FORUM STATISTICUM SLOVACUM Vol XVII (2021) No 1 pp 70-82

Možnosti tvorby náhodných výběrů použitím softwarového systému SAS
Possibilities of creating random samples using the SAS software system

Roman Pavelka
Štatistický úrad Slovenskej republiky, Odbor metód štatistických zisťovaní, Lamačská cesta 3, 840 05 Bratislava, Slovenská republika
Statistical Office of the Slovak Republic, Statistical Surveys and Methodology Department, Lamačská cesta 3, 840 05 Bratislava, Slovak Republic
Roman.Pavelka@statistics.sk

Abstrakt: *Tento článek ilustruje některé z mnoha algoritmů výběru vestavěných do procedury SURVEYSELECT programového prostředí SAS, a to zejména těch algoritmů výběru, které se týkají složitějších šetření. Mezi takové algoritmy výběru patří výběry systematické, výběry s pravděpodobností úměrnou velikosti vybírané jednotky, výběry stratifikované a skupinové (klastrové). Primárním cílem příspěvku je poskytnout základní informace o tom, proč se tyto techniky používají v praxi, a demonstrovat jejich použití na příkladech syntaxe. Nejedná se tedy o postup, jak navrhnout statisticky efektivní vzorek - tomuto předmětu jsou věnovány celé učebnice. V rámci stratifikovaného náhodného výběru článek pojedná o několika strategiích alokace výběrů - konkrétně o proporcionální, Neymanově a optimální alokaci. Příspěvek končí několika příklady, které ukazují, jak lze pomocí procedury SURVEYSELECT zvládnout vyskytující se problémy s návrhem jednodušších i složitějších výběrů, jako jsou alternativní metody náhodného výběru napříč vrstvami a vícestupňový skupinový (klastrový) výběr.*

Abstract: *This paper illustrates some of the many selection algorithms built into the SURVEYSELECT procedure of the SAS programming environment, especially those selection algorithms that involve complex investigations. Such selection algorithms include systematic selections, selections with a probability proportional to the size of the selected unit, stratified and group (cluster) selections. The primary goal of the paper is to provide basic information about why these techniques are used in practice and to demonstrate their use on syntax examples. It is therefore not a procedure for designing a statistically effective sample - entire textbooks are devoted to this subject. As part of the stratified random selection, the article discusses several strategies for allocating selections - namely proportional, Neyman and optimal allocation. The paper concludes with several examples that show how to use the SURVEYSELECT procedure to handle certain common selection design problems, such as alternative methods of random selection across layers and multilevel group (cluster) selection.*

Kľúčové slová: *alokace, náhodný výběr, SAS, SURVEYSELECT*

Key words: *allocation, random sample, SAS, SURVEYSELECT*

1 Úvod

Účelem tohoto příspěvku je zdůraznit některé možnosti procedury SURVEYSELECT v sytému SAS při vytváření náhodného výběru. I když různé přístupy k algoritmům náhodného výběru budou uvedeny s nezbytným

množstvím teoretických předpokladů a diskusí o tom, proč a jak jsou používány v praxi, záměrem článku je demonstrovat nezbytnou syntaxi procedury SURVEYSELECT k realizaci různých technik výběru. Pro úvod do problematiky a koncepce náhodných výběrů může čtenář nahlédnout do Kaltona (1983). Čtenář, který hledá podrobnější zpracování základní teorie komplexních výběrových šetření, je odkázán na jednu z mnoha vynikajících učebnic na toto téma, jako jsou Kish (1965), Cochran (1977), Scheaffer et al. (1996), Lohr (1999) a Valliant et al. (2013). Úvod do problematiky návrhů statisticky efektivních vzorků poskytuje článek Frankovič a Stankovičová (2020).

Článek je strukturován do několika kapitol. První kapitola je věnována základním technikám náhodného výběru, a to jednoduchému náhodnému výběru, systematickému náhodnému výběru a výběru s pravděpodobností úměrnou velikosti vybírané jednotky. V následujících částech jsou diskutovány pokročilé techniky náhodných výběrů, kdy pomocí příkazu STRATA, resp. CLUSTER v rámci procedury SURVEYSELECT lze náhodně vybírat jednotky uvnitř 2 nebo více nezávislých strat (vrstev) definovaných v opoře výběru, resp. náhodně vybírat celé skupiny (klastry) jednotek.

