Aplikace matematiky

Jaroslav Hájek
Oblastní výběr

Aplikace matematiky, Vol. 1 (1956), No. 2, 149–161

Persistent URL: http://dml.cz/dmlcz/102524

OBLASTNÍ VÝBĚR

JAROSLAV HÁJEK

(Došlo dne 3. listopadu 1955.) DT: 519.2

Z problémů, souvisících s oblastním výběrem je pojednáno o následujících: (1) sestrojení lineárního a poměrového odhadu a příslušných intervalů spolehlivosti; (2) optimální rozvržení výběru do oblastí vzhledem k dané nákladové funkci; (3) náhodné rozvržení výběru do oblastí.

1. Výběrové šetření

Výběrové šetření je statistické šetření na určitém počtu jednotek (předmětů, osob, správních jednotek a pod.) vybraných pomocí *pravděpodobnostního*[1]) *výběru* z daného základního souboru. Přitom termín „výběrové šetření" je užíván speciálně pro výběry ve velkém, často i celostátním měřítku, prováděné pro účely vlády, ministerstev a výzkumných ústavů. Zde vznikají osobité problémy, odlišující tyto výběry od výběrů v malém rámci, na př. uvnitř podniků při kontrole jakosti výroby. Samy tyto problémy sice nejsou předmětem matematického studia, avšak inspirují nalézání a zkoumání osobitých způsobů výběru a konstrukcí odhadů. Tak se v posledních desetiletích rozvinula theorie oblastního (stratifikovaného) výběru, mechanického (systematického) výběru, několikastupňového výběru, výběru s nestejnými pravděpodobnostmi, theorie poměrového odhadu a pod. V tomto článku je pojednáno o nejstarší a nejužívanější úpravě výběrového postupu — o oblastním výběru, a to jak ve spojení s lineárním tak i poměrovým odhadem.

2. Oblastní výběr

Oblastní způsob výběru bývá motivován buď tím, že základní soubor je představován několika seznamy, z nichž každý chceme zpracovat samostatně, nebo tím, že v určitých nám známých částech základního souboru se zkoumaná

[1]) Pravděpodobnostní výběr je opakem výběru úsudkového: jednotky se při něm nevybírají podle subjektivního úsudku resp. „cestou nejmenšího odporu", ale tak, že jednotlivým eventualitám je dána určitá pravděpodobnost, že právě ony budou vybrány.

veličina mění méně než v základním souboru jako celku, nebo do třetice tím, že v jedněch částech základního souboru naráží sběr údajů na větší obtíže než v jiných, takže máme zájem, vybrat v nich relativně méně pozorování. Ve všech těchto případech rozdělujeme, resp. máme už předem rozdělen základní soubor na několik samostatných částí — *oblastí*, a v každé zvlášť, nezávisle na sobě, provedeme výběr buď proporcionálního nebo co nejvýhodnější stanoveného počtu výběrových jednotek.

Vyjmenujme si stručně hlavní výsledky teorie oblastního výběru, týkající se lineárního[2]) odhadu průměru hodnot určité veličiny. Nuže, označme si počet oblastí L, počet výběrových jednotek v α-té oblasti N_α a hodnoty, sledované na těchto jednotkách — $a_{\alpha i}$, $\alpha = 1, \ldots, L$, $i = 1, \ldots, N_\alpha$; definujme si v každé oblasti oblastní rozptyl

$$\sigma_\alpha^2 = \frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} a_{\alpha i}^2 - \left(\frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} a_{\alpha i} \right)^2 , \quad \alpha = 1, \ldots, L . \tag{2.1}$$

V každé oblasti vyberme, nezávisle na výběru v ostatních oblastech, n_α jednotek pomocí náhodného výběru bez opakování a získaná pozorování si označme $x_{\alpha i}$, $\alpha = 1, \ldots, L$, $i = 1, \ldots, n_\alpha$. Položíme-li $N = \sum_{\alpha=1}^{L} N_\alpha$ a sestrojíme-li odhad

$$\bar{x} = \frac{1}{N} \sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} x_{\alpha i} , \tag{2.2}$$

pak střední hodnota $\bar{x}$ bude rovna základnímu průměru $\bar{a}$

$$\mathbf{E}\bar{x} = \bar{a} = \frac{1}{N} \sum_{\alpha=1}^{L} \sum_{i=1}^{N_\alpha} a_{\alpha i} \tag{2.3}$$

a rozptyl $\bar{x}$ bude roven

$$\mathbf{D}^2\bar{x} = \sum_{\alpha=1}^{L} \left(\frac{N_\alpha}{N} \right)^2 \frac{N_\alpha - n_\alpha}{N_\alpha - 1} \frac{\sigma_\alpha^2}{n_\alpha} ,$$

čili přibližně

$$\mathbf{D}^2\bar{x} = \sum_{\alpha=1}^{L} \left(\frac{N_\alpha}{N} \right)^2 \left(1 - \frac{n_\alpha}{N_\alpha} \right) \frac{\sigma_\alpha^2}{n_\alpha} . \tag{2.4}$$

