Fakulta informatiky a statistiky

STATISTICKÉ METODY SE ZAMĚŘENÍM NA KATEGORIÁLNÍ DATA

Vysoká škola ekonomická v Praze

Hana Řezanková

Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky

STATISTICKÉ METODY SE ZAMĚŘENÍM NA KATEGORIÁLNÍ DATA

Hana Řezanková

2024

VŠE / NAKLADATELSTVÍ OECONOMICA

Autorka:

prof. Ing. Hana Řezanková, CSc.

Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra statistiky a pravděpodobnosti

Hana Řezanková je absolventkou Vysoké školy ekonomické v Praze, kde získala též titul kandidáta ekonomických věd. V roce 1990 nastoupila na této vysoké škole na katedru statistiky a pravděpodobnosti Fakulty informatiky a statistiky. V letech 1998 až 2001 zastávala na této fakultě funkci proděkanky pro pedagogickou činnost. V roce 2008 byla jmenována profesorkou pro obor statistika. Je dlouholetou členkou České statistické společnosti, ve které v letech 2013 až 2017 zastávala funkci předsedkyně. V pedagogické a vědecko-výzkumné činnosti se prof. Řezanková zaměřuje na analýzu kategoriálních dat a na shlukovou analýzu.

Recenzenti:

doc. Ing. Eva Jarošová, CSc.
doc. Ing. Zdeněk Šulc, Ph.D.

Všechny obrázky, grafy, schémata a tabulky v této publikaci jsou dílem autorky.

ISBN 978-80-245-2521-1

Obsah

Předmluva

Tento učební text se zaměřuje zejména na analýzu datových souborů, v nichž jsou obory hodnot jednotlivých statistických proměnných tvořeny kategoriemi. Metody vícerozměrné analýzy takových dat často vycházejí z dvourozměrných tabulek četností, pro které se používá termín kontingenční tabulky. V nich se zobrazují četnosti pro všechny kombinace kategorií dvou sledovaných proměnných. Pro sledování vztahů více než dvou proměnných se vytvářejí posloupnosti těchto tabulek, případně má tabulka složitější strukturu.

Metody byly navrhovány zejména pro aplikace v biologii a společenských vědách. Používají se například při výzkumu trhu, výzkumu veřejného mínění, zdravotnickém výzkumu, sociálním a demografickém výzkumu. Jedním z cílů předkládané publikace je seznámit čtenáře se způsoby analýzy dat získaných některým z výše uvedených výzkumů. Na příkladu konkrétního průzkumu je naznačena problematika přípravy dotazníku, vysvětlována příprava datového souboru, vyhodnocení odpovědí na jednotlivé otázky a zjišťování vztahů mezi odpověďmi na zadané otázky.

Publikace je určena jako učební pomůcka k předmětu Statistické metody 2, v němž studenti analyzují data s využitím statistického programového systému SPSS. V textu jsou zařazeny výstupy z produktu IBM SPSS Statistics (verze 24), současně je uveden i návod k používání některých metod obsažených v této verzi programu. V jednodušších případech jsou uváděny též výsledky získané bez počítače, tj. dosazením do vzorců.

Text vychází z mé knihy Analýza dat z dotazníkových šetření, která vyšla ve čtyřech vydáních v nakladatelství Professional Publishing (poslední vydání [35]). Vzhledem k tomu, že jsou vysvětleny principy uvedených metod, mohou publikaci využít také ti, kteří pro analýzu používají jiné programy. Text je vhodný rovněž pro uživatele systému SPSS zaměřujících se na jiné typy aplikací, protože popisuje jednak obecnou přípravu datového souboru, jednak obvyklou strukturu výstupů.

Publikace se nejprve věnuje analýze jednotlivých proměnných. Poté jsou popsány principy analýzy závislostí dvou proměnných na základě kontingenční tabulky, naznačena je též analýza pro tři proměnné. Následuje kapitola o porovnávání souborů obsahujících např. hodnoty kvantitativních znaků nebo ordinálních znaků, které představují pořadí. Porovnávají se buď různé proměnné, nebo skupiny hodnot jedné proměnné vytvořené na základě kategorií druhé proměnné.

Závěrečné kapitoly jsou zaměřeny na komplexnější přístupy k analýze dat. Z velkého množství metod, které byly v minulosti navrženy pro vícerozměrnou statistickou analýzu a implementovány do programových systémů, byly vybrány některé postupy pro modelování jednostranné závislosti proměnných a pro zjišťování vztahů mezi kategoriemi. V prvním případě jsou hodnoty vysvětlované proměnné vyjadřovány pomocí jedné či více vysvětlujících proměnných. Metody zde zastupují klasifikační stromy a logistická regrese.

V druhém případě jde o grafické metody znázorňování vztahů mezi kategoriemi jedné či více proměnných. Uvedeny jsou principy shlukové analýzy, vícerozměrného škálování a korespondenční analýzy.

Pro úplné pochopení zařazených postupů by měl být čtenář seznámen se základy teorie pravděpodobnosti a měl by mít alespoň elementární znalosti z matematické statistiky. I když jsou všechny potřebné pojmy vysvětleny, místy je výklad zjednodušen a je určen spíše k připomenutí principů, na nichž je analýza dat založena. Některé učebnice statistiky obsahující výklad potřebné látky jsou uvedeny v seznamu literatury.

Publikace se soustřeďuje na klasické postupy, které vycházejí z prostého náhodného výběru. Při terénních průzkumech jsou však obvykle používány jiné typy výběrů. Uvedené metody jsou tedy úvodem do analýzy. Pro správnou interpretaci výsledků v případě jiných výběrů je nutné využít speciální nástroje. Jedním z nich je modul Complex Samples systému IBM SPSS Statistics, mezi jehož široké možnosti patří například posouzení vlivu výběru na výsledky.

Text byl původně určen pro knižní podobu a černobílý tisk. Z toho důvodu bylo při tvorbě základních grafů použito šrafování, které je nyní nahrazeno barvami. V elektronické verzi byla dále doplněna tato předmluva a opraveno několik typografických chyb. V publikaci nejsou zařazeny hypertextové odkazy ani v rámci textu, ani na použité zdroje. Doufám, že i při těchto omezeních bude elektronické vydání publikace užitečným pomocníkem při studiu a aplikaci statistických metod.

Hana Řezanková
V Praze dne 21. 5. 2024

Navržení dotazníku

1

S dotazníky se můžeme setkat při různých příležitostech. Studenti pomocí nich hodnotí výuku předmětů, absolventi vysokých škol charakterizují své uplatnění na trhu práce, hosté v hotelích jsou dotazováni na spokojenost z různých hledisek. Navržení dotazníku ovšem není triviální záležitost, jak se na první pohled zdá. V praxi někdy není zahrnuta možnost, že respondent na určitou otázku nezná odpověď nebo není ochoten odpovědět. Jsou známy situace, kdy v elektronickém dotazníku respondent musel odpovědět, aby mohl pokračovat ve vyplňování, a přitom v nabídkách nebyly obsaženy odpovědi typu „nevím“, „jinak“ apod.

V této kapitole se zaměříme na vztah mezi možnými odpověďmi a daty, která jsou zaznamenávána do počítače. Nezbytnou součástí přípravy dotazníkového šetření je také stanovení počtu, případně struktury respondentů (z hlediska věku, pohlaví apod.), jejichž odpovědi mají být předmětem analýzy. Následující text je pouze úvodem do dané problematiky, pro profesionální práci je nutné podrobnější studium. Kromě základních učebnic statistiky lze doporučit například literaturu [29] a [36].

1.1 Formulace dotazů

Každému šetření musí předcházet *formulace* jeho *cílů*. Například cílem může být zjištění, jaké hodnoty kladou lidé na přední místo svého hodnotového žebříčku, zda to závisí na náboženském vyznání, nebo na sociálně ekonomickém postavení. V některých případech se zkoumá, zda jsou vůbec určité informace dosažitelné. K tomu se používá tzv. pilotní studie, kdy se nejčastěji na malém výběru realizuje nestandardizovaný rozhovor. Pak se formulují dotazy pro respondenty, týkající se například důležitosti různých faktorů v životě (práce, rodiny, volného času), náboženského vyznání či sociálně ekonomického postavení. Otázky můžeme rozčlenit do dvou základních skupin:

- otázky týkající se názorů a chování respondentů,
- otázky za účelem získání jiných údajů, například demografických.

V první skupině jsou obsaženy především otázky zaměřené na zkoumaný problém, které se nazývají *meritorní*. Kromě nich se mohou vyskytovat otázky pomocné (kontaktní a větvící) a kontrolní. Druhá skupina obsahuje otázky *analytické* (třídicí a identifikační). V dotaznících se obvykle tyto otázky zařazují doprostřed nebo na konec. Není tedy vhodné, aby první dotaz zněl: „Jaký je Váš rodinný stav?“

Velice důležité je, aby otázky byly formulovány srozumitelně a jednoznačně. Respondentovi jsou u většiny dotazů nabízeny varianty odpovědí. Tyto otázky nazýváme *uzavřené.*

V rámci nich rozlišujeme otázky *alternativní* (nabízejí dvě varianty) a *selektivní* (více než dvě varianty). Je nutné, aby byly zahrnuty všechny možné odpovědi, aby tyto odpovědi byly jednoznačné a aby se nepřekrývaly.
Často proto bývá zařazena odpověď „nevím", „jiné" apod. Výjimkou jsou dotazy, které umožňují výběr většího počtu variant („Vyberte maximálně tři politiky, kteří jsou Vám nejsympatičtější."). V tom případě se odpověď respondenta nazývá *vícehodnotová* (jejím zpracováním se zabývá oddíl 3.1.3).

Kromě uzavřených otázek se v dotazníku mohou vyskytovat otázky *otevřené*. Škála hodnot se pak vytváří dodatečně na základě odpovědí respondentů. Existuje i možnost kombinovat uvedené dva typy a nabízet *polouzavřené* (polootevřené) otázky. Respondent si tak může buď vybrat některou z nabízených variant odpovědí, nebo uvést svou odlišnou variantu.

U odpovědí by měla být zajištěna jejich *validita* (co nejvěrnější zachycení skutečnosti) a *reliabilita* (spolehlivost), kterou můžeme charakterizovat jako opakovatelnost za stejných podmínek.

Vedle *přímých* otázek („Co se Vám nelíbí na městské hromadné dopravě?") lze z psychologických důvodů zařazovat i otázky *nepřímé* („Co se Vašim kolegům nelíbí na MHD?"), u nichž je ovšem třeba kontrolovat validitu. Sociologové v literatuře uvádějí ještě podrobnější členění dotazů dle způsobů jejich formulace. Můžeme se setkat například s pojmy *dokončovací* otázky, kdy má respondent dokončit naznačený výrok či dialog, nebo otázky *dialogové*, kdy se má respondent přiklonit k některému z nabízených dialogů. V určitých případech se doporučuje přidat k otázce komentář (vysvětlení) – pak se hovoří o otázkách *psychotaktických*.

Pozornost by měla být věnována také *pořadí dotazů*, jak již bylo naznačeno výše. Je vhodné, aby existoval logický sled otázek. Na druhé straně je třeba dbát, aby předchozí otázky neovlivňovaly odpovědi na otázky následující.

Kromě formulace dotazů a jejich pořadí je důležitá též *grafická úprava dotazníku*. V této kapitole se nebudeme zabývat elektronickým dotazníkem, kdy respondent či tazatel zaznamenává odpovědi přímo do počítače. Grafická úprava klasického dotazníku závisí na tom, zda budou odpovědi převáděny do počítače pomocí skeneru nebo zda budou vkládány pomocí klávesnice. První případ lze použít pro dotazníky, které jsou založeny na uzavřených otázkách. Výběry jsou zaznamenávány například pomocí křížků do předtištěných čtverečků, které mohou být umístěny buď vedle odpovědí, nebo na zvláštním listu spolu s kódy odpovědí.

Není-li použit skener, měl by být dotazník opatřen kódy, které budou vkládány do počítače. Jednou z možností je vytisknout vedle čtverečku (kroužku), který slouží pro vyznačení odpovědi, příslušný kód odpovědi.

Spolu s dotazníkem by měla být navrhována i struktura datového souboru (názvy a typy proměnných, škála hodnot, označení chybějících údajů). Dodatečné definování datového souboru je časově mnohem náročnější. Je třeba zohlednit, jaké postupy budou aplikovány a jaký programový systém k tomu bude využit. Některé systémy totiž pro některé operace vyžadují, aby hodnoty byly číselné, i když se jedná o kódy.

Jako příklad si uveďme několik uzavřených dotazů, včetně grafické úpravy.

1. Kde jste naposledy trávil(a) tu část dovolené, při níž jste si opravdu odpočinul(a)?

   1 ☐ na chatě (chalupě) v ČR
   2 ☐ jinde v ČR (mimo domov)
   3 ☐ mimo ČR, ale v Evropě
   4 ☐ mimo Evropu
   8 ☐ jiná odpověď (doma, neodpočívám, nemám dovolenou)

2. Celkově považujete svůj život za:

   1 ☐ velmi šťastný
   2 ☐ celkem šťastný
   3 ☐ spíše nešťastný
   4 ☐ velmi nešťastný
   8 ☐ ani šťastný, ani nešťastný

3.–5. Jaké je nejvyšší dosažené vzdělání

	Vaše,	Vašeho otce,	Vaší matky?
základní	☐ 1	☐ 1	☐ 1
vyučen, střední bez maturity	☐ 2	☐ 2	☐ 2
střední s maturitou	☐ 3	☐ 3	☐ 3
vysokoškolské	☐ 4	☐ 4	☐ 4

Seskupení dotazů 3 až 5 se nazývá *baterie otázek*. Na každý dotaz se předpokládá jedna odpověď. Do programového systému budou vkládány číselné kódy, uvedené u políček pro zaškrtnutí. Tyto kódy budou zaznamenány do proměnných, nazvaných například jako P1 až P5. U dotazů 1 a 5 jsou odpovědi mimo základní nabídku označeny kódem 8. Pokud respondent na některý dotaz neodpoví, může být do počítače vloženo číslo nevyužité pro konkrétní odpovědi, například 0 (o této problematice bude pojednáno v kapitole 2).

1.2 Škály odpovědí

Odpovědi respondentů jsou hodnotami z určité škály. Podle typů vztahů, které lze zjišťovat mezi hodnotami, rozlišujeme škály nominální, ordinální, intervalové a poměrové. Tato problematika bude podrobněji rozebrána v oddílu 2.1, zde budou naznačeny pouze základní principy jejich rozlišení. Dále můžeme škály klasifikovat podle cíle zjišťování (preferenční, hodnoticí), případně podle jejich formy (slovní, číselné, grafické).

U dotazů 1 až 5, uvedených v předchozím oddílu, představují nabízené odpovědi *slovní* škálu. Mnohem důležitější je však rozlišení otázek podle vztahů mezi odpověďmi. U otázky 1 mohou být odpovědi uspořádány zcela jinak („v Evropě mimo ČR", „v ČR na chatě", „v ČR, ale ne na chatě" atd.). I když uvedené uspořádání má určitý logický sled, nelze jednoznačně stanovit pořadí hodnot. Takovou škálu označujeme jako *nominální*.

U otázek 2 až 5 je pro odpovědi označené kódy od 1 do 4 již uspořádání zřejmé (u otázky 2 lze použít též opačné pořadí). Odpovědi tohoto typu označují určitou úroveň (štěstí, vzdělání) a jsou hodnotami z *ordinální* škály.

Kromě nominální a ordinální škály rozlišujeme škálu *kvantitativní* (číselnou), která je typická především pro otázky otevřené; je například výsledkem otázky na věk respondenta. Kromě toho, že můžeme říci, který respondent je starší (mladší) než jiní, můžeme též spočítat, o kolik let je určitý respondent starší (mladší) než jiný, případně kolikrát (jde o speciální typ kvantitativní škály, která se nazývá *poměrová*).

Při sestavování dotazníku je potřeba řešit celou řadu dalších otázek. Například u *preferenční* škály je důležité vhodně stanovit počet nabízených odpovědí. Má být počet základních kategorií lichý, nebo sudý? Má škála obsahovat neutrální bod (např. odpověď „smíšené pocity")? Má, či nemá být škála k tomuto bodu centrovaná?

Při hodnocení (výrobku, služeb) je možné použít *bodovací* nebo *známkovací škály*, k nimž lze připojit slovní hodnocení. Je třeba dobře zvážit, zda je výhodnější lepšímu ohodnocení přiřadit vyšší hodnotu (počet bodů), nebo nižší (známku). Druhý případ by měl být omezen na jevy, které souvisejí se školou (hodnocení výuky). Hodnoticí škála by měla obsahovat neutrální střed a měla by být k tomuto středu centrovaná. Nejvhodnějšími jsou pětibodové a sedmibodové stupnice.

Kromě číselných a slovních škál mohou být v dotazníku použity i *grafické škály* (respondent například vyznačí bod na úsečce vymezené minimem a maximem). Jejich cílem je snaha usnadnit respondentovi jeho vyjádření. Takto získané odpovědi zpracovatel dotazníku obvykle převádí na ordinální škálu.

Při velkých výzkumech by měl být vždy proveden tzv. *předvýzkum*, kdy se zjišťuje, zda jsou vhodně formulovány dotazy a zda jsou vhodně stanoveny škály. Mohou být zahrnuty i otevřené dotazy, u nichž nejsou nabízeny odpovědi. Na základě volných odpovědí lze pak navrhnout uzavřené dotazy.

Není dobré, když respondenti nejčastěji vybírají odpověď „nevím", nebo když naopak nevybírají některé odpovědi. Je tedy především vhodné zjistit rozdělení četností odpovědí na jednotlivé otázky a podle potřeby dotazník upravit. Kromě logické úvahy (vyřadit otázku, jestliže na ni odpovídá 90 % respondentů shodně) existují pro tyto účely též speciální metody, viz [27] a [29].

1.3 Příklad dotazníku

Jako příklad si uveďme část dotazníku inspirovaného šetřením absolventů vysokých škol REFLEX 2010[1] (níže uvedené dotazy obsahově pokrývají některé problematiky zkoumané v daném šetření, formulace jsou však odlišné). Předpokládejme, že tento dotazník vyplnili náhodně vybraní absolventi jisté vysoké školy, kteří ukončili magisterské studium před určitým počtem let. Nabídky odpovědí na některé dotazy byly redukovány na základě odpovědí získaných v předvýzkumu.

Dotazník

A Zaměstnání před studiem a v průběhu studia a informace o studiu

A1. Byl(a) jste před vstupem na vysokou školu zaměstnán(a)? (označte vše relevantní)

- ☐ ano, práce souvisela s mým následným studiem, trvala přibližně …… měsíců
- ☐ ano, práce nesouvisela s mým následným studiem, trvala přibližně …… měsíců
- ☐ ne

A2. Byl(a) jste v průběhu studia na vysoké škole zaměstnán(a)? (označte vše relevantní)

- ☐ ano, práce souvisela s mým studiem, trvala přibližně …… měsíců
- ☐ ano, práce nesouvisela s mým studiem, trvala přibližně …… měsíců
- ☐ ne

A3. Typ Vašeho magisterského studia byl:

1 magisterský dlouhý
2 magisterský navazující

B Hodnocení fakulty a studijního oboru

B1. Fakulta, kterou jste absolvoval(a), patřila podle Vás ve srovnání s jinými fakultami mezi:

1 vynikající
2 nadprůměrné
3 průměrné
4 podprůměrné

B2. Zhodnoťte (oznámkujte jako ve škole), jak byl Váš studijní obor přínosný pro:

a	vstup do práce	1	2	3	4	5
b	další učení v rámci práce	1	2	3	4	5
c	zvládání pracovních úkolů	1	2	3	4	5
d	osobní rozvoj	1	2	3	4	5
e	rozvoj podnikatelských schopností	1	2	3	4	5

1 REFLEX 2010: Zaměstnatelnost a uplatnění absolventů vysokých škol na pracovním trhu [online]. Praha: SVP Univerzita Karlova. [cit. 2016-03-28]. http://www.strediskovzdelavacipolitiky.info/default.asp?page=svp&KID=85

B3. Kdybyste se vrátil(a) v čase a měl(a) jste si vybrat studijní obor, pak (se zohledněním získaných zkušenosti se studiem) byste si vybral(a):
1 stejný studijní obor
2 jiný studijní obor na stejné vysoké škole
3 stejný či obdobný obor na jiné vysoké škole
4 jiný studijní obor na jiné vysoké škole
5 žádný studijní obor (nešel/nešla bych studovat)

C První zaměstnání a profesní historie

C1. Kdy jste nastoupil(a) do zaměstnání, kde jste působil(a) po absolvování vysoké školy?
1 před studiem vysoké školy nebo v průběhu studia vysoké školy
2 po absolvování vysoké školy; práci jsem si našel/našla po ……. měsících
3 dosud nepracuji (pokračujte částí E)

C2. Jaký typ pracovní smlouvy jste měl(a) v hlavním zaměstnání po absolvování VŠ?
1 na dobou neurčitou
2 na dobu určitou
3 byl(a) jsem pouze osoba samostatně výdělečně činná (OSVČ)

C3. Jaký studijní obor považujete za vhodný pro Vaše první (hlavní) zaměstnání?
1 vystudovaný obor
2 příbuzný studijní obor
3 zcela jiný studijní obor
4 zaměstnání nevyžaduje oborovou specializaci

C4. V kolika zaměstnáních (včetně samostatné výdělečné činnosti) jste od absolvování studia pracoval(a)? ……

C5. Jste v současné době v placeném zaměstnání (včetně samostatně výdělečné činnosti)?
1 ano
2 ne (pokračujte částí E)

D Současné (placené) zaměstnání

D1. Působíte stále v prvním zaměstnání?
1 ano (pokračujte od D4)
2 ne

D2. Jaký typ pracovní smlouvy máte v současném hlavním zaměstnání?
1 na dobou neurčitou
2 na dobu určitou
3 jsem pouze osoba samostatně výdělečně činná (OSVČ)

D3. Jaký studijní obor považujete za vhodný pro Vaše současné hlavní zaměstnání?
1 vystudovaný obor
2 příbuzný studijní obor
3 zcela jiný studijní obor
4 zaměstnání nevyžaduje oborovou specializaci

D4 Jakého typu je instituce, kde pracujete?
1 instituce ve veřejném sektoru
2 soukromá nezisková organizace
3 soukromá komerční společnost
4 instituce jiného typu

D5. Řídíte jiné pracovníky?
1 řídím
2 neřídím

D6. S Vaší současnou prací jste:
1 velmi spokojen(a)
2 spíše spokojen(a)
3 napůl spokojen(a)
4 spíše nespokojen(a)
5 velmi nespokojen(a)

E Demografické údaje

E1. Jste:
1 muž
2 žena

E2. Máte děti?
0 ne
1 ano uveďte počet dětí **…….**

E3. Uveďte nejvyšší dosažené vzdělání rodičů:

	otec	matka
1 bez maturity	☐	☐
2 středoškolské s maturitou	☐	☐
3 vysokoškolské	☐	☐
9 nevím	☐	☐

Vytvoření datového souboru

2

Nejsou-li při dotazníkovém šetření data zaznamenávána přímo do počítače, pak je třeba pro jejich vyhodnocení převést získané odpovědi do elektronické podoby. Odpovědi mohou být skenovány nebo jejich kódy ručně vkládány do *tabulky* určitého programového systému (databázového či statistického, případně do tabulkového procesoru). V některých případech může být vložení do tabulky provedeno zprostředkovaně pomocí elektronického formuláře.

Řádky tabulky jsou vymezeny pro odpovědi jednotlivých respondentů, sloupce obsahují odpovědi na jednotlivé otázky, resp. na jejich části. Řádky jsou ve statistických programových systémech nejčastěji označovány jako *případy* (*cases*), vyskytuje se též termín *pozorování* (*observation*); v databázových systémech se používá pojem *záznam*.

Cílem otázky je zjistit hodnotu určitého *statistického znaku*. Ve statistických programových systémech se používá ekvivalentní termín *proměnná* (*variable*). Zjištěné hodnoty proměnné jsou zaznamenávány do určitého sloupce tabulky, který se v databázových systémech označuje jako *pole* (můžeme se setkat též s vhodnějším pojmem *položka*). Nejčastější terminologii statistických programových systémů znázorňuje schéma 2.1.

Schéma 2.1 | Struktura základní datové matice

	1. proměnná	2. proměnná	…
1. případ			
2. případ			
…			

2.1 Typy proměnných

Odpovědím na uzavřené otázky jsou přiřazeny buď slovní, nebo číselné kódy. Odpovědi na otevřené otázky jsou zaznamenávány v původní podobě. Pro kódované odpovědi, případně i celočíselné hodnoty, se používá pojem *kategorie*. Proměnné, jejichž hodnotami jsou kategorie, se nazývají *kategoriální*. Jako příklady lze uvést:

- národnost (česká, slovenská, ...),
- úroveň vzdělání (základní, středoškolské, vysokoškolské),
- počet dětí (0, 1, 2, 3, 4, …).

Uvedené příklady zároveň ilustrují různou úroveň vztahů mezi kategoriemi – v prvním případě kategorie nelze uspořádat, ve druhém je můžeme uspořádat, ve třetím navíc můžeme vypočítat rozdíl. V tomto smyslu můžeme hovořit o škálách měření, jejichž základní dělení je následující:

- *škála nominální*, u jejíchž hodnot můžeme pouze určit, že jsou různé, nemůžeme stanovit jejich pořadí,
- *škála ordinální*, u jejíchž hodnot můžeme stanovit pořadí, nemůžeme však určit, o kolik je jedna hodnota větší či menší než druhá,
- *škála intervalová*, u jejíchž hodnot můžeme určit, o kolik je jedna hodnota větší či menší než druhá (jde o číselné hodnoty),
- *škála poměrová*, u jejíchž hodnot můžeme určit, o kolik i kolikrát je jedna hodnota větší než druhá (škála obsahuje pouze kladné hodnoty).

Každá odpověď respondenta musí být zaznamenána do samostatné proměnné. Je-li na určitý dotaz přípustná maximálně jedna odpověď, odpovídá tomuto dotazu jedna proměnná. Pokud je jich přípustných více, musí být rezervován potřebný počet proměnných. V případě polouzavřených (polootevřených) otázek se postupuje tak, že se na jednu otázku pohlíží jako na dvě. První je uzavřená (odpovědím jsou přiřazeny kódy) a druhá otevřená (odpovědi jsou zaznamenávány v původní podobě).

Typ škály je základem pro dělení proměnných na:

- **nominální,** např. typ absolvované střední školy, typ profese, druh výrobku,
- **ordinální** (*pořadové*), např. stupeň znalostí (klasifikace ve škole), dosažený stupeň vzdělání, stupeň důležitosti určitého faktoru, stupeň souhlasu s určitým výrokem, stupeň spokojenosti,
- **kvantitativní**, které obvykle dále členíme na:
  - *intervalové*, např. teplota udávaná ve stupních (má smysl rozdíl hodnot),
  - *poměrové*, např. počet členů domácnosti (do této skupiny jsou zařazovány proměnné buď s kladnými hodnotami, nebo u nichž nula znamená „žádný“).

  Podle jiného hlediska rozlišujeme kvantitativní proměnné
  - *diskrétní*, nabývající izolovaných, většinou celočíselných hodnot (počet počítačů v domácnosti) a
  - *spojité*, které mohou nabýt libovolných hodnot z určitého intervalu reálných čísel (věk respondenta, plocha bytu, měsíční výdaje domácnosti za potraviny).

Při určování typu proměnných se můžeme dále setkat s pojmem *kvalitativní*. V literatuře se vyskytuje dvojí výklad tohoto termínu. První přístup ztotožňuje uvedený typ proměnné s typem nominálním, druhý pod pojmem kvalitativní zahrnuje jak nominální, tak ordinální škálu.

Zvláštním typem je proměnná **dichotomická** (*alternativní*), která nabývá pouze dvou kategorií. Jako příklad lze uvést dvojice spokojen – nespokojen, ekonomicky aktivní – ekonomicky neaktivní, kuřák – nekuřák. (Pokud bude v dalším textu potřeba odlišit proměnné s více než dvěma kategoriemi, bude pro ně v dalším textu používán pojem *vícekategoriální*.)

U dichotomických proměnných můžeme dále rozlišit proměnné

- *symetrické*, které mají obě kategorie stejné důležitosti (muž, žena), a
- *asymetrické*, jejichž jedna kategorie je důležitější (pacient se uzdravil).

Při výpočtech se předpokládá, že jde o proměnné *binární*, nabývající hodnot 0 a 1 (např. číslo 1 znamená „ekonomicky aktivní" a číslo 0 „ekonomicky neaktivní").

Kromě kategorií, kterými jsou kódovány přípustné odpovědi, mohou proměnné obsahovat speciální kategorie pro zaznamenání informace, že údaj nebyl zjištěn (respondent neodpověděl, nebo byl vložen chybný údaj a nelze zjistit správný apod.). V případě číselného kódování kategorií jsou to obvykle hodnoty 0, 9, 999 aj., které nemají význam čísla a nelze s nimi tedy provádět aritmetické operace. Pro tyto hodnoty se používá název *chybějící údaje* (anglicky *missing values*).

Jako chybějící údaje můžeme kromě nezjištěných údajů někdy definovat i kategorie, které označují odpovědi mimo základní nabídku („platných") odpovědí (odpovědi typu „nevím", „jinak" apod.). U hodnotících ordinálních škál, které bývají maximálně sedmibodové, se v některých výzkumech (např. sociologických) používá číselný kód 8. Speciální kódy lze přiřadit též kategoriím, které jsou málo zastoupeny, či jsou pro sledování určitého problému nepodstatné. Může jít o extrémy, které by mohly zkreslit hodnoty některých charakteristik.

Rozlišujeme:

- *uživatelem definované chybějící údaje* (například kódy 0, 9, 999),
- *systémové chybějící údaje*, které vznikají především
  - při vstupu dat (nebyla zadána žádná hodnota, nebo byla vložena nepřípustná hodnota),
  - jako výsledky výpočtů, které nelze uskutečnit (dělení nulou).

Pro uživatelem definované chybějící údaje můžeme v některých případech zadat způsob zpracování (například zda mají být zahrnuty do grafu četností, či nikoli).

Údaje, které se zaznamenávají do počítače, jsou nejčastěji následujících typů:

- *číslo* (hodnoty kvantitativní a obvykle též ordinální proměnné, případně číselné kódy hodnot nominální proměnné),
- *datum* (speciální typ intervalových dat, s nímž se obvykle provádějí operace pouze jako s ordinálními daty),
- *řetězec znaků* (hodnoty nominální proměnné, především u otevřených otázek).

Kódy nominální proměnné se ve statistických programových systémech většinou zadávají jako číselné kódy, které ale nemají význam čísla, a tudíž s nimi neprovádíme aritmetické operace. Hodnoty ordinální proměnné se zadávají jako čísla z důvodu, aby bylo možno stanovit pořadí kategorií.

2.2 Popis proměnných a jejich hodnot

Ve výstupu 2.1 je uvedena ukázka popisu proměnných a jejich hodnot, které odpovídají odpovědím z dotazníku uvedeného v oddílu 1.3. Číselné kódy jsou však v některých případech změněny – např. hodnotě „ne“ při dvou možnostech odpovědí je přiřazeno číslo „0“ a pro ordinální proměnné platí, že nižší hodnota znamená nižší úroveň hodnocení. Navíc byla z dostupných údajů vytvořena proměnná *D0*.

Jde o výstup získaný pomocí systému *IBM SPSS Statistics*, dostupný od verze 17 pomocí procedury CODEBOOK (pro zadání požadavků uživatel volí nabízené položky *Analyze*, *Reports*, *Codebook*, nastaveny jsou pouze možnosti *Label*, *Measurement level* a *Value labels*, dostupné v listu *Output*). V tabulce 2.1 je uveden překlad základních anglických výrazů generovaných systémem.

Tabulka 2.1 | Překlad anglických výrazů uvedených ve výstupu 2.1

Anglický výraz	Český překlad
Value	Hodnota
Standard Attributes	Standardní atributy
Label	Popis (proměnné)
Measurement	Škála měření
Scale	Kvantitativní škála
Ordinal	Ordinální
Nominal	Nominální
Valid Values	Platné hodnoty

U některých proměnných byly zadány číselné hodnoty pro chybějící údaje. Ve výstupu 2.1 však tyto kódy nejsou uvedeny (vypisovaly by se pomocí možnosti *Missing values* a zobrazovaly na konci příslušné tabulky).

V systému *IBM SPSS Statistics* (ve verzi 24) se při vkládání řetězců znaků, které netvoří čísla, automaticky nastavuje nominální škála měření, při vkládání čísel se nastavuje neznámá (*unknown*) škála. Před použitím číselné proměnné je většinou třeba zadat konkrétní škálu, neboť systém po zadání statistické procedury či jiných výpočtů změní neznámou škálu na škálu nominální.

Na obrázcích 2.1 a 2.2 jsou ukázky dvou listů datového editoru systému *IBM SPSS Statistics*. List *Variable View* (obr. 2.1) slouží k popisu proměnných a jejich hodnot a k zobrazení tohoto popisu. Slovní popisy kódů, které si pro jednotlivé proměnné vytváří volitelně uživatel v tomto listu, usnadní identifikaci odpovědi, které je v datovém souboru přiřazen určitý číselný kód. List *Data View* (obr. 2.2) je určen ke vkládání a zobrazení zjištěných hodnot proměnných; obsahuje tedy vlastní data.

Výstup 2.1 | Popis proměnných a jejich kategorií

A11		Value
Standard Attributes	**Label**	Práce před studiem související se studiem
	Measurement	Nominal
Valid Values	**0**	Ne
	1	Ano

A11pm		Value
Standard Attributes	**Label**	Práce před studiem související se studiem – počet měsíců
	Measurement	Scale

A12		Value
Standard Attributes	**Label**	Práce před studiem nesouvisející se studiem
	Measurement	Nominal
Valid Values	**0**	Ne
	1	Ano

A12pm		Value
Standard Attributes	**Label**	Práce před studiem nesouvisející se studiem – počet měsíců
	Measurement	Scale

A21		Value
Standard Attributes	**Label**	Práce při studiu související se studiem
	Measurement	Nominal
Valid Values	**0**	Ne
	1	Ano

A21pm		Value
Standard Attributes	**Label**	Práce při studiu související se studiem – počet měsíců
	Measurement	Scale

Výstup 2.1 | Popis proměnných a jejich kategorií – pokračování

A22		Value
Standard Attributes	**Label**	Práce při studiu nesouvisející se studiem
	Measurement	Nominal
Valid Values	**0**	Ne
	1	Ano

A22pm		Value
Standard Attributes	**Label**	Práce při studiu nesouvisející se studiem – počet měsíců
	Measurement	Scale

A3		Value
Standard Attributes	**Label**	Typ studijního programu
	Measurement	Nominal
Valid Values	**1**	Magisterský dlouhý
	2	Magisterský navazující

B1		Value
Standard Attributes	**Label**	Hodnocení studované fakulty
	Measurement	Ordinal
Valid Values	**1**	Podprůměrné
	2	Průměrné
	3	Nadprůměrné
	4	Vynikající

B2a		Value
Standard Attributes	**Label**	Přínos oboru pro vstup do práce
	Measurement	Ordinal
Valid Values	**1**	1 vůbec ne
	2	2
	3	3
	4	4
	5	5 ve velké míře

Výstup 2.1 | Popis proměnných a jejich kategorií – pokračování

Obdobně **B2b** až **B2e.**

B3		Value
Standard Attributes	**Label**	Opětovný výběr oboru
	Measurement	Nominal
Valid Values	**1**	Stejný studijní obor
	2	Jiný studijní obor na stejné vysoké škole
	3	Stejný či obdobný obor na jiné vysoké škole
	4	Jiný studijní obor na jiné vysoké škole
	5	Žádný studijní obor

C1		Value
Standard Attributes	**Label**	Nástup do zaměstnání
	Measurement	Nominal
Valid Values	**1**	Před studiem nebo v průběhu studia
	2	Po absolvování vš
	3	Dosud nepracuji

C1pm		Value
Standard Attributes	**Label**	Počet měsíců do nalezení práce
	Measurement	Scale

C2		Value
Standard Attributes	**Label**	Typ smlouvy v prvním zaměstnání
	Measurement	Nominal
Valid Values	**1**	Na dobu neurčitou
	2	Na dobu určitou
	3	Pouze OSVČ

C3		Value
Standard Attributes	**Label**	Studijní obor vhodný pro první zaměstnání
	Measurement	Nominal
Valid Values	**1**	Vystudovaný obor
	2	Příbuzný studijní obor
	3	Zcela jiný studijní obor
	4	Zaměstnání nevyžaduje oborovou specializaci

Výstup 2.1 | Popis proměnných a jejich kategorií – pokračování

C4		Value
Standard Attributes	**Label**	Počet zaměstnání od absolvování studia
	Measurement	Scale

C5		Value
Standard Attributes	**Label**	Současné placené zaměstnání
	Measurement	Nominal
Valid Values	**0**	Nejsem zaměstnán(a)
	1	Jsem zaměstnán(a)

D0		Value
Standard Attributes	**Label**	Současný stav placeného zaměstnání
	Measurement	Nominal
Valid Values	**0**	Nejsem zaměstnán(a)
	1	Jsem zaměstnán(a) v prvním zaměstnání
	2	Jsem zaměstnán(a) v jiném než prvním zaměstnání

D1		Value
Standard Attributes	**Label**	Působení v prvním zaměstnání do současné doby
	Measurement	Nominal
Valid Values	**0**	Ne
	1	Ano

D2		Value
Standard Attributes	**Label**	Typ smlouvy v současném zaměstnání
	Measurement	Nominal
Valid Values	**1**	Na dobu neurčitou
	2	Na dobu určitou
	3	Pouze OSVČ

D3		Value
Standard Attributes	**Label**	Studijní obor vhodný pro současné zaměstnání
	Measurement	Nominal
Valid Values	**1**	Vystudovaný obor
	2	Příbuzný studijní obor
	3	Zcela jiný studijní obor
	4	Zaměstnání nevyžaduje oborovou specializaci

Výstup 2.1 | Popis proměnných a jejich kategorií – pokračování

D4		Value
Standard Attributes	**Label**	Typ instituce
	Measurement	Nominal
Valid Values	**1**	Instituce ve veřejném sektoru
	2	Soukromá komerční společnost
	3	Soukromá nezisková organizace
	4	Instituce jiného typu

D5		Value
Standard Attributes	**Label**	Řízení jiných pracovníků
	Measurement	Nominal
Valid Values	**1**	Řídím
	2	Neřídím

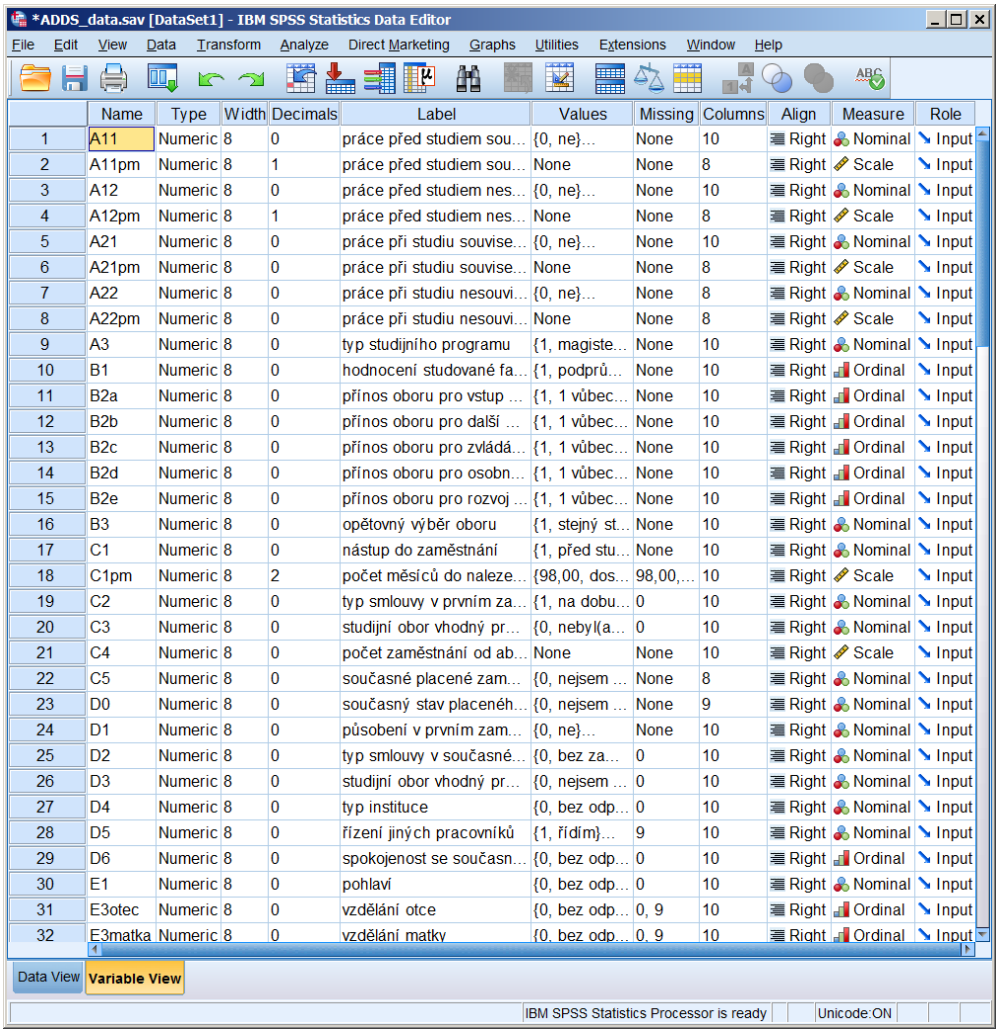
D6		Value
Standard Attributes	**Label**	Spokojenost se současnou prací
	Measurement	Ordinal
Valid Values	**1**	Velmi nespokojen(a)
	2	Spíše nespokojen(a)
	3	Napůl spokojen(a)
	4	Spíše spokojen(a)
	5	Velmi spokojen(a)

E1		Value
Standard Attributes	**Label**	Pohlaví
	Measurement	Nominal
Valid Values	**1**	Muž
	2	Žena

E3otec		Value
Standard Attributes	**Label**	Vzdělání otce
	Measurement	Ordinal
Valid Values	**1**	Bez maturity
	2	Středoškolské s maturitou
	3	Vysokoškolské

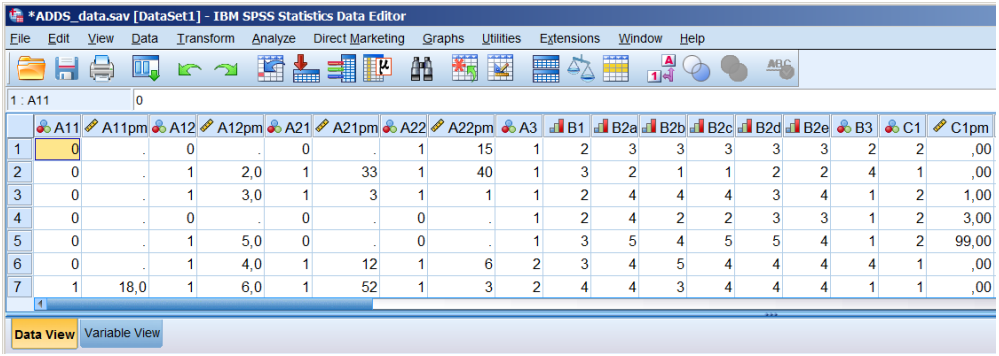
Obdobně **E3matka**; proměnné **E2** (děti) a **E2pd** (počet dětí) nejsou v knize analyzovány.

Obrázek 2.1 | List *Variable View* s popisem proměnných a jejich hodnot



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	A11	Numeric	8	0	práce před studiem sou...	{0, ne}...	None	10	Right	Nominal	Input
2	A11pm	Numeric	8	1	práce před studiem sou...	None	None	8	Right	Scale	Input
3	A12	Numeric	8	0	práce před studiem nes...	{0, ne}...	None	10	Right	Nominal	Input
4	A12pm	Numeric	8	1	práce před studiem nes...	None	None	8	Right	Scale	Input
5	A21	Numeric	8	0	práce při studiu souvise...	{0, ne}...	None	10	Right	Nominal	Input
6	A21pm	Numeric	8	0	práce při studiu souvise...	None	None	8	Right	Scale	Input
7	A22	Numeric	8	0	práce při studiu nesouvi...	{0, ne}...	None	8	Right	Nominal	Input
8	A22pm	Numeric	8	0	práce při studiu nesouvi...	None	None	8	Right	Scale	Input
9	A3	Numeric	8	0	typ studijního programu	{1, magiste...	None	10	Right	Nominal	Input
10	B1	Numeric	8	0	hodnocení studované fa...	{1, podprů...	None	10	Right	Ordinal	Input
11	B2a	Numeric	8	0	přínos oboru pro vstup ...	{1, 1 vůbec...	None	10	Right	Ordinal	Input
12	B2b	Numeric	8	0	přínos oboru pro další ...	{1, 1 vůbec...	None	10	Right	Ordinal	Input
13	B2c	Numeric	8	0	přínos oboru pro zvládá...	{1, 1 vůbec...	None	10	Right	Ordinal	Input
14	B2d	Numeric	8	0	přínos oboru pro osobn...	{1, 1 vůbec...	None	10	Right	Ordinal	Input
15	B2e	Numeric	8	0	přínos oboru pro rozvoj ...	{1, 1 vůbec...	None	10	Right	Ordinal	Input
16	B3	Numeric	8	0	opětovný výběr oboru	{1, stejný st...	None	10	Right	Nominal	Input
17	C1	Numeric	8	0	nástup do zaměstnání	{1, před stu...	None	10	Right	Nominal	Input
18	C1pm	Numeric	8	2	počet měsíců do naleze...	{98,00, dos...	98,00,...	10	Right	Scale	Input
19	C2	Numeric	8	0	typ smlouvy v prvním za...	{1, na dobu...	0	10	Right	Nominal	Input
20	C3	Numeric	8	0	studijní obor vhodný pr...	{0, nebyl(a...	0	10	Right	Nominal	Input
21	C4	Numeric	8	0	počet zaměstnání od ab...	None	None	10	Right	Scale	Input
22	C5	Numeric	8	0	současné placené zam...	{0, nejsem ...	None	8	Right	Nominal	Input
23	D0	Numeric	8	0	současný stav placeném...	{0, nejsem ...	None	9	Right	Nominal	Input
24	D1	Numeric	8	0	působení v prvním zam...	{0, ne}...	None	10	Right	Nominal	Input
25	D2	Numeric	8	0	typ smlouvy v současné...	{0, bez za...	0	10	Right	Nominal	Input
26	D3	Numeric	8	0	studijní obor vhodný pr...	{0, nejsem ...	0	10	Right	Nominal	Input
27	D4	Numeric	8	0	typ instituce	{0, bez odp...	0	10	Right	Nominal	Input
28	D5	Numeric	8	0	řízení jiných pracovníků	{1, řídím}...	9	10	Right	Nominal	Input
29	D6	Numeric	8	0	spokojenost se současn...	{0, bez odp...	0	10	Right	Ordinal	Input
30	E1	Numeric	8	0	pohlaví	{0, bez odp...	0	10	Right	Nominal	Input
31	E3otec	Numeric	8	0	vzdělání otce	{0, bez odp...	0, 9	10	Right	Ordinal	Input
32	E3matka	Numeric	8	0	vzdělání matky	{0, bez odp...	0, 9	10	Right	Ordinal	Input

Obrázek 2.2 | List *Data View* s daty



	A11	A11pm	A12	A12pm	A21	A21pm	A22	A22pm	A3	B1	B2a	B2b	B2c	B2d	B2e	B3	C1	C1pm
1	0	.	0	.	0	.	1	15	1	2	3	3	3	3	3	2	2	,00
2	0	.	1	2,0	1	33	1	40	1	3	2	1	1	2	2	4	1	,00
3	0	.	1	3,0	1	3	1	1	1	2	4	4	4	3	4	1	2	1,00
4	0	.	0	.	0	.	0	.	1	2	4	2	2	3	3	1	2	3,00
5	0	.	1	5,0	0	.	0	.	1	3	5	4	5	5	4	1	2	99,00
6	0	.	1	4,0	1	12	1	6	2	3	4	5	4	4	4	4	1	,00
7	1	18,0	1	6,0	1	52	1	3	2	4	4	3	4	4	4	1	1	,00

2.3 Problematika chybějících údajů

Existuje mnoho příčin, které mohou způsobit výskyt chybějícího údaje. Jako tři základní skupiny příčin lze uvést tyto:

- nezjištění příslušné hodnoty,
- chybná odpověď (mimo stanovený výčet či rozsah hodnot nebo odhalená logickou kontrolou dat),
- chyba při zápisu dat.

Příčiny nezjištění příslušné hodnoty mohou být následující. Respondent

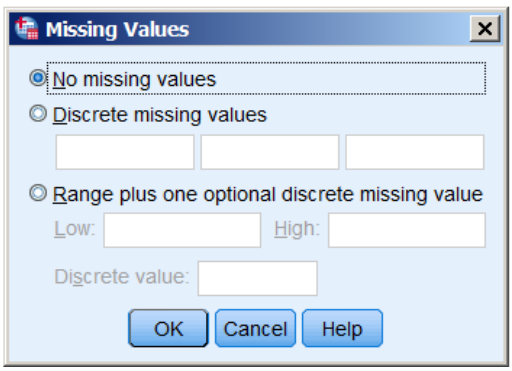
- nerozumí otázce,
- otázce rozumí, ale nabídka odpovědí neobsahuje možnost vhodné volby,
- odmítá odpovědět (nechce sdělovat osobní údaje, např. výši svého příjmu),
- ztratil zájem o vyplňování dotazníku,
- nemá dost času, aby vyplnil celý dotazník.

Podíl chybějících údajů je značně závislý na charakteru a kvalitě dotazníku. Je třeba vyhnout se především nejasným formulacím otázek.

V systému *IBM SPSS Statistics* se rozlišují *systémové chybějící údaje* (do souboru nebyly vloženy žádné hodnoty nebo nebylo možné vypočítat hodnoty odvozené proměnné) a *uživatelem definované chybějící údaje*. V prvním případě se v tabulce dat zobrazuje tečka. Ve druhém případě uživatel zadává hodnoty, které má program vyhodnotit jako chybějící, a v konkrétních statistických metodách lze specifikovat, zda mají, či nemají být tyto údaje zahrnuty do výpočtu (pro systémové chybějící údaje možnost volby neexistuje). Kromě jednotlivých hodnot lze uvést i celý interval hodnot, k dispozici jsou následující možnosti (viz též obrázek 2.3):

- 1 až 3 diskrétní hodnoty (např. odlišení, zda odpověď není vůbec, nebo je uvedena odpověď typu „nevím“, případně je vyznačeno více odpovědí místo jedné),
- interval hodnot, případně interval a 1 diskrétní hodnota.

Obrázek 2.3 | Panel pro specifikaci chybějících údajů



Existují dva základní postupy pro manipulaci s chybějícími údaji, kterými jsou:

- ponechání chybějících údajů, což vyžaduje speciální postupy při matematických výpočtech a při použití statistických metod,
- nahrazení chybějících údajů konkrétními hodnotami.

Ve druhém případě můžeme použít různé způsoby nahrazení. Při některých z nich je třeba znát určité statistické charakteristiky, které jsou popsány v oddílu 3.2. Jako příklady lze uvést následující metody:

Nahrazení průměrem. Spočívá v nahrazení chybějících údajů aritmetickým průměrem spočteným ze všech zjištěných hodnot dané proměnné. Tento postup je obvykle v programových systémech nabízen jako jediný možný. Metoda není vhodná, pokud chybí mnoho hodnot.

Nahrazení skupinovým (podmíněným) průměrem. Hodnoty proměnné, v níž se vyskytují chybějící údaje, jsou rozděleny do skupin podle hodnot jiné proměnné (má smysl, pokud se v těchto skupinách liší statistické charakteristiky). V každé skupině je vypočten aritmetický průměr, případně zjištěn modus, tj. nejčetnější hodnota (u kategoriální proměnné) a chybějící hodnota je nahrazena příslušnou charakteristikou. Jinou možností je nahrazení náhodně vybranou hodnotou z dané skupiny.

Nahrazení podle vzoru. Hodnoty vybraných proměnných u objektu (případu), u něhož se vyskytuje chybějící údaj, jsou porovnávány s hodnotami těchto proměnných u jiných objektů. Pokud je nalezen objekt se stejnými hodnotami, nahradí se chybějící údaj hodnotou příslušné proměnné vyskytující se u tohoto objektu. Není-li takový případ nalezen, pak lze zvolit jednu z následujících variant:

- postup se opakuje pro jiné proměnné,
- případ se náhodně vybere,
- hodnota se nenahrazuje.

Analýza jednotlivých proměnných

3

V této kapitole budeme nejprve uvažovat metody popisu datového souboru. Poté se zaměříme na úsudky o *základním souboru* (populaci) na základě pořízených dat (za předpokladu, že *výběrový soubor* respondentů byl pořízen prostým náhodným výběrem). K základním typům analýz patří zjištění rozdělení četností různých variant hodnot a výpočet souhrnných charakteristik pro každou proměnnou.

3.1 Rozdělení četností

Názorný přehled o zjištěných hodnotách kategoriální proměnné poskytuje rozdělení četností jednotlivých kategorií. To může být prezentováno ve dvou formách – v tabulce nebo v grafu. V sociologii je tato základní analýza označována jako *třídění I. stupně*.

3.1.1 Tabulky rozdělení četností

Nejprve uvažujme kategoriální proměnnou X a její kategorie označme symbolem x_i, kde $i = 1, 2, \ldots, K$, přičemž K je počet kategorií. Dále označme počet respondentů symbolem n. V tabulce rozdělení četností uvádíme pro každou kategorii především **absolutní četnost** n_i. Dále je možné uvádět **relativní četnost** p_i, která vyjadřuje podíl počtu výskytů dané kategorie na celkovém rozsahu souboru, tzn. $p_i = n_i / n$ (někdy jsou tyto hodnoty násobeny stem, tzn. vyjadřovány v procentech). U ordinálních a kvantitativních proměnných má navíc smysl počítat **kumulativní relativní četnost** P_i, přičemž její výpočet je prováděn podle následujícího postupu:

$$P_1 = p_1, P_2 = p_1 + p_2, \ldots, \quad P_K = \sum_{i=1}^{K} p_i = 1, \quad \text{tj. } P_i = \sum_{j=1}^{i} p_j \ .$$

Souhrnně popsanou symboliku znázorňuje schéma 3.1.

Schéma 3.1 | Symbolika pro tabulku rozdělení četností

Znak *X*	Četnost		
	Absolutní	Relativní	Kumulativní relativní
x_1	n_1	p_1	P_1
...	...	...	...
x_i	n_i	p_i	P_i
...	...	...	...
x_K	n_K	p_K	1
Celkem	n	1	

Systém *IBM SPSS Statistics* uvádí relativní četnosti v procentech a ve dvou sloupečcích, přičemž v prvním je uvažován celkový počet respondentů a ve druhém pouze počet respondentů, jejichž odpověď patří do tzv. platných hodnot.

Příklad 3.1

Vytvořme tabulku rozdělení četností pro proměnnou *B1* (hodnocení fakulty).

Četnosti jednotlivých možných hodnocení fakulty, na níž absolvent studoval, zobrazuje výstup 3.1, který byl získán volbou možností *Analyze*, *Descriptive Statistics a Frequencies*. V tabulce jsou uvedeny absolutní četnosti (*Frequency*), relativní četnosti v procentech (*Percent*), platné relativní četnosti v procentech (*Valid Percent*) a kumulativní relativní četnosti v procentech (*Cumulative Percent*). Z výstupu je zřejmé, že všech 635 respondentů otázku zodpovědělo (řádek *Total*).

Výstup 3.1 | Rozdělení četností pro proměnnou *B1*

Hodnocení studované fakulty					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Podprůměrné	14	2,2	2,2	2,2
	Průměrné	146	23,0	23,0	25,2
	Nadprůměrné	418	65,8	65,8	91,0
	Vynikající	57	9,0	9,0	100,0
	Total	635	100,0	100,0	

♦ ♦ ♦

Nabývá-li proměnná většího počtu hodnot, tabulka rozdělení četností pro jednotlivé varianty hodnot by již neposkytovala vhodný souhrnný popis. V takovém případě se hodnoty roztřídí do intervalů a zjišťují se četnosti hodnot v jednotlivých intervalech.

Příklad 3.2

Vytvořme tabulku rozdělení četností pro proměnnou *C1pm*, poté proveďme kódování do intervalů a vytvořme tabulku rozdělení četností pro novou proměnnou.

Postupem uvedeným v příkladu 3.1 získáme pro proměnnou *C1pm* výstup 3.2, který zobrazuje četnosti jednotlivých možných počtů měsíců do nalezení práce (u respondentů, kteří do zaměstnání nastoupili před studiem nebo v průběhu studia vysoké školy, byla do této proměnné dosazena hodnota 0). Z této tabulky četností se dozvídáme, že 30 respondentů tento dotaz nezodpovědělo, což je 4,7 % (část *Missing*). Provedeme kódování do pěti intervalů, které vyjadřují „0“, „ do 1 měsíce“, „2–3 měsíce“, „4–6 měsíců“ a „déle než 6 měsíců“.

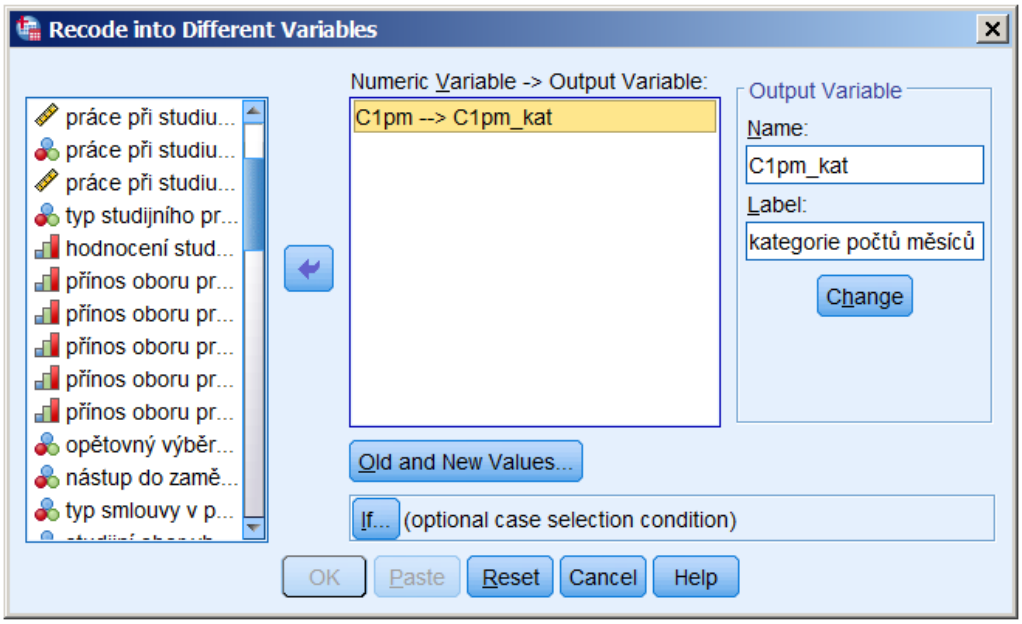
Překódování specifikujeme v prostředí, k němuž vedou nabízené položky *Transform*, *Recode into Different Variables*. Postup znázorňují obrázky 3.1 a 3.2. Nejprve vybereme výchozí proměnnou a specifikujeme proměnnou novou, a to názvem (*C1pm_kat*) a popisem (kategorie počtů měsíců do nalezení práce), viz obr. 3.1. Pak zadáváme jednotlivé hodnoty nebo intervaly výchozí proměnné, k nimž specifikujeme nové kódy (obr. 3.2).

Výstup 3.2 | Rozdělení četností pro proměnnou *C1pm*

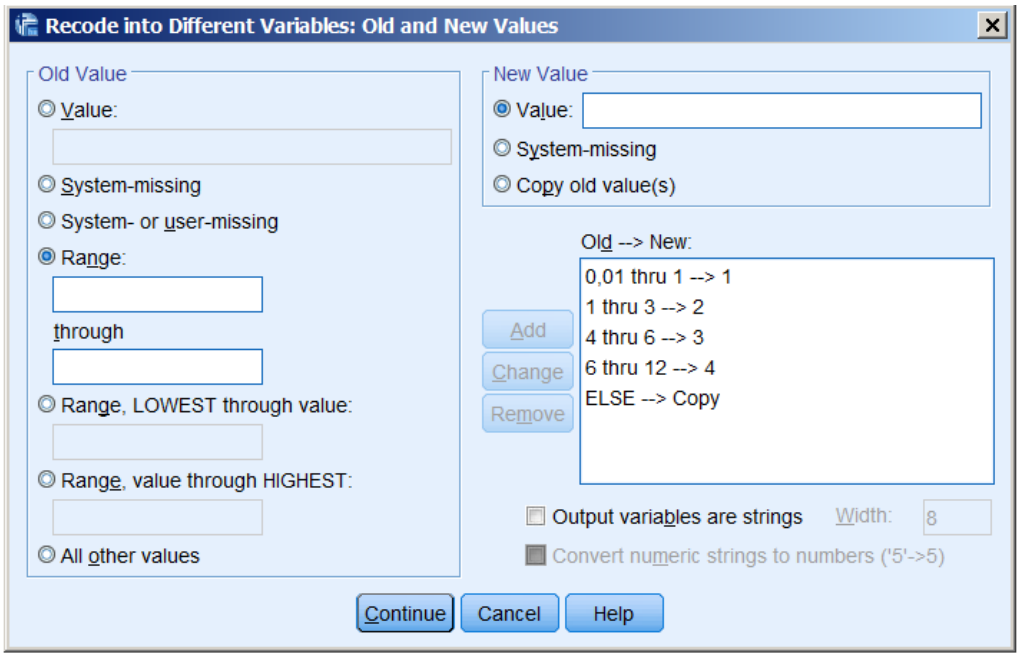
Počet měsíců do nalezení práce					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**,00**	327	51,5	55,3	55,3
	,03	1	,2	,2	55,5
	,25	1	,2	,2	55,7
	,50	2	,3	,3	56,0
	1,00	113	17,8	19,1	75,1
	2,00	76	12,0	12,9	88,0
	2,50	2	,3	,3	88,3
	3,00	39	6,1	6,6	94,9
	4,00	10	1,6	1,7	96,6
	5,00	7	1,1	1,2	97,8
	6,00	10	1,6	1,7	99,5
	8,00	1	,2	,2	99,7
	11,00	1	,2	,2	99,8
	12,00	1	,2	,2	100,0
	Total	591	93,1	100,0	
Missing	**Dosud nepracuji**	14	2,2		
	Bez odpovědi	30	4,7		
	Total	44	6,9		
Total		635	100,0		

Pro novou proměnnou zadáme popis jejich hodnot, včetně kódů pro chybějící údaje. K tomu je určen list *Variable View* v datovém editoru systému *IBM SPSS Statistics* (viz obrázek 2.1). Nová tabulka rozdělení četností je zobrazena ve výstupu 3.3.

