

LEKCE 5

STATISTICKÁ INFERENCE ANEB ZOBECŇOVÁNÍ VÝSLEDKŮ Z VÝBĚROVÉHO NA ZÁKLADNÍ SOUBOR

Empirický výzkum v sociálních vědách je velmi často založen na tom, že získává údaje jenom o části subjektů, tyto údaje analyzuje a poté je *generalizuje* (zobecňuje) na příslušnou populaci, z níž byly tyto subjekty vybrány. Hovoříme-li o generalizaci, mějme na paměti, že existují dva hlavní způsoby generalizace: generalizace statistická a teoretická (de Vaus 2002).

Teoretická generalizace

Tato generalizace znamená zobecňování z empirických dat do teorie. Využívá se hlavně v takových výzkumných designech jako experiment nebo případová studie, neboť oba tyto výzkumné postupy nejsou obvykle založeny na práci s reprezentativními vzorky případů, takže statistická generalizace zde nemá smysl. Teoretická generalizace je založena hlavně na replikaci (opakování) – nalezneme-li stejné výsledky, kdykoliv je experiment opakován, naše důvěra v jeho výsledky se stále zvyšuje. Pokud jsou výsledky experimentu navíc opakovaně nalézány i za různých podmínek a na různých souborech, naše důvěra se zvyšuje ještě více. Pak jsme schopni naše výsledky zobecnit a včlenit je do existující teorie. Jelikož předmětem našeho kursu je statistická analýza dat, je zřejmé, že dále se budeme zabývat pouze možnostmi generalizace statistické.

Statistická generalizace znamená zobecňování výsledků z výběrového souboru (vzorku) na soubor základní (populaci). Říká se jí také statistická inference (statistické usuzování).¹ Jelikož je založena na teorii pravděpodobnosti, je základní podmínkou statistické generalizace to, že výběrový soubor, z něhož chceme zobecňovat, musí být pravděpodobnostním výběrem, tedy výběrem založeným na pravděpodobnostních (náhodných) postupech. Součástí statistické generalizace je zjišťování, s jakou pravděpodobností výběrové výsledky odrážejí skutečné vlastnosti základního souboru. To nám umožňuje inferenční statistiky. Ty jsou dvojího druhu: a) bodové a intervalové odhady a b) testy statistické významnosti.

Bodové a intervalové odhady

Problematiku bodových a intervalových odhadů naleznete detailně popsánu na jiném místě.² Zde proto postačí sdělit, že ačkoliv ve výběrovém souboru jsme schopni vypočítat přesné statistické údaje např. o průměrném příjmu souboru, o jeho průměrné inteligenci, o korelaci mezi mírou tolerance a vzděláním, pro základní soubor musíme tyto veličiny (parametry) pouze odhadovat, neboť víme, že naše výběrové výsledky jsou vždy zatíženy (větší či menší) výběrovou chybou. Proto pro jejich stanovení vypočítáváme směrodatnou chybu a intervaly spolehlivosti.

Testy statistické významnosti

Tyto testy umožňují odhadnout, jak je pravděpodobné, že výsledky nalezené ve výběrovém souboru jsou způsobeny výběrovou chybou, jak je pravděpodobné, že např. zjistíme poměrně silnou korelaci ve výběrovém souboru, byť v souboru základním vůbec neexistuje.

Statistická generalizace má bohužel ale své limity. Postupy statistické inference lze použít pouze v případě, že:

¹ O statistické inferenci se lze česky dočíst např. v Hendl, Jan 2004. Přehled statistických metod zpracování dat. Portál, Praha, str. 167–202.

² Řehák, Jan 1971. „Poznámky k analýze sociologických dat.“ *Sociologický časopis*, (4), str. 425–433.

- výběrový soubor má charakter je dobrým reprezentantem populace,
- tato populace je náležitě definována,
- byl použit adekvátní pravděpodobnostní výběrový postup,
- míra neuskutečněných rozhovorů je nízká (počet odmítnutých rozhovorů byl nízký), takže tzv. *response rate* je vysoký.

Postupy statistické inference **nemá smyslu** používat, když:

- výběrový soubor je vybrán prostřednictvím nepravděpodobnostních metod;
- míra uskutečněných rozhovorů je nižší než 85 % (Blaikie 2003).³ Vysoká míra odmítnutých rozhovorů snižuje reprezentativitu souboru a zvyšuje tak pravděpodobnost, že náš výběr bude nespolehlivý pro generalizaci výsledků. Sociální vědy se ovšem s problémem odmítnutých rozhovorů setkávají dnes poměrně často. Co s tím? Mnoho se dělat nedá, lidé mají právo účast na výzkumném rozhovoru odmítnou. Jistým kompromisem je, že snažíme od těch, kdo odmítnou zjistit alespoň základní socio-demografické charakteristiky. Pokud socio-demografický profil odmítnutých je podobný těm, kdo se výzkumu neodmítli zúčastnit, můžeme výsledky generalizovat. Pokud jsou ale jejich charakteristiky odlišné, jsme ztraceni, neboť to je signálem, že určitá část vytipovaných respondentů systematicky odmítla se rozhovorů zúčastnit, takže nás výběrový soubor je určitým způsobem systematicky vychýlen.
- když pracujeme s celou populací (jakkoliv definovanou pro účely našeho výzkumu)

Populace a výběry

Každý sociálně vědní výzkum stojí před rozhodnutím, jakým způsobem získat dat pro své analýzy. Pokud je pro analýzu nezbytné získat o velkém množství jedinců, sociálních jednotek nebo sociálních artefaktů, pak se musíme vždy rozhodnout, zdali budeme sbírat data od celé populace těchto jedinců, jednotek nebo artefaktů, nebo zda provedeme sběr dat jenom od nějaké části z této populace, zda tedy provedeme výběr.