(Operátory $\mathbf{E}$ a $\mathbf{D}^2$ zde i v dalším označují střední hodnotu a rozptyl.) Kromě toho, pomocí oblastních výběrových rozptylů s_α^2

$$n_\alpha s_\alpha^2 = \sum_{i=1}^{n_\alpha} x_{\alpha i}^2 - \frac{1}{n_\alpha} \left(\sum_{i=1}^{n_\alpha} x_{\alpha i} \right)^2 , \quad \alpha = 1, \ldots, L , \tag{2.5}$$

bude možno sestrojit *interval spolehlivosti* pro $\bar{a}$ ve tvaru

$$\bar{x} \pm t \frac{1}{N} \sqrt{ \sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \left(\frac{N_\alpha}{n_\alpha} - 1 \right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha - 1} } , \tag{2.6}$$

[2]) Lineární odhad je lineární funkcí jednotlivých pozorování; v tom je zahrnut předpoklad, že součinitelé při jednotlivých pozorováních nezávisí na výsledku výběru.

kde velikost t se řídí velikostí pravděpodobnosti — t. zv. *spolehlivosti*, s níž chceme, aby interval (2.6) pokryl $\bar{a}$. Je-li celkový počet pozorování velký, řekněme aspoň 100, můžeme pro spolehlivost 0,95 vzít $t = 2$, což odpovídá Studentovu rozdělení s 60 stupni volnosti, a pro spolehlivost 0,99 můžeme vzít $t = 2{,}7$, což odpovídá Studentovu rozdělení se 40 stupni volnosti. Je-li celkový počet pozorování poměrně malý, nebo jsou-li oblasti neproporcionálně zastoupeny, nebo do třetice, jsou-li oblastní rozptyly příliš rozličné, vezmeme t raději ze Studentova rozdělení s počtem stupňů volnosti, rovném největšímu celému číslu, menšímu nebo rovnému výrazu[3])

$$\frac{\left[\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha} - 1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha - 1}\right]^2}{\sum_{\alpha=1}^{L}\left[\frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha} - 1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha - 1}\right]^2 \frac{1}{n_\alpha + 1}} - 2 . \tag{2.7}$$

Jak je vidět z (2.2), k sestrojení $\bar{x}$ je třeba aspoň jednoho pozorování v každé oblasti; jak je vidět z (2.6), k sestrojení intervalu spolehlivosti je třeba aspoň dvou pozorování v každé oblasti. Někdy se doporučuje učinit oblasti stejně velké, vybrat z každé *jedno* pozorování a při sestrojování rozptylů málo odlišné oblasti sdružit (viz [2]). Tím se skutečný rozptyl mírně nadhodnotí, což není příliš na závadu.

Jestliže předmětem našeho zájmu není průměr, ale úhrn $\Sigma\Sigma a_{\alpha i}$, pak vše zůstane v platnosti, jen s tím rozdílem, že v souhlase s rovnicí $\Sigma\Sigma a_{\alpha i} = Na$ musíme odhad i délku intervalu spolehlivosti, stanovené pro průměr, N-krát zvětšit. Zapišme obdržený výsledek ve formě

$$\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} x_{\alpha i} \pm t \sqrt{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha} - 1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha - 1}} , \tag{2.8}$$

z které je patrna jak forma odhadu, tak i intervalu spolehlivosti.

V právě přednesené „lineární" theorii oblastního výběru poněkud zaniká jeden důležitý předpoklad, jehož zanedbání bývá zdrojem těch nejhrubších chyb. Jde o to, že vyjmenované výsledky platí tehdy a jen tehdy, zajímá-li nás průměr vztažený skutečně na *výběrové* jednotky, a ne na některé jednotky drobnější. Velmi často tomu tak nebývá, a zatím co z organisačních důvodů výběrovými jednotkami jsou skupiny osob (školy, obce, rodiny, závody) nebo územní plošky, předmětem našeho zájmu jsou průměry vztažené na jednu osobu nebo na jeden hektar.

[3]) Toto přibližné pravidlo vede k uspokojivému výsledku při všech prakticky se vyskytujících výběrových šetřeních. Prvně o něm uvažoval patrně L. B. Welch v práci "The Generalisation of Student's problem, when several different population variances are involved", Biometrika XXXIV (1947), str. 28—35. Viz také L. B. Welch "Further note on Mrs Aspin's tables and on certain approximations to the tabled function", Biometrika XXXVI (1949), str. 293—296.

R. A. Fisherovu zásadu „jaký výběr — taková statistická analysa" je nutno dodržovat nejen při biologických experimentech, ale i při výběrových šetřeních. Výběr, spočívající v náhodném výběru 100 osob, není tentýž jako výběr, spočívající v náhodném výběru 25 rodin, průměrně po 4 členech. Oba tyto výběry by byly totožné jedině tehdy, kdyby rozdělení společnosti do rodin bylo náhodné, ale to se vzhledem k přítomnosti nejrůznějších sociálních, psychologických a biologických vztahů mezi příbuznými nedá předpokládat. Nahrazení výběru osob výběrem rodin se u různých znaků projeví různě, a alespoň teoreticky jsou myslitelné znaky, vzhledem k nimž budou oba výběry ekvivalentní. Na př. při statistice četnosti obou pohlaví bude výběr rodin vydatnější než výběr osob, neboť v rodiny se sdružují lidé opačného pohlaví. Na druhé straně, při statistice povolání bude výběr rodin méně vydatný než výběr osob, neboť osoby z jedné rodiny mají tendenci mít podobné povolání. Přitom menší vydatnost toho nebo onoho výběru znamená, že při jeho použití bude k dosažení požadované přesnosti nutno vybrat více osob. Tyto úvahy jsou ovšem jenom přibližné. Přesnější vodítko nám poskytne následující „poměrová" teorie oblastního výběru, zahrnující v sobě „lineární" teorii jako speciální případ.