Obrázek 3.1 | Vstupní panel pro překódování proměnné



Obrázek 3.2 | Panel pro zadání nových hodnot proměnné



Výstup 3.3 | Rozdělení četností pro proměnnou *C1pm_kat*

Kategorie počtů měsíců do nalezení práce					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 měsíců	327	51,5	55,3	55,3
	Do 1 měsíce	117	18,4	19,8	75,1
	2–3 měsíce	117	18,4	19,8	94,9
	4–6 měsíců	27	4,3	4,6	99,5
	Více než 6 měsíců	3	,5	,5	100,0
	Total	591	93,1	100,0	
Missing	Dosud nepracuji	14	2,2		
	Bez odpovědi	30	4,7		
	Total	44	6,9		
Total		635	100,0		

♦ ♦ ♦

Pokud nemáme k dispozici zdrojové údaje, ale již zpracovaná data ve formě tabulky četností, pak pro další analýzy zadáme do datového editoru systému *IBM SPSS Statistics* do jednoho sloupce jednotlivé kategorie a do druhého četnosti jejich výskytu. Poté zadaným kategoriím přiřadíme tyto četnosti (váhy) pomocí nabídky *Data*, *Weight Cases* (zvolíme možnost *Weight cases by* a do políčka *Frequency Variable* zadáme název sloupce, ve kterém jsou zadány četnosti).

Známe-li kategorie a četnosti z výstupu 3.1, pak do datového editoru zadáme hodnoty uvedené v tabulce 3.1.

Tabulka 3.1 Váhy pro jednotlivé kategorie proměnné *B*1 (*hodnocení studované fakulty*)

Hodnocení	Váha
1	14
2	146
3	418
4	57

3.1.2 Grafy rozdělení četností

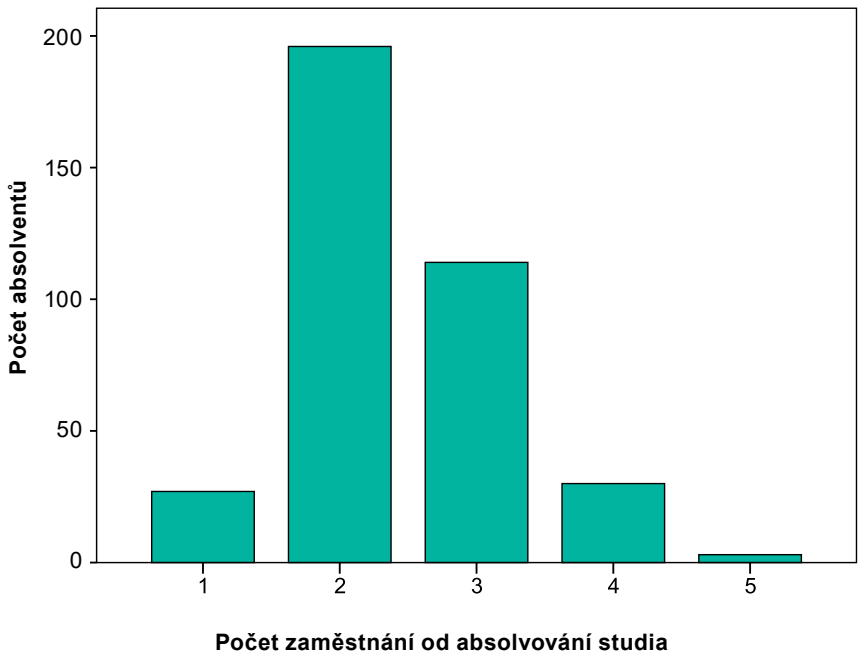
K zobrazení rozdělení četností hodnot kategoriální proměnné se používá zejména **graf sloupcový**, pro nominální proměnnou někdy též **graf výsečový**. V prvním případě výška sloupce představuje počet statistických jednotek, u nichž se hodnota sledovaného znaku rovná určité kategorii (výška může také reprezentovat relativní četnost těchto statistických jednotek). Ve druhém případě je k dispozici kruh rozdělený na výseče, přičemž podíly jejich ploch na celku odpovídají podílům četností jednotlivých kategorií proměnné na celkovém počtu statistických jednotek.

Příklad 3.3

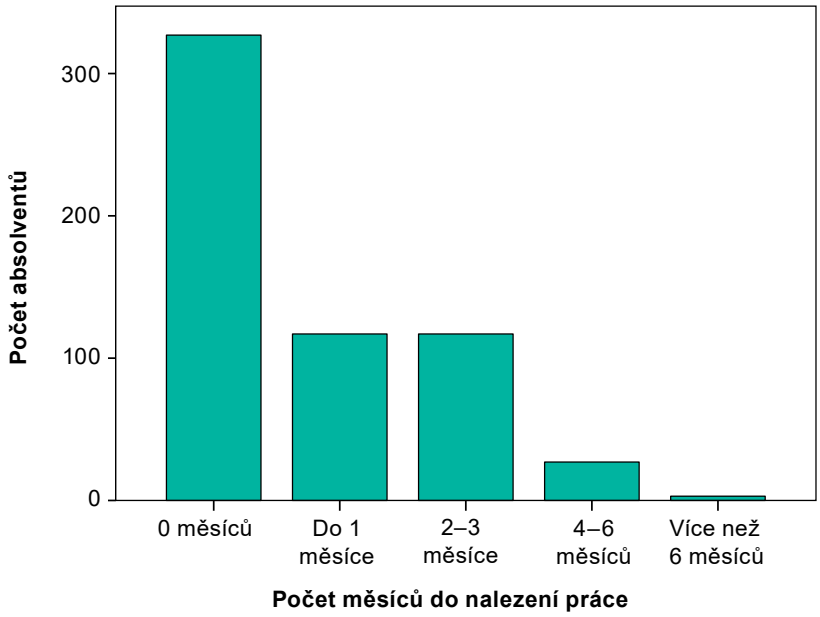
Graficky znázorněme rozdělení četností pro proměnné *C4* (počet zaměstnání od absolvování studia), *C1pm_kat* (počet měsíců do nalezení práce) a *C2* (typ pracovní smlouvy).

V systému *IBM SPSS Statistics* získáme grafy volbou nabízených položek *Analyze*, *Descriptive Statistics a Frequencies* a výběrem typu grafu v rámci možnosti *Charts* (pro sloupcový graf je určena položka *Bar charts*, pro výsečový graf položka *Pie charts*). Rozdělení četností pro první dvě proměnné znázorníme pomocí sloupcového grafu, podíly typů pracovní smlouvy v prvním zaměstnání pomocí výsečového grafu. Výsledné výstupy (upravené pomocí grafického editoru) zachycují grafy 3.1 až 3.3.

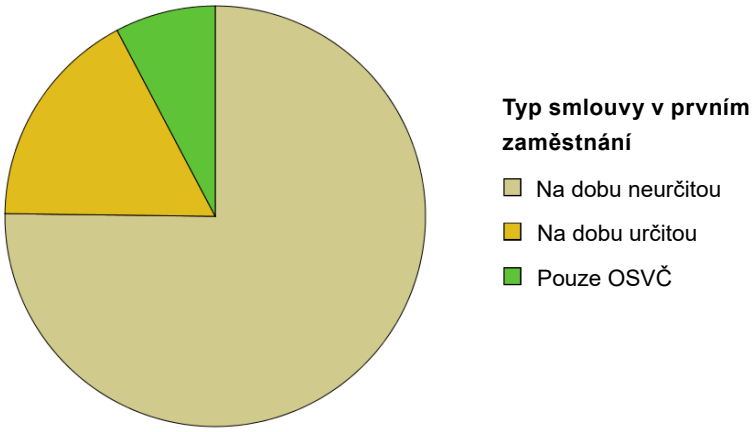
Graf 3.1 | Sloupcový graf pro proměnnou *C4*



Graf 3.2 | Sloupcový graf pro proměnnou *C1pm_kat*



Graf 3.3 | Výsečový graf pro proměnnou *C2*



3.1.3 Zacházení s vícehodnotovými odpověďmi

Při vyhodnocování dotazníků se vyskytují situace, kdy u některé otázky respondenti mohou zvolit více odpovědí. Například student mohl být před studiem zaměstnán u více zaměstnavatelů a některá práce mohla souviset s následným studiem a jiná ne. Tyto odpovědi jsou pak zaznamenávány do několika proměnných. Jejich zpracováním se zabývá *analýza vícehodnotových* (*vícenásobných*) *odpovědí* (*Multiple Response Analysis*).

Existují přitom dva základní přístupy, podle nichž máme možnost použít dva typy proměnných, a to:

- *dichotomické proměnné* (položená otázka je vlastně tvořena několika dílčími dotazy, na které lze odpovědět „ano“ nebo „ne“, přičemž počet proměnných odpovídá počtům dotazů; v modifikované podobě jsou příkladem otázky *A1* a *A2*),
- *vícekategoriální proměnné* (respondent obvykle vybírá určitý počet z možných odpovědí, otázka může například znít: „Uveďte maximálně tři možnosti týkající se zaměstnání před a při studiu, u nichž práce trvala nejdéle“, odpovědi jsou tedy zaznamenávány do tří proměnných).

Systém *IBM SPSS Statistics* umožňuje analýzu vícehodnotových odpovědí oběma způsoby. Podmínkou však je, že hodnoty proměnných musí být zadány jako čísla.

Příklad 3.4

Vytvořme tabulky rozdělení četností pro proměnné *A11*, *A12*, *A21* a *A22*, a to nejprve pro každou proměnnou zvlášť a poté jednu souhrnnou tabulku pro všechny proměnné.

Četnosti jednotlivých možných zaměstnání zobrazují výstupy 3.4–3.7 , které byly získány volbou možností *Analyze*, *Descriptive Statistics a Frequencies*. Procedura *Multiple Response* umožňuje spojit dichotomické proměnné do jedné skupiny. Pro definování této skupiny volíme položky *Analyze*, *Multiple Response* a *Define Variable Sets*. V dialogovém panelu *Define Multiple Response Sets* ponecháme nastavenu možnost *Dichotomies*, viz obrázek 3.3.

Tabulka četností (získaná volbou položek *Analyze*, *Multiple Response* a *Frequencies*) pro skupinu *$zaměstnání* je uvedena ve výstupu 3.8. Pro každou proměnnou se vypisují absolutní četnosti (*N*) zadané kategorie (zde hodnota 1), procentní podíl těchto četností (*Percent*) na celkovém počtu platných hodnot (*Responses*) a procentní relativní četnost vztažená k počtu respondentů (*Percent of Cases*), kteří odpověděli alespoň na jednu otázku kladně.

Počet absolventů, kteří byli před studiem nebo při studiu zaměstnáni, lze zjistit ze souhrnného přehledu, který je umístěn ve výstupu před vlastní tabulkou četností (ukázka zde není zařazena). V daném případě bylo těchto absolventů 618. Z výstupu 3.8 lze zjistit, že bylo zaznamenáno celkem 1 455 kladných odpovědí (sloupec *N*, řádek *Total*). Dále lze například vyčíst, že před studiem mělo 131 absolventů zaměstnání související se studiem, což představuje 21,2 % z celkového počtu absolventů, kteří byli před studiem nebo při studiu zaměstnáni (131/618 = 0,212). V posledním řádku a posledním sloupci je hodnota 235,4 %, která vyjadřuje procentuální zastoupení všech kladných odpovědí, tedy na jednoho respondenta 1455/618 = 2,354.

Výstup 3.4 | Rozdělení četností pro proměnnou *A11*

Práce před studiem související se studiem					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	504	79,4	79,4	79,4
	Ano	131	20,6	20,6	100,0
	Total	635	100,0	100,0	

Výstup 3.5 | Rozdělení četností pro proměnnou *A12*

Práce před studiem nesouvisející se studiem					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	201	31,7	31,7	31,7
	Ano	434	68,3	68,3	100,0
	Total	635	100,0	100,0	

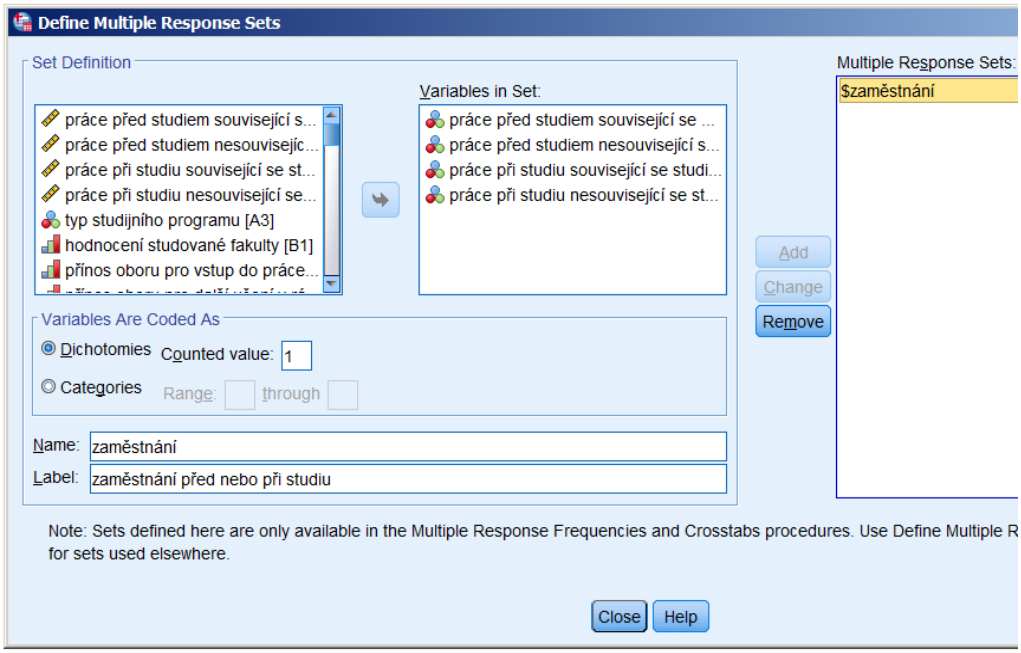
Výstup 3.6 | Rozdělení četností pro proměnnou *A21*

Práce při studiu související se studiem					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	205	32,3	32,3	32,3
	Ano	430	67,7	67,7	100,0
	Total	635	100,0	100,0	

Výstup 3.7 | Rozdělení četností pro proměnnou *A22*

Práce při studiu nesouvisející se studiem					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	175	27,6	27,6	27,6
	Ano	460	72,4	72,4	100,0
	Total	635	100,0	100,0	

Obrázek 3.3 | Ukázka definování skupiny dichotomických proměnných



Výstup 3.8 | Tabulka četností různých typů zaměstnání před studiem nebo při studiu

		Responses		Percent of Cases
		N	Percent	
Zaměst-nání před studiem nebo při studiu[a]	**Práce před studiem související se studiem**	131	9,0 %	21,2 %
	Práce před studiem nesouvisející se studiem	434	29,8 %	70,2 %
	Práce při studiu související se studiem	430	29,6 %	69,6 %
	Práce při studiu nesouvisející se studiem	460	31,6 %	74,4 %
Total		1 455	100,0 %	235,4 %

a. Dichotomy group tabulated at value 1.

Příklad 3.5

Vytvořme tabulky rozdělení četností pro čtyři proměnné vyjadřující různé možné přínosy oboru překódované do proměnných se dvěma kategoriemi (dvě nejnižší úrovně jsou označeny kódem 0 a tři úrovně vyjadřující alespoň částečný přínos kódem 1), tj. proměnných *B2a_2kat*, *B2b_2kat*, *B2c_2kat* a *B2d_2kat*. Opět nejprve pro každou proměnnou zvlášť a poté jednu souhrnnou tabulku pro všechny proměnné. Četnosti jednotlivých možných přínosů zobrazují výstupy 3.9–3.12.

Výstup 3.9 | Rozdělení četností pro proměnnou *B2a_2kat*

Přínos oboru pro vstup do práce					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	140	22,0	22,0	22,0
	Ano	495	78,0	78,0	100,0
	Total	635	100,0	100,0	

Výstup 3.10 | Rozdělení četností pro proměnnou *B2b_2kat*

Přínos oboru pro další učení v rámci práce					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	138	21,7	21,7	21,7
	Ano	497	78,3	78,3	100,0
	Total	635	100,0	100,0	

Výstup 3.11 | Rozdělení četností pro proměnnou *B2c_2kat*

Přínos oboru pro zvládání pracovních úkolů					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	156	24,6	24,6	24,6
	Ano	479	75,4	75,4	100,0
	Total	635	100,0	100,0	

Výstup 3.12 | Rozdělení četností pro proměnnou *B2d_2kat*

Přínos oboru pro osobní rozvoj					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ne	142	22,4	22,4	22,4
	Ano	493	77,6	77,6	100,0
	Total	635	100,0	100,0	

Tabulka četností pro skupinu *$obor* je uvedena ve výstupu 3.13.

Výstup 3.13 | Tabulka četností některých možných přínosů oborů

		Responses		Percent of Cases
		N	Percent	
Přínos oboru[a]	Přínos oboru pro vstup do práce	495	25,2 %	83,5 %
	Přínos oboru pro další učení v rámci práce	497	25,3 %	83,8 %
	Přínos oboru pro zvládání pracovních úkolů	479	24,4 %	80,8 %
	Přínos oboru pro osobní rozvoj	493	25,1 %	83,1 %
Total		1 964	100,0 %	331,2 %

a. Dichotomy group tabulated at value 1.

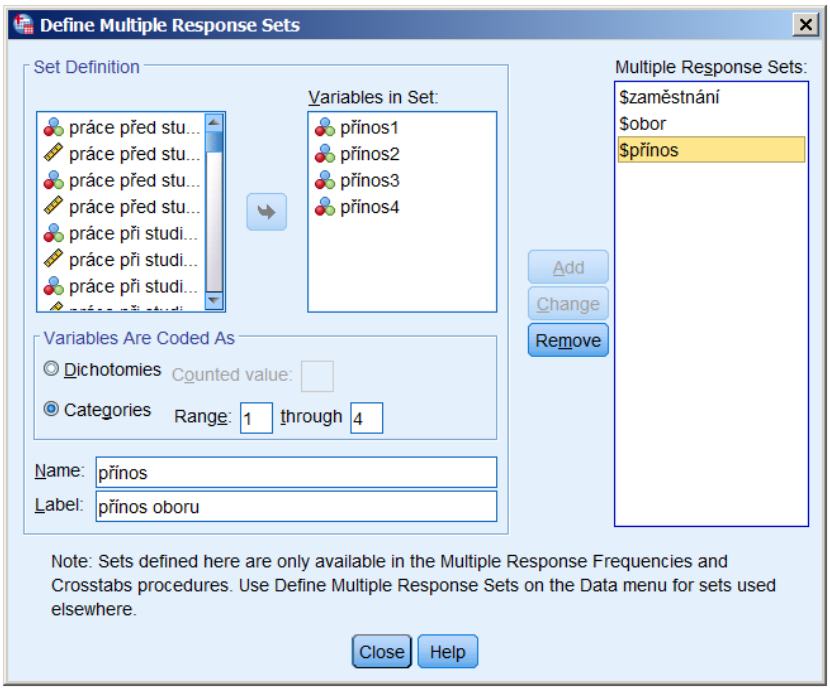
Jiným možným přístupem je zaznamenání odpovědí do čtyř vícekategoriálních proměnných (v případě, že by respondenti nevyužívali všechny možnosti, pak do menšího počtu proměnných), které by nabývaly hodnot od 1 (přínos oboru pro vstup do práce) do 4 (přínos oboru pro osobní rozvoj). Označme si takovéto proměnné názvy *přínos1*, *přínos2*, *přínos3* a *přínos4*. V dotaznících se mohou vyskytovat např. varianty odpovědí, které jsou zaznamenány v tabulce 3.2. Druhý řádek v této tabulce představuje odpovědi absolventů, pro které byl studijní obor přínosem pro vstup do práce a pro další učení v rámci práce atd.

Tabulka 3.2 | Ukázka dat pro analýzu vícekategoriálních proměnných

Varianta odpovědi	*přínos1*	*přínos2*	*přínos3*	*přínos4*
1	1	.	.	.
2	1	2	.	.
3	1	4	.	.
4	1	2	3	.
5	1	2	3	4
6	2	.	.	.
7	2	3	.	.
8	2	4	.	.
9	2	3	4	.
10	3	4	.	.
11	4	.	.	.

Nové proměnné sdružíme prostřednictvím procedury *Multiple Response* do nové skupiny *$přínos*. V daném případě se jedná o vícekategoriální typ s oborem hodnot od 1 do 4 (možnost *Categories* v dialogovém panelu *Define Multiple Response Sets*, viz obrázek 3.4).

Obrázek 3.4 | Ukázka definování skupiny vícekategoriálních proměnných



Tabulka četností (získaná volbou položek *Analyze*, *Multiple Response* a *Frequencies*) pro skupinu *$přínos* je uvedena ve výstupu 3.14. Vypisují se absolutní četnosti (*N*) jednotlivých kategorií, procentní podíl těchto četností (*Percent*) z celkového počtu platných hodnot (*Responses*) a procentní podíl těchto četností z počtu respondentů (*Percent of Cases*). Hodnoty v tabulce četností se shodují s hodnotami ve výstupu 3.13, přestože byl použit jiný způsob záznamu dat.

Výstup 3.14 | Tabulka četností některých možných přínosů oborů (skupina vícekategoriálních proměnných)

		Responses		Percent of Cases
		N	Percent	
Přínos oboru[a]	Přínos oboru pro vstup do práce	495	25,2 %	83,5 %
	Přínos oboru pro další učení v rámci práce	497	25,3 %	83,8 %
	Přínos oboru pro zvládání pracovních úkolů	479	24,4 %	80,8 %
	Přínos oboru pro osobní rozvoj	493	25,1 %	83,1 %
Total		1 964	100,0 %	331,2%

a. Group.

3.2 Popisné charakteristiky

Na základě rozdělení četností můžeme vypočítat statistické charakteristiky určitého znaku. Tento oddíl se zaměřuje především na *míry polohy* a *míry variability* používané pro různé typy proměnných, zmíněny budou též *míry koncentrace*.

3.2.1 Nominální proměnná

A. Míry polohy

Poloha (neboli úroveň) je u nominální proměnné charakterizována **modální kategorií**, což je kategorie s největší četností. Jsou-li kategorie označeny indexem i (i = 1, 2, …, K, kde K je počet kategorií), modální kategorie (*modus*) označena indexem Mo, n_i jsou absolutní četnosti a p_i relativní četnosti, pak $\max_i(n_i) = n_{\text{Mo}}$ a $\max_i(p_i) = p_{\text{Mo}}$. Tyto četnosti modální kategorie se rovněž nazývají modální.

U sledované proměnné se může vyskytovat buď jedna, nebo více modálních kategorií. V prvním případě jde o rozdělení četností *unimodální*, ve druhém případě o k-modální rozdělení, kde k je počet modálních kategorií. Konkrétně při výskytu dvou modálních kategorií se rozdělení nazývá *bimodální* a při třech těchto kategoriích jde o *trimodální* rozdělení.

Jestliže $p_{\text{Mo}} > 0{,}5$, pak můžeme modální kategorii označit též jako *majoritní* a četnosti n_{Mo} a p_{Mo} rovněž jako majoritní.

B. Míry variability

Základem pro zkoumání variability je zjištění *koncentrace*. Jako míru koncentrace pro nominální proměnné můžeme použít:

a) relativní četnost modální kategorie, tj. $p_{\text{Mo}} = n_{\text{Mo}}/n$, kde n je celkový rozsah základního souboru,

b) součet druhých mocnin relativních četností, tj. $\sum_{i=1}^{K} p_i^2$, kde K je počet kategorií.

Budeme uvažovat dva extrémní případy. V prvním bude nenulová četnost pouze u jedné kategorie, což znamená, že ostatní kategorie nejsou ve sledovaném souboru zastoupeny. Pak $p_{\text{Mo}} = 1$ a $\sum p_i^2 = 1$. Ve druhém případě budou kategorie v souboru rovnoměrně zastoupeny, takže $p_i = 1/K$ (pro i = 1, …, K). Potom též $p_{\text{Mo}} = 1/K$ a $\sum p_i^2 = 1/K$, neboť

$$\sum_{i=1}^{K} p_i^2 = \sum_{i=1}^{K} \left(\frac{1}{K}\right)^2 = K\frac{1}{K^2} = \frac{1}{K}\,.$$

Jako míry *variability* pak slouží

a) variační poměr v, který spočteme podle vzorce

$$v = 1 - p_{Mo} = 1 - n_{Mo}/n, \tag{3.1}$$

b) nominální rozptyl *nomvar* (Giniho koeficient), vyjadřující relativní počet všech dvojic, které nejsou ve stejné kategorii, a počítaný podle vzorce

$$nomvar = 1 - \sum_{i=1}^{K} p_i^2 = \sum_{i=1}^{K} (p_i(1 - p_i)), \tag{3.2}$$

nebo

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

c) **entropie** H, která je dána vzorcem

$$H = - \sum_{i=1}^{K} p_i \ln p_i, \tag{3.3}$$

je-li pro všechna i relativní četnost $p_i > 0$. V případě $p_i = 0$ se pro dané i položí příslušný sčítanec roven nule.

Platí, že $p_{Mo} \in \langle 1/K; 1 \rangle$ a $\sum p_i^2 \in \langle 1/K; 1 \rangle$. Tudíž $v \in \langle 0; (K - 1)/K \rangle$ a také $nomvar \in \langle 0; (K - 1)/K \rangle$. Dále $H \in \langle 0; \ln K \rangle$. Jestliže míra variability nabude hodnoty nula, pak hovoříme o nulovém rozptýlení čili úplné homogenitě. Platí, že čím vyšší je hodnota, která charakterizuje variabilitu, tím vyšší je heterogenita souboru (maximální variabilita nastává v případě, kdy jsou všechny kategorie rovnoměrně zastoupeny).

Míry variability mohou být vyjadřovány také pomocí hodnot z intervalu od 0 do 1, čehož dosáhneme tím, že hodnotu vypočítanou podle některého z výše uvedených vzorců dělíme maximálně možnou hodnotou, tj. $(K - 1)/K$, resp. $\ln K$. Používány jsou míry

d) **normalizovaný nominální rozptyl** $norm.\ nomvar = K \cdot nomvar/(K - 1)$,

e) **normalizovaná entropie** $H^* = H/\ln K$.

Příklad 3.6

Porovnejme polohu a variabilitu nominálních proměnných *C1* (*nástup do zaměstnání*) a *C2* (*typ pracovní smlouvy*), jejichž hodnoty souhrnně charakterizují výstupy 3.15 a 3.16 (jde o standardní výstupy ze systému *IBM SPSS Statistics*, které jsou vytvářeny bez ohledu na typ proměnné – pro nominální proměnné nemají kumulativní relativní četnosti význam).

Výstup 3.15 | Rozdělení četností pro proměnnou *C1*

Nástup do zaměstnání					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**Před studiem nebo v průběhu studia**	289	45,5	45,5	45,5
	Po absolvování VŠ	332	52,3	52,3	97,8
	Dosud nepracuji	14	2,2	2,2	100,0
	Total	635	100,0	100,0	

Výstup 3.16 | Rozdělení četností pro proměnnou *C2*

Typ smlouvy v prvním zaměstnání					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**Na dobu neurčitou**	467	73,5	75,2	75,2
	Na dobu určitou	106	16,7	17,1	92,3
	Pouze OSVČ	48	7,6	7,7	100,0
	Total	621	97,8	100,0	
Missing	**System**	14	2,2		
Total		635	100,0		

U proměnné *C1* je modální kategorií odpověď „po absolvování VŠ“ (četnost 332), u proměnné *C2* pak odpověď „na dobu neurčitou“ (četnost 467). Obě modální kategorie jsou navíc majoritní, neboť je pro odpověď vybralo více než 50 % absolventů. Protože ve druhém případě je četnost modální kategorie větší, je u proměnné *C2* variabilita hodnot menší než u proměnné *C1*.

Variační poměr je podle vzorce (3.1)

pro proměnnou *C1*: $v = 1 - 0{,}523 = 0{,}477$,

pro proměnnou *C2*: $v = 1 - 0{,}752 = 0{,}248$.

Nominální rozptyl je podle vzorce (3.2)

pro proměnnou *C1*: *nomvar* = 0,519, viz tabulka 3.3 (*norm. nomvar* = $3 \cdot 0{,}519/2 = 0{,}778$),

pro proměnnou *C2*: *nomvar* = 0,399, viz tabulka 3.4 (*norm. nomvar* = $3 \cdot 0{,}399/2 = 0{,}599$).

Entropie je podle vzorce (3.3)

pro proměnnou *C1*: $H = 0{,}781$, viz tabulka 3.5 ($H^* = 0{,}781/\ln 3 = 0{,}711$),

pro proměnnou *C2*: $H = 0{,}714$, viz tabulka 3.6 ($H^* = 0{,}714/\ln 3 = 0{,}65$).

Tabulka 3.3 | Pomocné výpočty pro nominální rozptyl (*nástup do zaměstnání*)

i	p_i	$p_i(1 - p_i)$
1	0,455	0,2480
2	0,523	0,2495
3	0,022	0,0215
Součet		**0,5190**

Tabulka 3.4 | Pomocné výpočty pro nominální rozptyl (*typ pracovní smlouvy*)

i	p_i	$p_i(1 - p_i)$
1	0,752	0,1865
2	0,171	0,1418
3	0,077	0,0711
Součet		**0,3993**

Tabulka 3.5 | Pomocné výpočty pro entropii (*nástup do zaměstnání*)

i	p_i	$p_i \ln p_i$
1	0,455	–0,3583
2	0,523	–0,3390
3	0,022	–0,0840
Součet		**–0,7813**

Tabulka 3.6 | Pomocné výpočty pro entropii (*typ pracovní smlouvy*)

i	p_i	$p_i \ln p_i$
1	0,752	–0,2143
2	0,171	–0,3020
3	0,077	–0,1974
Součet		**–0,7138**

3.2.2 Ordinální proměnná

A. Míry polohy

Kromě *modální kategorie* x_{Mo} (viz předchozí oddíl) patří u ordinální proměnné k mírám polohy též **mediánová kategorie** $x_{(\text{Me})}$. Je to kategorie, pro kterou je kumulativní četnost 0,5 nebo vyšší, když pro předchozí kategorii byla kumulativní četnost menší než 0,5. To znamená, že $P_{(\text{Me})-1} < 0{,}5$ a $P_{(\text{Me})} \geq 0{,}5$. Tato míra se obvykle používá, pokud kategorie vyjadřují např. stupeň určitého hodnocení a jsou označovány pořadovými čísly. V takovém případě je možno stanovit též **medián** $\tilde{x}$ a to následujícím způsobem. Pokud je $P_{(\text{Me})} > 0{,}5$, pak se medián rovná mediánové kategorii, tj. $\tilde{x} = x_{(\text{Me})}$. V případě, že $P_{(\text{Me})} = 0{,}5$, se medián spočte jako průměr z mediánové kategorie a kategorie následující, což lze v případě pořadových čísel vyjádřit jako $\tilde{x} = x_{(\text{Me})} + 0{,}5$. Tato charakteristika je tedy *střední hodnotou*.

B. Míry variability

Základní mírou variability pro ordinální proměnnou je **ordinální rozptyl** *dorvar* (diskrétní ordinální variance), určený vzorcem

$$dorvar = 2\sum_{i=1}^{K-1}(P_i(1-P_i)). \tag{3.4}$$

Platí $dorvar \in \langle 0; (K-1)/2 \rangle$. Na rozdíl od míry *nomvar*, která nabývá maxima v případě rovnoměrného rozdělení četností, míra *dorvar* nabývá maxima, právě když u 50 % objektů nabývá sledovaná proměnná hodnoty x_1 a u zbylých 50 % objektů hodnoty x_K za předpokladu, že proměnná obsahuje odpovědi na uzavřenou otázku s více než dvěma nabídkami a bereme v úvahu četnosti všech možných kategorií, tj. četnosti kategorií kromě první a poslední jsou nula. Kumulativní četnosti P_1 až P_{K-1} se rovnají ½, $P_K = 1$, takže $dorvar = 2 \cdot (K-1) \cdot ½ \cdot ½ = (K-1)/2$. Dělením maximální možnou hodnotou získáme **normalizovaný ordinální rozptyl** (nabývá hodnot z intervalu od 0 do 1)

$$norm.\ dorvar = 2 \cdot dorvar/(K-1).$$

Příklad 3.7

Porovnejme polohu a variabilitu ordinálních proměnných *B2a* (*přínos oboru pro vstup do práce*) a *B2c* (*přínos oboru pro zvládání pracovních úkolů*), jejichž hodnoty souhrnně charakterizují výstupy 3.17 a 3.18.

Výstup 3.17 | Rozdělení četností pro proměnnou *B2a*

Přínos oboru pro vstup do práce					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 vůbec ne	42	6,6	6,6	6,6
	2	98	15,4	15,4	22,0
	3	184	29,0	29,0	51,0
	4	197	31,0	31,0	82,0
	5 ve velké míře	114	18,0	18,0	100,0
	Total	635	100,0	100,0	

U obou proměnných je *modální kategorie* (nejčastější odpověď) označena kódem 4 (četnosti 197 a 216). *Mediánovou kategorií* je v obou případech třetí kategorie. U proměnné *B2a* vybralo první, druhou nebo třetí odpověď 51 % respondentů, u proměnné *B2c* je to 54,3 % pro první tři kategorie. V obou případech je tedy *mediánem* hodnota 3.

Z tabulek četností dále můžeme usoudit, že větší variabilita je u proměnné *B2a*, neboť první a poslední kategorie jsou četněji zastoupeny než u proměnné *B2c*. Tomu odpovídají též hodnoty *ordinálního rozptylu*, který je podle vzorce (3.4)

pro proměnnou *B2a*: *dorvar* = 2 · 0,6307 = 1,262, viz tabulka 3.7 (*norm. dorvar* = 0,631), pro proměnnou *B2c*: *dorvar* = 2 · 0,5916 = 1,183, viz tabulka 3.8 (*norm. dorvar* = 0,592).

Výstup 3.18 | Rozdělení četností pro proměnnou *B2c*

Přínos oboru pro zvládání pracovních úkolů					
		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**1 vůbec ne**	37	5,8	5,8	5,8
	2	119	18,7	18,7	24,6
	3	189	29,8	29,8	54,3
	4	216	34,0	34,0	88,3
	5 ve velké míře	74	11,7	11,7	100,0
	Total	635	100,0	100,0	

Tabulka 3.7 | Pomocné výpočty pro ordinální rozptyl (*přínos pro vstup do práce*)

i	P_i	$P_i (1 - P_i)$
1	0,066	0,0616
2	0,220	0,1716
3	0,510	0,2499
4	0,820	0,1476
Součet		**0,6307**

Tabulka 3.8 | Pomocné výpočty pro ordinální rozptyl (*přínos pro zvládání pracovních úkolů*)

i	P_i	$P_i (1 - P_i)$
1	0,058	0,0546
2	0,246	0,1855
3	0,543	0,2482
4	0,883	0,1033
Součet		**0,5916**

3.2.3 Kvantitativní proměnná

A. Míry polohy

Pro kvantitativní proměnnou lze kromě charakteristik určených pro ordinální proměnnou použít navíc **aritmetický průměr** $\bar{x}$. V případě diskrétní proměnné s malým počtem K variant hodnot se spočte podle vzorce

$$\bar{x} = \sum_{i=1}^{K} x_i p_i \,. \tag{3.5}$$

Vycházíme-li při výpočtu z původních pozorování (ne z tabulky četností), zapisujeme vzorec pro aritmetický průměr ve tvaru

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{3.6}$$

kde n je počet statistických jednotek.

V části 3.2.2 bylo vysvětleno zjištění **mediánu** $\tilde{x}$ na základě tabulky rozdělení četností. Z původních pozorování se medián stanovuje buď jako prostřední hodnota uspořádaného souboru hodnot (je-li celkový počet hodnot n lichý), nebo jako průměr dvou hodnot nacházejících se uprostřed uspořádaného souboru (pokud je počet hodnot sudý). Ve druhém případě je tedy soubor rozdělen na dvě poloviny.

Kromě mediánu lze jako míry polohy použít i jiné charakteristiky, které rozdělují soubor uspořádaných hodnot na dvě části v určitém poměru. Tyto charakteristiky se obecně nazývají *kvantily*. Nejčastěji uváděnými kvantily jsou *kvartily*, k nimž patří výše zmíněný medián, dále *dolní kvartil* $\tilde{x}_{25}$, který rozděluje soubor v poměru 25 % a 75 %, a *horní kvartil* $\tilde{x}_{75}$ používaný k rozdělení v poměru 75 % a 25 %. Polohu hodnot charakterizují též *minimální* a *maximální* hodnota, tj. $x_{\min}$ a $x_{\max}$.

B. Míry variability

Základní mírou pro hodnocení variability je **rozptyl** s^2. Vyjdeme-li v případě diskrétní proměnné s malým počtem K variant hodnot z tabulky četností, pak

$$s^2 = \sum_{i=1}^{K} (x_i - \bar{x})^2 p_i = \sum_{i=1}^{K} x_i^2 p_i - \bar{x}^2 \,. \tag{3.7}$$

Pro výpočet z původních pozorování zapisujeme

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \,. \tag{3.8}$$

Odvozenými mírami jsou **směrodatná odchylka** s a **variační koeficient** V_x:

$$s = \sqrt{s^2}, \qquad (3.9)$$

$$V_x = \frac{s}{\overline{x}}.$$

Dalšími ze způsobů hodnocení variability jsou **variační rozpětí**, dané intervalem mezi minimální a maximální hodnotou ($x_{max} - x_{min}$), a **mezikvartilové rozpětí**, dané rozdílem mezi dolním a horním kvartilem ($\tilde{x}_{75} - \tilde{x}_{25}$).

Poznámka

Statistické programové systémy poskytují odhady charakteristik základního souboru, neboť se předpokládá, že jsou obvykle analyzovány údaje zjištěné ve *výběrovém souboru*. Jako odhad rozptylu v tomto souboru se používá *výběrový rozptyl*, který se spočte tak, že se ve jmenovateli vzorce (3.8) použije hodnota ($n - 1$). Při použití vzorce (3.7) vynásobíme výsledek podílem $n/(n - 1)$. Druhá odmocnina z výběrového rozptylu se nazývá *výběrová směrodatná odchylka*. Podrobněji viz oddíl 3.3.

C. Míry koncentrace

Pojem *koncentrace* byl již zmíněn v souvislosti s nominální proměnnou. U ordinální proměnné nebyly míry koncentrace uvedeny, i když existují. Jsou založeny na kvantilech, jde tedy o *kvantilové charakteristiky*. Tyto míry lze použít i pro kvantitativní proměnnou, častěji se však používají míry založené na podobném principu, jako je výpočet aritmetického průměru a rozptylu. K hodnocení koncentrace hodnot slouží *míry šikmosti* a *míry špičatosti*.

Předpokládejme proměnnou buď diskrétní s menším počtem hodnot, nebo kategorizovanou, odvozenou z proměnné s větším počtem různých hodnot (kategorie tedy zastupují určité intervaly hodnot). Dále pro jednoduchost předpokládejme, že existuje jedna modální kategorie a četnost ostatních kategorií (nižších i vyšších, než je kategorie modální) se směrem od této kategorie postupně snižuje. Pak můžeme rozlišit rozdělení *symetrické* (modální kategorie je prostřední a u nižších i vyšších kategorií je stejné rozdělení četností sestupně seřazené směrem od modální kategorie), *kladně zešikmené* (s nižší koncentrací vyšších kategorií) a *záporně zešikmené* (s menší koncentrací nižších kategorií). Tuto vlastnost hodnotí **koeficient šikmosti**. Nabude-li hodnoty nula, pak jde o rozdělení symetrické. Kladné hodnoty nabývá v případě kladně zešikmeného rozdělení, záporné hodnoty, je-li rozdělení záporně zešikmené.

Koeficienty špičatosti hodnotí, zda je rozdělení spíše ploché (záporné hodnoty), či spíše špičaté (kladné hodnoty). Pokud bude nenulová četnost pouze u jedné kategorie, jde o nejvyšší možnou koncentraci hodnot. V takovém případě se rozdělení označuje jako *špičaté*. Jsou-li kategorie v souboru zastoupeny rovnoměrně, pak jde o rozdělení *ploché*. Mezi těmito dvěma extrémními variantami existují další možnosti, při kterých se rozdělení podobá buď více první, nebo více druhé variantě.

Z důvodu zjednodušeného výkladu, použitého v této knize, bude dále uveden pouze příklad použití koeficientu šikmosti.

Příklad 3.8

Charakterizujte proměnné *C4* (*počet zaměstnání od absolvování studia*) a *A1_počet* (*počet typů zaměstnání před studiem a při studiu*), jejichž rozdělení četností uvádějí výstupy 3.19 a 3.20, z hlediska polohy a variability. U první proměnné jsou uvažováni pouze respondenti, kteří již byli po absolvování studia zaměstnáni a kteří otázku zodpověděli Proměnná *A1_počet* vyjadřuje počet jedniček v proměnných *A11*, *A12*, *A21* a *A22* (byla vytvořena pomocí možnosti *Count* v rámci nabídky *Transform*). Obě proměnné vyjadřují počty, jsou tedy kvantitativní.

Výstup 3.19 | Rozdělení četností pro proměnnou *C4*

Počet zaměstnání od absolvování studia					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	27	4,3	7,3	7,3
	2	196	30,9	53,0	60,3
	3	114	18,0	30,8	91,1
	4	30	4,7	8,1	99,2
	5	3	,5	,8	100,0
	Total	370	58,3	100,0	

Výstup 3.20 | Rozdělení četností pro proměnnou *A1_počet*

Počet typů zaměstnání před studiem a při studiu					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	17	2,7	2,7	2,7
	1	95	15,0	15,0	17,6
	2	258	40,6	40,6	58,3
	3	216	34,0	34,0	92,3
	4	49	7,7	7,7	100,0
	Total	635	100,0	100,0	

Obě proměnné mají stejný počet kategorií, ale liší se minimální a maximální hodnotou. Polohu charakterizuje například *modus*, jehož hodnota pro proměnnou *C4* je 2 (četnost 196) a pro proměnnou *A1_počet* také hodnota 2 (četnost 258). Další charakteristikou je *mediánová kategorie*, jejíž hodnoty jsou u obou proměnných také 2. Protože zjištěné kumulativní četnosti jsou u mediánové kategorie větší než 0,5, je tato kategorie současně

také *mediánem*. Obdobně bychom mohli určit dolní a horní kvartil. U obou proměnných je medián současně dolním kvartilem (kumulativní relativní četnost dosáhne hodnoty vyšší než 0,25 u kategorie s hodnotou 2). Horním kvartilem je v obou případech hodnota 3. *Aritmetický průměr* proměnné *C4* spočteme podle vzorce (3.5) jako

$$\overline{x} = \sum_{i=1}^{K} x_i p_i = 1 \cdot 0,073 + 2 \cdot 0,53 + 3 \cdot 0,308 + 4 \cdot 0,081 + 5 \cdot 0,008 = 2,421,$$

stejným postupem bychom získali pro proměnnou *A1_počet* aritmetický průměr 2,29.

Z měr variability si uvedeme výpočet rozptylu a směrodatné odchylky pro proměnnou *C4*. *Rozptyl* můžeme spočítat buď podle vztahu

$$s^2 = \sum_{i=1}^{K} (x_i - \overline{x})^2 p_i = (1 - 2,421)^2 \cdot 0,073 + (2 - 2,421)^2 \cdot 0,53 + (3 - 2,421)^2 \cdot 0,308 +$$
$$+ (4 - 2,421)^2 \cdot 0,081 + (5 - 2,421)^2 \cdot 0,008 = 0,6,$$

nebo podle vztahu

$$s^2 = \sum_{i=1}^{K} x_i^2 p_i - \overline{x}^2 = 1^2 \cdot 0,073 + 2^2 \cdot 0,53 + 3^2 \cdot 0,308 + 4^2 \cdot 0,081 + 5^2 \cdot 0,008 - 2,421^2 = 0,6.$$

Směrodatná odchylka je druhá odmocnina z rozptylu, tj. hodnota 0,774. Výběrový rozptyl je 370/369 · 0,6 = 0,601 a výběrová směrodatná odchylka 0,775.

Pro proměnnou *A1_počet* je rozptyl 0,822 a směrodatná odchylka 0,907. Výběrový rozptyl je 635/634 · 0,822 = 0,823 a výběrová směrodatná odchylka 0,907 (zaokrouhleno).

IBM SPSS Statistics

V systému *IBM SPSS Statistics* získáme základní charakteristiky volbou *Analyze*, *Descriptive Statistics* a *Frequencies* a výběrem potřebných charakteristik v rámci možnosti *Statistics*. Vypisují se míry polohy, variability a šikmosti, viz výstup 3.21.

Ve výstupu jsou pro analyzované proměnné uvedeny hodnoty následujících charakteristik: rozsah souboru (*N*) v členění na platné hodnoty (*Valid*) a chybějící hodnoty (*Missing*), aritmetický průměr (*Mean*), medián (*Median*), modus (*Mode*), výběrová směrodatná odchylka (*Std. Deviation*), výběrový rozptyl (*Variance*), koeficient šikmosti (*Skewness*), směrodatná chyba odhadu koeficientu šikmosti (*Std. Error of Skewness*) a kvartily (*Percentiles*). Hodnoty výběrového rozptylu a výběrové směrodatné odchylky se liší od „ručního" výpočtu, při němž byly použity zaokrouhlené relativní četnosti. Medián je ve výstupní tabulce uveden dvakrát, a to jednak pod názvem *Median*, jednak pod označením *Percentiles 50*.

Počet zaměstnání od absolvování studia je charakterizován kladně zešikmeným rozdělením (menší koncentrací vyšších hodnot), *počet typů zaměstnání před studiem a při studiu* naopak záporně zešikmeným rozdělením (menší koncentrací nižších hodnot). Koncentraci hodnot lze dobře ilustrovat např. sloupcovým grafem (viz graf 3.1).

Výstup 3.21 | Popisné charakteristiky pro proměnné *C4 a A1_počet*

Statistics		Počet zaměstnání od absolvování studia	Počet typů zaměstnání před studiem a při studiu
N	Valid	370	635
	Missing	265	0
Mean		2,42	2,29
Median		2,00	2,00
Mode		2	2
Std. Deviation		,776	,907
Variance		,602	,822
Skewness		,578	–,200
Std. Error of Skewness		,127	,097
Minimum		1	0
Maximum		5	4
Percentiles	25	2,00	2,00
	50	2,00	2,00
	75	3,00	3,00

3.2.4 Grafické zobrazení

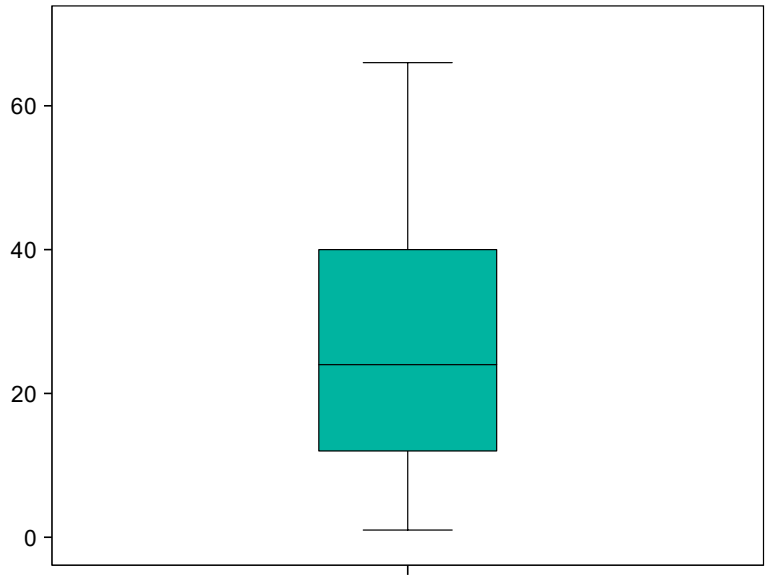
Stejně jako lze graficky znázornit rozdělení četnosti, je možné zobrazit též popisné charakteristiky. Některé grafy slouží k oběma účelům. Názornou představu o datech poskytuje *krabičkový graf*. V následujícím příkladu bude vysvětleno, které charakteristiky lze z tohoto grafu vyčíst.

Příklad 3.9

Charakterizujte pomocí krabičkového grafu hodnoty kvantitativních proměnných vyjadřujících počty měsíců, a to proměnných *A21pm* (*práce při studiu související se studiem*) a *A22pm* (*práce při studiu nesouvisející se studiem*).

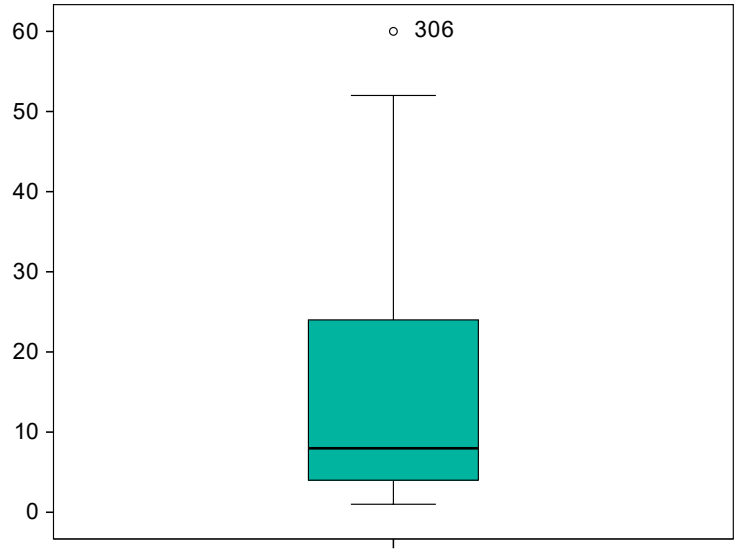
Tento graf získáme volbou *Analyze*, *Descriptive Statistics* a *Explore.* Výsledek zachycují grafy 3.4 a 3.5. Graf sestává z krabičky a „vousů". Je orientován vertikálně, takže potřebné údaje zjišťujeme na vertikální ose *Y*. Jsou to především *minimální* a *maximální hodnota*, dané začátkem dolního a koncem horního vousu (toto však platí pouze v případě, že neexistují odlehlé ani extrémní údaje, viz graf 3.4). Dále jsou v grafu zaznamenány dolní a horní kvartil, a to pomocí spodního a horního okraje krabičky, a medián, reprezentovaný úsečkou uvnitř krabičky (pokud tato hodnota není shodná s dolním nebo horním kvartilem).

Graf 3.4 | Krabičkový graf pro proměnnou *A21pm*



Práce při studiu související se studiem – počet měsíců

Graf 3.5 | Krabičkový graf pro proměnnou *A22pm*



Práce při studiu nesouvisející se studiem – počet měsíců

Graf 3.5 zobrazuje navíc jednu odlehlou hodnotu, která je od hodnoty horního kvartilu vzdálena o více než jeden a půl násobek kvartilového rozpětí (rozdíl mezi horním a dolním kvartilem). Číslo u této hodnoty označuje řádek datové matice, kde se hodnota nachází. Obě proměnné jsou kladně zešikmené, proměnná *A22pm* více.

Při větším rozsahu oboru hodnot bez odlehlých údajů můžeme říci, že v intervalu daném spodní a horní hranou krabičky (tj. dolním a horním kvartilem) se nachází 50 % hodnot, v intervalu od začátku dolního vousu ke spodní hraně krabičky 25 % hodnot a v intervalu od horní hrany krabičky ke konci horního vousu také 25 % hodnot.

3.3 Bodové a intervalové odhady

Cílem dotazníkových šetření je na základě odpovědí vybraných respondentů usoudit, jak lze charakterizovat celou populaci (*základní soubor*). Relativní četnosti a míry polohy a variability základního souboru se odhadují na základě hodnot zjištěných ve výběrovém souboru. Rozlišujeme odhady *bodové* a *intervalové*. K dané problematice je vhodné znát příslušnou teorii, stejně jako teorii pravděpodobnosti a základní typy rozdělení náhodné veličiny. Protože tato témata bývají zařazována do všech základních učebnic statistiky (viz např. [12], [20], [21] a [24]), budou dále uvedeny jen vzorce pro příslušné odhady.

Bodovým odhadem je jedna hodnota. **Intervalový odhad** může být buď oboustranný, nebo jednostranný (levostranný, pravostranný). V následujícím textu bude vždy uveden pouze interval oboustranný, který je vymezen dolní a horní mezí. Pravděpodobnost, že neznámá (populační) hodnota bude zahrnuta ve vymezeném intervalu, se nazývá *spolehlivost* a zapisuje se jako $(1 - \alpha)$. Hodnota α bude dále používána ve vzorcích bez dalšího komentáře.

3.3.1 Odhady relativních četností

Předpokládejme, že soubor o rozsahu n je výběrový a že v základním souboru jsou jednotlivé kategorie charakterizovány relativními četnostmi. Zaměřme se na jednu konkrétní kategorii – k-tou kategorii – s neznámou relativní četností π_k. Na základě výběrového souboru lze vypočíst *bodový odhad* této relativní četnosti. Je jím **relativní četnost** p_k, vypočtená z hodnot výběrového souboru, tj.

$$p_k = \frac{n_k}{n},$$

kde n_k je absolutní četnost zjištěná pro k-tou kategorii výběrového souboru.

Při výpočtu mezí *intervalového odhadu* záleží na velikosti výběrového souboru a na absolutní četnosti výskytu k-té sledované kategorie. Pro tzv. *velký výběr*, tj. (podle [31]) pro $n \geq 30$, $n_k \geq 5$ a $n - n_k \geq 5$, lze využít aproximaci binomického rozdělení rozdělením normálním a vypočíst dolní (p_D) a horní (p_H) mez oboustranného intervalu spolehlivosti podle vzorce (znaménko „+“ se vztahuje k p_H a „–“ k p_D):

$$p_{\mathrm{D,H}} = p_k \pm u_{1-\alpha/2} s_p ,$$

kde s_p je směrodatná chyba odhadu, která je dána vzorcem (pro výběr s opakováním)

$$s_p = \sqrt{\frac{p_k(1-p_k)}{n}} = \sqrt{\frac{n_k(n-n_k)}{n^3}} ,$$

a $u_{1-\alpha/2}$ je $100 \cdot (1 - \alpha/2)$procentní kvantil normovaného normálního rozdělení (způsob zjištění příslušné hodnoty je popsán v příloze této knihy). Postupy pro některé další typy výběrů podle velikosti jsou uvedeny v knize [31] a ve skriptech [34].

Příklad 3.10

Odhadněme podíl zaměstnaných absolventů, kteří mají druhé zaměstnání, a to bodovým a 95% intervalem spolehlivosti. Tabulka četností pro *počet zaměstnání od absolvování studia* je ve výstupu 3.19.

Bodovým odhadem je relativní četnost $p_2 = 0{,}53$, rozsah souboru je $n = 370$. Směrodatná chyba odhadu je

$$s_p = \sqrt{\frac{0{,}53 \cdot (1-0{,}53)}{370}} = 0{,}026 ,$$

dolní a horní mez 95% intervalu spolehlivosti pak

$$p_{\mathrm{D,H}} = p_3 \pm u_{0{,}975} s_p = 0{,}53 \pm 1{,}96 \cdot 0{,}026 = 0{,}53 \pm 0{,}05 , \text{ tj.}$$

$p_{\mathrm{D}} = 0{,}48$ a $p_{\mathrm{H}} = 0{,}58$. S pravděpodobností přibližně 0,95 je tedy populační podíl zaměstnaných absolventů, kteří mají druhé zaměstnání, zahrnut do intervalu od 0,48 do 0,58.

3.3.2 Odhady míry polohy

V tomto oddílu bude uveden odhad střední hodnoty $E(X)$ náhodné veličiny X za předpokladu, že rozsah výběru $n \geq 30$. *Bodovým odhadem* je **aritmetický průměr** $\bar{x}$, počítaný podle vzorce 3.5, resp. 3.6.

Při *intervalovém* odhadu se dolní ($\bar{x}_{\mathrm{D}}$)a horní ($\bar{x}_{\mathrm{H}}$) mez oboustranného intervalu spolehlivosti spočte podle vzorce

$$\bar{x}_{D,H} = \bar{x} \pm t_{1-\alpha/2}[n-1] \cdot s_{\bar{x}} ,$$

kde $t_{1-\alpha/2}[n-1]$ je $100 \cdot (1 - \alpha/2)$procentní kvantil Studentova t rozdělení s $(n - 1)$ stupni volnosti a $s_{\bar{x}}$ je směrodatná chyba odhadu, která je dána vzorcem

$$s_{\bar{x}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \frac{\hat{\sigma}}{\sqrt{n}} ,$$

kde $\hat{\sigma}^2$ je výběrový rozptyl, viz poznámka v oddílu 3.2.3, (též vzorec (3.12), resp. (3.13)), a $\hat{\sigma}$ je výběrová směrodatná odchylka, viz vzorec (3.14).

Příklad 3.11

Odhadněme střední hodnotu počtu měsíců práce při studiu související se studiem (proměnná *A21pm*), a to 95% intervalem spolehlivosti. Rozdělení četností je znázorněno pomocí grafu 3.4. Aritmetický průměr je $\bar{x}$ = 21,72, rozsah souboru je n = 430 a výběrová směrodatná odchylka je $\hat{\sigma}$ = 15,588. Směrodatná chyba odhadu je

$$s_{\bar{x}} = \frac{15,588}{\sqrt{430}} = 0,752\,,$$

dolní a horní mez 95% intervalu spolehlivosti pak

$$\bar{x}_{D,H} = \bar{x} \pm t_{0,975}[429] \cdot s_{\bar{x}} = 21,72 \pm 1,966 \cdot 0,752 = 21,72 \pm 1,478\,,$$

tj. $\bar{x}_D$ = 20,242 a $\bar{x}_H$ = 23,198. S pravděpodobností 0,95 se tedy neznámá střední honota počtu měsíců práce při studiu související se studiem nachází v intervalu od 20,242 do 23,198.

3.3.3 Odhady měr variability

Bodové odhady základních měr variability jsou následující. Pro nominální proměnnou počítáme základní výběrovou charakteristiku

$$\text{odhad } nomvar = M = \frac{n}{n-1}\sum_{i=1}^{K}(p_i(1-p_i))\,, \tag{3.10}$$

která se nazývá **míra mutability**. Obdobně pro ordinální proměnnou je

$$\text{odhad } dorvar = \frac{2n}{n-1}\sum_{i=1}^{K-1}(P_i(1-P_i))\,, \tag{3.11}$$

kde P_i jsou zjištěné kumulativní relativní četnosti.

Pro kvantitativní proměnnou odhadujeme rozptyl σ^2 pomocí **výběrového rozptylu**, který je definován jako

$$\hat{\sigma}^2 = \frac{n}{n-1}\sum_{i=1}^{K}(x_i - \bar{x})^2 p_i\,, \tag{3.12}$$

resp. v případě většího počtu variant hodnot jako

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\,. \tag{3.13}$$

Výběrovou směrodatnou odchylku můžeme vyjádřit pomocí vztahu

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \tag{3.14}$$

a **odhad variačního koeficientu** jako

$$\hat{V}_x = \frac{\hat{\sigma}}{\bar{x}}\,. \tag{3.15}$$

Intervalový odhad uvedeme pro rozptyl σ^2 za předpokladu, že jde o parametr normálního rozdělení. Pro něj platí, že s pravděpodobností (1 – α) se jeho hodnota nachází v intervalu daném vztahem

$$\frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}[n-1]} < \sigma^2 < \frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2}[n-1]},$$

kde $\chi^2_{1-\alpha/2}[n-1]$ je $100 \cdot (1-\alpha/2)$procentní kvantil chí-kvadrát rozdělení s $(n-1)$ stupni volnosti.

Příklad 3.12

Odhadněme variabilitu počtu měsíců práce při studiu související se studiem (proměnná *A21pm*) vyjádřenou rozptylem, a to bodovým a 95% intervalem spolehlivosti. Rozdělení četností je znázorněno v grafu 3.4. Bodovým odhadem je výběrový rozptyl $\hat{\sigma}^2 = 242{,}998$, rozsah souboru je $n = 430$. Dolní mez 95% intervalu spolehlivosti spočteme jako

$$\frac{(n-1)\hat{\sigma}^2}{\chi^2_{0{,}975}[429]} = \frac{429 \cdot 242{,}998}{488{,}281} = 213{,}496,$$

horní mez tohoto intervalu jako

$$\frac{(n-1)\hat{\sigma}^2}{\chi^2_{0{,}025}[429]} = \frac{429 \cdot 242{,}998}{373{,}507} = 279{,}101.$$

S pravděpodobností 0,95 se tedy neznámá populační hodnota rozptylu počtu měsíců práce při studiu související se studiem nachází v intervalu od 213,496 do 279,101.

