K problematice odmítnutí rozhovoru v sociologických výběrových šetřeních

JAN HERZMANN
Ústav pro výzkum veřejného mínění při FSÚ, Praha

JAN KÁBRT[1]
Matematicko-fyzikální fakulta UK, Praha

Získávání empirického materiálu v konkrétních sociologických výzkumech je velmi často založeno na metodách výběrového zjišťování. Jednou ze základních metodologických otázek výběrových postupů je problém jejich reprezentativity, jímž se zevrubně zabýval J. Řehák v Sociologickém časopise [1978]. V jeho článku je zmíněn závažný zdroj porušení reprezentativity — nepokrytí cílové populace. Tento problém je vlastní všem výběrovým postupům — pravděpodobnostnímu, kvótnímu i záměrnému — a je jednou z příčin vzniku takzvaných *nevýběrových chyb*, o nichž se více či méně podrobně zmiňuje řada učebnic výběrových šetření [například Cochran 1967; Konijn 1973]. Nevýběrovými se v literatuře nazývají takové chyby ve výpovědích o cílové populaci, které vznikají do určité míry nezávisle na zvoleném výběrovém postupu. H. S. Konijn uvádí čtyři zdroje takových zkreslení:

a) v době určené pro terénní sběr dat se nepodaří kontaktovat všechny osoby, od nichž měly být získány údaje;
b) některé kontaktované osoby odmítnou poskytnout údaje;
c) údaje zjištěné od jednotlivých osob mohou být záměrně či mimovolně nepřesné a zkreslené;
d) chyby mohou vzniknout při zpracování dat — při kódování, záznamu na počítačové medium, zpracování v počítači apod.

Všechny uvedené zdroje zkreslení jsou v praxi velmi závažné a chyby jimi způsobené nejsou ani zdaleka zanedbatelné. Proto jim je v poslední době věnována v odborné literatuře i v konkrétních výběrových šetřeních stále větší pozornost. Pro sociologická dotazníková šetření, jejichž typickým představitelem je výzkum veřejného mínění, jsou zvláště velkým nebezpečím chyby způsobené odmítnutím odpovědi, neboť lze opodstatněně předpokládat, že ta část populace, která odmítá poskytnout údaje, je z hlediska předmětu výzkumu podstatně odlišná od ostatních jedinců. Proto je při každém sociologickém výběrovém šetření nezbytné zjistit, jak velká část populace odmítá spolupracovat, nebo — řečeno v duálních pojmech — jaká část populace je šetřením pokryta.

Cílem našeho článku je ukázat možnost odhadu velikosti té části populace, která je při určitém tématu výzkumu a při daném výběrovém postupu (tj. při daném výběrovém plánu, dané tazatelské síti a daném předpisu pro práci tazatelů) ochotna poskytnout požadované informace. *Jedince náležející do této skupiny budeme nazývat*

(1) Autor *Dodatku.*

(potenciálními) respondenty.[2] Chceme se přitom věnovat problematice odhadu podílu respondentů na základě samotného výběrového šetření a stranou ponecháváme možnosti zjišťování tohoto údaje pomocí doplňkových výzkumů (např. nezávislého dotazníkového šetření, v němž je kladena otázka: „Kdybyste byl požádán, abyste řekl svůj názor na ....., byl byste ochoten odpovědět?"). Pro jednoduchost abstrahujeme v celém článku při odhadování tohoto podílu od zdroje zkreslení a - jinými slovy řečeno - předpokládáme, že příslušnost ke skupině respondentů je nezávislá na pravděpodobnosti kontaktování individua. Tento zjednodušující předpoklad se pravděpodobně v praxi nepotvrdí, což znamená, že odhady obsažené v této stati budou muset být korigovány s přihlédnutím k této pravděpodobnosti. Problémem korekcí se však prozatím nezabýváme.

Postupy a vzorce, které jsme pro tento článek vybrali z literatury nebo odvodili, jsou vypracovány pro pravděpodobnostní výběr, lze však očekávat, že budou používány i pro jiné výběrové postupy. Tak tomu ostatně je i v Ústavu pro výzkum veřejného mínění, kde se při kvótním výběru tradičně používá vzorců platných pro prostý náhodný výběr. Je proto nutné připomenout, že každé přenášení postupů z pravděpodobnostního výběru na kvótní nebo záměrný vyžaduje důkladně ověřit výběrové chování odhadů při konkrétním výběrovém postupu a usilovat o odhalení případné systematické chyby. Proto jsme do článku zařadili také empirické poznatky z výzkumů Ústavu pro výzkum veřejného mínění, na nichž demonstrujeme použití zkoumaných odhadů při kvótním výběru.

Teoretické aspekty odhadu podílu respondentů

Předpokládejme, že výběrovým postupem je prostý náhodný výběr s vracením, tj. že postupně vybíráme z cílové populace jednotky určené ke kontaktování a že při každé volbě jednotky je pravděpodobnost, že bude vybrána j-tá jednotka, $j = 1, \ldots, N$ rovna $1/N$, kde N je rozsah populace.[3] Označme dále počet kontaktovaných jednotek n, R počet respondentů mezi kontaktovanými jednotkami a O počet jednotek, které byly kontaktovány, avšak údaje neposkytly (počet odmítnutí). Platí tedy

$$R + O = n \,. \tag{1}$$

Označme dále p podíl respondentů v cílové populaci.

Podíl respondentů v cílové populaci (v základním souboru) budeme odhadovat na

---

(2) Termín „respondent" se běžně užívá pro označení jedince, od něhož byly získány údaje. V tomto článku chápeme pojem respondent šířeji, ani zde však nejde o abstraktní vymezení osoby ochotné poskytovat údaje. Je patrné, že osoba, která je potenciálním respondentem při zaručeně anonymním výzkumu, může odmítnout dotazování při neanonymním šetření, že jedinci, kteří jsou potencionálními respondenty při řízeném rozhovoru, mohou být odmítajícími při anketě apod. Současně je třeba mít na paměti, že příslušnost daného jedince do skupiny respondentů je závislá také na tématu výzkumu. V tomto článku budeme předpokládat, že tematika šetření a způsob sběru dat jsou dány.