Nuže, sledujme ne jednu, ale dvě veličiny, přiřazující výběrovým jednotkám hodnoty $A_{\alpha i}$ a $B_{\alpha i}$, $\alpha = 1, \ldots, L$, $i = 1, \ldots, N_\alpha$, a naším úkolem bude odhad poměru

$$\varphi = \frac{\Sigma\Sigma A_{\alpha i}}{\Sigma\Sigma B_{\alpha i}}, \quad i = 1, \ldots, N_\alpha, \quad \alpha = 1, \ldots, L. \tag{2.9}$$

Tato úloha pro $B_{\alpha i} = 1$ je totožná s výše probranou úlohou odhadu průměru na jednu výběrovou jednotku. Jestliže $B_{\alpha i}$ označuje počet drobnějších jednotek, obsažených v dané výběrové jednotce, a $A_{\alpha i}$ je součet na těchto drobnějších jednotkách zjištěných hodnot $a_{\alpha ij}$,

$$A_{\alpha i} = \sum_{j=1}^{B_{\alpha i}} a_{\alpha ij}$$

pak dostáváme úlohu o odhadu průměru na jednu jednotku, která je drobnější, než výběrová jednotka. Tím se ovšem řada možných interpretací naší úlohy nekončí: Hodnoty $A_{\alpha i}$ i $B_{\alpha i}$ se mohou týkat téže veličiny jen s tím rozdílem, že byly zjištěny v jiném časovém okamžiku či jinou metodou, a pod. Nejčastěji však budou $A_{\alpha i}$ a $B_{\alpha i}$ úhrnem hodnot, zjištěných na některých drobnějších jednotkách, což byl také důvod, proč jsme k jejich označení použili velkých písmen.

Vyberme nyní opět z každé oblasti náhodně bez opakování n_α výběrových jednotek a získané dvojice pozorování označme $X_{\alpha i}$, $Y_{\alpha i}$, $\alpha = 1, \ldots, L$, $i = 1, \ldots, n_\alpha$ (dvojice $X_{\alpha i}$, $Y_{\alpha i}$ označuje náhodně vybranou dvojici $A_{\alpha j}$, $B_{\alpha j}$).

Sestrojíme-li pro úhrny $\Sigma\Sigma A_{\alpha i}$ a $\Sigma\Sigma B_{\alpha i}$ lineární odhady podle vzoru (2.8) a dosadíme je pak do (2.9), dostaneme tak zvaný *poměrový odhad* f pro φ:

$$f = \frac{\sum \frac{N_\alpha}{n_\alpha} \sum X_{\alpha i}}{\sum \frac{N_\alpha}{n_\alpha} \sum Y_{\alpha i}}, \quad i = 1, \ldots, n_\alpha, \quad \alpha = 1, \ldots, L. \tag{2.10}$$

Sestrojení intervalu se středem f lze provést pomocí následujícího obratu, který pochází od R. A. Fishera, a který nám umožňuje vystačit v podstatě jen s prostředky „lineární" theorie: Nehledíce na to, že φ neznáme, uvažujme pozorování $X_{\alpha i} - \varphi Y_{\alpha i}$ a sestrojme z nich opět podle vzoru (2.8) statistiku[4])

$$\sum \frac{N_\alpha}{n_\alpha} \sum (X_{\alpha i} - \varphi Y_{\alpha i}) = \sum \frac{N_\alpha}{n_\alpha} \sum X_{\alpha i} - \varphi \sum \frac{N_\alpha}{n_\alpha} \sum Y_{\alpha i}. \tag{2.11}$$

Střední hodnota této statistiky bude rovna nule, neboť

$$\mathbf{E}\left(\sum \frac{N_\alpha}{n_\alpha} \sum X_{\alpha i} - \varphi \sum \frac{N_\alpha}{n_\alpha} \sum Y_{\alpha i}\right) = \sum\sum A_{\alpha i} - \varphi \sum\sum B_{\alpha i} = 0,$$

takže s příslušnou spolehlivostí můžeme tvrdit, že

$$\left| \sum \frac{N_\alpha}{n_\alpha} \sum X_{\alpha i} - \varphi \sum \frac{N_\alpha}{n_\alpha} \sum Y_{\alpha i} \right| \leq t \sqrt{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha} - 1\right) n_\alpha \psi_\alpha^2 \frac{n_\alpha}{n_\alpha - 1}}, \tag{2.12}$$

kde

$$n_\alpha \psi_\alpha^2 = \sum_{i=1}^{n_\alpha} (X_{\alpha i} - \varphi Y_{\alpha i})^2 - \frac{1}{n_\alpha}\left[\sum_{i=1}^{n_\alpha} (X_{\alpha i} - \varphi Y_{\alpha i})\right]^2, \quad \alpha = 1, \ldots, L. \tag{2.13}$$

