Kapitola 5.

Kolik vran musíme pozorovat?

Když nevíš co děláš, zeptej se někoho, kdo to ví.

Jerry Pournelle, téměř v každém čísle magazínu BYTE

Tohle je opět kapitola o redukci informací. Je to kapitola přece jen radostnější než ty předchozí. Redukce populace na vzorek má dobře propracovanou teorii i dobře vypracované a spolehlivé recepty. Některé operace tu nejsou snadné, ale je mnoho lidí, kteří je znají a mohou nám poradit. Buďte tedy zadobře se statistiky.

Touto kapitolou vstupujeme do spíše technické oblasti výzkumu. K tomu nám může hodit dobrý pomocník. Dovolte, abych vám představil Dr. Watsona.

Dr. Watson je svým způsobem chytrý muž na systemizovaném místě pitomce. Je to někdo, koho každý profesor touží mít ve třídě. Doktor Watson vždycky navrhne nějakou, zdánlivě zřejmou, ale ve skutečnosti pitomoučkou odpověď, čímž umožní profesorovi nabídnout správnou odpověď, a tak se zaskvíti svojí moudrostí a učeností. Budeme služeb Dr. Watsona hodně používat.

5.1. Vzorek z nouze

Začneme spíše stupidní otázkou: "Kolik vran musíme pozorovat, abychom mohli říci, že všechny vrány jsou černé?" Odpověď je tak jednoduchá, že po ní nemusíme pátrat na konci kapitoly a přirozeně zní "Všechny!" Na druhé straně asi nikdo nikdy nepozoroval všechny vrány. Nezbývá nám nic jiného, než se spokojit s tvrzením, že "většina vran je černých". Opět je to něco, co už známe: redukovaná analýza reality vede k tvrzením pravděpodobnostního charakteru.

Skupiny, o které se v sociologickém výzkumu zajímáme, nejsou malé. V kvantitativní verzi výzkumu jsme schopni zkoumat celou skupinu jenom výjimečně. Pravidelně jedině sčítání lidu je studií celé populace. Většinou studujeme jen některé členy skupiny a doufáme, že naše závěry budou aplikovatelné i na ostatní, na ty nestudované. To nás přivádí k dvěma základním termínům, které potřebujeme pro tuhle kapitolu: populace a vzorek (výběrový soubor). Jejich definice je jednoduchá:

VZOREK:	skupina jednotek, které skutečně pozorujeme
POPULACE	(neboli základní soubor) je soubor jednotek, o kterém předpokládáme, že jsou pro něj naše závěry platné

Náš stěžejní úkol je najít postup, aby výsledky, které získáme na vzorku, byly co nejvíce podobné těm, které bychom získali na celé populaci. První věc, která nám přijde na mysl, je, snažit se mít vzorek co největší. Ale naše následující pravdivá pohádka ním ukáže, že to není jen tak:

Pohádka pro odrostlejší děti 8.

O hodně velkém vzorku, aneb Jak to nevyšlo

Byl jednou v Americe velice rozšířený týdeník, který se jmenoval Literary Digest. Byl u svých čtenářů hodně oblíben. Byl proslulý také tím, že spolehlivě předpovídal výsledky presidentských voleb. Jeho předpovědi byly založeny na obrovském vzorku dvou milionů voličů. (Dnes jsou podobné předpovědi založeny na vzorku tisíckrát menším.) Vzorek byl zkonstruován z mnoha zdrojů. Literary Digest si opatřil adresy voličů z celých USA. Používal pro to zdroje jako telefonní seznamy, městské adresáře, adresy držitelů řidičských průkazů, členské seznamy organizací, seznamy předplatitelů novin a časopisů atd.

Předpovědi byly přesné a úspěšné ve volbách 1920, 1924, 1928, 1932, a pak přišly volby v roce 1936. Literary Digest předpověděl, že presidentský kandidát Landon porazí Roosevelta rozdílem 14%. Přišel volební den a s ním i konec slávy Literary Digestu: Franklin Delano Roosevelt zvítězil drtivou většinou.

Cvičení 4.1.

Reprezentoval vzorek použitý Literary Digestem dobře celou populaci voličů v USA?

To nebylo tak těžké, že? Trochu složitější je otázka, jak je možné, že vzorek, který prakticky vyloučil z výzkumu voliče náležející k nižším sociálním třídám, fungoval dobře v předchozích volbách? Klíčem k řešení je rok: v roce 1935 vrcholila v USA hospodářská krize, a to vedlo k ostré polarizaci podle vertikální stratifikační osy. Předtím sociálně ekonomický status nehrál příliš důležitou roli v otázce volebních preferencí. Daleko větší úlohu hrály takové faktory jako náboženství, zeměpisná poloha atd. Krize to všechno změnila: sociální status začal hrát důležitou funkci. Pravděpodobně nejdůležitější bylo to, že krize přivedla k volebním urnám příslušníky nižších sociálně ekonomických vrstev, kteří předtím příliš často nehlasovali. Můžeme tedy říci, že v letech 1920-1932 předpovědi Literary Digestu vyšly jenom náhodou. Abychom byli schopni z chování vzorku předpovídat chování populace, **musí struktura vzorku imitovat složení populace tak přesně, jak je to jen možné.**

Dr Watson: Ale to je přeci docela lehké! Když je v populaci řekněme 51% žen, tak vyberu také 51% žen do vzorku, a když je v populaci 12% osob nad 65 let věku, vyberu také stejné procento starých osob do vzorku, atd.

Tentokrát má Dr. Watson pravdu. Technika konstrukce vzorku, tak jak ji popsal,se opravdu používá. Říká se tomu **kvótní výběr.**

Kvótní výběr imituje ve struktuře vzorku známé vlastnosti populace.

Bohužel má tato technika některé nepříjemné vlastnosti. Jedna z nich souvisí se slovem "známé" v naší definici. Pro většinu populací není problém zjistit jejich skladbu podle

pohlaví, věku, vzdělání, povolání atd. Lze si snadno představit problém, pro který jsou důležitější jiné vlastnosti, takové, o kterých běžná statistická šetření údaje neshromažďují (kupř. věk, ve kterém se respondent poprvé zamiloval).

Na další problém snadno přijdete sami:

Cvičení 4.2.

Navrhněte prosím, kritéria pro konstrukci kvótního vzorku pro populaci veksláků.

Kvótní výběr může být použit jen na populaci, o které jsme dobře informováni, a to zdaleka není každá populace. Další obtíž je spojena s praktickou stránkou výběru přímo v terénu. Poslední krok obvykle závisí na tazateli, který vybírá jedince podle dané instrukce. Taková instrukce by mohla vypadat třeba takto:

Jméno tazatele: *Dr. Watson*

Respondent č.1.

muž, věk 30-40, dokončené středoškolské vzdělání, povoláním úředník, ženatý, ale bezdětný, bydlící v rodinném domku, žijící v našem městě alespoň 5 let, ale který se narodil v obci pod pětset obyvatel...

Respondent č.2.

žena, věk 60-65, alespoň s dokončeným základním vzděláním, důchodkyně, která pokud byla ještě ekonomicky aktivní, měla dělnické povolání, která žije sama, v bytě alespoň o dvou místnostech a bydlí od narození v našem městě...

