominálních proměnných.

4.4 Dvě kvantitativní proměnné

Pro simulaci míry asociace dvou kvantitativních proměnných jsou ze souboru dat zvoleny znaky *„How well does the following describe you: I am interested in how peple from various cultures see the world“* a *„How well does the following describe you: I am interest in finding out about the traditions of other cultures.“*. Tyto proměnné student hodnotil od 1 do 5. Tyto hodnoty jsou v SPSS slovně vydefinovány viz. obrázek 18.



Obrázek 18: Hodnotící hodnoty proměnné

V analýze je použit jen takový výběr z odpovědí, kde student hodnotil 1–5. Tedy hodnoty 95–99 jsou definovány jako *Missing*.

Pro zjištění intenzity závislosti odpovědí, jak respondenta zajímá vnímaní světa z pohledu jiných kultur a zájem o poznávání tradic ostatních kultur, je použit Pearsonův korelační koeficient. Vstupními daty pro výpočet je kontingenční tabulka absolutních četností (viz. výstup 13). Tuto tabulku vrátí SPSS ve výstupu i s vyhodnocením koeficientu (viz Výstup 14). Tedy po spuštění procedury CROSSTABS, kde ve *Statistics* je potřeba vybrat možnost *Correlations*.

How well does the following describe you: I am interested in how people from various cultures see the world. * How well does the following describe you: I am interested in finding out about the traditions of other cultures. Crosstabulation

Count

		How well does the following describe you: I am interested in finding out about the traditions of other cultures.					
		Very much like me	Mostly like me	Somewhat like me	Not much like me	Not at all like me	Total
How well does the following describe you: I am interested in how people from various cultures see the world.	Very much like me	12652	1837	675	184	136	15484
	Mostly like me	3205	13845	2916	550	183	20699
	Somewhat like me	1516	4913	18362	2296	512	27599
	Not much like me	476	1047	3747	6246	848	12364
	Not at all like me	354	335	1098	912	3936	6635
Total		18203	21977	26798	10188	5615	82781

Výstup 13: Absolutní četnosti dvou kvantitativních proměnných

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Interval by Interval	Pearson's R	,736	,002	312,614	,000[c]
Ordinal by Ordinal	Spearman Correlation	,737	,002	313,888	,000[c]
N of Valid Cases		82781			

a. Not assuming the null hypothesis.

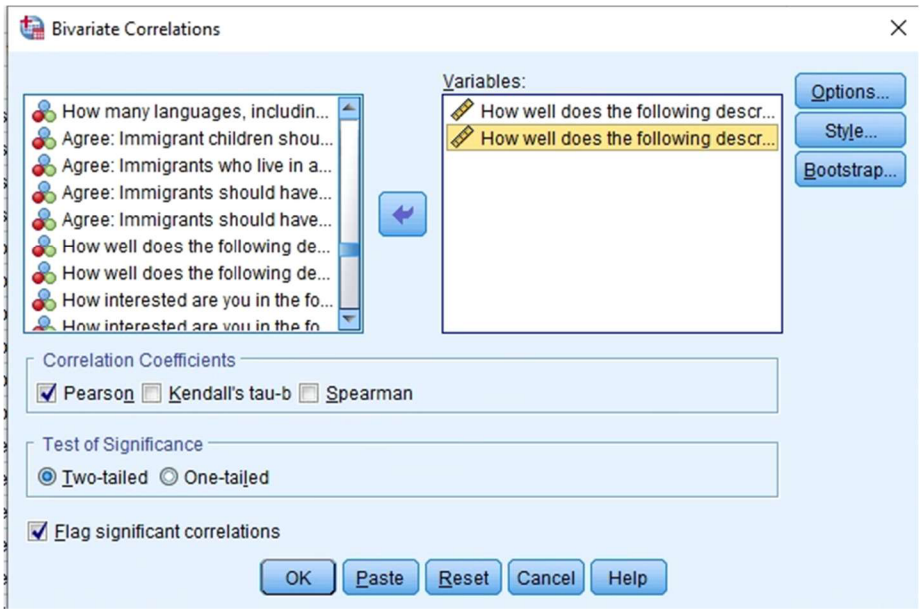
b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Výstup 14: Vyhodnocení korelačního koeficientu dvou kvantitativních proměnných

Ve výstupu 14 je mimo jiné uveden také Spearmanův korelační koeficient, zmíněn v kapitole 4.2. Tyto koeficienty se vyhodnocují společně při spuštění dané procedury s vybranou možností *Correlation*.

SPSS umožňuje i druhý způsob získání výsledku Pearsonovo korelačního koeficientu. Tím je procedura *Bivariate Correlations*, kterou lze spustit provedením volby *Analyze, Correlate, Bivariate*. Pro získání daného koeficientu je potřeba v sekci *Correlation Coefficients* vybrat možnost *Pearson* a v sekci *Variables* vybrat zkoumané veličiny (viz obrázek 19). Po spuštění této procedury SPSS vrací ve výstupu korelační matici vyjadřující vztah mezi hodnoceními (viz výstup 15).



Obrázek 19: Procedura „Bivariate Correlations“

Correlations

		How well does the following describe you: I am interested in how people from various cultures see the world.	How well does the following describe you: I am interested in finding out about the traditions of other cultures.
How well does the following describe you: I am interested in how people from various cultures see the world.	Pearson Correlation	1	,736**
	Sig. (2-tailed)		,000
	N	83280	82781
How well does the following describe you: I am interested in finding out about the traditions of other cultures.	Pearson Correlation	,736**	1
	Sig. (2-tailed)	,000	
	N	82781	83616

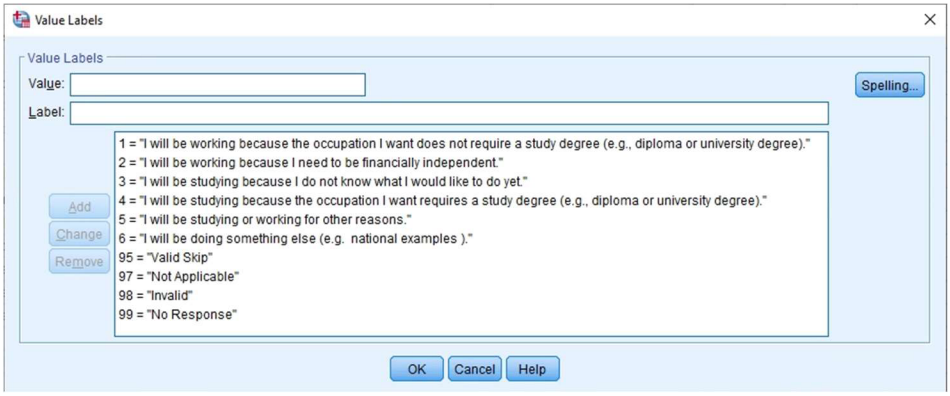
**. Correlation is significant at the 0.01 level (2-tailed).

Výstup 15: Korelační matice vyjadřující závislost dvou kvantitativních proměnných

Z obou postupů je ve výstupech získána stejná hodnota Pearsonova korelačního koeficientu, která se zásadně blíží 1. Jak již bylo zmíněno, tento koeficient může nabývat hodnoty z intervalu od -1 do 1. V tomto případě se jedná o pozitivní koeficient, který označuje přímou závislost. Podle kombinace interpretací v kapitole 3.9 lze sílu závislosti považovat za podstatnou až velmi silnou.

4.5 Kvantitativní vysvětlovaná proměnná

Pro výpočet koeficientu éta, který může nabývat hodnot z intervalu <0; 1>, je potřeba zvolit jako vysvětlovanou neznámou kvantitativní a jako vysvětlující neznámou nominální. Z datového souboru je jako nominální proměnná zvolena otázka znějící: *„What do you think you will be doing 5 years from now?“*. Možnosti odpovědí na položenou otázku jsou definovány podle obrázku 20. Hodnoty 95-99 jsou definovány jako *Missing*, tedy jsou vyřazeny z analýzy.

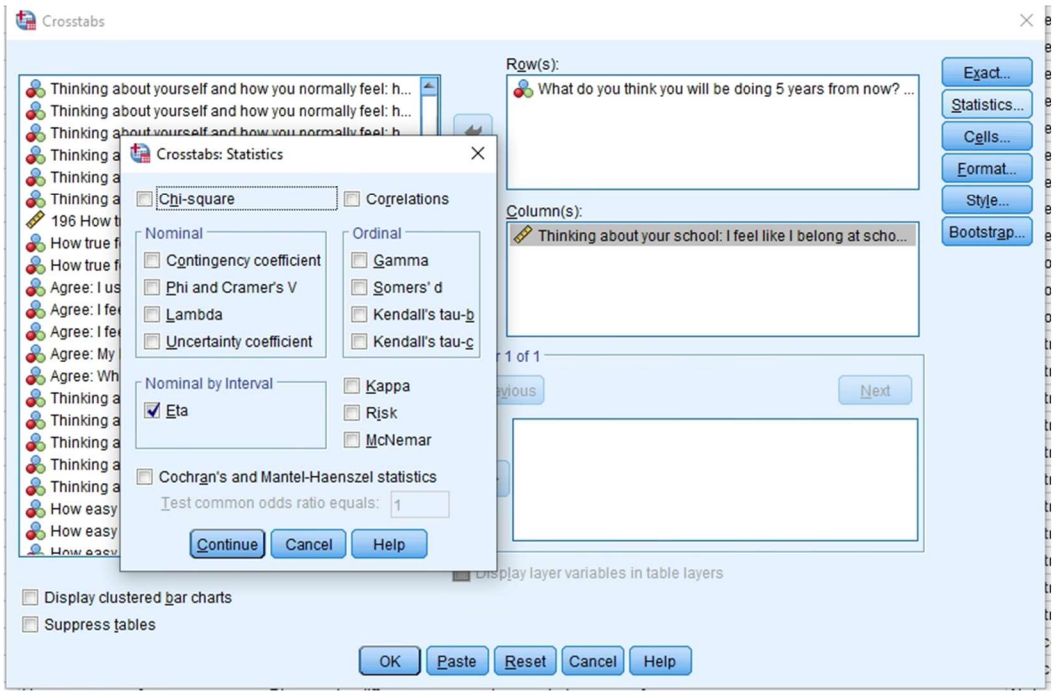


Obrázek 20: Možnosti vysvětlující nominální proměnné

Druhou otázkou, která je označována jako kvantitativní vysvětlovaná neznámá, je: *„Thinking about your school: I feel like I belong at school."* Respondenti tuto otázku mohli hodnotit 1-4. Slovně definované odpovědi lze pozorovat na obrázku 21. Přičemž z analýzy jsou vyřazeny odpovědi 5-9, tedy jsou definovány jako *Missing*. Tyto odpovědi označují nezodpovězenou otázku. Možnosti těchto odpovědí nejsou vhodné pro analýzu, při vyhodnocování kvantitativní neznámé.