Pro ukázky různých technik náhodného výběru je použit soubor dat HEART z knihovny SASHELP, který je systémovou součástí softwarového systému SAS (SAS Institute Inc., 2018). Soubor dat HEART obsahuje údaje o 5 209 pacientech s 17 proměnnými (viz tabulka 1).

Tab. 1 Proměnné souboru HEART *(Zdroj: SAS Institute Inc., 2018)*

#	Proměnná	Popis
1	Status	Živý / Mrtvý
2	DeathCause	Příčina smrti
3	AgeCHDdiag	Věk vzniku nemoci
4	Sex	Pohlaví
5	AgeAtStart	Věk na začátku sledování
6	Height	Výška
7	Weight	Váha
8	Diastolic	Diastolický tlak
9	Systolic	Systolický tlak
10	Smoking	Denní počet cigaret
11	AgeAtDeath	Věk úmrtí
12	Cholesterol	Míra cholesterolu v krvi
13	Chol_Status	Kategorie podle cholesterolu
14	BP_Status	Kategorie podle krevního tlaku
15	Weight_Status	Kategorie podle váhy
16	Smoking_Status	Kategorie podle kouření

S výše popsanými daty budou postupně realizovány návrhy výběrů, zpravidla každý (mimo výběrů skupinových) o předem dané velikosti n = 500. Zvolenou

velikost vzorku lze pro jednoduchost považovat za maximální velikost výběru povolenou rozpočtem firmy pro sběr dat v rámci průzkumu trhu.

2 Základní výběrové techniky

2.1 Jednoduchý náhodný výběr

Jak provést základní (také i výchozí) metodu v proceduře SURVEYSELECT, jednoduchý náhodný výběr bez vracení (angl. zkratka SRSWOR)[9], ilustruje níže uvedená příkazová syntaxe:

```
proc surveyselect data=heart out=sample_SRSWOR sampsize=500
seed=11111;
run;
```

Parametr DATA odkazuje na oporu výběru, příkaz OUT= označuje výstupní datový soubor, který obsahuje výsledný výběr. Požadovaná velikost výběru n = 500 je hodnotou parametru SAMPSIZE. Parametru SEED se přiřadí nezáporné celé číslo menší než 2^{31}-1. Použití stejné hodnoty parametru SEED zajistí, že bude vybrán přesný vzorek i při opakujících se náhodných výběrech (za předpokladu stejného nastavení ostatních parametrů procedury). Výstupem z procedury SURVEYSELECT je soubor SAMPLE_SRSWOR skládající se z 500 pozorování náhodně získaných z opory v datovém souboru HEART. V SAS reportu v SAS Enterprise Guide procedura SURVEYSELECT vygeneruje stručný přehled o provedené úloze. Tento přehled je ilustrován tabulkou 2. V přehledu se zobrazí všechny použité parametry včetně pravděpodobnosti výběru a výběrové váhy.

Tab. 2 Přehled o činnosti procedury SURVEYSELECT *(Zdroj: vlastní zpracování)*

The Status and DeathCause Variables

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	HEART
Random Number Seed	11111
Sample Size	500
Selection Probability	0,095988
Sampling Weight	10,418
Output Data Set	SAMPLE_SRSWOR

[9] SRSWOR je zkratkou pro anglické označení jednoduchého náhodného výběru bez vracení (simple random sampling without replacement).

Podobnou syntaxí příkazů lze zvolit jednoduchý náhodný výběr s vracením (angl. zkratka SRSWR)[10]. Toto lze dosáhnout nastavením parametru METHOD na hodnotu URS[11]. Příkladem pro náhodný výběr s vracením pomocí procedury SURVEYSELECT může být následující příkaz:

```
proc surveyselect data=heart out=sample_SRSWR sampsize=500
seed=11111 method=URS outhits;
run;
```

Je-li uživatelem specifikován náhodný výběr s vracením, skládá se výstupní náhodně vybraný soubor z řádků pro každý jedinečný náhodně vybraný datový záznam a číselné proměnné s názvem NUMBERHITS označující, kolikrát byl daný datový záznam vybrán. Parametrem OUTHITS se dosáhne toho, aby počet jedinečně vybraných datových záznamů odpovídal požadované velikosti výběru.

