Přehled koncepcí a stanovisek k problematice selhání výběrových jednotek

II. Statistické aspekty

JAN HERZMANN
Ústav pro výzkum veřejného mínění při FSÚ, Praha

V Sociologickém časopise č. 3/ 1984 byl podán přehled sociologicky a sociálně psychologicky zaměřené literatury pojednávající o selhání výběru, tedy o případu, kdy se nepodaří určitou osobu (obecně výběrovou jednotku) kontaktovat, nebo kdy tato osoba odmítne účast v prováděném sociologickém šetření, viz Herzmann [1984].

Příčiny selhání výběru jsou „společenské", „nestatistické" a také důsledky tohoto jevu jsou „společenské", neboť selhání může nepříznivě ovlivnit interpretační fázi celého výzkumu a způsobit, že do společenské praxe vstoupí poznatky do určité míry zkreslující skutečný stav věcí. Výběry a výběrová šetření jsou však povýtce záležitostí statistickou a statistické metody jsou vlastně také cestičkou, po níž se vliv selhání do výsledků šetření vkrádá. Není proto divu, že právě statistikové vyvinuli nemalé úsilí o nalezení, odstranění nebo alespoň zmenšení nepříznivého vlivu selhání.

Většina publikací, které budou dále citovány, pochází od amerických autorů. Důvod je objektivní: ve Spojených státech mají výběrová šetření lidských populací největší tradici i frekvenci a jejich teorií se tam zabývá mnoho odborníků. V socialistických státech se mimořádně zřídka odhalují a analyzují systematické chyby vznikající v procesu výběru. Je nepochybné, že teoretická a metodologická východiska buržoazních vědců, o jejichž práce se přehled opírá, jsou pro nás nepřijatelná, nicméně je nezbytné pečlivě a kriticky studovat současný stav konkrétní metodiky a techniky výzkumů v kapitalistických státech, a především v USA. Při sestavování tohoto přehledu jsem se snažil k takovému kritickému studiu prací buržoazních autorů alespoň malým dílem přispět. Abych zachoval „historický" ráz přehledu, ponechal jsem, pokud to bylo možné, původní značení jednotlivých citovaných prací.

Selhání jako zdroj vychýlení

Již v souvislosti se sociologickými aspekty selhání výběrových jednotek jsem se zmínil o tom, že toto selhání je jedním ze zdrojů nevýběrových chyb šetření. Vychýlení, k němuž vede, je důsledkem odlišnosti prošetřených a neprošetřených výběrových jednotek, o níž se zmiňují například Grušin [1972] nebo Linehan [1973]. Za klasický přístup k tomuto problému je dnes považován tzv. Cochranův model dvou strat, popsaný poprvé Birnbaumem a Sirkenem [1950a] a později Cochranem [1963] (první vydání Cochranovy knihy vyšlo v roce 1953). Tento model předpokládá, že v základním souboru existují dvě strata:

stratum I — „respondenti", tj. osoby, od nichž v případě, že budou vybrány, budou získány požadované údaje,

stratum II — „selhávající", tj. osoby, od nichž v případě vybrání požadované údaje získány nebudou.

Označíme-li

n počet výběrových jednotek,

n_1, n_2 počet výběrových jednotek náležejících po řadě do strat I a II,

$\bar{y}$ (výběrový) průměr hodnot sledované proměnné za všech n výběrových jednotek,

$\bar{y}_1$, $\bar{y}_2$ (výběrové) průměry hodnot sledované proměnné po řadě za n_1 jednotek ze strata I a za n_2 jednotek ze strata II,

pak pro provedení výběrového šetření máme k dispozici místo „požadované" hodnoty $\bar{y}$ pouze hodnotu $\bar{y}_1$. Tím ovšem dochází k vychýlení

$$(1) \qquad \bar{y} - \bar{y}_1 = \frac{n_2}{n} (\bar{y}_2 - \bar{y}_1) .$$

Je patrné, že vychýlení závisí jednak na relativní četnosti jednotek ze strata II ve výběru, tedy na podílu selhání, jednak na rozdílnosti obou strat z hlediska sledované proměnné. Šljapentoch [1976] doporučuje používat jako charakteristiky vychýlení, místo (1) tzv. *relativního zkreslení*

$$(2) \qquad \frac{\bar{y} - \bar{y}_1}{\bar{y}} = \frac{n_2}{n} \frac{\bar{y}_2 - \bar{y}_1}{\bar{y}} ,$$

které lze pro nezáporné náhodné veličiny považovat za lepší ukazatel, neboť vyjadřuje míru zkreslení v jednotkách samotné měřené veličiny. Stejný názor zastávají Platek, Singh a Stremblay [1978].

Podle Šljapentocha byl prvním, kdo se zabýval vychýlením způsobeným selháním, Yulm. Ten v roce 1911 vymezil interval, v němž při daném procentu selhání leží skutečná hodnota. Takový výpočet je možný, víme-li, že hodnoty sledované proměnné se pohybují v nějakých mezích y_d až y_h. Pak totiž podle (1) leží vychýlení v intervalu

$$(3) \qquad \left[\frac{n_2}{n} (y_d - \bar{y}_1) , \frac{n_2}{n} (y_h - \bar{y}_1) \right]$$

a platí tedy

$$(4) \qquad \bar{y}_1 + \frac{n_2}{n} (y_d - \bar{y}_1) \leq \bar{y} \leq \bar{y}_1 + \frac{n_2}{n} (y_h - \bar{y}_1) ,$$

neboli

$$(5) \qquad \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} y_d \leq \bar{y} \leq \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} y_h .$$

Tento vztah je užitečný zejména při odhadování proporce P jednotek s určitou vlastností v základním souboru. V takovém případě se obvykle píše p, p_1, p_2 místo $\bar{y}$, $\bar{y}_1$, $\bar{y}_2$, a protože všechny proporce leží v intervalu $\langle 0, 1 \rangle$, dostaneme

$$(6) \qquad \frac{n_1}{n} p_1 \leq p \leq \frac{n_1}{n} p_1 + \frac{n_2}{n} .$$

Vztahy (1) a (2) dobře charakterizují nevýběrovou chybu vzniklou selháním, řada autorů si však uvědomila, že tuto chybu je nutno vyhodnocovat ne izolovaně, ale společně s tzv. výběrovou chybou, vznikající v důsledku prošetření pouhé části (různorodého) základního souboru. Označíme-li $\overline{Y}$ odhadovaný průměr sledované veličiny v základním souboru, je výběrová chyba charakterizována v daném případě rozdílem $\overline{Y} - \bar{y}$.

První mně známou prací, v níž jsou výběrová a nevýběrová chyba dány do souvislosti, je stať Birnbauma a Sirkena [1950a]. Ti uvádějí explicitní vyjádření celkové chyby odhadu proporce P_y ve tvaru

(7) $$H = S + b\,,$$

kde

(8) $$b = P_{t,y} - P_y = P_n(P_{t,y} - P_{n,y})$$

je vychýlení způsobené selháním a

(9) $$S = N_{t,y}/N_t - P_{t,y}$$

je výběrová chyba estimátoru $N_{t,y}/N_t$, přičemž značení je následující:

P_y podíl jedinců se sledovanou vlastností v základním souboru,
P_n podíl strata „selhávajících" v základním souboru,
$P_{t,y}$, $P_{n,y}$ podíl jedinců se sledovanou vlastností po řadě ve stratu „respondentů" a ve stratu „selhávajících",
N_t počet skutečně prošetřených respondentů,
$N_{t,y}$ počet osob se sledovanou vlastností mezi prošetřenými.

Uvedený přístup lze zobecnit na případ odhadu libovolné náhodné veličiny s využitím cochranovské symboliky. Celková chyba $H = \overline{Y} - \bar{y}_1$ se zde rozkládá na výběrovou chybu $S = \overline{Y} - \bar{y}$ a vychýlení $b = \bar{y} - \bar{y}_1$. (V současné době se pojmem „vychýlení" označuje obvykle střední hodnota náhodné veličiny b.)

Cochran sám však šel do ještě vyšší úrovně obecnosti a definoval celkovou chybu libovolného estimátoru y^*, kterým má být odhadnuta hodnota $\overline{Y}$, jako tzv. střední čtvercovou chybu,

(10) $$MSE = E(y^* - \overline{Y})^2\,,$$

v níž střední hodnota je brána přes všechny možné výběry daného rozsahu. Kdyby tedy byla za estimátor y^* považována statistika $\bar{y}_1$, byl by základem pro výpočet střední čtvercové chyby rozdíl $\bar{y}_1 - \overline{Y}$, tedy celková chyba H ve smyslu Birnbaumově a Sirkenově. Rozkladem dostaneme

(11) $$MSE = ES^2 + Eb^2 + 2E(Sb)\,.$$

Zatímco Birnbaumova a Sirkenova formule přímo nabízí možnost vzájemného rušení výběrové chyby S a vychýlení b, ve vztahu (11) je tato možnost skryta v posledním členu: výběrová chyba a vychýlení se „v průměru" vzájemně kompenzují, je-li jejich korelace záporná.