Obrázek 21: Možnosti kvantitativní vysvětlované proměnné

Vhodnou metodou pro výpočet míry této závislosti je již zmíněný koeficient éta. Pro získání hodnoty tohoto koeficientu je potřeba vybrat možnost *Eta*, která se nachází ve *Statistics* procedury CROSSTABS a je umístěna v sekci *„Nominal by interval"* (viz obrázek 22).



Obrázek 22: Zvolení si Eta koeficientu v proceduře Crosstabs

Po spuštění procedury ve výstupu SPSS lze najít kontingenční tabulku (viz výstup 16) a výsledek koeficientu (viz výstup17).

What do you think you will be doing 5 years from now? * Thinking about your school: I feel like I belong at school. Crosstabulation

Count

		Thinking about your school: I feel like I belong at school.				
		Strongly agree	Agree	Disagree	Strongly disagree	Total
What do you think you will be doing 5 years from now?	I will be working because the occupation I want does not require a study degree (e.g., diploma or university degree).	6337	14557	6239	2553	29686
	I will be working because I need to be financially independent.	6175	16764	7227	2555	32721
	I will be studying because I do not know what I would like to do yet.	5623	17045	6947	2131	31746
	I will be studying because the occupation I want requires a study degree (e.g., diploma or university degree).	20679	45188	13805	4291	83963
	I will be studying or working for other reasons.	3576	9328	4076	1354	18334
	I will be doing something else (e.g. national examples).	2097	5120	2600	1145	10962
Total		44487	108002	40894	14029	207412

Výstup 16: Kontingenční tabulka absolutních četností kvantitativní vysvětlované proměnné

V kontingenční tabulce je absolutní četnost kvantitativního znaku, setříděné podle kategorií, které jsou na řádcích tohoto výstupu. Tedy odpovědi na otázku, co si respondent myslí že bude dělat do pěti let (volný překlad otázky).

Z výstupu 17, je důležitá hodnota na druhém řádku tabulky, která určuje požadovanou míru závislosti odpovědí respondenta, co si myslí že bude dělat za 5 let, a jestli si myslí že patří do školy. Z hodnoty vypovídá, že určitá závislost mezi těmito proměnnými je, ale lze ji považovat za velmi nízkou až triviální, protože z intervalu <0;1> je velmi vzdálená od hodnoty 1 a poměrně blízká 0.

Directional Measures

			Value
Nominal by Interval	Eta	What do you think you will be doing 5 years from now? Dependent	,030
		Thinking about your school: I feel like I belong at school. Dependent	,101

Výstup 17: Výsledek koeficientu Eta

4.6 Dvě dichotomické proměnné

Dichotomické neznámé se vyznačují tím, že respondenti mohou odpovídat pouze 2 způsoby. Pokud se porovnávají dvě dichotomické veličiny, jejich kontingenční tabulka absolutních četností je velikosti 2x2 a jedná se o tzv. čtyřpolní tabulku

4.6.1 Chí-kvadrát statistika

Chí-kvadrát statika byla již využita v případě analýzy dvou nominálních veličin, kdy kontingenční tabulka absolutních četností obsahovala více výběrů než 2, jako v případě dichotomických proměnných. Rozdílem je pouze zkrácená verze stejného vzorce.

Pro simulaci výpočtů jsou zvoleny znaky, kde respondenti odpovídali ano nebo ne. Jedná se o otázku *„Why did you study before or after school?"* Která obsahuje v dotazníku několik odpovědí, které mohl respondent označit, buď že s tvrzením souhlasí (Yes), nebo nesouhlasí (No). Ze sedmi možností byly zvoleny odpovědi *„I always study."* a *„All my classmates study before or after school".* Otázkou je, zda respondent vždy (ne)studuje, protože všichni jeho spolužáci (ne)studují před a po škole. Absolutní četnosti těchto odpovědí jsou zobrazeny ve výstupu 18. Ten představuje čtyřpolní kontingenční tabulku, získanou po spuštění procedury CROSSTABS.

Why did you study before or after school? I always study. * Why did you study before or after school? All my classmates study before or after school. Crosstabulation

Count

		Why did you study before or after school? All my classmates study before or after school.		
		Yes	No	Total
Why did you study before or after school? I always study.	Yes	52305	37097	89402
	No	24631	80037	104668
Total		76936	117134	194070

Výstup 18: Kontingenční tabulka dvou dichotomických proměnných

Před spuštěním zmiňované procedury, byla v části *Statistics* zvolena možnost *Chi-square*, díky níž je ve výstupu také upravený výstup 19, zachycující výsledky Pearsonovi statistiky chí-kvadrát („*Pearson Chi-Square*"), statistiku korigovanou na spojitost (*Continuity Correction*) a věrohodnostní poměr (*Likelihood Ratio*). Výsledky hodnot statistiky jsou ve sloupci *Value*. Dle nulového výsledku ve sloupci *Asymptotic Significance (2-sided),* která představuje minimální hladinu významnosti, lze

konstatovat, že se nulová hypotéza o nezávislosti zamítá na hladině významnosti 1% i 5%.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	24647,352[a]	1	,000
Continuity Correction[b]	24645,891	1	,000
Likelihood Ratio	25093,220	1	,000
N of Valid Cases	194070		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 35442,02.

b. Computed only for a 2x2 table

Výstup 19: Chi-square test dvou dichotomických proměnných

Aby se ve výstupu zobrazily také výpočty koeficientů, které vycházejí ze statistiky chí-kvadrát, je potřeba v proceduře CROSSTABS v části *Statistics* zvolit *Contingency coefficient*, *Phi and Cramer´s V* a *Correlations*.

Výsledky koeficientů jsou ve výstupu 20. Koeficient φ se shoduje s hodnotou korelačního koeficientu a také s Cramerovým V. Jediná odlišná hodnota je v případě kontingenčního koeficientu, přičemž se příliš neliší od výsledků ostatních koeficientů.

Dle všech výsledků lze konstatovat že asociace je pozitivní, jedná se tedy o závislost přímou. Sílu této závislosti lze popsat jako nízkou až střední, jelikož se hodnoty nacházejí v intervalu (0,2; 0,5). Jsou poměrně vzdálené od hodnoty 0 (absolutní nezávislosti), ale zdaleka se nepřibližují k hodnotě 1 (absolutní závislosti).

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Nominal by Nominal	Phi	,356			,000
	Cramer's V	,356			,000
	Contingency Coefficient	,336			,000
Interval by Interval	Pearson's R	,356	,002	168,026	,000[c]
N of Valid Cases		194070			

a. Not assuming the null hypothesis.

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

Výstup 20: Výsledky koeficientů dvou dichotomických proměnných

4.6.2 Další koeficienty dvou dichotomických proměnných

Dalším zajímavým koeficientem je určení **procentního rozdílu** odpovědí na dané otázky. Jsou pro něj zvoleny znaky pohlaví studenta („*Student (Standardized) Gender*“) a „*Involved in: I sign environmental or social petitions online.*“.

Pro dosazení hodnot do vzorce je potřeba kontingenční tabulka absolutních četností zvolených proměnných (viz výstup 21). Tuto tabulku lze získat spuštěním procedury CROSSTABS v defaultním nastavením.

Student (Standardized) Gender * Involved in: I sign environmental or social petitions online. Crosstabulation

Count

		Involved in: I sign environmental or social petitions online.		
		Yes	No	Total
Student (Standardized) Gender	Female	10315	32052	42367
	Male	10714	30268	40982
Total		21029	62320	83349

Výstup 21: Kontingenční tabulka pro výpočet procentního rozdílu

Dosazením do vzorce procentního rozdílu je získána zaokrouhlená hodnota

$$PR_{Y|X} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} = \frac{10315}{42367} - \frac{10714}{40982} = -0{,}02$$

Účast podpisu petice online je u žen o 2% procenta nižší než účast mužů.

Yuleovo Q nebo také koeficient **gama** určuje hodnocení závislosti. Pro ukázku v SPSS jsou zvoleny neznámé, které jsou v kontingenční tabulce výstupu 18. Pro získání hodnoty koeficientu gama je potřeba v proceduře CROSSTABS v části *Statistics* zaškrtnout možnost *Gamma*. Po následném spuštění je získán výstup 22.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Gamma	,642	,003	165,781	,000
N of Valid Cases		194070			

a. Not assuming the null hypothesis.

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

Výstup 22: Výsledek Yuleova Q

Na základě výsledku lze konstatovat středně až velmi silnou závislost, protože hodnota se nachází v intervalu <0,5; 0,8). Je poměrně vzdálená od 0 a také se zásadně přibližuje k hodnotě 1. Lze také pozorovat, že hodnota tohoto koeficientu je vyšší, než hodnoty koeficientů založených na chí-kvadrát statistice (viz výstup 20), které neberou v úvahu pořadí kategorií.

Pro výpočet **Yuleova koeficientu vazby** jsou vybrány stejné veličiny jako pro předchozí výpočet Yuleova Q. Dosazením do vzorce je získána zaokrouhlená hodnota

$$Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}} = \frac{\sqrt{52305 \cdot 80037} - \sqrt{37097 \cdot 24631}}{\sqrt{52305 \cdot 80037} + \sqrt{37097 \cdot 24631}} = 0{,}363.$$

Výsledek Yuleova koeficientu vazby je nižší než v případě Yuleova Q (viz výstup 22) a je srovnatelná s koeficienty založené na chí-kvadrát statistice (viz výstup 20).