2.2 Systematický náhodný výběr

Dalším široce používaným přístupem v aplikovaném výzkumu je systematický náhodný výběr. Základní myšlenkou je vybrat každou *k*-tou jednotku do výběru, kde *k* je typicky celé číslo (Taylor, 2013). Tato metoda je zvláště užitečná ve scénářích, kde by bylo velmi obtížné nebo zcela nemožné zkonstruovat oporu pro náhodný výběr. Příkladem může být výzkum spokojenosti zákazníků s obchodem s lahůdkami. Je nepravděpodobné, že by existoval vyčerpávající seznam zákazníků obchodu. Proto jedinou vhodnou metodou výběru je zapojení do výzkumu každého *k*-tého zákazníka vycházejícího z obchodu (navíc lze také i náhodně vybírat dny a časy, ve kterých se tento výzkum realizuje).

Tuto techniku výběru lze samozřejmě použít také v případě, když existuje dobře definovaná opora výběru. Následující příklad ukazuje základní syntaxi pro výběr systematického vzorku n = 500 pacientů. Systematický náhodný výběr inicializuje v proceduře SURVEYSELECT nastavení parametru METHOD=SYS. Procedura vypočítá vzorkovací interval k jako N/n, kde N je odvozeno z počtu pozorování v opoře výběru. Pokud k není explicitně celé číslo, použije se interval vyjádřený podílem N/n, takže se vrátí přesná požadovaná velikost výběru (další podrobnosti viz SAS Institute Inc., 2020). Pro jednoduchost se bude počítat s celočíselným intervalem k = 5000/500 = 10. Procedura začíná náhodným výběrem pozorování v opoře výběru mezi 1. a *k*-tým pozorováním. Označí-li se toto pozorování r, potom výsledný výběrový soubor bude složen z pozorování r-tého a (r+k)-tého, (r+2k)-tého, atd. až do konce opory výběry. Například, bude-li jako první vybrán 4. pacient, výsledný výběrový soubor by měl sestávat ze 4., 14.,

[10] SRSWR je zkratkou pro anglické označení jednoduchého náhodného výběru s vracením (simple random sampling with replacement).

[11] URS je zkratkou pro anglické označení jednoduchého náhodného výběr (unrestricted random sampling) a v systému SAS odpovídá jednoduchému náhodnému výběru s vracením.

24., 34.,..., 4994., … pacienta. Ukázka syntaxe příkazu pro systematický náhodný výběr je uvedena níže:

```
proc surveyselect data=heart out=sample_SYS sampsize=500
seed=11111 method=SYS;
run;
```

Pokud existuje určitá periodicita v uspořádání opory, která se shoduje se vzorkovacím intervalem, může dojít ke snížení reprezentativnosti systematického náhodného výběru. V zájmu zajištění správného výsledku systematického náhodného výběru, se opora výběru musí uspořádat podle 1 nebo více kontrolních proměnných. Toto uspořádání opory lze provést ještě buď před vstupem do procedury SURVEYSELECT pomocí procedury SORT anebo pomocí parametru CONTROL v proceduře SURVEYSELECT. Bude-li například parametr CONTROL=Height (výška pacienta), potom před samotným systematickým náhodným výběrem procedura uspořádá pacienty v opoře podle jejich výšky.

2.3 Náhodný výběr s pravděpodobnostmi úměrnými velikosti jednotky

Další populární algoritmus výběru je výběr s pravděpodobnostmi úměrnými velikosti - zkráceně výběr PPS[12]. K hlavním výhodám tohoto způsobu náhodného výběru je schopnost kontrolovat variabilitu odhadů z takto realizovaných výběrů. Například pokud výběrová opora obsahuje pomocnou proměnnou, která silně koreluje s klíčovou sledovanou proměnnou, náhodný výběr s pravděpodobností úměrnou velikosti této korelované pomocné proměnné může dramaticky snížit variabilitu odhadovaného úhrnu.