3.4 Testování hypotéz o četnostech kategorií

Důležitou součástí analýzy dat je testování statistických hypotéz. Stejně jako v případě odhadů se na základě odpovědí vybraných respondentů usuzuje, jak lze charakterizovat celou populaci (*základní soubor*). V této knize bude výklad zaměřen pouze na testy týkající se četností kategorií. Připomeňme si ve stručnosti základní principy testování hypotéz.

Stanovují se vždy dvě hypotézy, a to *testovaná* (*nulová*) H_0 a *alternativní* H_1. Nejjednodušší případy jsou takové, kdy testujeme, zda se některý parametr určitého rozdělení rovná zadané hodnotě. Vůči této jednoduché nulové hypotéze se staví hypotéza alternativní, která může být *oboustranná* (parametr se dané hodnotě nerovná), případně *jednostranná*. Tehdy se rozlišuje hypotéza *levostranná* (parametr je menší než zadaná hodnota) a *pravostranná* (parametr je větší než tato hodnota).

Cílem testování nulové hypotézy je dospět k úsudku, zda můžeme, či nemůžeme tuto hypotézu zamítnout vzhledem ke stanovené hypotéze alternativní (o konkrétním způsobu, jakým se k úsudku dospěje, bude pojednáno později). Vzhledem k tomu, že jsou možné dvě odpovědi na vytčenou otázku, mohou nastat dva druhy chyb. Zamítneme-li nulovou hypotézu,

která ve skutečnosti platí, pak se hovoří o *chybě prvního druhu*. Pokud nulovou hypotézu nezamítneme, ale tato hypotéza ve skutečnosti neplatí, jde o *chybu druhého druhu*.

Pravděpodobnost chyby prvního druhu se nazývá *hladina významnosti*. Značí se symbolem α, který již byl používán v předchozím oddílu. Pravděpodobnost chyby druhého druhu se označuje symbolem β. Pravděpodobnost, že testovanou hypotézu zamítneme správně, je dána vztahem 1 – β a nazývá se *síla testu*. Pro vlastní testování je třeba kromě dvou hypotéz stanovit též hladinu významnosti. Obvykle se připouští pravděpodobnost chyby prvního druhu $\alpha = 0{,}05$, případně $\alpha = 0{,}01$[2].

Hlavním prostředkem je *testové kritérium*, jímž je určitá vhodná statistika, která má při platnosti nulové hypotézy známé pravděpodobnostní rozdělení. Její definiční obor se rozdělí na dvě disjunktní části, kterými jsou *kritický obor* a *obor přijetí*. Kritický obor je ta část definičního oboru dané statistiky, pro jehož hodnoty platí, že pravděpodobnost jejich výskytu je velmi malá. V případě jednoduché nulové hypotézy pak podle typu alternativní hypotézy tento obor tvoří buď extrémně nízké i vysoké hodnoty (v případě oboustranné alternativní hypotézy), nebo pouze extrémně nízké (pro levostrannou hypotézu), resp. pouze extrémně vysoké (pro hypotézu pravostrannou). Pokud je nulová hypotéza složená (testuje se např. současně více relativních četností), pak se obvykle uvažují rozdělení, jejichž definičním oborem jsou nezáporná čísla a kritický obor tvoří extrémně vysoké hodnoty.

Hodnoty na rozhraní oboru přijetí a kritického oboru se nazývají *kritické hodnoty*. Při jejich zjišťování se bere v úvahu stanovená hladina významnosti. V případě jednoduché nulové hypotézy jde o $100 \cdot (1 - \alpha/2)\%$ kvantil daného rozdělení, resp. $100 \cdot (1 - \alpha)\%$ kvantil či $100\alpha\%$ kvantil pro jednostranné alternativní hypotézy. Pro složenou testovanou hypotézu se zjišťuje $100 \cdot (1 - \alpha)\%$ kvantil. Podstata testování pak spočívá ve zjištění, zda spočtená hodnota testového kritéria se nachází v oboru přijetí (tehdy H_0 nezamítáme), nebo v kritickém oboru (pak H_0 zamítáme).

V programových systémech bývá ve výsledcích uváděna tzv. *P-hodnota*, což je minimální hladina významnosti, od které můžeme zamítnout nulovou hypotézu. V dalším textu bude značena symbolem α'. Při jejím výpočtu se postupuje tak, že se zjistí, jakým kvantilem je hodnota x testového kritéria (zjistí se hodnota distribuční fukce $F(x)$). Na základě odpovídajícího vztahu, např. $(1 - \alpha'/2) = F(x)$, vypočteme hodnotu α'. Pokud $\alpha' \leq \alpha$, pak H_0 zamítneme, platí-li, že $\alpha' > \alpha$, H_0 nezamítáme.

3.4.1 Binomický test

Pro k-tou kategorii můžeme testovat hypotézu H_0: $\pi_k = \pi_{k,0}$ vůči oboustranné alternativní hypotéze H_1: $\pi_k \neq \pi_{k,0}$, příp. proti jednostranné alternativní hypotéze (levostranné či pravostranné). Sdružíme-li ostatní kategorie do jedné, získáme tak dvě kategorie. Dále budeme testovanou kategorii označovat pomocí indexu 1, druhou kategorii pomocí indexu 2. To znamená, že testujeme hypotézu H_0: $\pi_1 = \pi_{1,0}$.

2 Snížení pravděpodobnosti prvního druhu má za následek zvýšení pravděpodobnosti druhého druhu. Při volbě $\alpha = 0{,}05$ je také přijatelná pravděpodobnost chyby druhého druhu.

Pro $(n_1 + n_2) \leq 25$ se používá exaktní text s využitím binomického rozdělení, pro $(n_1 + n_2) > 25$ lze (dle dokumentace programového systému *IBM SPSS Statistics*) použít aproximaci normovaným normálním rozdělením. Je-li $\pi_{1,0} = 0{,}5$, pak můžeme nulovou hypotézu zapsat jako H_0: $\pi_1 = \pi_2$ a alternativní jako H_1: $\pi_1 \neq \pi_2$. V případě, že $n_1 \leq n_2$ a $p_1 < \pi_{1,0}$, je alternativní hypotéza stanovena jako H_1: $\pi_1 < \pi_{1,0}$. Při *exaktním testu* se spočte minimální hladina významnosti α', která se porovnává se zvolenou hladinou významnosti α. V případě *levostranné alternativní hypotézy* se α' spočte podle vztahu

$$\alpha' = \sum_{i=0}^{n_1} \binom{n}{i} (\pi_{1,0})^i (1-\pi_{1,0})^{n-i}. \tag{3.16}$$

Jde tedy o výpočet pravděpodobnosti, že náhodná veličina s binomickým rozdělením nabude hodnoty n_1 nebo menší, tzn. o hodnotu distribuční funkce v bodě n_1, což budeme značit jako $F(n_1)$[3]. Přitom pravděpodobnost, že nastane sledovaný náhodný jev, je $\pi_{1,0}$ a počet náhodných pokusů je n, tj. rozdělení, lze označit jako $Bi\ [n, \pi_{1,0}]$.

Pro $n_1 \leq n_2$ a $p_1 > \pi_{1,0}$ je alternativní hypotéza stanovena jako H_1: $\pi_1 > \pi_{1,0}$. Minimální hladina významnosti, od níž zamítáme nulovou hypotézu vůči *pravostranné alternativní hypotéze*, se exaktně počítá podle vztahu

$$\alpha' = 1 - F(n_1 - 1) = 1 - \sum_{i=0}^{n_1-1} \binom{n}{i} (\pi_{1,0})^i (1-\pi_{1,0})^{n-i}.$$

Jde tedy o výpočet pravděpodobnosti, že náhodná veličina s binomickým rozdělením nabude hodnoty n_1 nebo větší, což lze též psát též jako

$$\alpha' = \sum_{i=n_1}^{n} \binom{n}{i} (\pi_{1,0})^i (1-\pi_{1,0})^{n-i}. \tag{3.17}$$

V případě *oboustranné alternativní hypotézy* se minimální hladina významnosti, při které zamítáme nulovou hypotézu, spočte jako dvojnásobek minimální hladiny významnosti pro jednostrannou alternativní hypotézu, odpovídající jedné z výše uvedených variant.

Při testování *nulové hypotézy o shodě podílů* vůči oboustranné alternativní hypotéze je za předpokladu $n_1 < n_2$ tato hladina významnosti dána vztahem

$$\alpha' = 2F(n_1) = 2\sum_{i=0}^{n_1} \binom{n}{i} (0{,}5)^n. \tag{3.18}$$

Jde tedy o dvojnásobek minimální hladiny významnosti pro levostrannou alternativní hypotézu, neboť $\pi_{1,0} = 0{,}5$, a tudíž za uvedeného předpokladu platí, že $p_1 < \pi_{1,0}$.

Při *aproximaci normovaným normálním rozdělením* se vychází z toho, že střední hodnota binomického rozdělení je $n\pi$ a rozptyl $n\pi\ (1 - \pi)$, kde n a π jsou parametry tohoto rozdělení. Při velkých výběrech lze místo minima ze zjištěných četností uvažovat normovanou náhodnou veličinu. Ta je stanovena tak, že se od minima (příp. maxima) četností odečte střední hodnota a získaný rozdíl se dělí směrodatnou odchylkou (tzn. odmocni-

3 Hodnoty distribuční a pravděpodobnostní funkce binomického rozdělení lze také získat jako výsledky funkcí v systému *IBM SPSS Statistics*, viz příloha této knihy.

nou z rozptylu). Protože v systému *IBM SPSS Statistics* se tato aproximace využívá pro $(n_1 + n_2) > 25$, používá se tzv. oprava na spojitost, kdy se v čitateli přičítá nebo odečítá (viz níže) hodnota 0,5. Výsledná náhodná veličina má za platnosti nulové hypotézy přibližně normované normální rozdělení.

Pro $n_1 \leq n_2$ a $p_1 < \pi_{1,0}$, tedy při *levostranné alternativní hypotéze*, se používá veličina Z_1 s přibližně normovaným normálním rozdělením, daná vztahem

$$Z_1 = \frac{n_1 - n\pi_{1,0} + 0,5}{\sqrt{n\pi_{1,0}(1-\pi_{1,0})}}. \tag{3.19}$$

Jde o kvantil $u_{\alpha'}$, kde α' je minimální hladina významnosti. Platí tedy, že $\alpha' = \Phi(Z_1)$, kde $\Phi(Z_1)$ je hodnota distribuční funkce normovaného normálního rozdělení v bodě Z_1. Pro $n_1 < n_2$ a *pravostrannou alternativní hypotézu* se hodnota 0,5 odečítá, tj.

$$Z_2 = \frac{n_1 - n\pi_{1,0} - 0,5}{\sqrt{n\pi_{1,0}(1-\pi_{1,0})}}, \tag{3.20}$$

a minimální hladina významnosti se spočte na základě vztahu $\alpha' = 1 - \Phi(Z_2)$.

Pro $n_1 > n_2$ a $p_1 > \pi_{1,0}$, tedy pro *pravostrannou alternativní hypotézu*, se použije vzorec (3.20). Pro $n_1 > n_2$ a *levostrannou alternativní hypotézu* pak vzorec (3.19).

Při testování nulové hypotézy o shodě podílů vůči oboustranné alternativní hypotéze se pro $n_1 < n_2$ postupuje tak, že se spočte hodnota Z_1 podle vzorce (3.19). Jde o kvantil $u_{\alpha'/2}$, kde α' je minimální hladina významnosti, od které zamítáme nulovou hypotézu o shodě podílů u daných dvou kategorií. Tuto hladinu významnosti tedy vypočteme jako $\alpha' = 2\Phi(Z_1)$. Pro $n_1 > n_2$ se spočte hodnota Z_2 podle vzorce (3.20), což je kvantil $u_{1-\alpha'/2}$ a minimální hladina významnosti je $\alpha' = 2(1 - \Phi(Z_2))$.

Poznámka

Některé programové systémy nepoužívají korekci přičtením nebo odečtením hodnoty 0,5 v čitateli, takže se hodnoty získané pomocí binomického rozdělení a hodnoty pomocí aproximace více liší. Obvykle se ovšem předpokládá, že se aproximace používá pro výběr o rozsahu splňujícím buď podmínku uvedenou v části 3.3.1, nebo podmínku s využitím očekávané četnosti $\pi_{1,0}$, která je formulována jako $n\pi_{1,0}(1 - \pi_{1,0}) > 9$.

Příklad 3.13

Z pěti náhodně vybraných absolventů pouze jeden uvedl, že byl zaměstnán před studiem a že jeho práce souvisela se studiem. Označme odpověď *ano* na otázku týkající se zaměstnání před studiem související se studiem (s podílem 20 %) jako první kategorii a odpověď *ne* jako kategorii druhou.

A. Můžeme na základě tak malého počtu zamítnout hypotézu o shodě podílů pro dvě sledované kategorie v souboru všech absolventů v daném období?
B. Můžeme zamítnout H_0: $\pi_1 = 0,4$ vůči levostranné alternativní hypotéze?
C. Můžeme zamítnout H_0: $\pi_1 = 0,1$ vůči pravostranné alternativní hypotéze?

Řešení otázky A. Náhodná veličina vyjadřující počet výskytů odpovědí *ano* má binomické rozdělení s parametry $\pi_{1,0} = 0{,}5$ a $n = 5$. Dále víme, že $n_1 = 1$ a $n_2 = 4$. Hladinu významnosti spočteme podle vzorce (3.18) jako

$$\alpha' = 2\sum_{i=0}^{n_1}\binom{n}{i}(0{,}5)^n = 2\sum_{i=0}^{1}\binom{5}{i}(0{,}5)^5 = 0{,}375\,.$$

Protože tato hodnota je větší než 0,05, na 5% hladině významnosti nezamítáme nulovou hypotézu o shodě podílů.

IBM SPSS Statistics

V *IBM SPSS Statistics* připravíme data analogicky podle tabulky 3.1, tj. do nového souboru do prvního sloupce vložíme označení kategorií „1“ a „2“ a do druhého sloupce četnosti „1“ a „4“. První proměnnou definujeme jako nominální (sloupec *Measure* v datovém editoru). Jejím kategoriím přiřadíme váhy podle postupu popsaného v oddílu 3.1.1. Můžeme přiřadit rozšířený název proměnné a popisy kódů kategorií (1 = „ano“, 2 = „ne“).

Pro testování zadáme *Analyze*, *Nonparametric Tests* a *One Sample*. V listu *Fields* vybereme první proměnnou (v *Test Fields* bude pouze tato proměnná) a v listu *Settings* zvolíme *Customize tests*, *Compare observed binary probability to hypothesized (Binomial test)*. Vycházíme z původního nastavení s podílem 0,5, takže již stačí jen použít tlačítko *Run*.

Jako výsledky z příslušné procedury (viz výstup 3.22) získáme zadání úlohy a minimální hladinu významnosti, od které zamítáme hypotézu H_0 (sloupec *Sig.*). Tato hladina významnosti je 0,375; shoduje se s hodnotou vypočtenou podle vzorce. Ve sloupci *Decision* je odpověď pro 5% hladinu významnosti, nulovou hypotézu nezamítáme. Otevřením okna *Model Viewer* (dvojitým kliknutím na levé tlačítko myši, kurzor je umístěn na zobrazeném základním výstupu) lze obdržet další části výstupu.

Výstup 3.22 | Výsledek binomického testu k příkladu 3.13, otázka A

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by zaměstnání před studiem související se studiem = ano and ne occur with probabilities 0,5 and 0,5.	One-Sample Binomial Test	,375[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1] Exact significance is displayed for this test.

Řešení otázky B. Podle vzorce (3.16) spočteme hladinu významnosti

$$\alpha' = F(n_1) = \sum_{i=0}^{n_1}\binom{n}{i}(\pi_{1,0})^i(1-\pi_{1,0})^{n-i} = \sum_{i=0}^{1}\binom{5}{i}(0{,}4)^i(0{,}6)^{5-i} = 0{,}337\,.$$

Protože tato hodnota je větší než 0,05, na 5% hladině významnosti nezamítáme nulovou hypotézu vůči levostranné alternativní hypotéze.

IBM SPSS Statistics

V *IBM SPSS Statistics* postupujeme stejně jako u otázky A, a navíc v rámci možnosti (tlačítka) *Options* zadáme v části *Hypothesized proportion* hodnotu 0,4. Získáme výsledek uvedený ve výstupu 3.23. Minimální hladina významnosti se shoduje s hodnotou vypočtenou podle vzorce.

Výstup 3.23 | Výsledek binomického testu k příkladu 3.13, otázka B

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by zaměstnání před studiem související se studiem = ano and ne occur with probabilities 0,4 and 0,6.	One-Sample Binomial Test	,337[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1] Exact significance is displayed for this test.

Řešení otázky C. Podle vzorce (3.17) spočteme hladinu významnosti

$$\alpha' = \sum_{i=n_1}^{n} \binom{n}{i} (\pi_{1,0})^i (1-\pi_{1,0})^{n-i} = \sum_{i=1}^{5} \binom{5}{i} (0,1)^i (0,9)^{5-i} = 0,41.$$

Protože tato hodnota je větší než 0,05, na 5% hladině významnosti nezamítáme nulovou hypotézu vzhledem k pravostranné alternativní hypotéze.

IBM SPSS Statistics

Pomocí *IBM SPSS Statistics* (v části *Hypothesized proportion* zadáme hodnotu 0,1) získáme výsledek uvedený ve výstupu 3.24. Minimální hladina významnosti se shoduje s hodnotou vypočtenou podle vzorce.

Výstup 3.24 | Výsledek binomického testu k příkladu 3.13, otázka C

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by zaměstnání před studiem související se studiem = ano and ne occur with probabilities 0,1 and 0,9.	One-Sample Binomial Test	,410[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1] Exact significance is displayed for this test.

Příklad 3.14

U 533 zaměstnaných absolventů bylo zjištěno, že 212 (tj. 39,8 %) z nich působí v prvním zaměstnání do současné doby. Chceme zodpovědět následující otázky, týkající se základního souboru.

A. Můžeme zamítnout hypotézu o shodě podílů pro odpovědi *ne* (první kategorie) a *ano*?
B. Můžeme zamítnout H_0: $\pi_1 = 0{,}55$ vůči pravostranné alternativní hypotéze?
C. Můžeme zamítnout H_0: $\pi_1 = 0{,}65$ vůči levostranné alternativní hypotéze?

Řešení otázky A. Protože $n_1 > n_2$, použijeme vzorec (3.20), tj.

$$Z_2 = \frac{321 - 533 \cdot 0{,}5 - 0{,}5}{\sqrt{533 \cdot 0{,}5 \cdot (1 - 0{,}5)}} = 4{,}678.$$

Jde o kvantil $u_{0{,}9999986}$, tudíž $1 - \alpha'/2 = 0{,}9999986$. Minimální hladina významnosti, od které zamítáme hypotézu H_0, je tedy $\alpha' = 2 \cdot (1 - 0{,}9999986) = 0{,}0000028$. Protože tato hodnota je menší než 0,01, na 1% hladině významnosti zamítáme nulovou hypotézu o shodě podílů.

IBM SPSS Statistics

V souboru absolventů provedeme výběr pouze zaměstnaných absolventů, tj. zadáme *Data*, *Select Cases*, vybereme možnost *If condition is satisfied*, použijeme tlačítko *If* a do řádku zadáme „C5 = 1“. Pro testování zadáme *Analyze*, *Nonparametric Tests* a *One Sample*. V listu *Fields* zrušíme v části *Test Fields* výběr všech proměnných a ponecháme v ní pouze proměnnou *působení v prvním zaměstnání do současné doby* (označení proměnné *D1* není v nabídce standardně zobrazováno). V listu *Settings* zvolíme *Customize tests*, *Compare observed binary probability to hypothesized (Binomial test)*. Po zvolení možnosti *Options* v části *Define Success for Categorical Fields* zvolíme *Specify success values* a zadáme hodnotu 0 (kód kategorie *ne*). Vycházíme z původního nastavení s podílem 0,5, takže po návratu do listu *Settings* již stačí jen zvolit *Run*.

Jako výsledky z příslušné procedury (viz výstup 3.25) získáme zadání úlohy a minimální hladinu významnosti, od které zamítáme hypotézu H_0 (sloupec *Sig.*). Tato hladina významnosti se standardně zobrazuje pouze na tři desetinná místa, tj. výsledek je 0,000. Ve sloupci *Decision* je odpověď pro 5% hladinu významnosti, nulovou hypotézu zamítáme. Další výstup, který lze získat v okně *Model Viewer*, obsahuje například hodnotu testové statistiky (*Standardized Test Statistic*), která se shoduje s vypočtenou hodnotou Z_2 (4,678).

Výstup 3.25 | Výsledek binomického testu k příkladu 3.14, otázka A

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by působení v prvním zaměstnání do současné doby = (ne) and (ano) occur with probabilities 0,5 and 0,5.	One-Sample Binomial Test	,000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Řešení otázky B. Protože $n_1 > n_2$ a $p_1 > \pi_{1,0}$, použijeme vzorec (3.20), tj.

$$Z_2 = \frac{321 - 533 \cdot 0,55 - 0,5}{\sqrt{533 \cdot 0,55 \cdot (1 - 0,55)}} = 2,381 .$$

Jde o kvantil $u_{0,991}$, tudíž $\alpha' = 1 - 0,991 = 0,009$. Nulovou hypotézu $\pi_1 = 0,55$ vůči pravostranné alternativní hypotéze můžeme zamítnout na 1% hladině významnosti.

IBM SPSS Statistics

V *IBM SPSS Statistics* postupujeme stejně jako u otázky A, a navíc v rámci možnosti *Options* zadáme v části *Hypothesized proportion* hodnotu 0,55. V části *Define Success for Categorical Fields* by mělo být zvoleno *Specify success values* a zadána hodnota 0 (kód kategorie *ne*). Získáme výsledek uvedený ve výstupu 3.26. Minimální hladina významnosti se shoduje s hodnotou stanovenou bez použití programu. Další výstup v okně *Model Viewer* obsahuje hodnotu testové statistiky, která se shoduje s vypočtenou hodnotou Z_2 (2,381).

Výstup 3.26 | Výsledek binomického testu k příkladu 3.14, otázka B

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by působení v prvním zaměstnání do současné doby = (ne) and (ano) occur with probabilities 0,55 and 0,45.	One-Sample Binomial Test	,009	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Řešení otázky C. Protože $n_1 > n_2$ a $p_1 < \pi_{1,0}$, použijeme vzorec (3.19), tj.

$$Z_1 = \frac{321 - 533 \cdot 0,65 + 0,5}{\sqrt{533 \cdot 0,65 \cdot (1 - 0,65)}} = -2,266 .$$

Jde o kvantil $u_{0,012}$, tudíž $\alpha' = 0,012$. Na 5% hladině významnosti zamítáme nulovou hypotézu $\pi_1 = 0,65$ ve prospěch levostranné alternativní hypotézy. Na 1% hladině významnosti však tuto nulovou hypotézu zamítnout nemůžeme.

IBM SPSS Statistics

Pomocí *IBM SPSS Statistics* (v části *Hypothesized proportion* zadáme hodnotu 0,65) získáme výsledek uvedený ve výstupu 3.27. Minimální hladina významnosti se shoduje s hodnotou stanovenou bez použití programu (a hodnota testové statistiky *Standardized Test Statistic* v okně *Model Viewer* se shoduje s vypočtenou hodnotou Z_1).

Výstup 3.27 | Výsledek binomického testu k příkladu 3.14, otázka C

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories defined by působení v prvním zaměstnání do současné doby = (ne) and (ano) occur with probabilities 0,65 and 0,35.	One-Sample Binomial Test	,012	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

3.4.2 Chí-kvadrát test dobré shody

Testujeme hypotézu H_0: $\pi_i = \pi_{i,0}$, kde $i = 1, 2, ..., K$ (K je počet kategorií) a $\sum\pi_{i,0} = 1$, vůči alternativní hypotéze H_1: alespoň pro jedno i platí, že $\pi_i \neq \pi_{i,0}$. Pokud se konstanty $\pi_{i,0}$ rovnají, pak lze nulovou hypotézu zapsat jako H_0: $\pi_1 = \pi_2 = ... = \pi_K$. Pro $n\pi_{i,0} \geq 5$ se používá statistika chí-kvadrát daná vztahem

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - n\pi_{i,0})^2}{n\pi_{i,0}}, \tag{3.21}$$

kde $n\pi_{i,0}$ je teoretické (očekávané) obsazení i-té kategorie při výběru o rozsahu n.

Tato náhodná veličina má za předpokladu, že platí hypotéza H_0, přibližně chí-kvadrát rozdělení s $(K - 1)$ stupni volnosti, tj. $\chi^2 \approx \chi^2[K - 1]$. Vypočtená hodnota je kvantilem $\chi_{1-\alpha'}[K - 1]$. Získanou hodnotu α' porovnáme se zvolenou hodnotou α.

Příklad 3.15

Očekávali jsme, že 79 % absolventů bude mít v prvním zaměstnání smlouvu na dobu neurčitou, 15 % na dobu určitou a 6 % bude pouze OSVČ. Porovnáme nyní, zda zjištěné údaje (proměnná *C2*, viz výstup 3.16) jsou v souladu s naším předpokladem.

Hodnotu statistiky chí-kvadrát spočteme podle vzorce (3.21), tj.

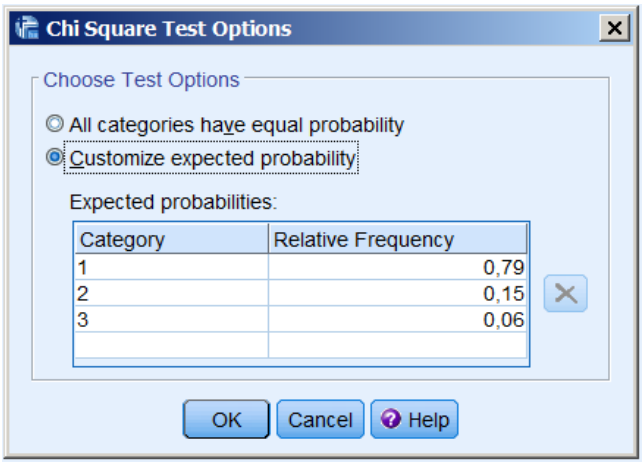
$$\chi^2 = \frac{(467 - 490{,}59)^2}{490{,}59} + \frac{(106 - 93{,}15)^2}{93{,}15} + \frac{(48 - 37{,}26)^2}{37{,}26} = 6{,}003\,.$$

Jde o kvantil $\chi^2_{0{,}95}$ [2], to znamená, že $1 - \alpha' = 0{,}95$, a tedy $\alpha' = 0{,}05$. Na 5% hladině významnosti zamítáme nulovou hypotézu o shodě rozdělení typu smlouvy v prvním zaměstnání s očekávaným předpokladem.

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests* a *One Sample*. V listu *Fields* zadáme proměnnou *typ smlouvy v prvním zaměstnání* (*C2*), v listu *Settings* zvolíme *Customize tests*, *Compare observed probabilities to hypothesized (Chi-Square test)* a využijeme možnost (tlačítko) *Options*. V zobrazeném panelu zvolíme *Customize expected probability* a zadáme očekávané četnosti podle návodu na obrázku 3.5.

Obrázek 3.5 | Ukázka definování očekávaných četností pro chí-kvadrát test dobré shody



Jako základní výsledky z příslušné procedury (viz výstup 3.28) získáme obecné zadání úlohy a minimální hladinu významnosti, od které zamítáme hypotézu H_0 (sloupec *Sig.*). Tato hladina významnosti je 0,05 a shoduje se tedy s hodnotou stanovenou bez použití programu. Ve sloupci *Decision* je odpověď pro 5% hladinu významnosti, nulovou hypotézu zamítáme.

Výstup 3.28 | Základní výstup pro chí-kvadrát test dobré shody

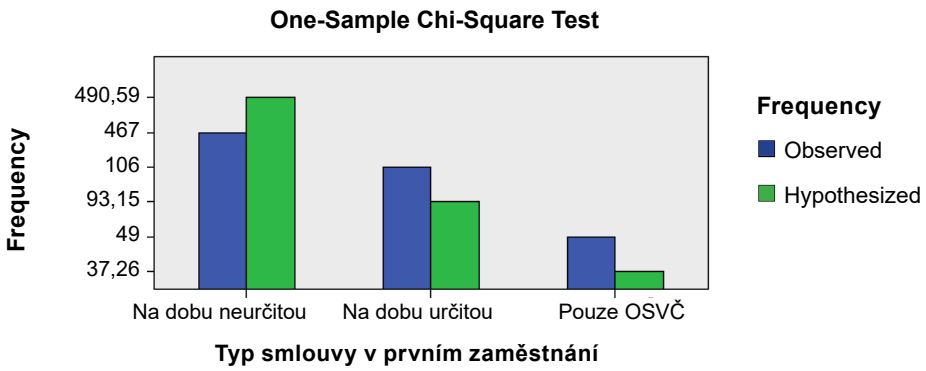
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories of typ smlouvy v prvním zaměstnání occur with the specified probabilities.	One-Sample Chi-Square Test	,050	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Další části výstupu (viz výstup 3.29) obdržíme v okně *Model Viewer* (přejdeme do něj dvojitým kliknutím na levé tlačítko myši, s kurzorem umístěným na zobrazeném základním výstupu). Zařazen je sloupcový graf s pozorovanými (*Observed*) a očekávanými (*Hypothesized*) absolutními četnostmi a tabulka obsahující celkový počet pozorování (*Total N*), hodnotu statistiky chí-kvadrát (*Test Statistic*), počet stupňů volnosti (*Degrees of Freedom*) a minimální hladinu významnosti (*Asymptotic Sig. (2-sided test)*).

Výstup 3.29 | Volitelný výstup pro chí-kvadrát test dobré shody



Total N	621
Test Statistic	6,003
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	,050

1. There are 0 cells (0%) with expected values less than 5. The minimum expected value is 37,260.

3.4.3 Testy porovnávající četnosti dvou kategorií

Testujeme hypotézu H_0: $\pi_i = \pi_j$ vůči oboustranné alternativní hypotéze H_1: $\pi_i \neq \pi_j$. Uvažujme pouze ty objekty (případy), které jsou z hlediska sledované proměnné zařazeny v *i*-té nebo *j*-té kategorii. V tomto redukovaném souboru označme pořadí kategorií 1 a 2. Platí, že $n_1 + n_2 = n$. Nulovou hypotézu pak zapíšeme ve tvaru H_0: $\pi_1 = \pi_2$, neboli H_0: $\pi_1 = 0{,}5$. Tím se dostáváme k binomickému testu, viz oddíl 3.4.1. Pro $(n_1 + n_2) > 25$ lze použít aproximaci normovaným normálním rozdělením, viz vzorec (3.19).

V případě $n \geq 30$ lze použít *aproximaci rozdělením chí-kvadrát*. Vztah (3.21) pro výpočet statistiky chí-kvadrát lze zjednodušit na

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_i - n \cdot 0{,}5)^2}{n \cdot 0{,}5} = \frac{\left(n_1 - \frac{n_1 + n_2}{2}\right)^2 + \left(n_2 - \frac{n_1 + n_2}{2}\right)^2}{\frac{n_1 + n_2}{2}} =$$

$$= \frac{\left(\frac{n_1 - n_2}{2}\right)^2 + \left(\frac{n_2 - n_1}{2}\right)^2}{\frac{n_1 + n_2}{2}} = \frac{\frac{2 \cdot (n_1 - n_2)^2}{4}}{\frac{n_1 + n_2}{2}} = \frac{(n_1 - n_2)^2}{n_1 + n_2}. \qquad (3.22)$$

Minimální hladinu významnosti spočteme podle vztahu $\alpha' = 1 - F(\chi^2)$, kde distribuční funkce $F(\chi^2)$ se zjišťuje pro rozdělení $\chi^2[1]$.

Jestliže $n > 36$ (v souladu se známou podmínkou $n\pi(1-\pi) > 9$), lze aplikovat *aproximaci normovaným normálním rozdělením*. Spočteme statistiku Z podle vztahu

$$Z = \frac{n_1 - n \cdot 0,5}{\sqrt{n \cdot 0,5 \cdot (1-0,5)}}. \tag{3.23}$$

Vypočtená hodnota uvedeného testového kritéria Z je kvantilem $u_{1-\alpha'/2}$, resp. $u_{\alpha'}$ či $u_{1-\alpha'}$ (v případě jednostranných alternativních hypotéz).

Umocněním na druhou této náhodné veličiny získáme náhodnou veličinu s rozdělením chí-kvadrát s jedním stupněm volnosti. Platí tedy

$$Z^2 = \left(\frac{n_1 - n \cdot 0,5}{\sqrt{n \cdot 0,5 \cdot (1-0,5)}} \right)^2 = \frac{\left(n_1 - \frac{n_1 + n_2}{2} \right)^2}{\frac{n_1 + n_2}{4}} = \frac{(n_1 - n_2)^2}{n_1 + n_2} = \chi^2 .$$

Minimální hladina významnosti, od které zamítáme nulovou hypotézu, je tedy jak v případě aproximace chí-kvadrát rozdělením, tak v případě aproximace rozdělením normovaným normálním stejná.

Vhodnější aproximací je však statistika chí-kvadrát ve tvaru

$$\chi^2 = \frac{\left(|n_1 - n_2| - 1 \right)^2}{n_1 + n_2}. \tag{3.24}$$

Pokud H_0 nezamítneme, můžeme *odhadnout společnou hodnotu* π. Bodovým odhadem je hodnota

$$p = \frac{1}{2}(p_1 + p_2),$$

intervalový odhad je dán mezemi

$$p_{D,H} = p \pm u_{1-\alpha/2} s_p,$$

kde s_p je směrodatná chyba odhadu, která se počítá podle vzorce

$$s_p = \sqrt{\frac{p(1-2p)}{2n}}.$$

Jestliže H_0 zamítáme, můžeme *odhadnout rozdíl* $\delta = \pi_i - \pi_j$. Bodovým odhadem je hodnota

$$d = p_1 - p_2,$$

intervalový odhad je dán mezemi

$$d_{D,H} = p \pm u_{1-\alpha/2} s_d,$$

kde s_d je směrodatná chyba odhadu, která se počítá podle vzorce

$$s_d = \sqrt{\frac{p_1 + p_2 - (p_1 - p_2)^2}{n}}.$$

Postupy pro některé další typy výběrů dle velikosti jsou uvedeny ve skriptech [34].

Příklad 3.16

Na základě dat z výše uvedeného dotazníku budeme zjišťovat, zda podíl absolventů, kteří nastoupili do zaměstnání před studiem nebo v průběhu studia, je stejný jako podíl absolventů, kteří nastoupili do zaměstnání po absolvování vysoké školy (viz výstup 3.15). Spočteme upravenou chí-kvadrát statistiku podle (3.24), tj.

$$\chi^2 = \frac{\left(|n_1 - n_2| - 1\right)^2}{n_1 + n_2} = \frac{\left(|289 - 332| - 1\right)^2}{289 + 332} = 2{,}841\,.$$

Jde o kvantil $\chi^2_{0{,}908}[1]$, to znamená, že $1 - \alpha' = 0{,}908$, a tedy $\alpha' = 0{,}092$. Na 5% hladině významnosti (stejně jako 1%) nezamítáme nulovou hypotézu o shodě podílů dvou uvedených skupin absolventů.

Poznámka

Výpočet chí-kvadrát statistiky podle vzorce (3.24) není v systému *IBM SPSS Statistics* aplikován v rámci chí-kvadrát testu dobré shody. Je ale využit při McNemarově testu, který bude popsán v oddílu 4.3.6.

Bodovým odhadem *společné hodnoty* π je

$$p = \frac{1}{2}(p_1 + p_2) = \frac{1}{2}(0{,}455 + 0{,}523) = 0{,}489\,,$$

intervalový odhad je dán mezemi

$$p_{D,H} = p \pm u_{1-\alpha/2}\sqrt{\frac{p(1-2p)}{2n}} = 0{,}489 \pm 1{,}96\sqrt{\frac{0{,}489 \cdot (1 - 2 \cdot 0{,}489)}{2 \cdot 621}} = 0{,}489 \pm 0{,}006\,,$$

tj. $p_D = 0{,}489 - 0{,}006 = 0{,}483$ a $p_H = 0{,}489 + 0{,}006 = 0{,}495$.

3.4.4 Znaménkové schéma odchylek

Jestliže zamítáme H_0 o shodě zjištěných četností s očekávanými, pak můžeme vytvořit znaménkové schéma odchylek, přičemž pro $n \geq 30$ a $n\pi_{i,0} > 5$ využijeme aproximaci normovaným normálním rozdělením. Pro každou *i*-tou kategorii vypočteme normovanou hodnotu

$$Z_i = \frac{n_i - n\pi_{i,0}}{\sqrt{n\pi_{i,0}(1 - \pi_{i,0})}} = \sqrt{n}\,\frac{p_i - \pi_{i,0}}{\sqrt{\pi_{i,0}(1 - \pi_{i,0})}}\,. \tag{3.25}$$

Tyto hodnoty pak porovnáváme s kvantily normovaného normálního rozdělení a určujeme stupeň odchýlení od hodnot $\pi_{i,0}$, s využitím čísla 0 (zjištěná četnost odpovídá očekávané četnosti) a znamének „plus“ (zjištěná četnost je významně větší než očekávaná) a „minus“ (zjištěná četnost je významně menší). Počet znamének „plus“ a „minus“ označuje menší či větší odchýlení. V [31] jsou uvedeny tři postupy přiřazení znamének; dále bude uveden pouze jeden z těchto způsobů, který vychází z následujících kritických hodnot: $u_{0{,}0005} = -3{,}29$, $u_{0{,}005} = -2{,}58$, $u_{0{,}025} = -1{,}96$, $u_{0{,}975} = 1{,}96$, $u_{0{,}995} = 2{,}58$, $u_{0{,}9995} = 3{,}29$. Znaménka přiřazujeme podle schématu 3.2.

Schéma 3.2 | Algoritmus pro přiřazení znamének

pro	$Z_i \leq -3{,}29$	přiřadíme $---$
pro	$-3{,}29 < Z_i \leq -2{,}58$	přiřadíme $--$
pro	$-2{,}58 < Z_i \leq -1{,}96$	přiřadíme $-$
pro	$-1{,}96 < Z_i < 1{,}96$	přiřadíme 0
pro	$1{,}96 \leq Z_i < 2{,}58$	přiřadíme $+$
pro	$2{,}58 \leq Z_i < 3{,}29$	přiřadíme $++$
pro	$3{,}29 \leq Z_i$	přiřadíme $+++$

Příklad 3.17

Vytvořme znaménkové schéma odchylek pro příklad 3.15. Potřebné hodnoty a výsledné schéma je uvedeno v tabulce 3.9.

Tabulka 3.9 | Postup při vytváření znaménkového schématu pro typ smlouvy

n_i	$n\pi_{i,0}$	$1-\pi_{i,0}$	$Z_i = \frac{n_i - n\pi_{i,0}}{\sqrt{n\pi_{i,0}(1-\pi_{i,0})}}$	Znaménko
467	490,59	0,21	–2,324	–
106	93,15	0,85	1,444	0
48	37,26	0,94	1,815	0

V případě první kategorie je zjištěná četnost nižší, než je očekávaná; u druhé a třetí kategorie zjištěná četnost přibližně odpovídá očekávané.

Analýza závislostí

4

Vstupem pro vícerozměrnou statistickou analýzu ve statistických programových systémech je obvykle *datová matice*, v níž řádky odpovídají statistickým jednotkám (objektům, kterými mohou být osoby, domácnosti apod.) a sloupce odpovídají statistickým znakům (proměnným). Některé programové systémy umožňují vycházet z již dříve získaných četností jednotlivých kombinací výskytů kategorií u analyzovaných proměnných. V systému *IBM SPSS Statistics* lze použít oba uvedené způsoby vstupu dat. V druhém případě je v každém řádku datové matice zaznamenána jedna kombinace výskytů kategorií a jejich četnost, přičemž tato četnost označuje váhu uvedené kombinace. Příklad zadání pro dvě proměnné je uveden v tabulce 4.2.

V této kapitole se budeme věnovat zkoumání statistické závislosti u dvojic proměnných, přičemž se zaměříme na analýzu kategoriálních dat (viz oddíl 2.1).

4.1 Dvourozměrné rozdělení četností

Také v případě dvou proměnných bývá prvním krokem zobrazení rozdělení četností, a to buď v tabulce, nebo v grafu. V sociálních vědách je tato základní analýza označována jako *třídění II. stupně*. U kategoriálních proměnných jsou četnosti zjišťovány pro všechny takové dvojice kategorií, kdy jedna kategorie z dvojice přísluší první proměnné a druhá kategorie druhé proměnné. Dostáváme tak dvourozměrnou tabulku četností (tzv. *kontingenční tabulku*), z jejíchž hodnot již případně můžeme usoudit na závislost či nezávislost mezi dvěma kategoriálními proměnnými.

V políčkách jsou uváděny buď absolutní, nebo relativní četnosti (často jsou udávány v procentech), které mohou být počítány třemi různými způsoby: podíly počítané na základě celého rozsahu souboru (jejich součet se rovná hodnotě 1, resp. 100 v procentním vyjádření), *řádkové podíly* (součet hodnot v každém řádku se rovná hodnotě 1) nebo *sloupcové podíly* (součet hodnot v každém sloupci se rovná hodnotě 1). Můžeme buď nechat zobrazit několik tabulek s různými typy četností, nebo zapsat několik hodnot do jednoho políčka (*IBM SPSS Statistics*). Kontingenční tabulka je základem pro testování nezávislosti a pro výpočet měr intenzity závislosti.

Označme si rozsah souboru (počet objektů se známými hodnotami obou sledovaných proměnných) symbolem n, počet kategorií proměnné X jako R a počet kategorií proměnné Y jako S. Zjištěné četnosti budeme dále označovat jako n_{ij}, kde $i = 1, 2, ..., R$ a $j = 1, 2, ..., S$. Schéma 4.1 znázorňuje značení v kontingenční tabulce.

Schéma 4.1 | Značení pro kontingenční tabulku absolutních četností

	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

Ve výše uvedené tabulce jsou kromě *sdružených absolutních četností* n_{ij} též *marginální četnosti* n_{i+} a n_{+j}, pro které platí, že $n_{i+} = \sum_{j=1}^{S} n_{ij}$ a $n_{+j} = \sum_{i=1}^{R} n_{ij}$. Při výpočtech můžeme dále vycházet z četností relativních, jejichž značení znázorňuje schéma 4.2.

Schéma 4.2 | Značení pro kontingenční tabulku relativních četností

	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_{R}	...	p_{RS}	p_{R+}
Celkem	p_{+1}	...	p_{+j}	...	p_{+S}	1

Jde-li o podíly počítané na základě celého souboru, pak pro $i = 1, 2, ..., R$ a $j = 1, 2, ..., S$ platí, že

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

Příklad 4.1

Charakterizujme vztah proměnných *C2* (typ pracovní smlouvy v prvním zaměstnání) a *E1* (pohlaví) pomocí dvourozměrného rozdělení četností.

IBM SPSS Statistics

V systému *IBM SPSS Statistics* vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*). V dialogovém okně zadáme do části *Row(s)* proměnnou *E1* (pohlaví bude uváděno v řádcích tabulky) a do části *Column(s)* zadáme proměnnou *C2* (typy smlouvy se budou nacházet ve sloupcích) a ponecháme standardně nastavené výstupy. Výsledná kontingenční tabulka je ve výstupu 4.1.

Výstup 4.1 | Kontingenční tabulka absolutních četností k příkladu 4.1

		Typ smlouvy v prvním zaměstnání			Total
		Na dobu neurčitou	Na dobu určitou	Pouze OSVČ	
Pohlaví	Muž	171	27	27	225
	Žena	268	73	18	359
Total		439	100	45	584

Pro zkoumání, zda se typ smlouvy liší v závislosti na pohlaví, je vhodné vypočítat řádkové relativní četnosti, tj. n_{ij}/n_{i+} (značení viz tabulka 4.1). V rámci možnosti *Cells* zadáme v části *Percentages* možnost *Row* (v části *Counts* zrušíme možnost *Observed*). Ve výsledné tabulce jsou relativní četnosti uvedeny v procentech, viz výstup 4.2.

Výstup 4.2 | Kontingenční tabulka řádkových relativních četností k příkladu 4.1

		Typ smlouvy v prvním zaměstnání			Total
		Na dobu neurčitou	Na dobu určitou	Pouze OSVČ	
Pohlaví	Muž	76,0 %	12,0 %	12,0 %	100,0 %
	Žena	74,7 %	20,3 %	5,0 %	100,0 %
Total		75,2 %	17,1 %	7,7 %	100,0 %

V uvedené tabulce se rozdělení relativních četností pro muže a ženy liší, muži jsou častěji osoby samostatně výdělečně činné a ženy jsou častěji zaměstnány na dobu určitou.

♦ ♦ ♦

Graficky lze hodnoty z kontingenční tabulky zobrazit jako *sloupcový graf*, přičemž četnosti pro dvojice kategorií mohou být vyjádřeny jako shluk sloupců (graf shlukový) nebo jako části jednoho sloupku (graf kumulativní). Výšky nebo části sloupců mohou představovat kterýkoli z výše uvedených typů četností.

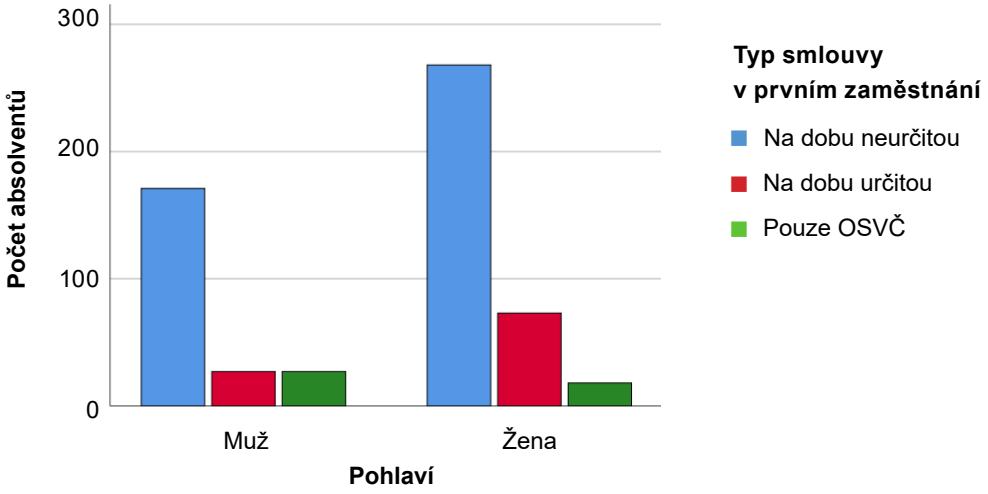
Příklad 4.2

Graficky znázorněme četnosti z kontingenční tabulky z příkladu 4.1.

IBM SPSS Statistics

Zobrazíme sloupcový graf shlukový s absolutními četnostmi. V systému *IBM SPSS Statistics* vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a zvolíme možnost *Display clustered bar charts*. Upravený výsledek zachycuje graf 4.1.

Graf 4.1 | Sloupcový graf četností absolventů podle typu smlouvy v prvním zaměstnání, členěno podle pohlaví



♦ ♦ ♦

Kromě absolutních a relativních četností lze ve dvourozměrné tabulce zobrazovat četnost očekávanou v případě nezávislosti (podrobněji viz následující oddíly) a rezidua, viz tabulka 4.1. Normalizovaná Habermanova rezidua poskytují hodnoty, pomocí nichž můžeme vytvořit znaménkové schéma odchylek podle návodu obsaženého ve schématu 3.1 (oddíl 3.4.4).

Pokud nemáme k dispozici vstupní datovou matici, ale pouze již dříve vytvořenou kontingenční tabulku absolutních četností (a chceme provádět další analýzy), pak je potřeba do datového editoru *IBM SPSS Statistics* zadat všechny možné kombinace kategorií, z nichž jedna přísluší řádkové proměnné a druhá proměnné sloupcové. Ke každé kombinaci zadáme počet jejich výskytů. Jednotlivým řádkům pak tyto četnosti (váhy) přiřadíme pomocí nabídky *Data*, *Weight cases* (zvolíme možnost *Weight cases by* a do políčka *Frequency Variable* zadáme název sloupce, ve kterém jsou zadány četnosti).

Kdybychom například získali tabulku uvedenou ve výstupu 4.1 a neměli k ní zdrojová data, zadáme četnosti způsobem uvedeným v tabulce 4.2.

Tabulka 4.1 | Charakteristiky políček kontingenční tabulky

Název charakteristiky	Vzorec
Řádková relativní četnost	$p_{j\|i} = \frac{n_{ij}}{n_{i+}}$
Sloupcová relativní četnost	$p_{i\|j} = \frac{n_{ij}}{n_{+j}}$
Očekávaná četnost při předpokladu nezávislosti	$m_{ij} = \frac{n_{i+}n_{+j}}{n}$
Reziduum	$n_{ij} - m_{ij}$
Standardizované reziduum	$\frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}}}$
Normalizované reziduum (Habermanovo)	$\frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}(1-p_{i+})\cdot(1-p_{+j})}}$

Tabulka 4.2 | Váhy pro kombinace kategorií proměnných *pohlaví* a *typ_smlouvy*

Pohlaví	Typ_smlouvy	Váha
1	1	171
1	2	27
1	3	27
2	1	268
2	2	73
2	3	18

4.2 Principy zjišťování závislosti dvou proměnných

Závislost dvou proměnných může být buď *symetrická* (*vzájemná*), nebo *asymetrická* (*jednostranná*). Zkoumáme například vzájemnou závislost názorů manžela a manželky či názorů dvou sourozenců a asymetrickou závislost názoru respondenta na jeho vzdělání (opačná závislost nemá logicky smysl). Možnost použití jednotlivých metod měření závislosti navíc souvisí s typy proměnných (viz oddíl 2.1).

Základním testem používaným ke zjišťování závislosti dvou kategoriálních znaků (bez zohlednění směru závislosti) je **chí-kvadrát test o nezávislosti**. Při něm vycházíme z úvahy, že pokud jsou dva znaky nezávislé, pak rozdělení četností v kontingenční tabulce

je úměrné řádkovým a sloupcovým marginálním četnostem. Tyto četnosti se nazývají *očekávané*; dále je budeme značit m_{ij} (vzorec pro jejich výpočet je uveden v tabulce 4.1). Stejně jako v chí-kvadrát testu dobré shody testujeme shodu zjištěných a očekávaných četností.

Pro měření intenzity (síly) závislosti jsou používány různé koeficienty, které obvykle nabývají hodnot z intervalu $0; 1\rangle$, případně $\langle -1; 1\rangle$, přičemž hodnota 0 znamená nezávislost. Dalšími testy jsou proto *testy o nulovosti* těchto *koeficientů*.

Pro speciální situace (např. tabulka 2×2 – tj. obě proměnné jsou dichotomické) existují míry, které v případě nezávislosti nabývají hodnoty 1. Tehdy testujeme, zda příslušný koeficient se rovná jedné. Jestliže není splněn předpoklad pro použití chí-kvadrát testu v kontingenční tabulce (očekávané četnosti v jednotlivých políčkách nejsou alespoň 5, podrobněji viz dále), jsou používány tzv. *exaktní testy*.

Koeficientů závislosti (asociace) je velké množství, v následujícím oddílu 4.3 je uvedeno přes 30 základních vzorců. Jde o tzv. výběrové koeficienty, které jsou bodovými odhady měr v základním souboru. Obvykle je klasifikujeme na základě:

- *rozměru tabulky* (tj. podle počtu kategorií u sledovaných proměnných),
- *typu proměnných* (zda jde o proměnné nominální, ordinální či kvantitativní),
- *typu závislosti* (zda jde o závislost symetrickou či asymetrickou).

Mnoho úloh z oblasti zkoumání vztahu kategoriálních proměnných se týká *závislosti asymetrické*. Pro dvě kategoriální proměnné lze vždy vypočítat hodnoty dvojice asymetrických koeficientů (často má věcný smysl jen jeden z nich), které posuzují míru závislosti jednak proměnné Y na X, jednak X na Y (na základě některých takovýchto dvojic můžeme vypočítat koeficient symetrický). V dalším textu budeme u asymetrických závislostí předpokládat, že sloupcová proměnná Y je vysvětlovaná, zatímco řádková proměnná X je vysvětlující.

Proměnná X statisticky působí na Y, jestliže se změnami kategorií x_i se mění statistické vlastnosti proměnné Y. Analýza asymetrické závislosti je založena na **principu analýzy rozptylu**, při níž se testuje shoda středních hodnot ve skupinách vytvořených na základě kategorií vysvětlující proměnné. Pokud proměnná Y není kvantitativní, pak místo středních hodnot zkoumáme jiné míry polohy, tj. u nominální proměnné *modus*, u ordinální proměnné *medián*. Testování shody mediánů bude vyloženo v kapitole 5.

Z metody nazývané analýza rozptylu (viz např. [17]) vyplývá, že variabilitu vysvětlované proměnné můžeme rozložit do dvou složek, a to na variabilitu vysvětlenou proměnnou X (meziskupinovou variabilitu) a tzv. zbytkovou (nevysvětlenou) variabilitu (vnitroskupinovou variabilitu). Matematickým zápisem může být tento rozklad vyjádřený jako

$$var\,(Y) = var\,(Y, X) + var\,(Y|X)\,,$$

přičemž pro zjištění variability lze použít základní míry uvedené v oddílu 3.2, samozřejmě vhodné pro daný typ vysvětlované proměnné.

Při analýze rozptylu měříme intenzitu závislosti pomocí poměru determinace, který se počítá jako podíl meziskupinové variability na celkové variabilitě. Obecně lze takovou míru zapsat jako

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

Jestliže tuto míru počítáme na základě kontingenční tabulky, pak

$$S_{Y|X} = \frac{var(Y) - \sum_{i=1}^{R} p_{i+} var(Y|x_i)}{var(Y)}, \tag{4.1}$$

kde $var(Y|x_i)$ je variabilita hodnot proměnné Y spočtená pro i-tý řádek tabulky (na základě četností v i-tém řádku).

V praxi se používá tato míra pro nominální nebo kvantitativní vysvětlovanou proměnnou. V prvním případě lze odvodit míry závislosti na základě všech tří měr variability uvedených v oddílu 3.2. V druhém případě se používá rozptyl a uvádí se také odmocnina z výsledné míry.

4.3 Testové statistiky a míry závislosti

V tomto oddílu budou probrány testové statistiky a míry intenzity závislosti jednak pro různé rozměry tabulky, jednak pro různé kombinace typů proměnných.

Míry závislosti určené pro dvě nominální proměnné můžeme použít též pro jiné typy proměnných, například pro jednu nominální a druhou ordinální apod. Pro dvě proměnné ordinální či kvantitativní nebo pro asymetrickou závislost proměnné kvantitativní na proměnné nominální však existují speciální míry, které jsou pro posouzení vztahu sledovaných znaků vhodnější.

4.3.1 Tabulka pro dvě nominální proměnné

Závislost dvou nominálních proměnných se nazývá *kontingence*. Základem pro zjišťování této závislosti je již zmíněný **chí-kvadrát test o nezávislosti.** Označíme-li relativní četnosti v základním souboru jako π_{ij} (jejich bodovými odhady jsou četnosti p_{ij}), pak nulovou hypotézu o nezávislosti zapíšeme ve tvaru H_0: $\pi_{ij} = p_{i+}p_{+j}$, kde $p_{i+}p_{+j}$ je relativní četnost očekávaná v případě nezávislosti. Tuto nulovou hypotézu testujeme vůči hypotéze H_1: $\pi_{ij} \neq p_{i+}p_{+j}$ alespoň pro jednu dvojici i, j ($i \neq j$). Jako testové kritérium lze použít **Pearsonovu statistiku chí-kvadrát**, která je vyjádřena vztahem (odvození viz [19])

$$\chi^2_P = n\sum_{i=1}^{R}\sum_{j=1}^{S}\frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}},$$

resp. vztahem

$$\chi_P^2 = \sum_{i=1}^{R}\sum_{j=1}^{S}\frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{4.2}$$

Statistika chí-kvadrát nabývá hodnot z intervalu $\langle 0; n\cdot(q-1)\rangle$, kde $q = \min\{R, S\}$. Tato náhodná veličina má za předpokladu platnosti nulové hypotézy přibližně chí-kvadrát rozdělení s $(R-1)(S-1)$ stupni volnosti, tj. $\chi_P^2 \approx \chi^2[(R-1)(S-1)]$. Pro vyhodnocení testu buď vypočtenou hodnotu testového kritéria porovnáváme s kvantilem $\chi_{1-\alpha}^2[(R-1)(S-1)]$, kde α je zvolená hladina významnosti, nebo zjistíme, jakým kvantilem je vypočtená hodnota, tj. zjistíme hodnotu distribuční funkce, a ze vztahu $F(\chi_P^2) = 1 - \alpha'$ vypočteme α′, které porovnáme se zvolenou hladinou α.

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 aspoň hodnoty 1 (v literatuře jsou uváděny i jiné požadavky, viz např. [31]). Není-li předpoklad splněn, používají se *exaktní testy*, viz oddíl 4.3.6.

Kromě statistiky chí-kvadrát se pro testování nezávislosti dvou nominálních proměnných využívá **věrohodnostní poměr**, který se v případě multinomického rozdělení spočte jako

$$G^2 = 2\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij} \ln\frac{n_{ij}}{m_{ij}}. \tag{4.3}$$

Uvedená statistika má asymptoticky chí-kvadrát rozdělení s $(R-1)(S-1)$ stupni volnosti. Při testování se tedy postupuje stejně jako v předchozím případě.

Příklad 4.3

Proveďme test o nezávislosti proměnných *C2* (*typ smlouvy v prvním zaměstnání*) a *C1* (*nástup do zaměstnání*). Kontingenční tabulka absolutních četností je uvedena ve výstupu 4.3, očekávané četnosti v případě nezávislosti pak ve výstupu 4.4.

Výstup 4.3 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.3

Zjištěné absolutní četnosti		Typ smlouvy v prvním zaměstnání			Celkem
		Na dobu neurčitou	Na dobu určitou	Pouze OSVČ	
Nástup do zaměstnání	Před studiem nebo v průběhu studia	223	34	32	289
	Po absolvování VŠ	244	72	16	332
Celkem		467	106	48	621

Výstup 4.4 | Kontingenční tabulka očekávaných absolutních četností k příkladu 4.3

Očekávané absolutní četnosti		Typ smlouvy v prvním zaměstnání			Celkem
		Na dobu neurčitou	Na dobu určitou	Pouze OSVČ	
Nástup do zaměstnání	Před studiem nebo v průběhu studia	217,3	49,3	22,3	289,0
	Po absolvování VŠ	249,7	56,7	25,7	332,0
Celkem		467,0	106,0	48,0	621,0

Bez použití statistického programu bychom výpočet Pearsonovy statistiky chí-kvadrát provedli podle vzorce (4.2), a to

$$\chi^2_P = \sum_{i=1}^{R}\sum_{j=1}^{S}\frac{(n_{ij}-m_{ij})^2}{m_{ij}} = \frac{(223-217{,}3)^2}{217{,}3} + \frac{(34-49{,}3)^2}{49{,}3} + \ldots + \frac{(16-25{,}7)^2}{25{,}7} = 17{,}004.$$

Výpočet věrohodnostního poměru bychom provedli podle vzorce (4.3), tj.

$$G^2 = 2\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij}\ln(n_{ij}/m_{ij}) = 2\cdot\left(223\cdot\ln(223/217{,}3) + \ldots + 16\cdot\ln(16/25{,}7)\right) = 17{,}332.$$

Při testování na 5% hladině významnosti vypočtenou hodnotu statistiky χ^2_P, resp. G^2, porovnáváme s kvantilem $\chi^2_{0,95}$ [(2 – 1)(3 – 1)] = $\chi^2_{0,95}$[2] = 5,99. V obou případech jsme získali hodnotu vyšší, než je uvedená kritická hodnota. Proto také v obou případech zamítáme nulovou hypotézu o nezávislosti proměnných *typ smlouvy v prvním zaměstnání* a *nástup do zaměstnání*.

IBM SPSS Statistics

V systému *IBM SPSS Statistics* provedeme volby *Analyze*, *Descriptive Statistics*, *Crosstabs*, dále zvolíme *Statistics* a vybereme *Chi-square*. Získáme výsledek uvedený na upraveném výstupu 4.5. První řádek (*Pearson Chi-Square*) se týká Pearsonovy chí-kvadrát statistiky, druhý (*Likelihood Ratio*) věrohodnostního poměru (ve třetím řádku je uveden celkový rozsah souboru). Pro obě statistiky je v prvním sloupci (*Value*) uvedena spočtená hodnota statistiky, ve druhém sloupci (*df*) počet stupňů volnosti a ve třetím (*Asymptotic Significance (2-sided)*) pak minimální hladina významnosti, od které zamítáme hypotézu H_0 o nezávislosti sledovaných proměnných. Pod tabulkou je poznámka, že v žádném z políček není očekávaná četnost menší než 5 (minimální očekávaná hodnota je 22,34), je tedy splněn předpoklad pro použití chí-kvadrát testu.

Získané výsledky se shodují s výpočty bez použití *IBM SPSS Statistics*. Minimální hladina významnosti je menší než tři tisíciny, tudíž na 1% hladině významnosti zamítáme hypotézu H_0 o nezávislosti. Stejný výsledek i interpretace platí pro věrohodnostní poměr.

Výstup 4.5 | Testy nezávislosti založené na chí-kvadrát statistikách k příkladu 4.3

Chi-Square Tests	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	17,004[a]	2	,000
Likelihood Ratio	17,332	2	,000
N of Valid Cases	621		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 22,34.