Výsledky koeficientů **Kendallovo tau-b a Somersovo d** pro dvě dichotomické proměnné lze pozorovat ve výstupech 23 a 24. Ty jsou získané po spuštění procedury CROSSTABS se zaškrtnutými požadovanými koeficienty v části *Statistics*. Zvolené proměnné jsou stejné jako v případě výpočtů síly asociace na základě koeficientů založené na chí-kvadrát statistice.

Hodnoty koeficientů určují slabě pozitivní míru závislosti daných proměnných. Výsledek symetrického koeficientu Somersova d je shodný se symetrickým koeficientem Kendallova tau-b a zároveň s koeficientem φ, korelačním koeficientem a Cramerovým V (viz výstup 20).

Directional Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d Symmetric	,356	,002	165,781	,000
	Why did you study before or after school? I always study. Dependent	,363	,002	165,781	,000
	Why did you study before or after school? All my classmates study before or after school. Dependent	,350	,002	165,781	,000

a. Not assuming the null hypothesis.

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

Výstup 23: Výsledek Somersova d dvou dichotomických proměnných

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-b	,356	,002	165,781	,000
N of Valid Cases		194070			

a. Not assuming the null hypothesis.

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

Výstup 24: Výsledek Kendallova tau-b dvou dichotomických proměnných

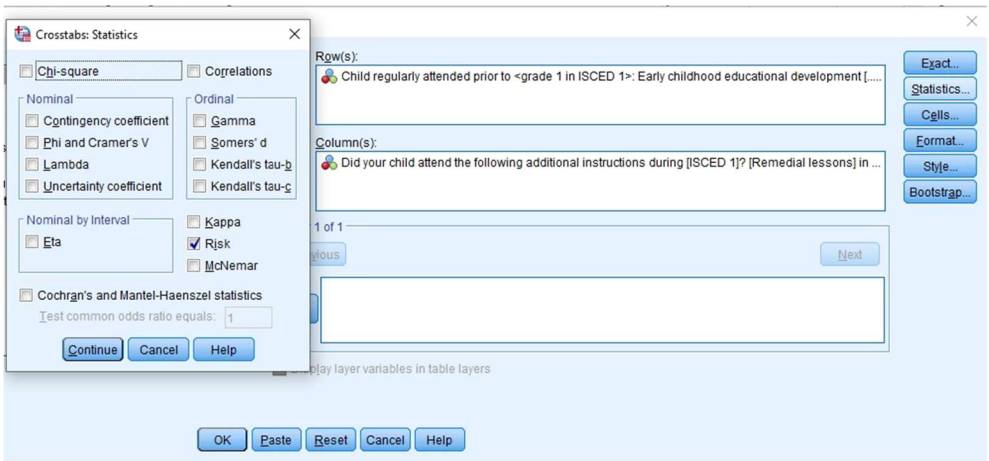
Dalším koeficientem této podkapitoly je **Hamannův koeficient**, kterým se měří míra souhlasu pro čtyřpolní tabulku. Pro dosazení do vzorce tohoto koeficientu je využita kontingenční tabulka výstupu 18. Tím je získán zaokrouhlený výsledek

$$\frac{(n_{11}+n_{22})-(n_{12}+n_{21})}{n} = \frac{(52305+80037)-(37097+24631)}{194070} = 0{,}364.$$

Dle výsledku koeficientu lze konstatovat, že se více hodnot nachází na hlavní diagonále. Jedná se o slabší až střední míru souhlasu těchto znaků, protože se výsledek mírně přibližuje k hodnotě 1 a vzdaluje se od hodnoty nulové.

Posledními koeficienty v této kapitole jsou koeficient relativního rizika a poměr šancí. Pro ně je položena otázka, pokud je dítě vzděláváno už před nástupem do školy, bude s nástupem na první stupeň potřebovat nápravné lekce jazyka. Odpovídající kontingenční tabulka absolutních četností je ve výstupu 25.

Pro získání výsledků koeficientů ve výstupu SPSS je potřeba v proceduře CROSSTABS v části *Statistics* zvolit možnost *Risk* (viz obrázek 23). Po spuštění procedury je získán výsledek požadovaných koeficientů ve výstupu 26.



Obrázek 23: Volba pro výpočet koeficientu relativního rizika

Child regularly attended prior to <grade 1 in ISCED 1>: Early childhood educational development [...] * Did your child attend the following additional instructions during [ISCED 1]? [Remedial lessons] in [test language] Crosstabulation

Count

		Did your child attend the following additional instructions during [ISCED 1]? [Remedial lessons] in [test language]		
		Yes	No	Total
Child regularly attended prior to <grade 1 in ISCED 1>: Early childhood educational development [...]	Yes	7670	24109	31779
	No	4180	19867	24047
Total		11850	43976	55826

Výstup 25: Kontingenční tabulka pro výpočet poměru šancí

Výstup 26 zobrazuje hodnoty poměru šancí (*odds Ratio*), a koeficienty relativního rizika (*For cohort = Yes* a *For cohor = No*). Pro tyto koeficienty se vypisuje nejen bodový odhad, samotná hodnota (*Value*), ale také 95% interval spolehlivosti prostřednictvím dolní (*Lower*) a horní (*Upper*) meze tohoto intervalu.

Risk Estimate

		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Child regularly attended prior to <grade 1 in ISCED 1>: Early childhood educational development [...] (Yes / No)	1,512	1,450	1,577
For cohort Did your child attend the following additional instructions during [ISCED 1]? [Remedial lessons] in [test language] = Yes	1,388	1,342	1,436
For cohort Did your child attend the following additional instructions during [ISCED 1]? [Remedial lessons] in [test language] = No	,918	,911	,926
N of Valid Cases	55826		

Výstup 26: Výsledek poměru šancí a relativních rizik

Jelikož je výsledná hodnota poměru šancí (viz hodnota *Odds Ratio* ve výstupu 26) větší než 1, vyplatí se dítě vzdělávat již před nástupem do školy.

Mimo tyto výsledky lze vyjádřit i jednotlivá rizika/šance. Příkladem jsou vyjádření šancí/rizik, pokud bylo dítě vzděláváno před nástupem do školy, zda nepotřebovalo/potřebovalo jazykové nápravné lekce. Vyjádřením rizika, že dítě potřebovalo zmíněné lekce i po tom, co bylo vzděláváno před nástupem do školy, je

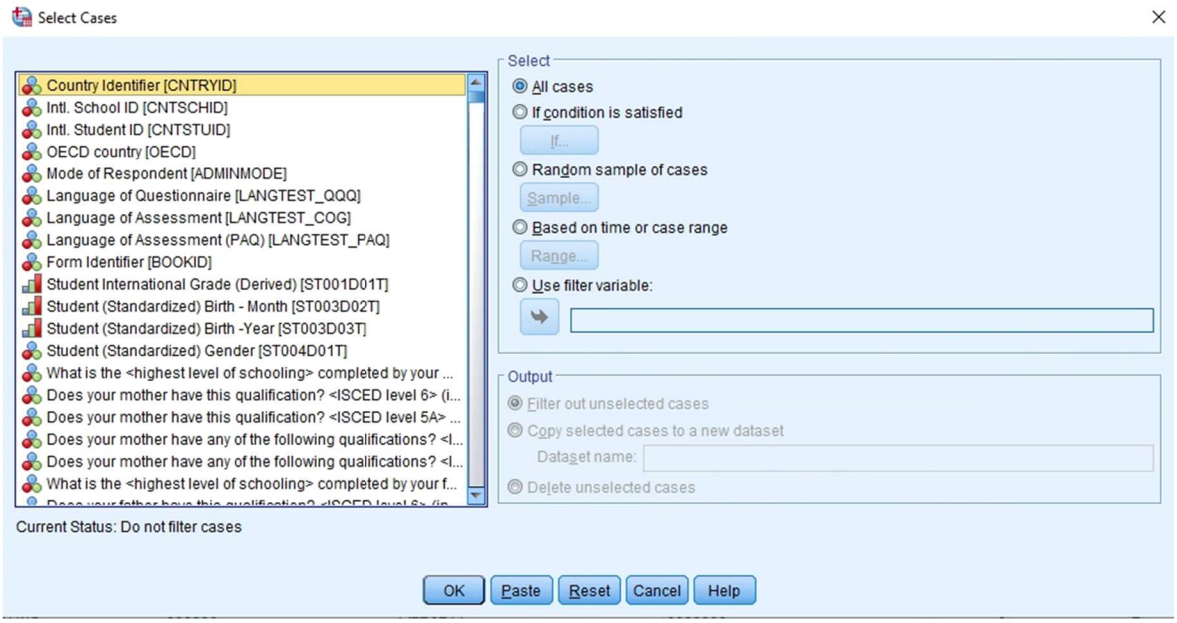
$n_{11}/n_{1+} = 7670/31779 = 0{,}24$. Potom šance, že dítě vzdělávané před nástupem na školu a nepotřebovalo nápravné lekce jazyka po nástupu na školu, je $n_{12}/n_{1+} = 24109/31779 = 0{,}76$.

4.6.3 Fisherův exaktní test

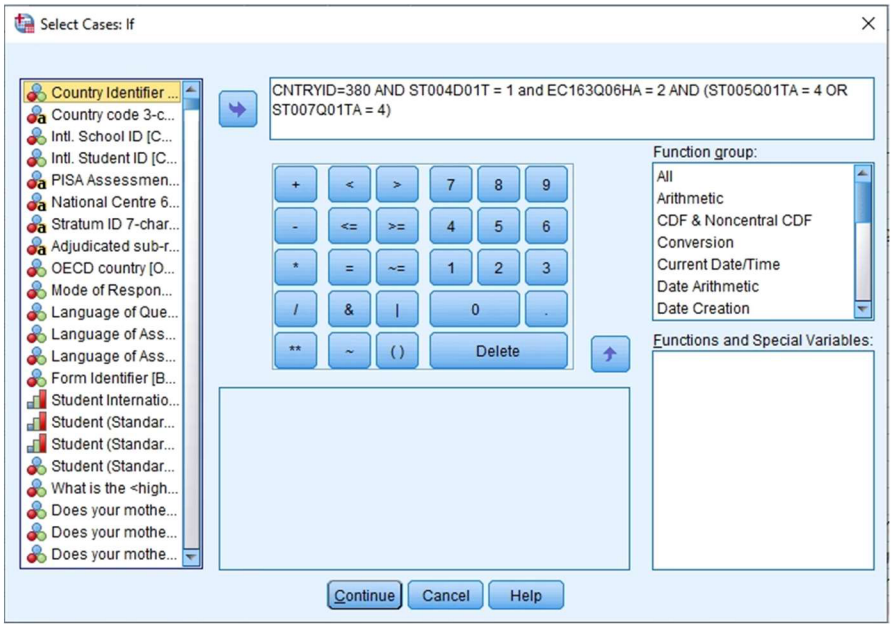
Fisherův exaktní test je vhodné použít, při potřebě (ne)zamítnout hypotézu nezávislosti dvou dichotomických proměnných při malém počtu pozorování nebo pokud alespoň 20% čtyřpolní tabulky je menší než 5 (tedy není splněn předpoklad pro použití chí-kvadrát statistiky).