Specifikace parametru METHOD=PPS v proceduře SURVEYSELECT vyvolá mechanismus výběru PPS bez vracení (angl. zkratka PPSWOR[13]). Číselná míra velikosti (pomocná silně korelovaná proměnná se sledovanou výstupní proměnnou) musí být uvedena v parametru SIZE=. K realizaci náhodného výběru PPSWOR pro stejnou cílovou velikost výběru n = 500 z opory HEART je příkazová syntaxe uvedena níže.

```
proc surveyselect data=heart out=sample_PPSWOR sampsize=500
seed=11111 method=PPS;
     size cholesterol;
run;
```

Procedura SURVESELECT po realizaci náhodného výběru metodou PPS vypočítá a automaticky uloží ke každé vybrané jednotce ve výběrovém souboru

[12] PPS je zkratkou pro anglické označení pro náhodný výběr s pravděpodobnostmi úměrnými velikosti (probability proportional to size sampling).

[13] PPSWOR je zkratkou pro anglické označení pro náhodný výběr s pravděpodobnostmi úměrnými velikosti bez vracení (probability proportional to size sampling without replacement).

pravděpodobnost výběru a odpovídající výběrovou váhu (převrácenou hodnotu pravděpodobnosti výběru). Ve výběrovém souboru se po provedeném náhodném tak objeví proměnná SelectionProb, resp. SamplingWeight obsahující pravděpodobnost výběru, resp. výběrovou váhu pro každé vybrané pozorování.

Náhodný výběr PPS pro výběry skupin byl dále rozvinut v textech Brewer (1963), Murthy (1957) a Sampford (1967).

3 Stratifikovaný náhodný výběr

Podstatou stratifikovaného náhodného výběru je rozdělení opory výběru do *H* disjunktních skupin jednotek na výběr (tj. $\sum_{h=1}^{H} N_h$, kde dolní index *h* označuje *h*-tou skupinu jednotek o počtu N_h) nazývaných strata (vrstva), přičemž v každé z nich se realizuje vzájemně nezávislý náhodný výběr. Technika výběru v každé stratě může být stejná nebo v každé stratě může být náhodný výběr realizován rozdílným způsobem.

3.1 Stratifikovaný jednoduchý náhodný výběr

V opoře výběru se nacházejí skupiny pacientů rozdělených do skupin podle krevního tlaku na pacienty s krevním tlakem vysokým, normálním a optimálním. V následujícím náhodném výběru má být do výběrového souboru vybráno 200 pacientů ze skupiny s vysokým krevním tlakem, 100 pacientů s krevním tlakem normálním a 200 pacientů s tlakem optimálním. Nejprve se opora výběru uspořádá podle předpokládané stratifikační proměnné BP_Status pomocí procedury SORT. Pro dosažení potřebného výběru je nutné oporu výběru rozvrstvit (stratifikovat) podle proměnné BP_Status na vrstvu (stratu) pacientů podle skupin hodnocení krevního tlaku. Dalším požadavkem je informovat proceduru SURVEYSELECT o velikostech výběrových souborů náhodně vybíraných v jednotlivých vrstvách (stratách) opory výběru. Pro tento účel lze použít v proceduře SURVEYSELECT parametr SAMPSIZE=(200 100 200). Preferovanou variantou s informacemi o velikosti výběru v jednotlivých vrstvách (stratách) opory výběru je vytvoření pomocného datového souboru se stejně nazvanou a stejně formátovanou stratifikační proměnnou, ve kterém je ke každé stratě opory přiřazena velikost výběru n_h do klíčové proměnné _NSIZE_. Jméno takto vytvořeného pomocného souboru se přiřadí jako hodnota parametru SAMPSIZE=. Pro takto koncipovaný stratifikovaný jednoduchý náhodný výběr lze využít následující syntaxi příkazů:

```
proc sort data=heart;
     by BP_Status;
run;

data sampsizes;
     length BP_Status $7;
```

```
     input BP_Status _NSIZE_;
     datalines;
High 200
Normal 100
Optimal 200
;
run;

proc surveyselect data=heart out=sample_STR_SRS
sampsize=sampsizes seed=11111;
     strata BP_Status;
run;
```

Podobně jako u náhodného výběru PPSWOR je do výsledného souboru vybraných jednotek připojena proměnná SelectionProb, resp. SamplingWeight obsahující pravděpodobnost výběru, resp. výběrovou váhu pro každé vybrané pozorování.