Sociální vědy většinou pracují s *výběry* (výběrovými soubory) – práce s celou populací není příliš častá (samozřejmě záleží zde na tom, jak je definována populace našich analytických jednotek). Ukazuje se, že počet 1000–2000 jednotek je obvykle dost velký na to, aby přinesl adekvátní informaci o jakékoli populaci.

Za to, že pracujeme s výběry a ne s celou populací a že si tedy ulehčujeme situaci, se ovšem platí jistá cena (ostatně nic není na tomto světě zadarmo):

1. Nikdy si nemůžeme být absolutně jisti, že to, co se ukazuje v našem vzorku (výběru), existuje také v populaci (v základním souboru);
2. práce s výběry velmi komplikuje analýzu

Populace (základní soubor)

Populace je vždy definována pro potřeby výzkumu, v souvislosti s problémem, který řešíme. Populace může být velká (obyvatelé ČR), ale také malá (studenti prvního roku na FSS), a nemusí to být populace osob. Příklady:

- obyvatelé ČR
- VŠ studenti ČR
- Studenti prvního roku studia na brněnských VŠ
- předplatitelé novin MF Dnes
- Senioři v Jihomoravském kraji
- Výrobní podniky s počtem zaměstnanců do 200 osob na Moravě
- Všechna čísla Lidových novin v roce 2002
- Pouze sobotní vydání Lidových novin v roce 2002

³ U tohoto pravidla ale pozor: pokud je tato míra nižší, ale my víme, že charakteristiky výběrového souboru odpovídají souboru základnímu, pak toto pravidlo neplatí.

- Články v Lidových novinách týkající se homosexuality

Pokud v našem výzkumu pracujeme s celou populací, provádíme census a statistická inference zde pak nemá žádný smysl. Pokud pracujeme s *výběry*, je ideálem získat takový soubor, který věrně odráží charakteristiky populace. Tohoto ideálu se v praxi dosahuje jen stěží. Existuje jediný postup, jak získat *reprezentativní výběrový soubor*: pravděpodobnostní (náhodný) výběr – ten je založen na tom, že každá jednotka populace má stejnou pravděpodobnost být vybrána do souboru výběrového. Jakýkoliv jiný postup výběru⁴ nevede k reprezentativnímu souboru a aplikace inferenční analýzy není možná.

Jelikož naším cílem je vypovídat o populaci, musíme na základě statistické analýzy odhadovat charakteristiky populace nebo vzorce vztahů v populaci z charakteristik a vztahů nalezených ve vzorku. Ve statistické terminologii tomu říkáme odhad populačních parametrů z výběrových statistik a činíme tak prostřednictvím inferenční statistiky.

Výběr nemá většinou charakter ideálního reprezentanta populace. Pokud by byl ideálním reprezentantem, to je byl by pouze zmenšenou replikou populace, nebylo by třeba používat inferenční statistiku. Jelikož tomu tak většinou není, musíme počítat s tím, že vždy existuje *výběrová chyba*, kterou se snažíme určit.

Velikost výběru

Velikost výběrového souboru (vzorku) je důležitá pro přesnost odhadu populačních parametrů, tedy údajů o populaci získaných na základě statistik zjištěných ve výběrovém souboru. Na otázku, jak velký by měl být výběrový soubor, existuje jednoduchá, ale poněkud vágní odpověď: čím větší, tím lepší. Ale ono „čím více“ má své jasné praktické limity. Přesnost odhadu se nezvyšuje lineárně. Zpočátku se velmi zvyšuje, od určité velikosti výběru však roste již jen zvolna a v jistém bodě nastává moment, kdy náklady spojené se zvyšováním velikosti výběru jsou již vyšší než výnosy – to je zvyšování přesnosti odhadu populačních parametrů. Proto platí již výše řečené: velikost výběru mezi jedním až dvěma tisíci jednotek je obvykle plně dostačující.

A jaká je minimální velikost výběrového souboru? Na tuto otázku neexistuje jasná odpověď, naše zkušenost říká (a je podepřena jinými), že 300 jednotek by mohlo být adekvátní, ovšem 500 je lepší a 1000 je ještě lepší. Při stanovování velikosti výběru jsou ovšem ještě i další faktory ve hře, o nichž je třeba vědět.

- Rozhodnutí o velikosti výběrového souboru nemá nic do činění s poměrem velikosti výběru a populace. Mezi výzkumníky se traduje, že výběr by měl představovat přibližně 10% populace. Takové pravidlo ovšem neexistuje.
- Rozhodnutí o velikosti výběru závisí na tom, jak důležitá je přesnost odhadu, jak velkou výběrovou chybu si můžeme dovolit tolerovat. Musíme-li mít malou výběrovou chybu (např. při testování nového léku), musíme mít velký výběrový soubor.
- Velikost vorku závisí také na postupech analýz a na typu dat, s nimiž pracujeme. Obecně platí, že nominální data vyžadují větší soubory než data ordinální a že data kardinální potřebují menší soubory než data ostatní. Proč? Obecné pravidlo říká, že při analýze nominálních dat, kdy hlavním způsobem práce jsou třídění, by v políčku tabulky mělo být v průměru 10 případů. Pak velikost z tohoto technického hlediska se dá odhadnout tak, že vezmeme v úvahu dvě proměnné s nejvyšším počtem kategorií. Např. pokud máme v datech proměnnou s šesti kategoriemi a proměnnou s pěti kategoriemi, je počet polí v tabulce vytvořené z těchto dvou znaků 30 (5 x 6) a velikost souboru by měla být 300 (5 x 6 x 10). Pokud předpokládáme třídění třetího stupně (to je tabulku se třemi proměnnými), tento počet musí být dále násoben počtem kategorií třetí proměnné. Pokud ta má, řekněme, také pět kategorií, měla by být velikost souboru 1500. Při plánovaném třídění ještě vyšších stupňů (a teprve třídění vyšších stupňů mnohdy odhalí v datech vztahy, které jsou výzkumně zajímavé a netriviální) se velikost vzorku samozřejmě dále zvyšuje. Jelikož v sociologických výzkumech pracujeme velmi často s nominálními proměnnými, leží zde odpo-