V tomto případě budeme pozorovat proměnné *„Why didn't you study before or after school? I never study.“* a *„Why did you study before or after school? My parents think studying is important.“*. Respondenti byly omezeni na německou národnost, ženské pohlaví, a jejichž odpověď na otázku „*Why did you study before or after school? I always study*“ je *„No“*. Zároveň jejich otec nebo matka mají nejvyšší dosažené vzdělání „*ISCED level 1*“.

Takový výběr v SPSS lze získat spuštěním procedury případů za pomoci podmínky (procedura *Select cases if*). K této proceduře se dá dostat pomocí voleb *Data, Select Cases*. Po otevření okna *Select Cases* (viz obrázek 24) je potřeba zvolit možnost *„If condition is satisfied“* a po kliknutí na tlačítko *„If...“* se otevře další okno, kde se vytvoří podmínka pro požadované možnosti již zmíněných proměnných (viz. obrázek 25)

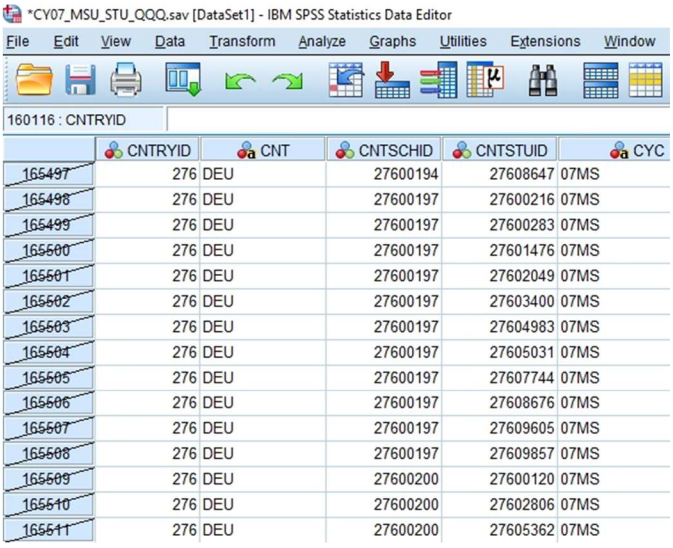


Obrázek 24: Výběr případů



Obrázek 25: Výběr případů za pomoci podmínky

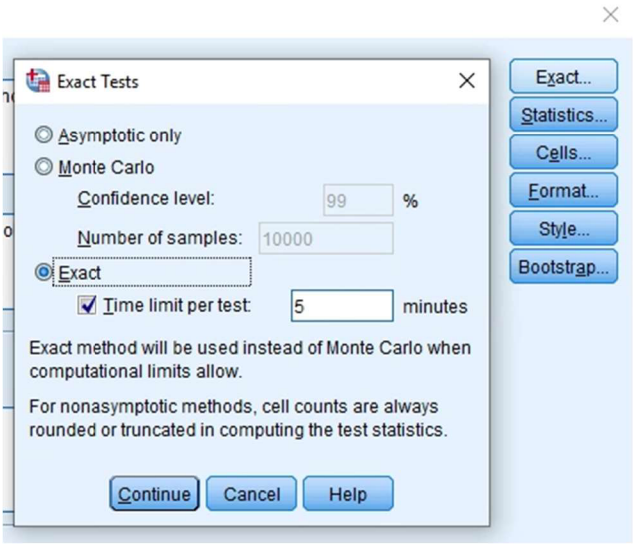
Po spuštění procedury se otevře „*Output*“, kde lze pozorovat úspěšné provedeno spuštěné procedury. „*Data view*“ obsahuje čísla některých řádků přeškrtnuta (viz obrázek 26). Tyto řádky (odpovědi respondentů) nebudou zahrnuty v následující analýze.



	CNTRYID	CNT	CNTSCHID	CNTSTUID	CYC
165497	276	DEU	27600194	27608647	07MS
165498	276	DEU	27600197	27600216	07MS
165499	276	DEU	27600197	27600283	07MS
165500	276	DEU	27600197	27601476	07MS
165501	276	DEU	27600197	27602049	07MS
165502	276	DEU	27600197	27603400	07MS
165503	276	DEU	27600197	27604983	07MS
165504	276	DEU	27600197	27605031	07MS
165505	276	DEU	27600197	27607744	07MS
165506	276	DEU	27600197	27608676	07MS
165507	276	DEU	27600197	27609605	07MS
165508	276	DEU	27600197	27609857	07MS
165509	276	DEU	27600200	27600120	07MS
165510	276	DEU	27600200	27602806	07MS
165511	276	DEU	27600200	27605362	07MS

Obrázek 26: Vyřazené řádky z analýzy

Pro získání výsledku Fisherova testu je potřeba spustit proceduru CROSSTABS, kde je v *Statistics* zvolen *Chi-square* a v části *Exact* je nastaven *Exact* (možnost *Monte Carlo* je zvolena pro případ, kdy by výběr byl příliš velký, a tedy výpočetně náročnější – viz obrázek 27).



Obrázek 27: Exact Tests v SPSS

Po spuštění procedury je získána kontingenční tabulka (viz výstup 27), dle které lze konstatovat, že podmínky pro použití Chí-kvadrát testu byly porušeny. Tuto skutečnost potvrzuje i poznámka ve výstupu 28, že 3 pole (tj. 75%) obsahují očekávanou hodnotu menší než 5.

Why didn't you study before or after school? I never study. * Why did you study before or after school? My parents think studying is important. Crosstabulation

Count

		Why did you study before or after school? My parents think studying is important.		
		Yes	No	Total
Why didn't you study before or after school? I never study.	Yes	3	2	5
	No	11	3	14
Total		14	5	19

Výstup 27: Čtyřpolní tabulka pro aplikaci Fisherova exaktního testu

Výstup 28 je upraveným výstupem z SPSS a zachycuje výsledky dvou variant Fisherova testu, pro oboustrannou alternativní hypotézu *Exact Sig. (2-sided)* a pravostrannou alternativní hypotézu *Exact Sig. (1-sided)*. Dle získaných výsledků obou alternativních hypotéz lze konstatovat že hypotéza o nezávislosti se nezamítá na hladině významnosti 1% a 5%.

Chi-Square Tests

	Value	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Fisher's Exact Test		,570	,397
N of Valid Cases	19		

3 cells (75,0%) have expected count less than 5. The minimum expected count is 1,32.

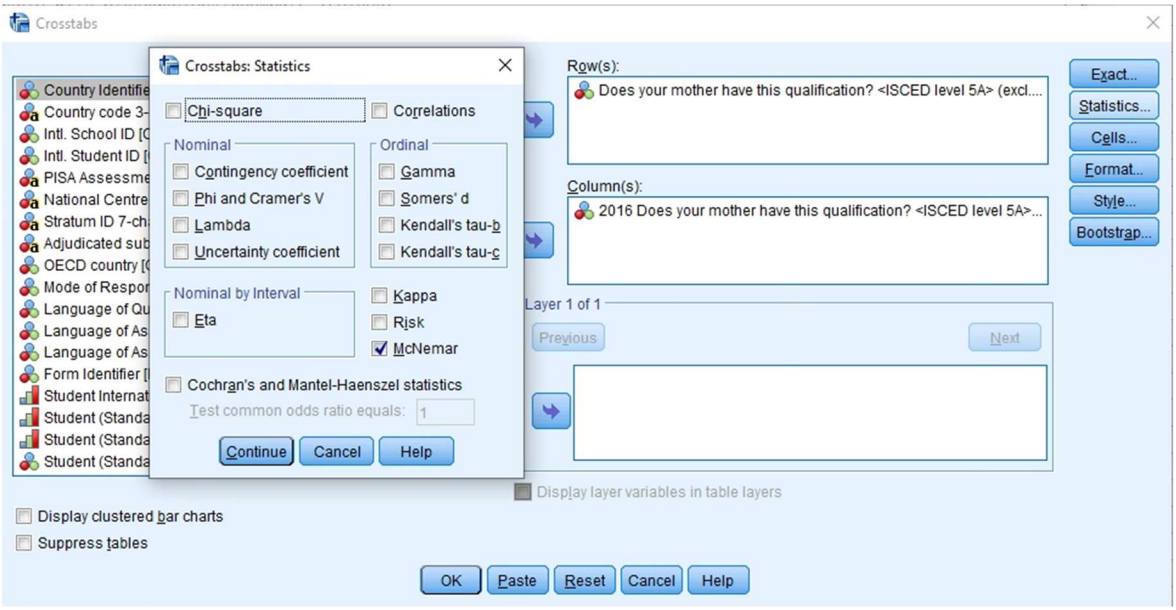
Výstup 28: Výsledek Fisherova exaktního testu

4.6.4 McNemarův test

Dalším speciálním testem je McNemarův test, vyjadřující speciální míru mezi dvěma stejnými znaky z různých období týchž respondentů.

V případě výběru dat bylo potřeba rozšířit výběr o data z roku 2016 od stejných respondentů, rozlišený podle jejich ID, za předpokladu, že ID jednotlivých studentů se pro rok 2018 neliší. Pro porovnání odpovědí je vybraná proměnná *„Does your mother have this qualification? <ISCED level 5A>“* na kterou byla možná odpověď ano a ne.