Nahradíme-li v (2.13) neznámé φ pomocí poměrového odhadu f, dostaneme

$$n_\alpha s_\alpha^2 = \sum_{i=1}^{n_\alpha} (X_{\alpha i} - f Y_{\alpha i})^2 - \frac{1}{n_\alpha}\left[\sum_{i=1}^{n_\alpha} (X_{\alpha i} - f Y_{\alpha i})\right]^2, \quad \alpha = 1, \ldots, L. \tag{2.14}$$

Zanedbáme-li rozdíl mezi ψ_α^2 a s_α^2,[5]) můžeme očekávat, že se stejnou pravděpodobností jako nerovnost (2.12) bude splněna i nerovnost

$$\left| \sum \frac{N_\alpha}{n_\alpha} \sum X_{\alpha i} - \varphi \sum \frac{N_\alpha}{n_\alpha} \sum Y_{\alpha i} \right| \leq t \sqrt{\sum \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha} - 1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha - 1}},$$

[4]) Statistikou rozumíme jakoukoliv funkci pozorování, ať už běží o nějaký speciální odhad nebo ne.

[5]) Aby mělo šetření praktickou cenu, musí být projektováno tak, aby relativní rozdíl mezi φ a f byl s velkou spolehlivostí zanedbatelný; potom však je zanedbatelný i rozdíl mezi ψ_α^2 a s_α^2. R. A. Fisher doporučuje získat interval spolehlivosti pro φ řešením kvadratické nerovnosti (2.12).

což dává pro φ tento odhad a interval spolehlivosti:

$$\frac{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} X_{\alpha i}}{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} Y_{\alpha i}} \pm \frac{t \sqrt{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha} - 1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha - 1}}}{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} Y_{\alpha i}} . \qquad (2.15)$$

Součinitel t lze i zde stanovit na základě úvah, jimiž jsme doprovodili interval spolehlivosti (2.6). Samozřejmě, při eventuálním použití zlomku (2.7), je nutno s_α^2 počítat podle (2.14).

Prostřednictvím poměrového odhadu lze často velmi jednoduše a účinně využít výsledků úplných šetření pro zpřesnění výběrových šetření, které po nich následují. Známe-li na př. jednotlivé hodnoty, a tedy i úhrn $\Sigma\Sigma B_{\alpha i}$ hodnot určité veličiny, pak odhad a interval spolehlivosti (2.15) pro φ nám bezprostředně poskytuje odhad a délku intervalu spolehlivosti pro $\Sigma\Sigma A_{\alpha i} = \varphi\Sigma\Sigma B_{\alpha i}$:

$$\sum_{\alpha=1}^{L} \sum_{i=1}^{N_\alpha} B_{\alpha i} \frac{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} X_{\alpha i}}{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} Y_{\alpha i}} \pm t \frac{\sum_{\alpha=1}^{L} \sum_{i=1}^{N_\alpha} B_{\alpha i}}{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} Y_{\alpha i}} \sqrt{\sum_{\alpha=1}^{L} \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha} - 1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha - 1}} , \qquad (2.16)$$

při čemž $n_\alpha s_\alpha^2$ jest počítati podle vzorce (2.14). Tento výsledek je sice složitější než odpovídající výsledek (2.8) při použití lineárního odhadu ($x_{\alpha i} = X_{\alpha i}$), avšak v případech, kdy hodnoty $A_{\alpha i}$ jsou až na náhodné odchylky přímo úměrné hodnotám $B_{\alpha i}$, bývá tento přírůstek počtářské práce plně vyvážen tím, že se délka intervalu spolehlivosti podstatně zkrátí. Tak je tomu na př. tehdy, jsou-li $A_{\alpha i}$ celkové výnosy určité plodiny na územních ploškách a $B_{\alpha i}$ jsou rozlohy těchto plošek, nebo je-li $A_{\alpha i}$ součtem hodnot, zjištěných na $B_{\alpha i}$ drobnějších jednotkách a pod.

Příklad. V tabulce jsou uvedeny údaje, zjištěné ve 20 závodech, vybraných ze 3 oblastí. V prvním sloupci jsou vybrané závody očíslovány, a to tak, že první cifra označuje oblast (1—2—3) a druhá cifra je pořadové číslo vybraného závodu uvnitř dané oblasti. Takové číslování je výhodné zejména při strojovém zpracování na děrných štítcích. Jak je vidět, z první oblasti bylo vybráno 9 závodů, z druhé 7 závodů a ze třetí 4, t. j. $n_1 = 9$, $n_2 = 7$, $n_1 = 4$. Čísla v druhém sloupci tabulky ($Y_{\alpha i}$) označují celkový počet zaměstnanců a čísla v třetím sloupci ($X_{\alpha i}$) označují kolik je z toho mužů ve věku 20—29 let.