♦ ♦ ♦

Míry intenzity vzájemné závislosti jsou založeny na již zmíněné **statistice chí-kvadrát,** dané vztahem (4.2). V praxi je používáno několik koeficientů, které v případě nezávislosti nabývají hodnoty 0. Příkladem je **Pearsonův kontingenční koeficient**, který budeme dále značit písmenem C_P. Počítá se podle vztahu

$$C_P = \sqrt{\frac{\chi_P^2}{\chi_P^2 + n}}. \tag{4.4}$$

Tento koeficient měří symetrickou závislost dvou proměnných. Nabývá hodnot z intervalu $\langle 0; \sqrt{(q-1)/q} \rangle$, kde $q = \min\{R, S\}$. Čím větší hodnotu získáváme při stejném n, R a S, tím je závislost silnější.

Další mírou intenzity vzájemné závislosti pro dvě proměnné je **koeficient φ** (fí), pro který platí vztah

$$\varphi = \sqrt{\frac{\chi_P^2}{n}}. \tag{4.5}$$

Jiným koeficientem určeným k měření symetrické závislosti je **Cramérovo V**, počítané podle vzorce

$$V = \sqrt{\frac{\chi_P^2}{n(q-1)}}, \tag{4.6}$$

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 $\langle 0; 1 \rangle$. Pro tabulku, v níž je alespoň jedna proměnná dichotomická (počet odpovídajících řádků a/nebo sloupců je 2), se shoduje s koeficientem φ.

Nakonec uveďme ještě **Čuprovův kontingenční koeficient** ve tvaru

$$C_T = \sqrt{\frac{\chi_P^2 / n}{\sqrt{(R-1)(S-1)}}}. \tag{4.7}$$

V případě čtvercové tabulky, která má stejný počet řádků a sloupců, platí, že

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

a hodnoty Cramérova V a Čuprovova kontingenčního koeficientu jsou shodné.

Pokud bychom měli vyhodnotit intenzitu závislosti pro jednu dvojici proměnných, pak je výhodné použít Cramérovo V, jehož hodnoty jsou z intervalu $\langle 0; 1\rangle$. K porovnání intenzit závislostí různých dvojic proměnných lze využít kterýkoli z uvedených koeficientů.

Příklad 4.4

Charakterizujme vztah proměnných *C2* (*typ smlouvy v prvním zaměstnání*) a *C1* (*nástup do zaměstnání*) pomocí symetrických koeficientů.

Bez použití statistického programu spočteme

$$C_P = \sqrt{\frac{\chi_P^2}{\chi_P^2 + n}} = \sqrt{\frac{17{,}004}{17{,}004 + 621}} = 0{,}163,$$

$$\varphi = V = \sqrt{\frac{\chi_P^2}{n}} = \sqrt{\frac{17{,}004}{621}} = 0{,}165,$$

$$C_T = \sqrt{\frac{\chi_P^2 / n}{\sqrt{(R-1)(S-1)}}} = \sqrt{\frac{17{,}004/621}{\sqrt{(2-1)(3-1)}}} = 0{,}139.$$

Pokud bychom chtěli vyhodnotit intenzitu závislosti na intervalu $\langle 0; 1\rangle$, použijeme hodnotu Cramérova V. V našem příkladu $V = 0{,}165$, takže bychom závislost ohodnotili jako slabou. Protože jedna z proměnných (řádková) nabývá pouze dvou kategorií, je hodnota Cramérova V shodná s hodnotou koeficientu φ.

IBM SPSS Statistics

V systému *IBM SPSS Statistics* vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnosti *Contingency coefficient* a *Phi and Cramér's V*. Výsledek zachycuje výstup 4.6. Součástí výstupu jsou minimální hladiny významnosti, od kterých zamítáme nulovou hypotézu o nezávislosti dvou sledovaných proměnných (sloupec *Approximate Significance*). Můžeme tedy konstatovat, že byla zamítnuta hypotéza o nezávislosti proměnných *typ smlouvy v prvním zaměstnání* a *nástup do zaměstnání*, a to jak na 5%, tak i na 1% hladině významnosti (viz chí-kvadrát test).

♦ ♦ ♦

Kromě výše uvedených symetrických měr byly navrženy *míry asymetrické*, které hodnotí intenzitu jednostranné závislosti vysvětlované proměnné na proměnné vysvětlující. Formálně můžeme sledovat jednak závislost proměnné Y na proměnné X, jednak závislost proměnné X na proměnné Y. K některým takovým dvojicím ještě existuje míra symetrická.

Výstup 4.6 | Míry závislosti založené na Pearsonově chí-kvadrát statistice (příklad 4.4)

Symmetric Measures		Value	Approximate Significance
Nominal by Nominal	Phi	,165	,000
	Cramer's *V*	,165	,000
	Contingency Coefficient	,163	,000
N of Valid Cases		621	

Budeme-li sledovat závislost sloupcové proměnné Y na řádkové proměnné X, pak mohou nastat dvě situace:

{1} sloupce jsou statisticky nezávislé na řádcích,
{2} sloupce jsou statisticky závislé na řádcích.

Mějme nový objekt, u kterého známe hodnotu znaku X, ale neznáme hodnoty znaku Y. Pokud předpokládáme situaci {1}, pak bychom pro nový objekt odhadovali hodnotu znaku Y nezávisle na X, zpravidla podle modální kategorie, pro kterou platí, že $p_{+\mathrm{Mo}} = \max_j(p_{+j})$. Pravděpodobnost chyby lze vyjádřit vztahem $P\{1\} = (1 - p_{+\mathrm{Mo}})$, což je variační poměr v znaku Y, viz (3.1). Když budeme předpokládat situaci {2}, budeme odhadovat hodnotu znaku Y tak, že vyhledáme maximum v řádku, který odpovídá známé hodnotě znaku X. Označíme-li toto maximum jako $p_{i\mathrm{Mo}} = \max_j(p_{ij})$, pak pravděpodobnost chyby je dána vztahem $P\{2\} = (1 - \Sigma_i p_{i\mathrm{Mo}})$. Proporcionální redukci chyby *PRE* (*Proporcional Reduction in Error*) počítáme podle schématu

$$PRE = \frac{P\{1\} - P\{2\}}{P\{1\}}. \tag{4.8}$$

Z tohoto vztahu vychází **Goodmanovo-Kruskalovo** λ (lambda), které v roce 1954 navrhli Goodman a Kruskal. Pro asymetrickou variantu lze odvodit vzorec

$$\lambda_{Y|X} = \frac{1 - p_{+Mo} - \left(1 - \sum_{i=1}^{R} p_{iMo}\right)}{1 - p_{+Mo}} = \frac{\sum_{i=1}^{R} p_{iMo} - p_{+Mo}}{1 - p_{+Mo}} = \frac{\sum_{i=1}^{R} n_{iMo} - n_{+Mo}}{n - n_{+Mo}}, \tag{4.9}$$

kde $n_{+\mathrm{Mo}} = \max_j(n_{+j})$ a $n_{i\mathrm{Mo}} = \max_j(n_{ij})$. Stejný vztah lze odvodit na základě analýzy variability při použití variačního poměru v, viz vzorec (3.1). Dosazením do obecného vzorce (4.1) dostáváme

$$\lambda_{Y|X} = \frac{v(Y) - \sum_{i=1}^{R} p_{i+} v(Y|x_i)}{v(Y)} = \frac{1 - p_{+Mo} - \sum_{i=1}^{R} p_{i+}\left(1 - \frac{p_{iMo}}{p_{i+}}\right)}{1 - p_{+Mo}} =$$

$$= \frac{1 - p_{+Mo} - \left(1 - \sum_{i=1}^{R} p_{iMo}\right)}{1 - p_{+Mo}} = \frac{\sum_{i=1}^{R} p_{iMo} - p_{+Mo}}{1 - p_{+Mo}}.$$

Podmínkou pro výpočet koeficientu λ je, aby se nenulové četnosti vyskytovaly ve více než jednom sloupci. K vlastnostem tohoto koeficientu patří, že nabývá hodnot z intervalu $\langle 0;\ 1\rangle$. Hodnoty 0 nabývá v případě, že kategorie řádkové proměnné žádným způsobem nepřispívají k predikci kategorie sloupcové proměnné. Hodnoty 1 nabývá, když každý řádek tabulky obsahuje nejvýše jedno políčko s nenulovou četností ($p_{ij} = p_{i+}$).

Při odvození symetrické varianty se uvažují pravděpodobnosti chyb

$$P\{1\} = 1 - \frac{1}{2}\left(p_{+Mo} + p_{Mo+}\right) \text{ a } P\{2\} = 1 - \frac{1}{2}\left(\sum_{i=1}^{R} p_{iMo} + \sum_{j=1}^{S} p_{Moj}\right).$$

Dosazením těchto pravděpodobností do vzorce (4.8) získáme symetrickou variantu koeficientu λ ve tvaru

$$\lambda_{\text{sym}} = \frac{\sum_{i=1}^{R} p_{iMo} + \sum_{j=1}^{S} p_{Moj} - p_{+Mo} - p_{Mo+}}{2 - p_{+Mo} - p_{Mo+}}, \tag{4.10}$$

případně s využitím absolutních četností ve tvaru

$$\lambda_{\text{sym}} = \frac{\sum_{i=1}^{R} n_{iMo} + \sum_{j=1}^{S} n_{Moj} - n_{+Mo} - n_{Mo+}}{2n - n_{+Mo} - n_{Mo+}}. \tag{4.11}$$

Je zřejmé, že tento koeficient zohledňuje pouze modální četnost a nepřihlíží k rozdělení ostatních četností. Často se tedy stává, že výsledkem je hodnota 0, ačkoliv jiné koeficienty nabývají hodnot vyšších. Koeficient λ vyjadřuje míru redukce chyby pro predikci, jestliže známe hodnotu vysvětlující proměnné. Tato míra redukce může být nulová a přitom mezi proměnnými může existovat závislost.

Dále vysvětlíme **Goodmanovo-Kruskalovo τ** (tau), které je založeno na principu analýzy rozptylu s použitím nominálního rozptylu *nomvar*, viz vzorec (3.2). Rozepišme si obecný vzorec (4.1) jako

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

$$= \frac{\sum_{i=1}^{R} \frac{1}{p_{i+}} \sum_{j=1}^{S} p_{ij}^2 - \sum_{j=1}^{S} p_{+j}^2}{1 - \sum_{j=1}^{S} p_{+j}^2} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{S} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}}}{1 - \sum_{j=1}^{S} p_{+j}^2}. \tag{4.12}$$

Pomocí absolutních četností můžeme psát

$$\tau_{Y|X} = \frac{n\sum_{i=1}^{R}\sum_{j=1}^{S}\frac{(n_{ij}-m_{ij})^2}{n_{i+}}}{n^2-\sum_{j=1}^{S}n_{+j}^2}. \tag{4.13}$$

Podmínkou pro výpočet koeficientu τ je, aby se nenulové četnosti vyskytovaly ve více než jednom sloupci. K vlastnostem tohoto koeficientu patří, že nabývá hodnot z intervalu $\langle 0; 1\rangle$. Hodnoty 0 nabývá v případě, kdy $p_{ij} = p_{i+}p_{+j}$ pro všechna i a j. Hodnoty 1 nabývá, když v každém řádku tabulky existuje takové p_{ij}, že $p_{ij} = p_{i+}$.

Chceme-li **testovat nulovost** $\boldsymbol{\tau}$ (H_0: $\tau_{Y|X} = 0$, H_1: $\tau_{Y|X} \neq 0$), pak vycházíme ze skutečnosti, že náhodná veličina $(n-1)(S-1)\,\tau_{Y|X}$ má za předpokladu platnosti nulové hypotézy přibližně chí-kvadrát rozdělení s $(S-1)(R-1)$ stupni volnosti. V případě ostatních zde uvedených koeficientů lze samozřejmě rovněž testovat jejich nulovost. Způsoby testování však dále nebudou uvedeny.

Pokud jako charakteristiku variability uvažujeme entropii H, viz vzorec (3.3), výsledkem dosazení do vzorce (4.1) bude **informační koeficient** (též **koeficient nejistoty**, resp. **koeficient neurčitosti**), pro který platí vztah

$$U_{Y|X} = \frac{H(Y)-\sum_{i=1}^{R}p_{i+}H(Y|x_i)}{H(Y)} = \frac{-\sum_{j=1}^{S}p_{+j}\ln p_{+j} - \sum_{i=1}^{R}p_{i+}\left(-\sum_{j=1}^{S}\frac{p_{ij}}{p_{i+}}\ln\frac{p_{ij}}{p_{i+}}\right)}{-\sum_{j=1}^{S}p_{+j}\ln p_{+j}} =$$

$$= \frac{-\sum_{j=1}^{S}p_{+j}\ln p_{+j} - \sum_{i=1}^{R}p_{i+}\left(-\frac{1}{p_{i+}}\sum_{j=1}^{S}p_{ij}\left(\ln p_{ij}-\ln p_{i+}\right)\right)}{-\sum_{j=1}^{S}p_{+j}\ln p_{+j}} =$$

$$= \frac{-\sum_{j=1}^{S}p_{+j}\ln p_{+j} + \sum_{i=1}^{R}\sum_{j=1}^{S}p_{ij}\ln p_{ij} - \sum_{i=1}^{R}\ln p_{i+}\sum_{j=1}^{S}p_{ij}}{-\sum_{j=1}^{S}p_{+j}\ln p_{+j}}.$$

Výsledný vzorec lze tedy psát jako

$$U_{Y|X} = \frac{-\sum_{i=1}^{R}p_{i+}\ln p_{i+} - \sum_{j=1}^{S}p_{+j}\ln p_{+j} + \sum_{i=1}^{R}\sum_{j=1}^{S}p_{ij}\ln p_{ij}}{-\sum_{j=1}^{S}p_{+j}\ln p_{+j}} = \tag{4.14}$$

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

Symetrický koeficient neurčitosti počítáme jako harmonický průměr[4] obou asymetrických koeficientů, vzorec tedy odvodíme následujícím způsobem:

$$U_{sym} = \frac{2}{\frac{H(Y)}{H(X)+H(Y)-H(XY)} + \frac{H(X)}{H(X)+H(Y)-H(XY)}} = \tag{4.15}$$

$$= \frac{2\big(H(X)+H(Y)-H(XY)\big)}{H(X)+H(Y)}.$$

Příklad 4.5

Charakterizujme vztah proměnných *C1* (*nástup do zaměstnání*) a *E1* (*pohlaví*). Četnosti kombinací kategorií jsou uvedeny ve výstupu 4.7.

Jde o jednostrannou závislost, kdy sledujeme, zda a případně do jaké míry je nástup do zaměstnání (z hlediska časového období) závislý na pohlaví absolventa vysoké školy. Můžeme tedy použít všechny tři uvedené koeficienty, kdy vysvětlovanou proměnnou je sloupcová a vysvětlující pak řádková.

Bez použití statistického programu počítáme *koeficient* λ podle vzorce (4.9), tj.

$$\lambda_{Y|X} = \frac{\sum_{i=1}^{R} n_{iMo} - n_{+Mo}}{n - n_{+Mo}} = \frac{(118+206)-313}{598-313} = 0{,}039.$$

Známe-li hodnotu vysvětlující proměnné (pohlaví), je očekávaná redukce chyby predikce hodnoty vysvětlované proměnné (doby nástupu do zaměstnání) 0,039, tj. 3,9 %.

Pro výpočet *koeficientu* τ potřebujeme znát teoretické četnosti, viz výstup 4.7.

Výstup 4.7 | Kontingenční tabulky zjištěných a očekávaných četností k příkladu 4.5

Zjištěné četnosti		Nástup do zaměstnání			Celkem
		Před studiem nebo v průběhu studia	Po absolvování VŠ	Dosud nepracuji	
Pohlaví	Muž	118	107	6	231
	Žena	153	206	8	367
Celkem		271	313	14	598

4 Obecný vzorec pro harmonický průměr n hodnot x_i je $\overline{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$.

Očekávané četnosti		Nástup do zaměstnání			Celkem
		Před studiem nebo v průběhu studia	Po absolvování VŠ	Dosud nepracuji	
Pohlaví	Muž	104,7	120,9	5,4	231,0
	Žena	166,3	192,1	8,6	367,0
Celkem		271,0	313,0	14,0	598,0

Použijeme vzorec (4.13), tj.

$$\tau_{Y|X} = \frac{n\sum_{i=1}^{R}\sum_{j=1}^{S}\frac{(n_{ij}-m_{ij})^2}{n_{i+}}}{n^2-\sum_{j=1}^{S}n_{+j}^2} = \frac{598\cdot\left(\frac{(118-104,7)^2}{231}+\ldots+\frac{(8-8,6)^2}{367}\right)}{598^2-(271^2+313^2+14^2)} = 0,0084\,.$$

Hodnota koeficientu je blízká nule, budeme tedy testovat hypotézu o nulovosti tohoto koeficientu. Přitom použijeme statistiku

$$(n-1)(S-1)\,\tau_{Y|X} = (598-1)\cdot(3-1)\cdot 0,0084 = 10,03\,.$$

Jde o kvantil $\chi^2_{0,993}$ [2]. To znamená, že $1-\alpha' = 0,993$, a tedy $\alpha' = 0,007$. Na 5% hladině významnosti (a rovněž na 1% hladině významnosti) zamítáme nulovou hypotézu o nulovosti koeficientu τ, tedy o tom, že doba nástupu do zaměstnání nezávisí na pohlaví.

K výpočtu *koeficientu neurčitosti* využijeme marginální a sdružené relativní četnosti, viz výstup 4.8. Použijeme vzorec (4.14), pro který připravíme mezivýpočty:

$$H(X) = -\sum_{i=1}^{R} p_{i+}\ln p_{i+} = -(0,386\cdot\ln 0,386 + 0,614\cdot\ln 0,614) = 0,667\,,$$

$$H(Y) = -\sum_{j=1}^{S} p_{+j}\ln p_{+j} = -(0,453\cdot\ln 0,453 + 0,523\cdot\ln 0,523 + 0,023\cdot\ln 0,023) = 0,784\,,$$

$$H(XY) = -\sum_{i=1}^{R}\sum_{j=1}^{S} p_{ij}\ln p_{ij} = -(0,197\cdot\ln 0,197 + \ldots + 0,013\cdot\ln 0,013) = 1,446\,.$$

Výsledná hodnota koeficientu neurčitosti je

$$_{Y|X} = \frac{H(X)+H(Y)-H(XY)}{H(Y)} = \frac{0,667+0,784-1,446}{0,784} = 0,006\,.$$

Jde tedy o hodnotu ještě nižší, než je hodnota koeficientu τ. Znamená to, že závislost je velmi slabá.

Výstup 4.8 | Kontingenční tabulka celkových relativních četností k příkladu 4.5

Relativní četnosti		Nástup do zaměstnání			Celkem
		Před studiem nebo v průběhu studia	Po absolvování VŠ	Dosud nepracuji	
Pohlaví	Muž	19,7 %	17,9 %	1,0 %	38,6 %
	Žena	25,6 %	34,4 %	1,3 %	61,4 %
Celkem		45,3 %	52,3 %	2,3 %	100,0 %

IBM SPSS Statistics

V systému *IBM SPSS Statistics* vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnosti *Lambda*, *Uncertainty coefficient*. Výsledek zachycuje výstup 4.9. Z dvojice asymetrických měr se zaměříme na druhou variantu (*nástup do zaměstnání Dependent*). Součástí výstupu jsou minimální hladiny významnosti, od kterých zamítáme nulovou hypotézu o nezávislosti dvou sledovaných proměnných (sloupec *Approximate Significance*). Pouze v případě koeficientu τ zamítáme nulovou hypotézu na 5% i na 1% hladině významnosti, že období nástupu do zaměstnání nezávisí na pohlaví.

Výstup 4.9 | Asymetrické míry k příkladu 4.5

Directional Measures			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,021	,029	,734	,463
		Pohlaví Dependent	,000	,000	.[c]	.[c]
		Nástup do zaměstnání Dependent	,039	,052	,734	,463
	Goodman and Kruskal tau	Pohlaví Dependent	,009	,008		,065[d]
		Nástup do zaměstnání Dependent	,008	,007		,007[d]
	Uncertainty Coefficient	Symmetric	,006	,005	1,172	,065[e]
		Pohlaví Dependent	,007	,006	1,172	,065[e]
		Nástup do zaměstnání Dependent	,006	,005	1,172	,065[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.

Příklad 4.6

Porovnejme nyní výsledky získané v příkladu 4.4 s dalšími symetrickými mírami a všimněme si měr asymetrických. Budeme zkoumat vztah proměnných *C2* (*typ smlouvy v prvním zaměstnání*) a *C1* (*nástup do zaměstnání*).

V systému *IBM SPSS Statistics* budeme postupovat tak, jak bylo popsáno v příkladu 4.5. Výsledek je uveden ve výstupu 4.10. Prorovnáme-li hodnoty symetrické varianty koeficientu λ a koeficientu neurčitosti (0,036 a 0,02) s hodnotami Pearsonova kontingenčního koeficientu a Cramérova V (0,163 a 0,165), jeví se závislost slabší. Je třeba ovšem vzít úvahu, že druhé dva koeficienty jsou odmocninou z navrženého podílu. Pokud bychom odmocnili hodnoty prvních dvou koeficientů, získali bychom ohodnocení 0,19 a 0,14, tedy řádově stejné. Koeficient λ hodnotící závislost *typu smlouvy v prvním zaměstnání* na *nástupu do zaměstnání* je nula, neboť modální kategorie vysvětlované proměnné pro obě kategorie vysvětlující proměnné je první, nedochází tudíž k její změně. Není proto proveden test o nulovosti (protože nula je přímo bodovým odhadem). Ve všech uvedených případech (kromě zmíněné hodnoty koeficientu λ) je minimální hladina významnosti, od které zamítáme nulovou hypotézu o nezávislosti dvou sledovaných proměnných, menší než 0,05. Proto na 5% hladině významnosti zamítáme hypotézu o nezávislosti.

Výstup 4.10 | Asymetrické míry k příkladu 4.6

Directional Measures			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,036	,015	2,319	,020
		Nástup do zaměstnání Dependent	,055	,023	2,319	,020
		Typ smlouvy v prvním zaměstnání Dependent	,000	,000	.[c]	.[c]
	Goodman and Kruskal tau	Nástup do zaměstnání Dependent	,027	,012		,000[d]
		Typ smlouvy v prvním zaměstnání Dependent	,009	,005		,003[d]
	Uncertainty Coefficient	Symmetric	,020	,009	2,121	,000[e]
		Nástup do zaměstnání Dependent	,020	,010	2,121	,000[e]
		Typ smlouvy v prvním zaměstnání Dependent	,020	,009	2,121	,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.

♦ ♦ ♦

Má-li tabulka stejný počet sloupců jako řádků, nazývá se *čtvercová*. Pokud dva sledované znaky nabývají stejných kategorií, pak lze sledovat *míru souhlasu*. Jako příklady můžeme uvést míru shody názorů manželů nebo rodiče a dítěte, rodinného stavu ženicha a nevěsty apod. K tomuto účelu používáme koeficient, který se nazývá **Cohenovo κ** (kappa). Počítá se podle vzorce

$$\kappa = \frac{\sum_{i=1}^{R} n_{ii} - \sum_{i=1}^{R} m_{ii}}{n - \sum_{i=1}^{R} m_{ii}}. \tag{4.16}$$

Tento koeficient nabývá svého maxima (hodnoty 1), jestliže se nenulové četnosti vyskytují pouze na diagonále. Hodnoty větší než 0,75 indikují výborný souhlas, hodnoty menší než 0,4 nesvědčí o souhlasu. Použití ukazuje příklad 4.7.

Příklad 4.7

Příkladem čtvercové tabulky pro dvě proměnné se stejnými kategoriemi je tabulka pro proměnné *C3* a *D3*, které vyjadřují, jaký studijní obor je vhodný pro první a pro současné zaměstnání. Kontingenční tabulka absolutních četností je uvedena ve výstupu 4.11.

Výstup 4.11 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.7

Zjištěné absolutní četnosti		Studijní obor vhodný pro současné zaměstnání				Celkem
		Vystudovaný obor	Příbuzný studijní obor	Zcela jiný studijní obor	Zaměstnání nevyžaduje oborovou specializaci	
Studijní obor vhodný pro první zaměstnání	Vystudovaný obor	16	9	5	2	32
	Příbuzný studijní obor	7	172	14	3	196
	Zcela jiný studijní obor	1	15	23	2	41
	Zaměstnání nevyžaduje oborovou specializaci	2	36	10	4	52
Celkem		26	232	52	11	321

Pro výpočet koeficientu κ potřebujeme znát teoretické četnosti, viz výstup 4.12.

Výstup 4.12 | Kontingenční tabulka očekávaných absolutních četností k příkladu 4.7

Očekávané absolutní četnosti		Studijní obor vhodný pro současné zaměstnání				Celkem
		Vystudovaný obor	Příbuzný studijní obor	Zcela jiný studijní obor	Zaměstnání nevyžaduje oborovou specializaci	
Studijní obor vhodný pro první zaměstnání	Vystudovaný obor	2,6	23,1	5,2	1,1	32,0
	Příbuzný studijní obor	15,9	141,7	31,8	6,7	196,0
	Zcela jiný studijní obor	3,3	29,6	6,6	1,4	41,0
	Zaměstnání nevyžaduje oborovou specializaci	4,2	37,6	8,4	1,8	52,0
Celkem		26,0	232,0	52,0	11,0	321,0

Dosazením do vzorce (4.16) dostáváme

$$\kappa = \frac{\sum_{i=1}^{R} n_{ii} - \sum_{i=1}^{R} m_{ii}}{n - \sum_{i=1}^{R} m_{ii}} = \frac{(16+172+23+4) - (2,6+141,7+6,6+1,8)}{321 - (2,6+141,7+6,6+1,8)} = 0,37 .$$

Můžeme konstatovat, že shoda mezi studijními obory vhodnými pro první a pro současné zaměstnání je slabá, koeficient kappa je pod hranicí, která indikuje souhlas.

IBM SPSS Statistics

V systému *IBM SPSS Statistics* vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnosti *Kappa*. Výsledek zachycuje výstup 4.13. Minimální hladina významnosti, od které zamítáme nulovou hypotézu o nulovosti koeficientu (sloupec *Approximate Significance*) říká, že shoda mezi proměnnými vyjadřujícími, jaký studijní obor je vhodný pro první a pro současné zaměstnání, je statisticky významná na 5% i na 1% hladině významnosti.

Výstup 4.13 | Míra souhlasu vhodnosti studijního oboru pro první a pro současné zaměstnání

Symmetric Measures		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	,370	,043	10,742	,000
N of Valid Cases		321			

a. Not assuming the null hypothesis.

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

4.3.2 Tabulka pro dvě ordinální proměnné

Zatímco u nominálních proměnných je statistická závislost označována jako kontingence, u ordinálních proměnných hovoříme o korelaci. Rozlišujeme přitom dva typy *korelace*, a to *pozitivní* (nízkým hodnotám jedné proměnné odpovídají nízké hodnoty proměnné druhé) a *negativní* (nízkým hodnotám jedné proměnné odpovídají vysoké hodnoty druhé proměnné).

K základním mírám patří **Spearmanův koeficient pořadové korelace**. Výpočet vychází z toho, že každé hodnotě proměnné X je přiřazeno pořadí a_l tak, že platí $\sum_{l=1}^{n} a_l = n\frac{n+1}{2}$, a každé hodnotě proměnné Y je přiřazeno pořadí b_l tak, že platí $\sum_{l=1}^{n} b_l = n\frac{n+1}{2}$.

Při výpočtu z četností kontingenční tabulky postupujeme následujícím způsobem:

a) kategoriím proměnné X přiřadíme postupně pomocné skóry a_i:

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

kategoriím proměnné Y přiřadíme pomocné skóry b_j:

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

b) spočteme hodnoty

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

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

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

které dosadíme do vzorce

$$r_S = \frac{\Omega_X + \Omega_Y - \delta^2}{2\sqrt{\Omega_X \Omega_Y}}. \tag{4.17}$$

Jestliže $\sum_{i=1}^{R} n_{i+}^3 = \sum_{j=1}^{S} n_{+j}^3$, pak $\Omega_x = \Omega_Y$. Získáváme tak jednodušší vzorec, a to

$$r_S = \frac{2\Omega_X - \delta^2}{2\sqrt{\Omega_X^2}} = 1 - \frac{\delta^2}{2\Omega_X}.$$

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

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

Pokud jde o prosté ordinální proměnné X a Y, které vyjadřují jednoznačné pořadí (žádná hodnota se neopakuje), pak není třeba počítat skóry a_i a b_j. Hodnotu δ^2 můžeme počítat přímo pomocí pořadí výrazem

$$\delta^2 = \sum (x_l - y_l)^2 .$$

Spearmanův koeficient nabývá hodnot z intervalu ⟨–1; 1⟩. Pokud jsou u každého respondenta v obou proměnných stejná pořadí, pak koeficient nabývá hodnoty 1 (*pozitivní korelace*, tzv. *přímá závislost*). Pokud seřadíme hodnoty proměnné X vzestupně a současně zjistíme sestupné pořadí u proměnné Y, koeficient je roven –1 (*negativní korelace*, tzv. *nepřímá závislost*). Hodnota 0 znamená nezávislost. Test na nulovost tohoto koeficientu (H_0: $\rho_S = 0$) se provádí pomocí statistiky

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

která má za předpokladu platnosti nulové hypotézy přibližně Studentovo t rozdělení s (n – 2) stupni volnosti (pro $n > 10$).

Kromě Spearmanova koeficientu existuje skupina měr, které vycházejí z porovnání dvojic objektů. Jsou-li ve sledované dvojici u jednoho objektu hodnoty u obou proměnných menší (resp. větší) než u druhého objektu, pak takový pár označujeme jako *konkordantní*. Je-li u jedné proměnné hodnota menší a u druhé proměnné větší, pak jde o pár *diskordantní*. V ostatních případech (hodnota u jedné proměnné nebo hodnoty u obou proměnných jsou shodné) hovoříme o párech *vázaných*.

Pro zjednodušení zápisů vzorců jsou používány 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 lze tyto četnosti vyjádřit podle následujících vzorců:

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

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

K symetrickým mírám patří **Goodmanovo-Kruskalovo γ** (gama), které je dáno vztahem

$$\gamma = \frac{C - D}{C + D}. \tag{4.18}$$

Pro tento koeficient platí, že $|\gamma| \in \langle 0; 1\rangle$, přičemž pro tabulku 2 × 2 nemusí nabýt hodnoty 0. Interpretace je obvyklá, tudíž hodnota 0 znamená nezávislost a hodnota 1 úplnou závislost. Hodnoty 1 je dosaženo tehdy, jestliže jsou nenulová políčka pouze na diagonále.

Další symetrickou mírou je **Kendallovo τ_b (**tau-*b*, též Kendallův koeficient pořadové korelace), dané vztahem

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X)(C + D + T_Y)}}. \tag{4.19}$$

Pro tento koeficient platí, že $\tau_b \in \langle -1; 1\rangle$. Pokud se žádná marginální četnost nerovná nule, pak může koeficient dosáhnout mezních hodnot pouze v tabulce $R \times R$.

K symetrickým mírám lze zařadit také **Kendallovo τ_c** (tau-*c*). Výpočet je založen na hodnotách C, D a dále $q = \min\{R, S\}$. Vzorec je následující:

$$\tau_c = \frac{2q(C - D)}{n^2(q - 1)}. \tag{4.20}$$

Pro tabulky typu R x S ($R \neq S$) platí, že $\tau_c \in \langle -1; 1\rangle$.

Asymetrickou mírou je **Somersovo *d***. Výpočet provedeme podle vzorce

$$d_{Y|X} = \frac{C - D}{C + D + T_Y}. \tag{4.21}$$

Symetrickou variantu počítáme jako harmonický průměr obou asymetrických variant; vzorec je odvozen následujícím způsobem:

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

K vlastnostem Somersova *d* patří, že geometrickým průměrem[5] jeho dvou asymetrických variant je koeficient τ_b, viz (4.19).

Příklad 4.8

Vypočtěme míry závislosti pro proměnné *přínos oboru pro vstup do práce* a *přínos oboru pro osobní rozvoj* překódované do proměnných se třemi kategoriemi, tj. pro proměnné *B2a_3kat* a *B2d_3kat*. Výsledné sdružené četnosti jsou ve výstupu 4.14.

5 Obecný vzorec pro geometrický průměr n hodnot x_i je $\overline{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$, kde Π je symbol pro součin hodnot.

Výstup 4.14 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.8

Zjištěné absolutní četnosti		Přínos oboru pro osobní rozvoj			Celkem
		Žádný nebo malý přínos	Střední přínos	Větší nebo velký přínos	
Přínos oboru pro vstup do práce	Žádný nebo malý přínos	78	40	22	140
	Střední přínos	53	80	51	184
	Větší nebo velký přínos	11	63	237	311
Celkem		142	183	310	635

Ilustrujme jednotlivé výpočty bez použití statistického programu. Pro výpočet *Spearmanova korelačního koeficientu* nejprve provedeme pomocné výpočty:

$$a_1 = \frac{n_{1+}+1}{2} = \frac{140+1}{2} = 70{,}5\,,\ a_2 = n_{1+} + \frac{n_{2+}+1}{2} = 140 + \frac{184+1}{2} = 232{,}5\,,$$

$$a_3 = n_{1+} + n_{2+} + \frac{n_{3+}+1}{2} = 140 + 184 + \frac{311+1}{2} = 484{,}5\,,$$

$$b_1 = \frac{n_{+1}+1}{2} = \frac{142+1}{2} = 71{,}5\,,\ b_2 = n_{+1} + \frac{n_{+2}+1}{2} = 142 + \frac{183+1}{2} = 234\,,$$

$$b_3 = n_{+1} + n_{+2} + \frac{n_{+3}+1}{2} = 142 + 1837 + \frac{310+1}{2} = 480{,}5\,,$$

$$\delta^2 = \sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij}(a_i - b_j)^2 = 78\cdot(70{,}5-71{,}5)^2 + \ldots + 237\cdot(484{,}5-480{,}5)^2 = 15\,134\,332\,,$$

$$\Omega_X = \frac{1}{12}\left(n^3 - \sum_{i=1}^{R} n_{i+}^3\right) = \frac{1}{12}\left(635^3 - \left(140^3 + 184^3 + 311^3\right)\right) = 18\,082\,845\,,$$

$$\Omega_Y = \frac{1}{12}\left(n^3 - \sum_{j=1}^{S} n_{+j}^3\right) = \frac{1}{12}\left(635^3 - \left(142^3 + 183^3 + 310^3\right)\right) = 18\,105\,425\,.$$

Výsledný výpočet je podle vzorce (4.17)

$$r_S = \frac{\Omega_X + \Omega_Y - \delta^2}{2\sqrt{\Omega_X \Omega_Y}} = \frac{18\,082\,845 + 18\,105\,425 - 15\,134\,332}{2\sqrt{18\,082\,845\cdot 18\,105\,425}} = 0{,}588\,.$$

Testové kritérium spočteme jako

$$t = r_S\sqrt{\frac{n-2}{1-r_S^2}} = 0{,}588\sqrt{\frac{635-2}{1-0{,}588^2}} = 18{,}3\,.$$

Protože kritická hodnota pro test o nulovosti korelačního koeficientu na 1% hladině významnosti je 1,96, můžeme jako výsledek t testu uvést, že na této hladině zamítáme nulovou hypotézu o nezávislosti.

Pro výpočty dalších koeficientů potřebujeme spočítat C, D, T_X a T_Y, tj.

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

$$= 80 \cdot 78 + 51 \cdot (78 + 40) + 63 \cdot (78 + 53) + 237 \cdot (78 + 40 + 53 + 80) = 79\,998 ,$$

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

$$= 53 \cdot (40 + 22) + 80 \cdot 22 + 11 \cdot (40 + 22 + 80 + 51) + 63 \cdot (22 + 51) = 11\,768 ,$$

$$T_X = \sum_{i=1}^{R}\sum_{j=2}^{S}\left(n_{ij} \sum_{h=i}\sum_{k<j} n_{hk} \right) = 40 \cdot 78 + 22 \cdot 118 + 80 \cdot 53 + 51 \cdot 133 + 63 \cdot 11 + 237 \cdot 74 = 34\,970 ,$$

$$T_Y = \sum_{i=2}^{R}\sum_{j=1}^{S}\left(n_{ij} \sum_{h<i}\sum_{k=j} n_{hk} \right) = 53 \cdot 78 + 80 \cdot 40 + 51 \cdot 22 + 11 \cdot 131 + 63 \cdot 120 + 237 \cdot 73 = 34\,758 .$$

Goodmanovo-Kruskalovo γ spočteme podle vztahu

$$\gamma = \frac{C - D}{C + D} = \frac{79\,998 - 11\,768}{79\,998 + 11\,768} = \frac{68\,230}{91\,766} = 0,744 .$$

Kendallovo τ_b spočteme podle vzorce

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X)(C + D + T_Y)}} = \frac{68\,230}{\sqrt{(91\,766 + 34\,970) \cdot (91\,766 + 34\,758)}} = 0,539 .$$

Kendallovo τ_c spočteme podle vztahu

$$\tau_c = \frac{2q(C - D)}{n^2(q - 1)} = \frac{2 \cdot 3 \cdot 68\,230}{635^2 \cdot (3 - 1)} = 0,508 .$$

Asymetrické *Somersovo d,* vyjadřující závislost sloupcové proměnné na řádkové, je

$$d_{Y|X} = \frac{C - D}{C + D + T_Y} = \frac{68\,230}{91\,766 + 34\,758} = 0,539$$

a druhou variantu, vyjadřující závislost řádkové proměnné na sloupcové, spočteme podle vzorce

$$d_{X|Y} = \frac{C - D}{C + D + T_X} = \frac{68\,230}{91\,766 + 34\,970} = 0,538 .$$

Symetrickou variantu spočteme podle vzorce

$$d_{sym} = \frac{2(C - D)}{2(C + D) + T_X + T_Y} = \frac{2 \cdot 68\,230}{2 \cdot 91\,766 + 34\,970 + 34\,758} = 0,539 .$$

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupy systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnosti *Correlations*, *Gamma*, *Somers'd*, *Kendall's tau-b* a *Kendall's tau-c*. Výsledek zachycují výstupy 4.15 (pouze symetrické míry) a 4.16 (asymetrické Somersovo *d*). Součástí výstupu jsou také minimální hladiny významnosti, od nichž zamítáme nulové hypotézy o nulovosti koeficientů (sloupec *Approximate Significance*, lze použít pouze pro soubory od $n = 20$).

Výstup 4.15 | Míry vzájemné závislosti přínosu oboru pro vstup do práce a přínosu oboru pro osobní rozvoj (příklad 4.8)

Symmetric Measures		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-*b*	,539	,028	18,830	,000
	Kendall's tau-*c*	,508	,027	18,830	,000
	Gamma	,744	,029	18,830	,000
	Spearman Correlation	,588	,030	18,268	,000[c]
Interval by Interval	Pearson's *R*	,583	,030	18,044	,000[c]
N of Valid Cases		635			

a. Not assuming the null hypothesis.

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

c. Based on normal approximation.

Výstup 4.16 | Míry jednostranné závislosti přínosu oboru pro vstup do práce a přínosu oboru pro osobní rozvoj (příklad 4.8)

Directional Measures			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' *d*	Symmetric	,539	,028	18,830	,000
		Přínos oboru pro vstup do práce Dependent	,538	,028	18,830	,000
		Přínos oboru pro osobní rozvoj Dependent	,539	,028	18,830	,000

a. Not assuming the null hypothesis.

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

Můžeme konstatovat, že ve všech případech zamítáme hypotézu o nulovosti daného koeficientu ve prospěch alternativní hypotézy, že proměnné *přínos oboru pro vstup do práce* a *přínos oboru pro osobní rozvoj* jsou závislé. Při volbě možnosti *Correlations* se počítají hodnoty korelačních koeficientů bez ohledu na typ proměnné, takže výstup obsahuje i Pearsonův korelační koeficient pro proměnné kvantitativní (*Pearson's R*). Koeficienty nabývají hodnot od 0,508 (Kendallovo tau-*c*) do 0,744 (gama). Závislost je středně silná a přímá (korelace je pozitivní).

Příklad 4.9

Vypočtěme míry závislosti pro nejvyšší dosažená vzdělání rodičů (otce a matky), tj pro proměnné *E3otec* a *E3matka*. Výsledné sdružené četnosti jsou ve výstupu 4.17.

IBM SPSS Statistics

V systému *IBM SPSS Statistics* postupujeme stejně jako v příkladu 4.8 Výsledek zachycují výstupy 4.18 (pouze symetrické míry) a 4.19 (asymetrické Somersovo *d*). Můžeme konstatovat, že ve všech případech zamítáme hypotézu o nulovosti daného koeficientu ve prospěch alternativní hypotézy, že vzdělání rodičů absolventů jsou závislá. Koeficienty nabývají hodnot od 0,466 (Kendallovo tau-*c*) do 0,747 (gama). Závislost je tedy i v tomto případě středně silná.

Výstup 4.17 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.9

Zjištěné absolutní četnosti		Vzdělání matky			Celkem
		Bez maturity	Středoškolské s maturitou	Vysoko-školské	
Vzdělání otce	Bez maturity	51	77	8	136
	Středoškolské s maturitou	33	108	22	163
	Vysokoškolské	10	112	170	292
Celkem		94	297	200	591

Výstup 4.18 | Míry vzájemné závislosti vzdělání rodičů (příklad 4.9)

Symmetric Measures		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-*b*	,504	,027	17,341	,000
	Kendall's tau-*c*	,466	,027	17,341	,000
	Gamma	,747	,033	17,341	,000
	Spearman Correlation	,549	,029	15,941	,000[c]
Interval by Interval	Pearson's *R*	,537	,029	15,434	,000[c]
N of Valid Cases		591			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

Výstup 4.19 | Míry jednostranných závislostí vzdělání rodičů (příklad 4.9)

Directional Measures			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers'*d*	Symmetric	,504	,027	17,341	,000
		Vzdělání otce Dependent	,512	,027	17,341	,000
		Vzdělání matky Dependent	,496	,029	17,341	,000

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

♦ ♦ ♦

V případě čtvercových tabulek s ordinálními proměnnými, které nabývají kategorií stejného významu, lze též počítat koeficient κ (kappa), viz vzorec (4.16).

Příklad 4.10

Pro sledování souhlasu kategorií ordinálních proměnných využijeme data z předchozího příkladu 4.9. Pro výpočet koeficientu κ potřebujeme znát očekávané četnosti, viz výstup 4.20.

Výstup 4.20 | Kontingenční tabulka očekávaných absolutních četností k příkladu 4.10

Očekávané absolutní četnosti		Vzdělání matky			Celkem
		Bez maturity	Středoškolské s maturitou	Vysokoškolské	
Vzdělání otce	Bez maturity	21,6	68,3	46,0	136,0
	Středoškolské s maturitou	25,9	81,9	55,2	163,0
	Vysokoškolské	46,4	146,7	98,8	292,0
Celkem		94,0	297,0	200,0	591,0

Dosazením do vzorce (4.16) dostáváme

$$\kappa = \frac{\sum_{i=1}^{R} n_{ii} - \sum_{i=1}^{R} m_{ii}}{n - \sum_{i=1}^{R} m_{ii}} = \frac{(51+108+170)-(21,6+81,9+98,8)}{591-(21,6+81,9+98,8)} = 0,326\,.$$

Míra souhlasu vzdělání otce a matky respondenta je menší než 0,4, souhlas je tedy slabý.

IBM SPSS Statistics

V systému *IBM SPSS Statistics* vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Kappa*. Výsledek zachycuje výstup 4.21. Součástí výstupu je minimální hladina významnosti, od které zamítáme nulovou hypotézu o nulovosti koeficientu (sloupec *Approximate Significance*). Shoda je statisticky významná na 5% i na 1% hladině významnosti.

Výstup 4.21 | Míra souhlasu stupně vzdělání rodičů (příklad 4.10)

Symmetric Measures		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	,326	,029	11,697	,000
N of Valid Cases		591			

a. Not assuming the null hypothesis.

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

4.3.3 Tabulka s ordinální vysvětlovanou proměnnou

Rozdělíme-li hodnoty ordinální vysvětlované proměnné *Y* do skupin podle kategorií vysvětlující proměnné *X* (obvykle nominální), pak pro porovnání skupin použijeme **Kruskalův-Wallisův test**. Nulová hypotéza předpokládá, že ve všech skupinách jsou shodná rozdělení, alternativní hypotéza říká, že alespoň v jedné skupině se rozdělení liší od ostatních.

Výpočet testového kritéria je založen na pořadích, která jsou přiřazena hodnotám v souboru, vzniklým spojením všech výběrů. Při výpočtu z četností kontingenční tabulky postupujeme tak, že kategoriím proměnné *Y* přiřadíme postupně pomocné skóry w_j podle vztahů

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

pro každou kategorii proměnné *X* vypočteme průměrné pořadí

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

a součet pořadí

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

vypočteme opravu na spojitost

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

a nakonec spočteme hodnotu Kruskalovy-Wallisovy statistiky

$$KW = \frac{\dfrac{12}{n(n+1)} \sum_{i=1}^{R} \dfrac{R_i^2}{n_{i+}} - 3(n+1)}{H_S}. \tag{4.21}$$

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

Příklad 4.11

Zjistěme, zda *spokojenost se současnou prací* (proměnná *D6* překódovaná do tříhodnotové škály; výsledkem je proměnná *D6_3kat*) závisí na *typu smlouvy v současném zaměstnání* (proměnná *D2*). Tabulka zjištěných četností je ve výstupu 4.22.

Výstup 4.22 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.11

Zjištěné absolutní četnosti		Spokojenost se současnou prací			Celkem
		Nespokojen(a) až napůl spokojen(a)	Spíše spokojen(a)	Velmi spokojen(a)	
Typ smlouvy v současném zaměstnání	Na dobu neurčitou	128	231	111	470
	Na dobu určitou	7	10	9	26
	Jsem pouze OSVČ	13	16	18	47
Celkem		148	257	138	543

Nejprve kategoriím sloupcové proměnné přiřadíme pomocné skóry w_j, tj.

$$w_1 = \frac{n_{+1}+1}{2} = \frac{148+1}{2} = 74{,}5\,,$$

$$w_2 = n_{+1} + \frac{n_{+2}+1}{2} = 148 + \frac{257+1}{2} = 277\,,$$

$$w_3 = \sum_{l=1}^{2} n_{+l} + \frac{n_{+3}+1}{2} = 148 + 257 + \frac{138+1}{2} = 474{,}5\,,$$

a pro každou kategorii řádkové proměnné vypočteme součet pořadí:

$$R_1 = \sum_{j=1}^{S} w_j n_{1j} = 74{,}5 \cdot 128 + 277 \cdot 231 + 474{,}5 \cdot 111 = 126\,192{,}5\,,$$

$$R_2 = \sum_{j=1}^{S} w_j n_{2j} = 74{,}5 \cdot 7 + 277 \cdot 10 + 474{,}5 \cdot 9 = 7\,562\,,$$

$$R_3 = \sum_{j=1}^{S} w_j n_{3j} = 74{,}5 \cdot 13 + 277 \cdot 16 + 474{,}5 \cdot 18 = 13\,941{,}5\,\cdot$$

Pak vypočteme opravu na spojitost

$$H_S = \frac{n^3 - \sum_{j=1}^{S} n_{+j}^3}{n(n^2-1)} = \frac{543^3 - (148^3 + 257^3 + 138^3)}{543 \cdot (543^2 - 1)} = 0{,}857\,317$$

a nakonec hodnotu testového kritéria podle vzorce (4.21), tj.

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

$$= \frac{\frac{12}{543 \cdot 544}\left(\frac{126\,192{,}5^2}{470} + \frac{7\,562^2}{26} + \frac{13\,941{,}5^2}{47}\right) - 3 \cdot 544}{0{,}857\,317} = 2{,}062\,.$$

Jde o kvantil $\chi^2_{0,643}$ [2], to znamená, že $1 - \alpha' = 0{,}643$, a tedy $\alpha' = 0{,}357$. Na 5% hladině významnosti nezamítáme nulovou hypotézu o shodě rozdělení v jednotlivých skupinách absolventů, vytvořených na základě typu smlouvy. Na této hladině významnosti tedy můžeme konstatovat, že spokojenost se současnou prací nezávisí na typu smlouvy.

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests*, *Independent Samples*. V listu *Fields* zadáme vysvětlovanou proměnnou do *Test Fields* a vysvětlující proměnnou do *Groups*. V listu *Settings* zvolíme *Customize tests* a v části *Compare Distributions across Groups* vybereme *Kruskal-Wallis 1-way ANOVA (k samples)*. Obdržíme výstup se zadáním úlohy, zvolenou metodou, minimální hladinou významnosti (0,357, shoduje se s hodnotou získanou bez použití programu) a závěrem. V okně *Model Viewer* získáme výstup 4.23 včetně testové statistiky *KW*, která je 2,062.

Výstup 4.23 | Výsledek Kruskalova-Wallisova testu (příklad 4.11)

Total N	543
Test Statistic	2,062
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	,357

4.3.4 Tabulka pro dvě kvantitativní proměnné

Jsou-li obě proměnné kvantitativní a zajímá nás vzájemná závislost, můžeme použít **Pearsonův korelační koeficient** ve tvaru

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

kde ve jmenovateli je pro vyjádření rozptylu použit druhý vztah z (3.7), nebo při vyjádření pomocí absolutních četností

$$r = \frac{\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij} x_i y_j - n\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{R} n_{i+} x_i^2 - n\overline{x}^2\right) \cdot \left(\sum_{j=1}^{S} n_{+j} y_j^2 - n\overline{y}^2\right)}}, \tag{4.22}$$

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

Tento koeficient nabývá hodnot z intervalu ⟨–1; 1⟩. Interpretace je obdobná jako u měr pro ordinální proměnné. Hodnota 0 znamená lineární nezávislost, hodnota 1 *pozitivní korelaci* (*přímou lineární závislost*) a hodnota –1 *negativní korelaci* (*nepřímou lineární závislost*).

Statistika pro test o lineární nezávislosti dvou proměnných založená na tomto koeficientu (testujeme jeho nulovost) se počítá stejným způsobem jako v případě Spearmanova korelačního koeficientu. Předpokladem tohoto testu však je, že jde o výběr z dvourozměrného normálního rozdělení. Pokud tato podmínka není splněna, lze použít test pro Spearmanův koeficient.

Příklad 4.12

Pro ukázku výpočtu Pearsonova korelačního koeficientu použijeme proměnnou *C4* (*počet zaměstnání od absolvování studia*) a proměnnou *A1_počet2* obsahující počty, které vyjadřují, zda byl(a) absolvent(ka) zaměstnán(a) před nebo při studiu (nebo obojí) a práce souvisela se studiem. Počet zaměstnání od absolvování studia je uvažován maximálně 3 (výběr absolventů se provede pomocí nabídek *Data* a *Select Cases*, kde zvolíme možnost *If condition is satisfied* a zadáme podmínku *If* ve tvaru „C4 < 4"). Uspořádání vstupních dat do kontingenční tabulky je ve výstupu 4.24.

Výstup 4.24 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.12

Zjištěné absolutní četnosti		Počet zaměstnání od absolvování studia			Celkem
		1	2	3	
Počet zaměstnání souvisejících se studiem	0	13	49	32	94
	1	9	109	60	178
	2	5	38	22	65
Celkem		27	196	114	337

Nejdříve spočteme aritmetické průměry:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{R} n_{i+}x_i = \frac{1}{337}(94\cdot 0 + 178\cdot 1 + 65\cdot 2) = \frac{308}{337} = 0{,}914\,,$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{S} n_{+j}y_j = \frac{1}{337}(27\cdot 1 + 196\cdot 2 + 114\cdot 3) = \frac{761}{337} = 2{,}258\,.$$

Poté dosadíme do vzorce (4.22), tj.

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

$$= \frac{(13\cdot 0\cdot 1 + 49\cdot 0\cdot 2 + \ldots + 22\cdot 2\cdot 3) - 337\cdot 0{,}914\cdot 2{,}258}{\sqrt{(94\cdot 0^2 + \ldots + 65\cdot 2^2 - 337\cdot 0{,}914^2)\cdot(27\cdot 1^2 + \ldots + 114\cdot 3^2 - 337\cdot 2{,}258^2)}} = 0{,}04\,.$$

Výsledná hodnota označuje velmi slabou přímou závislost.

IBM SPSS Statistics

V *IBM SPSS Statistics* můžeme korelační koeficienty získat dvěma způsoby. V prvním případě vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Correlations*. Výsledek zachycuje upravený výstup 4.25. Ve sloupci *Value* se nachází hodnota 0,040, která se shoduje s hodnotou vypočtenou bez použití programu. Druhým způsobem je volba *Analyze*, *Correlate, Bivariate* a zaškrtnutí položky *Pearson*. Výsledkem je korelační matice, kterou zachycuje výstup 4.26.

Hodnota korelačního koeficientu je vždy první hodnota v políčku matice.

Výstup 4.25 | Korelační koeficient vyjadřující závislost mezi počtem zaměstnání od absolvování studia a počtem zaměstnání souvisejících se studiem

Symmetric Measures		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Interval by Interval	Pearson's *R*	,040	,057	,738	,461[c]
N of Valid Cases		337			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

Výstup 4.26 | Korelační matice vyjadřující závislost mezi počtem zaměstnání od absolvování studia a počtem zaměstnání souvisejících se studiem

Correlations		Počet zaměstnání souvisejících se studiem	Počet zaměstnání od absolvování studia
Počet zaměstnání souvisejících se studiem	Pearson Correlation	1	,040
	Sig. (2-tailed)		,461
	N	337	337
Počet zaměstnání od absolvování studia	Pearson Correlation	,040	1
	Sig. (2-tailed)	,461	
	N	337	337

4.3.5 Tabulka s kvantitativní vysvětlovanou proměnnou

V případě kvantitativní vysvětlované proměnné s velkým počtem variant hodnot se ke zkoumání závislosti na proměnné s menším počtem kategorií používá buď analýza rozptylu (pokud jsou splněny určité předpoklady o náhodné složce), nebo Kruskalův-Wallisův test (viz oddíl 4.3.3). Zde se omezíme na měření intenzity statistické závislosti, při němž nemusíme zkoumat splnění výše zmíněných předpokladů, týkajících se kvantitativní proměnné a náhodné složky. Intenzitu závislosti lze vyjádřit pomocí poměru determinace nebo jeho odmocniny, označované jako **koeficient η** (éta). Tento koeficient nabývá hodnot z intervalu $\langle 0; 1 \rangle$.

Poměr determinace se počítá s využitím tzv. výpočtového tvaru rozptylu, stejně jako v případě Pearsonova korelačního koeficientu v předchozím oddílu 4.3.4, podle vzorce

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

$$= \frac{\sum_{j=1}^{S} p_{+j}(y_j - \bar{y})^2 - \sum_{i=1}^{R} p_{i+} \sum_{j=1}^{S} \frac{p_{ij}}{p_{i+}} (y_j - \bar{y}_i)^2}{\sum_{j=1}^{S} p_{+j}(y_j - \bar{y})^2} =$$

$$= \frac{\sum_{j=1}^{S} p_{+j} \left(y_j - \sum_{j=1}^{S} p_{+j} y_j \right)^2 - \sum_{i=1}^{R} p_{i+} \sum_{j=1}^{S} \frac{p_{ij}}{p_{i+}} \left(y_j - \sum_{j=1}^{S} \frac{p_{ij}}{p_{i+}} y_j \right)^2}{\sum_{j=1}^{S} p_{+j} \left(y_j - \sum_{j=1}^{S} p_{+j} y_j \right)^2} =$$

$$= \frac{\sum_{j=1}^{S} p_{+j} y_j^2 - \left(\sum_{j=1}^{S} p_{+j} y_j \right)^2 - \sum_{i=1}^{R} p_{i+} \left(\sum_{j=1}^{S} \frac{p_{ij}}{p_{i+}} y_j^2 - \left(\sum_{j=1}^{S} \frac{p_{ij}}{p_{i+}} y_j \right)^2 \right)}{\sum_{j=1}^{S} p_{+j} y_j^2 - \left(\sum_{j=1}^{S} p_{+j} y_j \right)^2},$$

což vede ke vztahu

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

Koeficient η vyjádřený pomocí absolutních četností lze psát ve tvaru

$$\eta_{Y|X} = \sqrt{I^2_{Y|X}} = \sqrt{\frac{\sum_{i=1}^{R} \frac{1}{n_{i+}} \left(\sum_{j=1}^{S} n_{ij} y_j \right)^2 - \frac{1}{n} \left(\sum_{j=1}^{S} n_{+j} y_j \right)^2}{\sum_{j=1}^{S} n_{+j} y_j^2 - \frac{1}{n} \left(\sum_{j=1}^{S} n_{+j} y_j \right)^2}} . \tag{4.23}$$

Příklad 4.13

Určeme intenzitu závislosti proměnné *C4* (*počet zaměstnání od absolvování studia*) na proměnné *E1* (*pohlaví*). Počet zaměstnání od absolvování studia je uvažován maximálně 3 (byl proveden výběr absolventů postupem uvedeným v příkladu 4.12). Uspořádání vstupních dat do kontingenční tabulky je ve výstupu 4.27.

Výstup 4.27 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.13

		Počet zaměstnání od absolvování studia			Celkem
		1	2	3	
Pohlaví	Muž	7	57	46	110
	Žena	19	129	61	209
Celkem		26	186	107	319

Pomocné výpočty ke koeficientu η jsou uvedeny v tabulkách 4.3 a 4.4. Výsledný koeficient, vyjadřující intenzitu závislosti počtu zaměstnání od absolvování studia na pohlaví, je

$$\eta_{Y|X} = \sqrt{\frac{\sum_{i=1}^{R} \frac{1}{n_{i+}} \left(\sum_{j=1}^{s} n_{ij} y_j \right)^2 - \frac{1}{n} \left(\sum_{j=1}^{S} n_{+j} y_j \right)^2}{\sum_{j=1}^{S} n_{+j} y_j^2 - \frac{1}{n} \left(\sum_{j=1}^{S} n_{+j} y_j \right)^2}} = \sqrt{\frac{\left(\frac{259^2}{110} + \frac{460^2}{209} \right) - \frac{719^2}{319}}{1\,733 - \frac{719^2}{319}}} = 0{,}123.$$

Jde tedy o velmi slabou závislost.

Tabulka 4.3 | Pomocné výpočty k příkladu 4.13

$y_j = j$	n_{+j}	$n_{+j} y_j$	$n_{+j} y_j^2$
1	26	26	26
2	186	372	744
3	107	321	963
Celkem	319	719	1 733

Tabulka 4.4 | Pomocné výpočty k příkladu 4.13

i	$y_j = j$	n_{ij}	$n_{ij}y_j$
1	1	7	7
1	2	57	114
1	3	46	138
2	1	19	19
2	2	129	258
2	3	61	183
Celkem		319	719

IBM SPSS Statistics

Získaný výsledek nyní porovnejme s výstupem ze systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Eta* (typ měr *Nominal by Interval*). Výsledek zachycuje výstup 4.28. Smysl má pouze druhá z uvedených hodnot, tj. pro případ, kdy vysvětlovanou proměnnou (*Dependent*) je *počet zaměstnání od absolvování studia*, která je 0,123.

Výstup 4.28 | Hodnocení závislosti počtu zaměstnání na pohlaví (příklad 4.13)

Directional Measures			**Value**
Nominal by Interval	**Eta**	**Pohlaví Dependent**	,129
		Počet zaměstnání od absolvování studia Dependent	,123

4.3.6 Čtyřpolní tabulka (pro dvě dichotomické proměnné)

Kontingenční tabulka pro dvě dichotomické proměnné má čtyři políčka a nazývá se *čtyřpolní*. Pro tento typ tabulky mohou být potřebné vzorce zapsány jednodušším způsobem než v případě jiných rozměrů. Vyjděme z tabulky relativních četností, kterou symbolicky vyjadřuje schéma 4.3.

Schéma 4.3 | Značení pro čtyřpolní tabulku relativních četností

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

Vzorec pro **chí-kvadrát statistiku** lze upravit následujícím způsobem:

$$\chi_P^2 = n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(p_{ij}-p_{i+}p_{+j})^2}{p_{i+}p_{+j}} =$$

$$= n\left(\frac{(p_{11}-p_{1+}p_{+1})^2}{p_{1+}p_{+1}} + \frac{(p_{12}-p_{1+}p_{+2})^2}{p_{1+}p_{+2}} + \frac{(p_{21}-p_{2+}p_{+1})^2}{p_{2+}p_{+1}} + \frac{(p_{22}-p_{2+}p_{+2})^2}{p_{2+}p_{+2}}\right) =$$

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

Pomocí absolutních četností můžeme zapsat

$$\chi_P^2 = n\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}. \tag{4.24}$$

Tato náhodná veličina má za předpokladu platnosti nulové hypotézy asymptoticky chí-kvadrát rozdělení pouze s jedním stupněm volnosti, tj. $\chi_P^2 \approx \chi^2[1]$.

V případě čtyřpolní tabulky se používá též korigovaná statistika. Pro výpočet *statistiky chí-kvadrát korigované na spojitost* **(Yatesova korekce)** můžeme využít některý z následujících dvou vzorců, a to buď

$$\chi_C^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}[\max(0,|n_{ij}-m_{ij}|-0,5)]^2 / m_{ij}, \tag{4.25}$$

nebo

$$\chi_C^2 = \frac{n\left(|n_{11}n_{22}-n_{12}n_{21}|-0,5n\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \quad \text{pro } |n_{11}n_{22}-n_{12}n_{21}| > 0,5n \tag{4.26}$$

(v opačném případě $\chi_C^2 = 0$).

Dále lze pro testování nezávislosti využít **věrohodnostní poměr,** který je počítán jako

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\ln(n_{ij}/m_{ij}). \tag{4.27}$$

Uvedená statistika má pro čtyřpolní tabulku asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti. Stejně jako u chí-kvadrát statistiky by měl být splněn požadavek na velikost očekávaných četností m_{ij} a rozsah výběru.