Tak, to si od nás Dr. Watson opravdu nezaslouží. Umíte si představit, na kolik dveří by musel zaklepat, než by našel osoby, odpovídající zmíněným charakteristikám. Třeba by je nenašel vůbec, možná, že vůbec neexistují. Ve skutečnosti je instrukce v kvótním výběru mnohem skromnější. Navrhuje jen několik málo proměnných, takových jako pohlaví, věk a povolání. Lokalita a typ obce je obvykle dán působištěm tazatele. Jinak nejsou tyto proměnné vázány do určitých kombinací. Instrukce by mohla znít takto: "Hovořte s deseti osobami, z toho se

šesti ženami a čtyřmi muži. Vyberte 3 osoby ve věku pod 20 let, 5 ve věku 21-50.." O ostatních, pro nás třeba daleko důležitějších proměnných můžeme jenom doufat, že budou ve vzorku dostatečně správně reprezentovány.

Dr. Watson: Co si s tím ale počneme?

Odpověď nám nabízí titul následujícího paragrafu.

5.2. Hodíme si korunou aneb Pravděpodobnost pro Dr. Watsona

Představme si, že máme velikou krabici, plnou kuliček, a že všechny kuličky jsou zelené. Dobře krabicí zatřepeme a poslepu vybereme jednu kuličku. Jakou máme šanci, že vybraná kulička bude zelená? To byla ale pitomá otázka, že ano? Tak si teď zkusme něco trochu složitějšího: Máme teď jinou populaci kuliček, sestávající ze zelených a červených kuliček. Těch zelených je 80% a těch červených je ovšem 20%. Ale počkejte, já se vás nebudu ptát, jaká je pravděpodobnost, že si náhodně vyberete červenou kuličku. To byla otázka jen o málo méně pitomá, než ta první, a všichni víme, že ta pravděpodobnost je 20%, a chceme-li to vyjádřit učeněji, můžeme říci, že p = 0,20.

My tu máme jiný úkol: zjistit, jaká je skladba populace, aniž bychom prohlíželi všechny kuličky. Jinými slovy, hledáme metodu, jak vytvořit vzorek, který by dobře reprezentoval celou populaci kuliček. Můžeme zkusit třeba toto: Opět začneme tím, že krabicí dobře zatřeseme. To není vtip, to je opravdu nutné: **každá kulička musí mít stejnou pravděpodobnost, že bude vybrána.** (Co kdyby všechny červené kuličky byly navrchu?) a teď vybereme poslepu 10 kuliček. Uvidíme třeba, že jsme vybrali 6 červených a 4 zelené. To je dost daleko od dobré reprezentativity. Perfektní vzorek by měl přeci obsahovat 20% červených a 80% zelených. Tedy vybereme opět poslepu dalších deset kuliček. Třeba 6 z nich bude zelených a 4 červené. Přidáme je k našemu původnímu vzorku. Nový, větší vzorek sestává z 10ti červených a 10ti zelených kuliček. Teď bychom odhadli, že v populaci je stejné procento červených, jako zelených kuliček. To ještě není vůbec dobré. Museli bychom tedy pokračovat, přidávat další a další kuličky. Brzy bychom zpozorovali zajímavou věc:

> Srostoucí velikostí vzorku se rozdíl mezi strukturou populace a vzorku zmenšuje.

Nejdříve rychle, pak pomaleji a pomaleji. Úplné shody mezi strukturou populace dosáhneme teprve tehdy, když jsme zahrnuli všechny elementy populace do vzorku.

Dr.Watson: "Ale to je všechno nesmysl! Když je to pravda, jak je potom možné, že obrovský vzorek použitý Literary Digestem vedl k tak nesprávným výsledkům?"

Asi už víte, co bychom mohli odpověď na tuhle námitku: "Ale to je přece elementární, Watsone. Ti lidé z Literary Digestu zapomněli pořádně zatřást krabicí." Voliči z nižších socioekonomických vrstev měli mnohem menší šanci býti vybráni do vzorku, než voliči ze středních a vyšších vrstev, což dramaticky zkreslilo výsledky.

My jsme tu totiž, aniž bychom o tom věděli, vytvořili náhodný vzorek "populace" kuliček. A náhodný vzorek, to je aristokrat mezi vzorky; má mnoho jedinečných, a pro nás důležitých, vlastností. Všechno, co budeme v tomto odstavci probírat, se týká jenom vzorků, které byly vytvořeny opravdu náhodným výběrem. Termín "náhodný" neznamená výběr nazdařbůh. I když náhodný výběr může být, jak brzy uvidíme, technicky velmi obtížný a často i nemožný, jeho definice je jednoduchá:

> Náhodný (pravděpodobnostní) výběr je takový výběr, ve kterém každý element populace má stejnou pravděpodobnost, že bude vybrán do vzorku.

To se lépe řekne než se to udělá. Ale dovolte, abych vás ještě dříve než budeme mluvit o řadě trampot, dobře naladil popisem pozoruhodných vlastností náhodného vzorku. Snad

nejdůležitější z nich, alespoň pro nás sociology - statistik by s námi možná nesouhlasil - je tato vlastnost:

> Náhodný vzorek reprezentuje všechny známé i neznámé vlastnosti populace.

A ještě dříve, než Dr. Watson začne namítat, uveďme si jednoduchý příklad. Máme teď novou populaci kuliček. Jsou opět červené a zelené. Ale mají ještě jednu zajímavou vlastnost, o které my nevíme: Jsou duté a uvnitř každé je malý papírek a na každém tom lístku je něco napsáno. (Znáte "fortune cookies" z čínských restaurací?) Třeba nějaké neslušné slovo. Když jsme vybrali dobrý náhodný vzorek kuliček, budou reprezentovat celou populaci kuliček nejen vzhledem k distribuci barev, ale i vzhledem k distribuci neslušných slov, i když o tom nevíme a třeba nikdy nebudeme vědět. Uveďme si jiný, užitečnější příklad. V náhodném vzorku obyvatelstva hlavního města Prahy budeme mít slušnou reprezentaci populace vzhledem k věku, pohlaví, vzdělání, povolání, politické orientaci, vzhledem ke všem postojům, ale i reprezentaci třeba vzhledem k oblíbeným jídlům, počtu zubních kazů, věku, kdy se lidé poprvé zamilovali, množství vypitého piva, počtu milenek, počtu veksláků, peněžní hodnotě nakradeného zboží, číslům bot, prostě vzhledem ke všemu. To neznamená, že tohle všechno budeme schopni měřit, to je jiný problém. Ale znamená to, že ať už je naším cílem cokoli, víme, že **proměnné, které jsou pro nás relevantní, budou mít v našem vzorku podobnou distribuci, jaká existuje v celé populaci a naše závěry jsou tedy na tuto populaci aplikovatelné.**

Náhodný výběr má ještě jednu pozoruhodnou vlastnost:

> U náhodného vzorku jsme schopni odhadnout, jak se vzorek liší od populace.