Úkolem jest sestrojit odhad a interval spolehlivosti (spolehlivost = 0,95) pro procento mužů ve věku 20—29 mezi všemi zaměstnanci. Jelikož výběrovými jednotkami jsou závody a nikoliv jednotlivé osoby, bude nutno úlohu řešit pomocí „poměrové" theorie.

Tabulka

	Y	X	$0{,}266Y$	$X - 0{,}266Y$	$(X - 0{,}266Y)^2$
11	321	76	85	— 9	81
12	116	34	31	3	9
13	486	81	129	—48	2304
14	420	84	112	—28	784
15	94	10	25	—15	225
16	213	60	57	3	9
17	286	41	76	—35	1225
18	650	159	173	—14	196
19	501	81	133	—52	2704
Σ	3087	626	821	—195	7537
21	356	103	95	8	64
22	883	312	235	77	5929
23	82	18	22	— 4	16
24	574	126	153	—27	729
25	327	93	87	6	36
26	205	91	54	37	1369
27	188	42	50	— 8	64
Σ	617	785	696	89	8207
31	1063	408	283	125	15625
32	509	256	135	121	14641
33	728	260	194	66	4356
34	493	223	131	92	8464
Σ	2793	1147	743	404	43086

Abychom mohli sestrojit poměrový odhad f podle vzorce (2.10), stačí **znát** tak zvané výběrové intervaly N_α/n_α v jednotlivých oblastech a součty hodnot $Y_{\alpha i}$ a $X_{\alpha i}$ pro jednotlivé oblasti. V daném případě bylo $N_1/n_1 = 30$, $N_2/n_2 = 20$, $N_3/n_3 = 10$ a oblastní součty jsou dány ve zvlášť pro ně vynechaných mezerách v tabulce. To znamená, že odhad zajímajícího nás procenta je roven

$$f = \frac{30 \,.\, 626 + 20 \,.\, 785 + 10 \,.\, 1147}{30 \,.\, 3087 + 20 \,.\, 2615 + 10 \,.\, 2793} = 0{,}266 \,.$$

Poněkud větší práci dá sestrojení intervalu spolehlivosti. Především je nutno vypočíst násobky oblastních rozptylů. Ve čtvrtém až šestém sloupci tabulky je proveden výpočet součtů $\sum_{i=1}^{n_\alpha} (X_{\alpha i} - fY_{\alpha i})$ a $\sum_{i=1}^{n_\alpha} (X_{\alpha i} - fY_{\alpha i})^2$, při čemž necelá čísla byla obvyklým způsobem zaokrouhlována na celky. Dosadíme-li tyto součty do vzorce (2.14), dostáváme

$$\begin{aligned} 9s_1^2 &= 7\,537 - \tfrac{1}{9}(-195)^2 = 3312 \,, \\ 7s_2^2 &= 8\,027 - \tfrac{1}{7}(89)^2 = 7075 \,, \\ 4s_3^2 &= 43\,086 - \tfrac{1}{4}(404)^2 = 2282 \,. \end{aligned}$$

To znamená, že

$$\sum_{\alpha=1}^{3} \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha}-1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha-1} = 30 \,.\, 29 \,.\, 3312 \,.\, \tfrac{9}{8} + 20 \,.\, 19 \,.\, 7075 \,.\, \tfrac{7}{6} + \\ + 10 \,.\, 9 \,.\, 2282 \,.\, \tfrac{4}{3} = (324 + 314 + 27)\, 10^4 \,, \qquad (2.17)$$

a tedy

$$\frac{\sqrt{\sum_{\alpha=1}^{3} \frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha}-1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha-1}}}{\sum_{\alpha=1}^{3} \frac{N_\alpha}{n_\alpha} \sum_{i=1}^{n_\alpha} Y_{\alpha i}} = \frac{\sqrt{(324 + 314 + 27) \,.\, 10^4}}{172\,840} = 0{,}015 \,.$$

Dosadíme-li tyto výsledky do (2.15), vidíme, že odhad a délka intervalu spolehlivosti jsou v daném případě rovny

$$0{,}266 \pm t \,.\, 0{,}015 \,.$$

Zbývá nyní již jen určit počet stupňů volnosti pro součinitele t. Jak je vidět z tvaru vzorce (2.7), je možno do něho místo čísel $\frac{N_\alpha}{n_\alpha}\left(\frac{N_\alpha}{n_\alpha}-1\right) n_\alpha s_\alpha^2 \frac{n_\alpha}{n_\alpha-1}$ dosadit jakákoliv jim úměrná čísla. Během výpočtů (2.17) jsme však již taková úměrná čísla našli, a to: 324, 314 a 27. To znamená, že za počet stupňů volnosti lze zvolit celky z

$$\frac{(324 + 314 + 27)^2}{(324)^2/10 + (314)^2/8 + (27)^2/5} - 2 = 17{,}2 \,,$$

t. j. 17. Pro 17 stupňů volnosti a spolehlivosti 0,95 máme $t = 2{,}11$. Konečný výsledek tedy zní

$$0{,}266 \pm 0{,}032 \,.$$

Skutečné procento tedy se spolehlivostí 0,95 leží někde mezi 23% a 30%.