Příklad 4.14

Budeme testovat nezávislost mezi proměnnými *přínos oboru pro vstup do práce* a *přínos oboru pro další učení v rámci práce*, překódovanými do proměnných se dvěma kategoriemi s významem „ne“ (původní kategorie „1“ a „2“) a „ano“ (původní kategorie „3“ až „5“), tj. mezi proměnnými *B2a_2kat* a *B2b_2kat*. Kontingenční tabulka je ve výstupu 4.29.

Výstup 4.29 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.14

Zjištěné absolutní četnosti		Přínos oboru pro další učení v rámci práce		Celkem
		Ne	Ano	
Přínos oboru pro vstup do práce	Ne	82	58	140
	Ano	56	439	495
Celkem		138	497	635

Pearsonovu chí-kvadrát statistiku spočteme podle vzorce (4.24) jako

$$\chi_P^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} = 635\frac{(82\cdot 439 - 58\cdot 56)^2}{140\cdot 495\cdot 138\cdot 497} = 143,294\,.$$

Výsledkem je velmi vysoká hodnota, tudíž α' je číslo blízké nule ($\chi^2_{0,95}[1]$, výsledná hodnota je tedy mnohem větší než kritická hodnota). Jako závěr chí-kvadrát testu o nezávislosti můžeme uvést, že na hladině významnosti 5 % (i 1 %) nulovou hypotézu o nezávislosti zamítáme. Usuzujeme, že analyzované proměnné jsou závislé.

Pro výpočet *korigované statistiky chí-kvadrát* podle vzorce (4.25) potřebujeme znát očekávané četnosti, viz výstup 4.30.

Výstup 4.30 | Kontingenční tabulka očekávaných absolutních četností k příkladu 4.14

Očekávané absolutní četnosti		Přínos oboru pro další učení v rámci práce		Celkem
		Ne	Ano	
Přínos oboru pro vstup do práce	Ne	30,43	109,57	140,00
	Ano	107,57	387,43	495,00
Celkem		138,00	497,00	635,00

Nejdříve provedeme pomocné výpočty pro použití vzorce (4.25), viz tabulka 4.5.

Tabulka 4.5 | Pomocné výpočty pro výpočet statistiky chí-kvadrát s opravou na spojitost

n_{ij}	m_{ij}	$b_{ij} = \lvert n_{ij} - m_{ij}\rvert - 0{,}5$	$[max(0, b_{ij})]2$	$[max(0, b_{ij})]^2/m_{ij}$
82	30,43	51,07	2 608,145	85,723
58	109,57	51,07	2 608,145	23,802
56	107,57	51,07	2 608,145	24,245
439	387,43	51,07	2 608,145	6,732

Výsledkem je 140,5 (při dosazení výše uvedených zaokrouhlených čísel), neboť

$$\chi_C^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}[\max(0,|n_{ij} - m_{ij}| - 0,5)]^2 / m_{ij} = 140,5 .$$

Můžeme také použít vzorec (4.26), pomocí něhož dostaneme stejný výsledek, tj.

$$\chi_C^2 = \frac{n\left(|n_{11}n_{22} - n_{12}n_{21}| - 0,5n\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}} = \frac{635\cdot\left(|82\cdot 439 - 58\cdot 56| - 0,5\cdot 635\right)^2}{140\cdot 495\cdot 138\cdot 497} = 140,529 .$$

Výsledkem je sice menší hodnota, než je Pearsonova chí-kvadrát statistika bez korekce, nicméně hodnota je stále vysoká, tudíž můžeme stejně jako v předchozím případě uvést, že na hladině významnosti 5 % (i 1 %) nulovou hypotézu o nezávislosti zamítáme.

Dále vypočtěme *věrohodnostní poměr*. Pomocné výpočty pro vzorec (4.27) jsou uvedeny v tabulce 4.6. Výsledkem je 125,414, neboť

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\ln(n_{ij} / m_{ij}) = 2\cdot 62,707 = 125,414 .$$

Na hladině významnosti 5 % i 1 % zamítáme nulovou hypotézu o nezávislosti, stejně jako v předchozích případech.

Tabulka 4.6 | Pomocné výpočty pro věrohodnostní poměr

n_{ij}	m_{ij}	n_{ij} / m_{ij}	$b_{ij} = \ln(n_{ij} / m_{ij})$	$n_{ij}b_{ij}$
82	30,43	2,695	0,991	81,300
58	109,57	0,529	–0,636	–36,900
56	107,57	0,521	–0,653	–36,559
439	387,43	1,133	0,125	54,865

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupem ze systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Chi-square*. Výsledky zachycuje upravený výstup 4.31, který obsahuje Pearsonovu statistiku chí-kvadrát (*Pearson Chi-Square*), statistiku korigovanou na spojitost (*Continuity Correction*) a věrohodnostní poměr (*Likelihood Ratio*). Je zřejmé, že se všechny výsledky shodují s výpočty podle vzorců, a to jak pokud jde o samotné hodnoty statistik (*Value*), tak pokud jde o minimální hladinu významnosti (*Asymptotic Significance (2-sided)*).

Výstup 4.31 | Výsledky testování nezávislosti proměnných přínos oboru pro vstup do práce a přínos oboru pro další učení v rámci práce (příklad 4.14)

Chi-Square Tests	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	143,294[a]	1	,000
Continuity Correction[b]	140,529	1	,000
Likelihood Ratio	125,414	1	,000
N of Valid Cases	635		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 30,43.
b. Computed only for a 2 × 2 table.

♦ ♦ ♦

Nyní odvodíme zjednodušený tvar pro **Pearsonův korelační koeficient**. Schéma 4.3 můžeme přepsat jako „jednorozměrnou" tabulku, v níž uvedeme jednotlivé kombinace kategorií a k nim četnosti, viz schéma 4.4.

Schéma 4.4 | Převod dvourozměrné tabulky do vstupní matice vážených kombinací kategorií

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

Pro výpočet korelačního koeficientu potřebujeme znát kovarianci a směrodatné odchylky, jejichž výpočet je založen na odchylkách od aritmetického průměru. Ten je počítán podle vztahu (3.5), tj. $\bar{x} = \sum x_i \cdot p_i$.

Na základě tohoto vztahu je *aritmetický průměr* proměnné X spočten jako

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

pro proměnou Y je průměr

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

Při výpočtu *rozptylu* můžeme vyjít ze vztahu (3.7), tj. $s^2 = \sum x_i^2 p_i - \bar{x}^2$. Protože pro binární proměnnou je $\sum x_i^2 p_i = \sum x_i p_i$, lze rozptyl zapsat ve tvaru $s^2 = \bar{x} - \bar{x}^2$ nebo též jako $s^2 = \bar{x}\,(1 - \bar{x})$. Pro proměnnou X je rozptyl

$$s_x^{\,2} = p_{1+}\, p_{2+},$$

pro proměnnou Y je rozptyl

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

Kovarianci vypočteme podle vztahu $s_{XY} = \overline{xy} - \bar{x} \cdot \bar{y}$, to znamená, že

$$\begin{aligned} s_{XY} &= p_{22} - p_{2+}p_{+2} = p_{22} - (p_{21} + p_{22})(p_{12} + p_{22}) = \\ &= p_{22} - (p_{12}p_{21} + p_{12}p_{22} + p_{21}p_{22} + p_{22}^2) = \\ &= p_{22}(1 - p_{12} - p_{21} - p_{22}) - p_{12}p_{21} = p_{11}p_{22} - p_{12}p_{21}. \end{aligned}$$

Korelační koeficient je podílem kovariance a součinu směrodatných odchylek, tj.

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1+}p_{2+}p_{+1}p_{+2}}},$$

což můžeme zapsat pomocí absolutních četností jako

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}. \tag{4.28}$$

Tato míra pro čtyřpolní tabulku bývá též označována jako **koeficient asociace**. Jde o míru symetrickou, která hodnotí intenzitu vzájemné závislosti a též její typ (zda je přímá, či nepřímá).

Pro čtyřpolní tabulku tedy platí, že $\chi_P^2 = nr^2$. To znamená, že míry závislosti vycházející z chí-kvadrát statistiky, viz (4.4) až (4.7), jsou v případě čtyřpolní tabulky funkcí korelačního koeficientu. Dále platí, že vzorec pro Cramérovo V je v tomto speciálním případě shodný se vzorcem pro koeficient φ. Navíc můžeme odvodit, že

$$|r| = \sqrt{\frac{\chi_P^2}{n}} = \phi = V. \tag{4.29}$$

Příklad 4.15

Navážeme na příklad 4.14. Mezi proměnnými *B2a_2kat* a *B2b_2kat* jsme zjistili závislost. Spočtěme nyní jednak míry závislosti vycházející z chí-kvadrát statistiky, jednak koeficient asociace (to znamená korelační koeficient).

Na základě vzorců (4.4) až (4.6) můžeme vypočítat *Pearsonův kontingenční koeficient* C_P a *koeficient φ*, jehož hodnota je stejná jako hodnota *Cramérova V*, tedy

$$C_P = \sqrt{\frac{\chi_P^2}{\chi_P^2 + n}} = \sqrt{\frac{143,294}{143,294 + 635}} = 0,429,$$

$$\phi = V = \sqrt{\frac{\chi_P^2}{n}} = \sqrt{\frac{143,294}{635}} = 0,475.$$

Koeficient asociace (tj. korelační koeficient) spočteme podle vzorce (4.28) jako

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}} = \frac{82 \cdot 439 - 58 \cdot 56}{\sqrt{140 \cdot 495 \cdot 138 \cdot 497}} = 0,475 .$$

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupem ze systému *SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnosti *Contingency coefficient*, *Phi and Cramér's V* a *Correlations*. Výsledek zachycuje upravený výstup 4.32, který obsahuje koeficient φ (*Phi*), Cramérovo V (*Cramer's V*), kontingenční koeficient (*Contingency Coefficient*) a korelační koeficient (*Pearson's R*). Získané hodnoty se shodují s výsledky získanými dosazením do vzorců.

Výstup 4.32 | Hodnocení míry závislosti přínosu oboru pro vstup do práce a přínosu oboru pro další učení v rámci práce (příklad 4.15)

Symmetric Measures		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Nominal by Nominal	Phi	,475			,000
	Cramer's *V*	,475			,000
	Contingency Coefficient	,429			,000
Interval by Interval	Pearson's *R*	,475	,042	13,582	,000[c]
N of Valid Cases		635			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

♦ ♦ ♦

Pokud bychom chtěli testovat nezávislost dvou znaků, ale nebyl by splněn požadavek týkající se očekávaných četností, pak je vhodný exaktní test. **Fisherův exaktní test** vychází z předpokladu, že marginální četnosti jsou považovány za neměnné, a že data jsou tudíž výběrem z hypergeometrického rozdělení. V takovém případě lze testovat pouze relativní četnost v levém horním políčku (odpovídající prvnímu řádku a prvnímu sloupci), neboť četnosti všech ostatních políček jsou potom jednoznačně určeny marginálními četnostmi. Testuje se tedy $H_0: \pi_{11} = p_{1+}p_{+1}$ vůči oboustranné, případně jednostranné alternativní hypotéze. Programové systémy obvykle zobrazují výsledky pouze pro jednu jednostrannou hypotézu, to v závislosti na vztahu četnosti n_{11} k odpovídající teoretické četnosti $m_{11} = (n_{1+}n_{+1})/n$. Platí-li, že $n_{11} > m_{11}$, pak jde o hypotézu $H_1: \pi_{11} > p_{1+}p_{+1}$, v opačném případě je $H_1: \pi_{11} < p_{1+}p_{+1}$.

Počítají se pravděpodobnosti výskytu všech možných variant četností v kontingenční tabulce, které dávají stejné marginální četnosti jako tabulka četností ze zjištěných dat. Tyto pravděpodobnosti lze získat vyhledáním hodnot pravděpodobnostní funkce hypergeometrického rozdělení (viz příloha na konci knihy). Při porovnání s výběrem bez vracení je počet prvků, z nichž vybíráme, roven počtu pozorování n, počet prvků se sledovanou vlastností je sloupcový marginální součet n_{+1} a počet výběrů je řádkový marginální součet n_{1+}. Pravděpodobnost, že četnost v políčku n_{11} nabude hodnoty t, je dána vztahem

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

Pro každou variantu četností lze spočítat pravděpodobnost

$$p = \frac{n_{1+}!\, n_{2+}!\, n_{+1}!\, n_{+2}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}. \tag{4.30}$$

Minimální hladinu významnosti, od níž zamítáme hypotézu H_0, spočteme pro pravostrannou a oboustrannou alternativní hypotézu podle vztahu

kde A je $\alpha' = \sum_A p$,

pro oboustranný test: množina variant četností, kde p je menší nebo rovno pravděpodobnosti zjištěných (pozorovaných) četností,

pro jednostranný test: množina případů 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.

Pro druhou variantu alternativní hypotézy spočteme minimální hladinu významnosti podle vztahu $\alpha' = 1 - \sum_A p$, kde A je množina variant četností shodná s množinou pro výše uvedený jednostranný test s výjimkou zjištěných četností.

Příklad 4.16

Analyzujme stejné proměnné jako v příkladu 4.14, ale vyberme hodnoty jen pro ženy, které řídí jiné pracovníky (výběr se provádí pomocí voleb *Data*, *Select Cases*). Protože těchto absolventek je pouze 13, četnosti v kontingenční tabulce nesplňují předpoklady pro použití chí-kvadrát testu o nezávislosti, viz výstup 4.33. Použijeme tedy Fisherův exaktní test.

Výstup 4.33 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.16

Zjištěné absolutní četnosti		Přínos oboru pro další učení v rámci práce		Celkem
		Ne	Ano	
Přínos oboru pro vstup do práce	Ne	4	1	5
	Ano	2	6	8
Celkem		6	7	13

Vzhledem k marginálním četnostem existuje pět variant sdružených četností. Tyto varianty a jejich pravděpodobnosti jsou uvedeny v tabulce 4.7 (zjištěné četnosti odpovídají druhé variantě, která je zvýrazněna tučným písmem). Hodnoty lze buď zjistit pomocí pravděpodobnostní funkce hypergeometrického rozdělení v bodech odpovídajících četnosti prvního políčka tabulky, nebo je lze spočítat podle vzorce (4.30), tj.:

$$p_1 = \frac{5!8!6!7!}{13!5!0!1!7!} = 0{,}004\ 66\,,\quad p_2 = \frac{5!8!6!7!}{13!4!1!2!6!} = 0{,}081\ 59\,,\quad p_3 = \frac{5!8!6!7!}{13!3!2!3!5!} = 0{,}326\ 34\,,$$

$$p_4 = \frac{5!8!6!7!}{13!2!3!4!4!} = 0{,}407\ 93\,,\quad p_5 = \frac{5!8!6!7!}{13!1!4!5!3!} = 0{,}163\ 17\,,\quad p_6 = \frac{5!8!6!7!}{13!0!5!6!2!} = 0{,}016\ 32\,.$$

Tabulka 4.7 | Pravděpodobnosti možných variant četností

Pořadí *l*-té varianty	(*i*,*j*)				p_l
	(1,1)	(1,2)	(2,1)	(2,2)	
1	5	0	1	7	0,004 66
2	**4**	**1**	**2**	**6**	**0,081 59**
3	3	2	3	5	0,326 34
4	2	3	4	4	0,407 93
5	1	4	5	3	0,163 17
6	0	5	6	2	0,016 32

Pro oboustrannou alternativní hypotézu spočteme minimální hladinu významnosti jako

$$\alpha' = 0{,}081\ 59 + 0{,}004\ 66 + 0{,}016\ 32 = 0{,}102\ 57\,.$$

Teoretická četnost $m_{11} = (n_{1+}n_{+1})/n = 2{,}31$ je menší než zjištěná četnost 4. Výsledek pro pravostrannou alternativní hypotézu zahrnuje případy $n_{11} \geq 4$ (nebo od jedničky odečteme hodnotu distribuční funkce v bodě $n_{11} = 3$), tedy

$$\alpha' = 0{,}081\ 59 + 0{,}004\ 66 = 0{,}086\ 25\,.$$

Pro obě alternativní hypotézy na 5% hladině významnosti nezamítáme hypotézu o nezávislosti. Pokud by zjištěné četnosti odpovídaly první variantě četností, pak bychom na 5% (i 1%) hladině významnosti mohli hypotézu o nezávislosti zamítnout.

Pro levostrannou alternativní hypotézu pak minimální hladinu významnosti spočteme jako (nebo zjistíme hodnotu distribuční funkce v bodě $n_{11} = 4$)

$$\alpha' = 1 - 0,004\,66 = 0,995\,34 .$$

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupem ze systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Chi-square*. Výsledek zachycuje upravený výstup 4.34. Jde pouze o část výstupu, neboť systém *IBM SPSS Statistics* zobrazuje hodnoty dalších statistik, jak bylo uvedeno v příkladu 4.14 (výstup 4.31). Ve výstupu jsou ponechány pouze výsledky dvou variant testů, a to pro oboustrannou alternativní hypotézu (*Exact Sig. (2-sided)*) a pravostrannou alternativní hypotézu (*Exact Sig. (1-sided)*). Zobrazené minimální hladiny významnosti odpovídají výše uvedeným zaokrouhleným hodnotám.

Výstup 4.34 | Výsledky testování nezávislosti proměnných přínos oboru pro vstup do práce a přínos oboru pro další učení v rámci práce pomocí Fisherova exaktního testu

Chi-Square Tests	Value	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Fisher's Exact Test		,103	,086
N of Valid Cases	13		

♦ ♦ ♦

Poznámka

Exaktní testy se používají i v případě tabulek větších rozměrů, pokud nejsou splněny předpoklady pro použití testů založených na chí-kvadrát statistikách. Systém *IBM SPSS Statistics* tyto testy zahrnuje, neprovádějí se ovšem automaticky, jak je tomu u čtyřpolní tabulky. Pro Fisherův exaktní test je třeba jednak zadat volbu *Chi-square* v části *Statistics*, jednak v části *Exact* zvolit možnost *Exact*. Je-li kontingenční tabulka rozsáhlá (proměnné mohou nabývat mnoha kategorií) a výpočet pomocí Fisherova exaktního testu by byl náročný na čas či paměť počítače, pak lze v systému *IBM SPSS Statistics* v části *Exact* vybrat možnost *Monte Carlo*, kdy se provádí aproximace metodou Monte Carlo.

Uveďme ještě jinou statistiku, která je stejně jako Pearsonova chí-kvadrát statistika funkcí korelačního koeficientu. Je to **Mantelova-Haenszelova chí-kvadrát statistika**, která vychází ze vztahu

$$Q_{MH} = \frac{(n_{11} - m_{11})^2}{v_{11}} ,$$

kde v_{11} je rozptyl pro četnosti v prvním políčku kontingenční tabulky. Stejně jako v případě Fisherova exaktního testu se vychází z předpokladu, že četnosti jsou výběrem z hypergeometrického rozdělení s rozptylem

$$v_{11} = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n^2(n-1)}.$$

Tato statistika má asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti.
Protože pro čtyřpolní tabulku platí

$$\left|n_{11} - m_{11}\right| = \left|\frac{n_{11}n_{22} - n_{12}n_{21}}{n}\right|,$$

lze odvodit, že

$$\chi_P^2 = \frac{n}{n-1}Q_{MH},$$

a tudíž

$$Q_{MH} = (n-1)r^2. \tag{4.31}$$

Příklad 4.17

Vraťme se k příkladu 4.14 a doplňme testy založené na chí-kvadrát statistikách o *Mantelovu-Haenszelovu statistiku*. Vyjdeme tedy z kontingenční tabulky ve výstupu 4.29. V příkladu 4.15 jsme vypočetli hodnotu Pearsonova korelačního koeficientu, a to 0,475 036 (výše byla uvedena zaokrouhlená hodnota). Dosazením do vzorce (4.31) dostáváme

$$Q_{MH} = (n-1)r^2 = (635-1)\cdot(0{,}475\,036)^2 = 143{,}068.$$

Výsledkem je velmi vysoká hodnota, tudíž α′ je číslo blízké nule. Stejně jako u testů uvedených v příkladu 4.14 nulovou hypotézu o nezávislosti zamítáme na 1% hladině významnosti. Usuzujeme, že analyzované proměnné jsou závislé.

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupem systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Chi-square*. Výsledek zachycuje upravený výstup 4.35. Jde pouze o část výstupu, neboť systém zobrazuje hodnoty dalších statistik, jak bylo uvedeno v příkladech 4.14 a 4.16. Mantelova-Haenszelova statistika je zde uvedena pod názvem *Linear-by-Linear Association*. Výsledky získané jak bez pomoci systému, tak s jeho využitím, jsou shodné.

Výstup 4.35 | Výsledky testování nezávislosti proměnných přínos oboru pro vstup do práce a přínos oboru pro další učení v rámci práce pomocí Mantelovy-Haenszelovy statistiky

Chi-Square Tests	Value	df	Asymptotic Significance (2-sided)
Linear-by-Linear Association	143,068	1	,000
N of Valid Cases	635		

♦ ♦ ♦

Speciálním testem pro čtyřpolní tabulku je **McNemarův test**[6]. Jde vlastně o párový test, který může být využit například při zjišťování, zda se shodují názory týchž respondentů ve dvou různých obdobích. Máme tedy k dispozici dvojice hodnot, u nichž nás zajímá vzájemný vztah. Testuje se nulová hypotéza o shodě četností v políčkách na vedlejší diagonále, což je odvozeno ze vztahu $\frac{n_{1+}}{n} = \frac{n_{+1}}{n}$ neboli $\frac{n_{11}+n_{12}}{n} = \frac{n_{11}+n_{21}}{n}$, čili $\frac{n_{12}}{n} = \frac{n_{21}}{n}$, odtud H_0: $\pi_{12} = \pi_{21}$, H_1: $\pi_{12} \neq \pi_{21}$.

Pro exaktní test je minimální hladina významnosti pro zamítnutí nulové hypotézy daná vztahem

$$\alpha' = 2 \sum_{i=0}^{\min\{n_{12}, n_{21}\}} \binom{n_{12}+n_{21}}{i} (0,5)^{n_{12}+n_{21}}. \tag{4.32}$$

Polovina této hodnoty vyjadřuje pravděpodobnost, že náhodná veličina s binomickým rozdělením (pravděpodobnost, že nastane sledovaný náhodný jev, je 0,5 a počet náhodných pokusů je $n_{12} + n_{21}$) nabude hodnoty $\min\{n_{12}, n_{21}\}$ nebo menší.

Je-li $n_{12} + n_{21} > 25$, lze použít aproximaci binomického rozdělení rozdělením chí-kvadrát. Statistika s opravou na spojitost, která má za platnosti nulové hypotézy přibližně chí-kvadrát rozdělení s jedním stupněm volnosti, je dána vztahem

$$Q_M = \frac{\left(|n_{12} - n_{21}| - 1\right)^2}{n_{12} + n_{21}}. \tag{4.33}$$

Příklad 4.18

Vyberme pouze absolventy, kteří změnili zaměstnání a nyní pracují v instituci ve veřejném sektoru. Výběr provedeme pomocí nabídek *Data*, *Select Cases*, možnosti *If condition is satisfied* (tlačítko *If*) a zadáním podmínky pro výběr ve tvaru „C5 = 1 & D1 = 0 & D4 = 1“ (absolvent má placené zaměstnání, nepůsobí v prvním zaměstnání a pracuje v instituci ve veřejném sektoru). Analyzujme proměnné *studijní obor vhodný pro první zaměstnání* a *studijní obor vhodný pro současné zaměstnání* překódované do proměnných se dvěma kategoriemi (proměnné *C3_2kat* a *D3_2kat*), viz výstup 4.36.

6 V IBM SPSS Statistics se používá jeho rozšíření na čtvercové tabulky větších rozměrů. Test se nazývá McNemarův-Bowkerův. V této knize je uvažována pouze čtyřpolní tabulka.

Výstup 4.36 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.18

Zjištěné absolutní četnosti		Studijní obor vhodný pro současné zaměstnání		Celkem
		Vystudovaný nebo příbuzný obor	Jiný nebo žádný studijní obor	
Studijní obor vhodný pro první zaměstnání	Vystudovaný nebo příbuzný obor	28	1	29
	Jiný nebo žádný studijní obor	10	0	10
Celkem		38	1	39

Protože součet sdružených četností na vedlejší diagonále je pouze 11, použijeme exaktní test, tedy budeme uvažovat binomické rozdělení Bi [11; 0,5]. Pro takové rozdělení je distribuční funkce $F(1) = 0{,}006$. Minimální hladina významnosti je pak dvojnásobkem této hodnoty, tj. $\alpha' = 2\ F(1) = 0{,}012$. Protože $0{,}012 < 0{,}05$, na 5% hladině významnosti hypotézu H_0 o shodě podílů zamítáme a usuzujeme, že pokud absolventi (s ohledem na uvedený výběr) změní zaměstnání, je to ve prospěch uplatnění vystudovaného nebo příbuzného studijního oboru. Ovšem na 1% hladině významnosti nulovou hypotézu zamítnout nemůžeme.

IBM SPSS Statistics

V *IBM SPSS Statistics* vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *McNemar*. Výsledek zachycuje upravený výstup 4.37. Jde pouze o část výstupu, neboť systém *IBM SPSS Statistics* zobrazuje hodnoty dalších statistik, jak bylo uvedeno v příkladech 4.14, 4.16 a 4.17. Výsledek získaný s využitím *IBM SPSS Statistics* je stejný jako výsledek s použitím binomického rozdělení.

Výstup 4.37 | Výsledek McNemarova testu porovnávajícího vhodnost studijního oboru v prvním a současném zaměstnání

Chi-Square Tests	Value	Exact Sig. (2-sided)
McNemar Test		,012[a]
N of Valid Cases	39	

a. Binomial distribution used.

Příklad 4.19

Analyzujme proměnné *C3_2kat* a *D3_2kat* pro absolventy, kteří změnili zaměstnání a nyní nepracují v instituci ve veřejném sektoru, viz výstup 4.38.

Výstup 4.38 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.19

Zjištěné absolutní četnosti		Studijní obor vhodný pro současné zaměstnání		Celkem
		Vystudovaný nebo příbuzný obor	Jiný nebo žádný studijní obor	
Studijní obor vhodný pro první zaměstnání	Vystudovaný nebo příbuzný obor	176	23	199
	Jiný nebo žádný studijní obor	44	39	83
Celkem		220	62	282

Při výpočtu můžeme postupovat buď *exaktně* pomocí binomického rozdělení, nebo aproximativně pomocí chí-kvadrát rozdělení. V prvním případě potřebujeme znát hodnotu distribuční funkce v bodě 23 pro binomické rozdělení Bi[282; 0,5]. Pro toto rozdělení lze zjistit, že F(23) = 0,007. Minimální hladinu významnosti pro zamítnutí shody četností políček na vedlejší diagonále spočteme podle vztahu $\alpha' = 2\ F(23) = 0{,}014$. Na hladině významnosti 5 % tedy zamítáme nulovou hypotézu o shodě četností. Můžeme proto učinit závěr, že v současném zaměstnání absolventi (s ohledem na uvedený výběr) uplatňují vystudovaný nebo příbuzný studijní obor ve větší míře než v prvním zaměstnání. Na 1% hladině významnosti ale takový závěr učinit nemůžeme.

S využitím aproximace chí-kvadrát rozdělením spočteme

$$Q_M = \frac{\left(|n_{12} - n_{21}| - 1\right)^2}{n_{12} + n_{21}} = \frac{\left(|23 - 44| - 1\right)^2}{23 + 44} = 5{,}97.$$

Jde o kvantil $\chi^2_{0{,}985}$ [1], to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je 0,015. Nulovou hypotézu o shodě četností můžeme zamítnout na 5% hladině významnosti.

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupem systému *SPSS*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *McNemar*. Výsledek zachycuje upravený výstup 4.39. Je na něm uvedeno, že bylo použito binomické rozdělení, což odpovídá výsledku zjištěnému pomocí distribuční funkce.

Výstup 4.39 | Výsledek McNemarova testu porovnávajícího vhodnost studijního oboru v prvním a současném zaměstnání

Chi-Square Tests	Value	Exact Sig. (2-sided)
McNemar Test		,014[a]
N of Valid Cases	282	

a. Binomial distribution used.

Další možností je použití procedury nabízené v rámci neparametrických testů. V tom případě zvolíme *Analyze*, *Nonparametric Tests* a *Related Samples*. V listu *Fields* zadáme dvě analyzované proměnné (definované jako nominální či ordinální), v listu *Settings* zvolíme *Customize tests* a v části *Test for Change in Binary Data* vybereme *McNemar's test (2 samples)*. Obdržíme výstup se zadáním úlohy, zvolenou metodou, minimální hladinou významnosti a závěrem. V okně *Model Viewer* získáme výstup 4.40 včetně testové statistiky Q_M, která je 5,97 a shoduje s hodnotou vypočtenou bez použití programu. Minimální hladina významnosti, od které zamítáme nulovou hypotézu, je shodná s hodnotou získanou výpočtem na základě aproximace chí-kvadrát rozdělením.

Výstup 4.40 | Výsledek McNemarova testu porovnávajícího uplatnění oboru

Total N	282
Test Statistic	5,970
Degrees of Freedom	1
Asymptotic Sig. (2-sided test)	,015

♦ ♦ ♦

Kromě měr závislosti založených na chí-kvadrát statistice jsou pro čtyřpolní tabulku používány také další míry. Nejjednodušší z nich je **procentní rozdíl**. Na rozdíl od korelačního koeficientu jde o míru asymetrickou, která hodnotí závislost vysvětlované proměnné na proměnné vysvětlující (například, jak je přijetí či nepřijetí na vysokou školu ovlivněno skutečností, zda uchazeč absolvoval střední školu státní či soukromou).

Pokud vytvoříme tabulku obsahující řádkové procentní podíly, pak procentní rozdíly v obou sloupcích budou stejné. Tento procentní rozdíl můžeme interpretovat jako hodnotu, která udává, o kolik procent se liší hodnota vysvětlované proměnné při porovnání dvou variant hodnot proměnné vysvětlující. Například o kolik procent se liší počty přijatých uchazečů na vysokou školu v závislosti na typu střední školy (státní či soukromé). Na základě absolutních četností počítáme procentní rozdíl podle vztahu

$$PR_{Y|X} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} = \frac{(n_{21}+n_{22})\,n_{11} - (n_{11}+n_{12})\,n_{21}}{(n_{11}+n_{12})\cdot(n_{21}+n_{22})} = \tag{4.34}$$

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

Příklad 4.20

Budeme analyzovat, zda se liší uplatnění studijního oboru podle pohlaví, tj. jak proměnná *D3* překódovaná do proměnné *D3_2kat* se dvěma kategoriemi závisí na proměnné *E1*. Kontingenční tabulka absolutních četností je ve výstupu 4.41.

Výstup 4.41 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.20

Zjištěné četnosti		Studijní obor vhodný pro současné zaměstnání		Celkem
		Vystudovaný nebo příbuzný obor	Jiný nebo žádný studijní obor	
Pohlaví	Muž	192	35	227
	Žena	224	64	288
Celkem		416	99	515

Dosazením do vzorce (4.34) dostáváme

$$PR_{Y|X} = \frac{192}{227} - \frac{224}{288} = \frac{192 \cdot 64 - 35 \cdot 224}{192 \cdot 64 + 35 \cdot 224 + 192 \cdot 224 + 35 \cdot 64} = 0,07 .$$

Mužů (absolventů), kteří uplatní vystudovaný nebo příbuzný obor, je tedy o 7 % více, než žen.

♦ ♦ ♦

Další mírou je **Yuleho *Q***, které je mírou symetrickou. Předpokládáme, že obě dvě analyzované dichotomické proměnné jsou stejně kódované, například pomocí hodnot 0 a 1. Součin četností na hlavní diagonále pak vyjadřuje počet párů objektů, z nichž jeden obsahuje u obou proměnných nuly a druhý jedničky. Součin četností na vedlejší diagonále vyjadřuje počet párů objektů, z nichž první obsahuje u jedné proměnné nulu a u druhé jedničku, zatímco u druhého objektu je to naopak. U ordinálních proměnných Yuleho Q měří relativní „přebytek" konkordantních párů nad páry diskordantními (viz oddíl 4.3.2), jde tedy o speciální název pro **koeficient γ**, viz vzorec (4.18), používaný v případě čtyřpolních tabulek.

Ze vzorců pro $PR_{Y|X}$ a $PR_{X|Y}$ je do výpočtu symetrické míry zahrnut čitatel a ze jmenovatele pouze část společná pro obě asymetrické míry, to znamená, že

$$Q = \gamma = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} . \tag{4.35}$$

Příklad 4.21

Vyjdeme ze stejného zadání jako v příkladu 4.14, viz výstup 4.29. To znamená, že budeme zjišťovat vztah mezi proměnnými *přínos oboru pro vstup do práce* a *přínos oboru pro další učení v rámci práce*, překódovanými do proměnných se dvěma kategoriemi (proměnné *B2a_2kat* a *B2b_2kat*). Dosazením do vzorce (4.35) dostáváme

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = \frac{82 \cdot 439 - 58 \cdot 56}{82 \cdot 439 + 58 \cdot 56} = 0,834 .$$

Na základě koeficientu Q můžeme odvodit, že jde o poměrně silnou závislost, přičemž počet stejných hodnot převažuje nad počtem různých hodnot (kladné znaménko koeficientu). Hodnota tohoto výsledku je vyšší, než jsou hodnoty koeficientů založených na chí-kvadrát statistice (viz příklad 4.15), nezohledňujících pořadí kategorií.

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupem ze systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Gamma*. Výsledek zachycuje výstup 4.42. Získaná hodnota koeficientu γ (0,834) odpovídá hodnotě spočtené podle vzorce (4.35).

Výstup 4.42 | Hodnocení závislosti přínosu oboru pro vstup do práce a přínosu oboru pro další učení v rámci práce pomocí koeficientu gama

Symmetric Measures		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Gamma	,834	,034	9,312	,000
N of Valid Cases		635			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

♦ ♦ ♦

Dále lze využít **Yuleho koeficient vazby**, daný vztahem

$$Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}}. \tag{4.36}$$

Příklad 4.22

Vyjdeme ze stejného zadání jako v příkladu 4.14 a 4.21, viz výstup 4.29. Budeme tedy zjišťovat vztah mezi proměnnými *přínos oboru pro vstup do práce* a *přínos oboru pro další učení v rámci práce*, překódovanými do proměnných se dvěma kategoriemi (*B2a_2kat* a *B2b_2kat*). Dosazením do vzorce (4.36) dostáváme

$$Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}} = \frac{\sqrt{82 \cdot 439} - \sqrt{58 \cdot 56}}{\sqrt{82 \cdot 439} + \sqrt{58 \cdot 56}} = 0,538 .$$

Výsledná hodnota je nižší než v případě Yuleho Q, spočteného v příkladu 4.21.

♦ ♦ ♦

Stejně jako lze zapsat jednodušším způsobem statistiku chí-kvadrát a Pearsonův korelační koeficient, lze zjednodušit i zápis dalších koeficientů. Pro **Kendallovo τ_b**, viz vzorec (4.19), dostáváme vztah

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

$$= \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11}n_{22} + n_{12}n_{21} + n_{11}n_{12} + n_{21}n_{22})(n_{11}n_{22} + n_{12}n_{21} + n_{11}n_{21} + n_{12}n_{22})}} =$$

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

To znamená, že pro čtyřpolní tabulku je Pearsonův korelační koeficient identický s Kendallovým τ_b.

Pro **Somersovo *d***, viz vzorec (4.21), platí:

$$d = \frac{2(C - D)}{2(C + D) + T_X + T_Y} =$$

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

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

Mírou souhlasu pro čtyřpolní tabulku je **Hamannův koeficient**, což je podíl

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

Tento koeficient nabývá hodnot v intervalu $\langle -1; 1\rangle$. Hodnoty 1 nabývá v případě, kdy se nenulové četnosti vyskytují pouze na hlavní diagonále, to znamená, že byly zaznamenány pouze dvojice shodných kategorií. Hodnoty –1 nabývá, pokud naopak nebyla zjištěna žádná dvojice shodných kategorií, a hodnoty 0 v případě stejného počtu shod jako případů neshod.

Kromě dosud uvedených měr, které jsou speciálními případy měr pro kontingenční tabulky větších rozměrů než 2 × 2, existují míry speciálně určené pouze pro čtyřpolní tabulky. Příkladem takové míry je **poměr šancí**. Používá se zejména v případech, kdy se máme rozhodnout pro jednu ze dvou možností, například zda se léčit, či neléčit (na základě zjištěných četností ve vztahu k počtu přeživších a zemřelých), zda se zúčastnit školení (na základě zjištěných četností ve vztahu k počtu úspěšných a neúspěšných respondentů v dané oblasti) apod., viz též [4].

Pokud jsou kategorie ve čtyřpolní tabulce vhodně uspořádány (například v prvním políčku je počet léčených, kteří přežili, příp. počet účastníků školení, kteří byli úspěšní), pak poměr součinů četností na hlavní a vedlejší diagonále je poměrem šancí, který indikuje, zda je lepší podniknout určitou aktivitu (nechat se léčit, zúčastnit se školení), či

nikoliv (je-li výsledná hodnota menší než 1). Hodnota 1 značí nezávislost analyzovaných proměnných. Matematicky poměr šancí zapíšeme jako

$$\psi = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{4.37}$$

Programové systémy obvykle provádějí hlubší analýzu vztahů. Výklad této analýzy si uvedeme současně s následujícím příkladem.

Příklad 4.23

Zajímá nás, zda *působení v prvním zaměstnání až do současné doby* (proměnná *D1*) závisí na *typu studijního programu* (ve smyslu, zda magisterské studium následovalo po bakalářském, resp. jiném vysokoškolském studiu, či nikoliv, proměnná *A3*). Odpovídající kontingenční tabulka je ve výstupu 4.43.

Výstup 4.43 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.23

Zjištěné absolutní četnosti		Působení v prvním zaměstnání do současné doby		Celkem
		Ne	Ano	
Typ studijního programu	Magisterský dlouhý	274	165	439
	Magisterský navazující	96	86	182
Celkem		370	251	621

Podíl absolventů dlouhého magisterského programu, kteří mění zaměstnání, můžeme vyjádřit poměrem $n_{11}/n_{1+} = 274/439 = 0{,}62$. Poměr tohoto podílu k podílu absolventů navazujícího magisterského programu, kteří mění zaměstnání, se nazývá *koeficient relativního rizika* (dále bude značen *RR*). Pro uvedené údaje vypočítáme jeho hodnotu jako

$$RR_1 = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}} = \frac{274/439}{96/182} = 1{,}183.$$

Podíl absolventů dlouhého magisterského programu, kteří nemění zaměstnání, je $n_{12}/n_{1+} = 165/439 = 0{,}38$. Poměr tohoto podílu k podílu absolventů ostatních vysokých škol, kteří neuplatnili vystudovaný obor v současném zaměstnání, je koeficient relativního rizika

$$RR_2 = \frac{n_{12}/n_{1+}}{n_{22}/n_{2+}} = \frac{165/439}{86/182} = 0{,}795.$$

Podíl dvou předchozích relativních rizik se nazývá *poměr šancí*. Může být vyjádřen následujícími vztahy:

$$\psi = \frac{RR_1}{RR_2} = \frac{\dfrac{n_{11}/n_{1+}}{n_{21}/n_{2+}}}{\dfrac{n_{12}/n_{1+}}{n_{22}/n_{2+}}} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Tato statistika nabývá hodnot od nuly do nekonečna, v případě nezávislosti mezi proměnnými nabývá hodnoty jedna. Pro daný příklad spočteme poměr šancí podle vzorce (4.37) jako

$$\psi = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{274 \cdot 86}{165 \cdot 96} = 1,488,$$

nebo jako

$$\psi = \frac{RR_1}{RR_2} = \frac{1,183}{0,795} = 1,488.$$

Výsledná hodnota poměru šancí je větší než jedna, což znamená, že u absolventů dlouhého magisterského studia je větší šance (téměř 1,5krát), že budou měnit zaměstnání (v porovnání s absolventy navazujícího magisterského studia).

IBM SPSS Statistics

Získané výsledky porovnejme s výstupem ze systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme možnost *Risk*. Výsledek zachycuje výstup 4.44. Tento výstup obsahuje výsledný poměr šancí (*Odds Ratio*), relativní rizika RR_1 (*For cohort Uplatnění = ano*) a RR_2 (*For cohort Uplatnění = ne*). Pro všechny tři koeficienty se vypisuje jednak samotná hodnota, tj. bodový odhad (*Value*), jednak 95% interval spolehlivosti (*95% Confidence Interval*) prostřednictvím dolní (*Lower*) a horní (*Upper*) meze tohoto intervalu. Získané hodnoty koeficientů jsou shodné s výsledky stanovenými bez systému *IBM SPSS Statistics*. Protože intervaly spolehlivosti neobsahují hodnotu 1, lze usoudit na závislost změny zaměstnání na typu programu.

Výstup 4.44 | Hodnocení závislosti změny zaměstnání na typu studijního programu pomocí poměru šancí

Risk Estimate	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for typ studijního programu (magisterský dlouhý / magisterský navazující)	1,488	1,049	2,109
For cohort působení v prvním zaměstnání do současné doby = ne	1,183	1,013	1,382
For cohort působení v prvním zaměstnání do současné doby = ano	,795	,654	,967
N of Valid Cases	621		

Příklad 4.24

Budeme analyzovat, zda *na typu studijního programu* (proměnná *A3*) závisí *doba nástupu do zaměstnání* (proměnná *C1*), přičemž nebudeme uvažovat absolventy, kteří dosud nepracují (nastavíme pomocí nabídek *Data* a *Select Cases*, kde zvolíme možnost *If condition is satisfied* a zadáme podmínku *If* ve tvaru „C1 < 3"). Vyjdeme z kontingenční tabulky ve výstupu 4.45.

Výstup 4.45 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.24

Zjištěné absolutní četnosti		Nástup do zaměstnání		Celkem
		Před studiem nebo v průběhu studia	Po absolvování VŠ	
Typ studijního programu	Magisterský dlouhý	193	246	439
	Magisterský navazující	96	86	182
Celkem		289	332	621

Poměr šancí vypočteme s využitím vzorce (4.37) jako

$$\psi = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{193 \cdot 86}{246 \cdot 96} = 0,703 .$$

Na základě výsledné hodnoty můžeme konstatovat, že u absolventů dlouhého magisterského studia byla menší šance nástupu do zaměstnání před studiem nebo v průběhu studia. Kolikrát menší tato šance byla, vyjádříme pomocí převrácené hodnoty poměru šancí, tj. spočteme 1/0,703 = 1,422. Šance je tedy přibližně 1,4krát menší v porovnání s absolventy navazujícího magisterského studia.

IBM SPSS Statistics

V systému *IBM SPSS Statistics* postupujeme stejně jako v příkladu 4.23. Výsledné hodnoty jsou uvedeny ve výstupu 4.46. Protože intervaly spolehlivosti neobsahují hodnotu 1, lze usoudit na závislost změny doby nástupu do zaměstnání na typu programu.

Výstup 4.46 | Hodnocení závislosti doby nástupu do zaměstnání na typu studijního programu pomocí poměru šancí

Risk Estimate	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for typ studijního programu (magisterský dlouhý / magisterský navazující)	,703	,497	,994
For cohort nástup do zaměstnání = před studiem nebo v průběhu studia	,833	,701	,991
For cohort nástup do zaměstnání = po absolvování VŠ	1,186	,996	1,412
N of Valid Cases	621		

♦ ♦ ♦

Kromě významu pro řešení specifického typu úloh má poměr šancí ještě další význam spočívající v tom, že některé jiné míry jsou jeho funkcí. Takovou mírou je především *Yuleho Q* (resp. koeficient γ), viz vzorec (4.35), které můžeme zapsat jako

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = \frac{n_{11}n_{22} / n_{12}n_{21} - 1}{n_{11}n_{22} / n_{12}n_{21} + 1} = \frac{\psi - 1}{\psi + 1}.$$

Obdobně můžeme zapsat *Yuleho koeficient vazby*, viz vzorec (4.36), tj.

$$Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}} = \frac{\sqrt{\psi} - 1}{\sqrt{\psi} + 1}.$$

4.3.7 Tabulka pro tři proměnné (2 dichotomické a 1 vícekategoriální)

Postupy používané při analýze dvourozměrných kontingenčních tabulek lze rozšířit na analýzu tří proměnných. V případě rozdělení čtyřpolní tabulky podle třetí proměnné o L kategoriích bude výsledkem soustava L čtyřpolních tabulek (nebo trojrozměrná kontingenční tabulka $2 \times 2 \times L$). Označme si celkový počet objektů zařazených do l-té tabulky ($l = 1, 2, ..., L$) jako n_l, sdružené četnosti jako n_{lij}, řádkové marginální četnosti jako n_{li+} a sloupcové marginální četnosti jako n_{l+j}. Při platnosti nulové hypotézy o nezávislosti dichotomických proměnných je očekávaná četnost (střední hodnota četnosti) v políčku 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. analogicky s výpočtem výběrového rozptylu jedné proměnné vztahem

$$v'_{lij} = \frac{n_{l1+}n_{l2+}n_{l+1}n_{l+2}}{n_l^2(n_l - 1)}.$$

Nezávislost dvou dichotomických proměnných podmíněnou další kategoriální proměnnou můžeme testovat pomocí Cochranovy nebo Mantelovy-Haenszelovy statistiky. V případě **Cochranovy statistiky** se používá vzorec[7]

7 Cochranova statistika je odmocninou z uvedeného vzorce (4.38) a má asymptoticky normované normální rozdělení. V praxi se používá její druhá mocnina.

$$Q_C = \frac{\left(\sum_{l=1}^{L} n_{l11} - \sum_{l=1}^{L} m_{l11} \right)^2}{\sum_{l=1}^{L} v_{l11}}. \tag{4.38}$$

Tato statistika má asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti.

Pro **Mantelovu-Haenszelovu statistiku** se používá vzorec[8]

$$Q_{MH} = \frac{\left(\sum_{l=1}^{L} n_{l11} - \sum_{l=1}^{L} m_{l11} + 0{,}5 \right)^2}{\sum_{l=1}^{L} v'_{l11}}. \tag{4.39}$$

Z porovnání vzorců (4.38) a (4.39) je zřejmé, že ve vzorci (4.39) je použita korekce na spojitost (přičtení hodnoty 0,5 se provádí v případě, že ve všech prvních políčkách tabulky je zjištěná četnost větší než očekávaná) a úprava při výpočtu rozptylu. Mantelova-Haenszelova statistika má asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti. Rovnají-li se součty zjištěných a očekávaných četností v prvním políčku tabulky, pak se korekce na spojitost nepoužívá, tj. nepřičítá se hodnota 0,5.

Pro stanovení míry asociace lze použít **Mantelův-Haenszelův odhad společného poměru šancí** pro L čtyřpolních tabulek, který je dán vztahem

$$\psi_{MH} = \frac{\sum_{l=1}^{L} \frac{n_{l11} n_{l22}}{n_l}}{\sum_{l=1}^{L} \frac{n_{l12} n_{l21}}{n_l}}. \tag{4.40}$$

Protože tato míra nabývá v případě nezávislosti hodnoty 1, pro testování o nezávislosti se vychází z přirozeného logaritmu vypočtené hodnoty. Příslušný test vyžaduje, aby se charakter závislosti v jednotlivých tabulkách příliš nelišil. Hypotézu homogenity v jednotlivých tabulkách lze testovat pomocí *Breslowy-Dayovy statistiky*, případně pomocí *Taroneho statistiky*. Z důvodu složitosti zápisu zde nebudou příslušné vzorce uvedeny. Breslowa-Dayova statistika je popsána například v [26].

Příklad 4.25

Budeme analyzovat, zda pozice absolventa v zaměstnání (ve smyslu, zda řídí, či neřídí jiné pracovníky) závisí na jeho pohlaví, tj. zda proměnná *D5* závisí na proměnné *E1*. Odpovídající kontingenční tabulka je ve výstupu 4.47.

8 Mantelova-Haenszelova statistika je odmocninou z uvedeného vzorce (4.39) a má asymptoticky normované normální rozdělení. V praxi se používá její druhá mocnina.

Výstup 4.47 | Kontingenční tabulka zjištěných absolutních četností k příkladu 4.24

Zjištěné četnosti		Řízení jiných pracovníků		Celkem
		Řídím	Neřídím	
Pohlaví	Muž	29	71	100
	Žena	13	95	108
Celkem		42	166	208

Pokud bychom analyzovali tyto dvě proměnné samostatně a použili testy založené na chí-kvadrát statistice, zamítneme nulovou hypotézu o nezávislosti na 1% hladně významnosti, viz výstup 4.48 (všechny uvedené statistiky dosahují ve sloupci *Asymptotic Significance* hodnot menších než 0,01). Můžeme usuzovat, že řízení jiných pracovníků a pohlaví jsou znaky závislé.

Výstup 4.48 | Výsledky testů založených na chí-kvadrát statistice k příkladu 4.24

Chi-Square Tests	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	9,271[a]	1	,002
Continuity Correction[b]	8,248	1	,004
Likelihood Ratio	9,426	1	,002
Linear-by-Linear Association	9,227	1	,002
N of Valid Cases	208		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 20,19.
b. Computed only for a 2 × 2 table.

Protože se jedná o závislost jednostrannou, použijeme také poměr šancí, viz upravený výstup 4.49. Výsledný 95% interval spolehlivosti (*95% Confidence Interval*) neobsahuje hodnotu 1, proto rovněž pomocí této statistiky usuzujeme na závislost na 5% hladině významnosti. Poměr šancí je přitom větší než 1, což znamená, že převažují počty četností na hlavní diagonále (muži častěji zastávají řídící funkce než ženy).

Výstup 4.49 | Bodový a intervalový odhad pro poměr šancí zjišťovaných u mužů a žen

Risk Estimate	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for pohlaví (muž/žena)	2,985	1,449	6,150
N of Valid Cases	208		

Budeme-li zkoumat závislost u skupin absolventů, kteří uvádějí odlišný přínos oboru pro vstup do práce (proměnná *B2a* překódovaná do proměnné *B2a_3kat* se 3 kategoriemi), sestavíme tři tabulky, viz výstup 4.50.

Výstup 4.50 | Kontingenční tabulka absolutních četností v členění podle přínosu oboru

Přínos oboru pro vstup do práce			Řízení jiných pracovníků		Celkem
			Řídím	Neřídím	
Žádný nebo malý přínos	Pohlaví	Muž	3	16	19
		Žena	5	20	25
	Celkem		8	36	44
Střední přínos	Pohlaví	Muž	9	15	24
		Žena	3	23	26
	Celkem		12	38	50
Větší nebo velký přínos	Pohlaví	Muž	17	40	57
		Žena	5	52	57
	Celkem		22	92	114

Na základě těchto tabulek můžeme porovnávat tři skupiny výsledků testů založených na chí-kvadrát statistice, viz výstup 4.51 (poznámka a. se vztahuje k celkové tabulce, která byla vynechána, protože je uvedena ve výstupu 4.47). V první skupině (žádný nebo malý přínos) nezamítáme nulovou hypotézu o nezávislosti řízení jiných pracovníků na pohlaví na 5% hladině významnosti. Stejně tak bychom nezamítli hypotézu o nezávislosti testovanou pomocí poměru šancí, neboť výsledný interval spolehlivosti ve výstupu 4.52 obsahuje hodnotu 1. Ve druhé skupině nulovou hypotézu můžeme zamítnout na 5% hladině významnosti (s výjimkou využití opravy na spojitost) a ve třetí skupině na 1% hladině významnosti. (Jednotlivé skupiny se sice liší počtem absolventů a v první skupině není splněn předpoklad pro použití chí-kvadrát testů, nicméně rozdíly ve skupinách jsou zřejmé.)

Výstup 4.51 | Výsledky chí-kvadrát testů v členění podle přínosu oboru

Přínos oboru pro vstup do práce		Value	df	Asymptotic Significance (2-sided)
Žádný nebo malý přínos	Pearson Chi-Square	,129[c]	1	,720
	Continuity Correction[b]	,000	1	1,000
	Likelihood Ratio	,130	1	,718
	Linear-by-Linear Association	,126	1	,723
	N of Valid Cases	44		
Střední přínos	Pearson Chi-Square	4,612[d]	1	,032
	Continuity Correction[b]	3,298	1	,069
	Likelihood Ratio	4,756	1	,029
	Linear-by-Linear Association	4,519	1	,034
	N of Valid Cases	50		
Větší nebo velký přínos	Pearson Chi-Square	8,111[e]	1	,004
	Continuity Correction[b]	6,815	1	,009
	Likelihood Ratio	8,486	1	,004
	Linear-by-Linear Association	8,040	1	,005
	N of Valid Cases	114		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 20,19.
b. Computed only for a 2 × 2 table.
c. 2 cells (50,0%) have expected count less than 5. The minimum expected count is 3,45.
d. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,76.
e. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 11,00.

Výstup 4.52 | Odhady pro poměr šancí v členění podle přínosu oboru

Přínos oboru pro vstup do práce		Value	95% Confidence Interval	
			Lower	Upper
Žádný nebo malý přínos	Odds Ratio for pohlaví (muž/žena)	,750	,155	3,623
	N of Valid Cases	44		
Střední přínos	Odds Ratio for pohlaví (muž/žena)	4,600	1,069	19,799
	N of Valid Cases	50		
Větší nebo velký přínos	Odds Ratio for pohlaví (muž/žena)	4,420	1,503	13,002
	N of Valid Cases	114		

Protože se výsledky testování závislosti změny profese na pohlaví liší podle toho, zda bereme, či nebereme v úvahu typ diplomu, provedeme test podmíněné závislosti pomocí Cochranovy a Mantelovy-Haenszelovy statistiky a spočteme Mantelův-Haenszelův odhad společného poměru šancí. Při postupu bez počítače spočteme nejdříve teoretické četnosti m_{l11} a rozptyly v_{l11}, resp. v'_{l11}, tj.

$$m_{111} = \frac{n_{11+}n_{1+1}}{n_1} = \frac{19 \cdot 8}{44} = 3{,}455\,,$$

$$m_{211} = \frac{n_{21+}n_{2+1}}{n_2} = \frac{24 \cdot 12}{50} = 5{,}76\,,$$

$$m_{311} = \frac{n_{31+}n_{3+1}}{n_3} = \frac{57 \cdot 22}{114} = 11\,,$$

$$v_{111} = \frac{n_{11+}n_{12+}n_{1+1}n_{1+2}}{n_1^3} = \frac{19 \cdot 25 \cdot 8 \cdot 36}{44^3} = 1{,}606\,,$$

$$v_{211} = \frac{n_{21+}n_{22+}n_{2+1}n_{2+2}}{n_2^3} = \frac{24 \cdot 26 \cdot 12 \cdot 38}{50^3} = 2{,}276\,,$$

$$v_{311} = \frac{n_{31+}n_{32+}n_{3+1}n_{3+2}}{n_3^3} = \frac{57 \cdot 57 \cdot 22 \cdot 92}{114^3} = 4{,}439\,,$$

$$v'_{111} = \frac{n_{11+}n_{12+}n_{1+1}n_{1+2}}{n_1^2\left(n_1 - 1\right)} = \frac{19 \cdot 25 \cdot 8 \cdot 36}{44^2(44-1)} = 1{,}643\,,$$

$$v'_{211} = \frac{n_{21+}n_{22+}n_{2+1}n_{2+2}}{n_2^2\left(n_2 - 1\right)} = \frac{24 \cdot 26 \cdot 12 \cdot 38}{50^2(50-1)} = 2{,}323\,,$$

$$v'_{311} = \frac{n_{31+}n_{32+}n_{3+1}n_{3+2}}{n_3^2\left(n_3 - 1\right)} = \frac{57 \cdot 57 \cdot 22 \cdot 92}{114^2(114-1)} = 4{,}478\,.$$

Cochranovu statistiku spočteme podle vzorce (4.38) jako

$$Q_{MH} = \frac{\left(\sum_{l=1}^{3} n_{l11} - \sum_{l=1}^{3} m_{l11}\right)^2}{\sum_{l=1}^{3} v_{l11}} =$$

$$= \frac{((3+9+17) - (3{,}455+5{,}76+11))^2}{(1{,}606+2{,}276+4{,}439)} = 9{,}276\,.$$

Jde o kvantil $\chi^2_{0{,}998}$ [1], to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je 0,002. Na hladině významnosti 1 % proto zamítáme nulovou hypotézu o nezávislosti řízení jiných pracovníků na pohlaví.

Mantelovu-Haenszelovu statistiku spočteme podle modifikace vztahu (4.39), tj.

$$Q_{MH} = \frac{\left(\sum_{l=1}^{3} n_{l11} - \sum_{l=1}^{3} m_{l11} - 0,5\right)^2}{\sum_{l=1}^{3} v'_{l11}} =$$

$$= \frac{((3+9+17)-(3,455+5,76+11)-0,5)^2}{(1,643+2,323+4,478)} = 8,13\,.$$

Jde o kvantil $\chi^2_{0,996}[1]$, to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je 0,004. Stejně jako v případě Cochranovy statistiky tedy na hladině významnosti 1 % zamítáme nulovou hypotézu o nezávislosti řídící funkce na pohlaví.

Mantelův-Haenszelův odhad společného poměru šancí spočteme podle (4.40), tj.

$$\psi_{MH} = \frac{\sum_{l=1}^{3} \frac{n_{l11} n_{l22}}{n_l}}{\sum_{l=1}^{3} \frac{n_{l12} n_{l21}}{n_l}} = \frac{\frac{3 \cdot 20}{44} + \frac{9 \cdot 23}{50} + \frac{17 \cdot 52}{114}}{\frac{16 \cdot 5}{44} + \frac{15 \cdot 3}{50} + \frac{40 \cdot 5}{114}} = 2,964\,.$$

Logaritmus této hodnoty je ln(2,964) = 1,087. Z výsledků usuzujeme, že řídící funkce závisí na pohlaví a ve větší míře zastávají řídící funkci muži.

IBM SPSS Statistics

Získané výsledky nyní porovnejme s výstupem ze systému *IBM SPSS Statistics*. Vybereme proceduru CROSSTABS (*Analyze*, *Descriptive Statistics*, *Crosstabs*) a v části *Statistics* zvolíme položku *Cochran's and Mantel-Haenszel statistics*. Výsledek zachycuje upravený výstup 4.53. Tento výstup zahrnuje tři tabulky.

První z nich se vztahuje k *testům homogenity poměru šancí* (*Tests of Homogeneity of the Odds Ratio*). V prvním řádku jsou výsledky pro Breslowův-Dayovův test, ve druhém pro Taroneho test. Minimální hladina významnosti je v obou případech 0,131, což znamená, že na 5% hladině významnosti nezamítáme hypotézu o shodě poměru šancí v jednotlivých tabulkách, vytvořených podle přínosu oboru pro vstup do práce.

Druhá tabulka obsahuje výsledky *testů podmíněné nezávislosti* (*Tests of Conditional Independence*), to znamená statistiky *Cochran's* a *Mantel-Haenszel*. V obou případech se vypisuje hodnota statistiky (*Chi-Squared*), počet stupňů volnosti (*df*) a minimální hladina významnosti, od které zamítáme nulovou hypotézu (*Asymptotic Significance*). Získané výsledky se shodují s výsledky spočtenými podle výše uvedených vzorců.

Třetí tabulka obsahuje výsledky vztahující se k *odhadu společného poměru šancí* (*Mantel-Haenszel Common Odds Ratio Estimate*). Jsou to bodový odhad poměru šancí (*Estimate*), jeho logaritmus (*ln(Estimate)*) a směrodatná chyba odhadu zlogaritmované hodnoty (*Std. Error of ln(Estimate)*). Dále je v tabulce uvedena minimální hladina

významnosti pro zamítnutí hypotézy o nezávislosti (*Asymptotic Significance*) a 95% oboustranné intervaly spolehlivosti (*Asymptotic 95% Confidence Interval*) pro společný poměr šancí (*Common Odds Ratio*) a pro logaritmus této hodnoty (*ln(Common Odds Ratio)*) pomocí dolních (*Lower Bound*) a horních (*Upper Bound*) mezí. Hodnota společného poměru šancí se shoduje s výsledkem získaným dosazením do příslušného vzorce. Na základě minimální hladiny významnosti lze konstatovat, že na 1% hladině významnosti můžeme zamítnout hypotézu o nulovosti logaritmu společného poměru šancí, tzn. o nezávislosti řízení jiných pracovníků na pohlaví. Dospěli jsme tedy ke stejnému výsledku jako v případě Cochranova a Mantelova-Haenszelova testu.

Výstup 4.53 | Testy homogenity a podmíněné nezávislosti a odhad společného poměru šancí

Tests of Homogeneity of the Odds Ratio	Chi-Squared	df	Asymptotic Significance (2-sided)
Breslow-Day	4,071	2	,131
Tarone's	4,070	2	,131
Tests of Conditional Independence	**Chi-Squared**	**df**	**Asymptotic Significance (2-sided)**
Cochran's	9,276	1	,002
Mantel-Haenszel	8,130	1	,004

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.

Mantel-Haenszel Common Odds Ratio Estimate			
Estimate			2,964
ln(Estimate)			1,087
Standard Error of ln(Estimate)			,368
Asymptotic Significance (2-sided)			,003
Asymptotic 95% Confidence Interval	Common Odds Ratio	Lower Bound	1,440
		Upper Bound	6,102
	ln(Common Odds Ratio)	Lower Bound	,365
		Upper Bound	1,809

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.

Porovnání souborů

5

Při analýze dat je velmi často řešenou úlohou též porovnání skupin hodnot z hlediska jejich vlastností, vyjádřených buď pomocí určité charakteristiky polohy (úrovně), nebo pomocí celkového rozdělení četností hodnot. Obecně bývají skupiny hodnot, příslušejících buď jedné, nebo několika proměnným, označovány pojmem *soubory*. V souvislosti s tímto porovnáváním se rozlišují nezávislé a závislé *výběry*.

Nezávislé výběry jsou takové, kdy sledujeme určitý statistický znak v členění podle skupin respondentů, přičemž tyto skupiny jsou vytvořeny na základě kategorií vysvětlující proměnné (musí se tedy jednat o proměnnou kategoriální). Jsou-li oborem hodnot vysvětlující proměnné dvě kategorie, hovoříme o *dvou nezávislých výběrech*. Pro více než dvě kategorie dostáváme *tři či více nezávislých výběrů*.