Jinými slovy, jsme schopni určit, jak dobrý je náš vzorek. Teď je na čase naučit se několik slov z odborné hantýrky, jednak abychom mohli oslnit přátele, jednak abychom rozuměli správně významu publikovaných statistických dat. Podívejme se na následující tabulku:

Tabulka 5.1.

Velikost vzorku a konfidenční interval

na 95% hladině významnosti pro alternativní znaky při distribuci 50:50

Velikost vzorku	Konfidenční interval
100	± 10%
400	± 5%
1600	± 2.5%

Adaptováno z Babbie: Social Research for Consumer, 1982

To vypadá dost učeně, že? Ale nebojte se. Pochopit princip, a vědět jak se taková věc aplikuje, není těžké. Trochu obtížnější je statistické zdůvodnění. Ale takové vysvětlení necháme pro někoho jiného, kdo vás uvede do zajímavého světa **skutečné** statistiky.

Řekněme, že jsme vybrali náhodně 400 kuliček a zjistili jsme, že ve vzorku (neboli ve výběrovém souboru) je 78% zelených kuliček. Protože jsme nevybrali všechny kuličky, musíme předpokládat, že jsme se dopustili určité chyby, že pozorovaná relativní četnost zelených kuliček ve vzorku se liší od procenta, které skutečně existuje v celé populaci (základním souboru). My však potřebujeme vědět, jak moc se mýlíme. A v tom nám pomůže ta nepřátelsky vyhlížející tabulka. Pozor! **Tahle tabulka je jen ilustrací a platí jen tehdy, je-li v populaci právě tolik zelených jako červených kuliček. Platí jen pro alternativní (binomické) proměnné,** to je pro takové znaky, které mají jen dvě kategorie, jako ANO a NE. V našem případě, zelená a "nezelená" kulička.

Velikost našeho vzorku je 400 a této velikosti vzorku odpovídá konfidenční interval (interval spolehlivosti) 5%. Odečteme tedy tuto hodnotu od pozorovaných 78% a dostaneme tedy 73%. Pak ji opět přičteme k pozorované hodnotě a dostaneme horní mez. a teď víme, že skutečná proporce zelených kuliček v celé populaci je mezi 73 a 83%.Jenomže to nevíme docela určitě, vždyť jsme nepozorovali všechny kuličky. Teď se dostáváme k tomu poněkud kryptickému výrazu v podtitulu naší tabulky: **hladina významnosti.**

V našem případě to znamená, že skutečná proporce, která existuje v populaci, se nalézá s 95% pravděpodobností uvnitř vypočítaného intervalu spolehlivosti. Kdybychom vytvořili 100 vzorků obdobné velikosti, jen v 5 vzorcích by bylo možné, že skutečná proporce zelených kuliček leží pod nebo nad vypočítaným konfidenčním intervalem. O tom, jakou hladinu zvolit, rozhodne výzkumník, a podle tohoto rozhodnutí je interval vypočítáván. Toto rozhodnutí je svobodné ovšem jen z hlediska statistické teorie; ve skutečnosti je vázán míněním, přijatým v příslušné vědecké komunitě. V sociologii je to obvykle 95 nebo 98%. (Vidíte, i v sociologii máme malý kousek paradigmatu.)

A teď se podívejme, jak by se takový interval **mohl** vypočítat. Není to tak, jak se to opravdu dělá. Ve skutečnosti neznáme distribuci proměnné, která existuje v populaci. Ale náš popis výpočtu nám dá alespoň nějaký vhled do logiky, která je skryta za pozoruhodnými vlastnostmi náhodného výběru. Protože jsem vám slíbil, že v naší knize nebudou (skoro) žádné vzorečky, popíšeme si výpočet slovně. Nejdříve musíme vypočítat veličinu, která má opravdu zajímavé vlastnosti a které se říká **směrodatná chyba**. Uvidíte, že je to nejen snadné vypočítat, ale také, že není těžké rozumět většině kroků v tomto výpočtu.

Výpočet směrodatné chyby:

CO UDĚLÁME	CO TO ZNAMENÁ
Nejdříve vynásobíme proporci zelených kuliček v populaci proporcí červených. Tato proporce musí být vyjádřena jako desetinný vzorek, ne v procentech. (Tedy, kdyby v populaci bylo 50% červených a 50% zelených budeme počítat 0.5 krát 0.5.)	Homogenita vzorku má vliv na velikost chyby. Čím nerovnoměrnější je distribuce ve vzorku, tím menší bude chyba a tím užší bude interval spolehlivosti. Kdyby na příklad v populaci bylo 90% zelených kuliček a velikost vzorku by byla 100, vypočítaný konfidenční interval by byl ± 6%. Kdyby ve stejně velkém vzorku byl stejný počet zelených jako červených kuliček, konfidenční interval by byl mnohem širší: ± 10%

Vypočítaný násobek vydělíme velikostí vzorku.	Čím větší vzorek, tím menší je směrodatná chyba a tím užší bude konfidenční interval. V případě, že by v populaci byla stejná proporce zelených a červených kuliček, ve vzorku 100 pozorování, by interval byl ± 10%; ve vzorku 400 pozorování by byl mnohem užší: ±5% a ve vzorku 1000: ±3%.
Nakonec vypočítáme druhou odmocninu z výsledku dělení.	To je transformace do čísla zajímavých vlastností. Ti, kdo jsou trošku seznámeni se statistikou, vidí už teď souvislost s konceptem směrodatné odchylky. My ostatní to pochopíme trochu lépe, až budeme mluvit o směrodatné odchylce v naší statistické kapitole.

A teď nám už zbývá jen jedno. Rozhodnout se, jakou hladinu významnosti chceme přijmout, a pak vypočítat interval spolehlivosti.

Směrodatná chyba má jednu pozoruhodnou vlastnost: do intervalu vymezeného ±1 standardní chybou od hodnoty pozorované ve vzorku připadne správná hodnota, existující v populaci, přibližně v 68 případech ze sta. Tak bychom dostali interval spolehlivosti na 68 % hladině významnosti. To ovšem není zdaleka dost vysoká pravděpodobnost. Abychom vypočítali interval spolehlivosti na úrovni, jaká je vyžadována v našem oboru, musíme přičíst a odečíst směrodatnou chybu dvakrát. Jinými slovy: **interval spolehlivosti na 95% hladině významnosti je dán rozmezím ± 2 směrodatné chyby od hodnoty, naměřené ve vzorku.** Rozmezí ±3 směrodatné chyby nám definuje ještě mnohem striktnější interval na hladině 99.9%. Ten je užíván zejména v přírodních vědách.

A teď už víme dost, abychom mohli představit další, opravdu překvapivou vlastnost náhodného výběru:

Velikost směrodatné chyby, a tedy i konfidenční interval (interval spolehlivosti) nezávisí vůbec na velikosti populace.

Jedině velikost vzorku a jeho homogenity ovlivňují velikost chyby.

Dr.Watson:

Počkejte, počkejte! Chcete mi namluvit, že řekněme vzorek 300 respondentů vykáže stejnou chybu, když reprezentuje populaci továrny s 800 dělníky, jako stejně velký vzorek, který reprezentuje město s 50.000 obyvatel, nebo dokonce zemi s 200,000.000 občanů? Já tomu prostě nevěřím!