Výsledek testu lze získat dvěma způsoby. Prvním je spuštění procedury CROSSTABS, kde je v části *Statistics* vybrána možnost *McNemar* (viz obrázek 28). Po spuštění je získán výstup 29 a 30.



Obrázek 28: Výběr McNemarova testu

Does your mother have this qualification? <ISCED level 5A> * 2016 Does your mother have this qualification? <ISCED level 5A> Crosstabulation

Count

		2016 Does your mother have this qualification? <ISCED level 5A>		
		Yes	No	Total
Does your mother have this qualification? <ISCED level 5A> (excl. higher qualifications at level 5A in some countries)	Yes	27926	55203	83129
	No	43092	116276	159368
Total		71018	171479	242497

Výstup 29: Kontingenční tabulka absolutních četností pro McNemarův test

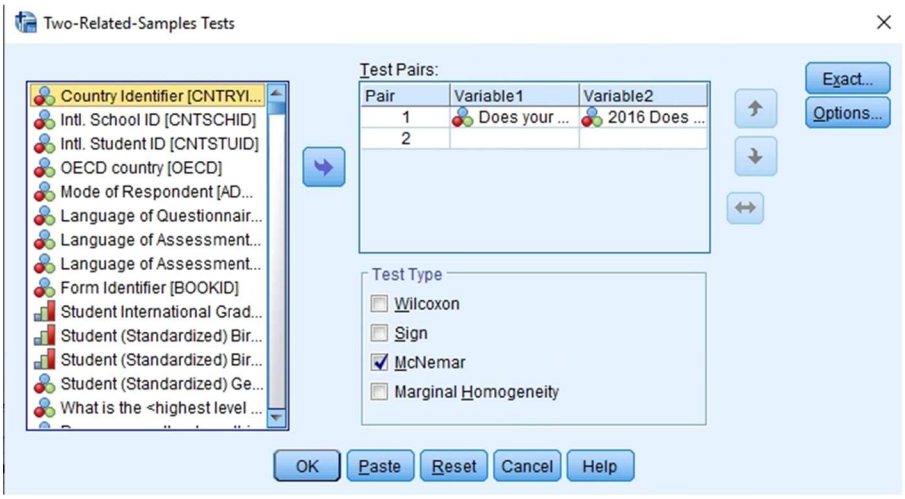
Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		,000[a]
N of Valid Cases	242497	

a. Binomial distribution used.

Výstup 30: Výsledek McNemarova testu 1

Druhým způsobem je v SPSS provést volbu *Analyze*, *Nonparametric Test*, *Legacy Dialogs*, *2 Related Samples*, zvolit si párový test a v sekci *Test Type* zvolit *McNemar* (viz obrázek 29). Po spuštění je ve výstupu jednak kontingenční tabulka absolutních četností se stejnými výsledky jako ve výstupu 29 a také výstup 31, kde mimo jiné lze vyčíst také hodnotu Chí-kvadrát. Dle poznámky *b.* byla hodnota získána použitím korekce na spojitost.



Obrázek 29: Druhá možnost výpočtu McNemarova testu pomocí SPSS

Test Statistics[a]

	Does your mother have this qualification? <ISCED level 5A> (excl. higher qualifications at level 5A in some countries) & 2016 Does your mother have this qualification? <ISCED level 5A>
N	242497
Chi-Square[b]	1491,959
Asymp. Sig.	,000

a. McNemar Test

b. Continuity Corrected

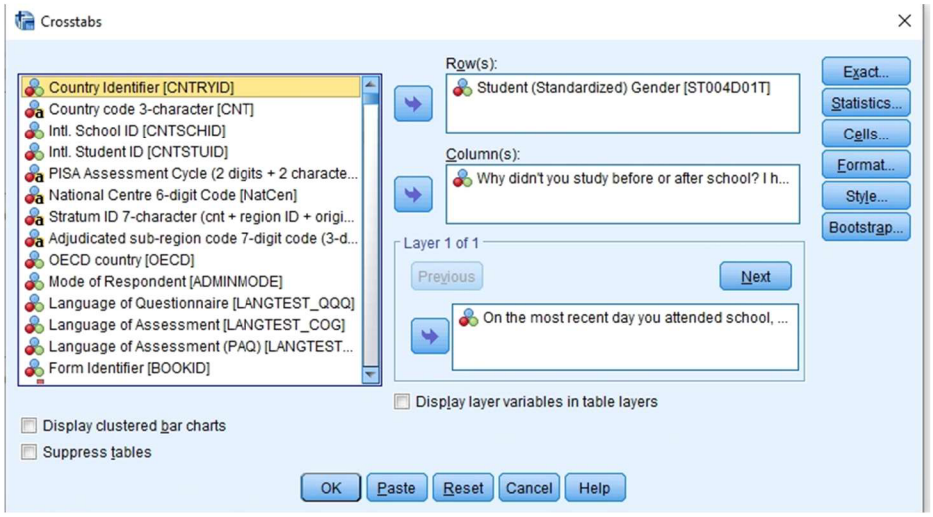
Výstup 31: Výsledek McNemarova testu 2

V obou výstupech vyšly stejné výsledky a podle minimální hladiny významnosti, lze konstatovat že se nulová hypotéza o nezávislosti nezamítá jak na hladině významnosti 1%, tak na 5%.

4.7 Dvě dichotomické a jedna vícekategoriální proměnná

Pro případ rozšíření dvourozměrné kontingenční tabulky o neznámých *„Student gender“* a *„Why didn't you study before or after school? I had no time to study.“* je použit třetí znak *„On the most recent day you attended school, how long did you study in the morning before going to school? Minutes“.* Zkoumanou otázkou je tedy závislost, kolik času tráví respondenti učením ráno před školou, v závislosti na jejich pohlaví, pokud uvedli že na učení nemají čas.

Základní čtyřpolní tabulka (viz výstup 32), s odpověďmi na otázku „*Why didn't you study before or after school? I had no time to study*.“ v závislosti na pohlaví, je ve výstupu 33 rozšířena o minuty strávené učením před školou. Vznikla tím soustava devíti matic. Takovou kontingenční tabulku lze v SPSS získat spuštěním procedury CORSSTABS. V té je potřeba zvolit proměnou řádkovou a sloupcovou. Dále neznámou rozšiřující proměnou řádkovou, která se vybírá v části „*Layer 1 of 1*“ (viz obrázek 30).



Obrázek 30: Zvolení si třetí zkoumané proměnné v proceduře Crosstabs

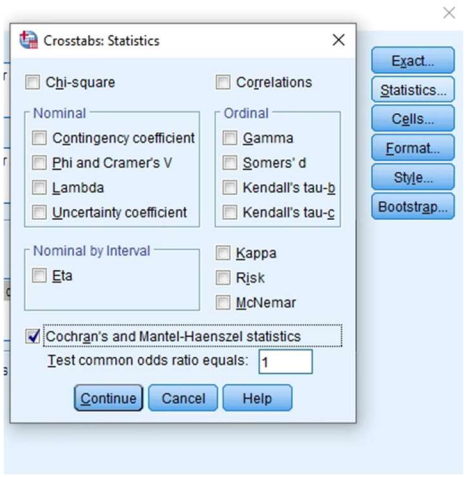
Student (Standardized) Gender * Why didn't you study before or after school? I had no time to study. Crosstabulation

Count

		Why didn't you study before or after school? I had no time to study.		
		Yes	No	Total
Student (Standardized) Gender	Female	10781	8450	19231
	Male	9004	9050	18054
Total		19785	17500	37285

Výstup 32: Čtyřpolní tabulka absolutních četností

Pro získání výstupů 34, 35 a 36 je potřeba v části *Statistics* zvolit možnost *Cochran's and Mantel-Haenszel statistiscs* (viz obrázek 31).



Obrázek 31: Volba Cochranovy a Mantelolvy-Haenszelovy statistiky

Student (Standardized) Gender * Why didn't you study before or after school? I had no time to study. * On the most recent day you attended school, how long did you study in the morning before going to school? Minutes Crosstabulation

Count

On the most recent day you attended school, how long did you study in the morning before going to school? Minutes			Why didn't you study before or after school? I had no time to study. Yes	No	Total
00	Student (Standardized) Gender	Female	512	356	868
		Male	556	506	1062
		Total	1068	862	1930
05	Student (Standardized) Gender	Female	69	51	120
		Male	79	73	152
		Total	148	124	272
10	Student (Standardized) Gender	Female	82	63	145
		Male	88	63	151
		Total	170	126	296
15	Student (Standardized) Gender	Female	69	68	137
		Male	67	68	135
		Total	136	136	272
20	Student (Standardized) Gender	Female	59	52	111
		Male	62	43	105
		Total	121	95	216
25	Student (Standardized) Gender	Female	30	33	63
		Male	43	22	65
		Total	73	55	128
30	Student (Standardized) Gender	Female	129	103	232
		Male	150	143	293
		Total	279	246	525
35	Student (Standardized) Gender	Female	20	24	44
		Male	27	29	56
		Total	47	53	100
40	Student (Standardized) Gender	Female	22	19	41
		Male	32	26	58
		Total	54	45	99
Total	Student (Standardized) Gender	Female	992	769	1761
		Male	1104	973	2077
		Total	2096	1742	3838

Výstup 33: Trojrozměrná kontingenční tabulka

Výsledky testů homogenity jsou obsahem výstupu 34. V prvním řádku se nachází výsledek Breslowa-Dayova testu, a ve druhém výsledky Taroneovy statistiky. Minimální hladina významnosti obou testů je 0,154, proto se na hladině významnosti 5% nezamítá nulová hypotézu o shodě poměru šancí v jednotlivých tabulkách vytvořených podle počtu minut, které respondent tráví nad učením, před odchodem do školy.

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymptotic Significance (2-sided)
Breslow-Day	11,939	8	,154
Tarone's	11,939	8	,154

Výstup 34: Výsledek testu homogenity

Výstup 35 obsahuje výsledky testů podmíněné nezávislosti a jedná se o statistiky *Cochran's* a *Mantel-Haenszel*. Pro obě statistiky lze z výstupu vyčíst hodnota statistiky, počet stupňů volnosti a minimální hladina významnosti od které (ne)zamítáme nulovou hypotézu o nezávislosti. Na prvním řádku je hodnota minimální hladiny významnosti rovna 0,049, takže v tomto případě se nulová hypotéza o nezávislosti zamítá na hladině významnosti 5%. V případě hodnoty hladiny pro Mantelovu-Haenszelovu statistiku nulovou hypotézu zamítnout nelze.