Při analýze *závislých výběrů* se porovnávají hodnoty dvou nebo více proměnných, např. buď údaje zjištěné v různých obdobích (názory na produkt před reklamou a po reklamě), nebo hodnocení z různých hledisek (hodnocení přínosu studijního oboru z hlediska vstupu do práce, osobního rozvoje apod.).

Pro *kvantitativní proměnné*, u kterých se předpokládá normální rozdělení, se používají především t testy (podle Studentova t rozdělení). Při nich se testuje shoda středních hodnot ve dvou souborech. Porovnáváme-li dva závislé výběry, hovoříme o *párovém t testu* (spočtou se diference hodnot v jednotlivých párech a testuje se, zda střední hodnota diferencí se rovná hodnotě 0 – úloha se tedy převede na jednovýběrový t test). Skutečným *dvouvýběrovým testem* je tedy pouze test porovnávající střední hodnoty ve dvou nezávislých souborech.

Má-li vysvětlující proměnná více než dvě kategorie, je třeba použít *analýzu rozptylu*, testující shodu středních hodnot ve třech či více souborech. Tato metoda byla již zmíněna v oddílu 4.2 v souvislosti se zkoumáním jednostranné závislosti.

Proměnné získané na základě dotazníkových šetření obvykle nejsou výběrem z normálního rozdělení a často jsou ordinální nebo dichotomické. V takových případech se pro porovnání souborů využívají *neparametrické testy*, viz též [8] a [10].

Pro *dva závislé výběry* se např. testuje, zda je medián diferencí párových hodnot roven nule (úloha se převede na jednovýběrový test). Dále můžeme testovat *shodu podílů* jedné ze dvou hodnot dichotomických proměnných. Pro *nezávislé výběry* testujeme buď shodu mediánů, nebo obecně shodu rozdělení.

V této kapitole budou uvedeny vybrané testy, které jsou zahrnuty v systému *IBM SPSS Statistics*. Budou naznačeny výpočty na základě charakteristik, které jsou součástí výstupů v *IBM SPSS Statistics*. K některým příkladům je připojena kontingenční tabulka, na jejímž základě lze vytvořit vstupní data zadáním kombinací kategorií a jejich četností.

5.1 Testy pro dva závislé výběry

V této části budou vysvětleny vybrané metody, které jsou implementovány v systému *IBM SPSS Statistics*. Pro kvantitativní nebo ordinální proměnné lze použít znaménkový a Wilcoxonův test, pro dichotomické proměnné pak McNemarův test.

A. Znaménkový test

Nulová hypotéza vyjadřuje, že podíl kladných diferencí je shodný s podílem záporných diferencí, alternativní hypotéza vyjadřuje nerovnost těchto podílů. Testovou statistikou je buď počet kladných, nebo počet záporných diferencí. V *IBM SPSS Statistics* je to počet kladných diferencí. Označme si menší hodnotu jako $n_{\min}$, větší jako $n_{\max}$ a celkový počet nenulových diferencí jako n_d. Pro $n_d > 25$ vyjdeme ze vzorce (3.19), resp. (3.20), a testovou statistiku můžeme zapsat pomocí vztahu

$$Z_1 = \frac{n_{\min} - n_d / 2 + 0{,}5}{\sqrt{n_d} / 2} \text{ , nebo } Z_2 = \frac{n_{\max} - n_d / 2 - 0{,}5}{\sqrt{n_d} / 2} \text{ .} \tag{5.1}$$

Za předpokladu platnosti nulové hypotézy má tato statistika asymptoticky normované normální rozdělení.

Pro $n_d < 26$ se používá binomické rozdělení. Minimální hladina významnosti pro zamítnutí nulové hypotézy se spočte na základě vzorce (3.18), tj.

$$\alpha' = 2F(n_{\min}) = 2\sum_{i=0}^{n_{\min}} \binom{n_d}{i} (0{,}5)^{n_d} = 2\left(1 - F(n_{\max} - 1)\right). \tag{5.2}$$

Příklad 5.1

Využijeme proměnné analyzované v příkladu 4.9. Budeme porovnávat nejvyšší dosažená vzdělání otce a matky pro absolventy vysokých škol. K výpočtům využijeme číselné označení kategorií. Kontingenční tabulka absolutních četností je uvedena ve výstupu 4.17, z něhož zjistíme, že počet kladných diferencí je 107 (součet četností nad diagonálou) a počet záporných diferencí 155 (součet četností pod diagonálou). Vypočteme $n_d = 107 + 155 = 262$. Protože $n_d > 25$ a počet kladných diferencí je $n_{\min}$, spočteme hodnotu testové statistiky Z_1, tj.

$$Z_1 = \frac{107 - 262 / 2 + 0{,}5}{\sqrt{262} / 2} = = 2{,}904.$$

Jde o kvantil $u_{0{,}00184}$, což znamená, že $\alpha'/2 = 0{,}00184$. Minimální hladina významnosti, od které zamítáme hypotézu H_0, je $\alpha' = 0{,}00184 \cdot 2 = 0{,}00368$. Nulovou hypotézu o nulovosti mediánu diferencí zamítáme na 1% hladině významnosti.

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests*, *Related Samples*. V listu *Fields* zadáme analyzované proměnné *vzdělání otce* a *vzdělání matky*, v listu *Settings* vybereme *Customize tests* a v části *Compare Median Difference to Hypothesized* zvolíme

Sign test (2 samples). Zobrazí se výstup 5.1, v němž lze otevřít okno *Model Viewer*, které navíc obsahuje výstup 5.2. V něm jsou v rámci grafu uvedeny počty různých typů diferencí, tj. kladné (*Positive Differences*), záporné (*Negative Differences*) a nulové (*Number of Ties*). Hodnotu testové statistiky (–2,904) nalezneme ve výstupu 5.2 (*Standardized Test Statistic*), minimální hladinu významnosti (0,004) ve výstupech 5.1 (*Sig.*) a 5.2 (*Asymptotic Sig. (2-sided test)*).

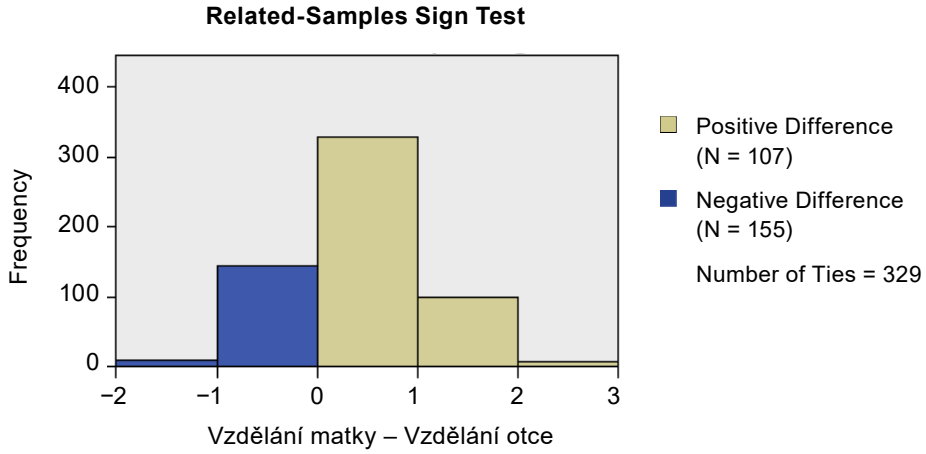
Výstup 5.1 | Základní výstup pro znaménkový test (příklad 5.1)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between vzdělání otce and vzdělání matky equals 0.	Related-Samples Sign Test	,004	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Výstup 5.2 | Volitelný výstup pro znaménkový test (příklad 5.1)



Total N	591
Test Statistic	107,000
Standard Error	8,093
Standardized Test Statistic	–2,904
Asymptotic Sig. (2-sided test)	,004

Příklad 5.2

Výstup 5.3 obsahuje kontingenční tabulku pro proměnné *B2b_3kat* a *B2d_3kat*, z níž lze zjistit, že počet kladných diferencí je 148 a počet záporných diferencí 122. Protože $n_d = 270$, tzn. $n_d > 25$, a počet kladných diferencí je n_{max}, využijeme testovou statistiku Z_2, tj.

$$Z_2 = \frac{148 - 270/2 - 0{,}5}{\sqrt{270}/2} = 1{,}521 .$$

Jde o kvantil $u_{0{,}936}$, výsledkem výrazu $2\,(1 - \Phi(Z_2))$ je hodnota 0,128, což je minimální hladina významnosti pro zamítnutí H_0. Na 5% hladině nezamítáme nulovou hypotézu o nulovosti mediánu diferencí mezi hodnoceními přínosů pro osobní rozvoj a další učení v rámci práce.

Výstup 5.3 | Kontingenční tabulka zjištěných absolutních četností k příkladu 5.2

Zjištěné absolutní četnosti		Přínos oboru pro další učení v rámci práce			Celkem
		Žádný nebo malý přínos	Střední přínos	Větší nebo velký přínos	
Přínos oboru pro osobní rozvoj	Žádný nebo malý přínos	75	51	16	142
	Střední přínos	39	63	81	183
	Větší nebo velký přínos	24	59	227	310
Celkem		138	173	324	635

V *IBM SPSS Statistics* postupujeme stejně jako v příkladu 5.1. Zobrazí se základní výstup 5.4, případně lze nechat zobrazit výstup 5.5. Standardizovaná testová statistika (1,521) a minimální hladina významnosti (0,128) se shodují s výsledky vypočtenými bez použití programu.

Výstup 5.4 | Základní výstup pro znaménkový test (příklad 5.2)

Hypothesis Test Summary

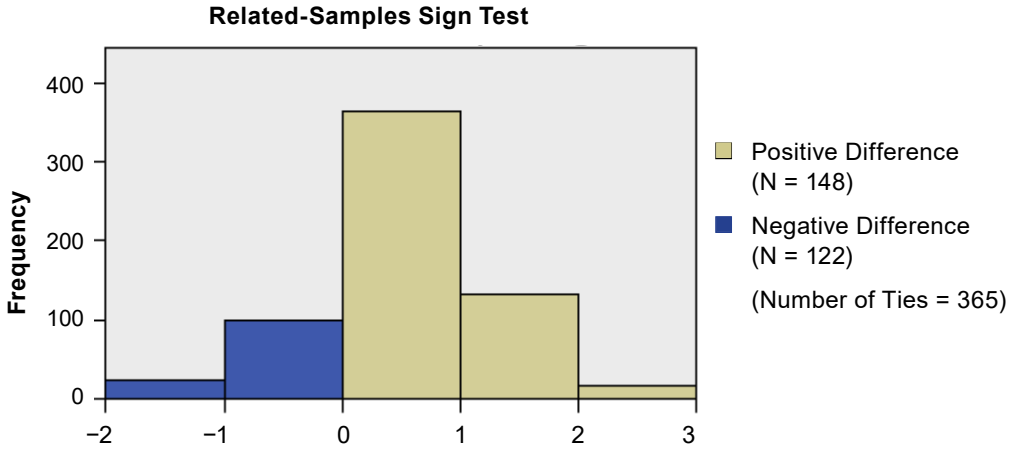
	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between přínos oboru pro osobní rozvoj and přínos oboru pro další učení v rámci práce equals 0.	Related-Samples Sign Test	,128	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Příklad 5.3

Výstup 5.6 obsahuje kontingenční tabulku vytvořenou na základě výběru absolventů, kteří byli v prvním zaměstnání osoby samostatně výdělečně činné. Porovnáváme proměnné zaměřené na přínos vystudovaného oboru, přičemž vycházíme z číselného označení kategorií od 1 do 3. Z tabulky lze zjistit, že počet kladných diferencí je 7 a počet záporných diferencí 12, takže $n_d = 7 + 12 = 19$. Protože $n_d < 26$, aplikujeme exaktní test s využitím binomického rozdělení. Předpokládáme, že při platnosti nulové hypotézy bude pravděpodobnost podílu kladných diferencí 0,5 a počet náhodných pokusů je 19, tj. rozdělení lze označit jako Bi [19; 0,5]. Hodnota distribuční funkce pro počet kladných diferencí n_{min} je $F(7) = 0{,}17964$. Minimální hladina významnosti je pak dvojnásobkem této hodnoty, tj. $\alpha' = 2\,F(7) = 0{,}35928$. Na 5% hladině významnosti proto nezamítáme nulovou hypotézu o shodě podílů kladných a záporných diferencí, tedy o tom, že medián diferencí párových hodnot je roven nule.

Výstup 5.5 | Volitelný výstup pro znaménkový test (příklad 5.2)



Přínos oboru pro další učení v rámci práce – Přínos oboru pro osobní rozvoj

Total N	635
Test Statistic	148,000
Standard Error	8,216
Standardized Test Statistic	1,521
Asymptotic Sig. (2-sided test)	,128

Výstup 5.6 | Kontingenční tabulka zjištěných absolutních četností k příkladu 5.3

Zjištěné absolutní četnosti		Přínos oboru pro osobní rozvoj			Celkem
		Žádný nebo malý přínos	Střední přínos	Větší nebo velký přínos	
Přínos oboru pro další učení v rámci práce	Žádný nebo malý přínos	10	1	2	13
	Střední přínos	7	4	4	15
	Větší nebo velký přínos	1	4	15	20
Celkem		18	9	21	48

V *IBM SPSS Statistics* zadáme test stejným způsobem, jak bylo uvedeno výše. Zobrazí se výstup 5.7, případně lze nechat zobrazit výstup 5.8 s počty nenulových diferencí a shod a s výsledky aplikace normovaného normálního rozdělení. Minimální hladinu významnosti pro exaktní test nalezneme ve výstupech 5.7 (*Sig.*) a 5.8 (*Exact Sig. (2-sided test)*).

Výstup 5.7 | Základní výstup pro znaménkový test (příklad 5.3)

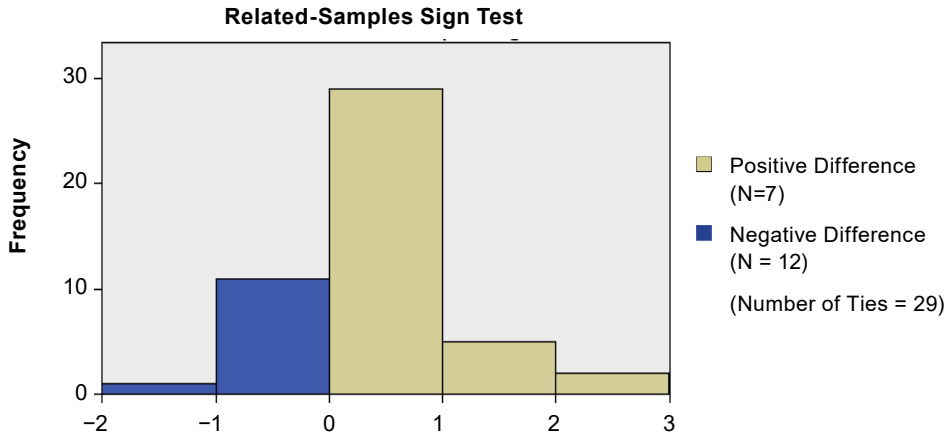
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between přínos oboru pro osobní rozvoj and přínos oboru pro další učení v rámci práce equals 0.	Related-Samples Sign Test	,359[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1] Exact significance is displayed for this test.

Výstup 5.8 | Volitelný výstup pro znaménkový test (příklad 5.3)



Total N	48
Test Statistic	7,000
Standard Error	2,179
Standardized Test Statistic	–,918
Asymptotic Sig. (2-sided test)	,359
Exact Sig. (2-sided test)	,359

1. The exact p-value is computed based on the binominal distribution because there are 25 or fewer cases.

B. McNemarův test

McNemarův test se používá pro dvě dichotomické proměnné, jak bylo popsáno v oddílu 4.3.6, viz příklady 4.18 a 4.19. Pokud proměnné obsahují číselné hodnoty, pak je McNemarův test speciálním případem znaménkového testu. Pokud se tyto číselné hodnoty liší o jedničku (například v případě binárních proměnných, nabývajících hodnot 0 a 1), porovnává se četnost takové kombinace kategorií dvou proměnných, pro které je jejich rozdíl –1, s četností kombinace kategorií, jejichž rozdílem je hodnota 1. Proto musí být výsledky McNemarova testu ve shodě s výsledky získanými pomocí znaménkového testu.

Rozdíl v postupu spočívá v tom, že při znaménkovém testu se používá statistika, která má za předpokladu nulové hypotézy přibližně normované normální rozdělení, zatímco při McNemarově testu je použita její druhá mocnina, která má asymptoticky chí-kvadrát rozdělení s jedním stupněm volnosti. Získaná hladina významnosti je však stejná, viz oddíl 3.4.3.

C. Wilcoxonův párový test

Tento test je rozšířením znaménkového testu. Kladné a záporné diference jsou charakterizovány součty pořadí absolutních hodnot. Z toho důvodu se po výpočtu diferencí každé z nich přiřadí pořadí na základě její absolutní hodnoty. Vyskytují-li se stejné absolutní hodnoty, pak jsou všem diferencím se stejnou absolutní hodnotou přiřazena průměrná pořadí (pořadí se sečtou a dělí počtem diferencí se stejnou hodnotou). Pro výpočet Wilcoxonovy statistiky potřebujeme znát součty pořadí zvlášť pro kladné a zvlášť pro záporné diference. Označme si součet pro kladné diference jako S_{poz} a počet nenulových diferencí opět jako n_d. Pro aproximaci normovaným normálním rozdělením (pro $n_d > 30$) získáme normovanou veličinu tak, že od součtu pro kladné diference odečteme součet, který bychom získali v případě, že by byl medián roven hodnotě 0, a získaný rozdíl dělíme směrodatnou odchylkou. Protože součet pořadí se spočte podle vzorce $n_d\,(n_d + 1) / 2$, je součet hodnot pod mediánem (nebo větších než medián) poloviční hodnotou tohoto celkového součtu. Pro test o nulovosti mediánu diferencí je tedy příslušná statistika daná vztahem

$$Z = \frac{S_{poz} - (n_d(n_d+1)/4)}{\sqrt{n_d(n_d+1)\cdot(2n_d+1)/24}}\,.$$

Rozptyl (hodnota pod odmocninou ve jmenovateli) se ještě redukuje v závislosti na počtu shodných diferencí, takže výsledný vzorec lze zapsat ve tvaru

$$Z = \frac{S_{poz} - \frac{n_d(n_d+1)}{4}}{\sqrt{\frac{n_d(n_d+1)\cdot(2n_d+1)}{24} - \frac{\sum_{i=1}^{n_s}(t_i^3 - t_i)}{48}}}, \tag{5.3}$$

kde t_i je počet shodných pořadí (pro jedno určité pořadí) a n_s je počet variant pořadí, které se vyskytují vícekrát. Tato statistika má za předpokladu, že platí nulová hypotéza o nulovosti mediánu diferencí, přibližně normované normální rozdělení. Minimální hladina významnosti se stanoví obvyklým způsobem, viz příklad 5.4.

Příklad 5.4

Budeme analyzovat stejné proměnné jako v příkladu 5.2 (výstup 5.3), tj. *přínos oboru pro osobní rozvoj* a *přínos oboru pro další učení v rámci práce*, přičemž vycházíme z číselného označení kategorií od 1 do 3. Četnosti všech kombinací jednotlivých kategorií a diference číselných označení kategorií jsou uvedeny v tabulce 5.1. Tabulka 5.2 obsahuje pomocné výpočty potřebné pro součty pro kladné diference (S_{poz}) a pro záporné diference (S_{neg}). Je zřejmé, že se vyskytují pouze diference 1, –1, 2 a –2 s četnostmi 132, 98, 16 a 24. Protože počet nenulových diferencí je 270, můžeme použít aproximaci normovaným normálním rozdělením. Absolutní hodnoty diferencí jsou pouze dvě, a to 1 a 2 s četnostmi 230 a 40. Pokud jedničkám přiřadíme pořadí od hodnoty 1, pak jejich součet bude (230 · 231)/2 = 26 565. Jejich průměrné pořadí je 26 565/230 = 115,5. Celkový počet pořadí pro všech 270 nenulových diferencí je (270 · 271)/2 = 36 585. Součet pořadí pro hodnoty 2 je 36 585 – 26 565 = 10 020. Jejich průměrné pořadí je 10 020/40 = 250,5. Kladným diferencím odpovídá pořadí 115,5 s četností 132 a pořadí 250,5 s četností 16 (celkový součet 19 254), záporným diferencím pořadí 115,5 s četností 98 a pořadí 250,5 s četností 24 (celkový součet 17 331). Pro výpočet testové statistiky je potřeba pouze jedna z těchto hodnot. V souladu s postupem implementovaným v systému *IBM SPSS Statistics* použijeme součet pořadí pro kladné diference, tj. 19 254. Dosadíme do vzorce (5.3), tj.

$$Z = \frac{19\,254 - \frac{270\cdot(270+1)}{4}}{\sqrt{\frac{270\cdot(270+1)\cdot(2\cdot270+1)}{24} - \frac{230^3 - 230 + 40^3 - 40}{48}}} = 0{,}814 \cdot$$

Jde o kvantil $u_{0,7922}$, výsledkem výrazu 2 (1 – 0,7922) je hodnota 0,4156, což je minimální hladina významnosti, od které zamítáme hypotézu H_0. Na 5% hladině významnosti tedy nezamítáme nulovou hypotézu o nulovosti mediánu diferencí.

Tabulka 5.1 | Zadání k příkladu 5.4

B2b_3kat	*B2d_3kat*	Četnost	Diference
1	1	75	0
2	2	63	0
3	3	227	0
1	2	51	1
2	3	81	1
2	1	39	–1
3	2	59	–1
1	3	16	2
3	1	24	–2

Tabulka 5.2 | Pomocné výpočty (s využitím nenulových diferencí) k příkladu 5.4

Označení diference	Hodnota diference	Absolutní diference	Četnost	Pořadí	Výpočet S_{poz}	Výpočet S_{neg}
1	1	1	132	115,5	15 246	–
2	–1	1	98	115,5	–	11 319
3	2	2	16	250,5	4 008	–
4	–2	2	24	250,5	–	6 012
Součet			**270**		**19 254**	**17 331**

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests*, *Related Samples.* V listu *Fields* zadáme analyzované proměnné, v listu *Settings* vybereme *Customize tests* a v části *Compare Median Difference to Hypothesized* zvolíme *Wilcoxon matched-pair signed-rank (2 samples)*.

Zobrazí se výstup 5.9 se zadáním úlohy a výsledkem testu. V okně *Model Viewer* lze získat výstup 5.10, v němž jsou v rámci grafu uvedeny počty různých typů diferencí. Dále je uvedena testová statistika S_{poz} (*Test Statistic*), standardizovaná testová statistika Z (*Standardized Test Statistic*) a minimální hladina významnosti (*Asymptotic Sig. (2-sided test)*). Výsledky získané bez použití programu a s jeho využitím se shodují.

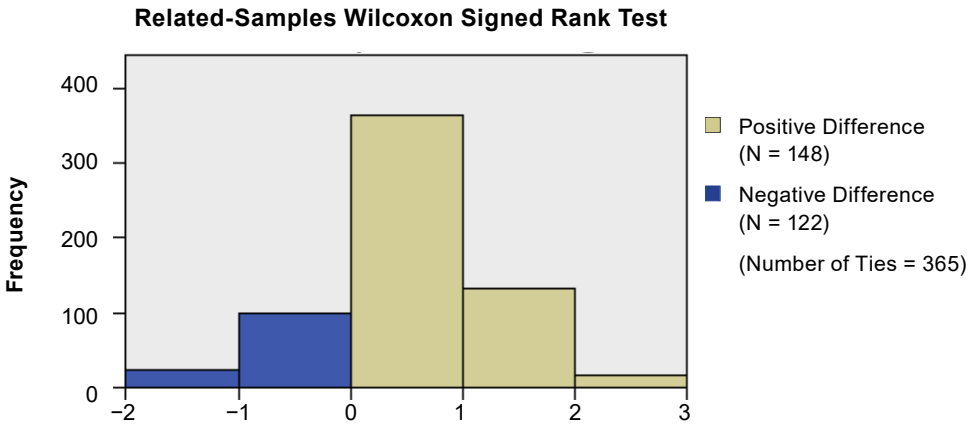
Výstup 5.9 | Základní výstup pro Wilcoxonův párový test (příklad 5.4)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between přínos oboru pro osobní rozvoj and přínos oboru pro další učení v rámci práce equals 0.	Related-Samples Wilcoxon Signed Rank Test	,416	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Výstup 5.10 | Volitelný výstup pro Wilcoxonův párový test (příklad 5.4)



Total N	635
Test Statistic	19 254,000
Standard Error	1 180,918
Standardized Test Statistic	,814
Asymptotic Sig. (2-sided test)	,416

5.2 Testy pro více než dva závislé výběry

V *IBM SPSS Statistics* jsou k dispozici Friedmanův test a testy založené na Kendallově koeficientu konkordance *W* a Cochranovu *Q*.

A. Friedmanův test

K posouzení shody rozdělení pro k proměnných ($k > 2$) slouží *Friedmanův test*, který je založen na průměrech, resp. součtech pořadí zvlášť pro každou proměnnou. Předpokládejme, že porovnáváme odpovědi respondentů na otázky se stejnou bodovou stupnicí hodnocení. Kódům odpovědí získaných od určitého respondenta přiřadíme pořadí k hodnot zjištěných u k sledovaných proměnných. Základní idea je taková, že pokud není rozdíl mezi výběry, pak není rozdíl mezi průměrnými pořadími. Friedmanova statistika F má při platnosti nulové hypotézy přibližně chí-kvadrát rozdělení s $(k - 1)$ stupni volnosti. Je dána vztahem

$$F = \frac{\frac{12}{nk\,(k+1)}\sum_{j=1}^{k} S_j^2 - 3n(k+1)}{1 - \frac{\sum_{i=1}^{n}(t_i^3 - t_i)}{nk\,(k^2-1)}}, \tag{5.4}$$

kde n je rozsah souboru, S_j je součet pořadových čísel pro j-tou proměnnou a t_i je počet proměnných, v nichž jsou zaznamenány stejné odpovědi i-tého respondenta.

Příklad 5.5

Testujme shodu rozdělení při hodnocení přínosů studijního oboru tří typů (proměnné *B2a*, *B2b* a *B2d* na původní pětibodové škále). Při výpočtu bez příslušné procedury bychom vytvořili tři odvozené proměnné s hodnotami vyjadřujícími pořadí hodnot tří původních proměnných (pro každého respondenta), přičemž součet těchto pořadí musí být vždy 6 (1 + 2 + 3). Shodují-li se pořadí u některých proměnných, pak se postupuje stejně jako v případě Wilcoxonova testu. Například trojice hodnot (5; 3; 5) bude ohodnocena pořadími (2,5; 1; 2,5), trojice hodnot (4; 2; 2) pořadími (3; 1,5; 1,5) atd. Hodnoty odvozených proměnných lze v programových systémech zadat pomocí podmíněných výpočtů (pro určitá pořadí se definují určité hodnoty).

Pro jednotlivé proměnné pak sečteme pořadová čísla. Z výstupu 5.12 víme, že průměrná pořadí (*Mean Rank*) pro jednotlivé proměnné jsou 2,02, 2,02 a 1,96. Součet pořadí získáme tak, že průměr vynásobíme počtem respondentů, tj. hodnotou 635 (souhrnný součet druhých mocnin těchto součtů lze vypočítat vynásobením součtu druhých mocnin průměrů druhou mocninou počtu respondentů). Hodnoty t_i, viz vzorec (5.4), jsou buď 2 (hodnoty jsou stejné u dvou proměnných), nebo 3 (hodnoty jsou stejné u všech tří proměnných). První případ se vyskytuje tolikrát, kolikrát se vyskytují pořadí 1,5 nebo 2,5; v analyzovaném souboru je to 358krát. Druhý případ je charakterizován pořadími (2; 2; 2), v analyzovaném souboru je to 184krát.

Hodnotu statistiky F spočteme podle vzorce (5.4) jako

$$F = \frac{\frac{12}{635 \cdot 3 \cdot (3+1)} 635^2 \cdot (2,02^2 + 2,02^2 + 1,96^2) - 3 \cdot 635 \cdot (3+1)}{1 - \frac{358 \cdot (2^3 - 2) + 184 \cdot (3^3 - 3)}{635 \cdot 3 \cdot (3^2 - 1)}} = 2,806 .$$

Jde o kvantil $\chi^2_{0,754}[2]$, to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je 0,246. Na 5% hladině významnosti tedy nezamítáme nulovou hypotézu o shodě rodělení u tří sledovaných proměnných.

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests*, *Related Samples*. V listu *Fields* zadáme analyzované proměnné (musí být definovány jako kvantitativní), v listu *Settings* vybereme *Customize tests* a v části *Compare Distributions* zvolíme *Friedman's 2-way ANOVA by ranks (k samples)*. Zobrazí se výstup 5.11. V okně *Model Viewer* necháme zobrazit výstup 5.12. V něm nalezneme u sloupcových grafů pro jednotlivé proměnné (vyjadřují četnosti jednotlivých pořadových čísel) průměrná pořadí (*Mean Rank*). Z tabulky přečteme počet pozorování (*Total N*), hodnotu testové statistiky (*Test Statistic*), počet stupňů volnosti (*Degrees of Freedom*) a minimální hladinu významnosti (*Asymptotic Sig. (2-sided test)*). Hodnota testové statistiky (2,806), a tedy i minimální hladina významnosti (0,246), se shodují s výsledky získanými výpočtem na základě vzorce a pomocí hodnoty distribuční funkce.

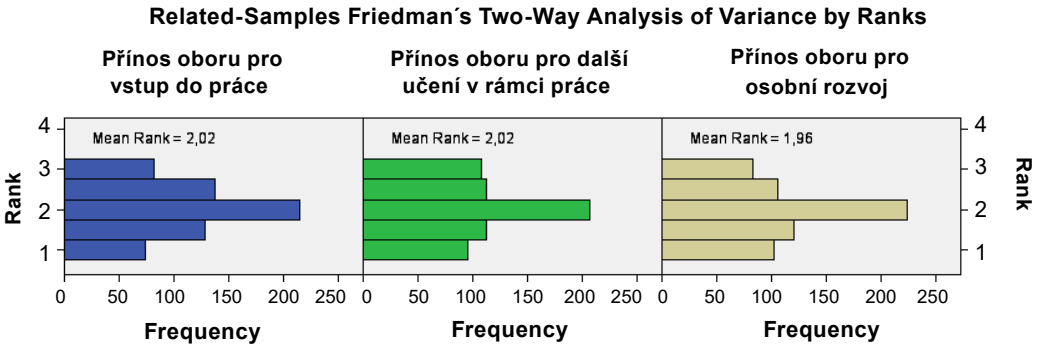
Výstup 5.11 | Základní výstup pro Friedmanův test (příklad 5.5)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of přínos oboru pro vstup do práce, přínos oboru pro další učení v rámci práce and přínos oboru pro osobní rozvoj are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	,246	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Výstup 5.12 | Volitelný výstup pro Friedmanův test (příklad 5.5)



Total N	635
Test Statistic	2,806
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	,246

1. Multiple comparisons are not performed because the overall test retained the null hypothesis of no differences.

B. Kendallovo W

Kendallův koeficient konkordance W je mírou souladu mezi pořadími získanými pro *k* sledovaných proměnných. Nabývá hodnot z intervalu od 0 do 1, přičemž dolní meze nabývá v případě nezávislosti. Test o nulovosti tohoto koeficientu je ekvivalentem Friedmanova testu. Výsledek testu je tedy stejný a navíc má Kendal-lovo *W* výše zmíněný význam pro určení míry souladu. Pro výpočet Kendallova *W* je potřeba nejprve znát hodnotu Friedmanovy statistiky *F*. Koeficient konkordance se spočte podle vzorce

$$W = \frac{F}{n\,(k-1)}\,. \tag{5.5}$$

Test o nulovosti Kendallova *W* se provádí pomocí statistiky $n\,(k-1)\,W$. Za platnosti nulové hypotézy má tato statistika chí-kvadrát rozdělení s $(k-1)$ stupni volnosti. Výsledná statistika je vlastně Friedmanovou statistikou *F*.

Příklad 5.6

Budeme analyzovat stejná data jako v předchozím příkladu 5.5. Testujme shodu rozdělení při hodnocení přínosů studijního oboru tří typů (proměnné *B2a*, *B2b* a *B2d*). Na základě znalosti hodnoty Friedmanovy statistiky *F* můžeme spočítat hodnotu

$$W = \frac{2{,}806}{635 \cdot (3-1)} = 0{,}002\,209\,.$$

Statistiku potřebnou pro provedení testu spočteme jako

$$n\,(k-1)\,W = 635\;\;(3-1)\cdot 0{,}002\,209 = 2{,}806.$$

Výsledek testu je shodný s výsledkem testu o shodě mediánů pomocí Friedmanova F.

IBM SPSS Statistics

Postupujeme stejně jako v příkladu 5.6 s tím rozdílem, že v listu *Settings* v části *Quantify Associations* zvolíme *Kendall's coefficient of concordance (k samples)*. Zobrazí se výstup 5.13; pomocí něj lze otevřít okno *Model Viewer*, které navíc obsahuje výstup 5.14. Ten se od předchozího výstupu 5.12 pro Friedmanův test liší pouze v tom, že obsahuje navíc hodnotu Kendallova W (0,002), která se shoduje s hodnotou vypočtenou podle vzorce (5.5). Výstup 5.13 se liší od výstupu 5.11 pouze názvem testu.

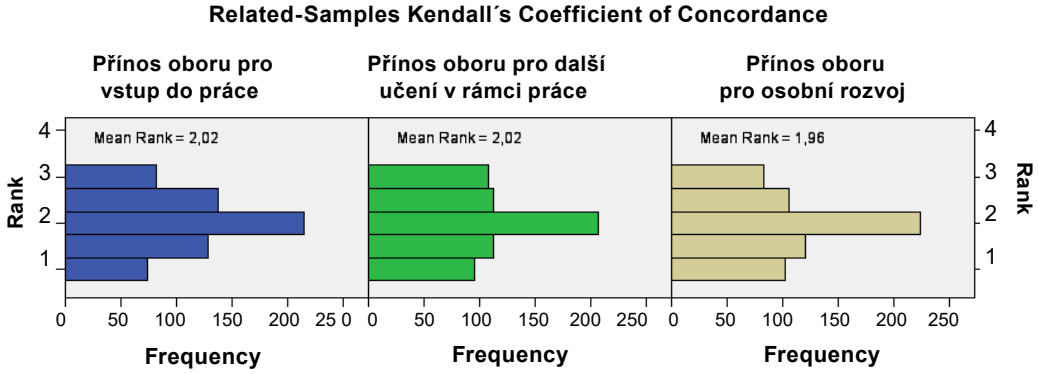
Výstup 5.13 | Základní výstup pro Kendallovo W (příklad 5.6)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of přínos oboru pro vstup do práce, přínos oboru pro další učení v rámci práce and přínos oboru pro osobní rozvoj are the same.	Related-Samples Kendall's Coefficient of Concordance	,246	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Výstup 5.14 | Volitelný výstup pro Kendallovo W (příklad 5.6)



Total N	635
Kendall's W	,002
Test Statistic	2,806
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	,246

1. Multiple comparisons are not performed because the owerall test retained the null hypothesis of no differences.

C. Cochranovo *Q*

Pomocí Cochranovy statistiky Q se testuje, zda pro k dichotomických proměnných je shodný podíl jedné ze dvou kategorií. Statistika je dána vztahem

$$Q = \frac{(k-1)\left(k\sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2\right)}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2}, \tag{5.6}$$

kde G_j je počet jedné ze dvou hodnot zjištěný u j-té proměnné a L_i je počet výskytů této hodnoty u i-tého objektu. Cochranovo Q má při platnosti nulové hypotézy o shodě rozdělení četností chí-kvadrát rozdělení s $(k-1)$ stupni volnosti.

Příklad 5.7

Zaměříme se na proměnné vyjadřující různé typy přínosů překódované do proměnných se dvěma kategoriemi (*B2a_2kat, B2b_2kat* a *B2d_2kat*). Budeme zjišťovat, zda je u nich stejné relativní zastoupení odpovědi *ne* (hodnoty 0). Pro výpočet potřebujeme znát zjištěné četnosti této hodnoty u jednotlivých proměnných. Přibližně lze vyčíst z výstupu 5.16, přesné četnosti jsou 140, 138 a 142 (viz výstupy 3.9, 3.10 a 3.12). Celkový počet nul (součet hodnot L_i) je tedy součtem těchto četností, výsledkem je hodnota 420. V programových systémech lze vypočítat odvozenou proměnnou obsahující počet odpovědí *ne* u jednotlivých respondentů. Další odvozenou proměnnou lze definovat jako druhé mocniny těchto četností. Na základě analyzovaných dat bychom jako součet druhých mocnin četností obdrželi hodnotu 890. Podle vzorce (5.6) spočteme

$$Q = \frac{(3-1)\cdot(3\cdot(140^2+138^2+142^2)-(140+138+142)^2)}{3\cdot 420-890} = 0{,}13\,.$$

Jde o kvantil $\chi^2_{0{,}063}[2]$, to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je 0,937. Na 5% hladině významnosti tedy nezamítáme nulovou hypotézu o shodě rozdělení četností u tří sledovaných proměnných.

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests* a *Related Samples*. V listu *Fields* zadáme analyzované proměnné (musí být definovány jako nominální), v listu *Settings* vybereme *Customize tests* a v části *Test for Change in Binary Data* zvolíme *Cochran's Q (k samples)*. Výsledkem jsou výstupy 5.15 a 5.16. První obsahuje zadání úlohy, minimální hladinu významnosti (*Sig.*) a vyjádření k testované hypotéze (*Decision*). Druhý výstup zahrnuje sloupcový graf rozdělení četností pro jednotlivé proměnné a tabulku obsahující počet pozorování (*Total N*), testovou statistiku (*Test Statistic*), počet stupňů volnosti (*Degrees of Freedom*) a minimální hladinu významnosti (*Asymptotic Sig. (2-sided test)*).

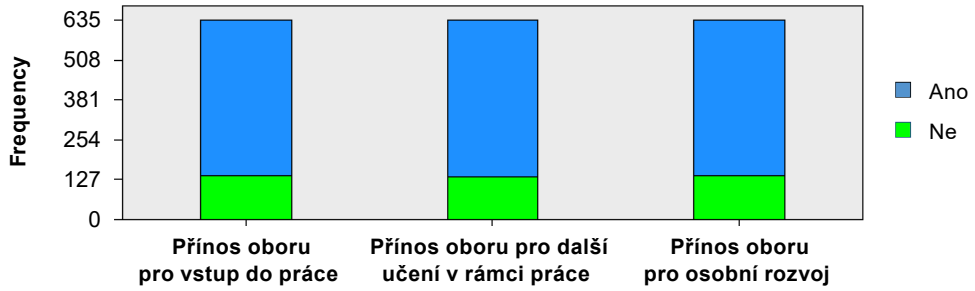
Výstup 5.15 | Základní výstup pro Cochranovo *Q* (příklad 5.7)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of přínos oboru pro vstup do práce, přínos oboru pro další učení v rámci práce and přínos oboru pro osobní rozvoj are the same for the specified categories.	Related-Samples Cochran's Q Test	,937	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Výstup 5.16 | Volitelný výstup pro Cochranovo Q (příklad 5.7)



Total N	635
Test Statistic	,130
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	,937

1. Multiple comparisons are not performed because the owerall test does not show significant differences across samples.

5.3 Testy pro dva nezávislé výběry

Pro testování shody rozdělení pro dva nezávislé výběry je v *IBM SPSS Statistics* několik testů. V tomto oddílu se z důvodu složitosti ostatních metod zaměříme pouze na **Mannův-Whitneyho test**, který je založen na Wilcoxonově statistice W, viz např. [3]. Ta se získá tak, že je každé hodnotě přiřazeno pořadí v rámci všech hodnot (oba výběry jsou spojeny do jednoho celku) a pořadová čísla se sečtou v rámci jednotlivých výběrů. Součet pro první výběr (daný kategorií vysvětlované proměnné pro první objekt v datovém souboru) o rozsahu n_1 označme S_1, součet pro druhý výběr o rozsahu n_2 označme S_2). *Wilcoxonovou statistikou W* je součet S_1. Mannova-Whitneyho statistika U se spočte podle vzorce

$$U = W - \frac{n_1(n_1+1)}{2}. \tag{5.7}$$

Kritické hodnoty se při malém rozsahu výběrů určují pomocí tabulek nebo simulační metodou. Je-li $n_1 n_2 > 400$ a $n_1 n_2/2 + \min\{n_1, n_2\} > 220$, lze provést aproximaci normálním rozdělením tak, že spočteme statistiku

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{n\,(n-1)}\left(\frac{n^3 - n}{12} - \frac{\sum_{j=1}^{n_s}(t_j^3 - t_j)}{12}\right)}}, \tag{5.8}$$

kde t_j je počet shodných pořadí (pro jedno určité pořadí) v obou výběrech a n_s je počet variant pořadí, které se vyskytují v obou výběrech. Tato statistika má za předpokladu, že platí nulová hypotéza o shodě rozdělení, přibližně normální rozdělení.

Příklad 5.8

Budeme zjišťovat, zda se hodnocení *přínosu studijního oboru pro osobní rozvoj* (*B2d*) liší v závislosti na *pohlaví* (*E1*). V programových systémech lze vytvořit odvozenou proměnnou obsahující pořadí hodnot původní proměnné. Protože se v analyzovaném souboru hodnoty opakují, získáme pouze pět různých pořadí, viz tabulka 5.3, sloupec *hodnota pořadí*. V této tabulce jsou dále uvedeny četnosti těchto pořadových čísel zvlášť pro muže a pro ženy (protože v datovém souboru je prvním respondentem žena, bude první skupina pro ženy a druhá pro muže). Na jejich základě vypočteme hodnoty S_1 a S_2. Hodnota Wilcoxonovy statistiky je $W = S_1 = 106\ 054$. Mannovu-Whitneyho statistiku U spočteme podle vzorce (5.7) jako

$$U = 106\,054 - \frac{367 \cdot 368}{2} = 38\,526.$$

Protože $n_1 n_2 = 91\ 707$ a $n_1 n_2/2 + \min\{n_1, n_2\} = 46\ 084{,}5$, lze provést aproximaci normálním rozdělením. K tomu potřebujeme znát četnosti pro shodná pořadová čísla pro muže a pro ženy. Získáme je z tabulky 5.3 (sloupec *četnost celkem*).

Tabulka 5.3 | Pomocné výpočty k příkladu 5.8

Původní hodnota	Hodnota pořadí	Četnost pro muže	Četnost pro ženy	Četnost celkem	Výpočet S_2	Výpočet S_1
1	12,5	3	21	24	37,5	262,5
2	78,5	42	66	108	3 297	5 181
3	221,5	63	115	178	13 954,5	25 472,5
4	410,0	86	113	199	35 260	46 330
5	554,0	37	52	89	20 498	28 808
Součet		**231**	**367**	**598**	**73 047**	**106 054**

Normováním statistiky U (odečtením střední hodnoty a dělením rozdílu směrodatnou odchylkou) se zohledněním počtu shodných pořadí získáme statistiku

$$Z = \frac{38\,526 - \frac{231 \cdot 367}{2}}{\sqrt{\frac{231 \cdot 367}{598 \cdot (598-1)} \cdot \left(\frac{598^3 - 598}{12} - \frac{24^3 - 24 + 108^3 - 108 + 178^3 - 178 + 199^3 - 199 + 89^3 - 89)}{12} \right)}} =$$

$$= -1{,}95\,.$$

Jde o kvantil $u_{0{,}0256}$, to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je $2 \cdot 0{,}0256 = 0{,}0512$. Na 5% hladině významnosti nezamítáme hypotézu vyjadřující, že v souborech vytvořených na základě pohlaví jsou shodná rozdělení proměnné *přínos studijního oboru pro osobní rozvoj,* a usuzujeme, že daná proměnná nezávisí na pohlaví.

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests*, *Independent Samples*. V listu *Fields* zadáme analyzované proměnné (proměnnou *B2d*, do *Test Fields* a proměnnou *E1* do *Groups*), v listu *Settings* vybereme *Customize tests* a v části *Compare Distributions across Groups* zvolíme *Mann-Whitney U (2 samples)*. Zobrazí se výstup 5.17 se zadáním úlohy a s výsledkem testu. Otevřením okna *Model Viewer* získáme výstup 5.18. V něm jsou v rámci grafu (rozdělení četností hodnot analyzované proměnné podle pohlaví) uvedena průměrná pořadí (*Mean Rank*) pro obě skupiny vytvořené podle pohlaví (tj. pro muže a ženy). Zobrazená tabulka obsahuje například Wilcoxonovu statistiku W (*Wilcoxon W*), Mannovu-Whitneyho statistiku U (*Mann-Whitney U*), standardizovanou testovou statistiku (*Standardized Test Statistic*) a minimální hladinu významnosti (*Asymptotic Sig. (2-sided test)*). Hodnoty standardizované statistiky Z, získané pomocí příslušné procedury v *IBM SPSS Statistics* a bez této procedury, se liší na třetím desetinném místě.

Výstup 5.17 | Základní výstup pro Mannův-Whitneyho test (příklad 5.8)

Hypothesis Test Summary

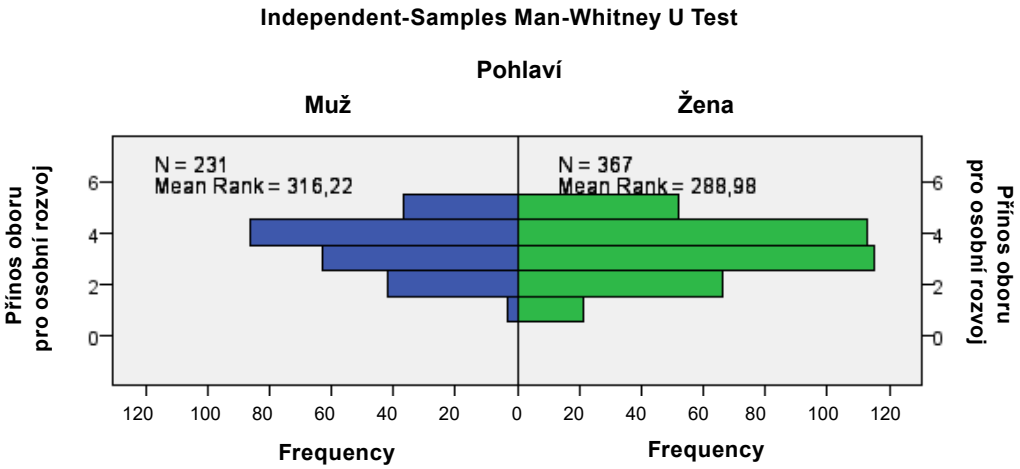
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of přínos oboru pro osobní rozvoj is the same across categories of pohlaví.	Independent-Samples Mann-Whitney U Test	,051	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Poznámka

Chceme-li v *IBM SPSS Statistics* vytvořit novou proměnnou s pořadími hodnot, je potřeba vybrat absolventy, kteří při šetření uvedli své pohlaví (hodnoty proměnné *E1* jsou různé od hodnoty 0).

Výstup 5.18 | Volitelný výstup pro Mannův-Whitneyho test (příklad 5.8)



Total N	598
Mann-Whitney U	38 526,000
Wilcoxon W	106 054,000
Test Statistic	38 526,000
Standard Error	1 981,185
Standardized Test Statistic	–1,950
Asymptotic Sig. (2-sided test)	,051

5.4 Testy pro více než dva nezávislé výběry

Nabývá-li proměnná *X*, jejíž hodnoty určují rozdělení proměnné *Y* do skupin, více než dvou kategorií, získáme více než dva nezávislé výběry. V tomto oddílu uvedeme dva testy implementované v systému *IBM SPSS Statistics*, a to Kruskalův-Wallisův a mediánový test.

A. Kruskalovo-Wallisovo *H*

Při *Kruskalově-Wallisově testu* nulová hypotéza předpokládá, že ve všech souborech jsou shodná rozdělení hodnot vysvětlované proměnné. Alternativní hypotéza říká, že alespoň jedno rozdělení se liší od ostatních. Tento test byl již popsán v oddílu 4.3.3. Nyní budeme uvažovat vysvětlovanou proměnnou *Y* obecnou, která může mít velký počet kategorií.

Výpočet testové statistiky je obdobně jako u testu pro dva výběry založen na pořadových číslech, která jsou přiřazena hodnotám v souboru o rozsahu *n*, vzniklým spojením všech výběrů. Pro každý výběr o rozsahu n_i pak vypočteme průměrné pořadí $\overline{R}_i$, kde $i = 1, 2, \ldots, K$, přičemž *K* je počet výběrů, tj. počet kategorií proměnné *X*. Kruskalova-Wallisova statistika je dána vztahem

$$KW = \frac{12}{n\,(n+1)} \sum_{i=1}^{K} n_i \overline{R}_i^2 - 3\,(n+1)\,. \tag{5.10}$$

Tato veličina má při platnosti nulové hypotézy o shodě rozdělení přibližně rozdělení chí-kvadrát s $(K - 1)$ stupni volnosti. Uvedený vzorec pro testové kritérium je platný pouze za předpokladu, že se všechny hodnoty proměnné *Y* liší. Programové systémy proto používají vzorec (4.21) s opravou na spojitost. Obecně lze pak vzorec pro Kruskalovu-Wallisovu statistiku zapsat ve tvaru

$$KW^* = \frac{\dfrac{12}{n\,(n+1)} \displaystyle\sum_{i=1}^{K} n_i \overline{R}_i^2 - 3\,(n+1)}{1 - \dfrac{\displaystyle\sum_{j=1}^{n_s} (t_j^3 - t_j)}{n^3 - n}}\,, \tag{5.11}$$

kde t_j je počet shodných pořadí (pro jedno určité pořadí) v různých výběrech (v kontingenční tabulce se sloupcovou vysvětlovanou proměnnou to jsou sloupcové marginální četnosti n_{+j}) a n_s je počet variant pořadí, které se ve výběrech vyskytují (počet kategorií vysvětlované proměnné).

Příklad 5.9

Provedeme obdobný test jako v příkladu 4.11. Budeme zjišťovat, zda *spokojenost se současnou prací* (*D6*) závisí na *typu smlouvy v současném zaměstnání* (*D2*). Tentokrát ale budeme analyzovat původní pětihodnotovou proměnnou *D6*. Ve třech skupinách (daných

kategoriemi proměnné *D2* s četnostmi 470, 26 a 47) spočteme průměrná pořadí, která jsou 268,0997, 292,519 a 299,66. Dále potřebujeme zjistit počet shodných pořadí pro jednotlivé varianty pořadí. Ve výběrech se vyskytuje pět variant pořadí, přičemž jejich četnosti jsou 4, 29, 115, 257 a 138. Jejich dosazením do jmenovatele vzorce (5.11) dostaneme hodnotu 0,868. Kruskalovu-Wallisovu statistiku pak spočteme podle vzorce (5.11) jako

$$KW = \frac{\frac{12}{543 \cdot (543+1)}(470 \cdot 268{,}099^2 + 26 \cdot 292{,}519^2 + 47 \cdot 299{,}66^2) - 3 \cdot (543+1)}{0{,}868} = 2{,}53.$$

Jde o kvantil $\chi^2_{0{,}718}[2]$, to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je 0,282. Na 5% hladině významnosti tedy nezamítáme nulovou hypotézu o shodě rozdělení ve třech skupinách vytvořených na základě typu smlouvy v současném zaměstnání. Nezjistili jsme tedy, že by spokojenost se současnou prací závisela na typu smlouvy v současném zaměstnání.

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests*, *Independent Samples*. V listu *Fields* zadáme analyzované proměnné (proměnnou *D6* do *Test Fields*, proměnnou *D2* do *Groups*). V listu *Settings* vybereme *Customize tests* a v části *Compare Distributions across Groups* zvolíme *Kruskal-Wallis 1-way ANOVA (k-samples)*. Zobrazí se výstup 5.19, v okně *Model Viewer* lze získat výstup 5.20. V něm jsou uvedeny jednak krabičkový graf pro porovnání úrovně a variability hodnot, jednak tabulka obsahující hodnotu testové statistiky (2,53) a minimální hladinu významnosti (0,282).

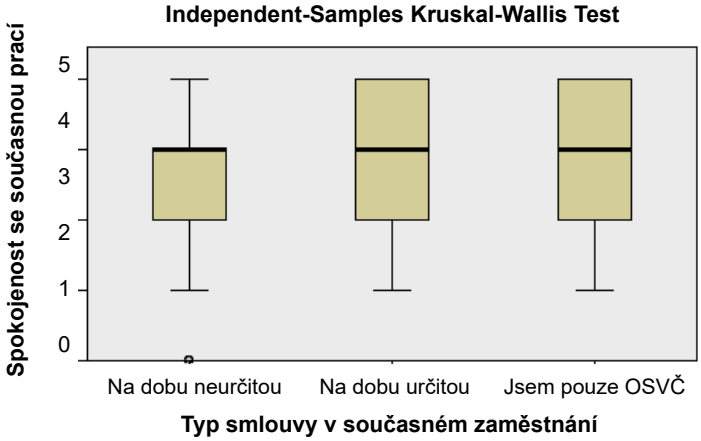
Výstup 5.19 | Základní výstup pro Kruskalův-Wallisův test (příklad 5.9)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of spokojenost se současnou prací is the same across categories of typ smlouvy v současném zaměstnání.	Independent-Samples Kruskal-Wallis Test	,282	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Výstup 5.20 | Volitelný výstup pro Kruskalův-Wallisův test (příklad 5.9)



Total N	543
Test Statistic	2,530
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	,282

1. The test statistic is adjusted for ties.
2. Multiple comparisons are not performed because the overall test does not show significant differences across samples.

B. Mediánový test pro K výběrů

Při mediánovém testu nevytváříme pořadí, ale zjistíme *medián* ze všech hodnot vysvětlované proměnné (postup viz oddíl 3.2.2). V každém výběru pak zjišťujeme, kolik hodnot je menších nebo rovno tomuto mediánu a kolik hodnot je větších. Výsledné počty můžeme znázornit v kontingenční tabulce, viz schéma 5.1.

Schéma 5.1 | Kontingenční tabulka absolutních četností pro mediánový test

	1. výběr	…	*j*-tý výběr	…	*K*-tý výběr	Celkem
Větší než medián	n_{11}		n_{1j}		n_{1K}	R_1
Menší než medián nebo rovno mediánu	n_{21}		n_{2j}		n_{2K}	R_2
Celkem	n_1		n_j		n_K	n

V případě platnosti nulové hypotézy o shodě mediánů by sdružené četnosti měly být úměrné marginálním četnostem tak, jak bylo popsáno ve čtvrté kapitole. Označme četnosti očekávané v případě nezávislosti shodně s předchozí kapitolou jako m_{ij}, přičemž platí, že $m_{ij} = R_i n_j / n$. Potřebnou statistiku chí-kvadrát spočteme podle vzorce

$$\chi^2_P = \sum_{i=1}^{2}\sum_{j=1}^{K}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}\,. \tag{5.12}$$

Jak je známo z předchozí kapitoly, tato statistika má za předpokladu nulové hypotézy asymptoticky chí-kvadrát rozdělení s $(K-1)$ stupni volnosti. Mělo by však platit, že žádná hodnota očekávané četnosti není menší než jedna a počet hodnot očekávaných četností, které jsou menší než pět, nepřesáhne 20 % z počtu políček tabulky.

Příklad 5.10

Budeme analyzovat stejná data jako v příkladu 5.9. Mediánem hodnot proměnné *D6* je hodnota 4. Podle schématu 5.1 můžeme vytvořit kontingenční tabulku, viz výstup 5.21. K výpočtu potřebujeme znát očekávané četnosti, viz výstup 5.22. Na základě zjištěných a očekávaných četností můžeme spočítat chí-kvadrát statistiku

$$\chi^2_P = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(n_{ij}-m_{ij})^2}{m_{ij}} = \frac{(111-119,4)^2}{119,4} + \frac{(9-6,6)^2}{6,6} + \ldots + \frac{(29-35,1)^2}{35,1} = 6,078\,.$$

Jde o kvantil $\chi^2_{0,952}[2]$, to znamená, že minimální hladina významnosti, od které zamítáme nulovou hypotézu, je 0,048. Na 5% hladině významnosti tedy zamítáme nulovou hypotézu o shodě mediánů ve třech skupinách vytvořených na základě typu smlouvy v současném zaměstnání a usuzujeme, že spokojenost se současnou prací závisí na typu smlouvy v současném zaměstnání. V porovnání s nezamítnutou nulovou hypotézou při použití Kruskalova-Wallisova testu jsme tedy dospěli k odlišnému závěru.

Výstup 5.21 | Kontingenční tabulka pro mediánový test k příkladu 5.10

		Typ smlouvy v současném zaměstnání		
		Na dobu neurčitou	Na dobu určitou	Jsem pouze OSVČ
Spokojenost se současnou prací	> medián	111	9	18
	<= medián	359	17	29

Výstup 5.22 | Očekávané četnosti pro mediánový test k příkladu 5.10

		Typ smlouvy v současném zaměstnání			Celkem
		Na dobu neurčitou	Na dobu určitou	Jsem pouze OSVČ	
Spokojenost se současnou prací	> medián	119,4	6,6	11,9	138,0
	<= medián	350,6	19,4	35,1	405,0
Celkem		470,0	26,0	47,0	543,0

IBM SPSS Statistics

V *IBM SPSS Statistics* zvolíme *Analyze*, *Nonparametric Tests*, *Independent Samples*. V listu *Fields* zadáme analyzované proměnné (proměnnou *D6* do *Test Fields*, proměnnou *D2* do *Groups*). V listu *Settings* vybereme *Customize tests* a v části *Compare Medians across Groups* zvolíme *Median test (k-samples)*. Zobrazí se výstup 5.23, v okně *Model Viewer* získáme výstup 5.24. V něm je uveden jednak krabičkový graf pro porovnání úrovně a variability hodnot, jednak tabulka obsahující hodnotu testové statistiky (6,078) a minimální hladinu významnosti (0,048).

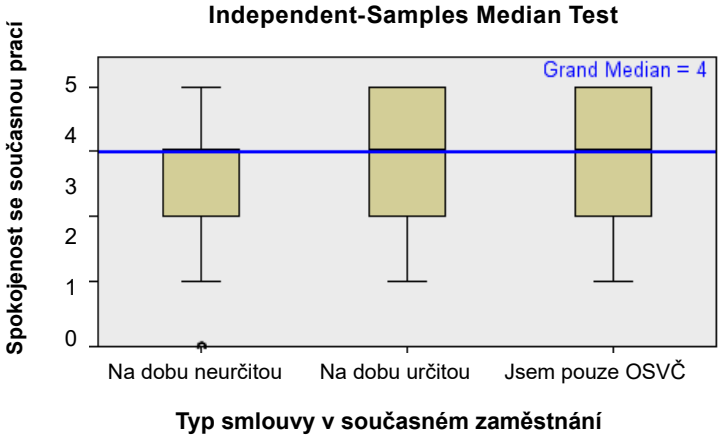
Výstup 5.23 | Základní výstup pro mediánový test (příklad 5.10)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The medians of spokojenost se současnou prací are the same across categories of typ smlouvy v současném zaměstnání.	Independent-Samples Median Test	,048	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Výstup 5.24 | Volitelný výstup pro mediánový test (příklad 5.10)



Total N	543
Median	4,000
Test Statistic	6,078
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	,048

Modely jednostranné závislosti

6

V předchozích dvou kapitolách byly probrány metody, které slouží ke zjištění závislosti především mezi dvěma proměnnými. V praktických aplikacích se však často navrhují a odhadují modely, v nichž je vyjadřována závislost určité (vysvětlované) proměnné na několika jiných (vysvětlujících) proměnných. Je-li vysvětlovaná proměnná kvantitativní spojitá, využívá se nejčastěji vícenásobná lineární regresní analýza. Na základě modelu pak lze při znalosti hodnot kvantitativních nebo dichotomických (binárních) vysvětlujících proměnných předpovědět neznámou hodnotu vysvětlované proměnné.

Při analýze dat získaných z dotazníkových šetření je situace poněkud odlišná. Obvykle je vysvětlovaná proměnná nominální nebo ordinální. Nejjednodušší je modelování v případě dichotomické vysvětlované proměnné. Při znalosti hodnot vysvětlujících proměnných se pak objekt (respondent) zařazuje do určité skupiny (klasifikuje), přičemž počet skupin je dán počtem kategorií vysvětlované proměnné. Pro tyto situace byly navrženy různé přístupy.

Také vysvětlující proměnné jsou obvykle kategoriální. Základním krokem při navržení modelu by měl být výběr těchto proměnných. Pokud by vysvětlovaná proměnná byla kvantitativní spojitá, mohli bychom pro výběr vysvětlujících proměnných využít analýzu rozptylu (a zjistit, na kterých kategoriálních proměnných je vysvětlovaná proměnná závislá).

Pro kategoriální vysvětlovanou proměnnou je potřeba aplikovat modifikovaný postup. Jednou z možností je výběr vysvětlujících proměnných na základě meziskupinové variability, která se stanoví tak, že se od celkové variability vysvětlované proměnné odečítá vnitroskupinová variabilita, kdy skupiny jsou vytvořeny podle kategorií vysvětlující proměnné. Pokud je variabilita vyjadřována pomocí entropie, viz vzorec (3.3), bývá meziskupinová variabilita označována jako *informační zisk*. Kromě entropie lze použít Giniho koeficient (nominální rozptyl), viz vzorec (3.2).

Protože výklad metod využívaných ke klasifikaci je již poněkud složitější, než byl výklad uvedený v předchozích kapitolách, budou v této knize pouze naznačeny principy vybraných postupů. Pro zájemce bude uvedena literatura, kde lze nalézt podrobnější vysvětlení a odvození jednotlivých postupů a vzorců. V této kapitole budou nejprve naznačeny možnosti klasifikačních stromů, poté bude následovat ukázka aplikací logistické regresní analýzy – vše pomocí *IBM SPSS Statistics*.

