Základní a výběrový soubor, metody výběru

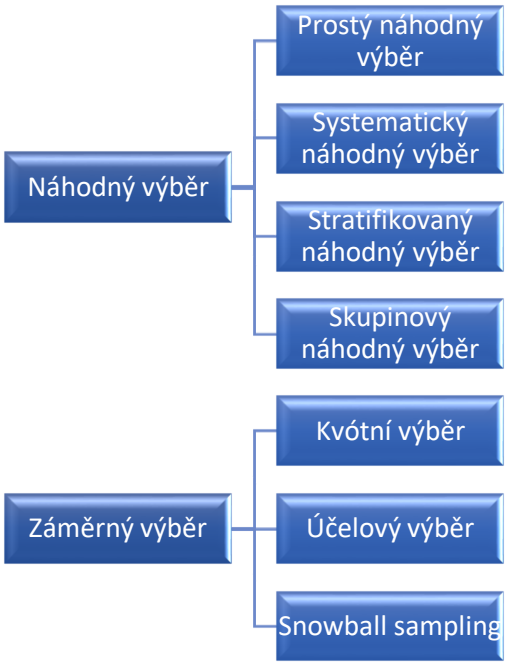
Když v našem výzkumu pracujeme s lidmi, často zjišťujeme, že není jednoduché získat data od všech subjektů, které spadají do skupiny, kterou chceme zkoumat. Těchto subjektů je buď příliš mnoho (např. všech obyvatel Česka nebo města Liberce), v tom případě není v našich silách zajistit data od všech, nebo je naopak obtížné zjistit, kdo všechno do dané skupiny náleží (např. cyklisté, trávící pravidelně svou letní dovolenou v geoparku Ralsko, amatérští sběratelé minerálů apod.). Naštěstí náš výzkum může být validní a zobecnitelný na celou skupinu, i když data budou pocházet pouze z její části. Tomuto procesu se říká **výběr vzorku** (angl. *sampling*).

Proč vzorku? Vzorek je totiž synonymum pro pojem **výběrový soubor**, což je množina jednotek, které v našem výzkumu zastupují tzv. **základní soubor**, což je množina všech jednotek, které spadají do naší definice zkoumaného objektu nebo subjektu.

Př. Pokud budeme studovat názory obyvatel Česka k problematice náboženství, základním souborem budou všichni obyvatelé Česka, výběrovým souborem (nebo také vzorkem) pak bude třeba 1000 obyvatel, kterých se reálně budeme dotazovat.

Asi vás napadne, že pokud budeme zobecňovat naše závěry na základě dat od malého množství respondentů, nebudou nejspíš příliš validní. Při výběru vzorku tak je důležitým kritériem, aby závěry byly zobecnitelné i na základní soubor – tomuto kritériu říkáme **reprezentativnost**. Ta je zajištěna vhodnou metodou výběru, která by měla zohledňovat důležité prvky struktury základního souboru (např. pohlaví, věk, vzdělání, geografické rozložení do regionů).

V některých případech není třeba výběr vzorku dělat, neboť máme k dispozici data za základní soubor. Je tomu tak v případech, kdy např. studujeme registrované členy nějaké organizace, kdy počet subjektů je relativně nízký, nebo kdy máme k dispozici data z šetření, které zahrnuje všechny jednotky základního souboru. Takové šetření se nazývá **cenzus** a jeho příkladem může být Sčítání lidu, domů a bytů, realizované Českým statistickým úřadem jednou za deset let.



Obr. 2 Základní druhy tvorby výběrového souboru

Vytvořeno s podporou projektu RoLiZ 4.0 - Rozvoj lidských zdrojů TUL pro zvyšování relevance, kvality a přístupu ke vzdělání v podmínkách Průmyslu 4.0, reg. č. CZ.02.2.69/0.0/0.0/16_015/0002329, operační program Výzkum, vývoj a vzdělávání.