Neuvěřitelné, a přece je to pravda, pokud ovšem distribuce zkoumané proměnné je ve všech těch populacích stejně homogenní. A pokud mi ještě nevěříte, podívejte se znovu na popis výpočtu směrodatné chyby. Najdete tam zmíněnou proporci zelených a červených kuliček, velikost vzorku a to je vše. Ani zmínka o populaci.

To, co víme, by nám mohlo dát dostatečnou informaci, abychom mohli **navrhnout velikost vzorku, jakou potřebujeme vzhledem k velikosti chyby, jakou jsme ochotni riskovat.** V praxi to však není snadné: pro výpočet směrodatné chyby potřebujeme znát homogenitu populace vzhledem k našim proměnným, rozptyl těchto proměnných. Většinou tuto znalost nemáme. Existují sice techniky, které nám umožní tuto informaci odhadnout, ale tyto techniky jsou buďto nákladné nebo nepřesné.

A tak v tvrdé praxi denního života výzkumníka spoléháme na zkušenost a na zdravý rozum. Můžeme se třeba zamyslit nad tím, které kombinace proměnných jsou pro nás nejdůležitější. Představíme si kolik polí bude mít tabulka (nebo tabulky) a navrhneme, kolik pozorování musí každé pole v těchto tabulkách obsahovat - prázdná pole, nebo pole s málo pozorováními mohou podstatně zkreslit výsledky statistické analýzy. Zaměřme se raději na dost vysoké minimum; někdy navrhovaný průměr 10 pozorování na jedno pole tabulek může být nezdravě optimistický. Data ve skutečnosti nebudou do všech polí rozdělena rovnoměrně; některá pole budou přeplněna a jiná téměř prázdná. Nadto v každém výzkumu máme mnoho proměnných, s různým počtem kategorií, někdy nevíme předem, které kombinace proměnných přinesou nějaké zajímavé výsledky, a tak si zaslouží hlubší analýzy atd. Zkrátka, teoretizování o velikosti vzorku patří spíše na stránky učebnic než do praxe sociologického výzkumu. Tam aplikujeme následující, velice nevědecké, ale velice praktické pravidlo: **Snažme se vytvořit**

co největší vzorek, jaký nám naše časové a finanční podmínky dovolují; ne však za cenu vážného narušení pravidel náhodného výběru. Doba pro aplikaci naší znalosti o intervalech spolehlivosti přichází v praxi teprve v etapě statistické analýzy sebraných dat. Pak je to ovšem velice důležité.

A teď ještě jedno důležité varování:

> Velikost směrodatné chyby se týká jen zkreslení, vyvolaného rozdíly mezi vzorkem a populací. Nevztahuje se, bohužel, na zkreslení vyvolané jinými typy redukce a transformace informací. Tato zkreslení jsou pro nás většinou mnohem nebezpečnější a my nemáme žádný nástroj, jak měřit velikost těchto omylů.

5.3. Jak správně házet korunou

Dr. Watson:

Já už vidím, že náhodný výběr je výborný. Hned to začnu používat. Vždycky jsem chtěl vědět, co si lidé v Praze myslí o mojí politické straně. Hned začnu pracovat na hypotézách a otázkách pro rozhovor. Od pondělí budu každé dopoledne na Václaváku a budu se vyptávat náhodně vybrané osoby...

Pokud náš pošetilý přítel doufá, že jeho výsledky budou reprezentovat mínění pražské populace, je ještě mnohem pošetilejší, než jsme si mysleli. Víme přece, že při náhodném výběru každý člen populace musí mít stejnou pravděpodobnost, že bude vybrán. Watsonův vzorek by byl silně zkreslený.

Cvičení 5.3.

Navrhněte prosím, jak by se Watsonův vzorek lišil od pražské populace.

Tedy jasně vidíme, že tento vzorek by snad mohl být reprezentativní pro populaci definovanou asi takto: osoby, které se nacházejí na Václaváku ve všední den dopoledne, v dané roční době. Pro nějaké speciální účely by mohla být taková populace zajímavá: kupř.

pro plánování obchodních strategií pro obchody na Václaváku, rozhodně však ne pro problémy spojené s politickou orientací obyvatel. Ale i tak by byla **náhodnost**, a tedy i reprezentativnost takového výběru problematická. Dr. Watson, protože je v podstatě konzervativní, by se mohl ostýchat oslovit méně konvenčně oblečené osoby. Kdyby takový výběr prováděl můj syn, půvabné mladé ženy by byly ve vzorku přereprezentovány. Kdybych prováděl výběr já, pak by byly podreprezentovány, protože jsem stydlivý. Ono se vůbec zdá, že lidská mysl není schopna pracovat opravdu náhodně.

Můžeme si to dost snadno vyzkoušet. Požádejte větší skupinu lidí - třeba třídu studentů - aby každý napsal na kousek papíru jakékoliv číslo mezi 1 a 10. Bez dlouhého přemýšlení musí napsat to, co jim přijde na mysl. Je-li skupina dost velká, je vysoká pravděpodobnost, že číslo 7 bude mít daleko nejvyšší frekvenci. Proč, to nevím, a předem můžete zavrhnout teorii vlivu sedmy v naší mariášnické kultuře; v Kanadě to funguje také, a jak! Snad to má něco dělat s tradiční mystikou čísel, ale v každém případě to krásně dokumentuje, že náš mozek je velice špatným generátorem náhodnosti. Musíme jej nahradit něčím neosobním. Hodit si korunou? Zatřepat krabicí?

Pomůcky, které v praxi při výběru náhodného vzorku používáme, skutečně imitují takové mechanismy. Mohli bychom třeba napsat jména všech členů populace na papírky, dát do klobouku, kloboukem pořádně zatřepat a pak poslepu vytáhnout tolik papírků, kolik osob potřebujeme do vzorku. Ovšem většinou by to musel být pěkně velký klobouk a v každém případě je to dost nepohodlný postup. Můžeme jej však dobře imitovat. Prostě jednotlivce v seznamu populace očíslujeme a pak použijeme "něco" co produkuje náhodná čísla a vybíráme ty jedince, jejichž číslo se s těmi náhodnými shoduje. Říká se tomu

prostý náhodný výběr

Jednoduchá však v tom není generace těch náhodných čísel. Kdysi se k tomu užívala taková podivná "kostka", mnohohran s deseti stejnými plochami, na každé z nich byla jedna z číslic od 0 do 9. Prý bylo obtížné vyrobit takovou "kostku", aby byla "poctivá", to je aby každá číslice měla stejnou pravděpodobnost, že "padne". Ještě do nedávna jsme používali tabulky náhodných čísel, dost tlusté knihy číselných skupin, o nichž nám matematici řekli, že v nich za takových a takových okolností nebyli schopni objevit žádnou pravidelnost. Dodnes jsou výtahy z těchto tabulek přetiskovány téměř v každé učebnici výzkumných metod. Jejich

správné používání rozhodně není nejzábavnější kratochvíle, ale někdy nám prostě nezbude nic jiného. Naštěstí dnes každá lepší kalkulačka a ovšem každý, i nejmenší osobní počítač umí produkovat náhodná (matematik by řekl "quasi-náhodná") čísla. Tenhle přístup má velikou výhodu: program produkuje náhodná čísla jenom v tom rozsahu, v jakém je potřebujeme. Řekneme počítači, jak je velká populace, třeba 300 a program pro nás vyprodukuje náhodná čísla jenom v rozsahu od 1 do 300. Tabulky náhodných čísel jsou nejméně pěticiferné. Pro naši velikost populace použijeme ovšem jen první nebo poslední tři sloupce číslic, ale i tak sedm z deseti nalezených nebudeme s to použít. Kalkulačka nebo počítač jsou mnohem efektivnější, a když si s tím nevíte rady, obraťte se na sousedova syna. a pokud by neměl takový program, většina těch chytrých holek a kluků, kteří vlastní třeba i ten nejmenší Sinclair, je schopna napsat takový program v Basicu za několik minut.