⁴ Výběrové postupy nejsou předmětem tohoto textu.

věd' na případnou otázku, proč některé výběry jsou vskutku velké a mají nezřídka velikost pohybující se kolem 4000 až 5000 jednotek. Při práci s kardinálními proměnnými se uvádí jako nejnížší možná velikost souboru z tohoto technického hlediska 30 jednotek.

Velikost výběrové chyby však není ovlivněna pouze velikostí souboru. Má na ni vliv také to, jak heterogenní je populace v parametru, který se snažíme generalizovat. Pokud je relativně homogenní (extremním případem by mohla být její naprostá uniformita – např. v postoji k trestu smrti. Pak by stačilo udělat výběr o jednom jediném prvku a my bychom z něj mohli směle svůj výsledek zobecnit), je výběrová chyba nižší, pokud je velmi heterogenní, míra výběrové chyby se zvyšuje. Potíž je ale v tom, že my v převážných případech heterogenitu populace neznáme (a proto je předmětem našeho výzkumu).

Vychýlený výběr a co s tím

Pro statistické generalizace je nesmírně důležité vědět, zdali náš výběrový soubor je či není vychýlen a pokud je vychýlen, tak jak mnoho. Vychýlenost výběru je relativním konceptem, vztahuje se totiž vždycky pouze k definici našeho základního souboru. Výběrový soubor je vychýlen tehdy, když je např. starší než soubor základní (máme v něm vyšší zastoupení seniorů než je v naší definované populaci) nebo vzdělanější (což se stává poměrně často – proč asi?) apod. Vychýlenost výběrového souboru zjistíme tak, že srovnáme některé jeho charakteristiky s údaji, které již máme o populaci z jiných zdrojů. Např. je-li náš vzorek reprezentantem celé dospělé populace ČR, pak můžeme srovnat jeho demografické charakteristiky se základním souborem, neboť ty o něm známe z výsledků sčítání lidu. Bohužel ne vždy máme takovouto informaci o základním souboru k dispozici. Týká se to především situace, kdy předmětem našeho výzkumného zájmu je nějaká specifická populace (např. nedobrovolně bezdětné páry), jejíž parametry nejsou známy. V takovém případě je stanovení vychýlenosti obtížné.⁵

Zjistíme-li, na základě srovnání příslušných charakteristik, že jsme získali vychýlený výběr, můžeme tuto situaci napravit tzv. **vážením** souboru. Vážení souboru ve statistice znamená, že kategorie, které jsou ve výběrovém souboru podreprezentovány, budeme počítat vícekrát a naopak kategorie, které jsou nadreprezentovány, budeme počítat méněkrát. Když např. zjistíme, že máme v souboru více respondentů s vysokoškolským vzděláním a méně respondentů se vzděláním základním, musíme pro každý výpočet každého respondenta se základním vzděláním započítat více než jedenkrát a každého respondenta s VŠ vzděláním započítat méněkrát než jednou.

Toto lze zařadit statisticky tak, že pro každou relevantní proměnnou (většinou se zajímáme o pohlaví, věk, vzdělání, popř. velikost místo bydliště a geografickou lokaci), v níž se náš výběrový soubor odchyluje od základního, stanovíme příslušné váhy a těmito vahami jednotky násobíme. Tím „vážíme soubor“. Vážení souboru závisí na tom, zdali jej vážíme pouze podle jedné proměnné, nebo podle několika proměnných.

Vážení souboru podle jedné proměnné

Vážení souboru podle jedné probíhá následovně.

1. Nejdříve získáme rozložení příslušné proměnné jak ve vzorku, tak v populaci. Dejme tomu, že ve vzorku byl poměr mužů a žen 40 : 60, zatímco v populaci, jak jsme zjistili z jiných zdrojů byl tento poměr 50 : 50. V tabulce (viz tab. 5.1) to tedy vypadá takto:

⁵ Mimochodem, v tomto případě bychom měli již problémy se samotným stanovením výběrového postupu, neboť náš základní soubor (naše populace) je sice jednoduše definovatelný (nedobrovolně bezdětní), avšak velmi obtížně bychom zde získávali oporu výběru, tedy seznam jednotek (a nejlépe ještě s jejich adresami), z nichž by bylo možné provést pravděpodobnostní výběr. Řešením by snad bylo, samozřejmě pouze v případě, pokud by neexistovalo lékařské tajemství a neplatil by zákon o ochraně osobních údajů, zkompileovat takový seznam od všech pracovišť v ČR, která léčí neplodnost.

Tab. 5.1: Podíl mužů a žen (v %) ve vzorku a v populaci

	Výběrový soubor	Populace
Muži	40 %	50 %
Ženy	60 %	50 %

Muži byli ve vzorku podreprezentováni, ženy naopak nadreprezentovány. Náš vzorek tedy musíme upravit (zvážit) tak, aby jeho proporce v proměnné pohlaví odpovídaly proporcím mužů a žen v populaci.