Na obr. 2 je uvedeno základní dělení postupu při tvorbě výběrového souboru. Dva hlavní přístupy se od sebe liší pohledem na to, jakým způsobem zajistit reprezentativnost dat. **Náhodný výběr** je založen na předpokladu, že každý prvek základního souboru má stejnou pravděpodobnost stát se prvkem výběrového souboru. Pokud tedy zajistíme, že výběr prvků bude opravdu náhodný, měl by být náš výběrový soubor reprezentativní. Skutečná náhodnost by měla být zaručena systémem výběru prvků, který by měl být založen na losování, nebo generování náhodných čísel, které vyberou ze seznamu prvků základního souboru ty, které budou součástí vzorku. Tento postup se nazývá **prostý náhodný výběr**.

Odlišnou strategií zajištění náhodnosti je **systematický náhodný výběr**. Ten probíhá tak, že ze seznamu prvků základního souboru vybereme každý n-tý případ. Tohoto využívají například firmy, které ze seznamu svých zákazníků vyberou např. každého desátého apod. Pokud jsou zákazníci seřazeni podle nějaké víceméně náhodné proměnné (což může být třeba jejich příjmení), je tato metoda výběru vhodná a jednoduše použitelná.

V humánní geografii se často setkáváme se situací, kdy odlišné sociální skupiny reagují na předmět výzkumu rozdílně. V takovém případě je vhodné využít **stratifikovaný náhodný výběr**. Při něm se nejprve základní soubor rozdělí do tzv. *strat* (jednotné číslo *stratum*, česky vrstva), což jsou podsoubory vymezené na základě jedné nebo více společných vlastností. Touto společnou vlastností může být např. věk, vzdělání, etnicita, víra apod. V druhém kroku se pak provede náhodný výběr v rámci každé straty. Rozdělení do strat nám zajišťuje, že všechny podstatné strukturní rozdíly budou ve výsledku zaznamenány, tedy že ve výběrovém souboru bude dostatečné zastoupení mladých lidí i seniorů, vysokoškoláků i lidí bez dokončeného vzdělání atd.

Podobným případem je i **skupinový náhodný výběr**, kdy základní soubor je rozdělen do skupin (klastrů), které se od sebe příliš neliší a jsou vnitřně heterogenní. To můžou být např. kraje, okresy, nebo obce. Náhodným způsobem vybereme některé z těchto skupin a u nich pak zkoumáme všechny jejich prvky. To v geografii je obvykle dost obtížné (těžko nám dotazník vyplní všichni obyvatelé určitého kraje), proto v tomto kontextu spíše využíváme **vícestupňový náhodný výběr**, který obvykle využívá dvě (a více) z výše zmíněných možností. Může např. zkombinovat skupinový a prostý náhodný výběr, kdy bychom nejprve náhodně vybrali skupiny (např. okresy) a v rámci těchto skupin bychom následně náhodně vybrali jednotlivé lidi, které bychom oslovili.

Náhodný výběr má jednu podstatnou vadu – je třeba mít k dispozici seznam jednotek základního souboru, což může být problém. Náhodnost totiž rozhodně není zaručena tím, že stojíte uprostřed města a oslovujete náhodné lidi. Vybíráte totiž pouze z množiny lidí, kteří se ve správný čas pohybují tam, kde sbíráte data, což řadu obyvatel této obce vylučuje, navíc podvědomě budete oslovovat spíše sympatické lidi, než podivná individua. Váš vzorek tedy příliš náhodný nebude. Při tomto druhu sběru dat proto spíše využíváme tzv. **záměrný výběr**, při kterém není reprezentativnost zajištěna náhodností, ale naopak systematičností.