Ke vztahu (11) je nutno poznamenat, že je-li y^* nevychýleným estimátorem $\overline{Y}$, pak

(12) $$ES^2 = \operatorname{var} y^*\,.$$

Tato skutečnost má svůj význam, porovnáme-li uvedené přístupy s Demingovou [1953] prací. Tento autor v podstatě odmítá možnost vzájemné kompenzace obou složek celkové chyby. K vysvětlení jeho ideje doplníme používanou symboliku o $\overline{Y}_1$, což je průměr hodnot sledované proměnné ve stratu I. Za vychýlení považuje Deming hodnotu

(13) $$B = Ey^* - \overline{Y}_1$$

a za celkovou chybu

(14) $$MSE = \operatorname{var} y^* + B^2\,.$$

Je patrné, že tento autor existenci strata II zanedbává, což ve své práci ostatně explicitně říká. To je jistě závažný nedostatek, který však Demingovu metodu měření chyby diskvalifikuje jen potud, pokud setrváme v klasickém modelu dvou strat. Jakmile si však uvědomíme, že selhání nemusí být deterministický jev (srovnej

Nargundkar—Joshi [1975]), že u určité osoby může vést část pokusů o získání údajů k responzi a část k selhání, nabývá Demingova metoda odhadování, o níž budu hovořit později, a s ní i Demingova charakteristika celkové chyby (14) na smysluplnosti. Jak již však bylo řečeno, chybí v ní postižení té složky celkové chyby, která je způsobena absencí jednotek s „nulovou pravděpodobností poskytnutí responze", a sporné je vyloučení možnosti vzájemné kompenzace jednotlivých složek celkové chyby.

Celkovou chybou estimátorů se zabývá také Kish [1965]. Obecně tuto chybu vyjádřil ve tvaru

$$\sqrt{VE^2 + B^2}\,, \tag{15}$$

kde VE je „chyba proměnné" (variable error) složená (opět „eukleidovsky") z výběrové variability a „nevýběrové chyby" chápané jako „nepřesnosti" v terénní práci a zpracování dat,

B je vychýlení, do něhož jsou zahrnuty také chyby vzniklé selháním.

V subkapitole věnované vlivu vychýlení na pravděpodobnostní tvrzení pak Kish říká, že intervaly spolehlivosti by v praxi neměly být určovány pouze v závislosti na výběrové variabilitě (tedy úměrně hodnotě σ^2/n, kde σ^2 je rozptyl sledované proměnné), ale i s přihlédnutím k vychýlení B, tedy úměrně celkové chybě $\sqrt{\sigma^2/n + B^2}$. (V tomto případě Kish ztotožňuje „chybu proměnné" s výběrovou variabilitou.) Současně však autor upozorňuje, že pečlivý odhad velikosti vychýlení může vést v konkrétním výzkumu k závěru, že vliv vychýlení na estimátor je zanedbatelný.

Pokračovatelem Kishova úsilí o vyjádření chyby vnesené do výsledků šetření prostřednictvím intervalů spolehlivosti je Rubin [1977]. Tento autor doporučuje využít ke konstrukci intervalů spolehlivosti kromě hodnot zjištěných samotným šetřením také apriorní informaci, kterou obvykle výzkumné pracoviště disponuje. Z hlediska statistické teorie vede toto úsilí k bayesovským přístupům. Rubin je konkrétně rozpracovává pro regresní model

$$\begin{cases} E(\bar{y}_R) &= \alpha_R + \beta_R \bar{X}'_R \\ E(\bar{y}_{NR}) &= \alpha_{NR} + \beta_{NR} \bar{X}'_{NR}\,. \end{cases} \tag{16}$$

kde $\bar{y}_R$, $\bar{y}_{NR}$ jsou výběrové průměry sledované proměnné po řadě ve stratu respondentů a nonrespondentů,

$\bar{X}_R$, $\bar{X}_{NR}$ jsou (řádkové) vektory výběrových průměrů „vysvětlujících" proměnných po řadě ve stratu respondentů a nonrespondentů,

α_R, α_{NR}, β_R, β_{NR} jsou příslušné regresní koeficienty, resp. jejich (rádkové) vektory.

Rubin předpokládá, že $\bar{X}_R$, $\bar{X}_{NR}$ jsou známá. Z výběru lze odhadnout $\bar{y}_R$ a tím i parametry α_R a β_R. Kdyby byly známy parametry α_{NR} a β_{NR}, bylo by možné odhadnout i $\bar{y}_{NR}$ a tím i celkový výběrový průměr $\bar{y}$. Apriorní, „subjektivní" znalosti o daném problému lze vyjádřit prostřednictvím určitých koeficientů, například

Θ_1 parametr vyjadřující míru podobnosti „sklonu" regresních funkcí,

Θ_2 parametr vyjadřující předpokládanou míru odlišnosti výběrového průměru sledované proměnné u respondentů a u nonrespondentů.

Meze intervalu spolehlivosti pro výběrový průměr $\bar{y}$ jsou pak v případě normálního rozdělení sledovaných proměnných dány výrazem

$$\bar{y}_R[1 + h_0 \pm z(\Theta_1^2 h_1^2 + \Theta_2^2 h_2^2 + h_3^2)^{1/2}]\,, \tag{17}$$

kde h_0, h_1, h_2, a h_3 jsou parametry odhadnuté z výběru.

Rubin kromě tohoto jednoduchého případu analyzuje i obecnou situaci, kdy hodnoty proměnných v jednotlivých stratech jsou vázány vektorovým parametrem Φ se známým (předpokládaným) apriorním rozdělením.

Měření vychýlení způsobeného selháním, popřípadě měření celkové chyby estimátoru, je ze statistického hlediska základem pro odstranění nebo alespoň zmenšení nepříznivého vlivu tohoto jevu na výsledky šetření. Postupy, které zde byly uvedeny, je možné obměňovat nebo modifikovat. Rozhodně však platí, že právě tak, jako by měla závěrečná zpráva výzkumu obsahovat informaci o podílu selhání, měla by obsahovat také alespoň orientační údaje o tom, jaké zkreslení vneslo selhání do použitých estimátorů. Nejvhodnější jsou k tomuto účelu pravděpodobně formule (5), (6), popřípadě (17), zatímco význam ostatních vzorců je spíše teoretický (srovnej Biggeri [1975]).

Úpravy vlivu selhání založené na klasické statistice

Prvním teoreticky podloženým přístupem k řešení problému selhání, přesněji řečeno problému selhání pokusů o kontakt, je tzv. Politz-Simmonsovo schéma, které v literatuře poprvé uvedl Hartley [1946] a později Politz a Simmons [1949], [1950]. Tuto metodu přehledně popsal také Cochran [1963]. Její hlavní ideou je snaha přisoudit velkou váhu údajům osob reprezentujících „těžko dostupné" skupiny a malou naopak jedincům snadno kontaktovatelným. Tím se má odstranit vliv toho, že v souboru skutečně kontaktovaných osob jsou „těžko dostupné" osoby zastoupeny nedostatečně a „lehko dostupné" nadměrně. Statistickým vyjádřením takové ideje je vážení odpovědi jednotlivců vahami úměrnými převrácené hodnotě pravděpodobnosti kontaktování, které jsou při vlastním výzkumu odhadovány prostřednictvím údaje o tom, kolikrát byl dotazovaný v uplynulých dnech v určitou dobu na místě dotazování. Deming [1953] tuto metodu popsal s použitím estimátoru

$$x(P) = \frac{\sum_{t=0}^{5} w_t R_t x_t}{\sum_{t=0}^{5} w_t R_t}, \tag{18}$$

kde R_t je počet responzí získaných od osob, které byly v uplynulých pěti dnech t-krát doma, $t = 0, 1, \ldots, 5$,

x_t je výběrový průměr sledované proměnné v této skupině osob,

$$w_t = 6/(1 + t)\,. \tag{19}$$

Ve stejné podobě převzal Politz-Simmonsovo schéma do své učebnice Cochran [1963].

Durbin a Stuart [1954] Politz-Simmonsovu metodu poněkud upravili tím, že zkoumané období rozšířili na 7 uplynulých dnů, takže do něj byl vždy zahrnut i víkend. Navíc Durbin [1954] a Simmons [1954] tvrdí, že není nutné provádět pokusy o kontakt v náhodně zvolených okamžicích ani zjišťovat přítomnost v náhodně zvolených okamžicích uplynulých dnů, jak původně Politz se Simmonsem doporučovali. Za předpokladu, že dotazování probíhá večer v bytě dotázaného[1]), postačuje podle Durbina a Simmonse, aby každý večer po celé období sběru dat byl prováděn stejný počet rozhovorů.

Jedním z hlavních nedostatků Politz-Simmonsova schématu je, že odmítající jsou ztotožněni s nekontaktovanými jedinci a stratum „0" je chápáno jako homogenní

(1) Většinou probíhá dotazování v bydlišti respondentů, místo dotazování však není pro danou metodu podstatné, pokud je pro všechny osoby stejné.

skupina neprošetřených. V rovině statistické teorie na tuto skutečnost upozorňují Greenlees, Reece a Zieschang [1982], kteří správně zařadili Politz-Simmonsovo schéma do skupiny metod založených na post-stratifikaci, při níž jsou chybějící údaje nahrazovány průměrem zjištěným v daném stratu. Na další nebezpečí aplikace této metody upozorňuje Kish [1965]: v určitých případech může být nekontaktování korelováno s předmětem výzkumu. V takovém případě pochopitelně vnáší Politz-Simmonsovo schéma do výsledků šetření nové specifické zkreslení.

Proti Politzově a Simmonsově metodě stojí postup založený na zdokonalení Leuvenem [1932] zavedené metody opakování pokusů: Hansen a Hurwitz [1946] navrhli vybrat z jednotek nepřošetřených při prvním pokusu (jejich podíl značili p a předpokládali o něm, že je apriorně znám) podvýběr rozsahu n_2, který navrhují prošetřit účinnějšími postupy. Pro odhad populačního průměru použili estimátoru

$$\bar{y}' = (1 - p)\,\bar{y}_1 + p\bar{y}_2\,, \tag{20}$$

kde $\bar{y}_1$ je výběrový průměr zjišťované proměnné u jednotek prošetřených v prvním pokusu,

$\bar{y}_2$ je výběrový průměr této proměnné u jednotek prošetřených ve druhém pokusu.[2])

Přitom se předpokládá, že počet neúspěšných druhých pokusů je zanedbatelný. Tento předpoklad je však, jak ukazují zkušenosti, neoprávněný. U poštovní ankety je dokonce znám případ, viz Cochran [1963], kdy po třech pokusech zůstalo 59 % vybraných osob nekontaktováno.[3]) Navíc Dalenius [1955] poukazuje na to, že Hansenův a Hurwitzův postup je náročný na čas, neboť po první „vlně" pokusů o získání údajů je nutné sestavit seznam neprošetřených jednotek, vybrat z něj soubor o rozsahu n_2 a poté provést druhou „vlnu" dotazování. Je ovšem třeba poznamenat, že uvedená metoda vznikla jako prostředek zlepšení informací poskytovaných poštovní anketou, kdy bývá návratnost (podíl prošetřených jednotek) obvykle dosti nízká a může být účelné pokusit se od části neprošetřenýcd osob získat údaje metodou interview, která má daleko větší procento úspěšných pokusů.[4]) Možnost využití jiné metody sběru dat předurčuje Hansenův a Hurwitzův postup k redukci podílu odmítnutí, nikoli jen k redukci vlivu selhání pokusů o kontakt. To je nespornou předností tohoto postupu oproti Politz-Simmonsově schématu.

Za přínos Hansenovy a Hurwitzovy práce z hlediska statistického lze v každém případě označit odvození optimálního výběrového podílu v „druhé vlně" pokusů při daných nákladech. Uvedenou metodu dále zobecnil na postup s více opakováním pokusů El-Badry [1956]. Další modifikaci metody podvýběru v druhé a dalších vlnách navrhl Srinath [1971], jehož snahou bylo odstranit závislost rozptylu hansen-hurwitzovského estimátoru na (předem neznámém) podílu selhání.

Hansen-Hurwitzova metoda byla v minulosti srovnávána s Politz-Simmonsovým schématem nejen co do nákladů časových i finančních, ale i co do výsledného efektu. Durbin a Stuart [1954] zjistili, že v případech, které prozkoumali, odpovídaly výsledky získané pomoci Politz-Simmonsova vážení spíše údajům z prvního pokusu než výsledkům založeným na kombinaci údajů z více pokusů.

---

(2) Statisticky je Hansenova a Hurwitzova metoda založena na teorii dvojitého výběru, kterou popsal Neyman [1938].

(3) Různými metodami snižování podílu selhání při druhém a dalších pokusech se zabývali např. Durbin a Stuart [1954], Sudman [1966] a další. Otázka snižování podílu selhání viz také Herzmann [1984].

(4) S Cochranovým poznatkem lze srovnat práce Dillmana [1972] nebo Blumberga, Fullera a Harea [1974]. Za výjimku lze označit případ popsaný Wisemanem [1972], který dosáhl u poštovní ankety s předchozím upozorněním 67 % úspěšných pokusů, zatímco u osobního interview pouze 60 %.

Z Cochranova [1963] srovnání vyplývá, že pokud

1. tazatelé nezvyšují pravděpodobnost kontaktování ve druhém a dalších pokusech využitím doplňkových informací o výběrových jednotkách,
2. výpovědi o přítomnosti nejsou zatíženy velkou výběrovou chybou,

je metoda s t pokusy o kontakt co do střední čtvercové chyby odhadu (10) rovnocenná Politz-Simmonsově schématu s otázkou na přítomnost v t uplynulých dnech. Protože však v praxi uvedené předpoklady (zejména první) nebývají splněny, je všude tam, kde to výběrová procedura, finanční prostředky a časový rozvrh dovolují, výhodnější používat metodu opakovaných pokusů. K tomuto doporučení je možno ještě připojit, že výhodou metody opakovaných pokusů je již zmíněná možnost získat responzi i od osob v dřívějších pokusech odmítnuvších.

Simmons [1954], pravděpodobně na základě provedených srovnání, navrhl kombinovat metodu vážení a opakování pokusů o kontakt.

„Statistické" kritice podrobili Hansenovu a Hurwitzovu metodu Birnbaum a Sirken [1950a], kteří postřehli, že tato metoda nepostihuje obě složky celkové chyby šetření — výběrovou i nevýběrovou. Proto tito autoři navrhli pojem celkové chyby šetření, který jsem objasnil ve vzorcích (7), (8) a (9) a v následujícím textu. Navíc odvodili rozptyl estimátoru $U = N_{i,y}/N_i$ za předpokladu výskytu selhání

$$\sigma^2(U) = \frac{P_{i,y}(1 - P_{i,y})}{(N + 1) P_i} \tag{21}$$

a zjistili, že tento rozptyl je největší, právě když maximální možné vychýlení

$$B = \max \{P_n(1 - P_{i'y}), P_n P_{i,y}\} \tag{22}$$

je nejmenší, totiž při $P_{i,y} = 1/2$. Blíží-li se $P_{i,y}$ k nule nebo k jedné, rozptyl (21) klesá a maximální možné vychýlení (22) roste. Za předpokladu, že P_n je známé, odvodili Birnbaum a Sirken potřebný rozsah prostého náhodného výběru, který zabezpečuje pro $\delta > P_n$ splnění nerovnosti

$$P[|U - P_y| > \delta] \leq \alpha\,. \tag{23}$$

Potřebný rozsah výběru $n(\delta, \alpha, P_n)$ musí splňovat nerovnost

$$n(\delta, \alpha, P_n) \geq \frac{t_\alpha^2}{4\delta(1 - P_n)(\delta - P_n) - 1}\,, \tag{24}$$

kde t_α je $(1 - \alpha/2)$-kvantil normovaného normálního rozdělení. Ve své druhé práci zabývající se danou metodou Birnbaum a Sirken [1950b] uvádějí tabulky $n(\delta, \alpha, P_n)$ pro různé hodnoty parametrů.

Nevýhodou navržené metody je, že všechny údaje, které se k určení rozsahu výběru používají, mají apriorní charakter, což při zvyšování požadované přesnosti (snižování δ) vede i při nevelkém P_n k prudkému růstu rozsahu výběru (viz Cochran [1963]). O vychýlení B se musí apriorně předpokládat, že je rovno P_n. Taková hodnota B pochopitelně často dominuje nad výběrovou chybou.

Kromě základního odvození pro výběr s jedním pokusem obsahuje citovaný článek rovněž odvození optimálního rozsahu výběrů v k po sobě jdoucích pokusech o kontakt při pevných celkových nákladech c a při známém podílu selhání v jednotlivých pokusech. Pro konkrétní výpočty však autoři neposkytli uspokojující algoritmus, určení optimálního počtu pokusů se děje prakticky zkusmo. Navíc má Birnbaumova a Sirkenova „nákladová" optimalizace ještě jednu slabinu, již si povšimli Stephan a McCarthy [1958]: v případě šetření prostřednictvím tazatelské sítě nejsou v praxi obvykle náklady na jednotlivá opakování pokusů o kontakt známy. Tato výhrada se

pochopitelně vztahuje také ke zmíněnému El-Badryho zobecnění Hansen-Hurwitzovy metody.

Počátkem padesátých let publikoval Deming [1953] stať, která je do jisté míry syntézou přístupu hansen-hurwitzovského a politz-simmonsovského v tom smyslu, že využívá jak opakování pokusů, tak pravděpodobnosti konstatování. Deming navrhuje roztřídit výběrové jednotky do tříd 0, 1, 2, 4, 6, 8 podle toho, kolik z u nich provedených pokusů by skončilo úspěchem — responzí. Hodnoty $\pi_i = 0, 1/8, 2/8, 4/8, 6/8, 1$ jsou vlastně aproximací pravděpodobností získání responze a Deming tvrdí, že zavedení většího počtu tříd (tj. jemnější aproximace) nepřináší významný zisk v přesnosti metody.

Svoji ideu autor přibližuje tak, že si každou výběrovou jednotku představí jako složenou z osmi buněk, z nichž pro jednotku ze třídy $i = 1, 2, 4, 6, 8$ je právě i typu R (responze) a zbylé typu NR (selhání, v originále „nonresponse"). Třídu „0" podle Deminga tvoří osoby dlouhodobě nepřítomné, trvale odmítající, opilé, když jsou konečně nalezeny, nebo neschopné rozumných odpovědí. Podíl této třídy na výběrovém souboru je prý pod 5 % a o většině takových osob lze podle Deminga získat údaje od sousedů, takže „skutečný podíl této třídy bude bezpochyby pod 1 %". S touto optimistickou tezí však, jak ukazují poznatky řady autorů i data citovaná samotným Demingem, nelze souhlasit. Kromě toho, například při výzkumu názorových struktur a vědomí, je možnost získání údajů „od sousedů nebo jiných osob" prakticky vyloučena.[5])

Pomiňme nepodloženost Demingových předpokladů a vraťme se k metodě samé. Ta předpokládá jakoby dvoustupňový výběr: v prvém stupni je vybrána jednotka a ve druhém v ní buňka. Pravděpodobnost responze při prostém náhodném výběru v obou krocích je pak $\pi_i p_i$, kde p_i je podíl třídy i na celé populaci. Dále Deming označil

a_i průměr sledované proměnné ve třídě i,
R_i počet responzí ve třídě i skutečně získaný na první pokus,
x_i výběrový průměr těchto R_i responzí

a zavedl estimátor

$$x(I) = \frac{\sum_{i=1}^{8} R_i x_i}{\sum_{i=1}^{8} R_i} \,. \tag{25}$$

Od estimátoru (18) používaného při Politz-Simmonsově schématu se liší vypuštěním vah strat.

Za vychýlení při prvním pokusu Deming označuje hodnotu

$$B(I) = E(x(I) - a^*) \,, \tag{26}$$

kde

$$a^* = \frac{\sum_{i=1}^{8} p_i a_i}{\sum_{i=1}^{8} p_i} \,. \tag{27}$$

V duchu předpokladů o třídě „0" tuto skupinu osob Deming zcela zanedbává (srovnej vzorec (20) a související text).

(5) Mám na mysli údaje tvořící předmět výzkumu. Sociodemografické charakteristiky nekontaktovaných a odmítajících osob často zjistit lze.

Po provedení prvého pokusu pak mají následovat další na souboru výběrových jednotek, od nichž dosud nebyly požadované údaje získány. Konkrétní vzorce Deming uvádí pro tři pokusy.[6]) Nevýhodou navrženého postupu je předpoklad apriorně známého rozdělení základního souboru do tříd 0, 1 2, 4, 6 a 8. V tomto směru Demingova metoda nedosahuje úrovně praktické použitelnosti svého „vzoru", Politz-Simmonsova schématu.

Mezi Politz-Simmonsovou, Hansen-Hurwitzovou, Birnbaum-Sirkenovou a Demingovou metodou je řada odlišností jak v přístupu autorů k problematice selhání, tak v aplikovatelnosti jednotlivých postupů, a především v konkrétních návodech a vzorcích. Přesto však mají všechny tyto práce, uveřejněné na přelomu čtyřicátých a padesátých let, určité společné rysy. Hlavním momentem spojujícím práce těchto autorů je, že všechny údaje, o nichž se předpokládá, že jsou známy předem, jsou využity jako koeficienty nebo váhy ve vážených průměrech. Výběry jsou výhradně prosté náhodné (popřípadě dvojité), v každém případě však se stejnou pravděpodobností. Složitější problémy se vlastně obcházejí — tak v Politz-Simmonsově schématu se výběrový charakter vah (převrácených hodnot počtů přítomnosti) zanedbává, Deming zase zanedbává skutečnost, že součin středních hodnot není roven střední hodnotě součinu příslušných veličin. Statistika je v uvedených postupech využita výhradně ve své deskriptivní funkci, a to pokud možno co nejjednodušeji. „Projekční" schopnosti statistické teorie se projevují pouze v nákladových optimalizacích rozsahu výběru a u Birnbauma a Sirkena ve stanovení rozsahu výběru na základě „vyvážení" výběrové a nevýběrové chyby. Pro svoji snadnou aplikovatelnost však jsou uvedené metody dodnes nejen používány, ale i dále rozvíjeny a modifikovány.

Mezi následovníky „klasiků" statistické redukce vlivu selhání je třeba počítat především Bartholomewa. Ten se ve své práci (Bartholomew [1961]) zabývá možností využít ke zlepšení výsledků změny vah při jednom opakování pokusu o kontakt. Mírně modifikuje Cochranův model dvou strat a dělí populaci následovně:

stratum 1 — osoby, u nichž vede první pokus o kontakt k responzi,
stratum 2 — osoby, u nichž vede první pokus o kontakt k selhání nekontaktování nebo odmítnutí,
substratum 2.1 strata 2 — osoby, u nichž vede druhý pokus o kontakt k responzi,
substratum 2.2 strata 2 — osoby, u nichž i druhý pokus o kontakt vede k selhání.

Jako estimátor populačního průměru $\bar{Y}$ pak Bartholomew doporučuje statistiku

$$\bar{y}_B = \frac{n_{11}}{n} h_{11} + \left(1 - \frac{n_{11}}{n}\right) \bar{y}_{21}, \tag{28}$$

kde pro $i = 1,2$ n_{i1} je počet jednotek, od nichž je při i-tém pokusu získána responze,
$\bar{y}_{i1}$ je výběrový průměr sledované proměnné u těchto n_{i1} jednotek.

Kdybychom přijali Hansenův a Hurwitzův předpoklad, že stratum 2.2 je prázdné, splynuly by estimátor (28) s (20). Bartholomew ovšem takový předpoklad nečiní.

Estimátor $\bar{y}_B$ odstraňuje vychýlení způsobené diferencí $\bar{Y}_{11} - \bar{Y}_{21}$ (označíme je A') a ponechává pouze složku vychýlení závislou na rozdílu $\bar{Y}_{21} - \bar{Y}_{22}$ (tu označíme B'), kde $\bar{Y}_{21}$, $\bar{Y}_{22}$ jsou po řadě populační průměry ve stratech 2.1 a 2.2. Bartholomewova idea je shodná s ideou Hansena a Hurwitze „reprezentovat" jednotky selhavší při prvním pokusu nějakou jejich podmnožinou prošetřenou při druhém pokusu. Autor tvrdí, že pokud tazatelé při neúspěšném prvním pokusu získá-

(6) x (II) a x (III) jsou definovány analogicky jako x (I).

vají informace o tom, kdy bývají vybrané osoby doma, lze získat často téměř nevychýlené estimátory založené pouze na jednom opakování pokusu. Platnost takového tvrzení však naráží na základní obsahový problém míry odlišnosti substrat 2.1 a 2.2. Hansen a Waksberg [1970] Bartholomewovi oponují: „Data získaná při opakovaných návštěvách nejsou pravděpodobně reprezentativní pro soubor selhavších jednotek a asi by vnesla větší vychýlení než současné úpravy dat".[7])

Formálně stejného efektu jako Bartholomew dosáhl o 12 let později Thomsen [1973], který také svoji metodu s Bartholomewovou porovnává (viz Thomsen [1975]). Jeho metoda je založena na předpokladu, že ve zkoumané populaci existuje L strat neznámých rozsahů N_i, $i = 1, \ldots, L$, z nichž každé obsahuje cochranovské substratum respondentů (o rozsahu N_{i1}) a substratum „selhávajících" (o rozsahu N_{i2}). Označíme-li $\overline{Y}_i'$ $\overline{Y}_i''$ populační průměry sledované proměnné po řadě v prvém a druhém substratu i-tého strata a $h_i = N_{i1}/N_i$ relativní četnost prvního substrata v i-tém stratu, pak estimátor $\bar{y}$ obvyklý při prostém náhodném výběru (průměr zjištěných hodnot) má střední vychýlení

$$(29) \qquad E(\bar{y} - \overline{Y}) = \frac{1}{h} \sum_{i=1}^{L} \overline{Y}_i' \, W_i(h_i - \bar{h}) + \sum_{i=1}^{L} W_i(1 - h_i)\,(\overline{Y}_i' - \overline{Y}_i''),$$

kde W_i je relativní četnost i-tého strata v základním souboru a kde

$$(30) \qquad \bar{h} = \sum_{i=1}^{L} W_i h_i$$

je podíl sjednocení prvních substrat v základním souboru. První sčítanec na pravé straně (29) označil Thomsen B, druhý A. Vychýlení B je velké, pokud se u velkých strat hodně liší h_i, lze ho však eliminovat použitím estimátoru

$$(31) \qquad \bar{y}_n = \sum_{i=1}^{L} W_i \bar{y}_i ,$$

pro nějž

$$(32) \qquad E(\bar{y}_n - \overline{Y}) = A .$$

Nejsou-li váhy W_i známé, doporučuje Thomsen užít estimátoru

$$(33) \qquad \bar{y}_n^* = \sum_{i=1}^{L} \frac{n_i}{n} \bar{y}_i ,$$

kde n je rozsah výběru,
n_i je náhodný počet jednotek vybraných z i-tého strata.

O estimátoru $\bar{y}_n^*$ dokázal Kish [1965 : 360], že platí-li

$$(34) \qquad P[\forall_i \, n_i > 0] = 1 ,$$

má $\bar{y}_n^*$ stejnou střední hodnotu, a tedy i stejné vychýlení jako $\bar{y}_n$. Pokud však není n velmi velké, má $\bar{y}_n^*$ daleko větší variabilitu, neboť n_i/n je estimátorem W_i a je zatíženo výběrovou chybou.

Thomsen uvádí, za jakých okolností přináší uvedený postup zmenšení celkového vychýlení. K tomu dojde, pokud A a B mají stejné znaménko nebo pokud platí $2|A| < |B|$. Redukce vychýlení je v obou případech

$$(35) \qquad ||B + A| - |A|| < |B| .$$

(7) Rozumí se vážení.

Dále autor uvádí nevychýlený estimátor hodnoty B:

$$(36) \qquad \hat{B} = \frac{n}{n'} \sum \bar{y}_i \frac{n_i}{n} \left(\frac{n_i'}{n_i} - \frac{n'}{n} \right),$$

kde n_i' je počet skutečně prošetřených jednotek náležejících do i-tého strata,
n' je celkový počet skutečně prošetřených jednotek.

Při srovnání své metody s Bartholomewovou Thomsen [1975] upozorňuje, že $|B|$ je podle zkušenosti Ústředního norského statistického úřadu značně menší než absolutní hodnota Bartholomewova B' a že až na jediné šetření (Výzkum spotřebitelských výdajů) byla vždy hodnota B (přesněji $\hat{B}$) zanedbatelná.

Aplikujeme-li Thomsenovu metodu na Bartholomewův model tak, že 1. skupina (v Thomsenově smyslu) odpovídá stratu jednotek v prvním pokusu selhávajících, máme $A = A'$, $B = B'$.

Na varšavském zasedání ISI v roce 1975 vystoupili Nargundkar a Joshi [1975] s modelem vycházejícím z chápání responze jako jevu, který má u i-té jednotky základního souboru pravděpodobnost výskytu u_i, $0 < u_i \leq 1$. Jde do jisté míry o rozvinutí a teoretické zdokonalení výše uvedeného Demingova přístupu. Je-li π_i pravděpodobnost zahrnutí i-té jednotky do výběru, má v daném modelu tzv. Horvitz-Thompsonův estimátor populačního úhrnu při jednostupňovém nestratifikovaném výběru tvar

$$(37) \qquad \hat{x} = \sum_i \frac{x_i}{\pi_i u_i},$$

kde x_i je pozorování u i-té výběrové jednotky a sumace se provádí přes všechny výběrové jednotky. Autoři uvádějí i rozptyl tohoto estimátoru.

Kardinálním problémem zmíněného postupu je, že vyžaduje znalost pravděpodobností u_i nebo alespoň jejich odhad pro každou jednotku z výběrového souboru. V tomto smyslu se Nargundkarova a Joshiho práce blíží pojetí Politzovu a Simmonsovu a Demingovu. Pravděpodobnosti u_i postihují nejen „dostupnost" jednotek (jako u Politze a Simmonse), ale i jejich ochotu poskytnout požadované informace (jako u Deminga). Návod na odhad těchto pravděpodobností práce bohužel neobsahuje, autoři pouze doporučují (v duchu Demingovy metody) vytvořit třídy výběrových jednotek se stejnou pravděpodobností u_i nějakým způsobem známou, tedy vycházející ze zkušenosti.

V jednom směru však práce Nargundkara a Joshiho překračuje rámec proudu statistického zacházení se selháním, který jsem označil za „klasický" — nepředpokládá totiž nic o výběrovém postupu a připouští libovolné (nenulové) pravděpodobnosti zahrnutí π_i i π_{ij}.

Historie tohoto netriviálního zobecnění sahá až do poloviny padesátých let. Tehdy přišel se zcela novou technikou zacházení s problémem nekontaktování Dalenius [1955], který navrhl využít pro zmenšení vlivu selhání výběru s nestejnými pravděpodobnostmi. Jeho metoda předpokládá náhodný výběr rodin domácností a v nich výběr náhodné osoby (srv. Kish [1965]) tak, že přítomné osoby mají vyšší pravděpodobnost vybrání proti nepřítomným. Pokud je vybrána nepřítomná osoba, tazatel se vrátí týž den znovu a pokusí se o kontakt. Cílem zavedení nestejných pravděpodobností vybrání je stratifikace a zmenšení počtu „zbytečných", tj. neúspěšných pokusů. Slabinou navrženého postupu je, že využívá estimátoru odvozeného pro dvojstupňový výběr a zbývající selhání v podstatě ignoruje. Proto se i na Daleniovu práci vztahují kritická slova řečená výše v souvislosti s Bartholomewovou úpravou estimátoru.

Nákladová optimalizace se zde netýká rozsahu výběru, ale pravděpodobnosti

zahrnutí ve druhém stupni výběru. I to je odlišnost ve srovnání s pracemi z přelomu čtyřicátých a padesátých let.

Všechny dosud uvedené postupy redukce vlivu selhání na výsledky výzkumu jsou založeny na „klasické" statistice, „klasickém" chápání výběrového postupu, při němž se předpokládá, že v základním souboru jsou pevně dané charakteristiky, tj. hodnoty všech zjišťovaných znaků, a odhady těchto pevných hodnot jsou založeny na znalosti rozdělení četností pro jednotlivé výběry. Údaje, které se nepodaří zjistit, jsou více či méně propracovanými postupy nahrazovány hodnotami vypočtenými ze zjištěných údajů. Použité výpočetní postupy jsou vesměs poměrně jednoduché a jsou založeny na doplnění poznatků, které byly předmětem výzkumu, o pomocná data (Politz-Simmonsovo schéma a obdobné postupy) nebo na dodatečném prošetření alespoň části jednotek, které v první vlně pokusů „selhaly" (hansen-hurwitzovský směr). Aplikace uvedených postupů nevyžaduje žádné předběžné předpoklady teoretického rázu, i když někdy je požadována apriorní znalost některých charakteristik. Implicitně se však předpokládá, že chybějící data jsou „v podstatě stejná" jako data zjištěná. Tento předpoklad může být v některých případech nesprávný, zjistit jeho oprávněnost či neoprávněnost však vyžaduje dlouhodobou pečlivou práci s empirickým materiálem a kombinování různých metod zjišťování předmětných údajů. Souhrnně však lze říci, že pro běžnou praxi lze mezi „klasickými" postupy najít řadu takových, které při minimální náročnosti přinášejí nezanedbatelná zpřesnění výsledků šetření.

Úpravy vlivu selhání založené na bayesovské statistice

Dalenius a Nargundkar s Joshim zavedli modely založené na pravděpodobnostním modelu selhání a je možno říci, že se tím poněkud vymykají z „klasického" proudu. Skutečný obrat však znamenala práce Ericsona [1967], který poprvé využil apriorních znalostí pomocí bayesovské statistiky. Teoreticky vychází Ericson z práce Raiffa-Schlaifer [1961], není však vyloučeno, že ho k novému přístupu inspirovala subkapitola Kishovy knihy [1965] věnovaná vlivu vychýlení na pravdepodobnostní tvrzení. Celý postup by z hlediska obecné teorie výběrových šetření bylo možno zařadit do třídy výběrových modelů založených na pojmu „superpopulace" (viz například Hájek [1981]).

