UNIVERZITA KARLOVA V PRAZE

FAKULTA SOCIÁLNÍCH VĚD

Institut ekonomických studií

Jan Krepl

Výběr z konečných populací v ekonomických úlohách

Bakalářská práce

Praha 2014

Autor práce: **Jan Krepl**
Vedoucí práce: **RNDr. Michal Červinka, Ph.D.**

Rok obhajoby: **2014**

Bibliografický záznam

KREPL, Jan. *Výběr z konečných populací v ekonomických úlohách.* Praha, 2014. 63 s. Bakalářská práce (Bc.) Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií. Vedoucí bakalářské práce RNDr. Michal Červinka, Ph.D.

Abstrakt

Výběrové šetření představuje základní metodu zjišťování populačních parametrů. Sociální vědy včetně ekonomie tuto techniku hojně využívají při získávání informací pro různé studie. Tato práce si dává za cíl seznámit čtenáře s problematikou výběrového šetření jako celku a následně popsat základní pravděpodobnostní techniky výběru. Pro empirickou část autor vyhledal vhodné práce studentů Institutu ekonomických studií FSV UK, v nichž data pochází z výběrového šetření. Tyto práce slouží jako podklad pro ilustraci popsaných metod z části teoretické. Na závěr autor diskutuje problematiku užití pravděpodobnostních výběrů v praxi.

Abstract

Survey sampling constitutes a basic method of obtaining values of population parameters. Social sciences including economics use survey sampling to collect information which is then used for research purposes. The goal of this thesis is to describe sample surveys in general and to focus on basic probability sampling schemes. For the empirical part, the author selected several suitable theses of IES FSV UK students where sample survey data was used. These theses serve as an illustration of described methods

in theoretical part. At the end, the possibility of applications of probability sampling is discussed.

Klíčová slova

pravděpodobnostní, výběr, dotazník, šetření, odhady, parametry.

Keywords

probability, sampling, survey, research, estimates, parameters.

Rozsah práce

71 886 znaků (včetně mezer)

Prohlášení

1. Prohlašuji, že jsem předkládanou práci napsal samostatně a výhradně s použitím citovaných pramenů.
2. Prohlašuji, že práce nebyla využita k získání jiného titulu.
3. Souhlasím s tím, aby práce byla zpřístupněna pro studijní a výzkumné účely.

V Praze dne 16. 5. 2014

Jan Krepl

Poděkování

Na tomto místě bych rád poděkoval RNDr. Michalu Červinkovi, Ph.D. za vedení této práce a rady při jejím zpracování. Dále děkuji mým rodičům za možnost studovat Karlovu univerzitu a pomoc v celém průběhu studia.

Teze bakalářské práce

Řešitel	Jan Krepl
Název (česky)	Výběr z konečných populací v ekonomických úlohách
Název (anglicky)	Sampling from finite population in economic problems
Akademický rok vypsání	2012/2013
Vedoucí / školitel	RNDr. Michal Červinka, Ph.D.
Datum zadání / přihlášení	30.05.2013

Zásady pro vypracování

Student nejprve popíše základní poznatky z teorie výběrových šetření. U výběrů z konečné populace není vhodné používat výsledky odvozené pro výběry z nekonečné populace. Student se proto zaměří na metody bodových a intervalových odhadů neznámých parametrů při pravděpodobnostních výběrech a zhodnotí přínos a nedostatky těchto metod. Zaměří se mimo jiné na roli a význam tzv. konečnostního násobitele. Student dále vybere malý počet ekonomických prací (například bakalářské nebo diplomové práce obhájené na IES FSV UK), ve kterých by bylo možné tyto metody použít. Součástí práce bude také navržení návodného postupu, jak identifikovat případy v ekonomii, kdy metody založené na výběrech z konečných populací je nutné nebo alespoň vhodné použít.

Předběžná náplň práce

1. Úvod
2. Odhady při pravděpodobnostních výběrech z konečných populací (přehled)
3. Aplikace metod založených na výběru z konečných populací na vybrané ekonomické úlohy
4. Závěr

Seznam odborné literatury

[1] Hájek, J. (1981). Sampling from a finite population. Marcel Dekker, inc., New York.

[2] Vorlíčková, D. (1985). Výběry z konečných populací. UK, Praha.

Obsah

1 Úvod

V ekonomické praxi se často snažíme získat informace o konečné skupině jednotek. Informací může být číselná hodnota (průměrný příjem, celoroční tržby atd.), stav (zaměstnanost, zájem o výrobek atd.) nebo názor (spokojenost zákazníků s novým produktem). Skupinou jednotek mohou být například všichni ekonomicky aktivní obyvatelé ČR, lékárny ve městě atd. Nabízí se dvě možnosti. Požadovanou vlastnost zkoumat u každé jednotky populace nebo jistým způsobem vybrat pouze část populace a na jejím základě vyvodit závěry o populaci celé. První možnost se nazývá *census*. V tomto případě žádné odhady či spekulace vytvářet nemusíme. O základním souboru a jeho charakteristikách máme veškeré informace a tedy úplnou znalost. Druhá možnost, v literatuře zpravidla nazývaná jako *výběrové šetření* (survey research nebo sample survey), má ovšem několik výhod. Cochran [3, s. 1] uvádí tyto: Menší finanční náklady, rychlost sběru a zpracování dat, realizovatelnost a přesnost.

V tomto textu se čtenář postupně seznámí s obecnou teorií výběrového šetření a následně budou vyloženy nejdůležitější pravděpodobnostní výběrové plány. Jmenovitě to jsou tyto: prostý náhodný výběr, stratifikovaný výběr a vícestupňový skupinový výběr. U každého z nich se autor pokusí popsat, kdy je vhodné jej použít, jaké jsou jeho kladné a záporné stránky a uvede teoretické výsledky týkající se odhadů. Dále autor nastíní problematiku výběru s nestejnou pravděpodobností zahrnutí a analyzuje problém neodpovědi. Volba uvedených témat[1] je motivována aplikací v části empirické. Tam budou analyzovány práce studentů

[1] Teorie pravděpodobnostních výběrů je širokou oblastí matematické statistiky. Každý její stručný přehled nutně vynechává jisté partie. Tato práce například neuvádí teorii poměrových a regresních odhadů a jiné.

IES FSV UK, ve kterých se s daty pocházejících z výběrového šetření pracovalo. Budou použity tři práce bakalářské a jedna diplomová. Tyto práce poslouží jako podklad pro ilustraci konkrétních výběrových metod.

Prvním cílem je poskytnout čtenáři základní přehled teorie výběru z konečných populací. Studenti ekonomických oborů často pro své akademické práce potřebují data. Pokud ovšem veřejné databáze požadované informace neobsahují, vlastní sběr dat může představovat jediné východisko. Data mohou sloužit k mnohým účelům. Často je ovšem snahou zjistit populační průměr jisté charakteristiky. Neznalost teorie výběrů z konečných populací nebo špatně zvolený způsob výběru mohou vést k značným nepřesnostem. Druhým cílem je u vybraných prací studentů IES nastínit možnost použití pravděpodobnostních a nepravděpodobnostních výběrů a popsat úskalí, která tomu brání.

2 Základní pojmy a techniky

Úplně na začátek uveďme jisté vymezení pole zájmu. Tento text se bude výhradně zabývat *deskriptivním výběrovým šetřením*. Při něm se odhadují předem definované charakteristiky populace. Příkladem může být zjišťování mediánu[2] měsíčních příjmů absolventů IES [12]. Dalším typem je *analytické výběrové šetření*. Má za cíl nacházet vztahy, příčiny a následky. Pro účely analýzy jsou často zvoleny dva výběrové soubory. Teoretickými příklady jsou experimenty, kohortní studie, studie případů a kontrol a jiné. Konkrétním příkladem experimentu může být behaviorální studie [15]. Byly vybrány dvě skupiny studentů. Každý člen první skupiny dostal hrnek, kdežto členové druhé skupiny dostali pero. Všem byla nabídnuta možnost svůj předmět vyměnit za jiný a k tomu navíc získat 5 centů. Experiment zkoumal stálost lidských preferencí.

2.1 Historie

V díle významného antického historika Herodota s názvem Dějiny (cca 440 př.n.l.) je popsána metoda, jakou perský král odhadl velikost své armády pří invazi do Řecka[3]. Velký soubor vojáků o předem zvoleném počtu se seskupil blízko sebe a po jejich stranách se postavil plot. Následně celá armáda prošla mezi těmito ploty. Počet skupin byl vynásoben předem zvoleným číslem a to počtem vojáku ve výchozím souboru. Výsledný odhad činil 1 700 000 vojáků. Toto číslo je dnešními historiky považováno za přehnané. Ani ne tak kvůli selhání metody odhadu, ale spíše kvůli autorově

[2]Medián představuje důležitý deskriptivní údaj o populaci. Tato práce se ovšem bude zaobírat výhradně metodami odhadu populačního průměru a úhrnu.

[3]Určování velikosti populace patří mezi partie teorie výběrů z konečných populací. My se jím ovšem v tomto textu zabývat nebudeme. U popsaných technik výběru se znalost velikosti populace buď předpokládá, nebo ji pro výběr nepotřebujeme.

tendenci zveličovat [22, s. 7].

Jedním z prvních doložených pokusů o využití výběrů bylo zjišťování počtu obyvatel Londýna v roce 1662. Strůjcem tohoto pokusu byl John Graunt (1620-1674), který je dnes považován za jednoho z prvních demografů. Jeho postup byl důmyslný. Zjistil, že v částech Londýna, kde se vedly záznamy o počtu pohřbů, připadaly zhruba 3 pohřby na 11 rodin ročně. Tento poměr následně předpokládal ve všech částech Londýna. Dále využil dostupné údaje o tom, že ročně se v Londýně koná 13 000 pohřbů. Tyto informace daly dohromady fakt, že v Londýně bylo okolo 48 000 rodin. Předpoklad, že rodina se v průměru skládá z 8 členů, dává výsledný odhad populace na 384 000. Jeho metoda ovšem nebyla pravděpodobnostní a tedy nemohl učinit žádné závěry o její přesnosti.

Významné aplikace se prováděly v průzkumech veřejného mínění. Důležitou roli v tomto oboru sehrál George Gallup (1901-1984). Ten v roce 1934 založil American Institute of Public Opinion (později Gallup organization). V roce 1936 se na základě kvotního výběru o velikosti 2 000 osob této agentuře povedlo správně předpovědět výsledek prezidentských voleb. Velkou pozornost to vyvolalo hlavně z důvodu, že časopis Literary Digest výsledek voleb předpověděl chybně. A to i přesto, že vyhodnotil přes 2 000 000 odpovědí. Problém byl nejspíš v nereprezentativitě vzorku a malé návratnosti [18, s. 8]. Také předpovědi Gallup organization však nebyly vždy správné. Jako důkaz lze uvést prezidentské volby v USA v letech 1948, 1976 a 2012 [9].

Díky mnohostrannému použití se začala kolem výběrů z konečných populací vyvíjet matematická teorie. Za základní kámen se dá považovat článek Jerzyho Neymana [19]. V něm byla popsána metoda prostého náhodného výběru a diskutována otázka

reprezentativního vzorku. V následujících letech se začaly zkoumat další typy výběrů a odhadů. Vznikala také nová paradigmata. Mezi významné osobnosti teorie pravděpodobnostních výběrů patří například William G. Cochran ([3]). V českém prostředí se o rozvoj této teorie výrazně zasloužil profesor MFF UK Jaroslav Hájek ([11]).

2.2 Fáze výběrového šetření

Výběrové šetření má několik fází. Cochran [3, s. 4] uvádí tyto:

1. Stanovení cílů
2. Určení základního souboru
3. Zvolení formální stránky
4. Stanovení požadované přesnosti
5. Určení způsobu sběru dat
6. Získání opory výběru[4]
7. Určení metody výběru vzorku
8. Pilotní studie
9. Uskutečnění a organizace výběrového šetření
10. Odhady, analýza dat a shrnutí
11. Doporučení pro budoucí šetření

Tato práce se bude hlavně zaměřovat na kroky 7 a 10. Společně se o těchto dvou fázích mluví jako o výběrovo-odhadové strategii

[4]Opora výběru znamená soupis populace, který je popřípadě doplněn o kontaktní informace.

[24, s. 49]. Vzhledem k tomu, že zvolení výběrové metody předchází tvorbě odhadů, budou právě různé výběrové plány hlavními tematickými okruhy v dalším textu.

Metody výběru lze rozdělit do dvou základních skupin. První skupinu představují *pravděpodobnostní výběry*. U nich se volba výběrového souboru převede na matematický model. Každému možnému výběrovému souboru je přiřazena konstantní pravděpodobnost, že bude zvolen. Následně se na základě těchto pravděpodobností jeden výběrový soubor zvolí (respektive vygeneruje náhodným/pseudonáhodným procesem) a s ním se dále pracuje. Druhou skupinu tvoří *nepravděpodobnostní výběry*. Nabízí se hned několik definic. Jeřábek [13, s. 44] jej popisuje jako výběr, u kterého nemáme pro odhad chyby matematicko-statistickou oporu. Disman [7, s. 111] uvádí, že výběr není založený na teorii pravděpodobnosti, ale na logickém úsudku. Levy [17, s. 20] jej definuje jako výběr, u kterého jednotky populace nemají předem stanovenou pravděpodobnost zahrnutí. S ohledem na výše uvedené definice budeme nepravděpodobnostní výběr chápat jako výběr, který není pravděpodobnostní.

2.3 Nepravděpodobnostní výběry

Nyní uvedeme základní a nejčastěji používané typy nepravděpodobnostních výběrů.

Kvótní výběr (Quota sampling)

Kvótní[5] výběr je založený na apriorní představě o složení populace. Často tato představa vychází z předchozích výzkumů. Složení výběrového souboru by pak mělo odpovídat složení populace. Při výběru se postupuje ve třech krocích [13, s. 47].

[5]Slovo *kvóta* znamená stanovený počet (absolutní kvóta) nebo podíl (relativní kvóta).

1. Zvolíme znaky (pohlaví, vzdělání, věk atd.)

2. Dle relevatních dostupných statistických údajů vyhledáme jejich zastoupení v základním souboru a stanovíme pro ně kvóty.

3. Vybereme požadovaný vzorek podle stanovených kvót. Pokud chceme, aby znaky byly vzájemně závislé, vybrané jednotky musí splňovat i vzájemné kvóty (lidé stejného pohlaví musí odrážet i rozložení populační vzdělanosti). Nebo nezávisle, kde stačí naplnit každou z kvót bez ohledu na jejich vztah.

Pokud původní představa o složení populace alespoň částečně odpovídá realitě, dá se považovat kvótní výběr za nejspolehlivější typ nepravděpodobnostního výběru [7, s. 111].

Účelový výběr (Purposive sampling)

Při tomto výběru je jediným kritériem libovůle výzkumníka. Ten volí do vzorku jednotky, které podle něj reprezentují celou populaci. Lze jej využívat i v situacích, kdy nemáme přesnou představu o populaci. Příkladem takové populace mohou být zákazníci jistého obchodu. Účelovému výběru nelze upírat jeho užitečnost, avšak příliš zobecňovat závěry získané jeho použitím by mohlo být zavádějící.

Náhodilý výběr (Haphazard sampling)

Jednotlivé jednotky populace jsou vybrány čirou náhodou. Nevyskytuje se zde žádné vědomé plánování. Nejedná se ovšem o případ, kdy by měly všechny výběrové soubory stejnou pravděpodobnost zvolení. Této podmínce vyhovuje prostý náhodný výběr (PNV), který bude podrobně popsán v podkapitole 3.2.

Pohodlnostní výběr (Convenience sampling)

Ten, jenž je pověřen sběrem dat, si vybírá jednotky populace, které jsou pro něj nejdostupnější. Jen těžko lze hodnotit reprezentativnost takto vybraného vzorku. Na druhé straně lze tímto výběrem velmi rychle získat data.

Anketa (Self-selecting poll)

V případě ankety[6] je do výběru zahrnut každý, kdo se jí chtěl zúčastnit. Jedná se o tzv. samovýběr respondentů. Ústřední otázkou je, na jakou populaci můžeme její výsledky generalizovat. Zpravidla platí, že ať už vztaženy na libovolnou populaci, výsledky ankety jsou velmi zavádějící.

Technika sněhové koule (Snowball sampling)

Metoda, při které se výběrový soubor určuje postupně. Již vybraná jednotka navrhne jednotky další. Tento proces se zpravidla opakuje dokud není výběrový soubor tzv. teoreticky nasycen [7, s. 114]. To odpovídá momentu, když všechny odkazy vedou k již dotázaným jednotkám. Využívá se v případech, kdy je cílová populace neznámá nebo by bylo pracné vytvořit její soupis.

2.4 Paradigmata a srovnání metod výběru

Na tomto místě je třeba se zmínit o dvou hlavních paradigmatech teorie výběrů z konečných populací[7]. Liší se v pohledu na to, kdy vstupuje do procesu výběru prvek náhody. Jedno možnost použití

[6] Pojem anketa lze také chápat v širším smyslu. Při něm anketa neznamená pouze metodu výběru, ale označuje jakékoliv dotazování standardizovanými technikami, většinou s použitím dotazníku [26, s. 76]. Takto je anketa chápána širokou veřejností. My ji budeme ovšem chápat v užším smyslu – jako metodu výběru popsanou výše.

[7] Existuje i paradigma, které je v podstatě spojením design-based a model-based přístupu. V literatuře se často nazývá *model assisted design-based* přístup.

matematických metod u nepravděpodobnostních výběrů zamítá, druhé ne.

- **Design-based přístup**

  Předpokládá, že populační hodnoty jsou konstanty. Prvek náhody zde představuje proces vybírání výběrového souboru. Tento přístup má svůj původ v článku [19] a v teorii výběrů z konečných populací se považuje za klasický [21]. O tomto přístupu bude pojednávat další text.

- **Model-based přístup**

  Chápe populační hodnoty sledovaných jednotek $y_1, y_2, ...y_N$ jako realizaci náhodných veličin. To odpovídá standardnímu statistickému modelu s „nekonečnou populací“. Není důležité, jak se vybírá výběrový soubor. Tedy to může být i nepravděpodobnostní metoda. Primární snahou je zjistit parametry sdruženého rozdělení a následně je využít pro predikci populačních charakteristik (úhrn, průměr), které jsou také náhodnými veličinami. V dalším textu se tímto přístupem řídit nebudeme. Důvodem jsou velmi silné předpoklady konkrétních modelů [18, s. 57], které často neodpovídají realitě.

S ohledem na to, že uvažujeme design-based přístup, pokusme se nyní porovnat obě skupiny – pravděpodobnostní a nepravděpodobnostní výběry. Vzhledem k tomu, že konkrétní pravděpodobnostní výběry budou uvedeny až v dalším textu, popíšeme jen rozdíly obecné. Nepravděpodobnostní výběr je mnohem častěji využíván v praxi. Důvodem je jeho nenáročnost – ať už po technické, logistické nebo personální stránce. Navíc základní typy pravděpodobnostních výběrů předpokládájí, že máme úplný seznam naší populace (oporu výběru). Častokrát je ovšem téměř nemožné tuto oporu sestavit. Za určitých podmínek může nepravděpodobnostní

výběr poskytnout užitečné výsledky a některé jeho formy jsou částečně odolné proti neodpovědi. Jeho největší slabinu ovšem představuje fakt, že o přesnosti odhadu populačních charakteristik nemáme žádné ponětí a může při něm docházet k systematickému vychýlení. Zde se nachází síla pravděpodobnostních výběrů. S těmito výběry lze pracovat jako s matematickými objekty. Následně lze vytvářet odhady (bodové i intervalové) o charakteristikách populace a stanovovat jejich spolehlivost. Dále je možné určovat požadovanou velikost vzorku, abychom s předem zvolenou spolehlivostí a tolerovanou chybou odhadli jistý populační parametr.

Rozlišují se dva druhy chyb, které mohou vzniknout v různých fázích výběrového šetření. Obě mají vliv na přesnost odhadů. Rozdíl skutečné populační hodnoty a výběrového odhadu se nazývá *výběrová chyba*. Výběrová chyba je pouhým odrazem toho, že se na základě výběrového souboru snažíme udělat závěr o celé populaci. Různé výběrové soubory mohou vést k jiným odhadům. U pravděpodobnostních výběrů lze tuto chybu zachytit v pojmech očekávaná hodnota, rozptyl a průměrná čtvercová odchylka. U výběrů nepravděpodobnostních lze o této chybě pouze spekulovat. Existují ovšem i *nevýběrové chyby*. Nejčastěji se jako jejich příčiny uvádějí chyby měření, chyby vyhodnocovací, nezískání informace, získání nepravdivé informace, špatná opora výběru (často undercoverage[8] nebo overcoverage[9]), nevhodně navržený dotazník, samovýběr atd. Některé tyto chyby mohou ovšem nastat i při censu. Zvětšování výběrového souboru zde může mít negativní efekt. Čím větší je náš výběrový soubou (a tím pádem

[8]K undercoverage dochází v situaci, kdy náš soupis populace je podmožinou skutečného soupisu populace.

[9]O overcoverage se jedná, když je náš soupis populace nadmnožinou skutečného soupisu populace.

rozsah výběrového šetření se všemi náležitostmi), tím větší je náchylnost na nevýběrové chyby [10, s. 133].

2.5 Matematický základ pro pravděpodobnostní výběry

Úkolem této sekce je vybudování matematického aparátu pro pravděpodobnostní výběry. Skrze celou práci bude uváděno několik důležitých výsledků, které ovšem nebudou doplněny důkazem. Tato kapitola bude částečně kopírovat strukturu matematických textů. V dalších kapitolách tomu tak nebude. Důvodem takto zvoleného výkladu pravděpodobnostních výběrů je cíl a rozsah práce. Tím je poskytnout základní přehled pravděpodobnostních metod, popsat jejich kladné a záporné stránky a diskutovat vhodné použití v praxi. Přesné znění teoretických výsledků doplněných o důkazy je možno dohledat v textech [3], [24], [11] a dalších. V následujícím textu se na ně již nebude jednotlivě odkazovat.

Pro porozumění se předpokládá znalost základních partií matematické statistiky a pravděpodobnosti. A to především pojmů pravděpodobnost, náhodná veličina a její rozdělení, očekáváná hodnota a rozptyl. U odhadu populačních parametrů se pracuje s pojmy jako bodový a intervalový odhad.

Středem zájmu bude konečná skupina jednotek. Pro snazší zacházení přiřadíme každé z nich jednoznačně přirozené číslo.

Definice 1 (Základní soubor)**.** Základním souborem rozumíme množinu $\mathbb{U} = \{1, 2, 3, ..., N\}$.

Alternativně se používá pojmu *populace*.

Definice 2 (Velikost základního souboru)**.** Počet jednotek N základního souboru nazveme velikost základního souboru.

U jednotek základního souboru nás bude zajímat jejich specifická vlastnost. Respektive číselná hodnota této vlastnosti. Označ-

me tyto hodnoty $y_1, y_2, y_3, ..., y_N$. Přirozeně se tedy zjišťuje průměr, úhrn nebo rozptyl.

Definice 3 (Populační úhrn)**.** Úhrnem rozumíme hodnotu $Y = \sum_{i=1}^{N} y_i$.

Definice 4 (Populační průměr)**.** Průměrem rozumíme hodnotu $\overline{Y} = \sum_{i=1}^{N} \frac{y_i}{N}$.

Definice 5 (Populační rozptyl)**.** Populačním rozptylem rozumíme $S^2 = \sum_{i=1}^{N} \frac{(y_i - \overline{Y})^2}{N-1}$.

Definice 6 (Množina všech výběrových souborů)**.** Potenční množinu základního souboru $\mathbb{U}$ nazveme množinu všech výběrových souborů a budeme ji značit symbolem $\mathbb{S}$.

Prvkům množiny $\mathbb{S}$ budeme říkat výběrové soubory a zpravidla je budeme značit písmenem s. Podle známé věty existuje 2^N různých výběrových souborů.

Definice 7 (Výběrový plán)**.** Výběrovým plánem rozumíme množinovou funkci $P : \mathbb{S} \to [0, 1]$ splňující podmínku $\sum_{s \subset \mathbb{U}} P(s) = 1$.

Věta 1. *Mějme pevně zadaný výběrový plán P a množinovou funkci $\mu : 2^{\mathbb{S}} \to [0, 1]$. Nechť dále platí, že $\mu(\{s_1, s_2, ..., s_k\}) = \sum_{i=1}^{k} P(s_i)$, kde $\{s_1, s_2, ..., s_k\} \in 2^{\mathbb{S}}$. A dále $\mu(\emptyset) = 0$. Potom μ je pravděpodobnost.*

Platí tedy, že trojice $(\mathbb{S}, 2^{\mathbb{S}}, \mu)$ je pravděpodobnostní prostor. Různé výběrové plány tedy definují různé pravděpodobnostní prostory.

V dalším textu budeme u různých metod výběru uvádět výběrový plán explicitně nebo implicitně. Explicitně tomu bude například u prostého náhodného výběru. Každému výběrovému souboru s bude přiřazena jeho pravděpodobnost $P(s)$. U jiných by

ale explicitní zápis byl zbytečně zdlouhavý a nepřehledný. Proto jej nadefinujeme implicitně. To v podstatě bude odpovídat realizaci tohoto plánu.

Definice 8 (Rozsah výběru)**.** Náhodnou veličinu $K : \mathbb{S} \to \mathbb{N}$ rovnou počtu prvků obsažených ve výběrovém souboru nazveme rozsah výběru.

Definice 9 (Pravděpodobnost zahrnutí jednotky i)**.** Nechť $i \in \mathbb{U}$. Číslo $\pi_i = \sum_{s \ni i} P(s)$ nazveme pravděpodobnost zahrnutí jednotky i.

Definice 10 (Pravděpodobnost zahrnutí dvou jednotek i,j)**.** Nechť $i, j \in \mathbb{U}$. Číslo $\pi_{ij} = \sum_{s \ni i,j} P(s)$ nazveme pravděpodobnost zahrnutí jednotek i, j.

Pravděpodobnost zahrnutí hraje důležitou roli při konstrukci odhadů a jejich rozptylů při různých výběrových plánech. Může se také stát, že právě pravděpodobnosti zahrnutí jsou jediným vstupem, dle kterého se výzkumník při sestavování výběrové metody řídí. V podstatě musí sestrojit výběrový plán zaručující požadované pravděpodobnosti zahrnutí. Tato úloha ovšem není jednoznačná.

Pro odhadování populačního parametrů se přirozeně dá volit z mnoha různých typů odhadů. My ovšem budeme uvažovat pouze třídu odhadů uvedených v další definici.

Definice 11 (Horvitzův-Thompsonův odhad)**.** Mějme výběrový plán s kladnými pravděpodobnostmi zahrnutí π_i, pro $i = 1, ..., N$. Náhodnou veličinu $\widehat{Y_{HT}} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ nazveme Horvitzův-Thompsonův (H-T) odhad úhrnu. A náhodnou veličinu $\overline{y_{HT}} = \sum_{i=1}^{n} \frac{y_i}{N\pi_i}$ nazveme Horvitzův-Thompsonův odhad průměru.

Věta 2. *H-T odhady jsou nestrannými odhady úhrnu a průměru.*

H-T odhad je tedy univerzálním způsobem, jak nalézt nestranné odhady populačních parametrů. Lze jej aplikovat na jakýkoliv výběrový plán s kladnými pravděpodobnostmi zahrnutí všech jednotek populace. Při výkladu konkrétních plánů budou H-T odhady využity. Jejich zápis se ovšem často zjednoduší.

Věta 3. *Pro rozptyl H-T odhadu úhrnu platí*

$$V(\widehat{Y_{HT}}) = \sum_{i=1}^{N} \frac{{y_i}^2}{\pi_i}(1-\pi_i) + \sum_{i=1}^{N}\sum_{j\neq i} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j)$$

a má-li výběrový plán pevnou velikost výběrového souboru, potom se vzorec redukuje na

$$V(\widehat{Y_{HT}}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2(\pi_i\pi_j-\pi_{ij}). \qquad (1)$$

Připomeňme, že rozptyl H-T odhadu průměru lze získat použitím základních pravidel pro počítání s rozptylem. Vztah (1) nám dává také „návod“, jak ideálně volit pravděpodobnosti zahrnutí, abychom minimalizovali rozptyl odhadů. Chceme, aby jednotky s vysokou hodnotou sledované proměnné měly menší pravděpodobnost zahrnutí. A také chceme, aby jednotky s nízkou hodnotou měly větší pravděpodobnost zahrnutí. Respektive požadujeme, aby poměr $\frac{y_i}{\pi_i}$ byl konstantní pro všechny jednotky populace. Jak je ze vzorce zřejmé, narážíme na dva problémy.

1. Populační hodnoty y_i neznáme. Jejich velikost lze pouze předvídat. Z toho důvodu je nutno i pro rozptyly odhadů (čísla) používat odhady (náhodné veličiny).

2. Pravděpodobnosti zahrnutí dvou prvků π_{ij} mohou být komplikované pro daný výběrový plán.

Nadefinujme si základní náhodné veličiny, které budeme v dalším textu často užívat přímo či nepřímo k odhadům populačních parametrů. Definiční obor těchto náhodných veličin je množina všech

možných výběrových souborů $\mathbb{S}$. Pro prostý náhodný výběr se první dvě shodují s H-T odhady průměru a úhrnu.

Definice 12 (Výběrový průměr)**.** Výběrovým průměrem rozumíme hodnotu $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$.

Definice 13 (Odhad populačního úhrnu)**.** Náhodnou veličinu $\widehat{Y} = N\overline{y}$ nazveme odhadem populačního úhrnu.

Následující náhodná veličina se také často bude vyskytovat v dalším textu.

Definice 14 (Výběrový rozptyl)**.** Výběrovým rozptylem rozumíme $s^2 = \sum_{i=1}^{n} \frac{(y_i - \overline{y})}{n-1}$

Připomeňme, že pro různé výběrové soubory $s \subset \mathbb{U}$ mohou výše nadefinované náhodné veličiny nabývat různých hodnot. V tomto ohledu se liší od svých populačních protějšků, jenž jsou pouhými konstantami, které se snažíme zjistit.

Na závěr úvodní kapitoly uveďme terminologické poznámky. Slova *vzorek* a *výběrový soubor* budou chápána jako synonyma. To stejné platí pro *populaci* a *základní soubor*. Naproti tomu slova *výběr* a *výběrový soubor* nechápeme jako synonyma. Výběr (sampling) je proces, kterým získáváme výběrový soubor (sample). Výběr lze též nazývat *vzorkování*. Na místě je také vyjasnění smyslu slova *odhad*. Odpovídá totiž dvěma různým anglickým pojmům. *Estimator* je náhodná veličina. Jedná se o předpis dávající návod, co se sesbíranými daty udělat. Kdežto *estimate* je číselná hodnota pro konkrétní realizaci náhodné veličiny.

3 Teorie pravděpodobnostních výběrů

Před vlastním popisem různých pravděpodobnostních výběrů bude stručně nastíněna problematika normální aproximace. Ta se týká všech výběrových plánů respektive jejich odhadů. Dále se jako první popíše *prostý náhodný výběr* (simple random sampling). Ten slouží jako teoretický základ pro téměř všechny uvedené plány a proto mu bude věnována největší pozornost[10]. V celém oddílu budeme uvažovat neexistenci nevýběrové chyby. A to hlavně v tom smyslu, že existuje správný soupis populace, pro vybraný vzorek lze získat potřebnou informaci od všech jeho jednotek a tato informace je pravdivá.

3.1 Normální aproximace

Pro konstrukci intervalů spolehlivosti při různých výběrových plánech je nutno znát rozdělení příslušných odhadů. Tato rozdělení jsou diskrétní, ale v naprosté většině případů nepatří do třídy základních diskrétních rozdělení. Používá se tedy normální aproximace. Teoretické výsledky týkající se normální aproximace jsou ovšem z důvodu technické náročnosti nad rámec této práce. Hlavní otázku v praxi představuje velikost výběrového soubou, která zaručí kvalitní aproximovaci. Lohr [18, s. 43] podotýká, že „magická“ hranice $n = 30$ často v kontextu konečných populací nestačí. To úzce souvisí se samotným rozdělením hodnot v populaci. Pokud populační hodnoty dané proměnné nemají rozdělení blízké normálnímu, přirozeně se musí pro normální aproximaci odhadu vy-

[10]Při výkladu PNV bude krom odhadu průměru a úhrnu podrobně popsána konstrukce intervalů spolehlivosti, odhad relativní a absolutní četnosti a určování velikosti výběrového souboru. Také bude vysvětleno tzv. odhadování rozptylu odhadů ze vzorku. U jiných výběrových plánů se ovšem omezíme jen na odhady průměru a úhrnu. Důvodem je analogie k PNV a cíle práce. Pro podrobnější výklad je čtenář odkázán na literaturu uvedenou v podkapitole 2.5.

užít většího vzorku. Tato situace není ryze teoretická. Často populační charakteristiky vykazují tendenci k šikmosti. Příkladem z oblasti financí mohou být výnosy akcií nebo jiných aktiv.

Přiložený Obrázek 1 udává četnosti tříročních výnosových měr podílových fondů společnosti Fidelity Investments. Data využitá při jeho konstrukci jsou zapsána v Tabulce 1. Celkový počet zahrnutých podílových fondů je $N = 264$. Průměrný výnos činí zhruba 7,8%. Kdežto medián je 7,25%. Koeficient šikmosti se rovná 0,42.

Jeden z explicitních vzorců pro výpočet požadované velikosti vzorku při PNV odhadu průměru je tzv. Cochranovo pravidlo[11] [3, s. 42]:

$$n_{min} = 25 \left(\frac{\sum_{i=1}^{N} (y_i - \overline{Y})^3}{NS^3} \right)^2 .$$

Existují ovšem i jiná doporučení. Lohr [18, s. 44] udává:

$$n^*_{min} = 28 + 25 \left(\frac{\sum_{i=1}^{N} (y_i - \overline{Y})^3}{NS^3} \right)^2 .$$

Číslo

$$\frac{\sum_{i=1}^{N} (y_i - \overline{Y})^3}{NS^3}$$

je populačním koeficientem šikmosti.

Vzhledem k tomu, že zpravidla populační průměr a populační rozptyl neznáme (snažíme se je odhadnout), tak pro výše uvedené vzorečky užíváme jejich výběrové protějšky. Kish [14, s. 16] uvádí,

[11] Cochran ve své knize uvádí vzorec s jinou mírou populačního rozptylu. Jedna se o $\sigma^2 = \sum_{i=1}^{N} \frac{(y_i - \overline{Y})^2}{N}$. Jeho odlišost od S^2 je ovšem zanedbatelná. Proto tyto dvě různé míry populačního rozptylu nerozlišujeme.

že v praxi je většinou chyba způsobená aproximací menší než jiné zdroje chyb.

3.2 Prostý náhodný výběr

Mějme pevně zadané $n \in \{1, 2, ..., N\}$. Položme

$$P(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{pokud } K(s) = n \\ 0 & \text{pokud } K(s) \neq n \end{cases}$$

pro $s \in \mathbb{U}$.

Snadno se lze přesvědčit, že daný předpis definuje výběrový plán. Dále platí, že $\forall i \in \mathbb{U} : \pi_i = \frac{n}{N}$. Takovou vlastnost nazýváme rovnoměrnost. Tento plán modeluje situaci, kdy chceme vybrat n jednotek z populace o velikosti N. Za předpokladu, že všech $\binom{N}{n}$ různých výběrových souborů o velikosti n má stejnou šanci být vybráno. Každý má pravděpodobnost vybrání $\frac{1}{\binom{N}{n}}$.

Existují v podstatě 3 způsoby, jak realizovat PNV. Každý z nich předpokládá existenci opory výběru.

- **Statistický software**

  Vzhledem k tomu, že základní populaci reprezentujeme přirozenými čísly, výběr prostého náhodného vzorku je ekvivalentní výběru n jednotek z N bez vracení a stejnými pravděpodobnostmi. Tato operace je součástí mnoha moderních statistických programů. Vygenerovaná čísla jsou pseudonáhodná, jelikož je generuje deterministickým způsobem jistý algoritmus.

- **Tabulka náhodných čísel**

  Jedná se v podstatě o sérii sloupců. Každý sloupec je složen z čísel o stejném množství cifer. Výběr počátečního sloupce a čísla je libovolný. Následně se postupuje po řadě. Do vzorku

jsou zahrnuty jednotky, které jsou reprezentovány těmito čísly.

- **Postupné vytahování**

  Pomocí PC, tabulky náhodných čísel nebo simulovaným náhodným pokusem (např. tahy čísel z klobouku, pokud to rozsah populace umožňuje) se n-krát provádí los čísel od 1 do N. Podmínkou ovšem je, že v každém tahu se již vytažené číslo nevrací a v každém tahu mají čísla stejnou pravděpodobnost vytažení. V prvním tahu je tedy tato pravděpodobnost $\frac{1}{N}$, v druhém $\frac{1}{N-1}$ atd.

Odhady průměru a úhrnu

Zkoumejme nyní metodu odhadu v kontextu prostého náhodného výběru. Zde si vystačíme s náhodnými veličinami zavedenými v podkapitole 2.5. Platí, že výběrový průměr $\overline{y}$, odhad populačního úhrnu $\widehat{Y}$ a výběrový rozptyl s^2 jsou při PNV nestranné odhady populačního průměru $\overline{Y}$, populačního úhrnu Y a populačního rozptylu S^2.

Máme tedy nástroj pro vytváření bodových odhadů. K sestrojení intervalových odhadů budeme potřebovat znát ještě rozptyl příslušných odhadů a použít normální aproximaci. Lze ukázat, že pro rozptyl $\overline{y}$ platí:

$$V(\overline{y}) = \frac{S^2}{n}(1 - \frac{n}{N}).$$

Z toho z pravidel pro počítání s rozptylem plyne:

$$V(\widehat{Y}) = \frac{N^2 S^2}{n}(1 - \frac{n}{N}).$$

Uvedené vzorce mají jeden velký nedostatek. Často neznáme populační rozptyl S^2. K jeho zjišťování používáme nestranný bo-

dový odhad s^2. Následně tedy platí, že náhodné veličiny

$$\widehat{V}(\overline{y}) = \frac{s^2}{n}(1 - \frac{n}{N})$$

a

$$\widehat{V}(\widehat{Y}) = \frac{N^2 s^2}{n}(1 - \frac{n}{N})$$

jsou nestrannými odhady rozptylu výběrového průměru $\overline{y}$ a $\widehat{Y}$.

Zkoumejme nyní faktory, které ovlivňují velikost rozptylu u obou odhadů $\overline{y}$ a $\widehat{Y}$.

- Velikost výběrového souboru n – Nepřímý vliv. Zvětšováním výběrového souboru snižujeme rozptyl.
- Velikost základní populace N – Přímý vliv. Větší populace znamená větší rozptyl.
- Populační rozptyl S^2 – Přímý vliv. Větší populační rozptyl způsobuje větší rozptyl našich odhadů.

Výraz $f = \frac{n}{N}$ nazveme *výběrový zlomek* (sampling fraction). Číslo $(1-f)$ nazveme *konečnostní násobitel* (finite population correction factor). Pokud je výběrový zlomek blízko 0 (do 0,1), lze jej v praxi ignorovat [3, s. 25]. Název konečnostní násobitel je odvozen z faktu, že ve standardních statistických úlohách s „nekonečnou populací“ je rozptyl výběrového průměru náhodného výběru roven $\frac{S^2}{n}$. V porovnání s konečnou populací se tedy liší o faktor $(1 - f)$. Pro konečnou populaci z toho plyne získání vesměs lepších odhadů průměru a úhrnu.

Předchozích poznatků o rozptylu využíváme k sestrojování intervalů spolehlivosti pro úhrn a průměr. Předpokládame-li, že $\frac{\overline{Y}-\overline{y}}{V(\overline{y})}$ a $\frac{\widehat{Y}-Y}{V(\widehat{Y})}$ mají normované normální rozdělené, potom příslušné konfideční intervaly vypadají následovně:

$$\left(\overline{y} - \frac{z_{\alpha/2}S}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}, \overline{y} + \frac{z_{\alpha/2}S}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}\right)$$

a

$$\left(\widehat{Y} - \frac{z_{\alpha/2}NS}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}, \hat{Y} + \frac{z_{\alpha/2}NS}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}\right).$$

Symbol $z_{\alpha/2}$ značí $(1 - \frac{\alpha}{2})$-kvantil normovaného normálního rozdělení. Symbolem S rozumíme odmocninu z populačního rozptylu S^2 a nazýváme ji *populační směrodatná odchylka*. Pokud populační směrodatnou odchylku neznáme, použijeme k jejímu odhadu výběrovou.

Odhady absolutní a relativní četnosti

Často se stává, že nějaká vlastnost rozděluje populaci na dvě části (muži/ženy, zaměstnaní/nezaměstnaní). V této situaci je naší snahou zjistit absolutní nebo relativní četnost jednotek, které tuto vlastnost mají/nemají. Zaveďme speciální značení.

A – populační absolutní četnost	a – výběrová relativní četnost
$P = \frac{A}{N}$ – populační relativní četnost	$p = \frac{a}{n}$ – výběrová relativní četnost

Přiřaďme každé jednotce populace, která má danou vlastnost hodnotu 1. A jednotkám nemající tuto vlastnost hodnotu 0. Díky tomu platí: absolutní četnost se rovná populačnímu úhrnu a relativní četnost se rovná populačnímu průměru. Toto pozorování nijak nezáviselo na typu výběrového plánu. Proto lze stejný „trik“ užít u všech pravděpodobnotních výběrů. V kontextu PNV jsme úhrn a průměr zkoumali v předchozí kapitole. Stačí tedy jen přepsat příslušné vzorečky pro odhady pomocí speciálního značení.

Populační rozptyl:

$$\begin{aligned} S^2 &= \sum_{i=1}^{N} \frac{(y_i - \overline{Y})^2}{N-1} = \sum_{i=1}^{N} \frac{(y_i - P)^2}{N-1} \\ &= \frac{\left(\sum_{i=1}^{N} y_i^2\right) - 2NP^2 + NP^2}{N-1} = \frac{\left(\sum_{i=1}^{N} y_i^2\right) - NP^2}{N-1} \\ &= \frac{\left(\sum_{i=1}^{N} y_i\right) - NP^2}{N-1} = \frac{N}{N-1} P(1-P). \end{aligned}$$

Výběrový rozptyl:

$$s^2 = \frac{n}{n-1} p(1-p).$$

Z poznatků v předchozí kapitole tedy plyne, že náhodná veličina p je nestranný odhad P a Np je nestraný odhad A. Pro rozptyl těchto odhadů platí:

$$V(p) = \frac{P(1-P)}{n} \frac{N-n}{N-1}$$

a

$$V(Np) = \frac{N^2 P(1-P)}{n} \frac{N-n}{N-1}.$$

Ve vzorcích pro rozptyl odhadů se objevuje neznámé P, proto je opět nutné tyto rozptyly odhadovat ze vzorku. Funkce $P(1-P)$ nabývá maxima pro $P = 0,5$. Z toho vyplývá: rozptyl odhadů bude tím větší, čím rovnoměrněji je populace rozdělena do dvou skupin podle dané vlastnosti.

Hlavním důvodem, proč se absolutní a relativní četností zabývat zvlášť je ten, že konfidenční intervaly příslušných odhadů lze v kontextu PNV teoreticky sestrojit přesně nebo při aproximaci použít jiného rozdělení než normálního. To plyne z faktu, že výběrová absolutní četnost a, má hypergeometrické rozdělení.

Určování velikosti výběrového souboru

V předchozím textu jsme n měli pevně zadané. Při výběrových šetřeních však určení velikosti vzorku představuje jeden z hlavních úkolů. Velikost vzorku se odvíjí podle toho, co očekáváme od našeho odhadu. Je několik požadavků, které charakterizují náš odhad. Pro ilustraci použijeme odhad populačního průměru – výběrový průměr. Pro odhady úhrnu, absolutní a relativní četnosti se postupuje analogicky. Je nutné ovšem předpokládat, že známe populační rozptyl S^2 a $\overline{y}$ má normální rozdělení.

Chceme, aby se bodový odhad populačního průměru $\overline{Y}$ od skutečné hodnoty lišil nejvýše o tolerovanou chybu d (margin of error) s předem zadanou spolehlivostí α. Zapsáno matematicky:

$$P(|\overline{y} - \overline{Y}| \leq d) = 1 - \alpha.$$

Výraz na levé straně předchozího vztahu lze ekvivalentně napsat jako $P(\frac{|\overline{y}-\overline{Y}|}{\sqrt{V(\overline{y})}} \leq \frac{d}{\sqrt{V(\overline{y})}})$. Z předpokladu normality $\overline{y}$ plyne, že $\frac{\overline{y}-\overline{Y}}{\sqrt{V(\overline{y})}}$ má normované normální rozdělení. Aby byla rovnost splněna stačí položit:

$$\frac{d}{\sqrt{V(\overline{y})}} = z_{\alpha/2}.$$

Což je to samé jako:

$$\frac{d}{\sqrt{\frac{S^2}{n}(1 - \frac{n}{N})}} = z_{\alpha/2}.$$

Vyjádřením n dostáváme:

$$n = \frac{z^2_{\alpha/2} S^2}{d^2 + \frac{z^2_{\alpha/2} S^2}{N}}. \tag{2}$$

Tento vzorec odpovídá intuitivní představě. Při pevně stanovené hladině spolehlivosti α se požadovaná velikost výběrového souboru zvětšuje s větším N a menším d.

Shrnutí a aplikace

Z teoretického hlediska představuje PNV základní stavební kámen pro další typy výběrů. V praxi se většinou užívá v případech, kdy o základním souboru nemáme žádné dodatečné informace. Jako příklad lze uvést soupis jmen studentů studujících jistý obor. Společne se systematickým výběrem je PNV v těchto typech úloh nejefektivnější [18, s. 58]. Díky jeho jednoduchosti se v porovnání s ostatními technikami výběru jeví jako intuitivní a příslušné vzorečky jsou snadné. Má ovšem několik úskalí.

- Jeho aplikace může být velmi drahá a časově náročná. V mnohých případech jsou jednotky populace rozmístěny v prostoru (například všichni důchodci v Německu atd.) a získávání informací o náhodně zvoleném výběrovém souboru se tak stává nákladným.

- PNV předpokládá, že máme k dispozici úplný seznam jednotek populace. To se ovšem často jeví jako nesplnitelný předpoklad [17, s. 75]. Zajímáme-li se při výzkumu i o podskupiny základního souboru, odhady charakteristik celé populace na základě PNV nám o těchto podskupinách nic neřeknou. Ve skutečnosti ani nemusí odpovídat skutečným hodnotám v těchto podskupinách.

- Rovnoměrné rozdělení pravděpodobností vybrání vzorku nemusí představovat nejefektivnější způsob sběru dat. To úzce souvisí s reprezentativitou. Příslušné odhady při PNV jsou sice nestranné, ale může dojít k vybrání extrémně nereprezentativního vzorku. Tato situace je krajně nežádoucí a pramení z velkého rozptylu odhadů. Způsob, jak proti těmto nereprezentativním vzorkům bojovat, je použití jiných výběrových plánů vylučujících jejich vybrání.

3.3 Stratifikovaný výběr

Stratifikovaný výběr[12] (stratified sampling) se provádí tak, že populaci rozdělíme do H podpopulací. Budeme užívat pojmu *stratum*. V každém z těchto strat vykonáme PNV. Celkově tedy dostáváme H výběrových souborů a jejich sjednocení je požadovaný finální výběrový soubor. Popsaný výběrový plán je přesněji nazýván prostý stratifikovaný výběr. Pro výběr v jednotlivých stratech lze použít i jiné výběrové plány než PNV. To ovšem teorii značně komplikuje a proto v tomto textu budeme uvažovat jen prostý stratifikovaný výběr (dále jen stratifikovaný výběr).

Mějme H strat $\mathbb{U}_i$, na které lze rozdělit populace. Všechna tato strata jsou navzájem disjunktní a $\mathbb{U} = \bigcup_{i=1}^{H} \mathbb{U}_H$. Označme N_i počet jednotek v i-tém podsouboru, pro $i = 1, ..., H$. Platí tedy, že $N_1 + ... + N_H = N$. Nechť n_i je požadovaná velikost vzorku s_i vybraného z i-tého podsouboru, pro $i = 1, ..., H$. Celková velikost vzorku tedy je $n = n_1 + ... + n_H$. Nechť každá jednotka populace má nějakou vlastnost y, která lze číselně vyjádřit. A nechť jsou jednotky v jednotlivých podsouborech očíslovány. Hodnotu j-té jednotky h-tého strata značíme

$$y_{jh}.$$

Populační úhrn h-tého strata

$$Y_h = \sum_{j=1}^{N_h} y_{jh}.$$

Nadefinujme náhodnou veličinu odhadující úhrn h-tého strata

$$\widehat{Y_h} = \sum_{j=1}^{n_h} \frac{N_h y_{jh}}{n_h}.$$

[12] V češtině je taky někdy označován jako *oblastní výběr*.

Populační průměr h-tého strata

$$\overline{Y_h} = \sum_{j=1}^{N_h} \frac{y_{jh}}{N_h}.$$

Výběrovým průměrem h-tého strata nazveme náhodnou veličinu

$$\overline{y_h} = \sum_{j=1}^{n_h} \frac{y_{jh}}{n_h}.$$

Populační rozptyl h-tého strata

$${S_h}^2 = \sum_{j=1}^{N_h} \frac{(y_{jh} - \overline{Y_h})^2}{N_h - 1}.$$

Výběrovým rozptylem h-tého strata nazveme náhodnou veličinu

$${s_h}^2 = \sum_{j=1}^{n_h} \frac{(y_{jh} - \overline{y_h})^2}{n_h - 1}.$$

Jak už víme z předchozí kapitoly, náhodné veličiny $\widehat{Y_h}, {s_h}^2, \overline{y_h}$ jsou při PNV nestranným odhadem $Y_h, {S_h}^2, \overline{Y_h}$.

U stratifikovaného výběru na rozdíl od PNV nemají obecně všechny jednotky populace stejnou pravděpodobnost zahrnutí π_i. Tato pravděpodobnost zahrnutí je shodná pouze v rámci strat a je rovna $\pi_i = \frac{n_h}{N_h}$. Pro odhad celkového populačního průměru budeme používat nasledující náhodnou veličinu:

$$\overline{y_{str}} = \sum_{i=1}^{H} \frac{N_i}{N} \overline{y_i}.$$

Budeme ji nazývat stratifikovaný výběrový průměr. Jedná se vlastně o vážený průměr výběrových průměrů přes všechny strata, kde jednotlivé váhy jsou $w_i = \frac{N_i}{N}$. Analogicky pro odhad celkového úhrnu používáme náhodnou veličinu:

$$\widehat{Y_{str}} = N\overline{y_{str}}.$$

V kontextu prostého stratifikovaného výběru je $\overline{y_{str}}$ je nestranným odhadem $\overline{Y}$ a $\widehat{Y_{str}}$ je nestranným odhadem Y.

Pro tyto odhady platí:

$$V(\overline{y_{str}}) = \sum_{i=1}^{H} \frac{{S_i}^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) \left(\frac{N_i}{N}\right)^2$$

a

$$V(\widehat{Y_{str}}) = \sum_{i=1}^{H} \frac{{S_i}^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) {N_i}^2.$$

Výše uvedené vzorce pro rozptyl odhadů ukazují, v čem je hlavní síla stratifikovaného výběru. Pokud se nám podaří základní soubor rozdělit na strata, ve kterých se hodnoty sledované proměnné výrazně neliší (rozptyl hodnot v rámci strat je nízký), pak se rozptyl odhadu sníží.

Proporcionální alokace

Zatím jsme měli velikosti vzorků ve stratech $n_1, ..., n_H$ pevně zadány. Ve skutečnosti je právě velikost těchto výběrových souborů jednou z prvních věci, které musí výzkumník určit. Toto určení probíhá na základě požadované přesnosti odhadu a lepší reprezentativity v porovnání s PNV. Proporcionální alokace představuje základní typ alokace. Popíšeme si ji a vyzdvihneme její výhody a nevýhody.

Tato alokace je založená na myšlence, že z větších strat bychom měli vybírat větší vzorky. Přesněji, aby velikost vzorku v každém stratu byla přesným odrazem podílu velikosti strata na celé populaci. Symbolicky tedy chceme docílit, aby pro všechna strata platilo:

$$\frac{N_i}{N} = \frac{n_i}{n}.$$

Což lze vyjádřit jako:

$$n_i = \frac{nN_i}{N}. \tag{3}$$

Ignorujme možnost vzniku nepřirozeného čísla. Při použití by se muselo zaokrouhlovat a analýza by byla složitější. U proporcionální alokace mají všechny jednotky populace stejnou pravděpodobnost zahrnutí. $\pi_i = \frac{n}{N}$. V tomhle ohledu se shoduje s PNV. Zkoumejme, jak se toto alokování odrazí na tvaru odhadu a rozptyl odhadů. Budeme uvádět jen vzorce pro průměr; u úhrnu se postupuje analogicky.

$$\begin{aligned}\overline{y_{str}} &= \sum_{i=1}^{H} \frac{N_i}{N} \overline{y_i} \\ &= \frac{\sum_{i=1}^{H} (\sum_{j=1}^{n_i} y_{ji})}{n}.\end{aligned}$$

$$\begin{aligned}V(\overline{y_{str}}) &= \sum_{i=1}^{H} \frac{{S_i}^2}{\frac{nN_i}{N}} \left(1 - \frac{n}{N}\right) \left(\frac{N_i}{N}\right)^2 \\ &= \sum_{i=1}^{H} \frac{{S_i}^2}{n} \left(1 - \frac{n}{N}\right) \left(\frac{N_i}{N}\right).\end{aligned}$$

Proporcionální alokace zaručuje reprezentativitu. Při PNV se může stát, že vybraný vzorek je naprosto nereprezentativní. Proto stratifikovaný výběr s proporcionální alokací představuje pravděpodobnostní protějšek kvótního výběru.

Alokace velikosti výběrových souborů závisí ryze na výzkumníkovi. Další základní alokací je tzv. *Neymanova alokace*. Pro ni platí:

$$n_h = \frac{N_H S_H}{\sum_{i=1}^{H} N_i S_i} n.$$

Na rozdíl od proporcionální alokace je v potaz brán i rozptyl v rámci jednotlivých strat a vede za jistých podmínek k nejnižšímu rozptylu odhadů. V podstatě se dá říci: při Neymanově alokaci volíme větší vzorky ve stratech, které jsou větší nebo mají

rozptýlenější hodnoty. V kontextu stratifikovaných výběrů určujeme požadovanou velikost celkového výběrového souboru n analogicky jako u PNV. To stejné platí o odhadech absolutní a relativní četnosti.

Shrnutí a aplikace

Užitím stratifikovaného výběru ve vhodných situacích lze docílit lepších odhadů. Vhodná situace nastává především tehdy, když se základní soubor dá rozložit na disjunktní straty, ve kterých by se hodnota sledované proměnné neměla příliš lišit.

Jako hlavní dvě výhody lze uvést:

- Stratifikovaný výběr představuje ochranu před velmi špatným/nereprezentativním vzorekem, který při PNV může být vybrán [18, s. 74].
- Může být méně logisticky a finančně náročný. Dále existuje možnost použít různé způsoby získání informací na různá strata.
- Jako důsledek vhodného rozřazení populace do podsouborů dostaneme odhady (průměru, úhrnu atd.) s menším rozptylem. Důvodem je nižší variablita proměnné v rámci jednotlivých podskupin ve srovnání s variabilitou proměnné v celé populaci.

Má ovšem i své nedostatky:

- Zvyšuje složitost celého procesu výběru.
- Nadefinování vhodných strat nemusí být jednoduché.
- Stejně jako u PNV se předpokládá existence úplného soupisu celé populace.

3.4 Vícestupňový skupinový výběr

Popišme nejprve *jednostupňový skupinový*[13] *výběr* (one-stage cluster sampling). Pro jednoduchost jej budeme označovat zkráceně skupinový výběr. Pří tomto výběru se nacházíme v situaci, kde základní populaci rozdělíme na H disjunktních podmnožin, jejichž sjednocení je celá populace. Nazvěme je skupiny. Obecněji jsou to primární výběrové jednotky – PVJ. Prvky základního souboru v tomto kontextu označme jako sekundární výběrové jednotky – SVJ. Dosud se tento postup shoduje se stratifikovaným výběrem. Zásadní rozdíl však leží ve způsobu selekce. Při skupinovém výběru prostým náhodným výběrem vybereme h PVJ. U všech vybraných PVJ provedeme census přes příslušné SVJ.

P-stupňový skupinový výběr (P-stage cluster sampling) v podstatě jen kopíruje výše uvedený proces, ale opakuje jej P-krát. Číslo P značí, na kolika úrovních provádíme PNV. Skutečnost, že v každé fázi používáme PNV, je pouze didaktické zjednoduššení. Vede k nejjednodušším odhadům a výpočtům. Někdy se používá v tomto případě název prostý vícestupňový skupinový výběr. Výzkumníkům nic nebrání v tom, volit různé výběrové plány v různých fázích. Mnohdy je to dokonce žádoucí.

Zkoumejme, jaké pravděpodobnosti zahrnutí mají jednotky základního souboru. Při skupinovém výběru platí: $\forall i \in \mathbb{U} : \pi_i = \frac{h}{H}$. U prostého dvoustupňového skupinového výběru je situace lehce komplikovanější. Odvíjí se od počtu sekundárních výběrových jednotek v příslušných PVJ, počtu terciárních výběrových jednotek v příslušných SVJ a velikosti vzorků v každém stupni. Pravděpodobnost zahrnutí je tedy násobkem výběrových zlomků v každém stupni. Jde tedy vidět, že až na speciální příklady, je prostý vícestupňový výběr nerovnoměrným výběrovým plánem.

[13] V některé literatuře se používá slovo *skupinkový*.

Popišme nyní odhady v kontextu skupinového výběru. Použijeme již zavedené značení u stratifikovaného výběru. Ostatní prosté vícestupňové skupinové výběry jsou komplikovanější. Zásadním poznatkem je ovšem fakt, že čím více stupňů má šetření, tím větší jsou rozptyly odhadů. Výsledný rozptyl vlastně sčítá variabilitu, jenž vzniká v každém stupni.

Skupinový výběr

Při skupinovém výběru se setkáváme s novým úkazem. Různé vzorky mohou mít různou velikost. Pokusme se zjednodušit Horvitzův-Thompsonův odhad úhrnu v kontextu skupinového výběru.

$$\widehat{Y_{sk}} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} \frac{y_i}{\frac{h}{H}} = \frac{H}{h} \sum_{i=1}^{n} y_i.$$

Pro H-T odhad průměru tedy platí:

$$\overline{y_{sk}} = \frac{H}{Nh} \sum_{i=1}^{n} y_i.$$

Dále platí:

$$V(\overline{y_{skup}}) = \left(\frac{H}{N}\right)^2 \frac{{S_a}^2}{h} \left(1 - \frac{h}{H}\right),$$

kde ${S_a}^2 = \frac{\sum_{i=1}^{H} \left(Y_i - \frac{Y}{H}\right)^2}{H-1}$.

Z předchozího vzorečku je zřejmé, že jsou-li úhrny podobné ve všech skupinách, rozptyl bude nízký. Při skupinovém výběru dostaneme tedy přesnější odhad, když jednotlivé skupiny jsou zmenšenými obrazy celé populace.

Systematický výběr

Předpokládejme, že $\frac{N}{n}$ je přirozené číslo a označme jej k. Budeme mu říkat *výběrový krok*. Při systematickém výběru postupujeme ve dvou krocích.

- Vybereme náhodně číslo z množiny $\{1, ..., k\}$. Označme jej b.

- Jako výběrový soubor zvolíme n-tici jednotek $\{b, b+k, ..., b+(n-1)k\}$

Celkový počet možných výběrových souborů je k, všechny jsou disjunktní a mají stejnou pravděpodobnost zvolení. Tento výběrový plán je ekvivalentní skupinovému výběru se stejně velkými skupinami. Platí tedy, že příslušný výběrový průměr

$$\overline{y_{sk}} = \frac{H}{Nh} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{y_i}{n}$$

je nevychýleným odhadem.

Porovnejme rozptyl odhadů při systematickém výběru a PNV. Vše se bude odvíjet od našeho seznamu jednotek – opory výběru. Záleží na vztahu pořadí jednotek a jejich hodnot [18, s. 196].

- Seznam je náhodný. V tomto případě nehraje pořadí na velikost charakteristik jednotek žádnou roli. V tomto případě se rozptyl shoduje s odhadem při PNV.

- Seznam je sestupný/vzestupný. To odpovídá situaci, kde s rostoucím číslem jednotky roste/klesá hodnota jejího sledovaného znaku. Systematický výběrový soubor se nemůže skládat z jednotek, jejichž číselné označení jsou blízko sebe. PNV takovou situaci ovšem nevylučuje. Je tedy vidět, že rozptýlení hodnot pro odhad při PNV bude větší.

- Seznam je periodický. Mohlo by se stát, že každá jednotka vzorku by měla stejnou hodnotu. To by v extrémním případě znamenalo, že rozptyl odhadu při systematickém výběru se rovná populačnímu rozptylu.

Obrovskou výhodou systematického výběru je jeho realizovatelnost pouze se znalostí velikosti populace N. Není potřeba opora výběru[14].

Shrnutí a aplikace

Požadavky na skupiny u skupinového výběru jsou opačné od požadavků na strata u výběru stratifikovaného. Snahou je, aby jednotlivé skupiny byly zmenšenými obrazy celé populace a aby se mezi sebou zásadně nelišily. Jako dvě hlavní silné stránky vícestupňového skupinového výběru lze uvést:

- Oproti PNV a stratifikovanému výběru k jeho uskutečnění není potřeba opora základního souboru. Stačí, když máme k dispozici seznam skupin populace. Kompletní seznam jednotek následně získame jen pro vybrané skupiny.
- Výrazně šetří finanční a časové náklady v případech, kdy jsou jednotlivé prvky populace rozmístěny v prostoru daleko od sebe. V tom případě vhodná volba skupin (sdružující jednotky sobě blízké) zaručí menší nároky na realizovatelnost.

Problémem je fakt, že rozptyl odhadů je často větší než u PNV a stratifikovaného výběru.

3.5 Další vybrané partie

Výběry s nestejnou pravděpodobností zahrnutí

V předchozí kapitolách jsme se setkali s výběrovými plány, které různým jednotkám základního souboru přiřadily ne nutně stejné pravděpodobnosti zahrnutí π_i. Byl to stratifikovaný výběr při neproporcionální alokaci nebo vícestupňový výběr. Zapříčínilo to

[14]Pro ilustraci lze uvést získávání odpovědí u vstupu do budovy. Zde jsou vybírány jednotky na základě jejich pořadí. Nutným předpokladem je, že celý základní soubor bude tímto vchodem vcházet.

použití PNV v různých skupinách a stratech. V těchto případech jsme nejdřív nadefinovali plán a potom odvodili pravděpodobnosti zahrnutí. V praktických i teoretických úlohách se ovšem často vyžaduje opačný postup. Máme předem zadané pravděpodobnosti zahrnutí a snažíme se najít výběrový plán, který tyto pravděpodobnosti zahrnutí zaručí. Tato úloha není jednoznačná. Jedním z hlavních důvodů, proč takovou úlohu vůbec řešit, je snaha snižovat rozptyly odhadů.

Vícestupňový skupinový výběr představuje ideální prostor pro aplikaci těchto metod. Například u prostého skupinového výběru může docházet ke značnému kolísání hodnoty odhadu z důvodu různých velikostí těchto skupin. Řešením je místo PNV použít jiný výběrový plán, který by místo rovnoměrné pravděpodobnosti zahrnutí pro všechny skupiny použil pravděpodobnosti úměrné velikosti skupin. Použitím H-T odhadu lze následně získat nestranný odhad populačnách parametrů. Rozptyl tohoto odhadu by byl značně nižší než kdybychom použili standardní odhad při prostém skupinovém výběru. Rozptyly odhadů se však stávájí komplikovanějšími a zde se jimi nebudeme zabývat.

Výběrových plánů, které dokáží zaručit předem zadané pravděpodobnostmi zahrnutí, existuje několik. Hodně z nich se snaží simulovat výběr vzorku o velikosti n jako vybírání po jedné jednotce. V každém tahu mají jednotky jistou pravděpodobnost vytažení α_i. Součet pravděpodobnosti vytažení přes všechny jednotky populace je 1. Díky tomu lze jednotky v každém tahu volit například na základě generování náhodného čísla z intervalu (0,1). Různé plány postupují jinak při opakovaném vybrání stejné jednotky. Nutno dodat, že právě výběr bez vracení s pevně zadanými pravděpodobnostmi zahrnutí upoutal pozornost teoretiků a obecně se nejedná o snadnou záležitost. V dnešní době existuje

mnoho způsobů, jak jej uskutečnit. Hlavními kritérii jsou v tomto ohledu jednoduchý vztah mezi pravděpodobnostmi zahrnutí a vytažení, časová nenáročnost a možnost spočítat/odhadovat rozptyl odhadů (znalost π_{ij}).

Neodpověď

Po výběru vzorku nastává moment, kdy je nutno získat od každé jednotky požadované informace. Právě tento krok je ovšem jednou z nejnáročnějších částí výběrového šetření. Pokud se nám nepovede získat potřebnou informaci, mluvíme o *neodpovědi*[15]. Běžný význam tohoto termínu nabádá, že jednotkami jsou lidé, obecně tomu však být nemusí.

Zkoumejme nyní, jak se neodpověď promítne do přesnosti odhadů. Postup ilustrujeme na příkladu průměru, pro ostatní charakteristiky lze postupovat analogicky. Mějme pro jednoduchost PNV a nechť $\overline{y}$ je výběrový odhad průměru (víme už, že je nevychýlený). Dále označme N_1 počet jednotek populace, které odpoví. Takové jednotky nazveme respondenty. Počet jednotek populace, které neodpoví, označme N_2. Zřejmě platí, že $N_1 + N_2 = N$. Dále označme $\overline{Y_1} = \sum_{i=1}^{N_1} \frac{y_i}{N_1}$ populační průměr respondentů. A $\overline{Y_2} = \sum_{i=1}^{N_2} \frac{y_i}{N_2}$ populační průměr nerespondentů. Pokud se při neodpovědi postupuje tak, že nehledáme náhradní jednotky a průměr počítáme jen přes respondendy ve vzorku, lze ukázat [17, s. 65] platnost:

$$\mathbb{E}(\overline{y}) = \overline{Y_1}.$$

Výše uvedeným postupem odhadu tedy dostaneme hodnoty, jenž se budou v průměru rovnat populačnímu průměru respondentů.

[15] Někdy se také používá pojem *neúčast*. Procentuální zastoupení jednotek ve vzorku, od kterých se podařilo získat požadovanou informaci nazýváme *návratnost*.

Zkoumejme tedy vychýlení, které vznikne v kontextu PNV při neodpovědi:

$$\mathbb{E}(\overline{y}) - \overline{Y} = \overline{Y_1} - \frac{\sum_{i=1}^{N} y_i}{N} = \overline{Y_1} - \frac{N_1\overline{Y_1} + N_2\overline{Y_2}}{N} = \frac{N_2}{N}(\overline{Y_1} - \overline{Y_2}).$$

Velikost vychýlení tedy závísí na dvou faktorech:

- $\frac{N_2}{N}$: Toto číslo představuje poměr nerespondentů v celé populaci. Platí, že při čím výšší je tento poměr, tím větší je vychýlení (za předpokladu, že druhý faktor není nulový).
- $\overline{Y_1} - \overline{Y_2}$: Jak je snadno vidět, tento výraz značí rozdíl populačních průměrů respondentů a nerespondentů. Čím větší je tento výraz, tím větší je vychýlení odhadu průměru při PNV (za předpokladu, že první faktor není nulový).

Důležitým poznatkem je také fakt, že vzorec pro vychýlení nikde neobsahuje počet respondentů v konkrétním vzorku.

U jiných výběrových plánů se tato analýza stavá komplikovanější a proto ji zde nebudeme uvádět. Obecně je ale dopad neodpovědi analogický. Neodpověď se týká jak pravděpodobnostních, tak nepravděpodobnostních výběrů. U nepravděpodobnostních výběrů ovšem vychýlení odhadu nelze kvantifikovat. A jakákoliv diskuze týkající se přesnosti odhadů se potom zakládá na intuici.

Lohr [18, s. 333] uvádí základní faktory ovlivňující míru neodpovědi u dotazníku:

- **Obsah**

  Míru odpovědi v případě citlivých témat (finanční situace, sexualní orientace atd.) lze zvýšit zaručením anonymity, znáhodněným dotazováním nebo vhodným uspořádáním otázek.

- **Načasování**

  Jiná období dotazování mohou vést k větší účasti. Je nutné brát v potaz různé svátky a prázdniny.

- **Tazatelé/sběratelé informací**

  Zvýšení kvality tazatelů lze docílit pomocí školení.

- **Způsob sběru dat**

  Je třeba brát zřetel na to, jakým kanálem se s respondentem komunikuje. Obecně platí, že internetové (včetně emailových) dotazníky mají vysokou neodpověď. Zároveň ale představují jeden z nejlevnějších způsobů sběru dat.

- **Forma dotazníku**

  Jednoznačnost a srozumitelnost otázek společně s estetickou stránkou dotazníku hrají významnou roli.

- **Zátěž na respondenta**

  Respondent odpovídáním ztrácí svůj čas. Proto je žádoucí pokládat co nejmenší počet otázek.

- **Uvedení**

  První dojem je často to nejdůležitější kritérium pro respondenta. V úvodu je nutné uvést motivace a cíle dotazníků. V jistých případech hraje důležitou roli zaručení anonymity.

- **Podněty**

  Zde je možné uvažovat dva druhy. Pozitivní podněty mohou například zahrnovat finanční či věcnou odměnu. Kdežto negativní podněty motivují respondenty odpovídat skrze potenciální pokuty nebo nevýhody.

- **Práce s nerespondenty**

  Neúspěch získat odpověď při prvním pokusu nemusí nutně

znamenat, že respondent odmítá spolupracovat. Opětovné žádosti a připomínání mohou vést k získání požadovaných informací.

4 Empirická část

4.1 Metodologie

Pro empirickou část byly zvoleny vybrané práce absolventů IES. Jako platforma pro vyhledávání sloužil Repozitář závěrečných prací UK a webové stránky IES FSV UK. Z prací využívajících metodu výběrového šetření byly vybírány jen deskriptivní. Ty mají za cíl v nějaké formě[16] zjišťovat hodnotu populačních parametrů (průměr, úhrn, relativní četnost atd.) pomocí výběrových odhadů. Stejně jako v teoretické části budeme uvažovat design-based přístup. To odpovídá situaci, kdy hodnoty sledované charakteristiky jsou pro všechny jednotky populace považovány za konstanty.

Předem je nutno zdůraznit, že cílem nebylo vyhledat všechny práce využívající metodu výběrového šetření. Práce byly voleny na základě autorova úsudku a možnosti ilustrace základních metod teorie výběrů z konečných populací. Celkově byly zvoleny tři bakalářské ([2], [20] a [25]) a jedna diplomová práce ([16]). U každé bude uveden obsahový souhrn a popis využitých metod. Následně se autor pokusí určit nejvhodnější pravděpodobnostní a nepravděpodobnostní výběr v daném kontextu. Snahou bude aplikovat teoretické poznatky z předchozích kapitol na konkrétní výběrová šetření.

[16] Některé práce sesbíraná data použily k ekonometrické analýze. Odhad populačních parametrů nebyl hlavním cílem. Pro naše účely však tato výběrová šetření představují prostor pro ilustraci výběrových metod z teoretické části. Proto budou techniky výběru použité autory prací analyzovány výhradně z pohledu odhadu populačních parametrů. Použitím získaných dat v ekonometrických modelech a splněním jejich předpokladů se nebudeme zabývat.