3. Optimální rozvržení výběru do oblastí

Je-li jednotkový náklad na sběr a zpracování zjišťovaného údaje v oblasti α roven c_α, pak budou celkové náklady rovny

$$C = \sum_{\alpha=1}^{L} n_\alpha c_\alpha \,. \qquad (3.1)$$

Stanovme n_α tak, abychom minimalisovali součin

$$C\left(N^2 \mathbf{D}^2 \bar{x} + \sum_{\alpha=1}^{L} N_\alpha \sigma_\alpha^2\right) = \left(\sum_{\alpha=1}^{L} c_\alpha n_\alpha\right)\left(\sum_{\alpha=1}^{L} \frac{N_\alpha^2 \sigma_\alpha^2}{n_\alpha}\right) , \qquad (3.2)$$

kde za $\boldsymbol{D}^2\bar{x}$ jsme pro jednoduchost dali aproximaci (2.4). Řešení řečeného minima podle n_α nám dá *optimální rozvržení* výběru do oblastí vzhledem k nákladové funkci (3.1), neboť při daném $\boldsymbol{D}^2\bar{x}$ povede k minimálnímu C a při daném C k minimálnímu $\boldsymbol{D}^2\bar{x}$.

Aplikací Buňakovského nerovnosti (viz 5) na pravou stranu rovnice (3.2) dostáváme, že

$$C\left(N^2\mathbf{D}^2\bar{x} + \sum_{\alpha=1}^{L} N_\alpha \sigma_\alpha^2\right) \geq \left(\sum_{\alpha=1}^{L} \sigma_\alpha N_\alpha \sqrt{c_\alpha}\right)^2 \tag{3.3}$$

a že rovnost nastává tehdy a jen tehdy, je-li

$$n_\alpha c_\alpha = \lambda^2 \frac{N_\alpha^2}{n_\alpha} \sigma_\alpha^2 , \quad \alpha = 1, \ldots, L ,$$

čili

$$n_\alpha = \lambda \frac{N_\alpha \sigma_\alpha}{\sqrt{c_\alpha}} , \quad \alpha = 1, \ldots, L . \tag{3.4}$$

Řečeno slovy, z každé oblasti při optimálním rozvržení vybíráme tím více jednotek, čím je v ní sběr lacinější a čím jsou v ní hodnoty variabilnější.

Konstantu λ určujeme buď z rovnice (2.4) nebo (3.1) podle toho, zda východiskem pro určení rozsahu výběru je požadovaný rozptyl $\mathbf{D}^2\bar{x}$ nebo přípustné náklady C:

$$\lambda = \frac{\sum_{\alpha=1}^{L} \sigma_\alpha N_\alpha \sqrt{c_\alpha}}{N^2\mathbf{D}^2\bar{x} + \sum_{\alpha=1}^{L} N_\alpha \sigma_\alpha^2} \tag{3.5}$$

nebo

$$\lambda = \frac{C}{\sum_{\alpha=1}^{L} \sigma_\alpha N_\alpha \sqrt{c_\alpha}} . \tag{3.6}$$

Odtud a z (3.4) je vidět, že v případě, kdy rozsah výběru je určován rozptylem $\mathbf{D}^2\bar{x}$, stačí ke stanovení optimálního rozvržení znát místo samotných čísel c_α kterákoliv čísla jim úměrná. Podobně v případě, kdy rozsah výběru je určován náklady C, stačí ke stanovení optimálního znát místo samotných čísel σ_α kterékoliv čísla jim úměrná.

Jedná-li se o poměrový odhad, a nahradíme-li řešení, minimalisující při dané nákladové funkci $\mathbf{D}^2 f$, řešení, které minimalisuje rozptyl statistiky (2.11), lze vzorec (3.4) použít s tou jedinou změnou, že σ_α^2 počítáme ne podle (2.1), ale podle vzorce

$$\sigma_\alpha^2 = \frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} (A_{\alpha i} - \varphi B_{\alpha i})^2 - \left(\frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} (A_{\alpha i} - \varphi B_{\alpha i})\right)^2 , \quad \alpha = 1, \ldots, L . \tag{3.7}$$

Výpočet λ pro dané $\mathbf{D}^2 f$ provádíme na základě aproximace[6])

[6]) Viz na př. [1] nebo [2] nebo [3].

$$\mathbf{D}^2 f = \frac{\sum_{\alpha=1}^{L} N_\alpha^2 \left(1 - \frac{n_\alpha}{N_\alpha}\right) \frac{\sigma_\alpha^2}{n_\alpha}}{\left(\sum_{\alpha=1}^{L} \sum_{i=1}^{N_\alpha} B_{\alpha i}\right)^2}, \tag{3.8}$$

kde σ_α^2 jest nutno chápat ve smyslu definice (3.7). Jestliže sem z (3.4) dosadíme a vyřešíme podle λ, dostáváme

$$\lambda = \frac{\sum_{\alpha=1}^{L} \sigma_\alpha N_\alpha \sqrt{c_\alpha}}{\left(\sum_{\alpha=1}^{L} \sum_{i=1}^{N_\alpha} B_{\alpha i}\right)^2 \mathbf{D}^2 f + \sum_{\alpha=1}^{L} N_\alpha \sigma_\alpha^2}. \tag{3.9}$$

V případě, že je dáno C, počítáme λ i pro poměrový odhad pomocí vzorce (3.6), v němž σ_α jest nutno chápat podle (3.7).