Dr. Watson:

Ale já nemám kalkulačku a všichni sousedi jsou bezdětní. Tak bych si to chtěl zjednodušit. Populace má 500 členů a já chci vzorek ve velikosti 100. Proč bych nemohl vzít jednoduše každou pátou osobu ze seznamu?

Tentokrát Dr. Watson promluvil pro změnu moudře. Technika, kterou navrhl se opravdu používá. Říká se jí systematický výběr. Nenechte se však zmást tím názvem; je to opět technika náhodného výběru.

Systematický výběr:

V systematickém výběru je do vzorku zahrnuta každá N-tá jednotka ze seznamu. Velikost kroku (N) dostaneme, když vydělíme velikost populace velikostí požadovaného vzorku. Důležité však je, aby první jedinec byl vybrán náhodně a teprve od tohoto výchozího bodu budeme vybírat každou N-tou jednotku.

Tento postup však nemůžeme použít, když jsou seznamy řazeny podle nějakého systematického schématu. Naše pohádka ilustruje něco, co se v praxi opravdu stává.

Pohádka pro odrostlejší děti 10.

O výběru, který byl příliš systematický

Bylo, nebylo, kdesi existovalo malé království, které se jmenovalo Org. Bylo to království, kde všechno bylo velice dobře zorganizováno, a přesto byl každý šťastný a spokojený. Každý, až na vojáky základní služby. Ti si stěžovali na plat, na stravu, na zacházení od představených, na všechno. a protože vše bylo dobře zorganizováno, vláda pozvala zahraničního odborníka, profesora P.I. Tomu, aby provedl výzkum postojů v armádě.

P.I. Toma přijel, zkonstruoval výborný dotazník a vyzkoušel jeho validitu. Protože to království bylo tak malé, že se tam ani počítač nevešel a místní knihovny neměly tabulku náhodných čísel, rozhodl se použít pro konstrukci vzorku techniku systematického výběru. Armáda toho malého království byla taky malá, důstojníci, poddůstojníci i mužstvo dohromady jen 12.000 osob. Profesor P.I. Toma odhadl, že vzorek 200 osob mu poskytne přijatelný interval spolehlivosti a zvolil tedy krok 60. Náhodně vybral prvého jedince. Byla to osoba č. 31 a pak vybíral každého dalšího šedesátého vojáka. Výsledky výzkumu byly prostě náramné. Ještě nikdo nikde nezkoumal tak spokojenou armádu. Každý byl šťastný v tom malém šťastném království - až do příštího jara, kdy začalo krvavé povstání vojáků základní služby.

Ale vy už víte, co se stalo: Prostě, v království Org vše bylo dobře organizováno. I seznamy členů armády byly uspořádány po četách, v každé četě nejdříve dva důstojníci, pak tři poddůstojníci, pak mužstvo základní služby a každá četa měla ne více, ne méně než 30 osob. a nás profesor měl smůlu, protože zvolený krok se shodoval přesně nejen s dvojnásobkem velikostí čety, ale také proto, že první náhodně vybraná osoba byl důstojník a tedy každá následující osoba musela být také důstojník. Poddůstojníci a vojáci základní služby nebyli zahrnuti do vzorku vůbec.

Nemysleme si, že takové zkreslení patří jen do absurdního světa pošetilých pohádek. Mnohé ze seznamů populací jsou systematicky uspořádány, kupř. žáci škol podle tříd, dělníci podle dílen atd. Někdy systém, podle kterého je seznam organizován, nemusí být na prvý pohled zřejmý. Kupř. byty na sídlištích ve velkých obytných budovách bývají identifikovány tříciferným čísly. Prvá číslice definuje podlaží, druhé dvě byt na podlaží. Protože půdorys se na každém podlaží opakuje, byty se stejnými posledními číslicemi budou mít obdobné vlastnosti, budou třeba větší či menší než byty ostatní, a to by opět při systematickém výběru mohlo produkovat zkreslení.

Podívejme se teď na jiný typ náhodného výběru, která by býval mohl zachránit profesora P.I. Tomu před zmíněnou blamáží:

Náhodný stratifikovaný výběr: Populace je rozdělena do skupin homogenních vzhledem k nějakému jasnému kritériu a jedinci jsou vybíráni do vzorku náhodně z těchto skupin.

Profesor Toma měl začít s třemi seznamy; se seznamem populace důstojníků, s jiným, zahrnujícím jen poddůstojníky, a konečně se seznamem vojáků základní služby. Z každé populace by pak byl vybrán náhodný vzorek, třeba technikou systematického výběru, a v našem malém království by k povstání třeba nedošlo. Ve skutečném světě, například při výzkumu studentů určité školy, bychom vybírali jedince zvlášť pro každý ročník. Při jiných výzkumech by populace mohla být stratifikována podle volebních obvodů, při výzkumu zaměstnanců továrny by mohl být výběr prováděn zvlášť mezi dělníky a zvlášť pro administrativu.

Stratifikovaný náhodný výběr má ještě jednu dodatečnou výhodu: snižuje velikost směrodatné chyby, a tedy i interval spolehlivosti. Třeba si ještě pamatujete, že chyba klesá s rostoucí velikostí vzorku a s přirůstající homogenností populace. Logika toho je zřejmá: když v populaci je pro kandidáta A 98% voličů a pro kandidáta B jen 2%, předpovědět, kdo vyhraje volby, je mnohem snadnější, než kdyby preference byle třeba 55% pro A a 45% pro B. Ve stratifikovaném výběru jsou vzorky podskupin zcela homogenní vzhledem k proměnné, podle které byly stratifikovány: ve skupině jsou jenom vojáci základní služby, nebo jenom posluchači druhého ročníku atd. Pro stratifikační proměnnou je tedy směrodatná chyba nulová a pro všechny jiné proměnné, které jsou s touto proměnnou asociovány, bude tato chyba podstatně menší.

A teď se podíváme na velmi zvláštní typ výběru, na vícestupňový náhodný výběr. Je to technika velice pracná, náročná a drahá, ale, jak hned uvidíme, velice důležitá a nenahraditelná.