(3) Připouštíme tedy, že některé jednotky mohou být pro daný výzkum vybrány vícekrát. Pravděpodobnost toho, že některá jednotka je ve výběrovém souboru zahrnuta vícekrát, je však při výběru malé části základního souboru zanedbatelná. Proto v článku nepracujeme s výběrem bez vracení, který přesněji popisuje prakticky používané postupy, ale s jednodušším výběrem s vracením.

základě údajů o počtu respondentů R mezi n kontaktovanými jednotkami. Přirozené je zavést odhad

$$\hat{p} = \frac{R}{n},$$

který budeme nazývat *prostým odhadem*. Vlastnosti tohoto odhadu však závisí na tom, jakým způsobem je určen počet kontaktovaných jednotek n. V této stati se budeme zabývat třemi základními možnostmi:

A. počet kontaktovaných jednotek, tj. počet pokusů o získání informací, je dán předem — této situaci odpovídá model s pevným počtem pokusů;
B. předem je stanoveno, kolik respondentů má být kontaktováno, od kolika jednotek mají být získány údaje, a počet kontaktovaných jednotek n je náhodný; zde hovoříme o modelu s pevným počtem responzí.
C. předem může být stanoven nejen požadovaný počet responzí, ale i požadovaný počet odmítnutí, n je opět náhodná veličina; v tomto případě budeme hovořit o modelu se stanoveným minimálním počtem responzí i odmítnutí.

Je zřejmé, že situace A je nejjednodušší a situace C nejsložitější. Tomu odpovídá i rostoucí složitost příslušných pravděpodobnostních modelů.

A. Model s pevným počtem pokusů

Jak již bylo řečeno, je model s pevným počtem pokusů nejjednodušší a v praxi také nejméně nákladný. Tyto dvě přednosti vedou k tomu, že je při výběrových šetřeních přes své značné nedostatky hojně používán. Jeho hlavní nevýhodou je, že v něm nelze předem určit, jaký počet responzí bude získán, tj. jaké množství dat bude k dispozici pro statistické zpracování. To znesnadňuje apriorní určování výběrových chyb použitých odhadů a vylučuje statisticky podložené určení velikosti výběrového souboru. Obvykle se chybějící informace nahrazují údaji zjištěnými v obdobných výzkumech uskutečněných dříve.[4]

V modelu s pevným počtem pokusů jsou tedy R a O náhodné veličiny, R má binomické rozdělení s parametry (n, p). Ze statistické literatury je známo, že prostý odhad $\hat{p}$ je v tomto modelu nevychýlený a je nejlepším odhadem parametru p.[5] Jeho rozptyl je

$$\operatorname{var} \hat{p} = \frac{p(1-p)}{n}. \tag{2}$$

Navíc má $\hat{p}$ při n rostoucím do nekonečna a při pevném p asymptoticky normální rozdělení se střední hodnotou p a s rozptylem (2).

B. Model s pevným počtem responzí

Nevýhodu předcházejícího výběrového postupu lze odstranit tím, že předem neurčujeme celkový počet kontaktů n, ale počet úspěšných kontaktů (responzí) R (tzv. inverzní výběr). V praxi se při tomto výběrovém postupu místo každé odmítající jednotky vybere pro kontaktování jednotka „náhradní" a v pokusech se po-

(4) Pokud hovoříme o výběrové chybě, máme na mysli chybu $t - T_r$ odhadu t nějaké charakteristiky T_r skupiny respondentů, neboť, jak jsme již uvedli, zjištěné údaje vypovídají prakticky pouze o této skupině. Charakteristiku T celé populace nelze na základě znalostí údajů o respondentech odhadovat, neboť informace o její druhé složce — charakteristice skupiny odmítajících T_o — chybí.

(5) Prostý odhad v tomto případě splývá s odhadem maximálně věrohodným i momentovým — viz [Anděl 1978].

kračuje tak dlouho, dokud R nedosáhne stanovené meze $r \geq 1$. V modelu s pevným počtem responzí lze pro stanovení hodnoty r využít všech poznatků z teorie prostého náhodného výběru a předem jsou známy i výběrové chyby odhadů (ve smyslu poznámky 4). Nevýhodou je, že nelze předem stanovit, kolik pokusů o kontakt bude nutno provést, tedy jaké budou finanční a časové náklady na výběrové šetření. K odhadu nákladů se využívá informace o parametru p získané v obdobných šetřeních. Hlavní předností tohoto modelu oproti modelu s pevným počtem pokusů tedy je, že informace získaná v jiných šetřeních se nepoužívá v souvislosti s odhady, které jsou cílem šetření, ale pouze v souvislosti s organizačním zabezpečením.

V modelu s pevným počtem responzí jsou náhodnými veličinami O a n, které se podle (1) liší pouze o konstantu $R = r$. Náhodná veličina O má záporně binomické rozdělení s parametry (r, p):

$$\mathrm{P}(O = k) = \binom{k + r - 1}{k} p^r (1 - p)^k, \quad k = 0, 1, 2, \ldots$$

J. Haldane [1943—46] ukázal, že prostý odhad $\hat{p}$ je v tomto případě vychýlený a že nevychýleným odhadem je statistika

$$\hat{p}_0 = \frac{r - 1}{O + r - 1}.$$

Při $O = 0$ je $\hat{p}_0 \equiv \hat{p} \equiv 1$ a při $O \geq 1$ zřejmě pro každé r platí

$$\hat{p}_0 < \hat{p},$$

prostý odhad tedy podíl respondentů systematicky nadhodnocuje, rozdíl obou odhadů je však velmi malý:

$$\hat{p} - \hat{p}_0 = \frac{n - r}{n(n - 1)} = \frac{O}{n(n - 1)} \leq \frac{1}{n} \quad (\text{při } O \geq 1).$$

Vliv vychýlení je řádově roven vlivu jedné jednotky ve výběrovém souboru. Pro praktické účely je takové vychýlení možno zanedbat. Střední hodnotu prostého odhadu v tomto modelu se nám nepodařilo odvodit. V Dodatku je ukázáno, že pro $r \geq 2$ platí

$$p < \mathrm{E}\hat{p} < p \frac{r}{r - 1 + p}.$$

$$\mathrm{var}\, \hat{p} < p^2 \frac{3r - 3pr - 2 + 3p}{(r - 1)(r - 2 + 3p)} \tag{3}$$

$$\mathrm{var}\, \hat{p}_0 < p^2 \frac{1 - p}{r - 2 + p}. \tag{4}$$

Nerovnosti (3) a (4) neumožňují rozhodnout, který z odhadů má menší variabilitu. Je ovšem možné dokázat, že při r rostoucím do nekonečna má odhad $\hat{p}$ asymptoticky normální rozdělení

$$\mathrm{N}\left(p, \frac{p^2(1 - p)}{r}\right).$$

Při velkém r se tedy rozptyl prostého odhadu příliš neliší od horní meze pro rozptyl odhadu $\hat{p}_0$ (pokud není p příliš malé — například při $p \geq 0{,}1$). Z těchto vlastností

obou odhadů a z jejich velmi podobných definičních vzorců lze soudit, že při dostatečně velkém r (např. $r \geq 100$) je rozdíl mezi nimi pro praxi zcela zanedbatelný.