Modifikace výpočtu λ při odhadu úhrnu jsou zřejmé a nebudeme se o nich rozepisovat.

Příklad. Máme-li $c_1 : c_2 : c_3 = 1 : 4 : 4$, $\sigma_1 = 5$, $\sigma_2 = 3$, $\sigma_3 = 2$, $\mathbf{D}\bar{x} = 0{,}2$, $N_1 = 621$, $N_2 = 468$, $N_3 = 1046$, pak budou výpočty postupovat takto: nejdříve stanovíme λ

$$\lambda = \frac{\sum_{\alpha=1}^{3} \sigma_\alpha^2 N_\alpha \sqrt{c_\alpha}}{N^2 \mathbf{D}^2 \bar{x} + \sum_{\alpha=1}^{3} N_\alpha \sigma_\alpha^2} =$$

$$= \frac{5 \,.\, 621 \sqrt{1} + 3 \,.\, 368 \sqrt{4} + 2 \,.\, 1046 \sqrt{4}}{(621 + 468 + 1046)^2 \,.\, (0{,}2)^2 + (621 \,.\, 5^2 + 468 \,.\, 3^2 + 1046 \,.\, 2^2)} = 0{,}049$$

a potom n_α

$$n_1 = \lambda \frac{N_1 \sigma_1}{\sqrt{c_1}} = 0{,}049 \frac{621 \,.\, 5}{\sqrt{1}} = 152\,,$$

$$n_2 = \lambda \frac{N_2 \sigma_2}{\sqrt{c_2}} = 0{,}049 \frac{468 \,.\, 3}{\sqrt{4}} = 34\,,$$

$$n_3 = \lambda \frac{N_3 \sigma_3}{\sqrt{c_3}} = 0{,}049 \frac{1046 \,.\, 2}{\sqrt{4}} = 51\,.$$

4. Náhodné rozvržení výběru do oblastí

Někdy se stává, že identifikaci oblastí není možno provést v době, kdy se provádí výběr, ale až v době, kdy se už sbírají pozorování a sestrojují odhady. Na př. pořídit si subjektivní (znalecké) odhady výnosů určité plodiny, a tedy i roztřídit podle nich zkoumaný soubor osetých ploch na oblasti, můžeme až v létě, kdežto výběr ploch pro výběrové šetření je nutno připravit již na jaře. V takových případech můžeme ze základního souboru jako celku, t. j. bez ohledu na jakékoliv oblasti, provést náhodný výběr bez opakování, a teprve potom, v příhodném okamžiku, zjistit rozsahy oblastí N_α a příslušnost pozorování do

oblastí, a na základě toho sestrojit kterýkoliv z probraných odhadů. Délku intervalu spolehlivosti přitom můžeme stanovit tak, jako kdybychom ne náhodně, ale záměrně vybrali z oblastí tolik jednotek kolik jsme jich vybrali, neboť — jak lze snadno dokázat, „podmíněný" náhodný výběr při fixovaných rozsazích výběru v jednotlivých oblastech, je totožný s odpovídajícím oblastním výběrem.

Rozvržení výběru do oblastí, získané pomocí náhodného výběru, nazývejme náhodným rozvržením. Nebude-li počet oblastí velký, řekněme $L \leq 6$, a bude-li střední počet jednotek vybraných při náhodném rozvržení z dané oblasti pro každou oblast roven aspoň 10, pak bude takřka vždy platit $n_\alpha \geq 2$, takže jak odhady tak i intervaly spolehlivosti, probrané v paragrafu 2, bude možno sestrojit. Kdyby se přece jen někdy stalo, že by na některou oblast připadla méně jak dvě pozorování, pak by patrně bylo nejlépe vybrat z této oblasti potřebná pozorování dodatečně resp. připojit ji k některé „sousední" oblasti.

Za podmínek, popsaných v předešlém odstavci, v důsledku zákona velkých čísel, nebude obvykle náhodné rozvržení výběru daleko od proporcionálního, a v souhlase s tím i dosažená přesnost se nebude příliš lišit od přesnosti, dosahované pri *proporcionálním* oblastním výběru, kdy

$$n_\alpha = \lambda N_\alpha , \quad \alpha = 1, \ldots, L$$

a

$$\mathbf{D}^2 \bar{x} = \frac{\sigma_\omega^2}{n} \frac{N - n}{N - L} .$$

Přitom $n = \Sigma n_\alpha$ je celkový rozsah výběru a σ_ω^2 je rozptyl uvnitř oblastí

$$\sigma_\omega^2 = \sum_{\alpha=1}^{L} \frac{\sigma_\alpha^2 N_\alpha}{N} .$$

Přibližný výpočet ukazuje, že při náhodném rozvržení je rozptyl $\bar{x}$ v průměru $[1 + (L - 1)(1 - n/N)/n]$-kráte větší než při proporcionálním rozvržení.