6.1 Klasifikační stromy

Klasifikační stromy patří do skupiny metod, které zahrnují alternativní postupy k diskriminační a regresní analýze a označují se jako *rozhodovací stromy*. Pro odhad hodnoty kvantitativní vysvětlované proměnné lze použít *regresní stromy*, k odhadům hodnot kategoriální proměnné se používají *klasifikační stromy*. Vysvětlující proměnné, které vstupují do analýzy, mohou být různých typů, pro vytváření stromu se však uvažují kategoriální proměnné (pokud jsou kvantitativní spojité, jsou navrhovány intervaly hodnot). Cílem modelování je vytvořit stromovou strukturu. Existuje řada různých algoritmů, jako příklady lze uvést CART (*Classification And Regression Tree*) nebo CHAID (*Chi-Square Automatic Interaction Detection*) pro kategoriální data, více viz [6] a [16].

Stromová struktura se graficky zobrazuje jako schéma, jehož prvky jsou uzly a větve (viz výstup 6.1). *Uzly* jsou přitom uspořádány do různých úrovní. Na nejvyšší úrovni se nachází jediný uzel, který se nazývá *kořen*. Uzly jsou dvou typů: *nelistové*, které se odkazují na nižší úrovně, a listové (*listy*), které jsou na nejnižší úrovni. Od kořene k listům vedou *větve*. Taková struktura se při analýze dat používá i pro jiné účely než zde uvedené.

Při využití stromové struktury pro modelování závislosti určité proměnné na proměnných vysvětlujících je kořenovým uzlem vysvětlovaná proměnná. Pro štěpení se vybere vysvětlující proměnná, která má největší vliv na hodnoty vysvětlované proměnné (v případě vícekategoriální proměnné je uvažován též menší počet kategorií). V IBM SPSS je v algoritmu CHAID standardně nastaven test pomocí Pearsonovy statistiky chí-kvadrát, viz vzorec (4.2) s hladinou významnosti 0,05 (jako alternativu lze zvolit věrohodnostní poměr, zadat lze i jinou hladinu významnosti).

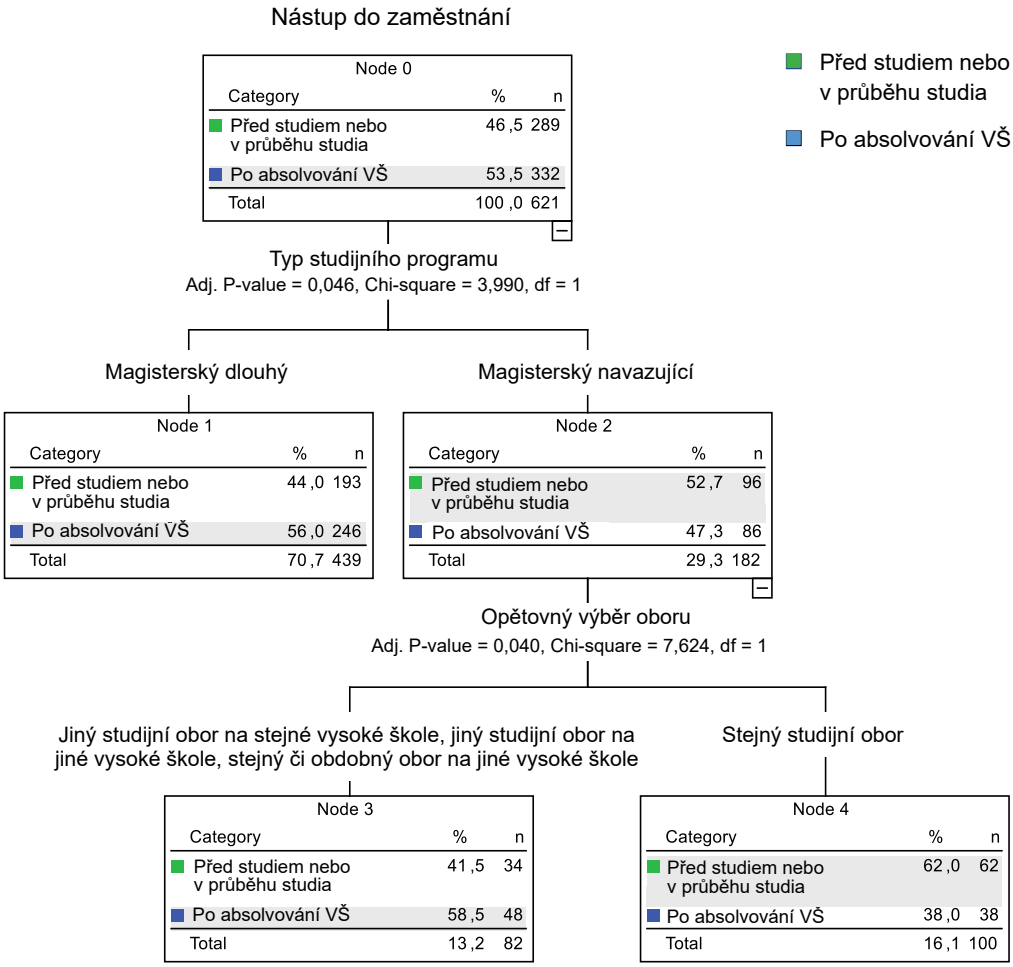
V dalších krocích se postupně v každém uzlu vyberou vysvětlující proměnné, u nichž byla zjištěna největší závislost ve vztahu k vysvětlované proměnné. Štěpení je ukončeno tehdy, když buď nelze vybrat vysvětlující proměnnou, na které by byla vysvětlovaná proměnná závislá, nebo jsou dosažena určitá kritéria, jejichž základem může být počet objektů nebo variabilita hodnot v uzlu. V systému IBM SPSS jsou pro algoritmus CHAID standardně nastaveny maximálně tři úrovně štěpení a minimální počet objektů v listu je 50 (tyto parametry lze změnit).

Příklad 6.1

Zajímá nás, zda doba nástupu do zaměstnání (proměnná *C1*, kategorie *před studiem nebo v průběhu studia* a kategorie *po absolvování VŠ*) může být vysvětlena typem studijního programu (*A3*) a vhodnou volbou studijního oboru (*B3*). Do analýzy nebudou zahrnuti absolventi, kteří dosud nepracují (nastavuje se pomocí nabídek *Data* a *Select Cases*).

V systému *IBM SPSS Statistics* s modulem *Decision Trees* vybereme nabízené možnosti *Analyze*, *Classify* a *Tree* (standardně nastavena metoda CHAID). Do části *Dependent Variable* zadáme proměnnou *C1* a do části *Independent Variables* proměnné *A3* a *B3*). Hlavní součástí výstupu je stromová struktura, viz výstup 6.1.

Výstup 6.1 | Klasifikační strom k příkladu 6.1



Výsledek můžeme interpretovat následujícím způsobem. Největší vliv na to, kdy absolvent nastoupil do zaměstnání, měl typ studijního programu. Pro absolventy magisterského navazujícího programu měla ještě vliv vhodnost výběru studijního oboru. Před studiem nebo v průběhu studia nastupovali do zaměstnání častěji absolventi magisterského navazujícího programu, kteří byli spokojeni se svým výběrem studijního oboru (62 % z těchto absolventů nastoupilo v uvedené době). Po absolvování vysoké školy častěji nastupovali absolventi dlouhého magisterského programu (56 % z nich nastoupilo po absolvování) a také absolventi magisterského navazujícího programu, kteří nebyli spokojeni se svým výběrem studijního oboru nebo výběrem vysoké školy (58,5 %).

Naznačme si postup při konstrukci stromu. Schéma se začíná vytvářet od kořenového uzlu, který představuje všechny sledované (zaměstnané) absolventy. V tomto uzlu se

vyznačí četnost variant odpovědí u vysvětlované proměnné. Zkoumáme postupně vysvětlující proměnné a pomocí chí-kvadrát testu o nezávislosti zjišťujeme, na které z nich je vysvětlovaná proměnná nejsilněji závislá. V případě vícekategoriálních vysvětlujících proměnných se prověřují též odvozené proměnné s menším počtem kategorií, které vznikly sloučením některých původních kategorií.

Tabulka 6.1 obsahuje hodnoty Pearsonovy chí-kvadrát statistiky, která je základem pro posouzení závislosti doby nástupu do zaměstnání na typu studijního programu a na vhodnosti výběru studijního oboru. V druhém případě je uvažováno několik variant počtu kategorií. V tabulce 6.1 jsou uvedeny pouze některé z možných variant. V nich je vždy spojena poslední kategorie *žádný studijní obor*, jejíž četnost je 2, tudíž není prováděn test pro původní proměnnou s pěti kategoriemi. Hlavním kritériem je minimální hladina významnosti pro zamítnutí nulové hypotézy o nezávislosti, která závisí na počtu stupňů volnosti. Na statisticky významnou závislost můžeme usuzovat pouze v případě typu studijního programu.

Tabulka 6.1 | Výsledek testů nezávislosti k příkladu 6.1 (zaměstnaní absolventi)

	Chí-kvadrát statistika	Počet stupňů volnosti	Minimální hladina významnosti
Typ studijního programu	3,990	1	0,046
Opětovný výběr oboru 4	3,849	3	0,278
Opětovný výběr oboru 3a	3,277	2	0,194
Opětovný výběr oboru 3b	3,847	2	0,147
Opětovný výběr oboru 2a	1,137	1	0,286
Opětovný výběr oboru 2b	3,246	1	0,072
Opětovný výběr oboru 2c	1,171	1	0,279

Z kořenového uzlu tedy vedou větve ke dvěma uzlům na nižší úrovni. Tyto uzly představují skupiny absolventů podle typu studijního programu. Jejich pořadí je určeno nejvyšším podílem modální kategorie vysvětlované proměnné. V každém uzlu se opět vyznačí četnost variant odpovědí u vysvětlované proměnné (v rámci jednotlivých skupin absolventů). Pro každou z těchto skupin se provedou testy o nezávislosti vysvětlované proměnné a proměnných dosud nepoužitých pro štěpení, tj. vhodnosti výběru studijního oboru, s různými počty a kombinacemi kategorií. Protože je prováděn statistický test, který navazuje na předchozí test, minimální hladina významnosti se přepočítává Bonferroniho metodou (výsledkem jsou vyšší hodnoty minimální hladiny významnosti).

Na základě zjištěných výsledků se v první skupině (dlouhý magisterský program) již štěpení neprovádí (minimální hladiny významnosti před přepočtem Bonferroniho metodou nabývají hodnot přibližně od 0,6 do 0,8), ve druhé skupině (magisterský navazující program) provede štěpení podle proměnné se dvěma kategoriemi s významy „stejný studijní obor“

a „ostatní možnosti“ (minimální hodnota před přepočtem 0,006, po přepočtu 0,04 – vztahuje se k hodnotě chí-kvadrát statistiky 7,624). Kategorie „ostatní možnosti“ ve výstupu 6.1 nezahrnuje původní kategorii *žádný studijní obor* z důvodu, že žádný absolvent magisterského navazujícího programu neoznačil tuto variantu odpovědi.

6.2 Logistická regrese

Jak již bylo zmíněno v úvodu této kapitoly, jednou z možností modelování vztahu vysvětlované proměnné na proměnných vysvětlujících je regresní analýza. Klasický lineární model mimo jiné předpokládá, že vysvětlovaná proměnná je kvantitativní spojitá. Pokud tento předpoklad není splněn, jak tomu bylo v příkladu 6.1, je třeba postupovat jiným způsobem.

Pro vysvětlovanou proměnnou dichotomickou nebo ordinální lze využít logistickou regresi. Ta vychází z přirozeného logaritmu šance, že vysvětlovaná proměnná nabude buď určité kategorie, nebo určité či nižší kategorie v případě ordinální proměnné. Nabývá-li proměnná jen dvou variant hodnot, zaměříme se na jednu z nich. Pravděpodobnost, že proměnná Y nabude kategorie označené kódem 1, označme písmenem π, tj. $P(Y = 1) = \pi$. Pravděpodobnost výskytu druhé kategorie pak označíme $1 - \pi$. Podíl těchto dvou hodnot vyjadřuje šanci, že vysvětlovaná proměnná nabude sledované kategorie. Logaritmus této šance se nazývá *logit*.

V logistické regresi se vychází ze zobecněného lineárního modelu, v němž se pomocí lineární transformace vysvětlujících proměnných vyjadřuje právě takový logit, bližší vysvětlení viz [19] a [33]. Uvažujeme-li k vysvětlujících (kvantitativních nebo binárních[9]) proměnných $X_1, X_2, \ldots, X_k$, pak pro konkrétní hodnoty $x_1, x_2, \ldots, x_k$ lze zapsat regresní funkci

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k , \tag{6.1}$$

kde $\beta_0, \beta_1, \ldots, \beta_k$ jsou parametry modelu a π je podmíněná střední hodnota vysvětlované proměnné. Ze vztahu (6.1) vyjádříme hodnotu π jako

$$\pi = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}} . \tag{6.2}$$

Není-li některá z vysvětlujících proměnných ani kvantitativní, ani binární, lze místo ní použít $K - 1$ pomocných proměnných (nejčastěji binárních), kde K je počet kategorií. O způsobech převedení bude pojednáno níže.

Pro jednu vysvětlující proměnnou X se vztah (6.1) zjednoduší na

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x . \tag{6.3}$$

9 Binární proměnná nabývá pouze hodnot 0 a 1.

Je-li vysvětlující proměnná dichotomická, potom může model vycházet ze vztahů probíraných v souvislosti s poměrem šancí, viz oddíl 4.3.6.

V kontingenční tabulce podmíněných relativních četností pro dvě binární proměnné můžeme zavést značení uvedené ve schématu 6.1. Vysvětlovaná proměnná je označena symbolem Y, $p_{1|1}$ je odhad pravděpodobnosti, že proměnná Y nabude hodnoty 1, pokud proměnná X nabude hodnoty 1, tj. $P(Y = 1 \mid X = 1)$, atd.

Schéma 6.1 | Značení podmíněných relativních četností pro dvě binární proměnné

X \ *Y*	1	0
1	$p_{1\|1}$	$p_{0\|1}$
0	$p_{1\|0}$	$p_{0\|0}$

Změnu šance nastoupení jevu vyjádříme vztahem (při změně hodnoty vysvětlované proměnné z 0 na 1)

$$\frac{\dfrac{p_{1|1}}{1-p_{1|1}}}{\dfrac{p_{1|0}}{1-p_{1|0}}} = \frac{\dfrac{n_{1,1}}{n_{0,1}}}{\dfrac{n_{1,0}}{n_{0,0}}} = \frac{n_{1,1}n_{0,0}}{n_{0,1}n_{1,0}},$$

což je již dříve zmíněný *poměr šancí*. Tento podíl vyjadřuje, kolikrát se zvýší (resp. sníží) šance nastoupení jevu při jednotkové změně hodnoty vysvětlující proměnné. Jeho zlogaritmováním získáme odhad koeficientu β_1, tj.

$$\hat{\beta}_1 = \ln\frac{p_{1|1}}{1-p_{1|1}} - \ln\frac{p_{1|0}}{1-p_{1|0}},$$

přičemž hodnota, kterou odečítáme, je stav, z něhož se vychází, tedy odhad koeficientu β_0. Pro $X = 0$ totiž platí, že

$$\hat{\beta}_0 = \ln\frac{p_{1|0}}{1-p_{1|0}}.$$

Nejsou-li hodnoty proměnných kódovány jako binární (nabývají jiných hodnot než 0 a 1), pak systém *IBM SPSS Statistics* vysvětlovanou proměnnou překóduje tak, že první kategorii (u číselných se bere vzestupné pořadí) přiřadí hodnotu 0, druhé hodnotu 1. Pro vysvětlující proměnnou lze nastavit, která kategorie bude *referenční* (základní), které se tedy přiřadí hodnota 0. Standardně program přiřadí kód 1 první kategorii. Binární vysvětlující proměnná se označuje jako *indikátorová*.

Příklad 6.2

Vyjdeme z příkladu 4.23, týkajícího se *působení v prvním zaměstnání do současné doby* v závislosti na *typu studijního programu*. Uveďme si tentokrát kontingenční tabulku řádkových relativních četností, viz výstup 6.2, pro který bylo u obou proměnných zaměněno pořadí kategorií (proměnné *A3_rev* a *D1_rev* s kódy kategorií „1“ a „2“).

Výstup 6.2 | Kontingenční tabulka řádkových relativních četností k příkladu 6.2

		Působení v prvním zaměstnání do současné doby		Celkem
		Ano	Ne	
Typ studijního programu	Magisterský navazující	47,3 %	52,7 %	100,0 %
	Magisterský dlouhý	37,6 %	62,4 %	100,0 %
Celkem		40,4 %	59,6 %	100,0 %

Pravděpodobnost, že absolvent magisterského navazujícího programu ($X = 1$) bude působit v době šetření v prvním zaměstnání, můžeme odhadnout relativní četností 0,473. Logit této četnosti (při použití uvedené zaokrouhledné hodnoty) je

$$\ln \frac{p_{1|1}}{1-p_{1|1}} = \ln \frac{0,473}{1-0,473} = -0,108 .$$

Pravděpodobnost, že v době šetření bude působit v prvním zaměstnání absolvent dlouhého magisterského programu, lze odhadnout relativní četností 0,376 (zaokrouhleno). Logitem je

$$\ln \frac{p_{1|0}}{1-p_{1|0}} = \ln \frac{0,376}{1-0,376} = -0,507 .$$

Pokud $X = 0$, pak ze vztahu (6.3) lze odvodit, že

$$\ln \frac{p_{1|0}}{1-p_{1|0}} = \hat{\beta}_0 ,$$

tj. $\hat{\beta}_0 = -0,507$. Pokud $X = 1$, platí, že

$$\ln \frac{p_{1|1}}{1-p_{1|1}} = \hat{\beta}_0 + \hat{\beta}_1 ,$$

a tedy $\hat{\beta}_1 = \ln \frac{p_{1|1}}{1-p_{1|1}} - \hat{\beta}_0 = -0,108 + 0,507 = 0,399$ (při přesném výpočtu vyjde 0,397).

Zvětší-li se vysvětlující proměnná o jednotku, má to za následek $e^{\hat{\beta}_1}$ krát větší šanci setrvání v prvním zaměstnání. V našem případě tedy $e^{0,399}= 1,49$, absolvent magisterského navazujícího programu má 1,49krát větší šanci, že déle setrvá v prvním zaměstnání než absolvent dlouhého magisterského programu. To odpovídá podílu

$$\frac{\frac{p_{1|1}}{1-p_{1|1}}}{\frac{p_{1|0}}{1-p_{1|0}}} = \frac{\frac{0,473}{1-0,473}}{\frac{0,376}{1-0,376}} = 1,49\,.$$

IBM SPSS Statistics Base* a *IBM SPSS Regression

Pro analýzu v systému *IBM SPSS Statistics* zadáme původní vysvětlovanou proměnnou *D1*. Prostřednictvím nabídek vybereme *Analyze*, *Regression* a *Binary Logistic*. Vysvětlovanou proměnnou zadáváme do políčka *Dependent*, vysvětlující do políčka *Covariates* a v rámci možnosti *Categorical* specifikujeme, že vysvětlující proměnná je kategoriální (přesuneme ji do políčka *Categorical Covariates*). Buď zadáme vysvětlující proměnnou *A3_rev* s opačným pořadím kategorií a ponecháme původní nastavení, nebo zadáme proměnnou *A3*, pomocí možnosti *Categorical* v části *Change Contrast* změníme referenční kategorii na první (First) a potvrdíme pomocí *Change*. Odhad parametrů je obsažen ve výstupu 6.3 ve sloupci *B* (jde pouze o jednu z několika tabulek, které se v systému zobrazují). Jaké násobky šance setrvání v prvním zaměstnání odpovídají jednotkové změně vysvětlující proměnné (navazující magisterské studium ve srovnání s dlouhým magisterským studiem), vyjadřuje první hodnota ve sloupci *Exp(B)*. Získané hodnoty se od výsledků spočtených bez systému IBM SPSS na základě zaokrouhlených relativních četností liší z důvodu zaokrouhlení.

Hodnoty ve sloupci *Sig.* indikují, zda je parametr modelu statisticky významný. Jde o minimální hladinu významnosti, od které zamítáme hypotézu u nulovosti, tj. nevýznamnosti parametru. Získaná hodnota 0,026 znamená, že na 5% hladině významnosti zamítáme hypotézu o nulovosti parametru β_1, a že je tedy tento parametr statisticky významný. Vysvětlení ostatních charakteristik, včetně jejich interpretace, lze nalézt v [19].

Výstup 6.3 | Odhad parametrů modelu logistické regrese k příkladu 6.2

Variables in the Equation		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Typ studijního programu(1)	,397	,178	4,968	1	,026	1,488
	Constant	–,507	,099	26,491	1	,000	,602

a. Variable(s) entered on step 1: typ studijního programu.

♦ ♦ ♦

Model

$$\ln \frac{\pi_{1|1}}{1-\pi_{1|1}} = \beta_0 + \beta_1$$

se označuje jako *logitový model* s jednou dichotomickou vysvětlující proměnnou. Obecně lze do logitového modelu zařadit jednak více vysvětlujících proměnných, jednak proměnné vícekategoriální, podrobněji viz např. [1], [2], [19] a [32].

Má-li vysvětlující proměnná více než dvě kategorie, pak se do modelu logistické regrese zařazuje $K - 1$ pomocných proměnných, kde K je počet kategorií. Je možné použít různé způsoby kódování. Dále bude uvažováno kódování, při kterém e^{β_j} vyjadřuje, kolikrát se zvětší šance, že vysvětlovaná proměnná nabude hodnoty 1 při změně vysvětlující proměnné z referenční (např. poslední) na j-tou kategorii.

Použité pomocné proměnné jsou binární, a tedy *indikátorové*. Například nabývá-li proměnná X tří kategorií, pak budou do modelu zařazeny dvě binární proměnné Z_1 a Z_2, přičemž $Z_1 = 1$ pro $X = 1$ a $Z_2 = 1$ pro $X = 2$, jinak $Z_j = 0$ ($j = 1, 2$). Regresní funkci s konkrétními hodnotami z_1 a z_2 lze zapsat ve tvaru

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 ,$$

nebo logitový model jako

$$\ln \frac{\pi_{1|j}}{1-\pi_{1|j}} = \beta_0 + \beta_j ,$$

kde $j = 1, 2$ (obecně $K - 1$).

Příklad 6.3

Zajímá nás, zda lze na *opětovný výběr oboru* (proměnná *B3* zredukovaná na dvě kategorie v členění „stejný studijní obor“ a „jiná možnost“, tj. proměnná *B3_2kat_a*) usoudit na základě hodnocení *přínosu oboru pro vstup do práce* (proměnná *B2a* zredukovaná na tři kategorie, tj. proměnná *B2a_3kat*). Rozdělme úlohu na dvě části, přičemž v každé budeme sledovat změnu šance opětovného výběru oboru při změně vysvětlující proměnné z referenční na jednu ze dvou ostatních kategorií. V systému *IBM SPSS Statistics* k ilustraci problematiky použijeme překódovanou proměnnou *B2a_3kat_rev* s opačným pořadím kategorií a provedeme nejprve výběr absolventů, kteří přínos oboru pro vstup do práce hodnotili jako „žádný nebo malý přínos“ a „střední přínos“, a poté absolventů s hodnocením „žádný nebo malý přínos“ a „větší nebo velký přínos“. Pro každou možnost spočteme poměr šancí.

Protože výpočet poměru šancí bez použití programového systému byl již výše naznačen, budou dále uvedeny pouze upravené výstupy ze systému *IBM SPSS Statistics*. Výstup 6.4 obsahuje tabulku řádkových relativních četností pro první výběr; odpovídající poměr šancí je ve výstupu 6.5. Obdobně tabulka řádkových relativních četností pro druhý výběr je ve výstupu 6.6 a odpovídající poměr šancí ve výstupu 6.7.

Výstup 6.4 | Kontingenční tabulka řádkových relativních četností k příkladu 6.3 (1. výběr)

		Opětovný výběr oboru		Celkem
		Stejný studijní obor	Jiný obor nebo jiná VŠ nebo žádná VŠ	
Přínos oboru pro vstup do práce	Střední přínos	53,8 %	46,2 %	100,0 %
	Žádný nebo malý přínos	27,9 %	72,1 %	100,0 %
Celkem	42,6 %	57,4 %	100,0 %	

Výstup 6.5 | Odhad poměru šancí k příkladu 6.3 (1. výběr)

Risk Estimate		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for přínos oboru pro vstup do práce (střední přínos / žádný nebo malý přínos)	3,016	1,886	4,824
N of Valid Cases	324		

Výstup 6.6 | Kontingenční tabulka řádkových relativních četností k příkladu 6.3 (2. výběr)

		Opětovný výběr oboru		Celkem
		Stejný studijní obor	Jiný obor nebo jiná VŠ nebo žádná VŠ	
Přínos oboru pro vstup do práce	Větší nebo velký přínos	78,8 %	21,2 %	100,0 %
	Žádný nebo malý přínos	27,9 %	72,1 %	100,0 %
Celkem	63,0 %	37,0 %	100,0 %	

Výstup 6.7 | Odhad poměru šancí k příkladu 6.3 (2. výběr)

Risk Estimate	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for přínos oboru pro vstup do práce (větší nebo velký přínos / žádný nebo malý přínos)	9,613	6,077	15,209
N of Valid Cases	451		

Výstup 6.5 obsahuje odhad poměru šancí, podle kterého můžeme usoudit, že absolventi, kteří hodnotí přínos oboru pro vstup do práce jako střední, mají ve vztahu k absolventům, kteří hodnotí přínos oboru pro vstup do práce jako malý nebo žádný, více než 3krát větší šanci, že budou také spokojeni se svým oborem tak, že by si ho znovu vybrali ke studiu. Obdobně na základě výstupu 6.7 lze zjistit, že absolventi, kteří hodnotí přínos oboru pro vstup do práce jako větší nebo velký, mají ve vztahu k absolventům, kteří hodnotí přínos oboru pro vstup do práce jako malý nebo žádný, 9,613krát větší šanci, že budou také spokojeni se svým oborem tak, že by si ho znovu vybrali ke studiu. Logaritmy uvedených odhadů poměru šancí jsou odhady parametrů modelu logistické regrese, tj. $\hat{\beta}_1 = \ln(3{,}016) = 1{,}104$ a $\hat{\beta}_2 = \ln(9{,}613) = 2{,}263$. Odhad konstanty vychází z relativní četnosti $p_{1|0} = 0{,}279$ (viz výstupy 6.4 a 6.6). Je to

$$\hat{\beta}_0 = \ln\frac{p_{1|0}}{1-p_{1|0}} = \ln\frac{0{,}279}{1-0{,}279} = -0{,}949\,.$$

Hodnota vyjadřuje logaritmus odhadu šance, že absolvent, který hodnotí přínos oboru pro vstup do práce jako malý nebo žádný, bude také spokojen se svým oborem tak, že by si ho znovu vybral ke studiu. Pro absolventy, kteří hodnotí přínos oboru pro vstup do práce jako střední, bude tento logit dán vztahem

$$\ln\frac{p_{1|1}}{1-p_{1|1}} = \hat{\beta}_0 + \hat{\beta}_1 = -0{,}949 + 1{,}104 = 0{,}155$$

a pro absolventy, kteří hodnotí přínos oboru pro vstup do práce jako větší nebo velký,

$$\ln\frac{p_{1|2}}{1-p_{1|2}} = \hat{\beta}_0 + \hat{\beta}_2 = -0{,}949 + 2{,}263 = 1{,}314\,.$$

Pokud by výpočty byly provedeny s nezaokrouhlenými relativními četnostmi (s odhadem konstanty –0,952), pak by odpovídaly logaritmům šancí, kdy by za $p_{1|1}$ byla dosazena hodnota 0,538 (viz výstup 6.4) a za $p_{1|2}$ dosazena hodnota 0,788 (viz výstup 6.6).

IBM SPSS Statistics Base* a *IBM SPSS Regression

Pro zadání analýzy v systému *IBM SPSS Statistics* je třeba překódovat vysvětlovanou proměnnou *B3_2kat_a* (opětovný výběr oboru) na binární tak, že hodnota „1“ bude označovat odpověď „ano“ (absolvent by si vybral stejný studijní obor). Pak prostřednictvím nabídek vybereme *Analyze*, *Regression* a *Binary Logistic*. Překódovanou vysvětlovanou proměnnou (*B3_2kat_a_bin*) zadáváme do políčka *Dependent*, vysvětlující (*B2a_3kat*) do políčka *Covariates* a volbou *Categorical* specifikujeme, že vysvětlující proměnná je kategoriální (přesunem do políčka *Categorical Covariates*). Zadáme, aby referenční kategorie byla první (volba *First* – je třeba potvrdit pomocí *Change*).

Místo proměnné *B2a_3kat* (přínos oboru pro vstup do práce) budou použity dvě indikátorové proměnné *B2a_3kat(1)* a *B2a_3kat(2)*. Způsob překódování automaticky provedeného systémem *IBM SPSS Statistics* (při zadání první kategorie jako referenční) je vysvětlen ve výstupu 6.8.

Výstup 6.8 | Způsob nahrazení vícekategoriální proměnné indikátorovými proměnnými

Categorical Variables Codings		Frequency	Parameter coding	
			(1)	(2)
Přínos oboru pro vstup do práce	Žádný nebo malý přínos	140	,000	,000
	Střední přínos	184	1,000	,000
	Větší nebo velký přínos	311	,000	1,000

Odhad parametrů je obsažen ve výstupu 6.9 ve sloupci *B*. Hodnoty ve sloupci *Exp(B)* vyjadřují, jaké násobky šance na opětovný výběr oboru odpovídají jednotkové změně vysvětlující proměnné (stupeň přínosu oboru pro vstup do práce) vzhledem k referenční kategorii (žádný nebo malý přínos). Získané hodnoty se shodují s výsledky spočtenými na základě poměrů šancí (pokud by výpočty byly provedeny s nezaokrouhlenými hodnotami).

Výstup 6.9 | Odhad parametrů modelu logistické regrese k příkladu 6.3

Variables in the Equation		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Přínos oboru pro vstup do práce			97,376	2	,000	
	Přínos oboru pro vstup do práce(1)	1,104	,240	21,232	1	,000	3,016
	Přínos oboru pro vstup do práce(2)	2,263	,234	93,508	1	,000	9,613
	Constant	–,952	,189	25,476	1	,000	,386

a. Variable(s) entered on step 1: přínos oboru pro vstup do práce.

Ve sloupci *Sig.* jsou minimální hladiny významnosti velmi nízké. To znamená, že na 1% hladině významnosti zamítáme jak hypotézu o nulovosti parametru β_1, tak hypotézu o nulovosti parametru β_2. Oba parametry jsou tedy v modelu statisticky významné.

♦ ♦ ♦

Jsou-li do modelu zařazeny kvantitativní vysvětlující proměnné, pak $e^{\beta j}$ vyjadřuje, kolikrát se zvětší šance, že vysvětlovaná proměnná nabude hodnoty 1 při změně *j*-té vysvětlující proměnné o jednotku, za předpokladu, že hodnoty ostatních vysvětlujících proměnných zůstanou nezměněny. Pro odhad parametrů se používá metoda maximální věrohodnosti. Aplikuje se iterativní postup, kdy se vychází z počátečních odhadů, které se postupně vylepšují. Více například viz [19].

Při zařazení několika vysvětlujících proměnných se zkoumá, zda mají být do modelu zařazeny všechny tyto proměnné, nebo jen některé z nich. Posuzování vhodnosti jednotlivých proměnných již bylo naznačeno v příkladech 6.2 a 6.3. Testuje se nulovost

jednotlivých regresních parametrů. V systému *IBM SPSS Statistics* se pro tento účel používá **Waldova statistika** (viz sloupec *Wald* ve výstupech 6.3 a 6.9), jejíž druhá mocnina má asymptoticky chí-kvadrát rozdělení. Podrobnější výklad lze nalézt v knize [19]. Pro výpočet Waldovy statistiky se používají dva přístupy – jeden v případě, kdy vysvětlující proměnná není zadána jako kategoriální, a druhý v případě, kdy je kategoriální.

Dále lze modely, obsahující různé vysvětlující proměnné, porovnávat podle správnosti zařazení objektu (respondenta) do jedné ze dvou skupin, odpovídajících hodnotě 0 nebo 1 vysvětlované proměnné. Zařazení se provádí pomocí odhadu pravděpodobnosti ($\hat{\pi}$), že vysvětlovaná proměnná nabude hodnoty 1. Tento odhad se porovnává s předem zadanou *prahovou hodnotou* (standardně nastaveno $\pi_0 = 0{,}5$). Jestliže $\hat{\pi} \geq \pi_0$, je vysvětlované proměnné přiřazena hodnota 1.

Pro vyhodnocení úspěšnosti modelu se sestrojí čtyřpolní *klasifikační tabulka*, na základě které lze určit podíl správně zařazených objektů (zvlášť pro kategorie 0 a 1 a též celkově). Kromě tohoto podílu se pro hodnocení úspěšnosti používají různé statistiky. Jako příklad lze uvést **Nagelkerkeho statistiku**, která nabývá hodnoty od 0 do 1, viz [19].

Pro testování shody modelu s daty lze použít chí-kvadrát test dobré shody, viz oddíl 3.4.2. Častější jsou však jiné postupy, například se využívá **Hosmerova-Lemeshowova statistika**. Základem pro její výpočet jsou odhady pravděpodobnosti, že vysvětlovaná proměnná nabude hodnoty 1, spočtené pro známé hodnoty vysvětlujících proměnných. Tyto očekávané pravděpodobnosti jsou vzestupně uspořádány. Seřazené hodnoty jsou rozděleny do tří nebo více přibližně stejně velkých skupin. K rozdělení se používají buď kvantily, nebo předem stanovené intervaly na definičním oboru od 0 do 1, například lze použít deset intervalů s horními hranicemi 0,1, 0,2, …, 1. V každé i-té skupině ($i = 1, 2, \ldots, g$) se sečte počet jedniček původní vysvětlované proměnné (budeme značit symbolem n_i) a dále se sečtou získané odhady pravděpodobnosti (označíme symbolem e_i). Poté se v každé skupině spočte aritmetický průměr očekávaných pravděpodobností. Označíme-li počet hodnot v i-té skupině jako m_i, pak je tento průměr dán vztahem $\overline{\pi} = e_i / m_i$. Hosmerova-Lemeshowova statistika se spočte podle vzorce

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(n_i - m_i \overline{\pi}_i)^2}{m_i \overline{\pi}_i (1 - \overline{\pi}_i)}. \tag{6.4}$$

Tato statistika má za předpokladu platnosti nulové hypotézy o shodě očekávaných a pozorovaných četností přibližně chí-kvadrát rozdělení s (g – 1) stupni volnosti. Předpoklad testu a interpretace získané hodnoty jsou stejné jako při aplikaci chí-kvadrát testu dobré shody, viz oddíl 3.4.2. Podrobněji viz např. [9].

Příklad 6.4

Rozšiřme příklad 6.3 o další vysvětlující proměnné, a to o *pohlaví* (*E1*) a o *počet zaměstnání od absolvování studia* (*C4*). Vzhledem k tomu, že logistická regresní analýza umožňuje do modelu zařadit kvantitativní spojité proměnné, použijeme konkrétní hodnoty počtu zaměstnání. V systému *IBM SPSS Statistics* v proceduře pro logistickou regresi zadáme

do políčka *Dependent* proměnnou *B3_2kat_a_bin* a do části *Covariates* proměnné *E1* (*pohlaví*), *B2a_3kat* (*přínos oboru pro vstup do práce*) a *C4* (*počet zaměstnání*).

Pomocí možnosti *Categorical* specifikujeme, že vysvětlující proměnné *E1* a *B2a_3kat* jsou kategoriální (přesuneme je do políčka *Categorical Covariates*). Zadáme, aby u *přínosu oboru pro vstup do práce* byla referenční kategorie první (volba *First*), tj. žádný nebo malý přínos; u pohlaví ponecháme druhou kategorii (volba *Last*), tj. ženu. Dále pomocí možnosti *Options* specifikujeme, aby výstup obsahoval kontingenční tabulku a výsledky pro Hosmerův-Lemeshowův test (zvolíme *Hosmer-Lemeshow goodness-of-fit*).

Odhad parametrů je obsažen ve výstupu 6.10 ve sloupci *B*. Na základě Waldovy statistiky (sloupec *Wald*) byly získány minimální hladiny významnosti (sloupec *Sig.*). Na 5% hladině významnosti tedy ve všech případech zamítáme hypotézu o nulovosti regresních koeficientů.

Výstup 6.10 | Odhad parametrů modelu logistické regrese k příkladu 6.4

Variables in the Equation		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Pohlaví(1)	,562	,258	4,754	1	,029	1,755
	Přínos oboru pro vstup do práce			55,758	2	,000	
	Přínos oboru pro vstup do práce(1)	,846	,325	6,751	1	,009	2,330
	Přínos oboru pro vstup do práce(2)	2,264	,320	50,189	1	,000	9,617
	Počet zaměstnání od absolvování studia	–,463	,161	8,315	1	,004	,629
	Constant	–,117	,440	,070	1	,791	,890

a. Variable(s) entered on step 1: pohlaví, přínos oboru pro vstup do práce, počet zaměstnání od absolvování studia.

Největší vliv na opětovný výběr oboru má větší nebo velký přínos oboru pro vstup do práce (9,617násobek šance vůči absolventům, kteří hodnotí přínos jako malý nebo žádný, viz sloupec *Exp(B)*), za ním následuje střední přínos oboru pro vstup do práce (2,33násobek šance vůči referenční kategorii). Dále má na opětovný výběr oboru vliv pohlaví; u mužů je 1,755krát větší šance než u žen. Pokud jde o počet zaměstnání od absolvování studia, pak zjišťujeme, že respondentovi s jedním zaměstnáním navíc šance klesá asi 1,6krát (1/0,629).

Na základě modelu můžeme odhadnout pravděpodobnost, že vysvětlovaná proměnná nabude hodnoty 1. Pro absolventa s jedním zaměstnáním, který hodnotí přínos oboru pro vstup do práce jako větší nebo velký, je tato šance podle (6.2)

$$\hat{\pi} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 E1(1)+\hat{\beta}_2 B2a\_3kat(1)+\hat{\beta}_3 B2a\_3kat(2)+\hat{\beta}_4 C4}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 E1(1)+\hat{\beta}_2 B2a\_3kat(1)+\hat{\beta}_3 B2a\_3kat(2)+\hat{\beta}_4 C4}} = \frac{e^{-0,117+0,562+2,264-0,463}}{1+e^{-0,117+0,562+2,264-0,463}} = 0,904\,.$$

Klasifikační tabulka je uvedena ve výstupu 6.11. Celková úspěšnost modelu je 71 % (*Overall Percentage*). Model lépe odhaduje variantu, že by absolvent zvolil stejný studijní obor (úspěšnost 73,2 %) než druhou variantu (68,2 %).

Protože byl požadován Hosmerův-Lemeshowův test, je součástí výstupu kontingenční tabulka, viz výstup 6.12. Z této tabulky je zřejmé, že bylo vytvořeno 10 intervalů zahrnujících určitý počet absolventů, a to od 23 do 61, viz sloupec *Total*. Pro každý z těchto intervalů je zvlášť pro obě kategorie vysvětlované proměnné uveden jednak zjištěný počet respondentů, kteří zvolili příslušnou odpověď (sloupec *Observed*), jednak očekávaná četnost (sloupec *Expected*). Výsledek testu zobrazuje výstup 6.13, který obsahuje hodnotu chí-kvadrát statistiky (sloupec *Chi-square*), počet stupňů volnosti (*df*) a minimální hladinu významnosti (*Sig.*). Protože je tato hladina 0,499, nezamítáme hypotézu o shodě modelu s daty na 5% hladině významnosti.

Výstup 6.11 | Klasifikační tabulka k příkladu 6.4 (upraveno)

Classification Table[a]		**Predicted**		
Step 1		**Opětovný výběr oboru**		
Observed		**Jiný obor nebo jiná VŠ nebo žádná VŠ**	**Stejný studijní obor**	**Percentage Correct**
Opětovný výběr oboru	**Jiný obor nebo jiná VŠ nebo žádná VŠ**	105	49	68,2
	Stejný studijní obor	52	142	73,2
Overall Percentage				71,0

a. The cut value is ,500.

Výstup 6.12 | Kontingenční tabulka pro Hosmerův-Lemeshowův test

	Contingency Table for Hosmer and Lemeshow Test				
Step 1	**Opětovný výběr oboru = Jiný obor nebo jiná VŠ nebo žádná VŠ**		**Opětovný výběr oboru = Stejný studijní obor**		
	Observed	**Expected**	**Observed**	**Expected**	**Total**
1	18	18,952	5	4,048	23
2	27	25,886	8	9,114	35
3	28	23,495	7	11,505	35
4	24	27,458	24	20,542	48
5	11	14,408	19	15,592	30
6	9	9,429	14	13,571	23
7	13	9,804	18	21,196	31
8	13	13,908	48	47,092	61
9	7	7,109	29	28,891	36
10	4	3,552	22	22,448	26

Výstup 6.13 | Kontingenční tabulka pro Hosmerův-Lemeshowův test

Hosmer and Lemeshow Test			
Step	**Chi-square**	**df**	**Sig.**
1	7,354	8	,499

Porovnání úspěšnosti modelů zahrnujících různé kombinace vysvětlujících proměnných shrnuje tabulka 6.5. Z hlediska Nagelkerkeho statistiky je „nejlepší" varianta poslední, zohledňující všechny tři vysvětlující proměnné, z hlediska celkové úspěšnosti pak hodnocení přínosu oboru pro vstup do práce a počet zaměstnání. Ovšem vzhledem k nízkým hodnotám Nagelkerkeho statistiky nejsou modely vhodným prostředkem pro odhady hodnot vysvětlované proměnné. Na druhou stranu lze modely použít k souhrnné analýze jednostranných závislostí. Model pouze s vysvětlující proměnnou pohlaví (*E1*) má sice 100% úspěšnost při zařazení do kategorie 1, ovšem nulovou úspěšnost při zařazení do kategorie 0.

Tabulka 6.5 | Výsledky klasifikačních tabulek pro kombinace vysvětlujících proměnných

Vysvětlující proměnné	Procentní úspěšnost zařazení do kategorie 1	Celková procentní úspěšnost	Nagelkerkeho statistika
B2a_3kat	89,8	70,1	0,219
E1	100,0	59,5	0,028
C4	92,3	56,5	0,014
B2a_3kat* + *E1	75,3	69,9	0,240
B2a_3kat* + *C4	82,7	71,4	0,245
E1* + *C4	83,0	59,8	0,035
E1* + *B2a_3kat* + *C4	73,2	71,0	0,257

♦ ♦ ♦

Porovnáme-li vypovídací schopnosti logistické regrese a klasifikačních stromů, pak výhodou klasifikačního stromu je členění úlohy do dílčích částí, v nichž mohou být zařazeny různé vysvětlující proměnné. Výsledkem logistické regrese jsou zase konkrétní hodnoty, udávající násobky šancí při jednotkových změnách vysvětlujících proměnných.

Je-li vysvětlovaná proměnná nominální *vícekategoriální*, definují se pro ni *bazické logity*, kdy se jednotlivé kategorie porovnávají se zadanou referenční kategorií, obdobně jako u vysvětlujících proměnných. V modulu *IBM SPSS Regression* lze v *multinomické*

logistické regresi jako referenční zadat první, poslední nebo i jinou konkrétní kategorii, specifikovanou jejím pořadovým číslem. Problémem tohoto přístupu je, že výskyt jednotlivých kategorií ve vztahu k referenční může být ovlivněn různými vysvětlujícími proměnnými. V takových případech je vhodnější vytvořit $K - 1$ indikátorových proměnných, kde K je počet kategorií vysvětlované proměnné, a použít binární logistickou regresi pro každou kategorii.

Příklad 6.5

Řešme modifikovanou úlohu z příkladu 6.4, kdy jako vysvětlovaná proměnná bude do modelu zařazena proměnná *opětovný výběr oboru* se třemi kategoriemi (*B3_3kat_a*). Jako vysvětlující proměnné uvažujme *přínos oboru pro vstup do práce* se třemi kategoriemi (*B2a_3kat*) a *počet zaměstnání* (*C4*).

V *IBM SPSS Statistics* prostřednictvím nabídek vybereme *Analyze*, *Regression* a *Multinomial Logistic*. Vysvětlovanou proměnnou (*B3_3kat_a*) zadáme do políčka *Dependent* (pomocí možnosti *Reference Category* lze nastavit jinou referenční proměnnou než poslední, toto nastavení však ponecháme). Kategoriální proměnnou *B2a_3kat* zadáme do části *Factor(s)* a proměnnou *C4* do části *Covariate(s)*.

Součástí výstupu je tabulka odhadů parametrů a souvisejících charakteristik, obdobně jako v případě binární logistické regrese. Má-li vysvětlovaná proměnná K kategorií, pak je tabulka rozdělena do $K - 1$ bloků odpovídajících jednotlivým kategoriím kromě referenční. V našem případě je referenční kategorie „jinou nebo žádnou VŠ“ a tabulka obsahuje dva bloky pro kategorie „stejný studijní obor“ a „jiný studijní obor na stejné VŠ“, viz výstup 6.14 (jsou vynechány intervaly spolehlivosti k bodovým odhadům *Exp(B)*). Statisticky významnými parametry byly shledány pouze parametry ve skupině „jiný studijní obor“. Měl by být tedy vybrán model pro vysvětlovanou proměnnou se dvěma kategoriemi, viz příklad 6.4.

Výstup 6.14 | Odhad parametrů modelu multinomické logistické regrese k příkladu 6.5

Parameter Estimates							
Opětovný výběr oboru[a]		**B**	**Std. Error**	**Wald**	**df**	**Sig.**	**Exp(B)**
Stejný studijní obor	**Intercept**	3,750	,616	37,110	1	,000	
	Počet zaměstnání od absolvování studia	–,597	,200	8,876	1	,003	,551
	[přínos oboru pro vstup do práce=1]	–2,388	,411	33,732	1	,000	,092
	[přínos oboru pro vstup do práce=2]	–1,478	,382	14,982	1	,000	,228
	[přínos oboru pro vstup do práce=3]	0[b]	.	.	0	.	.
Jiný studijní obor na stejné VŠ	**Intercept**	1,346	,654	4,240	1	,039	
	Počet zaměstnání od absolvování studia	–,294	,207	2,010	1	,156	,745
	[přínos oboru pro vstup do práce=1]	–,144	,421	,117	1	,733	,866
	[přínos oboru pro vstup do práce=2]	–,117	,426	,075	1	,784	,890
	[přínos oboru pro vstup do práce=3]	0[b]	.	.	0	.	.

a. The reference category is: jinou nebo žádnou VŠ.

b. This parameter is set to zero because it is redundant.

Zjišťování podobnosti kategorií

7

Při předzpracování dat, prováděném před vlastní analýzou, je jednou z možných transformací sdružování kategorií, což vede k překódování některých proměnných na menší počet kategorií. Transformované proměnné byly v této knize použity v mnoha příkladech. Překódovány byly jednak ordinální proměnné, kdy místo pětibodové stupnice byly vytvořeny tři nebo dvě kategorie (např. hodnocení přínosů oboru), jednak proměnné nominální, kdy z více kategorií byly opět vyvořeny tři nebo dvě (např. byly spojeny kategorie vyjadřující, zda pro současné zaměstnání je vhodný vystudovaný obor a zda je vhodný příbuzný studijní obor).

Důvodem překódování bylo například zajištění dostatečného zastoupení kombinací kategorií (vyjádřeného pomocí sdružených četností v kontingenční tabulce), porovnání vybraných skupin kategorií či možnost uspořádání kategorií. U ordinálních proměnných byly spojovány sousední kategorie (např. „větší přínos“ a „velký přínos“), u nominálních buď kategorie obsahově příbuzné (vystudovaný obor a příbuzný obor), nebo málo zastoupené kategorie s výsledným významem „ostatní“.

V některých případech u nominální proměnné není dána žádná apriorní informace, která by mohla být podkladem pro spojení kategorií, a přitom je potřeba kategorie spojit z důvodu malých sdružených četností. Někdy může mít počet kategorií vliv na výsledek testu jednostranné nezávislosti. Tehdy je vhodné použít metody pro odhalení skupin podobných kategorií vysvětlující proměnné ve vztahu k proměnné vysvětlované.

Ke zjišťování skupin podobných kategorií lze použít různé metody. Je to především *hierarchické aglomerativní shlukování*[10], případně *vícerozměrné škálování*. Jde o metody, které lze obecně použít pro zjišťování skupin podobných objektů (odhalení skupin respondentů, jejichž odpovědi se často shodují) či proměnných (pokud se na některé kombinace otázek vyskytují stejné kombinace odpovědí, lze vytvořit dotazník s menším počtem otázek). V této knize bude probrána pouze jedna z možných aplikací, zaměřená právě na spojování kategorií.

Metodou navrženou speciálně pro identifikaci podobných kategorií je *korespondenční analýza*. Pomocí ní se zjišťuje podobnost kategorií buď dvou, nebo více proměnných (tj. *vícenásobná korespondenční analýza*). V takovém případě lze zjistit též skupiny podobných proměnných. Tato problematika však zde nebude probírána.

10 V této kapitole jsou používány doslovné překlady z angličtiny jako *shluková analýza* (anglicky *cluster analysis*) a *shlukování* (*clustering*). V česky psané literatuře se však používají též jiné termíny, zejména *seskupovací analýza* a *seskupování*, které lépe vyjadřují podstatu procesu, jehož cílem je nalezení skupin. Používá se též počeštěný výraz *klastrování*.

7.1 Shluková analýza

Shluková analýza umožňuje buď identifikovat skupiny (shluky) podobných kategorií jedné proměnné na základě kategorií druhé proměnné, nebo zjišťovat vazby mezi kategoriemi obou proměnných (tzv. dvourozměrné shlukování). V obou případech jde obvykle o *hierarchické shlukování*. Nejčastěji používané je *shlukování aglomerativní*, kdy na počátku je každá kategorie samostatným shlukem. Nejprve se spojí dvě nejpodobnější kategorie do jednoho shluku a postup se opakuje tak dlouho, až jsou všechny kategorie zařazeny do jednoho shluku. Základní principy shlukové analýzy jsou popsány např. v knihách [37] a [41].

Vstupem pro shlukování kategorií je *kontingenční tabulka*, výstupem je identifikace *skupin podobných kategorií*. K posuzování vztahů mezi kategoriemi slouží *míry podobnosti*, resp. *nepodobnosti*. Při hierarchickém shlukování se obvykle vytváří *matice vzdáleností* (nepodobností), v níž je každá dvojice kategorií charakterizována číslem vyjadřujícím, do jaké míry jsou sledované kategorie *blízké*. Čím více se míra nepodobnosti blíží hodnotě 0, tím bližší jsou si kategorie. A samozřejmě čím vyšší hodnoty míra nepodobnosti dosahuje, tím jsou kategorie rozdílnější.

V tomto oddílu se budeme věnovat pouze jednorozměrnému shlukování. Budeme tedy posuzovat podobnosti všech dvojic kategorií určité proměnné. V systému *IBM SPSS Statistics* jsou k dispozici dvě míry nepodobnosti, *chí-kvadrát* a φ. Předpokládejme, že proměnná, u níž máme zjišťovat podobnost kategorií, je řádková. Pak na základě kontingenční tabulky vytvoříme všechny možné dílčí tabulky, které budou mít pouze dva řádky. Pro všechny tyto tabulky spočteme *Pearsonovu statistiku chí-kvadrát*, viz vzorec (4.2), případně *koeficient* φ, viz vzorec (4.5). Pro vlastní shlukování se v prvním případě používá druhá odmocnina ze statistiky chí-kvadrát, tj. pro i-tou a i'-tou kategorii zapíšeme míru nepodobnosti

$$\sqrt{\chi^2} = \sqrt{\sum_{j=1}^{S} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} + \sum_{j=1}^{S} \frac{(n_{i'j} - m_{i'j})^2}{m_{i'j}}}, \tag{7.1}$$

kde

$$m_{ij} = \frac{n_{i+}(n_{ij} + n_{i'j})}{n_{i+} + n_{i'+}} \quad \text{a} \quad m_{i'j} = \frac{n_{i'+}(n_{ij} + n_{i'j})}{n_{i+} + n_{i'+}}.$$

Koeficient φ pak spočteme podle vztahu

$$\phi = \sqrt{\frac{\chi^2}{n_{i+} + n_{i'+}}}. \tag{7.2}$$

V obou případech jde o míry nepodobnosti, pro které platí, že pokud $i = i'$, pak tyto míry nabývají hodnoty 0. Analogický postup lze použít pro sloupcovou proměnnou.

Příklad 7.1

Chceme zjistit vztah kategorií proměnné *opětovný výběr oboru* (překódovaná proměnná *B3_4kat* se čtyřmi kategoriemi) na základě kategorií proměnné *studijní obor vhodný pro první zaměstnání* (proměnná *C3*) – a naopak. Kontingenční tabulka absolutních četností je ve výstupu 7.1.

Výstup 7.1 | Kontingenční tabulka absolutních četností k příkladu 7.1

Zjištěné absolutní četnosti		Studijní obor vhodný pro první zaměstnání				Celkem
		Vystu-dovaný obor	Příbuzný studijní obor	Zcela jiný studijní obor	Zaměstnání nevyžaduje oborovou specializaci	
Opětovný výběr oboru	Stejný studijní obor	42	264	22	47	375
	Jiný studijní obor na stejné vysoké škole	10	84	39	23	156
	Stejný či obdobný obor na jiné vysoké škole	2	11	1	1	15
	Jiný obor na jiné VŠ nebo žádný obor	2	43	13	17	75
Celkem		56	402	75	88	621

Nejprve budeme zjišťovat vztah kategorií proměnné *opětovný výběr oboru*. Tato proměnná má čtyři kategorie, na základě nichž může být vytvořeno šest různých dvojic. Naznačíme postup pouze pro jednu takovou dvojici, a to *stejný studijní obor* a *jiný studijní obor na stejné vysoké škole*. Odpovídající kontingenční tabulka zjištěných četností je ve výstupu 7.2, tabulka teoretických četností ve výstupu 7.3.

Výstup 7.2 | Kontingenční tabulka zjištěných četností pro první dvě kategorie

Zjištěné absolutní četnosti		Studijní obor vhodný pro první zaměstnání				Celkem
		Vystu-dovaný obor	Příbuzný studijní obor	Zcela jiný studijní obor	Zaměstnání nevyžaduje oborovou specializaci	
Opětovný výběr oboru	Stejný studijní obor	42	264	22	47	375
	Jiný studijní obor na stejné vysoké škole	10	84	39	23	156
Celkem		52	348	61	70	531

Výstup 7.3 | Kontingenční tabulka očekávaných četností pro první dvě kategorie

Očekávané četnosti		Studijní obor vhodný pro první zaměstnání				Celkem
		Vystudovaný obor	Příbuzný studijní obor	Zcela jiný studijní obor	Zaměstnání nevyžaduje oborovou specializaci	
Opětovný výběr oboru	Stejný studijní obor	36,7	245,8	43,1	49,4	375,0
	Jiný studijní obor na stejné vysoké škole	15,3	102,2	17,9	20,6	156,0
Celkem		52,0	348,0	61,0	70,0	531,0

Chí-kvadrát míru nezávislosti spočteme podle vzorce (7.1) jako

$$\sqrt{\chi^2} = \sqrt{\sum_{j=1}^{S} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} + \sum_{j=1}^{S} \frac{(n_{i'j} - m_{i'j})^2}{m_{i'j}}} =$$

$$= \sqrt{\frac{(42-36,7)^2}{36,7} + \frac{(264-245,8)^2}{245,8} + \ldots + \frac{(23-20,6)^2}{20,6}} = \sqrt{42,704} = 6,535 .$$

Koeficient φ spočteme podle vztahu (7.2) jako

$$\phi = \sqrt{\frac{\chi^2}{n_{i+} + n_{i'+}}} = \sqrt{\frac{42,704}{531}} = 0,284 .$$

Další výsledky uvedeme pouze jako výstupy ze systému *IBM SPSS Statistics*.

IBM SPSS Statistics

Abychom získali matici nepodobností pro všechny dvojice kategorií proměnné *opětovný výběr oboru*, zkopírujeme sdružené absolutní četnosti do nové tabulky datového editoru *IBM SPSS Statistics* (lze také uložit celou tabulku do Excelu, vynechat řádek a sloupec s marginálními četnostmi a importovat do systému *IBM SPSS Statistics* včetně názvů řádků a sloupců).

Pro analýzu tabulky sdružených četností volíme *Analyze*, *Classify*, *Hierachical Cluster* a zadáme všechny „proměnné“ (pokud byla importována tabulka včetně názvů řádků, pak sloupec s těmito názvy zadáme do políčka *Label Cases by*). V části *Display* ponecháme pouze možnost *Statistics* a v rámci nabídky *Statistics* zadáme *Proximity matrix*. V rámci nabídky *Methods* zvolíme v části *Measure* možnost *Count* (standardně je nastavena míra *Chi-square measure*). Hlavní součástí výstupu je pak matice nepodobností, která obsahuje na diagonále nuly, viz výstup 7.4. Matici nepodobností na základě koeficientu φ získáme analogickým způsobem, jen v rámci možnosti *Count* zvolíme *Phi-square measure*. Odpovídající matice nepodobností je ve výstupu 7.5.

Výstup 7.4 | Matice nepodobností pro kategorie proměnné *B3_4kat* (míra chí-kvadrát)

Proximity Matrix	Chi-square between Sets of Frequencies			
	1: Stejný studijní obor	2: Jiný studijní obor na stejné vysoké škole	3: Stejný či obdobný obor na jiné vysoké škole	4: Jiný obor na jiné VŠ nebo žádný obor
1: Stejný studijní obor	,000	6,535	,703	4,615
2: Jiný studijní obor na stejné vysoké škole	6,535	,000	2,115	2,153
3: Stejný či obdobný obor na jiné vysoké škole	,703	2,115	,000	2,500
4: Jiný obor na jiné VŠ nebo žádný obor	4,615	2,153	2,500	,000

This is a dissimilarity matrix.

Výstup 7.5 | Matice nepodobností pro kategorie proměnné *B3_4kat* (koeficient φ)

Proximity Matrix	Phi-square between Sets of Frequencies			
	1: Stejný studijní obor	2: Jiný studijní obor na stejné vysoké škole	3: Stejný či obdobný obor na jiné vysoké škole	4: Jiný obor na jiné VŠ nebo žádný obor
1: Stejný studijní obor	,000	,284	,036	,218
2: Jiný studijní obor na stejné vysoké škole	,284	,000	,162	,142
3: Stejný či obdobný obor na jiné vysoké škole	,036	,162	,000	,263
4: Jiný obor na jiné VŠ nebo žádný obor	,218	,142	,263	,000

This is a dissimilarity matrix.

Vztah mezi kategoriemi *stejný studijní obor* a *jiný studijní obor na stejné vysoké škole* je pomocí míry chí-kvadrát ohodnocen číslem 6,535 a pomocí koeficientu φ číslem 0,284. To odpovídá výsledkům získaným bez systému *IBM SPSS Statistics*[11]. Jde vždy o největší vzdálenost v rámci dané matice, což znamená, že uvedené kategorie jsou nejvíce odlišné. Podle míry chí-kvadrát byla nejmenší hodnota zjištěna mezi kategoriemi *stejný studijní obor* a *stejný či obdobný obor na jiné vysoké škole* (0,703). Pomocí koeficientu φ bychom jako nejpodobnější vyhodnotili stejné kategorie (0,036). Na základě zjištěných vztahů bychom provedli spojení kategorií těchto dvou nejpodobnějších kategorií (viz proměnná *B3_3kat_b*).

11 Stejné tabulky lze získat též pomocí nabídek Analyze, Correlate, Distance.

Pro analýzu vztahů mezi kategoriemi proměnné *studijní obor vhodný pro první zaměstnání* na základě tabulky sdružených četností použijeme stejný postup, pouze v části *Cluster* vybereme možnost *Variables*. Výsledné matice jsou na výstupech 7.6 a 7.7.

Výstup 7.6 | Matice nepodobností pro kategorie proměnné *C3* (míra chí-kvadrát)

Proximity Matrix	Matrix File Input			
	Vystudovaný obor	**Příbuzný studijní obor**	**Zcela jiný studijní obor**	**Zaměstnání nevyžaduje oborovou specializaci**
Vystudovaný obor	,000	1,877	5,448	3,318
Příbuzný studijní obor	1,877	,000	6,364	2,785
Zcela jiný studijní obor	5,448	6,364	,000	3,573
Zaměstnání nevyžaduje oborovou specializaci	3,318	2,785	3,573	,000

Výstup 7.7 | Matice nepodobností pro kategorie proměnné *C3* (koeficient φ)

Proximity Matrix	Matrix File Input			
	Vystudovaný obor	**Příbuzný studijní obor**	**Zcela jiný studijní obor**	**Zaměstnání nevyžaduje oborovou specializaci**
Vystudovaný obor	,000	,088	,476	,277
Příbuzný studijní obor	,088	,000	,291	,126
Zcela jiný studijní obor	,476	,291	,000	,280
Zaměstnání nevyžaduje oborovou specializaci	,277	,126	,280	,000

Nejpodobnější jsou kategorie *vystudovaný obor* a *příbuzný studijní obor* (hodnota 1,877, resp. 0,088). Nejodlišnější jsou pomocí chí-kvadrát míry kategorie *příbuzný studijní obor* a *zcela jiný studijní obor* (hodnota 6,364) a pomocí koeficientu φ kategorie *vystudovaný obor* a *zcela jiný studijní obor* (hodnota 0,476). Počet kategorií bychom mohli snížit ze čtyř na tři tak, že bychom spojili kategorie *vystudovaný obor* a *příbuzný studijní obor*.