Navíc platí, že nepatrnou úpravou výběrové procedury lze dosáhnout toho, aby prostý odhad byl nevychýlený [Čermák 1980 : 99—101]. K tomu postačí, abychom výběrovou proceduru zastavili ne při získání r-té, ale až při získání $(r + 1)$-ní responze. Pokus, v němž k tomu došlo, označíme jako $(n + 1)$-ní, do odhadu jej však nezahrneme, tj. představíme si, že vůbec nebyl proveden. Jakmile tuto představu přijmeme, máme $R = r$, $R + O = n$ a prostý odhad je nevychýlený. Jeho rozptyl v daném případě není znám, v citované Čermákově učebnici však lze nalézt jeho nevychýlený odhad.

C. Model se stanoveným minimálním počtem responzí i odmítnutí

Při běžných výběrových šetřeních je snahou minimalizovat počet odmítnutí, přesněji řečeno minimalizovat počet pokusů nutných pro zabezpečení dostatečného počtu kontaktů s respondenty. Jinak tomu bude, stane-li se samo odmítání spolupráce jedním z předmětů výzkumu, tedy budeme-li chtít získat statisticky podložené informace o té skupině jednotek, které spolupráci odmítají. Můžeme si představit metodologický experiment zaměřený na hlubší poznání příčin odmítnutí ankety, při němž občany, kteří anketní list nevrátili, navštíví osobně pracovník výzkumného pracoviště. Údaje ze souboru anketních listů a ze souboru záznamů rozhovorů se pak porovnají a navíc je možno zjistit podrobnější charakteristiky odmítajících osob. Opačný postup je možno zvolit, obáváme-li se, že odmítnutí odpovědí při výzkumném rozhovoru je způsobováno obavami z porušení anonymity. Pak lze odmítajícím předat anketní listy, které sami odešlou výzkumnému pracovišti. Anketní listy slouží jednak ke zjištění důvodů odmítnutí rozhovoru, jednak k získání odpovědí na otázky shodné s otázkami použitými ve standardních dotaznících. Aby bylo možné důvody odmítnutí statisticky zpracovat a rozdělení odpovědí na otázky položené v rozhovoru i v anketě porovnávat, musí být vedle dostatečně velkého souboru respondentů k dispozici také dostatečně velký soubor odpovědí na anketu, tedy dostatečně rozsáhlý soubor odmítajících.

Požadavky takového zkoumání zabezpečuje výběrový postup, při němž určíme předem minimální počet responzí i odmítnutí (požadovaný počet responzí značíme r, požadovaný počet odmítnutí o). Při terénním sběru dat se pak jednotky vybírají tak dlouho, dokud počet responzí R není aspoň r a počet odmítnutí O aspoň o. Při tomto výběrovém postupu jsou R, O i n náhodné veličiny a platí (1). Tento vztah je v daném případě výhodné zapsat ve tvaru

$$(R - r) + (O - o) + r + o = n .$$

Zavedeme-li náhodnou veličinu Z označující počet „nadbytečných" pokusů o kontakt (bez ohledu na to, jde-li o responzi nebo odmítnutí[6]), tj.

$$Z = (R - r) + (O - o) ,$$

liší se náhodné veličiny Z a n pouze o konstantu:

$$Z + r + o = n .$$

(6) Je zřejmé, že při konkrétním výběrovém šetření jsou všechny výsledky „nadbytečných" pokusů stejné — buď jde o samé responze, bylo-li dříve dosaženo stanoveného počtu responzí než stanoveného počtu odmítnutí, nebo — v opačném případě — o samá odmítnutí.

Náhodná veličina Z má zobecněné záporně binomické rozdělení

$$\mathrm{P}(Z = z) = p^r(1-p)^o\left[\binom{k+r+o-1}{o-1}p^k + \binom{k+r+o-1}{r-1}(1-p)^k\right].$$

Prostý odhad parametru p má v tomto případě při provedení jednoho terénního šetření tvar

$$\hat{p} = \frac{R}{R+O}$$

Provedeme-li více terénních šetření, z nichž x skončí dosažením o odmítnutí a y skončí dosažením r responzí, a označíme-li počet „zbytečných" pokusů v i-tém šetření Z_i, můžeme prostý odhad parametru p založený na celé provedené sérii šetření zapsat ve tvaru

$$\hat{p} = \frac{(x+y)\,r + \sum_{i=1}^{x} Z_i}{(x+y)\,(r+o) + \sum_{i=1}^{x+y} Z_i}$$

(Předpokládáme takové přečíslování, po němž jsou šetření, v nichž se čekalo na o-té odmítnutí, na 1. až x-tém místě.) V *Dodatku* je ukázáno, že rozdíl mezi prostým a maximálně věrohodným odhadem konverguje k nule při r a o rostoucích do nekonečna. O momentech těchto odhadů však zatím není obecně nic známo. V případě odhadu z jednoho šetření jsme odvodili základní vlastnosti poněkud odlišného odhadu, který je obdobou nevychýleného odhadu v modelu s pevným počtem responzí

$$\hat{p}_0 = \frac{R-1}{R+O-1} = \frac{r-1}{r+o+Z-1}.$$

Stejně jako v předcházejícím modelu se při velkých r a o odhady p a p_0 liší jen nepatrně. Střední hodnota odhadu $\hat{p}_0$ je v případě, že $R = r$, tj. že stanoveného počtu odmítnutí bylo dosaženo dříve než stanoveného počtu responzí, rovna

$$\mathrm{E}\hat{p}_0 = p\left[1 - \frac{1}{pW}\binom{r+o-2}{r-1}\right],$$

kde

$$W = \frac{1}{p}\sum_{i=0}^{r-1}\binom{r+o-2-i}{r-1-i}p^{-i}.$$

Pro vychýlení D tohoto odhadu lze nalézt při $o \geq 1$ meze

$$0 < -D < p\,\frac{o}{r+o-1}, \tag{5}$$

(Při $o = 0$ se model redukuje na případ s pevným počtem responzí a $D = 0$.) Omezení (5) nedávají příliš optimistický výhled, pokud r není podstatně větší než o. Proto jsme vychýlení D spočítali v několika případech explicitně.[7]

[7] Při velkých r a o se nám vychýlení nepodařilo zjistit, neboť mezivýsledky přesahovaly rozsah konstant dostupné výpočetní techniky.