4.2 Financování terciárního vzdělávání v České republice ve srovnání se systémem fungujícím ve Švédsku

Popis práce a metod

Autorka práce [2] analyzuje stav financování terciárního systému vzdělávání v České republice a diskutuje možnost reforem. Pro účely vlastního výzkumu je vytvořen dotazník s názvem „Postoje studentů vysokých škol 2010“. Autorka nepopsanou metodou vybrala vzorek 851 studentů soukromých a veřejných vysokých škol. Cílem bylo mimo jiné[17] určit relativní četnosti studentů s jistou charakteristikou (ochota vzít si půjčku, názor na spravedlivost přijímacího řízení atd.) a dále určení průměrných hodnot (velikost kapesného/vlastních příjmů, výše školného atd.).

Analýza a návrh výběrové strategie

Autorka sama upozorňuje, že data mohou být zkreslená z důvodu malého počtu respondentů. Dle údajů ČSÚ [4] studovalo v akademickém roce 2009/2010 vysokou školu 389 044 studentů (včetně doktorského studia). Skutečnost, že základní soubor je v porovnání se vzorkem značně větší, nečiní zásadní problém. Největším nedostatkem při výběrovém šetření bylo to, že autorka neuvedla způsob vybrání vzorku. Dá se tedy předpokládat, že použitým typem výběru byla anketa nebo pohodlnostní výběr. Vypovídací hodnota provedeného dotazníku je tedy neurčitelná a jeho závěry lze jen stěží vztahovat na celou populaci.

Autorka si zadala velmi složitý úkol se svým dotazníkem. Pravděpodobnostní výběr by šel užít teoreticky v několika formách. Nepřístupnost opory výběru (seznamu všech vysokoškolských stu-

[17] Mnoho otázek se snažilo zjistit, jak jsou studenti rozdělení do více skupin (např. „Jaké máte vlastní zdroje příjmů?“). Vzhledem k tomu, že této problematice jsme se v teoretické části nevěnovali, nejsou tyto typy odhadů naším hlavním zájmem.

dentů) by nebyl překážkou, jelikož užití PNV by bylo v tomto kontextu nerealizovatelné. To samé lze říci o výběru stratifikovaném. Šlo by ovšem postupovat s využitím seznamu všech veřejných a soukromých vysokých škol a metody vícestupňového skupinového výběru. Jednotlivé školy by představovaly primární výběrové jednotky, jejich fakulty sekundární výběrové jednotky. Terciární výběrové jednotky by mohly být například studenti stejných ročníků. Popřípadě by bylo nutné jít i o stupeň dál. Pro výběr škol (v prvním stupni) by byl použit výběrový plán s nestejnými pravděpodobnostmi zahrnutí. Tyto pravděpodobnosti by byly odvozeny od počtu studentů. Jinak by totiž školy s velkým počtem studentů a ty s malým počtem studentů měly stejnou šanci být vybrány. Což by pravděpodobně značně zvýšilo rozptyl odhadů. Analogicky by se volil výběrový plán ve stupni druhém (mezi fakultami) a třetím (mezi ročníky). Jak ovšem vidno, realizace takového šetření zdaleka přesahuje časové, finanční a plánovací možnosti jednoho člověka. Odhad rozptylu odhadů a tvorba intervalů spolehlivosti by také nebylo snadné.

Nepravděpodobnostní výběr by představoval méně finančně a časově náročnou alternativu. Nabízí se kvótní výběr. Jednotlivé kvóty by se mohly týkat zastoupení mužů a žen, vysokých škol, typů studijních programů atd. Rozložení studentů podle vysokých škol a typu studijního programu je uvedeno v [5] a [6].

4.3 Vliv sportu na kouření a spotřebu alkoholu

Popis práce a metod

Bakalářská práce [20] zkoumá význam sportu pro moderního člověka a společensko-ekonomické dopady sportovní neaktivity. Dále analyzuje nákladovou efektivnost různých investic zdravotního systému. Autor využívá vlastního sběru dat pro účely následné

ekonometrické analýzy. Respondenty byli studenti pardubických středních škol. Jako formu šetření zvolil dotazník a nemalý prostor věnuje teoretickým poznatkům o dotaznících. Jako výběrový plán byl podle autora použit stratifikovaný výběr. Následně podává popisné statistiky o napozorovnaných proměnných. Jde o průměry (počet vykouřených cigaret denně, počet hodin sportu týdně atd.) a relativní četnosti (zastoupení kuřáků, zastoupení studentů dostávajících kapesné atd.). Dále s daty provádí ekonometrickou analýzu a snaží se zjistit, jaký vliv má sport na kouření a konzumaci alkoholu.

Analýza a návrh výběrové strategie

Z textu není jasné, jaký byl základní soubor/cílová skupina výzkumu. Výběrové šetření autor prováděl mezi pardubickými středoškoláky. Kdežto příklady jiných studií zahrnovaly různé věkové kategorie. Na dalších místech textu ovšem autor mluví o výzkumu (všech) středoškoláků a v závěru práce uvádí:

> „Hlavní přínos práce vidíme v tom, že se nám podařilo zmapovat situaci v ČR a získat dobrý důkaz, že existuje negativní vztah mezi sportem a kouřením.“ [20, s. 38]

Pokud jsou cílovou skupinou čeští středoškoláci, provedené výběrové šetření lze potom považovat za pohodlnostní výběr (jeden z typů nepravděpodobnostních výběrů). Autor byl mylně přesvědčen, že vykonává stratifikovaný výběr. Správná aplikace jakéhokoliv pravděpodobnostního výběru by byla pro jednoho člověka časově a finančně prakticky nemožná. Dalo by se postupovat zcela analogicky jako u navrženého postupu v práci v podkapitole 4.2. Vícestupňový skupinový výběr by byl opět vhodnou volbou. Jen zaměníme vysokoškoláky za středoškoláky.

Pokud data a analýza měly vypovídat jen o pardubických středoškolácích, situace se částečně mění. Pokusme se nastínit, jak by bylo možné stratifikovaný výběr vykonat. Konkrétně použijeme proporcionální alokaci (podkapitola 3.3). Předpokládájme znalost velikosti základního souboru a možnost získat oporu výběru. Rozdělme základní soubor na tři strata – studenti pardubických gymnázií, studenti pardubických odborných škol a studenti pardubických učilišť. Nyní můžeme využít uvedených údajů pro zastoupení jednotlivých pardubických středoškolských studentů v poměru 25% gymnázia, 25% střední odborná učiliště a 50% střední odborné školy. Určení požadované velikosti celkového výběrového souboru lze docílit pomocí modifikace postupu z podkapitoly 3.2. Tento krok má ovšem dvě úskalí. Dotazník zjišťuje hned několik populačních parametrů najednou, vzorec pro výpočet velikosti vzorku by se správně tedy musel aplikovat pro každou charakteristiku zvlášť. Navíc populační rozptyly každé z charakteristik v jednotlivých stratách jsou neznámé, k jejich zjištění by například pomohla pilotáž. Následující krok obnáší vypočítání velikosti vzorků v jednotlivých stratách pomocí vzorce (3) a vykonání PNV v každé stratě.

Výše uvedený postup opět značně přesahuje možnosti jednoho člověka. Méně časově náročné pravděpodobnostní výběry by vyžadovaly vícestupňový přístup, ale to celou teorii odhadu komplikuje. Jako vhodnou nepravděpodobnostní alternativu lze užít kvótní výběr, kde hlavním znakem by byl typ střední školy. Výběr, kterého užil autor, se dá považovat za pohodlnostní. Nepopsanou metodou zvolil jedno gymnázium, jedno odborné učiliště a dvě odborné školy. Následně v jejich rámci vybral (opět nepopsanou metodou) třídy. Finálním krokem byl census přes všechny žáky ve třídách. Získané odhady z takto provedeného šetření by mohly

být systematicky vychýlené. Důvodem je například fakt, že žáci stejných tříd spolu sdílí mnohé aktivity – ať už sportovní nebo i ty nežádoucí.

4.4 Factors that influence the success of small and medium enterprises

Popis práce a metod

Hlavním cílem práce [16] je najít faktory, které ovlivňují úspěch malých a středních podniků v českém ICT sektoru (sektor informačních a komunikačních technologií). Práce nejdříve seznamuje čtenáře s ICT trhem v České republice. Následně se zaobírá různými přístupy hodnocení úspěšnosti ICT firem a klade si otázku, zda-li existují indikátory měřící úspěch. Z důvodu nedostupnosti relevantních informací si autor pro empirický výzkum zvolil dotazníkové šetření skrze email. Seznam ICT firem a jejich kontaktní údaje se autorovi podařilo získat. Respektive jej vytvořil na základě uvedeného emailu. Celkově tak získal populaci o velikosti 7 979. Všechny tyto firmy oslovil emailem. Konečný výběrový soubor obsahoval 105 firem. Firmám bylo položeno 30 otázek, z nichž několik bylo kvantitativního charakteru – počet zaměstnanců, tržby atd. Následuje deskriptivní popis sesbíraných dat a ekonometrická analýza.

Analýza a návrh výběrové strategie

Prvním problémem může být undercoverage. Pravděpodobně existují malé a střední firmy, které autor neoslovil. Důvodem je neexistence kontaktní informace. Ignorujeme-li tuto možnost, autor se pokusil o census. Vzhledem k formě výběrového šetření – internetový dotazník rozeslaný skrze email, se dalo očekávat hodně neodpovědí. To se následně prokázalo. A tím pádem se jednalo

o nepravděpodobnostní formu výběrového šetření (v podstatě jde o anketu). Opět zde docházelo k samovýběru respondentů a nelze vyloučit jejich systematickou odlišnost od nerespondentů. Autor práci s malým vzorkem ospravedlnil odkazem na článek [1]. Tento článek shrnuje problematiku velikosti výběrového souboru při požadovaném stupni spolehlivosti. A dále pojednává o neodpovědi. Obě témata jsou ovšem probírána v kontextu PNV a systematického výběru. Dotazník byl pouze pokus o census s obrovským počtem neodpovědí, tento článek a metody v něm popsané nejsou tedy aplikovatelné na daný problém.

V tomto případě by šlo využít jednoho ze základních pravděpodobnostních výběrů. Nabízí se PNV nebo systematický výběr. Obrovskou výhodou je dostupnost opory výběru. Jednotlivé firmy ve vzorku mohou být rozprostřeny „chaoticky“ po celé České republice. To ovšem nečiní žádný problém, jelikož se dá předpokládat, že s ohledem na jejich zaměření preferují elektronický způsob komunikace. Při PNV by se podle požadované spolehlivosti a tolerované chyby zvolila velikost výběrového souboru. Nutno dodat, že dotazník zjišťuje několik populačních charakteristik. Tím pádem by měla být správně pro každou z nich určena požadováná velikost souboru zvlášť a následně zvolena ta nejvyšší. Klasickým problémem zůstává odhadování populačního rozptylu S^2 ve vzorci (2). Prostým náhodným výběrem se ovšem nevyhneme neodpovědím. Konkrétní kroky a návrhy pro prevenci a boj proti neodpovědi jsou uvedeny v podkapitole 3.5.

4.5 An Analysis of Households Expenditure of Vietnamese Community living in the Czech Republic

Popis práce a metod

Práce se snaží analyzovat výdaje vietnamských rodin v České republice. Zkoumá rozložení nakupovaných statků a služeb a následně zjištění porovnává s průměrnou českou domácností. Z důvodu neexistence potřebných dat autorka pro jejich získání používá výběrové šetření. Část respondentů je osloveno osobně, zbytek elektronicky přes sociální sítě pomocí dotazníku. Výsledný výběrový soubor čítá 151 vietnamských domácností. Kladené otázky se mimo jiné týkají kvantitativních charakteristik jako počet členů domácnosti, délka pobytu v ČR a hlavně průměrné měsíční výdaje na konkrétní zboží a služby. Průměrné hodnoty a poměrná zastoupení jsou nejčastější ukazetele, které tvoří deskriptivní část údajů. Data jsou následně použita k porovnání českých a vietnamských domácností. Pro české domácnosti byly statistiky získány z ČSÚ. Finální částí práce je ekonometrický model využívající sesbíraná data.

Analýza a návrh výběrové strategie

Autorka v práci uvádí, že dle ČSÚ žilo v roce 2012 v ČR 58 205 Vietnamců. Použijme pro jednoduchost odhad, že průměrná domácnost má 4 členy. Následně tedy dostáváme základní soubor o velikost 14 551 domácností. Pro získání výběrového souboru bylo použito nepravděpodobnostního výběru. Podle popisu postupu sbírání se tento výběr dá považovat za kvótní nebo účelový. Fakt, že jsou do vzorku zahrnuty domácnosti z různých částí ČR může přispět k reprezentativitě. Rozdělení domácností ve vzorku dle místa bydliště však autorka neuvádí. Není tedy jasné, jestli byly

kvóty naplněny. Častým problémem byla neodpověď nebo podhodnocování.