3.2 Výběrové alokační strategie

Ve stratifikovaném jednoduchém náhodném výběru byly počty vybíraných jednotek do vrstev (strat) n_h zvoleny náhodně tak, aby platilo $\sum_{h=1}^{H} n_h = n$. Při aplikovaných výzkumech však takové náhodné rozdělení vybíraných jednotek přestalo vyhovovat a vyvstala potřeba o co nejvhodnější rozdělení pevného počtu *n* vybraných jednotek do jednotlivých vrstev (strat) ve výběrovém souboru. Procedura SURVEYSELECT umožňuje při realizaci stratifikovaných náhodných využití známých metod alokace jednotek do strat, které budou prezentovány prostřednictvím několika jednoduchých příkladů.

3.2.1 Proporcionální alokace

První prezentovanou alokační strategií je proporcionální alokace. Principem je realizovat výběr z dané vrstvy v poměru k jejímu podílu na celkové velikosti populace. To znamená, že počet vybíraných jednotek n_h pro *h*-tou vrstvu (stratu) lze vyjádřit jako $n_h = nN_h/N$, kde *N* je celková velikost opory výběru, N_h je velikost opory výběru v *h*-té vrstvě (stratě). Jelikož platí $\sum_{h=1}^{H} N_h/N = 1$, potom velikost celkového výběru je $\sum_{h=1}^{H} n_h = n$.

Následující příklad ukazuje syntaxi pro proceduru SURVEYSELECT k realizaci proporcionální alokace. Velikost vybraného vzorku $n = 500$ se nastaví parametrem SAMPSIZE=. Proporcionální alokace do vrstev (strat) definovaných proměnnou BP_Status (počet strat v opoře $H = 3$) se vyvolá zadáním příkazu ALLOC=PROP za lomítkem v příkazu STRATA. Při použití příkazu NOSAMPLE (opět za lomítkem v příkazu STRATA) se ve výstupním datovém souboru SAMPSIZES_PROP pro každou vrstvu (stratu) uloží vypočítané výběrové četnosti, za platnosti kterých jednotky ve výběru budou rozděleny proporcionálně. Tento

mezikrok dovoluje prozkoumat vypočítané výběrové četnosti k ujištění, že jsou použitelné pro další zpracování. Například může nastat situace, procedura SURVEYSELECT vrátí počet vybraných jednotek $n_h = 1$. To by mohlo představovat problémy během fáze analýzy, protože metody odhadu rozptylu procedurami v systému SAS vyžaduje, aby počet jednotek vybraných z každé vrstvy byl $n_h \geq 2$. V proceduře lze také zadat minimální velikost výběru ve vrstvě (stratě), a to pomocí parametru ALLOCMIN= také po lomítku za příkazem STRATA. Za předpokladu, že velikosti výběru ve stratách jsou považovány za vhodné, výstupní datový soubor s vypočítanými četnostmi lze v následujícím kroku využít k realizaci náhodného výběru. Strukturu datového souboru SAMPSIZES_PROP ilustruje tabulka 3.

Tab. 3 Struktura datového souboru SAMPSIZES_PROP *(Zdroj: vlastní zpracování)*

BP_Status	Total	AllocProportion	SampleSize	ActualProportion
High	2267	0,435208293	217	0,434
Normal	2143	0,41140334	206	0,412
Optimal	799	0,153388366	77	0,154

Každá strata definovaná proměnnou BP_Status má 1 řádek s následujícími proměnnými:

- Total – populační četnost N_h odvozenou z datového souboru (opory) HEART,
- AllocProportion – cílový podíl alokace v *h*-té stratě,
- SampleSize – aktuální výběrová četnost n_h alokovaná do *h*-té straty,
- ActualProportion – aktuální alokovaný podíl v *h*-té stratě.

Cílový podíl alokace, jejíž hodnoty jsou uloženy v proměnné AllocProportion, nemusí být celé číslo, zatímco počet vybíraných jednotek celé číslo být musí. Z tohoto důvodu procedura SURVESELECT při výpočtu alokací do strat vypočítané podíly zaokrouhluje, aby počet jednotek alokovaných do strat byl celé číslo. Zaokrouhlené hodnoty cílového podílu alokace jsou ukládány do proměnné ActualProportion. Podobný mechanismus zaokrouhlování používá procedura i pro ostatní alokační strategie.

3.2.2 Neymanova alokace

Proporcionální alokace je dobrá strategie, pokud jsou rozptyly ukazatelů jednotek v každé vrstvě (stratě) přibližně stejné. Pokud se rozptyly mezi stratami podstatně liší, je preferovanou alokační strategií tzv. Neymanova alokace. Použití Neymanovy alokace vyžaduje znalost populačních rozptylů σ_h^2 ukazatelů ve všech definovaných stratách. Pokud tyto populační rozptyly známy nejsou, lze použít jejich odhady s_h^2 z předešlých zjišťování nebo pilotní studie. Pro pevnou velikost výběrového souboru *n* při stratifikovaném náhodném výběru lze minimalizovat

varianci výběrového průměru, pokud jsou velikosti náhodných výběrů n_h v jednotlivých stratách definovány jako $n_h = nN_h\sigma_h / \sum_{h=1}^{H} N_h\sigma_h$ (Cochran, 1977).

Při Neymanově alokaci velikosti náhodných výběrů v jednotlivých vrstvách (stratách) závisí nejen na velikosti populace, ale také od variability. V proceduře SURVEYSELECT lze tuto alokační strategii vyvolat zadáním parametru ALLOC=NEYMAN za lomítkem v příkazu STRATA.

Nechť jsou k dispozici výsledky srovnatelného výzkumu provedeného už v minulém období, který odhalil, že pacienti s vysokým tlakem mají větší varianci hodnot cholesterolu než pacienti s krevním tlakem normálním a optimálním. Konkrétně se rozptyl cholesterolu u pacientů s vysokým tlakem pohybuje na úrovni 300 mg/l, u ostatních skupin pacientů byl zjištěn ve výši 70 mg/l a 30 mg/l. Pro alokaci jednotek do vrstev (strat) je tedy nutné brát do úvahy také varianci jednotlivých skupin pacientů. Aby bylo možné procedurou SURVEYSELECT alokovat jednotky do výběru v souladu s Neymanovou alokační strategií, je nutné do procedury předat informaci o rozptylech hodnot v jednotlivých stratách. Toho je dosaženo odkazem na datový soubor obsahující informaci o rozptylech ve stratě pomocí parametru VAR= procedury umístěný za lomítkem příkazu STRATA. Klíčová proměnná z tohoto datového souboru, ve kterém bude procedura hledat informace o rozptylech ve stratě se nazývá _VAR_. Syntaxe příkazu je podobná jako u stratifikovaného výběru s proporcionální alokací. Parametrem v příkazu STRATA lze také vložit velikost přípustné chyby odhadu průměru a hladinu spolehlivosti (implicitní hodnota je 95%).

Výstupem procedury je datový soubor SAMPSIZES_NEYMAN, který obsahuje alokaci jednotek do strat výběrového souboru s ohledem na variabilitu hodnot ve stratách definovaných proměnnou BP_Status. Struktura souboru je obdobná jako u výstupů z proporcionální alokace s výjimkou samotných hodnot alokací. Je patrné, že největší počet vybraných jednotek je alokováno do skupiny pacientů s vysokým krevním tlakem (BP_Status=High), což je odrazem největší variability. Struktura souboru SAMPSIZES_NEYMAN je ilustrována Tabulkou 4.

Tab. 4 Struktura datového souboru SAMPSIZES_NEYMAN *(Zdroj: vlastní zpracování)*

BP_Status	Total	Variance	AllocProportion	SampleSize	ActualProportion
High	2267	300	0,637723288	319	0,638
Normal	2143	70	0,291199966	146	0,292
Optimal	799	30	0,071076746	35	0,07

3.2.3 Optimální alokace

Optimální alokační strategie je rozšířením Neymanovy alokace o náklady na zjišťování specifické každé vrstvě (stratě). Podle (Lohr, 1999) optimální alokace je výsledkem minimalizace očekávaného rozptylu odhadu populačního průměru za použití nákladového modelu ve tvaru $C = C_0 + \sum_{h=1}^{H} c_h$, kde C_o jsou konstantní náklady zjišťování a c_h jsou náklady vynaložené na zjišťování v h-té vrstvě (stratě). Za platnosti výše uvedeného nákladového modelu a známých rozptylů ve stratách, alokování jednotek do h-té straty v souladu s optimální alokací se realizuje na základě vztahu $n_h = n\frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H}N_h\sigma_h/\sqrt{c_h}}$. Syntaxe příkazů procedury SURVEYSELECT je velmi podobná jako v předešlé alokaci. Rozdíl spočívá pouze v tom, že do parametru COST= za lomítko příkazu STRATA se vloží jméno pomocného datového souboru, který v proměnné _COST_ obsahuje náklady na zjišťování specifické jednotlivým stratám a v proměnné _VAR_ rozptyl ve stratách. Současně se za lomítko příkazu STRATA vloží parametr VAR, který proceduru informuje o tom, že proměnná _VAR_ (obsahující rozptyly strat) se také nachází v pomocném datovém souboru.

3.2.4 Realizace různých technik náhodného výběru pro různé straty

Diskuse o stratifikovaném náhodném výběru bude uzavřena jednoduchým příkladem, který ukazuje, jak je možné realizovat různé metody náhodného výběru pro různé vrstvy (straty).

Pokud bude při plánovaném výzkumu podmínek života pacientů kladen speciální důraz na reprezentativnost výběru pacientů s vysokým krevním tlakem (BP_status= High), je výhodné provést v této skupině pacientů náhodný výběr s pravděpodobnostmi úměrnými jejich hladině cholesterolu v krvi z proměnné Cholesterol. U ostatních skupin pacientů bude vyhovovat jednoduchý náhodný výběr. Počet vybraných pacientů s vysokým krevním tlakem má být 300, pacientů z ostatních skupin jen 200. Syntaxe příkazů pro realizaci tohoto výběru je uvedena následovně:

```
proc surveyselect data=heart (where=(BP_Status='High'))
out=sample_PPS_part1 sampsize=300 seed=11111 method=PPS;
     size cholesterol;
run;

proc surveyselect data=heart (where=(BP_Status<>'High'))
out=sample_STR_part2 sampsize=200 seed=11111 stats;
run;

data sample_STR_all;
     set sample_PPS_part1 sample_STR_part2;
run;
```

4 Skupinový (klastrový) náhodný výběr

V aplikovaném průzkumném výzkumu někdy mohou být primární jednotky výběru (PSU[14]) tvořené skupinami elementárních jednotek. K tomu může dojít přirozeně nebo záměrně, aby se usnadnil proces sběru dat. V datovém souboru o pacientech HEART lze identifikovat skupiny pacientů například podle kategorie kouření, tj. podle proměnné Smoking_Status. V některých výzkumech může být vhodnější namísto jednotlivých pacientů náhodně vybírat celé skupiny.

4.1 Jednostupňový skupinový (klastrový) náhodný výběr

V opoře výběru se nacházejí skupiny pacientů rozdělených do skupin podle hmotnosti v librách. Hmotnost (proměnná Weight) se pohybuje od hodnoty 67 lb až po 200 lb, tj. od 30 kg až po 90 kg. Primárními jednotkami výběru jsou pacienti rozdělení do skupin podle své hmotnosti. V následujícím příkladu náhodného výběru má skupinový náhodný výběr zajistit náhodně vybraný soubor o velikosti 20 skupin pacientů seskupených podle své hmotnosti s pravděpodobností úměrnou velikostem systolického krevního tlaku (hodnoty v proměnné Systolic). To znamená, že největší pravděpodobnost výběru mají skupiny pacientů s vyšší hodnotou systolického krevního tlaku. Syntaxe příkazu je následující:

```
proc surveyselect data=heart out=sample_cluster sampsize=20
seed=11111;
     cluster Weight;
     size Systolic;
   run;
```

4.2 Vícestupňový skupinový (klastrový) náhodný výběr

Použití procedury SURVEYSELECT pro vícestupňový skupinový (klastrový) náhodný výběr představuje následující příkazová syntaxe:

```
proc sort data=heart;
     by BP_Status;
run;

data sampsizes_clusters2;
     length BP_Status $7;
     input BP_Status _NSIZE_;
     datalines;
High 50
Normal 25
Optimal 20
;
run;
```

[14] PSU je zkratkou pro anglické označení pro primární jednotky výběru (Primary Sampling Unit).

```
* Krok 1) náhodný výběr primárních výběrových jednotek
(skupiny pacientů podle hmotnosti) ;
proc surveyselect data=heart seed=11111
sampsize=sampsizes_clusters2
     out=sample_primarystage (rename=(SelectionProb=SelectProb1
SamplingWeight=SamplingWeight1));
     strata BP_Status;
     cluster Weight;
run;

* Krok 2) náhodný výběr sekundárních výběrových jednotek
(pacientů uvnitř skupin);
proc surveyselect data=sample_primarystage seed=11111 rate=.5
     out=sample_secondarystage
(rename=(SelectionProb=SelectProb2
SamplingWeight=SamplingWeight2));
     strata BP_Status Weight;
run;
```

Jako primární výběrové jednotky se v prvním stupni výběru (označený jako Krok 1) náhodně vybírají skupiny pacientů o stejné hmotnosti. Jedná se o náhodný výběr stratifikovaný, kdy opora výběru bude rozvrstvena podle hodnoty krevního tlaku (proměnná BP_Status). Z tohoto důvodu musí opora výběru na prvním stupni podle této proměnné být uspořádána procedurou SORT. Podle hodnot proměnné _NSIZE_ v pomocném datovém souboru SAMPSISEZES_CLUSTER2 bude náhodně vybráno celkem 50 + 25 + 20 = 95 skupin pacientů. Vybrané skupiny na prvním stupni výběru jsou ukládány do souboru SAMPLE_PRIMARYSTAGE. Současně s uložením výstupu náhodného výběru jsou proměnné SelectionProb a SamplingWeight z tohoto souboru přejmenovány na SelectionProb1 a SamplingWeight1.

Pro druhý stupeň výběru jako výběrová opora je použit datový soubor obsahující v prvním stupni náhodně vybrané skupiny SAMPLE_PRIMARYSTAGE. Stratifikovaným náhodným výběrem (podle proměnných BP_STatus a Weight) se náhodně vybere soubor pacientů a tento výběr se uloží do výsledného souboru SAMPLE_SECONDARYSTAGE. Velikost náhodného výběru na druhém stupni je určena parametrem RATE=0,5. který definuje míru výběru pacientů ve skupinách. DATA krok za sekundárním výběrem počítá ve výsledném datovém souboru SAMPLE_MULTISTAGE celkovou hodnotu pravděpodobností výběru a výběrových vah pro každou náhodně vybranou jednotku.

5 Závěr

Cílem tohoto příspěvku bylo představit některé běžné techniky realizace náhodných výběrů používané v aplikovaném průzkumu a ilustrovat základní příklady syntaxe procedury SURVEYSELECT, kterými jsou tyto techniky

implementovány. Kromě diskuze o několika základních přístupech k náhodnému výběru byly prezentovány aplikace stratifikovaného a skupinového (klastrového) náhodného výběru. Vhodně realizovaný náhodný výběr včetně zohlednění všech otázek techniky výběru (pravděpodobnosti výběru, výběrové váhy, stratifikace a shlukování) poskytuje záruku správné statistické inference.

6 Literatura

Brewer, K. (1963). A Model of Systematic Sampling with Unequal Probabilities. Australian Journal of Statistics, 5, pp. 93–105.

Cochran, W. (1977). Sampling Techniques. 3. vyd. New York: Wiley.

Frankovič, B., & Stankovičová, I. (2020). Určenie veľkosti vzorky pre aplikovaný výskum. Forum Statisticum Slovacum, Vol. XVI (2020), No. 1, pp.57-50.

Kalton, G. (1983). Introduction to Survey Sampling. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-035. Newbury Park, CA: Sage.

Kish, L. (1965). Survey Sampling. New York, NY: Wiley.

Lohr, S. (1999). Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press.

Murthy, M. (1957). Ordered and Unordered Estimators in Sampling without Replacement. Sankhya, 18, pp. 379–390.

Sampford, M. (1967). On Sampling without Replacement with Unequal Probabilities of Selection. Biometrika, 54, pp. 499–513.

SAS Institute Inc. (2018). Sashelp Data Sets. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2020). SAS/STAT® 15.2 User's Guide. Cary, NC: SAS Institute Inc.

Scheaffer, R., Mendenhall, W., & OTT, R. (1996). Elementary Survey Sampling. 5. vyd. Belmont, CA: Duxbury Press.

Taylor, L. (2013) "PROC SURVEYSELECT as a Tool for Drawing Random Samples", MidWest SAS User´s Group, pp.1-12.

Valliant, R., Dever, J., and Kreuter, F. (2013). Practical Tools for Designing and Weighting Survey Samples. New York, NY: Springer.