Nejčastějším druhem záměrného výběru je **kvótní výběr**. U něj je reprezentativnost výběrového souboru zaručena tzv. *kvótami*, které definují rozsah zastoupení určitých klíčových parametrů prvků. Pokud např. víme, že v reálné populaci je poměr mužů a žen 49:51, můžeme si např. nastavit kvótu, že jejich vzájemný poměr v našem vzorku bude v rozsahu 40:60 – 60:40. Podobně můžeme vymezit i zastoupení jednotlivých věkových skupin, kategorií vzdělání apod. Kvótní výběr tedy předpokládá, že známe údaje o základním souboru, které však v případě demografických kategorií v Česku známe velmi přesně. Pokud však určitou charakteristiku neznáme, logicky nemůžeme ani nastavovat kvóty. Kvótní výběr je často užívanou možností při různých dotazníkových šetřeních, kdy jde o rozumnou alternativu náhodného výběru. Na druhou stranu při jeho použití existuje určité úskalí, že pokud si již během

Vytvořeno s podporou projektu RoLiZ 4.0 - Rozvoj lidských zdrojů TUL pro zvyšování relevance, kvality a přístupu ke vzdělání v podmínkách Průmyslu 4.0, reg. č. CZ.02.2.69/0.0/0.0/16_015/0002329, operační program Výzkum, vývoj a vzdělávání.

procesu sběru dat nedáme pozor na průběžné plnění kvót, můžeme se k jeho konci nadít situace, kdy budeme hledat naprosto marginální případy (např. ženu nad 70 let s vysokoškolským vzděláním technického zaměření na malé obci). Je tedy dobré si průběžně plnění kvót kontrolovat.

Při kvalitativním výzkumu můžeme použít techniky záměrného výběru, které nezaručují reprezentativnost výběrového souboru. Pokud z dat nebudeme chtít získat kvantitativní data, ale budeme chtít pouze proniknout do určité zajímavé problematiky, není třeba, aby náš vzorek odpovídal základnímu souboru. Obvyklou volbou je v takovém případě **účelový výběr**, nazývaný také *ad hoc výběr*. V rámci něj výzkumník znalý tématu či problému vybere prvky sám. Obvykle tímto procesem vybíráme takové prvky, které jsou důležitější než ostatní – vybíráme např. starosty obcí, experty, vhodné lokality apod.

Určitou variantou účelového výběru pak je **snowball sampling**, který využíváme v případě, že neznáme dostatečný počet vhodných prvků (subjektů). Při snowball samplingu oslovíme několik prvních respondentů, kteří nám kromě odpovědí poradí i další kontakty na vhodné respondenty. Takto se dostáváme ke stále většímu počtu vhodných subjektů a náš výzkum je rozsáhlejší a kvalitnější. Tímto způsobem je vhodné shánět různé experty, lidi se společným koníčkem aj.

Poslední otázkou, kterou je třeba při tvorbě výběrového souboru zohlednit, je jeho velikost. Ve kvalitativním výzkumu další sběr dat ukončíme ve chvíli, kdy se přestaneme dozvídat nové informace. V kvantitativním výzkumu je potřeba dopředu zvážit, jaké množství dat potřebujeme. Přitom musíme vzít v potaz heterogenitu základního souboru (např. nakolik se preference lidí mohou lišit) a počet možných variant (např. různé odpovědi na otázky v dotazníku). Z matematického pohledu pak můžeme ještě zvážit hladinu významnosti a velikost vzorku určit výpočtem. Z praktického hlediska platí, že čím větší výběrový soubor, tím větší přesnost výzkumu. Rozumnou horní hranicí pro humánně geografické výzkumy je 1000 – 2000 obyvatel. Dolní hranice závisí na povaze výzkumu a výše zmíněné heterogenitě a počtu variant. U běžného dotazníku však postačuje 300 – 400 odpovědí.

Vytvořeno s podporou projektu RoLiZ 4.0 - Rozvoj lidských zdrojů TUL pro zvyšování relevance, kvality a přístupu ke vzdělání v podmínkách Průmyslu 4.0, reg. č. CZ.02.2.69/0.0/0.0/16_015/0002329, operační program Výzkum, vývoj a vzdělávání.