Tests of Conditional Independence

	Chi-Squared	df	Asymptotic Significance (2-sided)
Cochran's	3,867	1	,049
Mantel-Haenszel	3,731	1	,053

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Výstup 35: Výsledky testů podmíněné nezávislosti

Poslední výstup spuštěné procedury obsahuje výsledky, které se vztahují k odhadu společného poměru šancí (viz výstup 36). Ve výstupu lze tedy vyčíst hodnota

Mantelova-Haenszelova odhadu společného poměru šancí (*Estimate*) a jeho logaritmu, podle kterého lze zhodnotit závislost velmi slabou až triviální. Podle hodnoty minimální hladiny významnosti se může konstatovat, že na 5% hladině významnosti se zamítá hypotéza o nulovosti logaritmu společného poměru šancí tzn. o nezávislosti nedostatku času pro učení na pohlaví. Dále jsou ve výstupu uvedeny výsledky pro 95% oboustranný interval spolehlivosti, a to jak pro společný poměr šancí, tak pro logaritmus této hodnoty pomocí dolních a horních mezí.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1,137
ln(Estimate)			,128
Standard Error of ln(Estimate)			,065
Asymptotic Significance (2-sided)			,050
Asymptotic 95% Confidence Interval	Common Odds Ratio	Lower Bound	1,000
		Upper Bound	1,291
	ln(Common Odds Ratio)	Lower Bound	,000
		Upper Bound	,256

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1,000 assumption. So is the natural log of the estimate.

Výstup 36: Výsledky odhadu společného poměru šancí

4.8 Přehled použitých metod

Proměnné	**Seznam použitých metod**	**Způsob získání hodnot**
Dvě nominální proměnné	Chí-kvadrát test Pearsonův kontingenční koeficient Cramerovo V Koeficient φ Čuprovův kontingenční koeficient Goodmanova-Kruskalova τ (tau) Goodmanova-Kruskalova λ Cohenovo κ	Většina hodnot byla získána spuštěním procedury CROSSTABS a zvolením si vhodné metody ve *Statistics*. V případě získání hodnoty Čuprovova kontingenčního koeficientu bylo potřeba dosadit do vzorce $C_T = \sqrt{\frac{\chi_P^2/n}{\sqrt{(R-1)(S-1)}}}$
Dvě ordinální proměnné	Spearmanův koeficient pořadové korelace Goodmanovu-Kruskalovu gamu Kendallovo tau-b a tau-c Somersovo d	Veškeré hodnoty zmíněných koeficientů byly získány spuštěním procedury CROSSTABS za zvolení si vhodné metody ve *Statistics*.

Ordinální vysvětlovaná proměnná	Kruskalův-Wallisův test	Hodnota testu byla získána spuštěním procedury CROSSTABS. Pro zjištění míry by se mohl použít koeficient Cramerovo V nebo Goodmanova-Kruskalova λ.
Dvě kvantitativní proměnné	Spearmanův korelační koeficient Pearsonův korelační koeficientu.	Hodnoty obou koeficientu byly získány za pomocí spuštění procedury CROSSTABS, kde je potřeba zvolit možnost *Correlation* ve *Statistics*. K Pearsonovu korelačnímu koeficientu se lze dopracovat také spuštěním procedury *Bivariate Correlations,* kde je potřeba sekci *Correlation Coefficients* vybrat možnost *Pearson*
Kvantitativní vysvětlovaná proměnná	Koeficient η (éta)	Hodnotu koeficientu pro kvantitativní vysvětlovanou proměnnou je potřeba zvolit si danou metodu ve *Statistics* v proceduře CROSSTABS a poté tuto proceduru spustit.
Dvě dichotomické proměnné	Chí-kvadrát test Pearsonův korelační koeficient Cramerovo V Yuleovo Q Hemannův koeficient Poměr šancí Procentní rozdíl Yuleův koeficient vazby Kendallovo τ_b (tau-b) Somersovo d Fisherův exaktní test McNemarův test	Většina zmíněných koeficientu a výsledky všech zmíněných testů lze získat spuštěním SPSS procedury CROSSTABS, kde je potřeba ve *Statistics* zvolit vhodnou metodu a v případě Fisherova exaktního testu pro malé výběry je potřeba v části *Exact* vybrat vhodný exaktní test. Následující koeficienty bylo potřeba vypočítat za pomocí dosazení do vzorců: Procentní rozdíl $PR_{Y\|X} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}$

		Yuleův koeficient vazby: $Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}}$ Hamannův koeficient $\frac{(n_{11}+n_{22}) - (n_{12}+n_{21})}{n}$
Dvě dichotomické a jedna více kategoriální proměnná	Cochranova statistika Matelova-Haenszelova statistika Mantelův-Haenzelův odhad společného poměru šancí	Spuštěním procedury CROSSTABS při zvolením si položky *Cochran's and Mantel-Haensyel statistics* v části *Statistics,* se ve výstupu zobrazí výsledky všech zmíněných koeficientů a také hodnoty testu homogenity (*Breslow-Day* a *Tarone's*).

Tabulka 6: Vyhodnocení podle typu proměnných

5 Shrnutí a závěr

Cílem diplomové práce bylo jednak zpracování přehledu výpočtových možností měr asociací dle různých typů znaků, popsání jejich vlastností a možností využití. Následně je provedena aplikace vhodných metod nad zvolenými daty z veřejně dostupné databáze za využití softwaru IBM SPSS Statistic a vyhodnocení výsledků výstupů.

Jak bylo zmíněno v úvodu v práce, je důležité nejprve stanovit cíle analýzy, určit proměnné se kterými se pracuje, a poté zvolit vhodnou statistickou metodu. Problematika proměnných je rozebrána v prvních kapitolách, spolu s jejich různými typy, a také způsoby, jak tyto proměnné rozlišit a určit. Následuje vysvětlení problematiky hypotéz a jejich testování, které předpokládá vhodný výběr testu.

Následně se práce podrobněji věnuje popisu testů o nezávislosti a koeficienty měr asociací dle testovaných veličin. Jedná se o testy: (a) dvou nominálních proměnných, kde je vysvětlena statistika chí-kvadrátu a z ní odvozené symetrické koeficienty, které jsou uvedeny s dalšími možnostmi koeficientů jako je například koeficient nejistoty nebo míra souhlasu; (b) dvou ordinálních neznámých, pro které jsou popsány koeficienty symetrické (Spearmanův koeficient pořadové korelace, Goodmanova-Kruskalova γ, Kendallovo τ_b) a asymetrické (Somersovo d); (c) dvou kvantitativních proměnných, pro které lze použít Pearsonův korelační koeficient; (d) dvou dichotomických neznámých, kde je mimo jiné uveden také Fisherův exaktní test pro malé výběry; (e) ordinální vysvětlované proměnné, jejíž nulovou hypotézu lze testovat pomocí Kruskal-Wallis testu; (f) kvantitativní vysvětlované neznámé, pro kterou lze určit sílu asociace za pomocí koeficientu éta; (g) dvou dichotomických a jedné vícekategoriální proměnné, pro které lze použít statistiky Cochranovu a Mantelovu-Haenszelovu a mírou asociace je Mantelův-Haenszelův odhad společného poměru šancí.

V metodice práce je také rozebrán základní pohled na systém IBM SPSS Statistic a jeho základní obrazovky *Data View, Variable View* a *Output.* Obrazovka *Data view* zobrazuje konkrétní hodnoty znaků, které jsou uvedeny ve sloupcích a každý řádek odpovídá odpovědím jednotlivých respondentů (v tomto případě studentů). Tyto znaky a jejich hodnoty jsou pak konkrétněji definovány ve *Variable view*. Obrazovka *Output* zobrazuje jednotlivé výsledky a je automaticky otevřena po spuštění procedury. V případě výsledků této práce je nejčastěji využita procedura CROSSTABS.

Primární část metodiky se zabývá konkrétními výpočtovými metodami jednotlivých testů a měr asociací. Zmíněné metody jsou přejaty z publikace paní Hany Řezankové, *Analýza dat z dotazníkových šetření (2007/2017)*. Jednotlivé metody jsou také podpořeny v teoretické části, kde je více popsán jejich význam a jsou zde odlišeny dle typu dvou neznámých. Toto odlišení je následně použito i ve vlastních výsledcích, kde z dat **PISA 2018 Database** (publikované na webu www.oecd.org) byly zvoleny vhodné znaky určitého typu. Následně na nich byla spuštěna vhodná výpočetní metoda dané procedury.

Kapitola s vlastními výsledky se opírá o metody použité ve zmiňované publikaci paní Hany Řezankové, doplněné o detailnější postup při práci se softwarem IBM SPSS Statistic. Výstupy z SPSS dávají jasné závěry o síle závislosti, která se opírá o znalosti získané z článku Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography* od Khamise (2008). Autor článku zde uvádí tabulku o určování síly lineárního vztahu, kde je interval <-1; 1> rozdělen na podintervaly, kterým je slovně určena síla závislosti vztahu.

Další metodou pro vyhodnocování síly měr asociací byla využita interpretace souvislostí hodnot korelace. Tu uvádí ve své publikaci Mareš, Rabušic a Soukup (2015). Obě jmenované metody interpretací jsou uvedeny v metodice práce, a ve většině případů se výsledky vyhodnotily pomocí kombinace obou metodik.