Příslušné váhy stanovíme podle jednoduchého vzorce:

$$\text{váha} = \frac{\text{populace}(\%)}{\text{vzorek}(\%)} \quad (1)$$

Pro muže bude váha: $50 / 40 = 1,25$, pro ženy $50 / 60 = 0,83$

Nyní musíme v SPSS vytvořit novou proměnnou, která bude tyto váhy obsahovat a pak mu dát příkaz, aby tyto váhy při každém výpočtu použil. Váhovou proměnnou nazvěme *vaha_1*. Vytvoříme prostřednictvím příkazu *Compute*, proměnná pohlaví necht' má v našem souboru jméno *pohl*, muži necht' mají hodnotu 1, ženy 2. Doporučuji pracovat v režimu syntaxe.

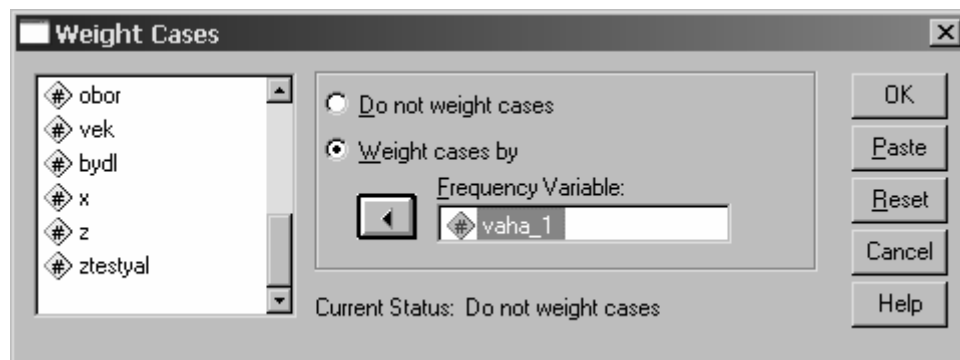
Nejdříve si otevřeme editor pro syntaktické příkazy: *File – New – Syntax*

Syntax bude mít podobu:

```
COMPUTE VAHA_1 = 0.
IF (POHL = 1) VAHA_1 = 1.25.
IF (POHL = 2) VAHA_1 = 0.83.
```

Pozn. Když vytváříte novou váhovou proměnnou příkazem *Compute*, je dobré nastavit počáteční hodnotu na 0. Další dva řádky tuto hodnotu změní buď na 1,25 (pokud to bude muž) nebo na 0,83 (pokud to bude žena). Těm, u nichž nemáme záznam o pohlaví, zůstane nula a při vážení zmizí ze souboru. A pozor, desetinnou čárku je třeba v SPSS psát jako desetinnou tečku!

Spustíme syntaktický příkaz, čímž vytvoříme váhovou proměnnou. Váhy poté zapneme tak, že spustíme proceduru: *Data – Weight cases* a v dialogovém okně klikneme na příkaz *Weight cases by*. Do okénka *Frequency variable* vložíme váhovou proměnnou *vaha_1* a klikneme na O.K.



Tím jsme zapnuli váhy, které při jakémkoliv výpočtu budou jednotky započítávat s příslušnou hodnotou. Všimněte si, že zapnutí vah je indikováno v pravém dolním rohu datového spreadsheetu nápisem *Weight On*.

Obvykle ale musíme soubor vážit nejenom podle jedné proměnné, ale podle více proměnných. Postup, jak to uděláme je podobný, pouze syntaktický příkaz bude poněkud delší a komplikovanější.

Vázení souboru podle více proměnných

Pro jednoduchost předpokládejme, že musíme náš soubor upravit váhami podle tří proměnných: podle pohlaví (*pohl*), národnosti (*narod*) a vzdělání (*vzdel*). Pro jednoduchost příkladu předpokládejme, že jsme národnost měli pouze dichotomickou (1. Češi, 2. Slováci) a rovněž vzdělání že jsme měřili dichotomicky (1. nižší, 2. vyšší.). Nejdříve tedy musíme zjistit, jaké byly proporce jednotlivých proměnných a jejich kategorií ve vzorku a v populaci. To uvádí tabulka 5.2.

Tab. 5.2: Podíl kategorií pohlaví, národnost a vzdělání (v %) ve vzorku a v populaci

(A) Výběrový soubor

	Muži		Ženy	
	Češi	Slováci	Češi	Slováci
Nižší vzdělání	20	10	15	10
Vyšší vzdělání	10	15	10	10

(B) Populace

	Muži		Ženy	
	Češi	Slováci	Češi	Slováci
Nižší vzdělání	25	10	20	5
Vyšší vzdělání	20	5	10	5

Pozn.: Procentuální podíly jsou v každé tabulce vypočteny z celkového N. Součet tedy musí dávat 100%

Váhy se vypočtou podle stejného vzorce, jako v případě, kdy jsme vážili soubor podle jedné proměnné (viz rovnici 1). Vycházejí následovně (viz tab. 5.3)

Tab. 5.3: Váhy založené na proměnných pohlaví, národnost a vzdělání

	1. Muž		2. Žena	
	1. Češi	2. Slováci	1. Češi	2. Slováci
1. Nižší vzdělání	1,25	1,0	1,33	0,50
2. Vyšší vzdělání	2,00	0,33	1,00	0,50

Syntax pro nastavení jednotlivých hodnot nové váhy, nazvané *vaha_2*:

```
COMPUTE VAHA_2 = 0.
IF ((POHL = 1) AND (NAROD = 1) AND (VZDEL = 1)) VAHA_2 = 1.25.
IF ((POHL = 1) AND (NAROD = 1) AND (VZDEL = 2)) VAHA_2 = 2.00.
IF ((POHL = 1) AND (NAROD = 2) AND (VZDEL = 1)) VAHA_2 = 1.00.
IF ((POHL = 1) AND (NAROD = 2) AND (VZDEL = 2)) VAHA_2 = 0.33.
IF ((POHL = 2) AND (NAROD = 1) AND (VZDEL = 1)) VAHA_2 = 1.33.
IF ((POHL = 2) AND (NAROD = 1) AND (VZDEL = 2)) VAHA_2 = 1.00.
IF ((POHL = 2) AND (NAROD = 2) AND (VZDEL = 1)) VAHA_2 = 0.50.
IF ((POHL = 2) AND (NAROD = 2) AND (VZDEL = 2)) VAHA_2 = 0.50.
```

Tento příkaz pak je třeba nechat proběhnout v SPSS a nastavit váhy tak, aby soubor převažovaly podle proměnné *vaha_2*.

Máme-li náš soubor ošetřen tak, že je relativně dobrým reprezentantem populace, můžeme začít se statistickými generalizacemi. Podívejme se nejdříve, jak se pracuje s intervalovým odhadem, to je jak se určují intervaly spolehlivosti.

Intervaly spolehlivosti

Nyní již víme, že ať pracujeme se sebelepším výběrovým souborem, nikdy si nemůžeme být jisti, že charakteristika vypočtená ze vzorku bude mít tutéž hodnotu také v souboru základním, neboť naše výběrové hodnoty jsou zatíženy výběrovou chybou. Nemá proto valného smyslu očekávat, že když např. náš výběrový soubor v otázce na předvolebních stranické preference ukáže, že ODS by získala 32 % hlasů a ČSSD 17 %, bude přesně taková proporce i v populaci. Je mnohem lepší strategií náš výběrový výsledek (výběrový odhad se tomu také říká) vzít jako základ pro odhad příslušného populačního výsledku (populačního parametru). Tento populační parametr nestanovujeme jedním číslem (bodovým odhadem), ale intervalově (intervalovým odhadem), to je vypočtením pravděpodobné dolní a horní hodnoty tohoto parametru. Pak si můžeme být docela jisti, že náš interval je *spolehlivým* rámcem, v němž se bude hodnota populačního parametru. Proto se tomuto intervalu říká **interval spolehlivosti**.

Pro velikost intervalu spolehlivosti jsou důležité tři věci: 1) velikost výběrového souboru, 2. velikost rozptylu v základním souboru a 3) míra jistoty (míra spolehlivosti), kterou chceme mít, že naše výběrová statistika se bude v tomto intervalu pohybovat. Ve statistice je obvyklé, že tato úroveň spolehlivosti je stanovena na 95 %. Což znamená, že pokud bychom např. v našem výše uvedeném příkladu předvolebních preferencí vypočítali, že interval spolehlivosti pro preference ODS je s 95 % jistotou 29–35 %, věděli bychom, že když provedeme 100 různých výběrů, pouze v pěti z nich by skutečný podíl preferencí ODS byl mimo tento interval.

Pro výpočet intervalu spolehlivosti musíme:

1. znát výběrovou charakteristiku (může jí být průměr, procento, procentuální rozdíl mezi skupinami, mohou to být ale i korelační koeficienty nebo regresní koeficienty; tu získáme výpočtem z dat výběrového souboru.
2. Vypočítat směrodatnou chybu, která měří velikost výběrové chyby.
3. Se rozhodnout, jak velkou úroveň spolehlivosti požadujeme. Obvykle pro sociální vědy nám stačí jistota 95 %, ale někdy požadujeme i úroveň 99 %, popř. 99,9 % (s tak vysokou jistotou pracují především v biologii a medicíně).

Ukažme si vše prakticky.

Příklad P5.1: Z příkladu P2.3, kde jsme z dat EVS spočítali, jaká je průměrná hodnota postoje k důležitosti Boha v životě českých respondentů (na desetibodové stupnici byl průměr 3,63) chceme nyní stanovit interval spolehlivosti (*confidence interval, CI*) pro základní soubor. Jelikož výběrový soubor je reprezentativní pro populaci ČR starší 18 let – k tomu viz článek Jana Řeháka v časopise *Sociální studia* 6 z roku 2001, str. 16 – má toto úsilí smysl. Zjišťujeme tedy, jaké hodnoty průměru můžeme očekávat v celé dospělé populaci České republiky. Intervaly spolehlivosti pro hodnotu průměru vypočítá SPSS v proceduře *Analyze — Descriptive Statistics — Explore* pro proměnnou q33. Výsledek:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Q33 Bůh - důležitost v životě	1846	96,8%	62	3,2%	1908	100,0%

Descriptives

		Statistic	Std. Error	
Q33 Bůh - důležitost v životě	Mean	3,63	,07	
	95% Confidence Interval for Mean	Lower Bound	3,49	
		Upper Bound	3,77	
	5% Trimmed Mean	3,43		
	Median	2,00		
	Variance	9,345		
	Std. Deviation	3,06		
	Minimum	1		
	Maximum	10		
	Range	9		
	Interquartile Range	5,00		
	Skewness	,858	,057	
	Kurtosis	-,614	,114	

Vzhledem k tomu, že chceme mít poměrně velkou jistotu o hodnotě průměru základního souboru, je v SPSS nastavena standardně úroveň spolehlivosti na 95 %. V tabulce *Descriptives* vidíme, že dolní hranice intervalu spolehlivosti je 3,49 (*lower bound* v zeleném rámečku) a jeho horní hranice 3,77 (*upper bound*). Tato čísla tedy říkají, že s 95% jistotou můžeme očekávat, že průměrná hodnota odpovědi na otázku o důležitosti Boha v našem životě by se v celé české populaci pohybovala mezi 3,49–3,77.