Základní ideou Ericsonovy práce je zdokonalení Hansen-Hurwitzovy metody v tom smyslu, že rozsah výběru pro první i pro druhý pokus (značí se po řadě n, m) jsou určovány na základě apriorního, hypotetického rozdělení pravděpodobností náhodného vektoru (μ_1, μ_2, π), kde

μ_1 je populační průměr sledované charakteristiky ve stratu „respondentů", tj. osob, které poskytnou údaje při prvním pokusu,
μ_2 je populační průměr této charakteristiky ve stratu ostatních osob,
π je podíl prvního strata na populaci.

Označíme dále μ hodnotu sledované charakteristiky v celém základním souboru. Konstrukce jejího estimátoru $\hat{\mu}$ vychází rovněž z uvedeného apriorního rozdělení a z kvadratické ztrátové funkce

$$K(\hat{\mu} - \mu)^2 + c_1 n + c_2 r + c_3 m \,, \tag{38}$$

kde K je ocenění poměru mezi chybou odhadu a náklady,

c_1 jsou náklady na vybrání a pokus o kontaktování jedné jednotky prošetřené v první „vlně" pokusů,
c_2 jsou náklady na zjištění a zpracování údajů od jednotky v první „vlně" pokusů,

c_3 jsou náklady na výběr jednotky pro druhou „vlnu" pokusů a na zjištění a zpracování údajů od ní,
n je rozsah výběrového souboru pro první „vlnu",
r je počet responzí v první „vlně",
m je počet jednotek vybraných a prošetřených ve druhé „vlně".

Stejně jako Hansen a Hurwitz [1946] Ericson předpokládá, že všechny jednotky vybrané pro druhý pokus budou prošetřeny.

Uvedená ztrátová funkce vede k tomu, že optimálním estimátorem je střední hodnota aposteriorního rozdělení statistiky

$$\pi\mu_1 + (1 - \pi)\,\mu_2 \tag{39}$$

po provedení obou „vln" pokusů. Autoři předpokládají, že náhodná veličina π má apriorní rozdělení typu beta s parametry n', r' (interpretovat je lze jako předpokládaný rozsah výběru a předpokládaný počet responzí v první „vlně") a že náhodný vektor (μ_1, μ_2) je na π nezávislý a má známé normální rozdělení. Za těchto předpokladů lze odvodit optimální rozsah výběru v obou krocích (n, m) a dokázat, že estimátor (39) má tvar

$$\mu_0 = \frac{r + r'}{n + n'}\,\mu_1'' + \frac{n + n' - r - r'}{n + n'}\,\mu_2'' , \tag{40}$$

kde μ_1'' a μ_2'' jsou aposteriorní střední hodnoty náhodných veličin μ_1, μ_2 po provedení obou „vln" pokusů. Konkrétní vzorce a zejména postup určení optimálního n jsou velmi složité a jejich uvádění by si vyžádalo zavedení rozsáhlého aparátu, který by přesahoval meze tohoto přehledu. Pokud však změníme předpoklad o apriorním rozdělení podílu „respondentů" a budeme předpokládat, že π má v intervalu $\langle 0, 1\rangle$ rovnoměrné rozdělení (odpovídá to nulové informaci o podílu selhání), a pokud μ_1 i μ_2 mají velmi velkou variabilitu, lze (39) aproximovat výrazem

$$\frac{n}{r}\,\bar{X}_r + \frac{n - r}{n}\,\bar{Y}_m , \tag{41}$$

kde $\bar{X}_r$ je výběrový průměr dat zjištěných v první „vlně",
$\bar{Y}_m$ je výběrový průměr dat zjištěných ve druhé „vlně".

Vzorec (41) splývá s Hansenovým a Hurwitzovým estimátorem (20).

Ericson také zavádí model pro zacházení s vychýlením, které může vzniknout „přinucením" vybraných osob k odpovědi ve druhé „vlně" pokusů. Tato modifikace vede ke změně rozsahu výběrového souboru ve druhém kroku. Tím ovšem nelze vzniklé vychýlení odstranit, možné je pouze zmenšit celkovou chybu estimátoru.

V podstatě stejný přístup jako u Ericsona lze nalézt u Kaufmana a Kinga [1973], kteří ovšem opouštějí předpoklad normality a jako apriorní rozdělení všech složek uvažovaného náhodného vektoru zavádějí rozdělení typu beta. Omezují se přitom na jednodušší případ odhadu proporcí, tedy případ, kdy jednotlivá měření jsou nula-jedničkové náhodné veličiny.

Jiným směrem rozvíjejí Ericsonovu práci Singh-Sedransk [1978]. Ti si povšimli, že Ericsonův model je odvozen pro nekonečné základní soubory. Tuto nevýhodu odstranili[8]) ve výběrovém postupu probíhajícím ve dvou krocích tak, že první výběr n' jednotek (pilotáž) se provádí z celého základního souboru rozsahu N a druhý (n jednotek) ze „zbývajících" $N - n'$ jednotek. V prvém kroku je vybráno n_1' respondentů a u n_2' jednotek dojde k selhání, v druhém je n_1 respondentů a n_2

(8) Teoreticky je uvedený postup založen na práci Rao – Ghangurde [1972]

selhání. Při prvním výběru se podaří dodatečně získat údaje od m' z n_2' osob, ve druhém od m z n_2 osob.

Označíme-li

p'' podíl jednotek poskytujících údaje na první pokus (tedy podíl $n_1 + n_1'$ na $n + n'$),

$\bar{x}_{s1}''$ výběrový průměr sledované proměnné pro $n_1 + n_1'$ jednotek, které poskytly údaje na první pokus,

$\bar{x}_{s2}''$ výběrový průměr sledované proměnné pro $m + m'$ jednotek, které se podařilo prošetřit teprve na druhý pokus,

pak aposteriorní střední hodnota pro populační průměr po provedení obou prostých náhodných výběrů je

$$\bar{x}_s'' = p''\bar{x}_{s1}'' + (1 - p'')\,\bar{x}_{s2}'' , \tag{42}$$

což je analogie Hansenova a Hurwitzova estimátoru (20). Výhodou uvedeného postupu však je, že rozptyl statistiky $\bar{x}_s''$ nezávisí na neznámém parametru π udávajícím hypotetický podíl responzí v celém základním souboru.[9]) Za předpokladu normality odvozují Singh a Sedransk optimální hodnotu m a aproximaci pro optimální rozsah druhého výběru n při ztrátové funkci obdobné (38). Na první pohled se může zdát nevýznamné znát optimální počet úspěšných druhých pokusů v druhém výběru, je třeba si však uvědomit, že úspěšnost opakovaných pokusů lze ovlivnit vhodným postupem zjišťování údajů.

Z citované práce je třeba ještě upozornit na modifikaci celé procedury, která nepředpokládá provedení prvního výběru (pilotáže). Statistickou „kompenzací" za informaci, která je tím ztracena, je předpoklad, že výběrový průměr sledované proměnné u selhavších jednotek $\bar{x}_{s2}$ se od stejné hodnoty u prošetřených jednotek $\bar{x}_{s1}$ liší pouze o aditivní konstantu γ:

$$\bar{x}_{s2} = \bar{x}_{s1} + \gamma . \tag{43}$$

Aposteriorní střední hodnota pro populační průměr má v daném modelu tvar

$$\bar{x}_s = \frac{n_1}{n}\,\bar{x}_{s1} + \left(1 - \frac{n_1}{n}\right)\bar{x}_{s2} + \frac{mn_1}{N}\,\frac{\bar{x}_{s2} - \bar{x}_{s1} - \gamma}{m + n_1} , \tag{44}$$

kde význam jednotlivých symbolů je analogický vzorci (42). Uvedený postup lze srovnat s jednofázovou procedurou navrženou Rubinem [1977], v níž aposteriorní střední hodnota populačního průměru má přesně tvar Hansen-Hurwitzova estimátoru (20), tedy postrádá poslední sčítanec ve (44), který představuje „korekci" původního předpokladu (43).

Modely Ericsona a jeho následovníků jsou velmi přínosné. Umožňují využít poznatků, jimiž výzkumné pracoviště disponuje ještě před provedením vlastního šetření, a zabezpečují teoreticky opodstatněné korekce zjištěných údajů s ohledem na podíl selhání, které je důsledně chápáno jako náhodný jev. Předpokládají však velmi zjednodušeně, že selhání nezávisí na zjišťované proměnné. Lze vůči nim tedy vznést výhradu, která již byla vyslovena ve vztahu ke „klasickým" postupům. Na tuto skutečnost upozorňují Little [1982] a Greenlees s Reecem a Zieschangem [1982]. V článku prvého z nich lze kromě postupu využívajícího bayesovsky známou informaci o základním i výběrovém souboru nalézt i model založený na předpokladu, že responze je od i-té jednotky získána, pokud náhodná veličina λ_i nabyde hodnoty

(9) V tomto směru je poměr Singhovy a Sedranskovy práce k původnímu Ericsonovu modelu stejný jako vztah Srinathovy [1971] úpravy k původnímu Hansen – Hurwitzovu postupu.

menší než c (konstanta), přičemž λ_i má se zjišťovanou náhodnou veličinou v_i sdružené normální rozdělení s kovariancí β. Metodou maximální věrohodnosti je za tohoto předpokladu možno odvodit estimátor populačního průměru zjišťované proměnné ve tvaru

$$\hat{\mu} = \bar{v} + \left(\frac{n-m}{m}\right) \hat{\beta}\, \Phi(\hat{c})\, (1 - \Phi(\hat{c}))^{-1}, \tag{45}$$

kde n je rozsah výběru,
m je počet prošetřených jednotek,
$\bar{v}$ je výběrový průměr sledované proměnné pro m prošetřených jednotek,
$\hat{\beta}, \hat{c}$ jsou odhady pro β a c získané metodou maximální věrohodnosti.

V článku je uvedena i modifikace daného postupu pro simultánní odhad rozdělení četnosti pro kategorizovaná data.

Greenless, Reece a Zieschang navrhují obdobný model, v němž se však předpokládá, že platí regresní vztah zjišťované proměnné Y vůči skupině proměnných $x_1, \ldots, x_k$, jejichž hodnoty jsou pro všechny jednotky ve výběrovém souboru známy, a že pravděpodobnost responze je logistickou funkcí hodnoty zjišťované proměnné Y a dalších proměnných, jejichž hodnoty jsou pro všechny jednotky v základním souboru známy — $Z_1, \ldots, Z_m$.[10])

Parametry $\alpha, \beta_1, \ldots, \beta_k, \gamma, \delta_1, \ldots, \delta_m$ modelu mohou být odhadnuty metodou maximální věrohodnosti a pro jednotky, od nichž nejsou k dispozici údaje, může být určena „náhradní" hodnota pomocí věty o podmíněné střední hodnotě. Celý postup je však výpočetně velmi pracný a jeho podrobnější uvedení by přesáhlo rámec této práce. Rozhodně se však jedná o postup, který velmi dobře odpovídá řadě reálných případů, s nimiž se v sociologických výzkumech můžeme setkat.

Bayesovské postupy, jejichž proniknutí do problematiky výběru koresponduje s velkým rozmachem teorie bayesovské statistiky v šedesátých letech, mají ze statistického hlediska nesporné přednosti, spočívající zejména v přesnějším a plnějším využití apriorních znalostí i v tom, že mají optimalizační povahu. Teprve Ericsonova práce přinesla efektivní algoritmus pro určování optimálního výběrového podílu v několika po sobě jsoucích „vlnách" pokusů. Za přesnost a obsažnost modelu se však platí teoretickou i výpočetní složitostí, která proniknutí těchto postupů do praxe brzdí. Navíc nároky kladené na apriorní informace prakticky předurčují bayesovské postupy pouze pro ta výzkumná pracoviště, která disponují poměrně rozsáhlými znalostmi o zkoumané populaci. Domnívám se však, že přes uvedené obtíže je bayesovský přístup k problematice selhání velmi perspektivní, i když od něj pochopitelně nelze očekávat více než „umně vypočtené korekce" zjištěných údajů.

Závěr

Zamyslíme-li se nad celou plejádou postupů „řešení" problému selhání výběru, můžeme rozdělit uvedené metody do pěti skupin (prvé tři uvádějí Nargundkar a Joshi [1975], další dvě však, podle mého názoru, nelze do jimi vymezených kategoriií zahrnout):

1. úpravy estimátorů s využitím vah úměrných podílu počtu výběrových jednotek k počtu responzí,
2. nahrazení chybějící informace informací získanou od respondentů se stejnými charakteristikami (popř. údaji od jinak určených respondentů),
3. Politz-Simmonsovo schéma,

(10) Podobný model navrhli Cassel, Särndal a Wretman [1979], kteří však uvažovali závislost responze pouze na $x_1, \ldots, x_k$.

4. využití apriorních znalostí pomocí bayesovské statistiky,
5. opakované pokusy o získání responze.

Metody spadající do první až čtvrté skupiny implicitně předpokládají, že informace získaná od respondentů je co do charakteru stejná jako údaje, které zůstaly nedostupné (viz Nargundkar-Joshi [1975]). Výjimkou z tohoto pravidla jsou nejnovější bayesovsky orientované práce, v nichž se připouští závislost selhání na zjišťovaných údajích. Podle Thomsena [1973] jsou skupiny 1, 3 a 4 charakterizovány úsilím o zmenšení vlivu selhání ve fázi zpracování dat, skupina 5 soustřeďuje metody redukce podílu selhání. Do skupiny 2 by ovšem bylo možno zařadit vedle metod „úprav dat" také jeden z postupů orientovaných na redukci podílu selhání — metodu náhradníků, o níž jsem se zmínil v prvé části přehledu.

Jednotlivé postupy mohou být, jak bylo z některých citovaných prací patrné, vzájemně kombinovány, cílem kombinací bývá co největší redukce vychýlení při zachování organizační jednoduchosti terénní práce a minimalizaci nákladů. Považuji však za nutné znovu připomenout, že statistické korekční metody nemohou problém selhání vyřešit, neboť vždy upravují získané údaje zase jen na základě těchto údajů samotných, neberou ohled na příčiny selhání, strukturu skupiny neprošetřených osob a často ani na to, zda selhání vzniklo odmítnutím nebo nekontaktováním. V praxi by měla mít proto vždy hlavní slovo obsahová, sociologická analýza, na jejímž základě je pak možné aplikovat postupy ke zpřesnění numerických výsledků šetření.

Literatura

Anderson, H.: *On nonresponse bias and response probabilities.* Scandinavian Journal of Statistics, Theory Appel 6 (1979), s. 107—112.

Bailar, B. A. — Bailey, L. — Corby, C.: *A comparison of some adjustement and weighting procedures for survey data.* In: Namboodhi, N. K. ed.: *Survey sampling and measurement.* New York, Academic Press 1978, s. 175—198

Bartholomew, D. J.: *A method of allowing for "not-at-home" bias in sample survey.* Applied Statistics 10 (1961), s. 52—59.

Biggeri, L.: *On the distinction between sampling and non-sampling errors and the use of mathematical models for their evaluation.* BISI/BIIS 46 (1975), book 3, s. 114—120.

Birnbaum, Z. W. — Sirken, M. G.: *Bias due to non-availability in sampling surveys.* Journal of the American Statistical Association 45 (1950a), s. 98—111.

Birnbaum, Z. W. — Sirken, M. G.: *On the total error due to noninterview and to random sampling.* International Journal of Opinion and Attitude Research 4 (1950b), s. 179—191.

Blumberg, H. H. — Fuller, C. — Hare, A. P.: *Response rates in postal surveys.* Public Opinion Quarterly 38 (1974), s. 113—123.

Cassel, C. M. — Särndal, C. E. — Wretman, J.: *Prediction theory for finite population when model-based and design-based principles are combined.* Scandinavian Journal of Statistics, Theory Appel 6 (1979), s. 97—106.

Cochran, W. G.: *Sampling techniques.* 2nd edition, New York, Wiley 1963.

Crossley, H. M. — Fink, R.: *Response and non-response in probability sample.* International Journal of Opinion and Attitude Research 5 (1951), s. 1—19.

Dalenius, T.: *The problem of not-at-home.* Statistik Tidskrift 4 (1955), s. 208—211.

Dalenius, T.: *Treatment of the non-response problem.* Journal of Advertising Research 9 (1961), s. 1—7.

Dalenius, T.: *Automatic estimation of missing values in censuses and sample surveys.* Statistik Tidskrift 11 (1962), s. 359—400.

Dalenius, T.: *Non-sampling errors in surveys.* Conference on mathematical methods in economic research, Prag 1966, s. 19—57.

Dalenius, T.: *Bibliography on non-sampling errors in surveys.* International Statistical Review 45 (1977), s. 71—89, 181—197, 303—317.

Deming, W. E.: *On errors in surveys.* American Sociological Review 9 (1944), s. 359—369.

Deming, W. E.: *On a probability mechanism to attain economic balance between the resultant error of response and the bias of nonresponse.* Journal of the American Statistical Association 48 (1953), s. 743—772.
Dillman, D. A.: *Increasing mail questionnaire response in large samples of the general public.* Public Opinion Quarterly 36 (1972), s. 254—257.
Durbin, J.: *Non-response and call-backs in surveys.* BISI/BIIS 26 (1954), s. 72—86.
Durbin, J. — Stuart, A.: *Callbacks and clustering in sample surveys: An experimental study.* Journal of the Royal Statistical Society, series A, 117 (1954), s. 388—428.
El-Badry, M. A.: *A sampling procedure for mailed questionnaires.* Journal of the American Statistical Association 51 (1956), s. 209—227.
Ericson, W. A.: *Optimal sample design with nonresponse.* Journal of the American Statistical Association 62 (1967), s. 63—67.
Greenlees, W. S. — Reece, J. S. — Zieschang, K. D.: *Imputation of missing value when the probability of response depends upon the variable being imputed.* Journal of the American Statistical Association 77 (1982), s. 251—261.
Grušin, V. A.: *Sociologický výskum.* Bratislava, Pravda 1972.
Hájek, J.: *Sampling from a finite population.* New York, Dekker 1981.
Hansen, M. H. — Hurwitz, W. N.: *The problem of non response in sample surveys.* Journal of the American Statistical Association 41 (1946), s. 517—529.
Hansen, M. H. — Hurwitz, W. N. — Bershad, M.: *Measurement errors in censuses and surveys.* BISI/BIIS 33 (1963), s. 359—374.
Hansen, M. H. — Waksberg, J.: *Research on non-sampling errors in censuses and surveys.* International Statistical Review 38 (1970), s. 317—332.
Hartley, H. O.: *Discussion of paper by F. Yates.* Journal of the Royal Statistical Society, series A, 109 (1946), s. 44—46.
Hawkins, D. F.: *Estimation of nonresponse bias.* Sociological Methods and Research 3 (1975), s. 462—485.
Hendrics, W. A.: *Adjustment for bias by nonresponse in mailed surveys.* Agricultural Economics Research 1 (1949), s. 52—56.
Herzmann, J.: *Přehled koncepcí a stanovisek k problematice selhání výběrových jednotek. I. Sociálně psychologické aspekty.* Sociologický časopis 3 (1984), s. 165.
Herzmann, J. — Kábrt, J.: *K problematice odmítnutí rozhovoru v sociologických výběrových šetřeních.* Sociologický časopis 1 (1982), s. 74—90.
Houseman, E. E.: *Statistical treatment of the nonresponse problem.* Agricultural Economics Research 5 (1953), s. 12—19.
Kaufman, F. M. — King, B.: *A bayesian analysis of nonresponse in dichotomous process.* Journal of the American Statistical Association 68 (1973), s. 670—678.
Kish, L.: *Survey Sampling.* New York, Wiley 1965.
Leuven, M.: *The income of physicians.* Chicago, Chicago University Press 1932.
Linehan, T. P.: *Problems of confidentiality with particular references to population censuses.* BISI/BIIS 45 (1973), Bok 3, s. 214—227.
Little, R. J. A.: *Models for nonresponse in sample surveys.* Journal of the American Statistical Association 77 (1982), s. 237—249.
Moser, C. A. — Kalton, G.: *Survey methods in social investigation.* 2nd edition, London, Heinemann 1971.
Nargundkar, M. S. — Joshi, G. B.: *Non-response in sample surveys.* Contributed papers of the 40th session of the ISI, Warszawa 1975, s. 626—628.
Neyman, J.: *Contribution to the theory of sampling human populations.* Journal of the American Statistical Association 33 (1938), s. 101—116.
Neyman, J.: *Bias in surveys due to non-response.* In: Johanson, N. L. — Smith, H. ed.: *New development in survey sampling.* New York, Wiley 1969.
Platek, R. — Singh, M. P. — Stremblay, V.: *Adjustment for nonresponse in surveys.* In: Namboodhi, N. K.: *Survey sampling and measurement.* New York, Academic Press 1978, s. 157—174.
Politz, A. N. — Simmons, W. R.: *An attempt to get the "not at homes" into sample without callbacks.* Journal of the American Statistical Association 44 (1949), s. 9—31.
Politz, A. N. — Simmons, W. R.: *Note on "An attempt to get the not-at-homes into the sample without callbacks".* Journal of the American Statistical Association 45 (1950), s. 136—137.
Raiffa, H. — Schlaifer, R. O.: *Applied statistical decision theory.* Boston, Harvard University 1961.
Rao, J. N. K.: *Some nonresponse sampling theory when the frame contains an unkown amount of duplications.* Journal of the American Statistical Association 63 (1968), s. 41—63.

Rao, J. N. K. — Ghangurde, P. D.: *Bayesian optimization in sampling finite populations.* Journal of the American Statistical Association 67 (1972), s. 439—443.
Rubin, D. B.: *Formalizing subjective notions about the effect of nonrespondents in sample surveys.* Journal of the American Statistical Association 72 (1977), s. 538—543.
Rubin, D. B.: *Using multivariate matched sampling and regression adjustment to control bias in observational studies.* Journal of the American Statistical Association 74 (1979), s. 318—327.
Sethi, I. C. — Singh, D.: *A study of nonresponse in "repeated survey". Abstract.* Journal of the Indian Society of Agricultural Statistics 23 (1971), s. 95.
Simmons, W. R.: *A plan to account for "not-at-homes" by combining weighting and callbacks.* The Journal of Marketing 11 (1954), s. 42—54.
Singh, B. — Sedransk, J.: *A two-phase sample design for estimating the finite population mean when there is nonresponse.* In: Namboodhi, N. K. ed.: *Survey sampling and measurement.* New York, Academic Press 1978, s. 143—155.
Srinath, K. P.: *Multiphase sampling in nonresponse problem.* Journal of the American Statistical Association 66 (1971), s. 583—586.
Stephan, F. — McCarthy, P.: *Sampling opinions.* New York, Wiley 1958.
Sudman, S.: *Probability sampling with quotas.* Journal of the American Statistical Asociation 61 (1966), s. 749—771.
Šljapentoch, V. E.: *Problemy reprezentativnosti sociologičeskoj informacii.* Moskva, Statistika 1976.
Thomsen, J.: *A note on the efficiency of weighting subclass means to reduce the effect of non-response when analysing survey data.* Statistik Tidskrift 11 (1973), s. 278—283.
Thomsen, J. B.: *Evaluating the efficiency of two weighting procedures to reduce non-response bias: an application of Cochran's non-response stratum approach.* Contributed papers of the 40th session of the ISI, Warzsawa 1975, s. 826—831.
Wiesman, F.: *Methodological bias in public opinion surveys.* Public Opinion Quarterly 36 (1972), str. 105—108.

Резюме

Я. Герцманн: Обзор концепций и позиций по проблематике неудачи в использовании выборочных единиц. II. Статистические аспекты

В «Социологическом журнале» № 3 за 1984 год был опубликован перечень работ, посвященных социологическим и социально-психологическим аспектам неудачи выбора при проведении социологического обследования. Эта статья исходит из вышеприведенной статьи. В ней ведется речь о статистических методах, служащих для оценки и уменьшения отклонения результатов обследований (анкетирования), вызванных неудачей выбора. При подготовке этой статьи автор ставил своей целью дать в распоряжение читателей исторически упорядоченный обзор, дополненный небольшими примечаниями и сравнением. Поэтому отдельные методы не представлены в деталях — подчеркиваются прежде всего их главные идеи. Автор стремился оказать помощь читателям, интересующимся данной проблематикой, в поисках целесообразных статистических методов в случае неудачи в выборе единиц при проведении социологических обследований.

Во введении к статье можно найти образцы, описывающие величину отклонения, возникшего при неудаче выбора, выведенную как с помощью классических, так и с помощью статистических методов Байеса. В этой связи приводятся также различные методы подсчета так называемой «общей ошибки», первым компонентом которой является выборочная ошибка, а второй — отклонение, возникшее вследствие неудачи (осечки).

Во второй части статьи говорится о методах редуцирования отклонения, вызванного неудачей в выборе, основанных на классической статистике.

Байесовские модели решения неудачи в выборе подытожены в последней части статьи. При упрощенных предпосылках байесовский метод сопоставляется с классическим, причем обнаруживается, что байесовские оценки в определенном смысле являются обобщением оценок, предложенных в 50-е и в начале 60-х годов.

Статья дополнена обширным перечнем литературы.

Summary

J. Herzmann: Survey of Conceptions and Standpoints Relating to the Problem Area of Non-response. II. Statistical Aspects

In the Sociological Review No. 3/1984, a survey of works dealing with the sociological and social psychological aspects of nonresponse was published. The present paper discusses statistical methods applied in the estimation and reduction of the nonresponse bias. Its aim was not to give methods in full detail but merely to bring forward the author's ideas, and perhaps to encourage the interested reader to look for an adequate statistical treatment of nonresponse in a social survey.

In the first part the nonresponse bias is studied both in the classical and in the bayesian form. Some methods combining the nonresponse bias with the sampling error into a "total error" are mentioned here.

The second part of the paper deals with methods of reducing the nonresponse bias based on classical statistics.

Bayesian models of nonresponse may be found in the last part of the paper. Some simplifying assumption allow us to compare bayesian approaches with classical ones, and in this way bayesian estimates are shown to be a kind of generalization of classical estimates.

The paper is supplemented by a large list of references.