Pokud se výsledek ve výstupu nacházel v intervalu <0,2; 0,5) byl vztah označen za slabě pozitivní a síla závislosti slabší až střední. V případě intervalu <0,5; 0,8) byl vztah označen za mírně pozitivní a sila závislosti střední až velmi silná. Nastane-li situace kdy se hodnota koeficientu rovná číslu z intervalu <0,8; 1), pak se vztah považuje za silně pozitivní. Síla této závislosti bude považována za velmi silnou až téměř perfektní. Pokud by nastala situace, kdy se hodnota koeficientu rovná 1, jednalo by se o absolutní závislost. V případě intervalu (0; 0,2) je míra asociace velice malá a nedá se uvažovat o zásadní existenci vztahu. Síla asociace v tomto intervalu je ve výstupu práce označovaná za nízkou až triviální. V takovém případě je vhodné analyzovat odchylky, zda asociace nevzniká jen v užší skupině respondentů, tedy jestli se tito respondenti nedají seskupit dle stejné vlastnosti. Pak by byl vztah zajímavý právě v souvislosti s touto vlastností.

Stejné síly závislostí lze použít i na intervaly v záporných hodnotách. V takovém případě by se jednalo o závislost nepřímou, která se při aplikaci metod na zvolených datech se neprojevila.

Na konci práce je uveden souhrn všech metod, které byly použity pro vlastní interpretaci. Souhrn je členěn dle testovaných proměnných, u kterých jsou uvedeny veškeré metody využity při testování, spolu se způsobem, jak bylo hodnot koeficientů a testů dosaženo. Z největší části se hodnoty získaly použitím procedury CROSSTABS, kde je možné zvolit většinu metod.

Výstup práce může sloužit jako podklad pro jiné oborné práce ve kterých by se autor chtěl věnovat výzkumu analýzy souvislostí a zjišťování jejich míry. V práci jsou uvedeny nejen testy a koeficienty pro zkoumání měr asociací, ale také podmínky, které musí zkoumané veličiny splňovat. Pro aplikaci těchto statistických metod byl použit program IBM SPSS Statistics, čímž vznikl také jednoduchý návod pro práci s tímto nástrojem, se zaměřením na statistické metody související s mírou asociací.

6 Literatura

[1] HEBÁK, Petr. *Statistické myšlení a nástroje analýzy dat*. Praha: Informatorium, 2013. ISBN 978-80-7333-105-4.

[2] ŘEZANKOVÁ, Hana. Analýza dat z dotazníkových šetření. (Čtvrté přepracované vydání). Praha: Professional Publishing, 2017. ISBN 978-80-906594-8-3.

[3] ŘEZANKOVÁ, Hana. *Analýza dat z dotazníkových šetření*. (První vydání) Praha: Professional Publishing, 2007. ISBN 978-80-86946-49-8.

[4] RABUŠIC, Ladislav, Petr SOUKUP a Petr MAREŠ. Statistická analýza sociálněvědních dat (prostřednictvím SPSS). 2., přepracované vydání. Brno: Masarykova univerzita, 2019. ISBN 978-80-210-9248-8.

[5] PECINA, Pavel. *LEXICAL ASSOCIATION MEASURES: SURES Collocation Extraction*. Praha, 2009. Dostupné také z: https://ufal.mff.cuni.cz/books/preview/pecina_preview.pdf. Institute of Formal and Applied Linguistics.

[6] KHAMIS, Harry. Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography*. 2008, **24**(3), 155-162. DOI: 10.1177/8756479308317006. ISSN 8756-4793. Dostupné také z: http://journals.sagepub.com/doi/10.1177/8756479308317006

[7] Gonzalez-Chica DA, Bastos JL, Duquia RP, Bonamigo RR, Martínez-Mesa J. Tests of association: which one is the most appropriate for my study? An Bras Dermatol. 2015;90(4):523-8.

[8] GINGRICH, Paul. Introductory Statistics for the Social Sciences: Chapter 11 Association Between Variables [online]. 1993 [cit. 2021-02-21]. Dostupné z: http://uregina.ca/~gingrich/text.htm

[9] NOVÁK, Tomáš a Miloslav KOPEČEK. Typy proměnných. Mladí psychiatři. 2010, 11(4), 176-177. Dostupné také z: https://www.psychiatriepropraxi.cz/pdfs/psy/2010/04/12.pdf

[10] Hypotéza. Stemmark [online]. 2013 [cit. 2020-04-12]. Dostupné z: https://www.stemmark.cz/encyklopedie-hypoteza/

[11] BOŘIL, Tomáš. Lineární regrese a korelace [online]. 2015 [cit. 2020-04-13]. Dostupné z: https://fu.ff.cuni.cz/STATf/18_linearni_regrese_korelace.html

[12] Základy statistiky. Matemetika.cz [online]. Nová media, s.r.o, 2014 [cit. 2020-12-26]. Dostupné z: https://matematika.cz/zaklady-statistiky

[13] HOLČÁK, Lukáš. *Statistická analýza souborů s malým rozsahem* [online]. Brno, 2008 [cit. 2021-5-7]. Dostupné z: http://hdl.handle.net/11012/25225. Diplomová práce. Vysoké učení technické v Brně. Fakulta strojního inženýrství. Ústav matematiky. Vedoucí práce Zdeněk Karpíšek.

[14] ŘEZANKOVÁ, H. a kolektiv: IASTAT - Interaktivní učebnice statistiky. 2001. www.http://iastat.vse.cz/.

[15] Testování nezávislosti (Pearsonův chí-kvadrát test). *Https://portal.matematickabiologie.cz* [online]. Institut biostatistiky a analýz Lékařské fakulty Masarykovy univerzity [cit. 2021-8-12]. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--analyza-a-management-dat-pro-zdravotnicke-obory--testovani-hypotez-o-kvalitativnich-promennych--analyza-kontingencnich-tabulek--testovani-nezavislosti-pearsonuv-chi-kvadrat-test

[16] EVERT, Stefan. Association measures. *Collocations* [online]. Institut biostatistiky a analýz Lékařské fakulty Masarykovy univerzity, 2010 [cit. 2021-8-12]. Dostupné z: http://www.collocations.de/AM/index.html

[17] Association. Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001 [cit. 2021-8-12]. Dostupné z: https://en.wikipedia.org/wiki/Association

[18] Wikisofia [online]. 2013 [cit. 2021-8-12] ISSN 2336-5897. Dostupné z: https://wikisofia.cz/wiki/Porovn%C3%A1v%C3%A1n%C3%AD_%C4%8Detnost%C3%AD

[19] NĚŠPOR, Zdeněk R. Tabulky kontingenční. Encyklopedie.soc.cas [online]. 2017 [cit. 2021-8-12]. Dostupné z: https://encyklopedie.soc.cas.cz/w/Tabulky_kontingen%C4%8Dn%C3%AD

[20] CVRČEK, Václav. Asociační (kolokační) míry. Wiki.korpus [online]. 2019 [cit. 2021-8-12]. Dostupné z: https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry

[21] *Testy korelace* [online]. [cit. 2019-04-26]. Dostupné z: https://wikisofia.cz/wiki/Statistick%C3%A9_testy_korelace#cite_note-analyza-1

[22] PISA 2018 Database [online]. OECD, 2018 [cit. 2021-8-12]. ISSN 2336-5897. Dostupné z: https://www.oecd.org/pisa/data/2018database/

7 Seznam schémat

8 Seznam tabulek

9 Seznam obrázků

10 Seznam výstupů z SPSS

11 Přílohy

			What is the <highest level of schooling> completed by your mother?					
			ISCED level 3A	ISCED level 3B, 3C	ISCED level 2	ISCED level 1	Not compet ISCED level 1	Total
Country code 3-character	Albania	Count	2988	560	2535	122	20	6225
		Expected Count	3393,6	1202,1	1071,1	366,6	191,6	6225,0
	United Arab Emirates	Count	13674	2072	1969	433	528	18676
		Expected Count	10181,4	3606,5	3213,5	1100,0	574,7	18676,0
	Argentina	Count	6421	0	2553	1743	556	11273
		Expected Count	6145,6	2176,9	1939,7	663,9	346,9	11273,0
	Australia	Count	8996	1053	1901	517	110	12577
		Expected Count	6856,5	2428,7	2164,0	740,7	387,0	12577,0
	Austria	Count	2729	3066	484	97	96	6472
		Expected Count	3528,3	1249,8	1113,6	381,2	199,2	6472,0
	Belgium	Count	5089	1749	720	205	168	7931
		Expected Count	4323,7	1531,5	1364,6	467,1	244,1	7931,0
	Bulgaria	Count	2327	1906	634	113	73	5053
		Expected Count	2754,7	975,8	869,4	297,6	155,5	5053,0
	Bosnia and Herzegovina	Count	842	3992	1266	198	44	6342
		Expected Count	3457,4	1224,7	1091,2	373,5	195,2	6342,0
	Belarus	Count	2968	2095	596	49	7	5715
		Expected Count	3115,6	1103,6	983,3	336,6	175,9	5715,0
	Brazil	Count	4667	1032	2053	1294	1154	10200
		Expected Count	5560,6	1969,7	1755,0	600,7	313,9	10200,0
	Brunei Darussalam	Count	5191	509	485	357	175	6717
		Expected Count	3661,8	1297,1	1155,8	395,6	206,7	6717,0
	Canada	Count	19212	1665	202	171	0	21250
		Expected Count	11584,6	4103,6	3656,3	1251,6	653,9	21250,0
	Colombia	Count	1964	2375	1354	899	582	7174

	Expected Count	3911,0	1385,4	1234,4	422,5	220,8	7174,0
Costa Rica	Count	2079	802	1625	2004	620	7130
	Expected Count	3887,0	1376,9	1226,8	419,9	219,4	7130,0
Czech Republic	Count	4712	1306	477	294	12	6801
	Expected Count	3707,6	1313,3	1170,2	400,6	209,3	6801,0
Germany	Count	1590	400	2025	14	233	4262
	Expected Count	2323,5	823,0	733,3	251,0	131,1	4262,0
Denmark	Count	4783	1419	855	130	73	7260
	Expected Count	3957,9	1402,0	1249,2	427,6	223,4	7260,0
Dominican Republic	Count	2408	1070	1211	414	355	5458
	Expected Count	2975,5	1054,0	939,1	321,5	168,0	5458,0
Spain	Count	18597	5154	7690	2296	749	34486
	Expected Count	18800,4	6659,5	5933,8	2031,1	1061,2	34486,0
Estonia	Count	3410	894	781	73	7	5165
	Expected Count	2815,7	997,4	888,7	304,2	158,9	5165,0
Finland	Count	5159	0	227	52	51	5489
	Expected Count	2992,4	1060,0	944,5	323,3	168,9	5489,0
France	Count	3501	1559	711	91	159	6021
	Expected Count	3282,4	1162,7	1036,0	354,6	185,3	6021,0
United Kingdom	Count	6024	4991	670	92	100	11877
	Expected Count	6474,9	2293,6	2043,6	699,5	365,5	11877,0
Georgia	Count	3912	919	549	30	39	5449
	Expected Count	2970,6	1052,2	937,6	320,9	167,7	5449,0
Greece	Count	4517	1093	574	146	21	6351
	Expected Count	3462,3	1226,4	1092,8	374,1	195,4	6351,0
Hong Kong	Count	2508	1249	1142	634	232	5765
	Expected Count	3142,8	1113,3	991,9	339,5	177,4	5765,0