Vícestupňový náhodný výběr

se provádí ve dvou nebo více krocích. Nejdříve jsou náhodně vybrána určitá přirozená seskupení, a pak teprve jsou náhodně vybíráni jedinci z oněch vybraných seskupení.

K čemu je to dobré? Pro ilustraci jednoho aspektu vás pozvu na výlet na jiný kontinent. Představte si, že bychom měli dělat výzkum na náhodném vzorku reprezentujícím dospělé obyvatelstvo Kanady. Kanada má něco přes dvacet milionů obyvatel, ale její plocha je téměř

10,000.000 čtverečných kilometrů. Řekněme, že velikost vzorku by byla 1.000 jedinců, a tak bychom teoreticky měli jednoho respondenta na deset tisíc čtverečných kilometrů. Ve skutečnosti by to bylo mnohem méně, obrovské rozlohy země jsou prázdné. Ale i tak jsou rozměry země obrovské a takové by byly i náklady. Při dané velikosti vzorku bychom měli nejmenší potíže s nejlidnatějšími provinciemi. V Quebecku bychom měli asi 290 respondentů, v Ontariu přibližně 350. Ale v Northwest Territories jednoho, nebo dva a ti by nás přišli pěkně draho. Pokud bychom neměli velké štěstí, museli bychom, abychom je zastihli, najmout hydroplán, helikoptéru nebo psí spřežení. Ale i v nejlidnatějších provinciích, a nebo i v prostorově malé zemi s tak vysokou hustotou obyvatelstva jako má Československo, rozptyl populace v prostoru podstatně zvyšuje náklady a nesmírně ztěžuje organizaci výzkumů. (Kupř. tazatelské týmy jsou organizovány a školeny lokálně; to snižuje cestovní náklady. Ale je jen omezený počet terénních center, které jsme schopni organizovat a financovat.) Tady je právě oblast uplatnění vícestupňového náhodného výběru. Můžeme postupovat třeba takto:

1. Nejdříve vybereme náhodně reprezentativní soubor okresů.
2. Pak v každém z vybraných okresů provedeme náhodný výběr obcí.
3. Ve velkých vybraných obcích zařadíme ještě další mezistupeň výběru: vybereme náhodně menší prostorové jednotky, třeba volební obvody.
4. Teprve pak vybíráme jedince.

Tímto způsobem obdržíme mnohem kompaktnější vzorek. Respondenti nejsou rozptýleni po celém teritoriu, ale jsou koncentrováni do zvládnutelného počtu regionů. Je-li takový výběr proveden správně, žádné závažné zkreslení reprezentativnosti nehrozí.

Nicméně existuje ještě jedna, dokonce důležitější doména použití tohoto výběru. Největším problémem pro použití pravděpodobnostního výběru v sociologii je fakt, že pro mnoho zajímavých populací žádný seznam neexistuje. Pro mnoho těchto situací je vícestupňový náhodný výběr jediným řešením. Řekněme, že bychom chtěli vytvořit pravděpodobnostní vzorek celé země a žádné spolehlivé seznamy obyvatelstva buď neexistují, nebo nejsou výzkumníkovi dostupné. To je mimochodem situace ve většině zemí světa.
Postup by byl shodný v prvých třech krocích s předchozí tabulkou, ale pak by následovaly dva další, logicky jednoduché, ale pracovně náročné kroky:

4. Ve vybraných malých obcích, nebo městských obvodech, je proveden soupis všech sídelních jednotek (bytů, rodinných domků).
5. Pak je vytvořen náhodný vzorek těchto jednotek.
6. Je vytvořen seznam osob žijících ve vybraných jednotkách a pak jsou opět náhodně vybráni jedinci (nebo obvykle jedinec) do vzorku.

Nejnáročnější je ovšem krok č.4. Představuje obsáhlou práci jak v přípravě, tak i v terénu; záznamy se obvykle opožďují za skutečností, nemusí rozlišovat mezi jednotkami, které jsou obydleny a těmi, které jsou používány pro jiné účely atd. Poslední krok je obvykle prováděn tazatelem přímo v terénu. Náhodnost musí být zaručena i při tomto kroku. Záznamový arch pro interview obsahuje instrukci, v jakém pořadí mají být členové domácnosti zaznamenáváni, a náhodně generované pořadové číslo osoby, která má být interviewována. Bez takové instrukce by tazatel vybral osobu, která je právě dosažitelná, aby se tak vyhnul nutnosti další návštěvy, nebo osobu, která je mu sympatická. Tak by byly kupř. podreprezentovány osoby, které během dne pracují mimo dům.

Někdy aplikace vícestupňového výběru nemusí být obtížná a je přitom velice užitečná. Chtěli bychom třeba studovat na celostátním vzorku mínění studentů dvou nejvyšších ročníků střední školy. Ústřední seznam středoškolských studentů asi neexistuje, ale existuje seznam všech středních škol a každá škola má seznam žáků, sestavený pravděpodobně podle ročníků. Výběr by mohl být prováděn třeba takto: Náhodně by byly vybrány okresy, pak vzorek škol v těchto okresech a jedinci do vzorku by byli náhodně vybráni ze seznamu žáků posledních dvou ročníků.

Před časem jsme zkoumali postoje starších osob k možnosti vstoupit do institucí pro staré občany (Disman & Disman, 1989). Naším cílem bylo sledovat vliv etnické kultury na tyto postoje; porovnávali jsme postoje Portugalců a Italů žijících v Torontu, ve věku 65 nebo starších, s postoji stejně starých Kanaďanů, jejichž mateřským jazykem je angličtina.

Vytvoření vzorku nebylo snadné. Osoby starší než 65 let představují 11% torontské populace, z těchto starších osob je jen 5% Italů a 1% Portugalců. (To znamená, že Portugalci ve věku 65 a více představují asi 0.11% z torontské populace.) Kdybychom tedy chtěli interviewovat

100 Italů a 100 Portugalců, museli bychom kontaktovat asi 100.000 domácností, a to je ovšem nemožné přinejmenším z finančních důvodů. Naštěstí jsme měli k dispozici seznamy osob pro daňové účely a tyto seznamy zahrnují prakticky všechny dospělé občany. Nadto tyto seznamy zahrnovaly také informaci o věku. Tato informace podstatně zúžila velikost vzorku pro vyhledávací fázi výzkumu. Ale i tak, abychom vyhledali vzorek 100 portugalských respondentů, museli bychom kontaktovat asi 10.000 domácností a i to by bylo nemožné.

Zůstala pro nás tedy otevřena jediná možnost: kontaktovat osoby ze seznamu, jejichž jména znějí italsky nebo portugalsky. Jistě, tato metoda má některé nevýhody. Kupř. portugalské jméno může mít britská manželka portugalského manžela, ale tyto případy byly vyloučeny v předběžném rozhovoru. Do vzorku nebyly zahrnuty osoby s etnicky netypickými jmény, italské nebo portugalské manželky mužů jiného etnického původu atd. Nicméně toto zkreslení - zejména vzhledem k silné tendenci obou národnostních skupin uzavírat sňatek uvnitř etnické skupiny (endogamy) nebylo příliš vážné. Ale i tak - zejména vzhledem k úmrtnosti mezi staršími osobami, vinou nepřesnosti záznamů, a vzhledem ke značné horizontální mobilitě - bylo nutno kontaktovat 652 portugalských adres, s výtěžkem 161 jmen respondentů, odpovídajících naší definici populace.