Jedna důležitá poznámka.: Všimněte si hodnoty směrodatné chyby – *Std. Error* = 0,07. Násobte tuto hodnotu dvěma⁶ a postupně ji odečtete a přičtete k hodnotě průměru. Jaký bude výsledek? No přesně takový, jaký vypočítal SPSS. Interval spolehlivosti se tedy, pokud znáte směrodatnou chybu, dá lehce spočítat i ručně. A jak vypočítáme směrodatnou chybu? I tu lze lehce spočítat a to tak, že podělíme směrodatnou odchylku druhou odmocninou velikosti výběrového souboru (N). Zkontrolujme si: velikost výběrového souboru je, jak vidíme z tabulky *Case Processing Summary*, 1846 – je třeba pracovat pouze s údajem o platných odpovědích, ti, kdo na tuto otázku neodpověděli, nebyli do výpočtu průměru zahrnuti. Druhá odmocnina tohoto čísla je 42,96. Směrodatná odchylka (*Std. Deviation* v tab. *Descriptives*) je 3,06, pak $3,06/42,96 = 0,07$, což je hodnota směrodatné chyby.

Pokud bychom chtěli mít interval spolehlivosti stanoven s jistotou 99 %, nastavíme v dialogovém okně v proceduře *Explore – Statistics* hodnotu intervalu spolehlivosti na 99 %. Lze ji ovšem vypočítat i ručně. Ruční výpočet je nesmírně jednoduchý. Hodnotu směrodatné chyby násobíme třemi (viz poznámku 6) a výsledek přičteme a odečteme od hodnoty průměru. Takže v našem případě:

$$0,07 * 3 = 0,21.$$

$$3,63 + 0,21 = 3,84$$

$$3,63 - 0,21 = 3,42$$

99% interval spolehlivosti je tedy 3,42–3,84. Zkontrolujme výsledek z výpočtu v SPSS:

⁶ Dvěma násobíme proto, že víme, do dvou směrodatných odchylek na každou stranu od průměru v normálním rozložení leží 95 % případů. A 95 % je přesně ta jistota, kterou požadujeme. Do tří směrodatných odchylek na každou stranu pak leží 99 % případů, takže pokud bychom chtěli jistotu 99 %, násobili bychom průměr 3x.

který mně tyto triky poradil. Záleží přitom na tom, zdali hledáme interval spolehlivosti pro proměnnou, která je dichotomická (má jenom dvě varianty znaku, např. muž–žena, je spokojen–je nespokojen atd.), nebo polytomická (má více variant znaku). Na rozdíl od průměru, kdy stanovujeme interval spolehlivosti pouze k jedné jediné hodnotě, u výpočtu intervalu spolehlivosti pro procenta to je jiné. Zde musíme intervaly spolehlivosti vypočítávat pro jednotlivé varianty kategorické proměnné zvlášť.

Příklad P5.2: Interval spolehlivosti pro dichotomické proměnné.

Trik spočívá v tom, že hodnoty dichotomie (ať byly kódovány jako 0 a 1, nebo jako 1 a 2) převedeme (rekódujeme procedurou *Recode*) na hodnoty 0 a 100. Pro takto upravenou proměnnou pak již v proceduře *Explore* spočteme normální průměr (t.j. procento) a jeho interval spolehlivosti, který je v dané situaci hledaným intervalem spolehlivosti pro procenta.

Ukázka výpočtu. Chceme zjistit, jaký je v souboru EVS-ČR1999 interval spolehlivosti pro rozložení odpovědí na otázku q42: *Myslíte si, že žena musí mít děti, aby se splnila její poslání, nebo to není nutné?* Jelikož je to dichotomická proměnná, můžeme uplatnit Řehákův trik. Nejdříve tedy musíme rekódovat původní hodnoty 1 a 2 na hodnoty 0 a 100. Provedme:

Původní proměnná:

Tab. A

Q42 Žena musí mít děti, aby splnila poslání

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1 ano	795	41,7	44,1	44,1
2 není to nutné	1007	52,8	55,9	100,0
Total	1803	94,5	100,0	

Rekódovaná proměnná:

```
RECODE
  q42 (1=0) (2=100) .
EXECUTE.
```

Tab. B

Q42 Žena musí mít děti, aby splnila poslání

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 0	795	41,7	44,1	44,1
100	1007	52,8	55,9	100,0
Total	1803	94,5	100,0	

Tab. C

Descriptives

		Statistic	Std. Error	
Q42 Žena musí mít děti, aby splnila poslání	Mean	55,87	1,17	
	95% Confidence Interval for Mean	Lower Bound	53,58	
		Upper Bound	58,17	
	5% Trimmed Mean	56,52		
	Median	100,00		
	Variance	2466,89		
	Std. Deviation	49,67		
	Minimum	0		
	Maximum	100		
	Range	100		
	Interquartile Range	100,00		
	Skewness	-,237	,058	
	Kurtosis	-1,946	,115	

Vidíme, že vypočítaný průměr 55,87 odpovídá podílu respondentů, kteří se domnívají, že není nutné, aby žena měla děti (55,9 ve sloupci *Valid Percent* v tab. A nebo B). Proto pro tento údaj můžeme údaje o horní a dolní hranici 95% intervalu spolehlivosti pro průměr (v tabulce C) chápat jako údaje o horní a dolní hranici intervalu spolehlivosti pro toto procento. Tudiž v základním souboru, to je mezi dospělou populací ČR, se pohybuje podíl lidí, kteří si myslí, že není nutné, aby žena měla děti k naplnění jejího poslání, mezi 53,6 a 58,2 %.

Pro výpočet intervalu spolehlivosti pro podíl lidí, zastávají názor, že žena musí mít děti k naplnění poslání, již musíme použít kalkulačky – stačí ale pouze hodnoty intervalů spolehlivosti odečíst od 100: $100 - 53,58 = 46,42$ a $100 - 58,17 = 41,83$. Podíl respondentů s tímto postojem se bude tak v základním souboru pohybovat mezi 41,8 a 46,2 %.

Příklad P5.3: Interval spolehlivosti pro polytomické proměnné.

U vícehodnotového znaku se postupuje, pokud nechceme interval spolehlivosti počítat ručně, jinak. Tabulku, kterou v SPSS dostaneme z *Frequencies*, zkopírujeme (prostřednictvím příkazu *Copy*) a vložíme ji do Excelu. V něm si připravíme příslušný vzorec pro výpočet směrodatné chyby pro procento:

$$\sqrt{\frac{p(100-p)}{N}}$$

a pak už jen dosazujeme příslušná data. A pokud si tento excelovský soubor uložíme jako matici, můžeme se k němu opakovaně vrátit a vypočítat velmi rychle interval spolehlivosti pro jakoukoliv polytomickou proměnnou.

Ukázka:

V příkladu P2.1 jsme se zajímali o rozložení proměnné q46_3. Vypočítejme pro jednotlivá procenta intervaly spolehlivosti. Nejdříve si tedy v SPSS udělejme znovu třídění prvního stupně této proměnné a použijme k tomu proceduru *Frequencies*. Vypočtený výsledek zkopírujeme a vložíme do tabulkového procesoru Excel. Bude to vypadat takto:

Nyní si vložíme příslušné vzorce pro výpočet intervalu spolehlivosti (jeho matematickou podobu jsme jen pro připomenutí přidali do volného prostoru). Vidíte, že jsme si v nové tabulce zkopírovali údaje o absolutních četnostech (*Frequency*), z nichž ovšem pro výpočet použijeme, jak říká vzorec, pouze jeden údaj, jímž je celková velikost souboru (1 780). To je ve vzorci ono N . Dále jsme si zkopírovali údaje o platných procentech (*Valid Percent*), neboť to jsou hodnoty p ve vzorci. Přidali jsme tři nové sloupce. Do sloupce **Std. error 95** jsme za použití excelovské syntaxe vepsali celý vzorec pro výpočet směrodatné chyby, kterou stanovujeme pro 95% jistotu. Tento vzorec je vepsán do buňky D16 a způsob zápisu je zobrazen v dialogovém okně.

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid				
1 rozhodně souhlasí	213	11,2	12,0	12,0
2 souhlasí	1070	56,1	60,1	72,1
3 nesouhlasí	474	24,9	26,6	98,7
4 rozhodně nesouhlasí	22	1,2	1,3	100,0
Total	1780	93,3	100,0	
Missing				
-2 neodpověděla	8	0,4		
-1 neví	120	6,3		
Total	128	6,7		
Total	1908	100,0		

	Freq.	Valid %	std. Error 95%	CI dolní	CI horní
1 rozhodně souhlasí	213	12,0	1,54	10,5	13,5
2 souhlasí	1070	60,1	2,32	57,8	62,4
3 nesouhlasí	474	26,6	2,10	24,6	28,7
4 rozhodně nesouhlasí	22	1,3	0,53	0,7	1,8
Total	1780	100,0			

Další dva přidání jsou již přímo hodnoty dolního a horního intervalu spolehlivosti. Pod údajem $10,4$ je vzorec = C14-D14 a pod údajem $13,5$ je vzorec = C14+D14, tedy operace, kdy od 12 % těch, kdo rozhodně souhlasí s výrokem, že *Většina žen touží po domově a dětech*, nejdříve odečítáme velikost směrodatné chyby (1,5) a pak ji k 12 % přičítáme. Tím získáváme interval spolehlivosti (10 – 14 %) pro podíl obyvatel ČR, kteří rozhodně souhlasí s tímto výrokem.

Vidíme, že směrodatná chyba je v každém řádku jiná a je to pochopitelné. Pro 1,3 % musí být menší než pro 60,1 %. Proto uvádějí-li někdy agentury pro výzkum veřejného mínění velikost výběrové chyby (což je samo o sobě velmi chválný fakt) a tvrdí, že např. velikost výběrové chyby je 2 %, není to informace tak úplně přesná (proč?).

Máte-li takto připravenou matici, pak při výpočtu dalších intervalů spolehlivosti z jiných výpočtů SPSS stačí přepsat údaje o velikosti vzorku (buňka B18) a do buněk C14...C17 dosadit příslušná validní procenta. Excel (a v tom je jeho kouzlo), okamžitě přepočítá nově dosazené údaje a vy máte k dispozici nové intervaly spolehlivosti.⁷ Pokud bude vaše nová proměnná mít vyšší počet variant než

⁷ Tuto matici naleznete jako samostatný soubor pod názvem *int-spol.xls* na dokumentovém serveru informačního systému MU.

4, budete si muset přidat příslušný počet řádků, do nichž vepíšete patřičné vzorce (dávejte přitom velký pozor, abyste – pokud budete vzorce kopírovat – v nich měli správně označeny všechny odkazy na buňky).