♦ ♦ ♦

V předchozím textu byly uvedeny příklady, jak by mohly být některé kategorie sloučeny. Je-li kategorií více, vhodnější než „ruční" vyhledávání nejnižších hodnot je grafické znázornění vztahů. Nejpoužívanějším grafem je zřejmě *dendrogram*, v němž se nejprve graficky spojí dva nejbližší (nejpodobnější) objekty (zde kategorie). V dalších krocích se vychází vždy z rozměrově menší matice vzdáleností, která vzniká vynecháním příslušných

dvou řádků a dvou sloupců a doplněním řádku a sloupce, obsahujících vzdálenosti ostatních objektů či shluků od právě vytvořeného shluku. V této matici se opět najde nejnižší hodnota a spojí se odpovídající shluky. Postup se opakuje až do spojení všech shluků do jediného.

Pro **výpočet vzdáleností mezi shluky** existují různé metody. V systému *IBM SPSS Statistics* mohou být s chí-kvadrát mírou nepodobnosti (resp. koeficientem φ) použity následující metody spojení[12]:

- *metoda průměrného spojení pro mezishlukové vzdálenosti*, kdy se vzdálenost mezi dvěma shluky spočte jako aritmetický průměr vzdáleností pro všechny dvojice objektů, z nichž jeden patří do prvního shluku a druhý do druhého shluku,
- *metoda průměrného spojení pro vnitroshlukové vzdálenosti*, kdy se objekty dvou uvažovaných shluků spojí do jednoho shluku a pak se spočte aritmetický průměr vzdáleností pro všechny dvojice objektů,
- *metoda jednoduchého spojení* (*nejbližšího souseda*), kdy je vzdálenost shluků dána minimální vzdáleností dvou objektů, z nichž jeden patří do prvního shluku a druhý do druhého shluku,
- *metoda úplného spojení* (*nejvzdálenějšího souseda*), pro kterou je určující maximální vzdálenost objektů patřících do různých shluků.

Příklad 7.2

Na základě dat matic nepodobností na výstupech 7.4 a 7.5 vytvořme pro kategorie proměnné *opětovný výběr oboru* schéma spojování pomocí *jednoduchého* a *úplného spojení*.

Nejprve spojíme dvě nejbližší kategorie podle *chí-kvadrát míry*, tj. *stejný studijní obor* a *stejný či obdobný obor na jiné vysoké škole* (hodnota 0,703). Poté bude matice nepodobností obsahovat pouze tři řádky a tři sloupce. Na základě metody jednoduchého spojení bude v políčku v prvním řádku a druhém sloupci hodnota 2,115, v políčku v prvním řádku a třetím sloupci hodnota 2,5 a v políčku ve druhém řádku a třetím sloupci hodnota 2,153. Nejmenší nenulová hodnota v nové matici je 2,115, která označuje odlišnost vytvořeného shluku od kategorie *jiný studijní obor ma stejné vysoké škole*. Vytvoří se nový shluk obsahující navíc tuto kategorii. Matice nepodobností tak již bude obsahovat pouze dva řádky a dva sloupce. Zbývá tedy připojit kategorii *jiný obor na jiné VŠ nebo žádný obor* (nepodobnost 2,153). Další postupy budou uvedeny pomocí programového systému.

IBM SPSS Statistics

Postupujeme stejně jako v příkladu 7.1 s tím, že v rámci možnosti *Statistics* specifikujeme, že požadujeme *Agglomeration schedule* (standardně nastaveno). Dále v části *Display* ponecháme volbu *Plots* a v rámci možnosti *Plots* zvolíme *Dendrogram* (v části *Icicle* zvolíme *None*). V rámci možnosti *Methods* v části *Cluster Method* vybereme *Nearest neighbor* (metodu nejbližšího souseda).

12 Systém IBM SPSS Statistics nabízí více metod spojení, další jsou popsány například v [19] a [37].

V upraveném výstupu 7.8 jsou popsány jednotlivé kroky spojování kategorií. Celkem bylo spojování provedeno ve třech krocích (sloupec *Stage*). Ve druhém a třetím sloupci (*Cluster Combined*) jsou uvedena pořadí kategorií – buď těch, které jsou spojovány samostatně, nebo těch, které jsou dle pořadí první v daném shluku (sloupce *Cluster 1* a *Cluster 2*). V jakých vzdálenostech je spojení provedeno, se uvádí ve sloupci *Coefficients*. V ostatních sloupcích jsou informace, ve kterých krocích se shluk vyskytl poprvé (*Stage Cluster First Appears*) a ve kterém následujícím kroku se bude vyskytovat (*Next Stage*).

Výstup 7.8 | Postup spojování kategorií proměnné *B3_4kat* (jednoduché spojení)

Agglomeration Schedule						
Stage	**Cluster Combined**		**Coefficients (chi-square)**	**Stage Cluster First Appears**		**Next Stage**
	Cluster 1	**Cluster 2**		**Cluster 1**	**Cluster 2**	
1	1	3	,703	0	0	2
2	1	2	2,115	1	0	3
3	1	4	2,153	2	0	0

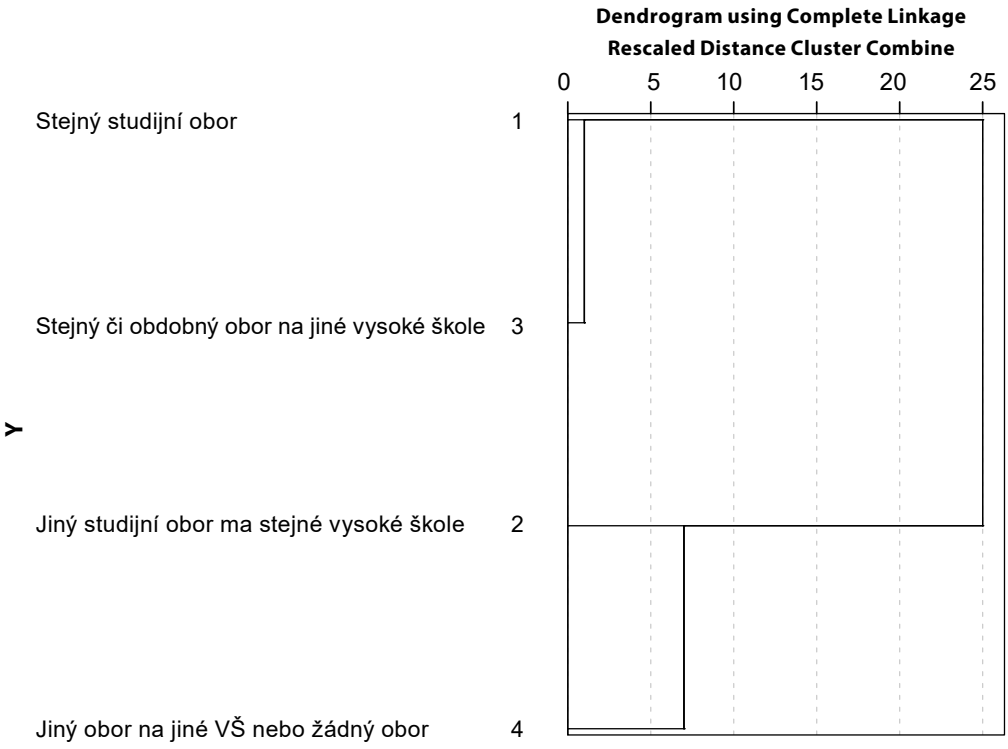
Postup spojování kategorií na základě chí-kvadrát míry s využitím *metody úplného spojení* je uveden v upraveném výstupu 7.9. Pro tento typ spojení vybereme položku *Furthest neighbor* (ve výstupu je uvedeno *Complete Linkage*). V prvním kroku se vytvoří shluk z první a třetí kategorie a ve druhém kroku shluk ze druhé a čtvrté kategorie. Tyto dva shluky se spojí ve třetím kroku.

Výstup 7.9 | Postup spojování kategorií proměnné *B3_4kat* (úplné spojení)

Agglomeration Schedule						
Stage	**Cluster Combined**		**Coefficients (chi-square)**	**Stage Cluster First Appears**		**Next Stage**
	Cluster 1	**Cluster 2**		**Cluster 1**	**Cluster 2**	
1	1	3	,703	0	0	3
2	2	4	2,153	0	0	3
3	1	2	6,535	1	2	0

Výstup 7.10 obsahuje dendrogram. V nadpisu grafu je uvedeno, že jde o metodu úplného spojení spojení (*Complete Linkage*). Systém *IBM SPSS Statistics* používá v dendrogramu transformované vzdálenosti (*Rescaled Distance*) na škále od 0 do 25.

Výstup 7.10 Dendrogram spojování kategorií proměnné *B3_4kat* (úplné spojení)



S pomocí *koeficientu φ* a obou typů spojení obdržíme stejné posloupnosti spojování jako při použití chí-kvadrát míry a úplného spojení, viz postupy spojování v upravených výstupech 7.11 a 7.12.

Výstup 7.11 | Postup spojování kategorií proměnné *B3_4kat* (jednoduché spojení)

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients (phi-square)	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	3	,036	0	0	3
2	2	4	,142	0	0	3
3	1	2	,162	1	2	0

Výstup 7.12 | Postup spojování kategorií proměnné *B3_4kat* **(úplné spojení)**

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients (phi-square)	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	3	,036	0	0	3
2	2	4	,142	0	0	3
3	1	2	,284	1	2	0

Příklad 7.3

Na základě dat matice nepodobností ve výstupu 7.6 vytvořme pro kategorie proměnné *studijní obor vhodný pro první zaměstnání* schéma spojování metodami *průměrného spojení* pro meziskupinové a vnitroskupinové vzdálenosti.

Nejprve spojíme dvě nejbližší kategorie podle *chí-kvadrát míry*, tj. *vystudovaný obor* a *příbuzný studijní obor* se vzdáleností 1,877. Musí být přepočteny vzdálenosti mezi nově vzniklým shlukem a zbylými kategoriemi. Pomocí *meziskupinových vzdáleností* se v prvním případě spočte aritmetický průměr ze vzdáleností dvojic (*vystudovaný obor – zcela jiný studijní obor*) a (*příbuzný studijní obor – zcela jiný studijní obor*), tj. (5,448 + 6,364) / 2 = 5,906, ve druhém případě průměr ze vzdáleností (*vystudovaný obor – zaměstnání nevyžaduje oborovou specializaci*) a (*příbuzný studijní obor – zaměstnání nevyžaduje oborovou specializaci*), tj. (3,318 + 2,785) / 2 = 3,052. V nově vzniklé matici je nejmenší hodnotou 3,052, vyjadřující vzdálenost mezi shlukem vytvořeným v prvním kroku a čtvrtou kategorií. Vytvoříme tedy nový shluk z první, druhé a čtvrté kategorie. Zbývá připojit k tomuto shluku třetí kategorii. Výsledná vzdálenost bude vyjádřena jako průměr ze vzdáleností pro tři dvojice kategorií, z nichž jedna bude vždy třetí kategorie, tj. (5,448 + 6,364 + 3,573) / 3 = 5,128.

Při použití *vnitroskupinových vzdáleností* začínáme stejně, tj. spojením prvních dvou kategorií. Při výpočtech vzdáleností mezi nově vzniklým shlukem a zbylými kategoriemi vezmeme v úvahu navíc vzdálenost prvních dvou kategorií. Ze získaných vzdáleností bude nejmenší pro vytvořený shluk a čtvrtou kategorii, tj. (1,877 + 3,318 + 2,785) / 3 = 2,66. Zbývá tedy ke shluku vytvořenému z první, druhé a čtvrté kategorie připojit třetí kategorii. Výsledná vzdálenost bude (1,877 + 5,448 + 3,318 + 6,364 + 2,785 + 3,573) / 6 = 3,894.

IBM SPSS Statistics

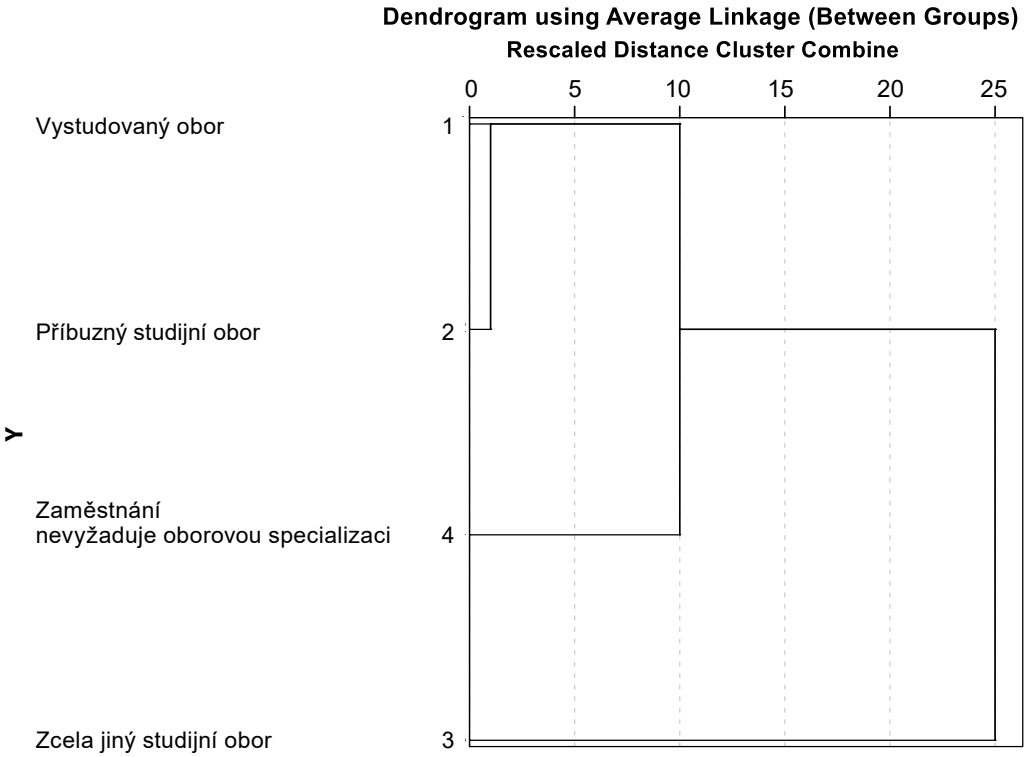
V systému *IBM SPSS Statistics* postupujeme analogicky jako v příkladu 7.2 s tím, že v části *Cluster* vybereme možnost *Variables*. Pro průměrné spojení s mezishlukovými vzdálenostmi vybereme *Between-groups linkage* (pro vnitroshlukové vzdálenosti bychom vybrali *Within-groups linkage*). Ve výstupu je uvedeno *Average Linkage (Between Groups)*. Výstup 7.13 popisuje postup spojování pro mezishlukové vzdálenosti. Hodnoty uvedené

ve sloupci *Coefficients* se shodují s výsledky získanými bez použití programu. Ve výstupu 7.14 je zobrazen dendrogram vytvořený na základě metody průměrného spojení pro meziskupinové vzdálenosti.

Výstup 7.13 | Postup spojování kategorií proměnné *C3* (průměrné spojení mezi shluky)

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	2	1,877	0	0	2
2	1	4	3,052	1	0	3
3	1	3	5,128	2	0	0

Výstup 7.14 Dendrogram spojování kategorií proměnné C3 (průměrné spojení mezi shluky)



7.2 Vícerozměrné škálování

Matice vzdáleností, popsaná v předchozím oddílu, může být využita též při vícerozměrném škálování. To je určeno především k redukci počtu proměnných, výsledkem je projekce objektů do prostoru nižší dimenze. Při projekci do dvou či tří dimenzí mohou být všechny vztahy znázorněny graficky (v případě více dimenzí jsou do grafu vybrány jen některé z nich).

Vícerozměrné škálování lze však využít též pro znázornění vztahů mezi objekty, případně kategoriemi. Každá sledovaná kategorie pak může být zobrazena jako bod ve dvourozměrném prostoru. Podle vzdálenosti těchto bodů můžeme usuzovat na podobnost daných kategorií. Čím jsou body bližší, tím je větší podobnost mezi odpovídajícími kategoriemi. Dále nás zajímají skupiny podobných kategorií vzhledem k poloze bodů vůči hlavním osám (procházejících nulou). Ve dvourozměrném prostoru můžeme rozlišit buď dvě skupiny, a to jednak podle svislé osy (*Y*), jednak podle vodorovné (*X*), nebo skupiny čtyři podle jednotlivých kvadrantů.

Matematické odvození předpokládá podstatně hlubší znalosti týkající se operací s maticemi. Z toho důvodu bude v tomto oddílu uvedena pouze ukázka získání hodnot pro jednotlivé souřadnice a grafické zobrazení pomocí programového systému. Způsoby výpočtů jsou popsány například v [11] a [19].

Příklad 7.4

Pomocí vícerozměrného škálování zobrazme ve dvourozměrném prostoru kategorie proměnné *studijní obor vhodný pro první zaměstnání*, a to pomocí koeficientu φ na základě kategorií proměnné *opětovný výběr oboru*. Stejně jako u shlukové analýzy, popsané v předchozím oddílu, vyjdeme z tabulky sdružených absolutních četností (viz výstup 7.6), zkopírované do nové tabulky datového editoru *IBM SPSS Statistics*.

V *IBM SPSS Statistics* pro analýzu volíme *Analyze*, *Scale*, *Multidimensional Scaling (PROXSCAL)*, v části *Data Format* zvolíme možnost *Create Proximities from Data* a potvrdíme pomocí *Define*. Pak zadáme všechny „proměnné" (kategorie proměnné *studijní obor vhodný pro první zaměstnání*) a v rámci možnosti *Measure* specifikujeme výpočet matice nepodobností zcela stejně, jak bylo popsáno v předchozím oddílu u shlukové analýzy. To znamená, že v části *Measure* zvolíme možnost *Count* a míru *Phi-square measure*)[13].

Ve dvourozměrném prostoru se standardně zobrazují proměnné, tj. v našem případě kategorie sloupcové proměnné (pro zobrazení kategorií řádkové proměnné je potřeba v části *Measure* zvolit pro výpočet matice vzdáleností (*Create Distance Matrix*) řádky, tj. možnost *Between cases*). Výsledkem jsou jednak souřadnice kategorií uvedené v tabulce (viz výstup 7.15), jednak vlastní grafické znázornění kategorií jako bodů v rovině, viz graf 7.1. V grafu

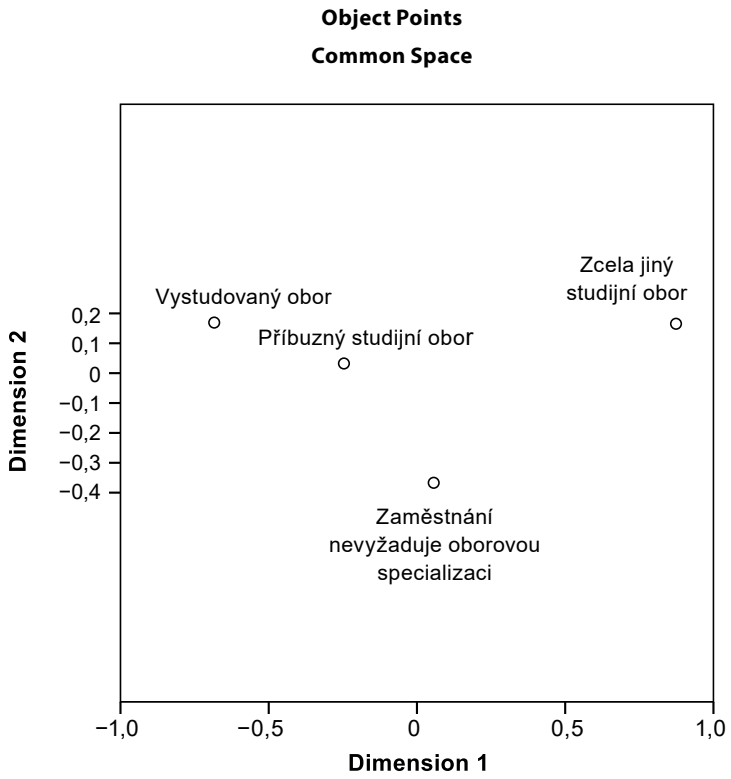
13 Stejné výstupy bychom obdrželi, pokud bychom vyšli přímo z matice nepodobností, tj. do nové tabulky datového editoru bychom zkopírovali matici nepodobností, získanou například jakou součást výstupu shlukové analýzy. V části *Data Format* bychom ponechali standardně nastavenou možnost *The data are proximities.*

můžeme rozlišit dvě skupiny kategorií, které se nacházejí v oblastech oddělených vertikálně podle hodnoty 0 na ose *X* (*Dimension 1*). V jedné se nacházejí kategorie *vystudovaný obor* a *příbuzný studijní obor* (záporné hodnoty v první dimenzi) a ve druhé *zbylé kategorie*. Horizontálně podle hodnoty 0 na ose *Y* (*Dimension 2*) je oddělena kategorie *zaměstnání nevyžaduje oborovou specializaci* (záporná hodnota ve druhé dimenzi).

Výstup 7.15 | Hodnoty souřadnic ve dvourozměrném prostoru pro *studijní obor vhodný pro první zaměstnání* (koeficient φ)

Final Coordinates	Dimension	
	1	2
Vystudovaný obor	–,684	,169
Příbuzný studijní obor	–,246	,032
Zcela jiný studijní obor	,874	,165
Zaměstnání nevyžaduje oborovou specializaci	,056	–,367

Graf 7.1 | Bodový graf zobrazující výsledek vícerozměrného škálování (koeficient φ)



7.3 Korespondenční analýza

Jak již bylo zmíněno v úvodu, existuje speciální metoda pro zjišťování skupin podobných kategorií, a to *korespondenční analýza*. Na rozdíl od shlukové analýzy a vícerozměrného škálování, jejichž principy byly naznačeny v předchozích oddílech, vychází tato metoda přímo ze zdrojové datové matice. To znamená, že v systému *IBM SPSS Statistics* není třeba vytvářet kontingenční tabulku a kopírovat ji do datového editoru. Dále při použití dvou proměnných umožňuje korespondenční analýza znázorňovat souvislost kategorií obou proměnných současně. Navíc lze tento přístup rozšířit na analýzu více než dvou proměnných.

Stejně jako vícerozměrné škálování, také korespondenční analýza umožňuje zobrazovat kategorie v redukovaném souřadném systému, který se nazývá *korespondenční mapa*. Vztahy mezi kategoriemi dvou proměnných zkoumá *jednoduchá korespondenční analýza*, vztahy mezi kategoriemi více než dvou proměnných pak *vícenásobná korespondenční analýza*.

V případě *dvou proměnných* je základem dvourozměrná tabulka sdružených relativních četností, viz schéma 4.2. Při matematickém odvození se vychází z hodnot této tabulky, reprezentovaných jako *korespondenční matice*, která bude značena symbolem **P** a její prvky p_{ij}, kde $i = 1, 2, \ldots, R$ a $j = 1, 2, \ldots, S$.

Řádkové marginální relativní četnosti p_{i+} se nazývají *řádkové zátěže*; v dalším textu budou značeny jako r_i (podle anglického výrazu „row“). Obdobně sloupcové marginální relativní četnosti p_{+j} se nazývají *sloupcové zátěže* a v dalším textu budou značeny jako c_i (podle anglického výrazu „column“). Řádkové relativní četnosti (viz tabulka 4.1) se nazývají *řádkové profily*, obdobně sloupcové relativní četnosti představují *sloupcové profily*.

Označme R-členný vektor řádkových zátěží symbolem **r** a S-členný vektor sloupcových zátěží jako **c**. Dále označme matici řádkových profilů symbolem **R** a matici sloupcových profilů jako **C**. Tyto matice lze vyjádřit pomocí vztahů

$$\mathbf{R} = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P} = \begin{bmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \vdots \\ \mathbf{r}_R^T \end{bmatrix}, \tag{7.3}$$

kde $\mathbf{D}_{\mathbf{r}}$ je diagonální matice s prvky vektoru **r** na diagonále, $\mathbf{r}_i = [p_{1\,i}, p_{2\,i}, \ldots, p_{Si}]$,

$$\mathbf{C} = \mathbf{D}_{\mathbf{c}}^{-1}\mathbf{P}^T = \left[\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_S\right], \tag{7.4}$$

kde $\mathbf{D}_{\mathbf{c}}$ je diagonální matice s prvky vektoru **c** na diagonále, $\mathbf{c}_j = [p_{1\,j}, p_{2\,j}, \ldots, p_{Rj}]$.

Celou *korespondenční tabulku* pak můžeme schematicky vyjádřit jako

$$\begin{bmatrix} \mathbf{P} & \mathbf{r} \\ \mathbf{c}^T & 1 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1S} & r_1 \\ p_{21} & p_{22} & \cdots & p_{2S} & r_2 \\ \vdots & \vdots & & \vdots & \vdots \\ p_{R1} & p_{R2} & \cdots & p_{RS} & r_R \\ c_1 & c_2 & \cdots & c_S & 1 \end{bmatrix}.$$

Pro vektor řádkových zátěží platí, že

$$\mathbf{r} = \sum_{j=1}^{S} p_{+j}\mathbf{c}_j \ ,$$

a pro vektor sloupcových zátěží platí vztah

$$\mathbf{c} = \sum_{i=1}^{R} p_{i+}\mathbf{r}_i$$

(je použito značení řádkových a sloupcových relativních četností podle tabulky 4.1).

Stejně jako v případě shlukové analýzy či vícerozměrného škálování se také v případě korespondenční analýzy vyjadřují nepodobnosti mezi kategoriemi řádkové, resp. sloupcové proměnné. Lze samozřejmě použít *chí-kvadrát míru*, tak jak bylo popsáno v oddílu 7.1 v souvislosti se shlukovou analýzou, viz vztah (7.1). Pomocí nově zavedené symboliky zapíšeme pro *i*-tou a *i'*-tou řádkovou kategorii tuto míru nepodobnosti jako

$$\sqrt{\chi^2} = \sqrt{\sum_{j=1}^{S} \frac{(r_{ij} - r_{i'j})^2}{c_j}} \ , \tag{7.5}$$

kde r_{ij}, resp. $r_{i'j}$ jsou prvky matice řádkových profilů **R** a c_j jsou prvky vektoru sloupcových zátěží **c**. Analogicky se postupuje při výpočtu nepodobnosti mezi sloupcovými kategoriemi.

Cílem korespondenční analýzy je redukovat vícerozměrný prostor vektorů řádkových a sloupcových profilů do prostoru menší dimenze. Obvykle se uvažuje dvourozměrný prostor, tj. rovina. Bod roviny, který je nejblíže určitému bodu ve vícerozměrném prostoru, se nazývá *projekce*. V korespondenční analýze se hledají souřadnice bodů, které nejlépe reprezentují původní data. Řešení vychází z matice standardizovaných reziduí **Z** s prvky

$$z_{ij} = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}$$

a jejího singulárního rozkladu podle vztahu

$$\mathbf{Z} = \mathbf{U} \cdot \mathbf{\Gamma} \cdot \mathbf{V}^{\mathrm{T}}, \tag{7.6}$$

kde $\mathbf{\Gamma}$ je diagonální matice a kde platí, že $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, což je jednotková matice (s jedničkami na diagonále). Podrobněji viz například [19] a [28].

Před vlastním odhadem souřadnic jednotlivých kategorií se provádí volba *normalizační metody*, tj. způsobu zobrazení bodů v korespondenční mapě. Pro upřednostnění vztahů mezi řádkovými kategoriemi je určena *analýza řádkových profilů*, k preferenci vztahů mezi sloupcovými kategoriemi pak *analýza sloupcových profilů*. Dále existují přístupy umožňující vzájemně srovnávat řádkové i sloupcové kategorie. V systému *IBM SPSS Statistics* je k tomu určena například *symetrická normalizace*.

Při *analýze řádkových profilů* se souřadnice řádkových kategorií nacházejí ve sloupcích matice **F**, která se spočte podle vztahu

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Gamma}, \tag{7.7}$$

a souřadnice sloupcových kategorií ve sloupcích matice **Y,** dané vztahem

$$\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}\,. \tag{7.8}$$

Obdobně při *analýze sloupcových profilů* lze souřadnice sloupcových kategorií nalézt ve sloupcích matice **G**, která se spočte podle vztahu

$$\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Gamma}, \tag{7.9}$$

a souřadnice řádkových kategorií ve sloupcích matice **X**, dané vztahem

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}\,. \tag{7.10}$$

Při *simultánní analýze* řádkových i sloupcových profilů (možnost *Principal* v systému *IBM SPSS Statistics*) se souřadnice řádkových kategorií nacházejí ve sloupcích matice **F**, která se spočte podle vztahu (7.7), a souřadnice sloupcových kategorií ve sloupcích matice **G,** dané vztahem (7.9).

Pokud jde o interpretaci umístění bodů v korespondenční mapě, pak platí to, co již bylo uvedeno v předchozím oddílu v souvislosti s vícerozměrným škálováním. Čím jsou body bližší, tím je větší podobnost mezi odpovídajícími kategoriemi. Dále se často podaří interpretovat skupiny podobných kategorií vzhledem k poloze bodů vůči hlavním osám, zejména vůči vertikální ose *Y*.

Příklad 7.5

S využitím systému *IBM SPSS Categories* zobrazme korespondenční mapy pro kategorie proměnných *opětovný výběr oboru* (*B3_4kat*) a *studijní obor pro první zaměstnání* (*C3*). Příslušná procedura v *SPSS* vyžaduje, aby kategorie byly zaznamenány pomocí číselných kódů. Textově kódované proměnné lze převést na číselně kódované pomocí nabídek *Transform* a *Automatic Recode*.

Zadání pro korespondenční analýzu lze provést po zvolení *Analyze*, *Dimension Reduction*, *Correspondence Analysis*. Specifikujeme řádkovou (*Row*) a sloupcovou (*Column*) proměnnou a pro každou z nich zadáme rozpětí (*Define Range*), v našem případě je minimální

hodnota 1 a maximální 4 (je potřeba zvolit *Update*). V rámci možnosti *Model* nejprve zadáme *analýzu řádkových profilů* (*Row principal* v části *Normalization Method*). V rámci *Statistics* specifikujeme výpis řádkových a sloupcových profilů (*Row profiles*, *Column profiles*) a v rámci *Plots* v části *Scatterplots* přidáme grafy *Row points* a *Column points*.

Korespondenční tabulka absolutních četností je ve výstupu 7.16. Obsahově je shodná s kontingenční tabulkou uvedenou ve výstupu 7.1.

Výstup 7.16 | Korespondenční tabulka absolutních četností k příkladu 7.5

Correspondence Table					
Opětovný výběr oboru	**Studijní obor vhodný pro první zaměstnání**				
	Vystudovaný obor	**Příbuzný studijní obor**	**Zcela jiný studijní obor**	**Zaměstnání nevyžaduje oborovou specializaci**	**Active Margin**
Stejný studijní obor	42	264	22	47	375
Jiný studijní obor na stejné vysoké škole	10	84	39	23	156
Stejný či obdobný obor na jiné vysoké škole	2	11	1	1	15
Jiný obor na jiné VŠ nebo žádný obor	2	43	13	17	75
Active Margin	56	402	75	88	621

Dále je výsledkem výše uvedeného zadání tabulka řádkových profilů (viz výstup 7.17) a tabulka sloupcových profilů (viz výstup 7.18). Ve srovnání například s výstupem procedury CROSSTABS pro analýzu kontingenčních tabulek nejsou relativní četnosti uváděny v procentech.

Korespondenční mapa je uvedena třikrát (jsou automaticky kráceny názvy kategorií). V první jsou zobrazeny řádkové kategorie (viz graf 7.2), ve druhé sloupcové kategorie a ve třetí kategorie obou proměnných současně (viz graf 7.4). V grafu 7.2 je zachycena největší podobnost kategorií *stejný studijní obor* a *stejný či obdobný obor na jiné vysoké škole*, které jsou podle vertikální osy odlišeny od zbylých dvou kategorií, což odpovídá dendrogramu ve výstupu 7.10.

Graf 7.4 zachycuje vzájemné vztahy kategorií obou proměnných. Vertikální osa odděluje kategorie vztahující se k vystudovanému oboru či příbuznému studijnímu oboru od kategorií týkajících se jiného oboru nebo netýkajících se žádného oboru.

Pro *analýzu sloupcových profilů* v rámci možnosti *Model* zadáme *Column principal* (v části *Normalization Metod*). Výsledkem jsou opět tři korespondenční mapy. Sloupcové kategorie jsou zobrazeny v grafu 7.3, v němž jsou zachyceny obdobné vztahy jako pomocí vícerozměrného škálování (graf 7.1).

Výstup 7.17 | Tabulka řádkových profilů k příkladu 7.5

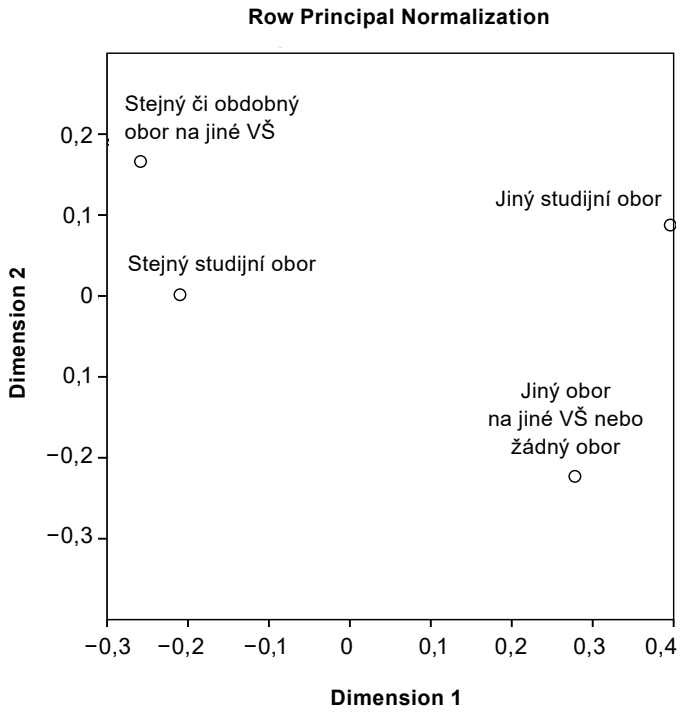
Row Profiles					
Opětovný výběr oboru	Studijní obor vhodný pro první zaměstnání				
	Vystudovaný obor	Příbuzný studijní obor	Zcela jiný studijní obor	Zaměstnání nevyžaduje oborovou specializaci	Active Margin
Stejný studijní obor	,112	,704	,059	,125	1,000
Jiný studijní obor na stejné vysoké škole	,064	,538	,250	,147	1,000
Stejný či obdobný obor na jiné vysoké škole	,133	,733	,067	,067	1,000
Jiný obor na jiné VŠ nebo žádný obor	,027	,573	,173	,227	1,000
Mass	,090	,647	,121	,142	

Výstup 7.18 | Tabulka sloupcových profilů k příkladu 7.5

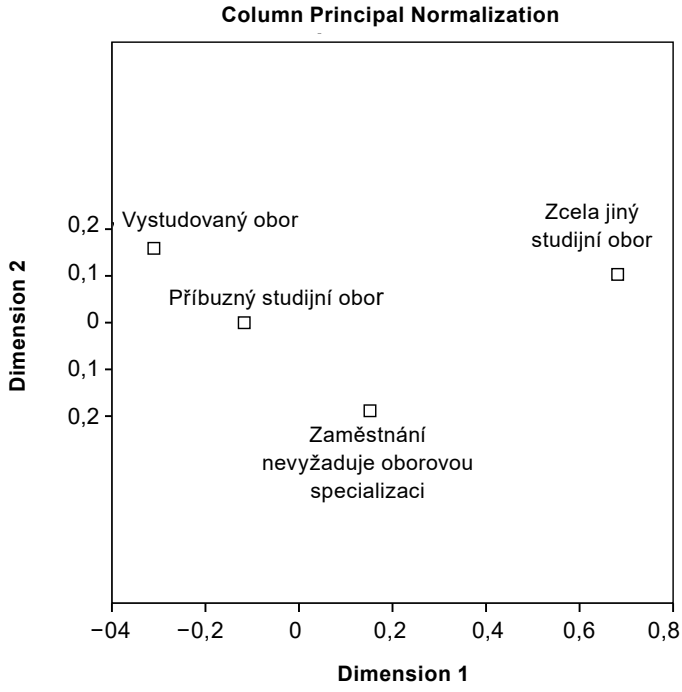
Column Profiles					
Opětovný výběr oboru	Studijní obor vhodný pro první zaměstnání				
	Vystudovaný obor	Příbuzný studijní obor	Zcela jiný studijní obor	Zaměstnání nevyžaduje oborovou specializaci	Mass
Stejný studijní obor	,750	,657	,293	,534	,604
Jiný studijní obor na stejné vysoké škole	,179	,209	,520	,261	,251
Stejný či obdobný obor na jiné vysoké škole	,036	,027	,013	,011	,024
Jiný obor na jiné VŠ nebo žádný obor	,036	,107	,173	,193	,121
Active Margin	1,000	1,000	1,000	1,000	

Vzájemné vztahy kategorií jsou zachyceny v grafu 7.5. Stejně jako v grafu 7.4 vertikální osa odděluje kategorie vztahující se k vystudovanému oboru či příbuznému studijnímu oboru od kategorií týkajících se jiného oboru nebo netýkajících se žádného oboru.

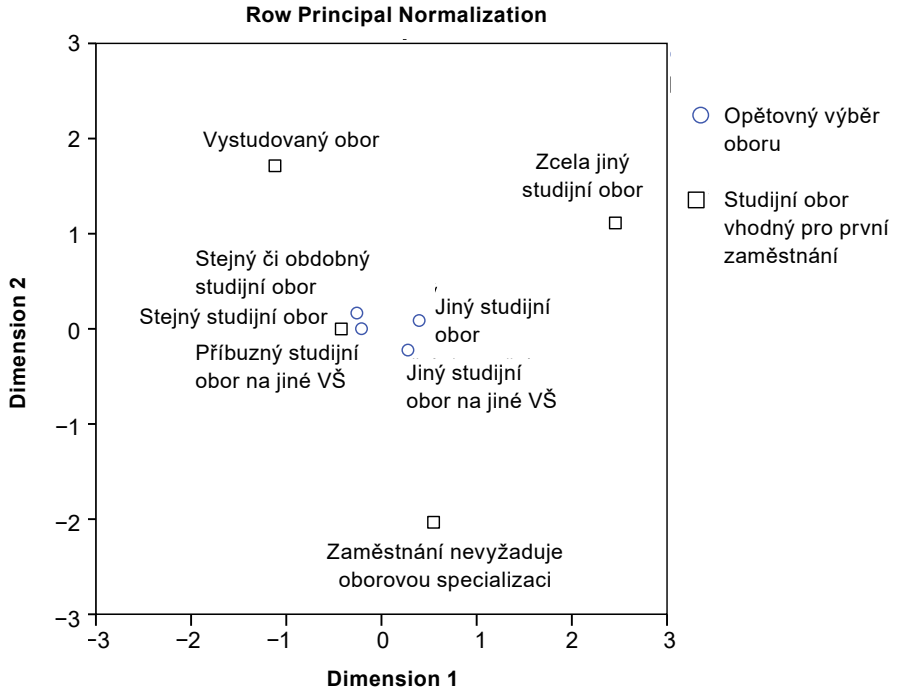
Graf 7.2 | Korespondenční mapa – analýza řádkových profilů (řádkové kategorie)



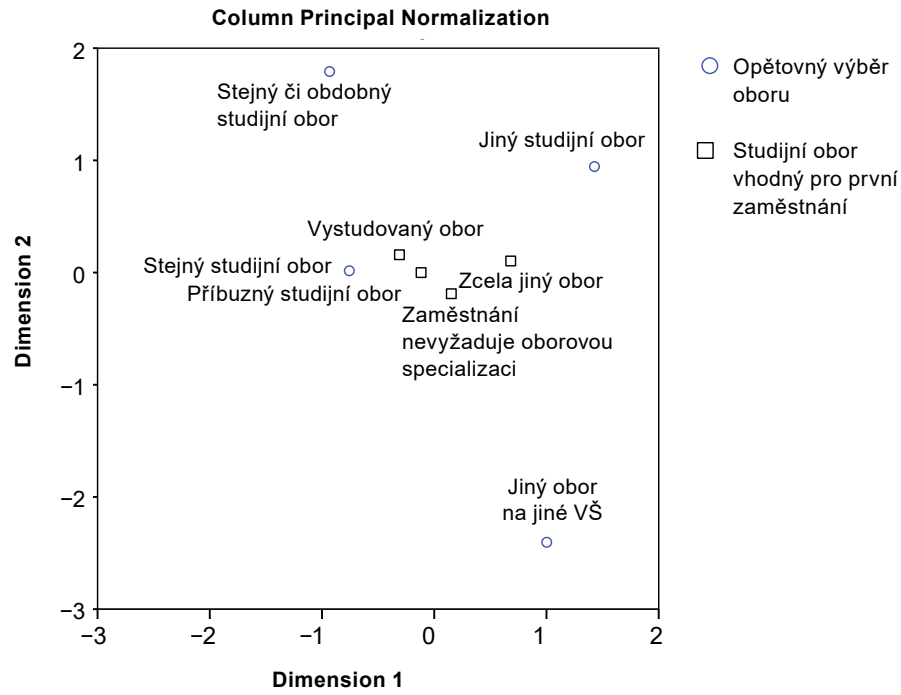
Graf 7.3 | Korespondenční mapa – analýza sloupcových profilů (sloupcové kategorie)



Graf 7.4 | Korespondenční mapa – analýza řádkových profilů (všechny kategorie)



Graf 7.5 | Korespondenční mapa – analýza sloupcových profilů (všechny kategorie)

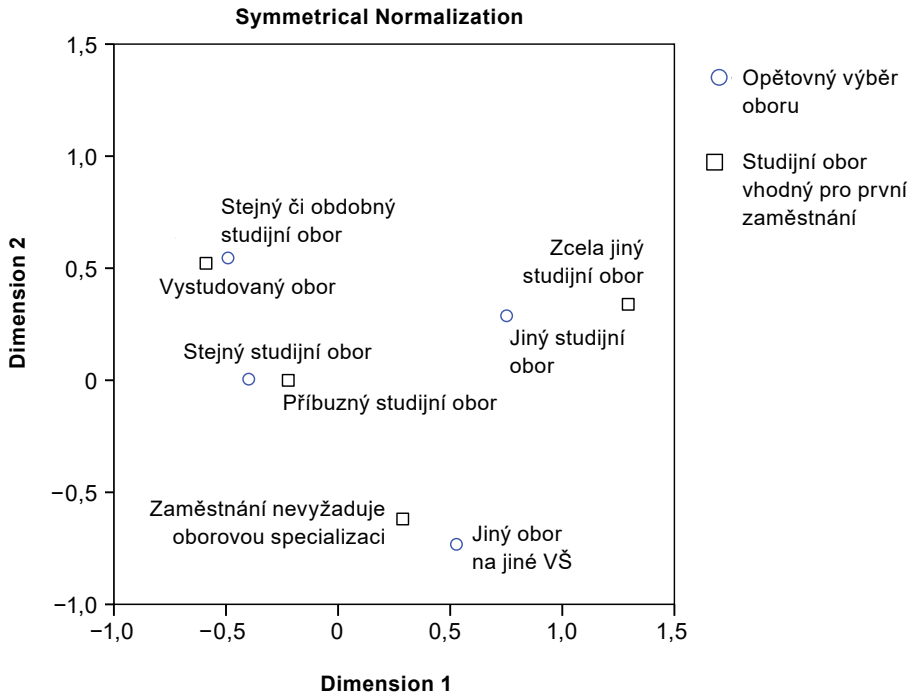


Pokud v rámci možnosti *Model* zadáme *symetrickou normalizaci* (možnost *Symmetrical* v části *Normalization Method*), pak dostaneme zobrazení všech kategorií pomocí korespondenční mapy v grafu 7.6. Je zřejmé rozlišení čtyř skupin kategorií: *stejný či obdobný obor na jiné vysoké škole* a *vystudovaný obor*, *stejný studijní obor* a *příbuzný studijní obor*, *jiný studiní obor na stejné vysoké škole* a *jiný studijní obor*, *jiný obor na jiné vysoké škole nebo žádný obor* a *zaměstnání nevyžaduje oborovou specializaci*.

♦ ♦ ♦

V systému *IBM SPSS Statistics* je implementována také *vícenásobná korespondenční analýza*, která je dostupná pomocí nabídek *Analyze*, *Data Reduction* a *Optimal Scaling*. Její grafické výstupy znázorňují především vztahy mezi proměnnými, případně mezi objekty. Vztahy mezi kategoriemi jsou zobrazovány zvlášť pro jednotlivé proměnné. Pokud bychom analyzovali kategorie tří proměnných, výstup by obsahoval tři korespondenční mapy odpovídající těmto proměnným. Z nich lze vyčíst jednak větší či menší podobnost kategorií, jednak skupiny podobných kategorií vzhledem k jejich polohám k hlavním osám. Pokud by byly projekce kategorií prvních dvou proměnných zaneseny do jednoho grafu, jejich uspořádání by se podobalo rozmístění kategorií v grafu 7.6.

Graf 7.6 | Korespondenční mapa – symetrická normalizace (všechny kategorie)



Příloha

Vyhledávání kvantilů a hodnot pravděpodobnostní a distribuční funkce vybraných pravděpodobnostních rozdělenípomocí systému IBM SPSS Statistics

I když pro tyto hodnoty není třeba žádný datový soubor, systém *IBM SPSS Statistics* vyžaduje pro jakoukoli operaci vstupní data. Stačí tedy zadat jednu hodnotu do prvního políčka tabulky a pak je již možno zjišťovat kvantily, a to v rámci procedury ***Transform***, ***Compute Variable***. Ta vyžaduje zadat jméno cílové proměnné (sloupce, v jehož prvním políčku se bude požadovaná hodnota zobrazovat) v části *Target Variable* a vlastní výraz v části *Numeric Expression*. Tento výraz je tvořen funkcí, kterou je možné vybrat v části *Functions*. Součástí funkce jsou její parametry, které je třeba zadat oddělené čárkou. Desetinná místa se oddělují desetinnou tečkou.

A. Binomické rozdělení

a) *Hodnoty pravděpodobnostní funkce*

Pravděpodobnost, že náhodná veličina s binomickým rozdělením nabude určité hodnoty, zjistíme pomocí funkce PDF.BINOM, která má syntaxi

PDF.BINOM(*quant*,*n*,*prob*),

kde *quant* je hodnota náhodné veličiny (určitý počet výskytů náhodného jevu), jejíž pravděpodobnost chceme znát, *n* je počet náhodných pokusů a *prob* je pravděpodobnost výskytu jednoho určitého jevu. Pokud má náhodná veličina binomické rozdělení s parametry $\pi = 0{,}5$ a $n = 5$, pak pravděpodobnost, že se náhodný jev vyskytne jedenkrát, získáme zadáním funkce

PDF.BINOM(1,5,0.5).

Výsledkem bude hodnota 0,156.

b) *Hodnoty distribuční funkce*

Pravděpodobnost, že náhodná veličina s binomickým rozdělením nabude určité hodnoty nebo hodnoty menší, zjistíme pomocí funkce CDF.BINOM, která má syntaxi

CDF.BINOM(*quant*,*n*,*prob*),

kde význam parametrů je shodný jako v předchozím případě. Pokud má náhodná veličina binomické rozdělení s parametry = 0,5 a $n = 5$, pak pravděpodobnost, že se náhodný jev vyskytne nejvýše jedenkrát (tj. nevyskytne se nebo vyskytne jedenkrát), získáme zadáním funkce

CDF.BINOM(1,5,0.5).

Výsledkem bude hodnota 0,188.

B. Hypergeometrické rozdělení

a) *Hodnoty pravděpodobnostní funkce*
Pravděpodobnost, že náhodná veličina s hypergeometrickým rozdělením nabude určité hodnoty, zjistíme pomocí funkce PDF.HYPER, která má syntaxi

PDF.HYPER(*quant*,*total*,*sample*,*hits*),

kde *quant* je hodnota náhodné veličiny (určitý počet výskytů náhodného jevu), jejíž pravděpodobnost chceme znát. Náhodná veličina s hypergeometrickým rozdělením vyjadřuje počet výskytů náhodného jevu při výběrech bez vracení. Dalšími parametry jsou počet prvků, ze kterých vybíráme (*total*), počet vybraných prvků (*sample*) a počet příznivých jevů (*hits*). Pokud vybíráme z 11 prvků, z nichž 6 má vlastnost, na kterou se zaměřujeme, a provedeme 7 výběrů, pak pravděpodobnost, že se náhodný jev vyskytne pětkrát, získáme zadáním funkce

PDF.HYPER(5,11,7,6).

Výsledkem bude hodnota 0,182.

b) *Hodnoty distribuční funkce*
Pravděpodobnost, že náhodná veličina s binomickým rozdělením nabude určité hodnoty nebo hodnoty menší, zjistíme pomocí funkce CDF.HYPER, která má syntaxi

CDF.HYPER(*quant*, *total*,*sample*,*hits*),

kde význam parametrů je shodný jako v předchozím případě. Pokud vybíráme z 11 prvků, z nichž 6 má vlastnost, na kterou se zaměřujeme, a provedeme 7 výběrů, pak pravděpodobnost, že se náhodný jev vyskytne nejvýše pětkrát, získáme zadáním funkce

CDF.HYPER(5,11,7,6).

Výsledkem bude hodnota 0,985.

C. Normální rozdělení

a) *Kvantily*
Pro zjištění kvantilů normálního rozdělení použijeme funkci IDF.NORMAL, která má syntaxi

IDF.NORMAL(*prob*,*mean*,*stddev*),

kde *prob* je číslo z intervalu $\langle 0; 1\rangle$, které určuje, jaký kvantil požadujeme, *mean* je střední hodnota normálního rozdělení a *stddev* je směrodatná odchylka normálního rozdělení. Parametry jsou tedy odděleny čárkou a místo desetinné čárky je třeba použít desetinnou tečku. Například pokud požadujeme 95% kvantil normovaného normálního rozdělení, které má střední hodnotu 0 a směrodatnou odchylku 1, zadáme

IDF.NORMAL(0.95,0,1).

Výsledkem bude hodnota 1,64.

b) *Hodnoty distribuční funkce*
Příslušná funkce má název CDF.NORMAL a syntaxi

CDF.NORMAL(*quant*,*mean*,*stddev*),

kde *quant* je hodnota, pro kterou chceme zjistit hodnotu distribuční funkce. Například chceme-li znát, kolikaprocentní kvantil je hodnota 1,96 pro normální rozdělení se střední hodnotou 0 a směrodatnou odchylkou 1, zadáme

CDF.NORMAL(1.96,0,1)

a obdržíme výsledek 0,975, což znamená, že jde o 97,5% kvantil.

D. Studentovo t rozdělení

a) *Kvantily*
Zajímají-li nás kvantily Studentova t rozdělení, použijeme funkci IDF.T, která má syntaxi

IDF.T(*prob*,*df*),

kde *prob* je číslo z intervalu ⟨0; 1⟩, které určuje, jaký kvantil požadujeme, a *df* je počet stupňů volnosti. Například pokud požadujeme 95% kvantil Studentova t rozdělení se 100 stupni volnosti, zadáme

IDF.T(0.95,100).

Výsledkem bude hodnota 1,66.

b) *Hodnoty distribuční funkce*
Příslušná funkce má název CDF.T a syntaxi

CDF.T(*quant*,*df*),

kde *quant* je hodnota, pro kterou chceme zjistit hodnotu distribuční funkce, a *df* je počet stupňů volnosti. Například chceme-li zjistit, kolikaprocentní kvantil je hodnota 1,66 pro Studentovo t rozdělení se 100 stupni volnosti, zadáme

CDF.T(1.66,100).

Obdržíme výsledek 0,95, jde tedy o 95% kvantil.

E. Chí-kvadrát rozdělení

a) *Kvantily*

Zajímají-li nás kvantily chí-kvadrát rozdělení, použijeme funkci IDF.CHISQ, která má syntaxi

IDF.CHISQ(*prob*;*df*),

kde *prob* je číslo z intervalu ⟨0; 1⟩, které určuje, jaký kvantil požadujeme, a *df* je počet stupňů volnosti. Například pokud požadujeme 95% kvantil chí-kvadrát rozdělení s 5 stupni volnosti, zadáme

IDF.CHISQ(0.95,5).

Výsledkem bude hodnota 11,07.

b) *Hodnoty distribuční funkce*

Příslušná funkce má název CDF.CHISQ a syntaxi

CDF.CHISQ(*quant*;*df*),

kde *quant* je hodnota, pro kterou chceme zjistit hodnotu distribuční funkce, a *df* je počet stupňů volnosti. Například jestliže chceme zjistit, kolikaprocentní kvantil je hodnota 11,07 pro chí-kvadrát rozdělení s 5 stupni volnosti, zadáme

CDF.CHISQ(11.07,5).

Obdržíme výsledek 0,95, jde tedy o 95% kvantil.

Literatura

[1] AGRESTI, A. *Categorical Data Analysis*. *Second Edition*. Hoboken: John Wiley & Sons, 2002.

[2] AGRESTI, A. *An Introduction to Categorical Data Analysis*. *Second Edition*. Hoboken: John Wiley & Sons, 2007.

[3] ANDĚL, J. *Statistické metody*. Praha: Matfyzpress, 1998.

[4] ANDĚL, J. Statistické modely. *Statistika*, 2/2003, 1–17.

[5] BARTHOLOMEW, D. J., STEELE, F., MOUSTAKI, I. a J. I. GALBRAITH. *The Analysis and Interpretation of Multivariate Data for Social Science*. Boca Raton: Chapman & Hall/CRC, 2002.

[6] BERKA, P. *Dobývání znalostí z databází*. Praha: Academia, 2003.

[7] BISHOP, Y. M., FIENBERG, S. E. a P. W. HOLLAND. *Discrete Multivariate Analysis. Theory and Practice*. Cambridge: The MIT Press, 1980.

[8] BLATNÁ, D. *Neparametrické metody*. Praha: Nakladatelství Oeconomica, VŠE, 1996.

[9] COLLETT, D. *Modelling Binary Data. Second Edition*. Boca Raton: Chapman & Hall/CRC, 2003.

[10] CONOVER, W. J. *Practical Nonparametric Statistics. Third Edition*. New York: John Wiley & Sons, 1999.

[11] COX, T. F. and M. A. A. COX. *Multidimensional Scaling*. *Second Edition*. New York: Chapman & Hall/CRC, 2001.

[12] CYHELSKÝ, L., KAHOUNOVÁ, J. a R. HINDLS. *Elementární statistická analýza*. Praha: Management Press, 2001.

[13] FIELD, A. P. *Discovering Statistics Using SPSS. Second Edition*. London: SAGE Publications, 2005.

[14] FIENBERG, S. E.: *The Analysis of Cross-Classified Categorical Data*. *Second Edition*. New York: Springer, 2007.

[15] HABERMAN, S. J. *Advanced Statistics*. New York: Springer, 1966.

[16] HAN, J. a M. KAMBER. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers, 2001.

[17] HEBÁK, P., HUSTOPECKÝ, J., JAROŠOVÁ, E. a I. Pecáková. *Vícerozměrné statistické metody (1)*. 2. vydání. Praha: Informatorium, 2007.

[18] HEBÁK, P., HUSTOPECKÝ, J. a I. MALÁ. *Vícerozměrné statistické metody (2)*. Praha: Informatorium, 2005.

[19] HEBÁK, P. a kol. *Vícerozměrné statistické metody (3)*. 2. vydání. Praha: Informatorium, 2007.

[20] HINDLS, R., HRONOVÁ, S. a I. NOVÁK. *Metody statistické analýzy pro ekonomy*. Praha: Management Press, 2000.

[21] HINDLS, R., HRONOVÁ, S., SEGER, J. a J. FISCHER. *Statistika pro ekonomy*. 8. vydání. Praha: Professional Publishing, 2007.

[22] LUHA, J. Analýza nominálnych a ordinálnych znakov. *EKOMSTAT 2000*. Trenčianske Teplice: SŠDS, 2000, 120–150.

[23] LUHA, J. Aplikácie metód štatistickej analýzy kvalitatívnych znakov vo výskumoch verejnej mienky. Tajov: *FernStat 2004*. SŠDS, 2004. 36–49.

[24] MAREK, L. a kol. *Statistika pro ekonomy – aplikace*. 2. vydání. Praha: Professional Publishing, 2007.

[25] NOVÁK, I. a I. PECÁKOVÁ. Měření souvislostí kategoriálních proměnných. *Statistika*, **38** (2001), 35–48.

[26] PECÁKOVÁ, I. Mantelovy-Haenszelovy statistiky v SPSS. *Statistika*, **39** (2002), 103–110.

[27] PECÁKOVÁ, I. *Statistika v terénních průzkumech*. *2. vydání*. Professional Pub-lishing, Praha 2011.

[28] PECÁKOVÁ, I.: The possibilities of correspondence analysis in a two-way contingency table. *Applications of Mathematics and Statistics in Economy*. Praha: Professional Publishing, 2004.

[29] PECÁKOVÁ, I., NOVÁK, I. a J. HERZMANN. *Pořizování a vyhodnocování dat ve výzkumech veřejného mínění.* Praha: Nakladatelství Oeconomica, VŠE, 2004.

[30] POWERS, D. A. a Y. XIE. *Statistical Methods for Categorical Data Analysis*. San Diego: Academic Press, 2000.

[31] ŘEHÁK, J. a B. ŘEHÁKOVÁ. *Analýza kategorizovaných dat v sociologii*. Praha: Academia, 1986.

[32] ŘEHÁK, J. a B. ŘEHÁKOVÁ, B. Logitové modely: analýza vlivu exogenních faktorů u kategorizovaných dat. *Sociologický časopis*, **28** (1992), 84–102.

[33] ŘEHÁKOVÁ, B. Nebojte se logistické regrese. *Sociologický časopis*, **36** (2000), 475–492.

[34] ŘEZANKOVÁ, H. *Analýza kategoriálních dat*. Praha: Nakladatelství Oeconomica, VŠE, 2005.

[35] ŘEZANKOVÁ, H.: *Analýza dat z dotazníkových šetření. 4. vydání.* Praha: Professional Publishing, 2017.

[36] ŘEZANKOVÁ, H. a S. HRONOVÁ. *Statistická data a databázový systém MS Access*. Praha: Nakladatelství Oeconomica, VŠE, 2006.

[37] ŘEZANKOVÁ, H., HÚSEK, D. a V. SNÁŠEL. *Shluková analýza dat. 2. vydání*. Praha: Professional Publishing, 2009.

[38] SARIS, W. E. and I. N. GALLHOFER. *Design, Evaluation and Analysis of Questionnaires for Survey Research*. Hoboken: John Wiley & Sons, 2007.

[39] SIMONOFF, J. S. *Analyzing Categorical Data*. New York: Springer, 2003.

[40] STANKOVIČOVÁ, I. Logistická regresia a jej využitie v ekonomickej praxi. *Forum Statisticum Slovakum*, 1/2007, 42–54.

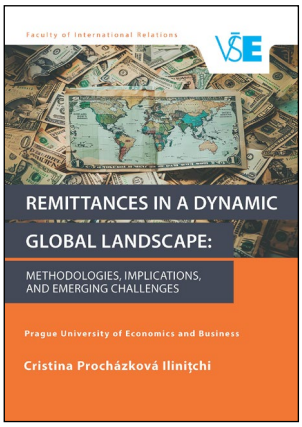
[41] STANKOVIČOVÁ, I. a M. VOJTKOVÁ. *Viacrozmerné štatistické metódy s aplikáciami*. Bratislava: Iura Edition, 2007.

[42] ZVÁROVÁ, J. *Základy statistiky pro biomedicínské obory*. Praha: Karolinum, 2001.

Rejstřík

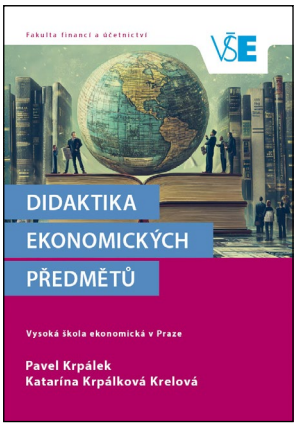
Z produkce Nakladatelství Oeconomica

více informací na https://oeconomica.vse.cz/



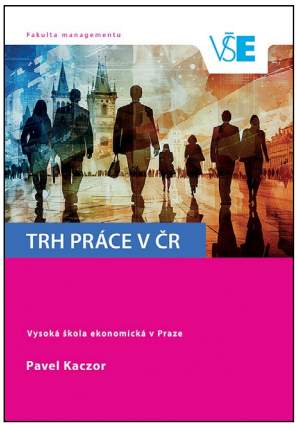
Cristina Procházková Ilinitchi: Remittances in a Dynamic Global Landscape

ISBN 978-80-245-2511-2, VŠE, 1. vydání, 2024, 238 stran, 496 Kč



Pavel Krpálek, Katarína Krpálková Krelová: Didaktika ekonomických předmětů

ISBN 978-80-245-2516-7, VŠE, 1. vydání v elektronické podobě, 2024, 254 stran, 239 Kč, (e-kniha)



Pavel Kaczor: Trh práce v ČR

ISBN 978-80-245-2513-6, VŠE, 1. vydání, 2024, 216 stran, 401 Kč

Název	**Statistické metody se zaměřením na kategoriální data**
Autorka	prof. Ing. Hana Řezanková, CSc.
Vydavatel	Vysoká škola ekonomická v Praze Nakladatelství Oeconomica
Vydání	1. vydání v elektronické podobě
Jazyková a redakční úprava	Mgr. Ludmila Doudová
Grafický návrh	Daniel Hamerník, DiS.
Počet stran	210
DTP	Vysoká škola ekonomická v Praze Nakladatelství Oeconomica
Doporučená cena	Zdarma ke stažení

ISBN 978-80-245-2521-1