V tomto případě jméno jako kritérium pro výběr - doufejme - nezpůsobilo vážné zkreslení. Ale nemusí tomu tak být vždycky. Mezi americkými sociology koluje hezká historka, kterou uvedeme v naší pohádce č.11.

Pohádka pro odrostlejší děti 11.

O zrádném písmenu

Bylo před místními volbami v jednom velkém městě na východním pobřeží U.S.A. a skupina politiků si objednala výzkum, předpověď výsledků voleb. V té době mělo město dobrý seznam voličů řazený abecedně. Kartotéky zaplňovaly několik místností. Pro konstrukci vzorku byla použita technika vícestupňového náhodného výběru. Nejdříve byla vybrána náhodně místnost, pak kartotéční skříň a ze zásuvek této skříně byli vybráni technikou systematického výběru jedinci do vzorku.

Výzkum skončil neslavně: jako vítěze vyhlásil kandidáta, který skončil daleko vzadu v poli poražených. Prostě výzkumník se dopustil omylu, ale zejména měl smůlu. Náhodně vybral začátek písmena M, a tak se stalo, že voliči irského a skotského původu, jejichž jména velice často začínají na Mac a Mc, byli silně přereprezentováni. Hlasování ve volbách v USA a Kanadě velmi často sleduje etnickou linii. Není proto divu, že výzkum mylně předpověděl vítězství irského kandidáta. Tomuto zkreslení bylo snadné zabránit, kdyby byl stejný počet voličů vybrán z více kartotéčních skříní. Je ale také pravda, že kdyby bylo vybráno jiné písmeno, ke zkreslení by asi nedošlo.

5.4. Když koruna nepracuje

Zatím jsme viděli členy dvou rodin výběrových technik. Nejdůležitější jsou pravděpodobnostní techniky, založené na náhodném výběru. Jsou velice mocné, zajišťují, že budou dobře reprezentovány všechny známé i neznámé vlastnosti populace. Nadto jen u nich jsme schopni prostředky statistiky odhadnout, nakolik se vzorek liší od populace. Bohužel, zdaleka ne vždy jsme schopni tyto techniky použít. Někdy třeba proto, že pracnost a nákladnost těchto technik přesahuje rámec našich možností. Jindy proto, že neexistuje žádný seznam cílové populace. Nejčastější překážkou je však kombinace obou těchto důvodů. Speciální populace, o kterou se zajímáme, může být rozptýlena mezi celou populací a mít velice nízkou frekvenci. Teoreticky by bylo jistě možné vytvořit veliký vzorek celé populace a pak, po předběžných rozhovorech, vybrat jen ty jedince, kteří odpovídají definici naší cílové populace. Jak jsme si ilustrovali na příkladu výběru starých Portugalců, z hlediska nákladů by to bylo prostě nemožné. My jsme měli štěstí, byli jsme schopni zimprovizovat seznam populace, ale to se stává spíše výjimečně.

Jako prvou techniku tvorby vzorku jsme v této kapitole diskutovali kvótní výběr. Reprezentuje druhou skupinu výběrových technik, které nejsou založeny na teorii pravděpodobnosti, ale na logickém úsudku. Kvótní výběr je pravděpodobně nejspolehlivější mezi těmito technikami,

ale opět ne vždy je možno jej použít. Může být aplikován jen tehdy, když máme dostatečnou znalost o populaci, abychom její strukturu mohli imitovat ve struktuře vzorku. Do téže skupiny patří účelový výběr, někdy i anketa, a bývá sem zařazována i technika sněhové koule (snowball sampling).

Účelový výběr
je založen pouze na úsudku výzkumníka o tom, co by mělo být pozorováno a o tom, co je možné pozorovat.

Jak vidíte, není to příliš vědecký přístup, ale velice často jediný, který nám zbývá. Je používán i profesionálními agenturami, které provádějí za úplatu výzkum trhu. Řekněme, že byste při sobotním nákupu v Bílé labuti byli osloveni mladým mužem a dotazováni na to, co si myslíte o určité skupině výrobků. Na jakou populaci se výsledky takového výzkumu vztahují?

Dr.Watson
Ale to je přece jednoduché! Na lidi, kteří nakupují v obchodních domech!

Jako obvykle, Dr. Watson je příliš optimistický. Takto konstruovaný vzorek by reprezentoval přinejlepším populaci osob, které nakupují v Bílé labuti v sobotu dopoledne a právě v této roční době. A kdybychom měli být opravdu přesní, museli bychom ještě dodat, že se závěry vztahují jen na ty jedince z takto definované populace, kteří jsou ochotni odpovídat na otázky daného typu. Není to příliš široká a dobře definovaná populace, ale pro účely výzkumu trhu by mohly být takto získané informace určitě užitečné.

Účelový výběr nám téměř nikdy neumožní nějakou opravdu širokou generalizaci našich závěrů, ale to neznamená, že tyto závěry nejsou užitečné. Jen nesmíme předstírat jiným, a především ne sobě, že tyto závěry platí pro každého jedince ve vesmíru.

Při použití účelového výběru musí výzkumník jasně, přesně a otevřeně definovat populaci, kterou jeho vzorek **opravdu** reprezentuje.

Užití účelového výběru je pro některé populace jediným řešením. To platí kupř. pro etnické minority; snad v žádné zemi neexistují spolehlivé a vyčerpávající seznamy takových skupin. Pak nezbývá nic jiného, než použít jako výchozí bod seznamy členů etnických organizací. Takový vzorek jistě nebude reprezentovat ty příslušníky etnické skupiny, kteří nejsou v žádné z takových skupin organizováni. Je pak na výzkumníkovi, aby posoudil, s použitím znalosti skupiny, jak dalece jsou jeho závěry zobecnitelné. Kupř. Pejovič (1990) zkoumal vzdělávací aspirace středoškolských studentů chorvatského původu, žijících v Torontu. Jeho závěry jsou velice zajímavé a závažné. Zdá se, že pro tuto skupinu neplatí obvyklé socioekonomické determinanty aspirací, které americká sociologie má tendenci považovat za univerzální. Pejovič užil techniku účelového výběru. Východiskem bylo členstvo různých chorvatských kulturních a sociálních organizací, účastníci různých chorvatských společenských akcí atd. Pejovič nikde nepředstírá, že jeho závěry platí mimo jeho vzorek. Nicméně síla zjištěných souvislostí a známá fakta o kulturní a socioekonomické homogenitě této etnické skupiny naznačují, že je silně pravděpodobné, že podobné výsledky bychom mohli dostat i pro většinu jiných mladých Chorvatů, žijících ve velkých kanadských městech. **Důkaz** pro to by ovšem mohl být získán jen opakováním výzkumu na vzorcích vytvořených z dalších populací. Tedy i technicky vzato nereprezentativní vzorek může někdy poskytnout hodnotné výsledky. Ne však vždycky a ne automaticky a musíme si být vědomi toho, že je jen náhražkou za pravděpodobnostní výběr.