Tabulka 1. Vychýlení odhadu $\hat{p}_0$ při vybraných hodnotách p, r a o (v % hodnoty p)

a) $p = 0{,}80$

r \ o	100	200	300	500	1000
100	38,0	58,7	69,0	79,4	88,7
200	17,2	37,8	50,2	64,4	79,3
300	7,1	25,2	37,7	53,3	71,2
500	0,1	11,0	22,0	37,6	.
1000	0,0	0,0	4,3	.	.
1500	0,0	0,0	.	.	.

b) $p = 0{,}85$

r \ o	100	200	300	500	1000
100	41,6	61,1	70,8	80,6	89,4
200	21,9	41,1	53,1	66,5	80,5
300	12,1	29,6	41,3	56,0	72,9
500	2,8	16,1	26,6	41,3	.
1000	0,0	2,5	9,6	.	.
1500	0,0	0,0	.	.	.

c) $p = 0{,}90$

r \ o	100	200	300	500	1000
100	44,8	63,2	72,4	81,6	90,0
200	26,2	44,6	55,7	68,4	81,6
300	16,9	33,5	44,6	58,4	74,4
500	7,7	20,7	30,6	44,5	.
1000	0,3	7,3	14,6	.	.
1500	0,0	2,2	.	.	.

Hodnoty uvedené v tabulce ukazují, že odhadu $\hat{p}_0$ nebo nepříliš odlišného prostého odhadu $\hat{p}$ lze v daném modelu alespoň zčásti oprávněně užívat, pokud

$$r \geq 300 \tag{6}$$

$$r \geq 3o\,. \tag{7}$$

Pravděpodobnost, že výběr skončí r-tým kontaktem s respondenty a nikoli o-tým výskytem odmítnutí, je pochopitelně tím větší, čím větší je poměr $r/(r + o)$ oproti p.

Protože vlastnosti odhadu $\hat{p}_0$ jsou odvozeny pouze v takovém případě, není reálné plánovat použití tohoto odhadu, pokud neplatí

$$r \geq \frac{op}{1-p},$$

což je při $p \geq 0{,}8$ ještě přísnější požadavek než (7). Například pro $p = 0{,}9$ a $o = 300$ by r mělo být podle této podmínky alespoň 2700, aby bylo možno očekávat, že nastane situace $R = r$. Z tabulky 1 je patrné, že potom je již vychýlení zanedbatelné.

Je-li v modelu se stanoveným minimálním počtem responzí i odmítnutí provedeno jediné šetření a je-li v tomto šetření $O = o$, tj. stanoveného počtu responzí bylo dosaženo dříve než stanoveného počtu odmítnutí, je odhad $\hat{p}_0$ vždy kladně vychýlený. Tento případ lze převést na předchozí, spokojíme-li se s odhadem parametru $q = 1 - p$, který vyjadřuje podíl odmítajících v populaci.

Podíl responzí a odmítnutí ve výzkumech ÚVVM

Ústav pro výzkum veřejného mínění v Praze používá ve svých výzkumech metodu kvótního výběru. Při terénním sběru dat je místo každého odmítajícího jednotlivce vybrán náhradník a výběr pokračuje tak dlouho, dokud každý tazatel, který se na výzkumu podílí, nezíská předepsaný počet responzí. Až na předpoklad prostého náhodného výběru jde tedy o aplikaci modelu B. Pro stanovený podíl respondentů se užívá prostého odhadu. Ve výzkumech ÚVVM z let 1978—80, v nichž byla základním souborem populace občanů ČSSR starších 15 let,[8] se odhad podílu respondentů pohyboval od 0,827 do 0,882, průměrná hodnota odhadů je $\overline{p} = 0{,}8548$, příslušná střední čtvercová odchylka $s^2 = 0{,}000371$ a směrodatná odchylka $s = 0{,}0193$. Horní meze pro rozptyl odhadů $\hat{p}$, resp. $\hat{p}_0$ v modelu B jsou podle (3) a (4) při $r = 2000$ a $p = 0{,}8548$

$$h_1 = 0{,}0001095 < s^2$$

$$h_2 = 0{,}0000538 < s^2$$

a rozptyl normálního rozdělení, k němuž rozdělení odhadu $\hat{p}$ v modelu s pevným počtem responzí konverguje, je

$$\sigma^2 = 0{,}0000531 < s^2 .$$

Tato zjištění nasvědčují tomu, že platí hypotéza o závislosti parametru p na tématu výzkumu. Vzhledem k nedostatku empirického materiálu jsme tuto hypotézu nemohli prověřit.

Údaje o podílu respondentů a odmítajících jsou uváděny v literatuře vzácně. Moser a Stuart [1953] hovoří o podílu odmítnutí při šetřeních v Británii, který byl při kvótním výběru 13,5 % a při náhodném výběru 6,6 %. Gallup [1948] uváděl, že podíl respondentů v jeho výzkumech ve čtyřicátých letech činil asi 90 % a podobné číslo vyslovil v souvislosti s výzkumem veřejného mínění v letech 1947—50 v Čechách také Č. Adamec.[9] V obou případech se jednalo o pravděpodobnostní výběry. Podíl respondentů při výzkumech ÚVVM je tedy přibližně stejný jako při obdobném výběrovém postupu v Británii a nižší, než se uvádí u náhodných výběrů.

ÚVVM provedl v roce 1979 výzkum názorů obyvatel Prahy, při němž bylo experimentálně použito prostého náhodného výběru s pevným počtem responzí ($r = 460$). Na 459 skutečně kontaktovaných respondentů v něm připadlo 59 odmítnutí, tedy

(8) Populace občanů ČSSR starších 15 let je nejčastěji zkoumaným subjektem veřejného mínění.

(9) Č. Adamec se takto vyslovil při své přednášce v ÚVVM v Praze dne 17. 6. 1981.

$\hat{p} = 0{,}8861$, což se dosti blíží Gallupovým a Adamcovým údajům. Rozptyl odhadu $\hat{p}$ je v modelu B za předpokladu $p = 0{,}9$ asymptoticky roven 0,000176 a směrodatná odchylka 0,0133. Zjištěná hodnota $\hat{p}$ při prostém náhodném výběru v Praze by tedy tedy nasvědčovala tomu, že hypotéza $p = 0{,}9$ platí (na 95% hladině významnosti). Při kvótních výběrech byla průměrná zjištěná hodnota odhadu $\hat{p}$ pro populaci Prahy rovna 0,8195, což je podstatně méně než při zmíněném kvótním výběru (směrodatná odchylka činí pro Prahu 0,0233). Uvedené skutečnosti nás nutí pochybovat o hypotéze ekvivalentnosti kvótního výběru ÚVVM prostému náhodnému výběru.

Prostého odhadu jsme použili pro porovnání empirických poznatků z výzkumů ÚVVM s dalším údajem uváděným v literatuře. Gray a Corlett [1950] píší, že odmítnutí rozhovoru jsou častější u žen než u mužů a že podíl respondentů klesá s věkem. Hodnoty odhadů $\hat{p}$ pro tyto demografické skupiny zjištěné ve výše uvedených výzkumech ÚVVM shrnujeme v tabulce.

Tabulka 2. Průměrná hodnota a variabilita odhadu podílu respondentů v závislosti na pohlaví a věku

Skupina	$\hat{p}$	s^2	s
muži	0,8437	0,000126	0,0112
ženy	0,8592	0,000515	0,0227
15—29 let	0,8711	0,000988	0,0314
30—44 let	0,8283	0,000400	0,0200
45—59 let	0,8367	0,000389	0,0184
60 a více let	0,8559	0,000962	0,0310

Z tabulky vyplývá, že při výběrovém postupu ÚVVM a při jeho současné tazatelské síti se Grayovo a Corlettovo tvrzení nepotvrzuje. Je však třeba si uvědomit, že odmítnutí v sociologickém výzkumu není závislé jen na apriorním postoji individua k šetření, ale i na aktuální situaci rozhovoru. Proto nelze uvedené zjištění automaticky přenášet na výzkumy ostatních sociologických pracovišť.

Závěr

Běžnou součástí zpráv z výběrových šetření by podle našeho názoru měl být údaj o tom, nakolik data reprezentují cílovou populaci, aby bylo možno učinit si představu o možné nevýběrové chybě výsledků. V teoretické části článku jsme ukázali, že při nejběžnějších typech výběrových postupů — výběru v modelu s pevným počtem pokusů nebo s pevným počtem responzí — je ze statististického hlediska opodstatněné užívání prostého odhadu podílu respondentů jako jednoho z ukazatelů reprezentativity výběrového souboru. To umožňuje standardizovat informaci o podílu respondentů při různých typech výzkumů. Statistické vlastnosti prostého odhadu při složitějších výběrových postupech je však ještě třeba zkoumat.

V článku jsme také ukázali, že podíl respondentů ve výzkumech ÚVVM odpovídá údajům uváděným v literatuře pro kvótní výběry, ne však pro výběry náhodné. To spolu s dalšími empirickými údaji odporuje dlouho tradované hypotéze, že výběr ÚVVM se chová jako prostý náhodný. Kromě toho uvádíme materiál protiřečící hypotéze o menším podílu respondentů ve skupině žen a o poklesu podílu respondentů v závislosti na věku. Obě hypotézy by však měly být dále zkoumány, neboť empirický materiál, který nám byl k dispozici, zdaleka nepostačoval k jejich potvrzení či

vyvrácení. Malý objem empirického materiálu také způsobil, že jsme museli upustit od zkoumání závislosti podílu respondentů na tématu výzkumu, na roční době terénního sběru dat a na dalších okolnostech, které podle našeho názoru charakteristiku p ovlivňují. Proto svůj článek chápeme spíše jako vstup do problematiky, které bude z teoretického i empirického hlediska nutné věnovat ještě značnou pozornost.

Dodatek[10]

Při náhodném výběru (s vracením) lze na kontakty s výběrovými jednotkami pohlížet jako na posloupnost bernoulliovských pokusů, v nichž responze představuje zdar a odmítnutí nezdar. Pravděpodobnost zdaru je přitom rovna podílu respondentů v základním souboru p.

ad *B. Model s pevným počtem responzí (inverzní výběr)*

Označme r_i pevně stanovený počet zdarů v i-té sérii a O_i počet nezdarů v ní, $i = 1, 2, \ldots, s$. $O_1, \ldots, O_s$ tvoří s-tici nezávislých náhodných veličin se záporně binomickým rozdělením s parametry $(r_1, p), \ldots, (r_s, p)$. Označme

$$O = \sum_{i=1}^{s} O_i$$

$$r = \sum_{i=1}^{s} r_i ;$$

O má záporně binomické rozdělení s parametry (r, p).

Lemma 1: Pro $r \geq 2$ a $m \geq -1$ zaveďme funkci celočíselné proměnné m

$$g(m) = \mathrm{E}\,\frac{1}{O + r + m},$$

pak platí:

(i) $$g(-1) = \frac{p}{r - 1}$$

(ii) $$g(m) = \frac{p}{r + m} + \frac{1 + m}{r + m}(1 - p)\, g(m + 1)$$

a pro $m \geq 0$ navíc

(iii) $$\frac{p}{r + mp} < g(m) < \frac{p}{r - 1 + (m + 1)\, p}.$$

Důkaz. Tvrzení (i) lze snadno dokázat výpočtem příslušné střední hodnoty. Abychom dokázali (ii), upravíme náhodnou veličinu

$$\frac{1}{O + r + m}$$

(10) Podrobněji je látka zpracována v práci J. Kábrta: *Odhad pravděpodobnosti v bernoulliovském pokusu.* (Studentská vědecká práce) Matematicko-fyzikální fakulta UK, Praha 1981, nepublikováno

na tvar

$$\frac{r-1}{(r+m)(O+r-1)} + \frac{(1+m)O}{(r+m)(O+r-1)(O+r+m)}. \tag{8}$$

Platí

$$\mathrm{E}\,\frac{O}{(O+r-1)(O+r+m)} = (1-p)\,g(m+1)\,,$$

$$\mathrm{E}\,\frac{r-1}{(r+m)(O+r-1)} = \frac{r-1}{r+m}\,g(-1) = \frac{p}{r+m}\,,$$

což spolu s (8) tvrzení (ii) dokazuje. Důkaz tvrzení (iii) je triviální.

Věta 1. Prostý odhad $\hat{p}$ parametru p je v modelu s pevným počtem responzí $r \geq 2$ vychýlený a platí

$$p < \mathrm{E}\hat{p} < p\,\frac{r}{r-1+p}. \tag{9}$$

Nevychýleným odhadem je v daném případě

$$\hat{p}_0 = \frac{r-1}{O+r-1}.$$

Důkaz. Nejprve ukážeme, že $\hat{p}_0$ je nevychýlený odhad; s využitím tvrzení (i) lemmatu 1 dostáváme

$$\mathrm{E}\hat{p}_0 = (r-1)\,g(-1) = p\,.$$

Dále máme

$$\mathrm{E}\hat{p} = rg(0)\,,$$

což spolu s tvrzením (iii) lemmatu 1 dokazuje nerovnost (9).

Lemma 2. Pro $r \geq 3$ a $m, n \geq -2$ zaveďme funkci celočíselných proměnných m, n

$$h(m,n) = \mathrm{E}\,\frac{1}{(O+r+m)(O+r+n)}.$$

Pak platí

(i) $$h(-2,n) = \frac{p^2}{(r-2)(r+n)} + \frac{1+n}{r+n}(1-p)\,h(-1,n+1)$$

(ii) $$h(-1,n) = \frac{p^2}{(r-1)(r+n)} + \frac{2+n}{r+n}(1-p)\,h(-1,n+1)$$

(iii) $$h(0,0) < \frac{p^2}{(r-1)(r-2+3p)}$$

(iv) $$h(-1,-1) < \frac{p^2}{(r-1)(r-2+p)}$$

Důkaz je obdobný jako u lemmatu 1.

Věta 2. Pro rozptyly odhadů $\hat{p}$ a $\hat{p}_0$ v modelu inverzního výběru při $r \geq 3$ platí

(i) $$\operatorname{var} \hat{p} < p^2 \frac{3r - 3pr - 2 + 3p}{(r-1)(r-2+3p)}$$

(ii) $$\operatorname{var} \hat{p}_0 < p^2 \frac{1-p}{r-2+p} .$$

Důkaz. Rozptyl odhadu $\hat{p}_0$ rozepíšeme pomocí funkcí $g(m)$ a $h(m, n)$:
$\operatorname{var} \hat{p}_0 = \mathrm{E}(\hat{p}_0 - p)^2 = (r-1)^2 h(-1, -1) - 2(r-1) pg(-1) + p^2$.
Podle lemmat 1 a 2 pak máme

$$\operatorname{var} \hat{p}_0 < \frac{r-1}{r-2+p} p^2 - 2p^2 + p^2 ,$$

odkud přímo plyne (ii). Pro důkaz tvrzení (i) využijeme nerovnosti

$$\operatorname{var} \hat{p} < \mathrm{E}(\hat{p} - p)^2 ;$$

pro střední hodnotu na pravé straně máme

$$\mathrm{E}(\hat{p} - p)^2 = r^2 h(0, 0) - 2rpg(0) + p^2 < r^2 \frac{p^2}{(r-1)(r-2+3p)} - 2p^2 + p^2 ,$$

což dokazuje tvrzení (i).

Věta 3. Prostý odhad $\hat{p}$ konverguje v modelu inverzního výběru k hodnotě p podle pravděpodobnosti a má asymptoticky normální rozdělení $\mathrm{N}\left(p, \frac{p^2(1-p)}{r}\right)$

Důkaz. Věrohodnostní rovnice v daném modelu má tvar

$$\frac{r}{p} - \frac{O}{1-p} = 0 .$$

Odtud plyne, že $\hat{p}$ je maximálně věrohodný odhad parametru p, který je jediným řešením dané věrohodnostní rovnice. Důkaz pak plyne z věty 10 uvedené na s. 268 práce [Anděl 1978] a z vlastností geometrického rozdělení s parametrem $p \in (0, 1)$.

ad *C. Model se stanoveným minimálním počtem resposní i odmítnutí*

Věta 4. V bernoulliovské posloupnosti pokusů definujme náhodnou veličinu Z jako počet pokusů nutných pro dosažení alespoň $r \geq 1$ zdarů a alespoň $o \geq 1$ nezdarů zmenšený o $r + o$, tj. počet „zbytečných" pokusů. Pak Z má zobecněné záporně binomické rozdělení s parametry (r, o, p):

$$\mathrm{P}(Z = k) = p^r (1-p)^o \left[\binom{k + r + o - 1}{k + r} p^k + \binom{k + r + o - 1}{k + o} (1-p)^k \right] .$$

Důkaz. Zaveďme náhodnou veličinu J jako výsledek, na nějž se čekalo, aby platilo $R \geq r$, $O \geq o$. Pak lze psát

$$\mathrm{P}(Z = k) = \mathrm{P}(Z = k, J = [\text{nezdar}]) + \mathrm{P}(Z = k, J = [\text{zdar}]) . \qquad (10)$$

Přitom $P(Z = k, J = [\text{nezdar}])$ je rovno pravděpodobnosti, že před o-tým nezdarem předchází $r + k$ zdarů, tedy

$$P(Z = k, J = [\text{nezdar}]) = \binom{k + r + o - 1}{k + r} p^{r-k} (1 - p)^o .$$

Podobně lze odvodit tvar druhého sčítance v (10). Tím je věta dokázána.

Lemma 3. Označme

$$W = \sum_{i=0}^{\infty} \binom{i + r + o - 1}{i + o} (1 - p)^i ,$$

pak

(i) $$W = \frac{1}{p} \sum_{i=1}^{r-1} \binom{r + o - 2 - i}{r - 1 - i} p^{-i}$$

(ii) $$P(Z = k/J = [\text{zdar}]) = \frac{1}{W} \binom{k + r + o - 1}{k + o} (1 - p)^k .$$

Důkaz. Pravděpodobnost $P(J = [\text{zdar}])$ lze vyjádřit jednak jako součet pravděpodobností $P(Z = k, J = [\text{zdar}])$ přes všechna celá nezáporná k, tedy ve tvaru

$$p^r (1 - p)^o \sum_{k=0}^{\infty} \binom{k + r + o - 1}{k + o} (1 - p)^k ,$$

jednak jako pravděpodobnost, že o nezdarů nastane dříve než r zdarů, tj. že před o-tým nezdarem předchází nejvýše $r - 1$ zdarů. V druhém případě je vzorec pro tuto pravděpodobnost

$$(1 - p)^o \sum_{k=1}^{r-1} \binom{k + o - 1}{k} p^k .$$

Rovnost obou výrazů dává

$$W = \sum_{k=0}^{r-1} \binom{k + o - 1}{k} p^{k-r} = \frac{1}{p} \sum_{k=0}^{r-1} \binom{k + o - 1}{k} \left(\frac{1}{p}\right)^{r-1-k} .$$

Položíme-li v posledním výrazu $k' = r - 1 - k$, je (i) dokázáno. Tvrzení (ii) lze snadno dokázat dosazením explicitních vyjádření pravděpodobností na pravou stranu rovnosti

$$P(Z = k/J = [\text{zdar}]) = \frac{P(Z = k, J = [\text{zdar}])}{P(J = [\text{zdar}])} .$$

Věta 5. Při jedné realizaci série pokusů v modelu se stanoveným minimálním počtem responzí i odmítnutí je střední hodnota odhadu

$$p_0 = \frac{r - 1}{r + o + Z - 1}$$

za předpokladu, že se čeká na r-tý zdar,

$$E p_0 = p \left[1 - \frac{1}{pW} \binom{r + o - 2}{r - 1}\right]$$

a pro vychýlení D tohoto odhadu při $o \geq 1$ platí

$$0 < -D < p\frac{o}{r+o-1}.$$

Důkaz. Střední hodnota $\mathrm{E}\hat{p}_0$ je za předpokladu, že se čeká na r-tý zdar, dána vztahem

$$\mathrm{E}\hat{p}_0 = \sum_{k=0}^{\infty} \frac{r-1}{r+o+k-1} \mathrm{P}(Z = k/J = [\text{zdar}]).$$

Dosadíme-li na pravou stranu výraz z tvrzení (ii) lemmatu 3 a použijeme-li dále rovnost

$$\frac{r-1}{r+o+k-1} = 1 - \frac{k+o}{r+o+k-1},$$

dostáváme s použitím tvrzení (i) téhož lemmatu

$$\mathrm{E}\hat{p}_0 = \frac{1}{W}\left[W - \sum_{k=0}^{\infty} \frac{k+o}{r+o+k-1}\binom{k+r+o-1}{r-1}(1-p)^k\right] =$$

$$= 1 - \frac{1-p}{W}\left[W + \binom{r+o-2}{r-1}\frac{1}{1-p}\right],$$

což dokazuje prvé tvrzení. Z toho hned plyne, že při $o \geq 1$ platí

$$D^{-1} = \frac{(r-1)!\,(o-1)!}{(r+o-2)!} \sum_{i=0}^{\infty} \binom{r+o+i-1}{r-1}(1-p)^i =$$

$$= \sum_{i=0}^{\infty} (1-p)^i \prod_{j=0}^{i} \frac{r+o+j-1}{o+j} >$$

$$> \sum_{i=0}^{\infty} \frac{r+o-1}{o}(1-p)^i = \frac{r+o-1}{o} \cdot \frac{1}{p}.$$

Tím je věta dokázána.

Věta 6. Rozdíl maximálně věrohodného odhadu $\hat{p}_{\mathrm{MV}}$ a prostého odhadu $\hat{p}$ parametru p v modelu se stanoveným minimálním počtem responzí (r) i odmítnutí (o) konverguje bodově k nule při r a o rostoucích do nekonečna.

Důkaz. Seřaďme realizace daného modelu tak, že v 1. až x-té je $J = [\text{nezdar}]$ a v $(x+1)$-ní až $(x+y)$-té $J = [\text{zdar}]$. Věrohodnostní rovnice má v daném případě tvar

$$\sum_{i=1}^{x} \frac{\hat{c}}{\hat{c}p} \ln \frac{\binom{Z_i+r+o-1}{Z_i+r} p^{Z_i}(1-p)^o}{\sum_{k=0}^{o-1}\binom{k+r-1}{k}(1-p)^k} +$$

$$+ \sum_{=x+1}^{x+y} \frac{\hat{c}}{\hat{c}p} \ln \frac{\binom{Z_i+r+o-1}{Z_i+o} p^r(1-p)^{Z_i}}{\sum_{k=0}^{r-1}\binom{k+o-1}{k}p^k} = 0$$

Po provedení derivací lze rovnici upravit na tvar

$$\frac{\sum_{i=0}^{x} Z_i}{p} - \frac{xo}{1-p} + \frac{x}{1-p} \frac{\sum_{k=0}^{o-1} k \binom{k+r-1}{k} p^r (1-p)^k}{\sum_{k=0}^{o-1} \binom{k+r-1}{k} p^r (1-p)^k} +$$

$$+ \frac{yr}{p} - \frac{\sum_{i=x+1}^{x+y} Z_i}{1-p} - \frac{y}{p} \frac{\sum_{k=0}^{r-1} k \binom{k+o-1}{k} (1-p)^o p^k}{\sum_{k=0}^{r-1} \binom{k+o-1}{k} (1-p)^o p^k} = 0$$

Označíme podíly sumací v této rovnici $A(p)$ a $B(p)$. Je patrné, že

$$\lim_{o \to \infty} A(p) = r \frac{1-p}{p}$$

$$\lim_{r \to \infty} B(p) = o \frac{p}{1-p}.$$

Rovnici (11) přepíšeme do tvaru

$$p = \frac{x \frac{p}{1-p} A(p) + yr + \sum_{i=1}^{x} Z_i}{x \frac{p}{1-p} A(p) + xo + yr + y \frac{1-p}{p} B(p) + \sum_{i=1}^{x+y} Z_i}.$$

Pravá strana konverguje při r a o rostoucích do nekonečna k výrazu

$$\frac{(x+y) r + \sum_{i=1}^{x} Z_i}{(x+y)(r+o) + \sum_{i=1}^{x+y} Z_i},$$

což při dané sérii realizací odpovídá prostému odhadu $\hat{p}$. Tím je věta dokázána.