5. Buňakovského nerovnost

Nakonec si uveďme „pravděpodobnostní" důkaz Buňakovského nerovnosti spolu s podmínkou, kdy v ní nastává případ rovnosti. Jak známo z theorie pravděpodobnosti, pro jakákoliv čísla $z_1, z_2, \ldots, z_n$ a pravděpodobnosti p_1, $p_2, \ldots, p_n$, platí tato průhledná algebraická identita a z ní vyplývající nerovnost:

$$\sum_{i=1}^{h} z_i^2 p_i = \left(\sum_{i=1}^{h} z_i p_i \right)^2 + \sum_{i=1}^{h} \left(z_i - \sum_{i=1}^{h} z_i p_i \right)^2 p_i \geq \left(\sum_{i=1}^{h} z_i p_i \right)^2 . \tag{5.1}$$

Jestliže jsou všechny pravděpodobnosti kladné, pak v této nerovnosti nastává případ rovnosti zřejmě tehdy a jen tehdy, když

$$z_i = \lambda , \quad i = 1, .., h , \tag{5.2}$$

kde λ je některá konstanta. Mějme nyní dvě libovolné posloupnosti čísel $a_1, a_2, \ldots, a_h$ a $b_1, b_2, \ldots, b_h$, vesměs různých od nuly. Položíme-li

$$z_i = \frac{a_i}{b_i}, \quad p_i = \frac{b_i^2}{\sum_{j=1}^{h} b_j^2}, \quad i = 1, \ldots, h$$

a dosadíme do nerovnosti (5.1), dostáváme nerovnost

$$\frac{\sum \left(\frac{a_i}{b_i}\right)^2 b_i^2}{\sum b_j^2} \geq \left(\frac{\sum \frac{a_i}{b_i} b_i^2}{\sum b_j^2}\right)^2,$$

to jest Buňakovského nerovnost

$$\left(\sum_{i=1}^{h} a_i^2\right)\left(\sum_{i=1}^{h} b_i^2\right) \geq \left(\sum_{i=1}^{h} a_i b_i\right)^2. \tag{5.3}$$

Nutná a postačující podmínka pro nastoupení případu rovnosti v této nerovnosti vyplývá z podmínky (5.2) v tomto tvaru:

$$a_i = \lambda b_i, \quad i = 1, \ldots, h. \tag{5.4}$$

Z platnosti nerovnosti (5.3) pro c_i a d_i, $i = 1, \ldots, h$, vesměs různá od nuly snadno vyplývá, že platí pro jakákoliv c_i a d_i. Také podmínka (5.4) zůstane nutnou a postačující, jakmile v každé posloupnosti je aspoň jedno číslo různé od nuly.

Při odvození nerovnosti (3.3) bylo položeno $a_\alpha = \sqrt{n_\alpha c_\alpha}$ a $b_\alpha = \sqrt{N_\alpha^2 \sigma_\alpha^2 / n_\alpha}$.

KNIŽNÍ LITERATURA V ČSR

[1] *W. E. Deming:* Some Theory of Sampling, New York, John Wiley and Sons, London, Chapman Hall, 1950.

[2] *W. G. Madow, M. H. Hansen, W. N. Hurwitz:* Sample Survey Methods and Theory, New York 1951, I. Methods and Applications, II. Theory.

[3] *J. Hájek:* Theorie výběrových šetření, scripto na Vysoké škole ekonomické v Praze, 1955.

Резюме

ВЫБОРКА ПО ГРУППАМ (СЕРИЯМ)

ЯРОСЛАВ ХАЕК (Jaroslav Hájek)

(Поступило в редакцию 3/XI 1955 г.)

Выборкой по группам называется последовательность независимых случайных выборок, совершаемых в отдельных группах, причем группы — любые непересекающиеся части основной совокупности. В этой работе рассмотрены следующие вопросы и видоизменения:

1. Линейная и „пропорцийная“[1]) оценки и соответствующие доверительные интервалы.

2. Оптимальные численности выборок из отдельных групп, когда издержки в разных группах разны.

3. Случайные численности выборок из отдельных групп.

Доверительный интервал, соответствующий „пропорцийной“ оценке получен с помощью видоизмененного метода Р. А. Фишера. Оптимальные численности выборок получены с помощью неравенства Буняковского. Случайными численностями выборок можно воспользоваться тогда, когда численности групп не известны во время выборки, но известны во время исчисления оценок.

Summary

STRATIFIED SAMPLING

JAROSLAV HÁJEK

(Received November 3, 1955.)

In this paper the following problems arising with stratified sampling are considered:

1. Linear and ratio estimates and corresponding confidence intervals.

2. Optimum allocation of sample to strata with respect to a given cost function.

3. Random allocation of sample to strata.

Confidence interval corresponding to ratio estimate is derived by means of a modified method of R. A. Fisher. Optimum allocation is derived by means of inequality of Buňakovski. Random allocation of sample to strata is of importance, when strata are not identifiable at the time of sampling, but are so at the time of computing estimates.

[1]) По английски „ratio estimate“.