Nejlepší volbou z kategorie nepravděpodobnostních výběrů by byl výběr kvótní. Jako hlavní znaky lze použít místo pobytu, vzdělání a jiné. Rozdělení vietnamských rodin podle místa pobytu je dostupnou informací a autorka jej dokonce v práci uvádí. Díky tomu lze snadno stanovit kvóty pro výběrový soubor. Osobní sběr dat by tedy byl finačně (náklady na cestování) a časově náročný. Internetový sběr tyto náklady nemá, předpokládá ovšem dostupnost kontaktních údajů nebo věrohodnost při použití sociálních sítí.

Co se týká výběru pravděpodobnostního, základní problém představuje nedostupnost opory výběru. Nabízí se tedy vícestupňový skupinový výběr. Kde primární jednotky mohou být reprezentovány městy. Následně podle rozložení vietnamských občanů v ČR lze sestrojit výběrový plán, kde jednotlivé pravděpodobnosti zahrnutí budou úměrné počtu vietnamských domácností ve městě. Sekundární jednotky mohou být městské čtvrti. Jak je ovšem zřejmé, výběrové šetření tohoto rozsahu má několik nároků. Hlavním je získání opory výběru v příslušných oblastech. Jeho uskutečnění by nebylo po technické a výpočetní stránce snadné. Pro přesné odhady by musely být použity velké výběrové soubory, což by bylo časově velmi náročné. A těžko by šlo realizovat pouze skrze internetové dotazníky.

5 Závěr

Prvním cílem této práce bylo vyložit základní poznatky z teorie výběru z konečných populací. Volba témat byla odvozena od možnosti použití v části empirické. Popsané pravděpodobnostní a nepravděpodobnostní výběry jsou ovšem aplikovatelné i v mnoha jiných úlohách. Volba metody výběru předchází sběru dat a díky její znalosti můžeme určit některé chyby, které při odhadech vznikají. I přesto, že pravděpodobnostní výběr je některými autory považován za jedinou vědeckou metodu výběrového šetření, často je možné (nutné) se bez něj obejít. Kvótní výběr představuje nepravděpodobnostní alternativu, která v jistém smyslu zaručuje reprezentativní vzorek. Vzhledem k mnohem méně náročnější realizaci umožnuje získat potřebné odhady rychleji a pohodlněji. Rozhodně by se jeho použití nemělo odsuzovat. Zásadním nedostatkem, kterého by se výzkumník ovšem neměl dopouštět, je neuvedení metody výběru a neupozornění na možné chyby.

Analýza konkrétních akademických prací studentů IES představovala druhý cíl. Každá z nich použila výběrové šetření k získání dat. Odhad populačních parametrů nebyl u všech prací hlavním záměrem, jejich povaha ovšem umožnila ilustraci konkrétních metod. U každé z prací byl uveden obsahový souhrn, popsány použité techniky a následně byl uveden návodný postup. V kontextu uvedených prací se pravděpodobnostní výběry většinou jevily časově, technicky a finančně náročné. Získání opory výběru, prostorové vzdálenosti jednotek základní populace, sběr dat a potýkání se s neodpovědí představují hlavní úskalí. Další překážkou při pravděpodobnostních výběrech je neznalost populačního rozptylu a jeho různých obměn. Lze jej odhadovat ze vzorku, celý proces to ovšem komplikuje. Komplikace také přichází v podobě

počtu odhadovaných parametrů. Dotazníky totiž jen zřídkakdy pokládají jen jednu otázku. Při sestrojování požadované velikosti vzorku a intervalových odhadů je však nutno postupovat parametr po parametru.

Na úplný závěr je třeba připomenout, že analyzované práce byly zvoleny pouze na základě libovůle autora (účelový výběr). Nelze dělat závěry o celkovém povědomí a kvalitě výběrových šetření prováděných v pracech studentů IES. To si taky práce jako svůj přínos neklade. Hlavní snahou bylo ilustrovat na konkrétních úlohách pravděpodobnostní a nepravděpodobnostní metody výběru. Případně pomoci budoucím studentům bakalářských a magisterských programů porozumět základním technikám teorie výběru z konečných populací.

Použitá literatura

[1] BARTLETT, J.E., J.W. KOTRLIK a C.C. HIGGINS. Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*. 2001, roč. 19, č. 1, s. 43-50.

[2] BUREŠOVÁ, Nikola. *Financování terciárního vzdělávání v České republice ve srovnání se systémem fungujícím ve Švédsku*. Praha, 2010. 84 s. Bakalářská práce. Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií.

[3] COCHRAN, William Gemmell. *Sampling techniques*. 3rd ed. New York: Wiley, 1977, xvi, 428 s. ISBN 04-711-6240-X.

[4] ČESKÝ STATISTICKÝ ÚŘAD. *Vysoké školy v České republice* [online]. 2013 [cit. 2014-04-17]. Dostupné z: `http://www.czso.cz/cz/cr_1989_ts/1208.pdf`

[5] ČESKÝ STATISTICKÝ ÚŘAD. *Vysoké školy soukromé - studenti, absolventi; rok 2012* [online]. 2012 [cit. 2014-04-17]. Dostupné z: `http://www.czso.cz/csu/2013edicniplan.nsf/t/2E0028BC8A/$File/33011332.pdf`

[6] ČESKÝ STATISTICKÝ ÚŘAD. *Vysoké školy veřejné - základní přehled za rok 2012* [online]. 2012 [cit. 2014-04-17]. Dostupné z: `http://www.czso.cz/csu/2013edicniplan.nsf/t/2E0028BC8C/$File/33011331.pdf`

[7] DISMAN, Miroslav. *Jak se vyrábí sociologická znalost: Příručka pro uživatele*. 3.vyd. Praha: Karolinum, 2000, 374 s. ISBN 978-80-246-0139-7.

[8] FIDELITY INVESTMENTS. *Mutual funds research* [online]. 2014 [cit. 2014-02-26]. Dostupné z: https://www.fidelity.com/fund-screener/research.shtml

[9] GALLUP, Inc. *Election Polls – Accuracy Record in Presidential Elections* [online]. 2013 [cit. 2014-04-12]. Dostupné z: http://www.gallup.com/poll/9442/election-polls-accuracy-record-presidential-elections.aspx

[10] HAIR, Joseph F. *Essentials of marketing research*. 2nd ed. New York, NY: McGraw-Hill Irwin, 2010, xvii, 398 s. ISBN 00-734-0482-9.

[11] HÁJEK, Jaroslav a Václav DUPAČ. *Sampling from a finite population*. New York: M. Dekker, 1981, v, 247 s. ISBN 08-247-1291-9.

[12] CHYTILOVÁ, Julie a Michal BAUER. *IES GRADUATES' EARNINGS* [online]. 2009 [cit. 2014-04-15]. Dostupné z: http://ies.fsv.cuni.cz/default/file/download/id/12408

[13] JEŘÁBEK, Hynek. *Úvod do sociologického výzkumu*. Dot. Praha: Karolinum, 1993, 162 s. ISBN 80-706-6662-5.

[14] KISH, Leslie. *Survey sampling*. Classics ed. New York [u.a.]: Wiley, 1965. ISBN 04-714-8900-X.

[15] KNETSCH, Jack L. Preferences and nonreversibility of indifference curves. *Journal of Economic Behavior*. 1992, vol. 17, issue 1, s. 131-139. DOI: 10.1016/0167-2681(92)90082-M. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/016726819290082M

[16] KREJČÍ, Martin. *Factors that influence the success of small and medium enterprises*. Praha, 2013. 89 s. Diplomová práce. Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií.

[17] LEVY, Paul S a Stanley LEMESHOW. *Sampling of populations: methods and applications*. 3rd ed. New York: Wiley, 1999, xxxi, 525 s. ISBN 04-711-5575-6.

[18] LOHR, Sharon L. *Sampling: design and analysis*. 2nd ed. Boston, Mass.: Brooks/Cole, 2010, xi, 596 s. ISBN 04-951-1084-1.

[19] NEYMAN, Jerzy. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*. 1934, vol. 97, issue 4, s. 558-625. DOI: 10.2307/2342192. Dostupné z: `http://www.jstor.org/stable/10.2307/2342192?origin=crossref`

[20] MICHLIAN, Štefan. *Vliv sportu na kouření a spotřebu alkoholu*. Praha, 2012. 53 s. Bakalářská práce. Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií.

[21] SÄRNDAL, Carl-Erik. Design-Based and Model-Based Inference in Survey Sampling [with Discussion and Reply]. *Scandinavian journal of statistics* [online]. 1978, vol. 5, issue 1, s. 27-52 [cit. 2014-04-13]. Dostupné z: `http://www.jstor.org/stable/4615682`

[22] THOMPSON, Steven K. *Sampling*. 3rd ed. Hoboken, N.J.: Wiley, 2012, xxi, 436 s. Wiley series in probability and statistics. ISBN 04-704-0231-8.

[23] VÍŠEK, Jan Ámos. *Selected statistical methods*. 1st ed. Praha: Karolinum, 2003, 171 s. Učební texty (Univerzita Karlova). ISBN 80-246-0726-3.

[24] VORLÍČKOVÁ, Dana. *Výběry z konečných souborů*. Praha: Univerzita Karlova, 1985.

[25] VU, Thi My. *An Analysis of Households Expenditure of Vietnamese Community living in the Czech Republic*. Praha, 2013. 62 s. Bakalářská práce. Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií.

[26] WINKLER, Jiří a Miloslav PETRUSEK. *Velký sociologický slovník*. Praha: Karolinum Praha, 1997. 598 s. Academia. ISBN 80-7184-164-1.

Přílohy

Tabulka 1: Tříroční výnosové míry podílových fondů společnosti Fidelity [8]

44.32	16.61	15.3	14.3	12.8	10.38	8.74	7.88	7.1	6.9	4.87	3.97	2.84	.08	.01	.01
28.37	16.56	15.3	13.94	12.25	10.33	8.67	7.87	6.96	6.6	4.79	3.86	2.76	.05	.01	.01
23.58	16.48	15.1	13.88	12.21	10.26	8.51	7.83	6.96	5.86	4.66	3.76	2.54	.03	.01	.01
22.77	16.47	14.94	13.88	11.91	10.24	8.49	7.72	6.88	5.8	4.64	3.64	2.49	.03	.01	.01
21.49	16.46	14.91	13.83	11.8	10.21	8.47	7.53	6.81	5.76	4.58	3.63	2.35	.02	.01	-.03
21.16	16.41	14.82	13.75	11.5	10.16	8.29	7.53	6.79	5.7	4.58	3.61	1.56	.02	.01	-3.35
20.64	16.38	14.78	13.58	11.47	10.8	8.28	7.46	6.78	5.65	4.52	3.54	1.32	.02	.01	-7.2
18.93	16.35	14.76	13.54	11.43	10.7	8.17	7.45	6.75	5.46	4.49	3.54	1.22	.02	.01	-11.79
18.86	16.17	14.64	13.43	11.22	9.37	8.14	7.42	6.61	5.4	4.4	3.53	1.21	.02	.01	-23.43
18.2	15.84	14.58	13.3	11.13	9.34	8.7	7.27	6.59	5.3	4.39	3.49	.89	.02	.01	
17.87	15.62	14.53	13.28	10.98	9.24	8.2	7.26	6.53	5.24	4.39	3.44	.34	.02	.01	
17.5	15.45	14.5	13.26	10.91	9.24	8	7.25	6.44	5.15	4.39	3.2	.18	.02	.01	
17.5	15.23	14.12	13.2	10.88	9.19	7.99	7.24	6.36	5.3	4.31	3.13	.14	.01	.01	
16.88	15.9	14.1	13.7	10.48	9.13	7.99	7.21	6.3	4.99	4.29	3.1	.14	.01	.01	
16.75	15.5	14.8	12.86	10.47	8.78	7.97	7.1	6.23	4.91	3.97	2.85	.13	.01	.01	

Obrázek 1: Rozdělení četností tříroční výnosové míry podílových fondů