Tento postup je, jak jistě uznáte, poněkud krkolomný. Proto kolega Řehák sepsal v rámci jazyka SPSS malý prográmeček (v jazyce SPSS se mu říká *script*), který intervaly spolehlivosti doplňuje přímo do tabulky z *Frequencies*. Postup je následující:

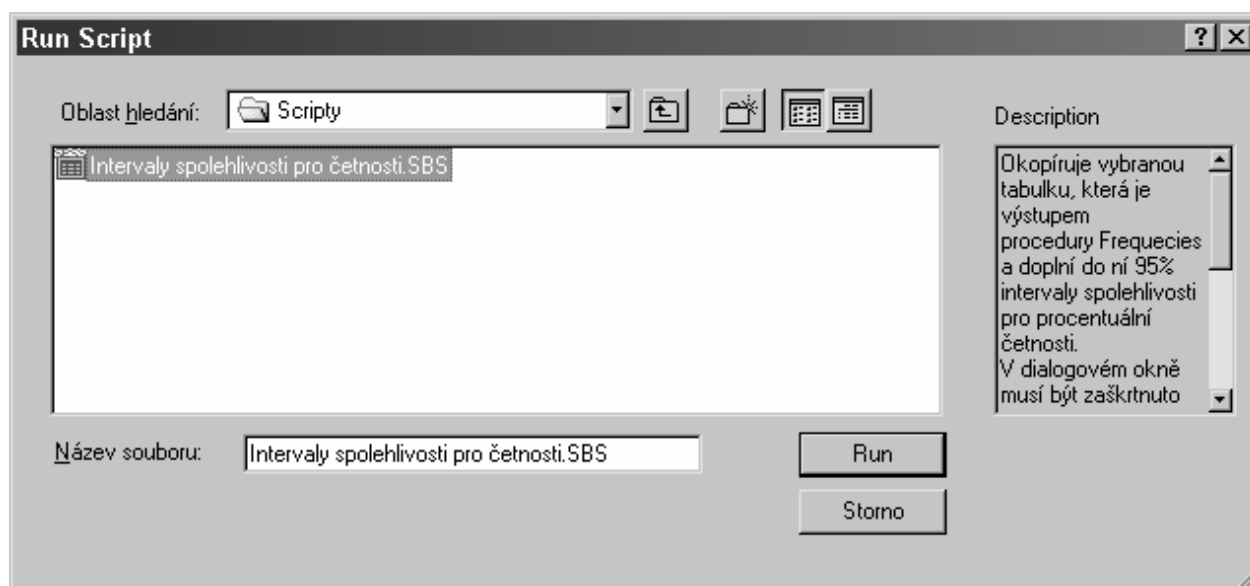
1. Necháte si spočítat *Frequencies* příslušné proměnné.
2. V Outputu SPSS na tuto tabulku 1x kliknete a tím ji označíte (viz malou šipku u tabulky na obrázku).

The screenshot shows the SPSS Output Viewer window. The main area displays a frequency table for the variable 'Q46_3 Většina žen touží po domově a dětech'. The table is as follows:

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 rozhodně souhlasí	213	11,2	12,0	12,0
	2 souhlasí	1070	56,1	60,1	72,1
	3 nesushlasí	474	24,9	26,6	98,7
	4 rozhodně nesouhlasí	22	1,2	1,3	100,0
Total		1780	93,3	100,0	
Missing	-2 neodpovědě/a	8	,4		
	-1 nev	120	6,3		
Total		128	6,7		
Total		1908	100,0		

A small arrow points to the table from the left sidebar, indicating the step described in the text.

3. Klikněte na tlačítko *Utilities* a v něm na příkaz *Run Script*.
4. V dialogovém okně naved'te SPSS tam, kde máte uložen script pro intervaly spolehlivosti.



Když si kliknete na jméno tohoto *scriptu* (jmenuje se „Intervaly spolehlivosti pro četnosti“), vkopíruje se do okénka „Název souboru“ a v okně vpravo nazvaném *Description* se objeví popis procedury. Klikněte na příkaz *Run*. Objeví se nová tabulka, která uvádí příslušné horní a dolní meze intervalu spolehlivosti pro jednotlivé varianty znaku (viz tab. 5.4, kterou jsme editovali tak, že jsme odmazali sloupce pro intervaly spolehlivosti kumulovaných četností).

Tab. 5.4: Intervaly spolehlivosti vypočtené z Řehákova *scriptu*

Četnostní tabulka s intervaly spolehlivosti proměnné Q46_3 Většina žen touží po domově a dětech

Hodnoty	Statistiky						
	Četnost	Relativní četnost	Dolní mez ^a	Horní mez ^a	Rel. četnost platných hodnot	Dolní mez ^a	Horní mez ^a
Platné							
1 rozhodně souhlasí	213	11,18%	9,77%	12,60%	11,99%	10,48%	13,50%
2 souhlasí	1070	56,09%	53,86%	58,31%	60,11%	57,84%	62,39%
3 nesouhlasí	474	24,86%	22,92%	26,80%	26,65%	24,59%	28,70%
4 rozhodně nesouhlasí	22	1,17%	,69%	1,65%	1,25%	,73%	1,77%
Celkem	1780	93,30%	92,18%	94,42%	100,00%		
Vynechané							
-2 neodpověděl/a	8	,43%	,14%	,73%			
-1 neví	120	6,27%	5,18%	7,35%			
Celkem	128	6,70%	5,58%	7,82%			
Celkem	1908	100,00%					

a. 95%ní interval spolehlivosti. K výpočtu je použita asymptotická metoda, která předpokládá, že celkový počet pozorování je větší než 30 a v každé kategorii se vyskytuje alespoň 5 případů.

V této tabulce nás zajímají sloupce pro validní četnosti (sloupce platných hodnot), které jsou vyznačeny žlutě. Když tyto intervaly spolehlivosti zkontrolujete s intervaly, které jsme vypočítali v excelovské tabulce výše, uvidíte, že jsou totožné.

Literatura

de Vaus, David 2002. *Analyzing Social Science Data*. SAGE Publications, London, str. 147 – 165, 187–193.