Některé techniky vytváření účelového vzorku jsou velice problematické. Bohužel, stále je hojně používána anketa.

V anketě je výběr jedinců založen na rozhodnutí respondenta zodpovědět otázky uveřejněné v masových sdělovacích prostředcích.

Definovat populaci, ke které se nálezy ankety vztahují, je skutečně nemožné. Nejsou to čtenáři určitých novin nebo časopisu. To by bylo ještě dobré. Vzorek se však liší od celé populace právě tím, že to jsou ti, kteří zodpověděli anketu. Maximálně můžeme říci, že lidé ve vzorku jsou více motivováni, než ostatní čtenáři. a to je velice slabá definice. Ani velikost vzorku

nepomůže. Správně konstatuje Zich (1976, str. 207) že anketa Rudého práva, která získala vzorek větší než 110 tisíc není reprezentativní, i když v základních demografických ukazatelích se dosti shodovala se strukturou celé dospělé populace. Problém je v samovýběru respondentů. Ale to už známe z našeho příkladu vojáků, filmu a postojů k U.S.A. Poznávací hodnota ankety je podle mého názoru pod hodnotou dobře a zodpovědně napsaného fejetonu.

A konečně tu máme techniku "snowball sampling", techniku sněhové koule. Podle mého názoru tato technika vůbec do této kapitoly nepatří. Je to technika identifikace populace, a ne vytvoření reprezentativního vzorku. Ale všechny učebnice, které znám, ji zařazují mezi výběrové techniky, a tedy i my sledujeme toto schéma. Ale posuďte to sami.

"Snowball Technique" spočívá na výběru jedinců, při kterém nás nějaký původní informátor vede k jiným členům naší cílové skupiny.

Nejlépe si to ukážeme na jednoduchém příkladu. Třeba bychom chtěli studovat mocenskou strukturu v malé obci. Identifikovat oficiální vlivné osoby, jejichž pozice je formálně definována by nebylo těžké. U nás by to byl třeba starosta, v nedávné minulosti straničtí funkcionáři, SNB atd. Ale vliv v obci mohou mít osoby, jejichž vliv není definován funkcí, a tato část souboru vlivných osob se liší podle místních okolností. V některé obci může být vlivnou osobou ředitel školy či továrny, v jiné obci mohou být osoby v takových funkcí bez vlivu a významný vliv na rozhodování může mít kněz, nebo vlivný a mocný rodák, který v obci již dlouhá léta nežije. To vše je pro výzkumníka, který přichází z venku, neviditelné. Tady je na místě uplatnit výběr technikou sněhové koule.

Výzkumník začne rozhovorem s jasně definovanou osobou, třeba starostou. V tomto rozhovoru požádá respondenta, aby jmenoval další vlivné osoby. Ty jsou pak interviewovány a každá z nich dostane i stejnou otázku o vlivných lidech. Po určitém počtu rozhovorů se již jména nových vlivných osob neobjevují. Výzkumník může prohlásit, že vzorek je "teoreticky nasycen". Populace vlivných osob v obci byla jasně identifikována a náš vzorek je totožný s touto populací.

Technika sněhové koule, kde jména dalších osob se v řetězci rozhovorů "nabalují" jako sněhová koule (taková, kterou je znázorňována lavina v kreslených vtipech) je

nenahraditelným nástrojem pro zkoumání populací, které existovaly jen dočasně: účastníci určitých demonstrací, svědkové katastrofy nebo jiné řídké události atd. Zde většinou teoretické nasycenosti vzorku nedosáhneme a aplikace této techniky má opravdu charakter konstrukce účelového vzorku.

Termín "teoretická nasycenost" byl uveden Glaserem a Strausem (1967) v souvislosti s jejich konceptem "grounded theory", snad nejdůležitějším epistemologickým nástrojem pro kvalitativní výzkum. Technika sněhové koule hraje pod jménem "teoretický výběr" velice důležitou úlohu. Má zde do jisté míry funkci ověřování validity. Ale k tomu se ještě vrátíme s celou řadou podrobností. Doufám, že to bude docela zajímavé.

5.5. Koruna přece jen není všechno

Techniky náhodného výběru opravdu produkují nejlepší možnou reprezentaci populace. Jenomže je reprezentativní pouze za předpokladu, že všichni vybraní jedinci se opravdu na výzkumu zúčastnili, to jest, že například zodpověděli naše otázky. Výpočet směrodatné chyby je plně založen na tomto očekávání. V příští kapitole uvidíme, že je to příliš optimistický předpoklad. V současné době procento osob, které odmítly tazatele nebo nevrátily dotazník, téměř všude roste. U dotazníku návratnost často nedosáhne ani padesáti procent.

Dr.Watson

Ale to přece vůbec není problém. Já vím, že nám v jedné z dalších kapitol řeknete, že dotazník je jednou z nejlevnějších technik sběru dat. Tak když potřebuji ve vzorku 300 jedinců, prostě rozešlu 900 dotazníků a tak dostanu vzorek i větší, než skutečně potřebuji.

Jistě už víte, proč by tento recept nefungoval: populace, která odpověděla, není totožná s tou, která odmítla odpovědět. Liší se v něčem, co bylo důvodem pro toto rozhodnutí, a pravděpodobně ono "něco" je silně spojeno s problémy, na které je výzkum zaměřen. Obvykle jsme o těchto důvodech schopni jenom spekulovat. Obávám se, že tu musím uvést nový typ nepříjemné redukce informací:

Redukce negativním samovýběrem vzniká tehdy, když část jedinců, vybraných do vzorku, odmítla na výzkumu participovat. Tento typ redukce může vážně ohrozit reprezentativnost vzorku.

Toto je vážný problém. Tak vážný, že před několika lety byl ústředním tématem výročního zasedání Americké statistické společnosti. Vidíte, na začátku této kapitoly jsme si pochvalovali, že redukce populace na vzorek je logicky, technicky a metodologicky dobře propracovanou operací, kde riziko zkreslení je menší, než v jiných výzkumných operacích. Je to stále pravda, ale přece i zde máme zranitelné místo. Neznáme žádný univerzální lék na tento neduh. Jediné řešení je usilovat o co nejvyšší návratnost. U některých technik sběru informací je to snadnější, u některých je to téměř nemožné. Ale tohle už patří do příští kapitoly.

Řešení úkolů z kapitoly 5.

Cvičení 5.1.

Jistěže ne. Lidé, kteří neměli telefon (a těch bylo v roce 1936 velice mnoho), ti kteří nevlastnili auto, nebyli členy organizací, tedy lidé náležející do nižších socioekonomických vrstev byli ze vzorku opravdu vyloučeni.

Cvičení 5.2.

Tohle nebyla poctivá otázka. Kvótní výběr může být aplikován jen na populaci, jejíž vlastnosti relativně dobře známe. V našem případě bychom mohli nanejvýše navrhnout něco o taxikářích, vrátných v hotelech a prý i příslušnících bezpečnostních orgánů, ale rozhodně by to nebylo dost pro konstrukci kvótního vzorku.