Literatura

Anděl, J.: *Matematická statistika.* Praha, SNTL — Alfa 1978.

Cochran, W.: *Sampling Techniques.* New York, Willey 1967.

Čermák, V.: *Výběrové statistické zjišťování.* Praha, SNTL—Alfa 1980.

Gallup, J.: *Průvodce po výzkumu veřejného mínění.* Praha, Orbis 1948.

Gray, P. G. — Corlett, T.: *Sampling for the Social Survey.* JRSS, Series A, Vol. 113, 1950.

Haldane, J. B. S.: *On a Method of Estimating Frequences.* Biometrika, Vol. 33, 1943—46.

Konijn, H. S.: *Statistical Theory of Sample Survey Design and Analysis.* Amsterdam, North-Holland 1973.

Moser, C. A. — Stuart, A.: *An Experimental Study of Quota Sampling.* JRSS, Series A, Vol. 116, 1953.

Řehák, J.: *K pojmu „reprezentativita" v sociologických výzkumech.* Sociologický časopis 1978, č. 5.

Резюме

Я. Герцманн - Я. Кабрт: К проблематике отказа участвовать в интервью в социологических выборочных обследованиях

Получение эмпирического материала в конкретных социологических исследованиях весьма часто основано на применении методов выборочного обследования. Одной из причин нарушения репрезентативности и возникновения т. н. невыборочных погрешностей является тот факт, что часть контактируемых лиц не желает предоставить устанавливаемые данные и отказывается принять участие в интервью. Такой отказ представляет большую опасность для социологических анкетных обследований, так как можно со всем основанием предполагать, что та часть обследуемой социальной группы, которая умышленно не предоставит требуемые данные, с точки зрения предмета исследования существенно отличается от остальных индивидов.

Статья ставит своей целью показать возможности статистической оценки размера той части населения, которая покрыта выборочным обследованием, т. е. которая потенциально готова предоставить требуемую информацию. Внимание уделяется проблематике оценки на основе самого выборочного обследования, т. е. без использования априорной информации. Такая оценка имеет практическое значение прежде всего для тех подразделений, которые не проводят систематическое исследование одной и той же социальной группы, но она может служить также для уточнения раннее полученной информации о доле потенциальных опрашиваемых, готовых отвечать и тех, кто не пожелают участвовать в обследовании, в основной совокупности.

В статье изучаются свойства простой оценки доли потенциальных опрашиваемых, готовых предоставить необходимую информацию, в социальной группе посредством соотношения количества фактически конктактированных опрашиваемых и общего количества попыток установить контакт. Теоретическая часть статьи разработана для простого случайного выбора с возвратом (для возвратной выборки). Изучаются три модели выборочной процедуры: А-модель с твердо установленным количеством попыток, Б-модель с твердо установленным числом фактически контактированных опрашиваемых (т. н. инверсионная выборка) и В-модель с установленным минимальным количеством контактированных опрашиваемых и отказов. В первых двух моделях простая оценка сопоставляется с другими оценками, полученными на основе применения методов математической статистики, причем удалось доказать, что отличия являются незначительными. В модели В удалось вывести основные свойства лишь для оценки, возникшей вследствие простого вычета единицы в числителе и знаменателе и равной, таким образом, для практики простой случайной выборке. Автор показывает, что эта оценка в модели В является смещенной, однако цифровое исследование свидетельствует о том, что это смещение при не особо ограничивающих предпосылках незначительно.

Кроме теоретической части, в статье имеется раздел о применении простой оценки к выборочной процедуре Института по исследованию общественного мнения в Праге, которая является, правда, квотной, но считают ее эквивалентной простой случайной выборке. Эмпирические данные исследований Института по исследованию общественного мнения сопоставляются как с теоретической моделью, так и с эмпирическими данными, приводимыми в литературе.

Математическая теория, касающаяся более сложных моделей выборки (Б и В), сосредоточена в дополнении и основана на понимании простой случайной возвратной выборки как последовательности попыток Бернулли.

Summary

J. Herzmann — J. Kábrt: One the Problem of Response Refusals in Sociological Surveys

Acquisition of empirical material in concrete sociological surveys is very often based on the methods of sampling surveys. One of the reasons for lower representativeness and the occurence of so called non-sampling errors is the fact that part of the persons interviewed is unwilling to provide appropriate data or to participate. Refusals represent a great danger to sociological questionnaire surveys because it can be safely predicted that the part of the population which deliberately declines to provide required data is — from the viewpoint of the subject of the opinion poll in question — substantially different from other individuals.

The purpose of the article is to point to the possibility of estimating the size of that portion of the population which is covered by sampling surveys, i. e. the part potentially willing to provide required information. Attention is payed to the problems of estimates based on the sampling survey itself, i. e. without the use of a priori information. Such estimates are of great practical significance particularly for institutions which do not carry out systematic opinion surveys but they may serve to specify previously obtained information relating to the share of potential respondents and refusals in the basic population.

The article goes on to examine the properties of a simple estimate of the proportion of potential respondents in the population defined by the ratio of genuinely interviewed persons and the total number of attempts at contact. The theoretical part of the article is designed for simple random sampling with returns. The authors study three models of selective procedure: A-model with a fixed number of attempts, B-model with a fixed number of genuinely contacted respondents (so called inverse sampling) and C-model with a set of minimum number of both contacted respondents and refusals. As far as the first two models are concerned, the simple estimate is compared with other estimates obtained by the methods common in mathematical statistics. The authors managed to prove that the difference is negligible. As regards the C model the authors derived basic properties only for the estimate which originates from a simple substraction of one both in numerator and denominator and thus is approximately equal to simple estimate as far as practice goes. The article shows that in model C this estimate is biased but the subsequent numerical study demonstrates that the bias is negligible under certain not so limiting preconditions.

In addition to the theoretical part, the article contains a section dealing with the application of simple estimate to sampling surveys as practised by the Public Opinion Research Institute in Prague which are quota-samplings but generally regarded as being equivalent to simple random sampling. Empirical findings obtained by the Institute are compared with the theoretical model and empirical data given in literature.

The Appendix includes mathematical theories concerning more complex models of selection (B and C) and is based on the interpretation of the simple random sampling with returns as a Bernoulli sequence of attempts.