	Croatia	Count	4267	1761	490	22	11	6551
		Expected Count	3571,3	1265,1	1127,2	385,8	201,6	6551,0
	Hungary	Count	3396	1052	532	20	15	5015
		Expected Count	2734,0	968,4	862,9	295,4	154,3	5015,0
	Switzerland	Count	1735	2478	882	311	131	5537
		Expected Count	3018,5	1069,2	952,7	326,1	170,4	5537,0
	Chile	Count	3483	2197	1106	195	172	7153
		Expected Count	3899,5	1381,3	1230,8	421,3	220,1	7153,0
	Indonesia	Count	5214	921	2785	2341	613	11874
		Expected Count	6473,2	2293,0	2043,1	699,3	365,4	11874,0
	Ireland	Count	4068	601	622	101	30	5422
		Expected Count	2955,9	1047,0	932,9	319,3	166,8	5422,0
	Iceland	Count	2238	393	539	0	28	3198
		Expected Count	1743,4	617,6	550,3	188,4	98,4	3198,0
	Israel	Count	4837	924	352	137	105	6355
		Expected Count	3464,5	1227,2	1093,5	374,3	195,6	6355,0
	Italy	Count	5618	2204	3234	174	78	11308
		Expected Count	6164,7	2183,7	1945,7	666,0	348,0	11308,0
	Jordan	Count	5281	846	1887	336	241	8591
		Expected Count	4683,5	1659,0	1478,2	506,0	264,4	8591,0
	Japan	Count	4799	985	178	0	0	5962
		Expected Count	3250,2	1151,3	1025,8	351,1	183,5	5962,0
	Kazakhstan	Count	11578	4940	2753	66	19	19356
		Expected Count	10552,1	3737,8	3330,5	1140,0	595,6	19356,0
	Korea	Count	5233	723	553	93	4	6606
		Expected Count	3601,3	1275,7	1136,7	389,1	203,3	6606,0
	Kosovo	Count	1763	1139	1470	518	99	4989

	Expected Count	2719,8	963,4	858,4	293,8	153,5	4989,0
Lebanon	Count	1781	1315	872	624	452	5044
	Expected Count	2749,8	974,0	867,9	297,1	155,2	5044,0
Lithuania	Count	3485	1672	1222	86	37	6502
	Expected Count	3544,6	1255,6	1118,8	382,9	200,1	6502,0
Luxembourg	Count	2447	796	814	443	259	4759
	Expected Count	2594,4	919,0	818,9	280,3	146,4	4759,0
Latvia	Count	3148	1494	446	37	9	5134
	Expected Count	2798,8	991,4	883,4	302,4	158,0	5134,0
Macao	Count	1556	447	1138	465	145	3751
	Expected Count	2044,9	724,4	645,4	220,9	115,4	3751,0
Morocco	Count	1079	190	1602	973	2643	6487
	Expected Count	3536,4	1252,7	1116,2	382,1	199,6	6487,0
Moldova	Count	2195	1176	1740	75	18	5204
	Expected Count	2837,0	1004,9	895,4	306,5	160,1	5204,0
Mexico	Count	2072	874	2210	930	358	6444
	Expected Count	3513,0	1244,4	1108,8	379,5	198,3	6444,0
North Macedonia	Count	2212	1787	966	184	83	5232
	Expected Count	2852,3	1010,3	900,2	308,1	161,0	5232,0
Malta	Count	1325	436	1415	61	16	3253
	Expected Count	1773,4	628,2	559,7	191,6	100,1	3253,0
Montenegro	Count	2177	3780	429	122	38	6546
	Expected Count	3568,6	1264,1	1126,3	385,5	201,4	6546,0
Malaysia	Count	4380	247	861	342	209	6039
	Expected Count	3292,2	1166,2	1039,1	355,7	185,8	6039,0
Netherlands	Count	2198	0	2111	155	115	4579
	Expected Count	2496,3	884,2	787,9	269,7	140,9	4579,0

Norway	Count	3490	1425	423	39	66	5443
	Expected Count	2967,3	1051,1	936,5	320,6	167,5	5443,0
New Zealand	Count	2637	0	2876	110	53	5676
	Expected Count	3094,3	1096,1	976,6	334,3	174,7	5676,0
Panama	Count	1304	1459	1733	999	432	5927
	Expected Count	3231,2	1144,6	1019,8	349,1	182,4	5927,0
Peru	Count	3350	750	1018	730	177	6025
	Expected Count	3284,6	1163,5	1036,7	354,9	185,4	6025,0
Philippines	Count	3198	1220	1225	998	467	7108
	Expected Count	3875,0	1372,6	1223,0	418,6	218,7	7108,0
Poland	Count	4087	896	489	0	0	5472
	Expected Count	2983,1	1056,7	941,5	322,3	168,4	5472,0
Portugal	Count	2335	703	1471	621	373	5503
	Expected Count	3000,0	1062,7	946,9	324,1	169,3	5503,0
Qatar	Count	7723	2681	1696	590	530	13220
	Expected Count	7207,0	2552,9	2274,7	778,6	406,8	13220,0
Baku (Azerbaijan)	Count	4888	0	1447	101	149	6585
	Expected Count	3589,9	1271,6	1133,0	387,8	202,6	6585,0
B-S-J-Z (China)	Count	4428	2191	3675	1153	535	11982
	Expected Count	6532,1	2313,8	2061,7	705,7	368,7	11982,0
Moscow Region (RUS)	Count	948	700	343	3	2	1996
	Expected Count	1088,1	385,4	343,4	117,6	61,4	1996,0
Tatarstan (RUS)	Count	2742	1821	1092	16	9	5680
	Expected Count	3096,5	1096,9	977,3	334,5	174,8	5680,0
Romania	Count	3420	524	740	254	58	4996
	Expected Count	2723,6	964,8	859,6	294,2	153,7	4996,0
	Count	3762	2245	1393	14	15	7429

	Russian Federation	Expected Count	4050,0	1434,6	1278,3	437,5	228,6	7429,0
	Saudi Arabia	Count	1689	1273	1353	718	846	5879
		Expected Count	3205,0	1135,3	1011,6	346,3	180,9	5879,0
	Singapore	Count	5809	0	324	328	146	6607
		Expected Count	3601,9	1275,9	1136,8	389,1	203,3	6607,0
	Serbia	Count	1525	4466	436	48	20	6495
		Expected Count	3540,8	1254,2	1117,6	382,5	199,9	6495,0
	Slovak Republic	Count	4450	967	339	71	33	5860
		Expected Count	3194,6	1131,6	1008,3	345,1	180,3	5860,0
	Slovenia	Count	1549	4296	414	25	13	6297
		Expected Count	3432,9	1216,0	1083,5	370,9	193,8	6297,0
	Sweden	Count	3618	745	580	90	96	5129
		Expected Count	2796,1	990,5	882,5	302,1	157,8	5129,0
	Chinese Taipei	Count	2886	2874	938	335	81	7114
		Expected Count	3878,3	1373,8	1224,1	419,0	218,9	7114,0
	Thailand	Count	3333	1381	1537	1977	311	8539
		Expected Count	4655,1	1649,0	1469,2	502,9	262,8	8539,0
	Turkey	Count	1580	1079	1519	1936	704	6818
		Expected Count	3716,9	1316,6	1173,1	401,6	209,8	6818,0
	Ukraine	Count	3808	882	1193	22	13	5918
		Expected Count	3226,3	1142,8	1018,3	348,6	182,1	5918,0
	Uruguay	Count	1996	240	1616	1007	132	4991
		Expected Count	2720,9	963,8	858,8	294,0	153,6	4991,0
	United States	Count	4040	0	509	117	77	4743
		Expected Count	2585,7	915,9	816,1	279,3	146,0	4743,0
	Vietnam	Count	1437	178	2079	976	625	5295

UNIVERZITA HRADEC KRÁLOVÉ
Fakulta informatiky a managementu
Akademický rok: 2018/2019

Studijní program: Systémové inženýrství a informatika
Forma studia: Prezenční
Obor/kombinace: Informační management (im2-p)

Podklad pro zadání DIPLOMOVÉ práce studenta

Jméno a příjmení:	**Bc. Marie Brixiová**
Osobní číslo:	**I1800128**
Adresa:	Nová Ves u Bakova 55, Nová Ves u Bakova, 29401 Bakov nad Jizerou, Česká republika
Téma práce:	Míry asociace mezi znaky a možnosti jejich využití.
Téma práce anglicky:	Measures of Association and Their Applications
Vedoucí práce:	prof. RNDr. Hana Skalská, CSc. Katedra informatiky a kvantitativních metod

Zásady pro vypracování:

Cíl práce
Přehled výpočnových možností míry asociace různých typů znkaků, popis jejich vlastností a možností využití.
Pomocí zvolených vhodných typů dat (z veřejně dostupných databází) a SW (např. SPSS Statistic) zpracovat přehledy (vlastnosti, vysvětlení, omezení pro aplikace apod.)
Osnova:

1. Úvod
2. Literární rešerše
3. Metodika
4. Vlastní popis a výsledky
5. Shrnutí a závěry
6. Literatura

Seznam doporučené literatury:

Řezanková: Analýza dat dotazníkových šeření
Petr Hebák a kolektiv: Statistické myšlení a nástroje analýzy dat
HARRY KHAMIS, PhD: Measures of Association How to Choose?

Podpis studenta: Datum:

Podpis vedoucího práce: Datum: