УДК: 303.5

АНТОН ПІГІДА,
аспірант факультету соціології Київського національного університету імені Тараса Шевченка

Про важливість коректного округлення кількості респондентів при побудові вибірки

Анотація

У роботі розглянуто проблеми, пов'язані з необхідністю округлювати кількість респондентів у стратах при побудові вибірки. Аналізуються відхилення від запланованого обсягу вибірки та зміщення у структурі вибіркової сукупності, що можуть виникати внаслідок застосування загальновживаних правил округлення чисел. На прикладах демонструється необхідність застосування спеціальних методів випадкового округлення. Пропонується алгоритм, що дає змогу мінімізувати похибки, які виникають в результаті округлення, та зберігає запланований обсяг вибірки. Аналізуються переваги такого випадкового округлення порівняно із загальновживаними правилами.

Ключові слова: *вибірка, квота, округлення чисел, похибка, випадкове округлення*

Проектуючи випадкову вибірку для емпіричного дослідження, зазвичай працюють з дробовими числами, адже обсяг страти обчислюється як пропорційна частка вибірки, що є, як правило, дійсним (дробовим) числом. На останньому кроці, при переході до кількості респондентів вочевиднюється, що всі числа потрібно округлити до натуральних, бо ми не можемо планувати страти, наприклад, із 12,6 респондента. Для цього здебільше застосовують класичне заокруглення до ближчого натурального числа або до ближчого більшого натурального числа [Turner, 2003; Suhr, 2009; Westfall, 2011; Chaudhuri, 2003]. Проте застосування звичайних правил дає незадовільний результат. Унаслідок їх застосування обсяг вибірки може змінитися — ми можемо отримати меншу вибірку (що економить витрати на дослідження), але гіршу репрезентативність вибірки, або ж більшу вибірку,

що спричинює подорожчання дослідження. У разі обох підходів буде різнитися запланований обсяг вибірки і може з'явитися похибка як результат зміщення структури вибіркової сукупності стосовно генеральної. У сучасній літературі цій проблемі не приділяють достатньої уваги; цю тему не розглядають, оскільки сприймають проблему як очевидну чи неважливу. Однак на практиці зустріч з цією проблемою є неминучою і її подолання є не надто очевидним.

У цій статті пропоновано алгоритм обчислення обсягу страт для випадкової вибірки з найменшим відхиленням від заданої.

Сформулюю саму проблему. Для цього розгляньмо тривіальний приклад: потрібно спроектувати пропорційну вибірку обсягом 400 респондентів для генеральної сукупності "населення старше за 18 років включно, яке проживає у містах із населенням понад 500 тисяч" (табл. 1).

Таблиця 1

Вибірка для генеральної сукупності

Місто	Населення міста	Населення міста віком 18+	Частка	Кількість об'єктів	Кількість об'єктів, округлена звичайним методом
Київ	2799199	2181161	0,3045	121,8019	122
Харків	1446500	1110006	0,1550	61,9857	62
Одеса	1009145	765917	0,1069	42,7709	43
Дніпропетровськ	1004853	750693	0,1048	41,9207	42
Донецьк	962049	713514	0,0996	39,8446	40
Запоріжжя	776535	584886	0,0817	32,6616	33
Львів	732009	559940	0,0782	31,2686	31
Кривий Ріг	665080	496859	0,0694	27,7460	28
Усього				**400**	**401**

Як бачимо, план вибірки відрізняється від очікуваного: замість запланованих 400 респондентів після округлення отримано 401 респондента. Це лише приклад із 8 стратами. Але чим більше значень у сукупності неокруглених чисел, тим вища ймовірність, що сума округлених чисел буде відрізнятися від суми неокруглених і тим більшою буде ця різниця.

Звичайно, завжди можна вручну поправити вибірку після округлення для будь-якого міста, щоб вийти на заданий сумарний обсяг респондентів, але в разі великих вибірок доведеться досить багато правити вручну. Крім того, при проектуванні вибірки краще якомога менше втручатися у вибірку ручними правками — по-перше, через порушення принципу випадковості, по-друге, через банальну можливість зробити ще більше помилок.

Отже, перша вимога — зберегти запланований обсяг. Але відхилення від заданої суми — не найбільша проблема, яка на нас чатує при округленні кількості респондентів.

Розгляньмо вибірку для сіл Київської області із квотами на стать і вік; оскільки статистика окремо по кожному селу чи району недоступна або ж відсутня, є лише стосовно сільського населення області в цілому, тому квоти у відсотковому вимірі однакові (табл. 2).

Таблиця 2

Вибірка для сіл Київської області із квотами за статтю та віком

Село	Чоловіки						Жінки						Усього
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Нові Петрівці	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Тарасівка	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Новосілки	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Красилівка	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Щасливе	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Вишеньки	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Рогозів	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Стайки	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Маслівка	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Підгірці	0,67	0,74	2,09	1,89	2,96	1,5	0,64	0,67	1,94	1,86	3,06	1,99	**20**
Усього	**6,66**	**7,39**	**20,87**	**18,88**	**29,63**	**15,01**	**6,36**	**6,65**	**19,4**	**18,6**	**30,62**	**19,95**	**200**

Якщо застосувати звичайне округлення, то отримаємо вибірку, подану в таблиці 3.

Таблиця 3

Вибірка для сіл Київської області із квотами за статтю та віком після звичайного округлення

Село	Чоловіки						Жінки						Усього
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Нові Петрівці	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Тарасівка	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Новосілки	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Красилівка	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Щасливе	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Вишеньки	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Рогозів	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Стайки	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Маслівка	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Підгірці	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Усього	**10**	**10**	**20**	**20**	**30**	**20**	**10**	**10**	**20**	**20**	**30**	**20**	**220**

Як видно з таблиці 3, не тільки обсяг вибірки збільшився на 20 респондентів, а й повністю змінилися квоти (табл. 4).

Таблиця 4

Різниця між округленою та неокругленою вибірками до та після округлення

Стать	Вік	До округлення	Після округлення	Різниця
Чоловіки	12–15	6,66	10	3,34
	16–19	7,39	10	2,61
	20–29	20,87	20	–0,87
	30–39	18,88	20	1,12
	40–54	29,63	30	0,37
	55–65	15,01	20	4,99
Жінки	12–15	6,36	10	3,64
	16–19	6,65	10	3,35
	20–29	19,40	20	0,60
	30–39	18,60	20	1,40
	40–54	30,62	30	–0,62
	55–65	19,94	20	0,06

Звичайно, отримана після округлення вибірка не зовсім відповідає структурі генеральної сукупності. Отже, її репрезентативність є гіршою порівняно із запланованою вибіркою.

Для розв'язання цієї проблеми сформулюю алгоритм для округлення сукупності чисел зі збереженням суми їх.

Після округлення числа в нього "зникає" чи "з'являється" частина, що є різницею між початковим числом та округленим числом, тобто залишок заокруглення. Накопичення таких залишків і призводить до загальної різниці між сумою початкової сукупності чисел і результувальної округленої. Цей алгоритм не іґнорує накопичення таких різниць, а після кожного округлення числа додає цю різницю до наступного, ще не округленого числа. А щоби позбавитися можливого систематичного зсуву при округленні чисел, що йдуть поруч у таблиці квот, введено випадковий вибір елемента, який будемо округляти наступним.

Зображу алгоритм у вигляді блок-схеми (рис. 1).

Операції додавання різниці з попереднього округлення, обчислення нової різниці, округлення числа та випадковий добір наступного числа повторюємо, доки буде округлено всі елементи.

Через те, що алгоритм містить випадковий добір, кожне застосування алгоритму буде давати трохи відмінний результат. Один із таких результатів роботи цього алгоритму запишемо у таблицю 5.

Порівняємо між собою суми за стовпчиками та рядками, щоб подивитися, наскільки відрізняється структура вибіркової сукупності до і після округлення (табл. 6).

Позначення:

r — змінна, що зберігає різницю між неокругленим і округленим числом;

a — пронумерована сукупність чисел;

a[i] — *i*-й елемент сукупності *a*;

round() — функція звичайного (математичного – до найближчого натурального) округлення числа.

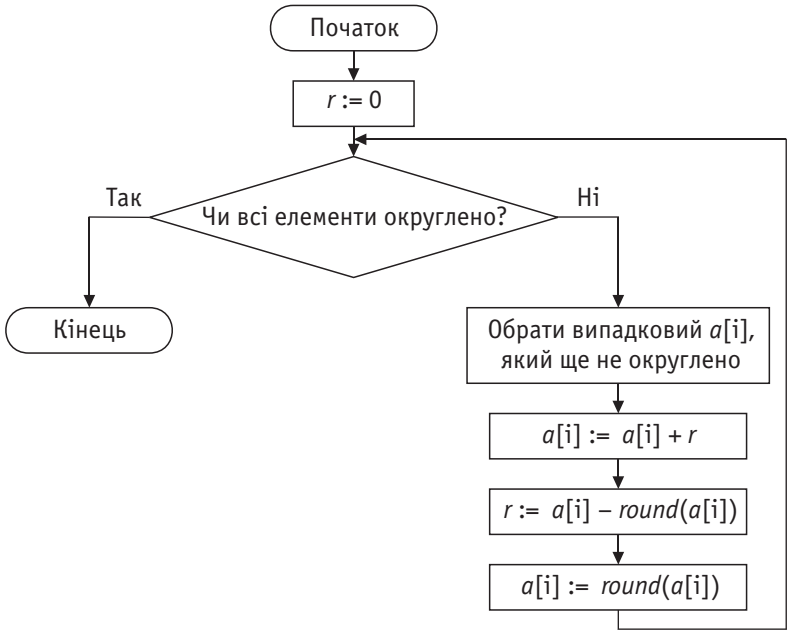

Рис. 1. Алгоритм округлення сукупності чисел зі збереженням суми

Таблиця 5

Квоти за статтю та віком після округлення спеціальним алгоритмом

Село	Чоловіки						Жінки						Усього
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Нові Петрівці	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Тарасівка	0	1	2	2	3	2	0	1	2	2	3	2	**20**
Новосілки	0	1	2	2	3	2	1	1	2	2	3	2	**21**
Красилівка	0	0	2	2	3	1	0	0	2	2	3	2	**17**
Щасливе	1	1	2	2	2	1	1	1	2	2	3	2	**20**
Вишеньки	1	0	2	2	3	1	0	1	2	2	3	2	**19**
Рогозів	1	1	3	2	3	1	0	1	2	1	3	2	**20**
Стайки	1	1	2	2	3	2	1	0	2	2	3	2	**21**
Маслівка	0	0	2	2	3	2	1	0	2	2	3	2	**19**
Підгірці	1	1	2	1	3	2	1	1	2	2	3	2	**21**
Усього	**6**	**7**	**21**	**19**	**29**	**16**	**6**	**7**	**20**	**19**	**30**	**20**	**200**

Таблиця 6

Порівняння різних типів округлення

Параметри		Вибірка			Квадрат різниці	
		Оригінальна (1)	Округлена у звичайний спосіб (2)	Округлена за спеціальним алгоритмом (3)	Між (1) і (2)	Між (1) і (3)
Нові Петрівці		20	22	22	4	4
Тарасівка		20	22	20	4	0
Новосілки		20	22	21	4	1
Красилівка		20	22	17	4	9
Щасливе		20	22	20	4	0
Вишеньки		20	22	19	4	1
Рогозів		20	22	20	4	0
Стайки		20	22	21	4	1
Маслівка		20	22	19	4	1
Підгірці		20	22	21	4	1
Чоловіки	12–15	6,66	10	6	11,17	0,43
	16–19	7,39	10	7	6,81	0,15
	20–29	20,87	20	21	0,75	0,02
	30–39	18,88	20	19	1,26	0,02
	40–54	29,63	30	29	0,14	0,40
	55–65	15,01	20	16	24,93	0,99
Жінки	12–15	6,36	10	6	13,28	0,13
	16–19	6,65	10	7	11,24	0,12
	20–29	19,40	20	20	0,36	0,36
	30–39	18,60	20	19	1,96	0,16
	40–54	30,62	30	30	0,39	0,39
	55–65	19,94	20	20	0	0
Сума квадратів різниці					112,29	21,17

Як бачимо, використаний алгоритм значно зменшив відхилення структури вибіркової сукупності після округлення за спеціальним алгоритмом стосовно неокруленої. Як міру відхилення обрано суму квадратів різниці. За кожним стовпчиком та рядком обчислюється сума (окремо для вибірки до і після округлення), потім обчислюється різниця цих сум для кожного рядка та стовпчика та підноситься до квадрата. Після цього підраховуємо загальну суму квадратів різниці за всіма рядками та стовпчиками. Піднесення до квадрата дає змогу, по-перше, позбавитися одного знака, по-друге, велика різниця нелінійно сильніша за меншу різницю (інакше кажучи, різниця між 20 і 22 та 20 і 21 є не вдвічі, а в 4 рази більшою), оскільки сильне відхилення для однієї категорії для вибірки є гіршим, ніж декілька невеликих відхилень

для кожної з категорій (адже, наприклад, аби виправити сильно занижену в розмірі категорію стосовно початкового плану вибірки, доведеться вводити ваги з великим коефіцієнтом).

Проте можна досягти ще менших відхилень за сумами стовпчиків та рядків, тобто відтворити цілими числами потрібні нам пропорції зі ще більшою точністю. Для цього потрібно трохи модифікувати алгоритм (рис. 2).

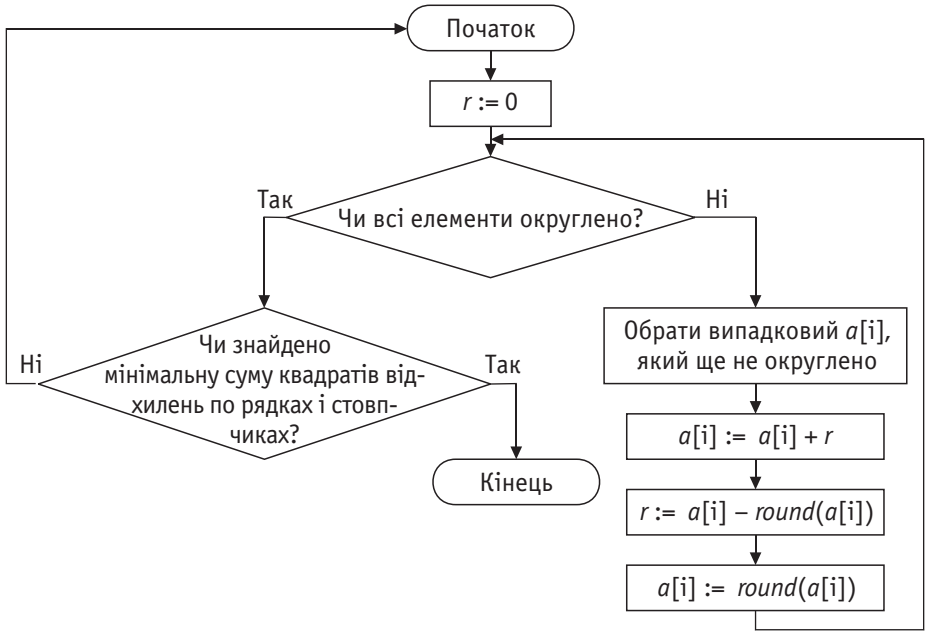

Рис. 2. Модифікована версія алгоритму

Додалася одна умова: "Чи знайдено мінімальну суму квадратів відхилень за рядками та стовпчиками?" Ясна річ, наперед неможливо сказати, яка сума квадратів відхилень для певної сукупності чисел буде мінімальною. Потрібно застосувати певний критерій для визначення достатньої суми квадратів різниці. Він може різнитися залежно від конкретної реалізації алгоритму.

Так, можна порівнювати суму квадратів різниці даної ітерації алгоритму з найкращою на даний момент ітерацією. Якщо, наприклад, упродовж 1000 ітерацій жодна із сум квадратів різниці не є меншою за найкращу, то видати той результат округлення, який був під час найкращої ітерації. Якщо ж знайшовся менший результат, то оголосити його найкращим і зробити ще 1000 ітерацій.

Узагалі найкращий результат можна знайти, просто перебираючи всі можливі комбінації. Але для цього доведеться зробити $n!$ ітерацій (де n — кількість чисел). Тому краще застосовувати випадкові перестановки. Чим більше ітерацій робить такий алгоритм, тим сильніше сума квадратів різниці буде наближатися до справжньої мінімально можливої суми квадратів різниці для даної сукупності чисел. Це відбувається завдяки тому, що кож-

ного разу ми відбираємо випадкові елементи для округлення з усієї сукупності.

Необхідна кількість ітерацій для досягнення практично достатньої суми квадратів різниці залежить від кількості елементів у сукупності, кількості стовпчиків та рядків. Для даних, застосовуваних у цьому прикладі, збільшення кількості ітерацій у 10 раз приводить у середньому до зменшення суми квадратів різниці у 1,37 раза. Тобто за 100 ітерацій у середньому буде досягнуто суму квадратів різниці 10,4, за 1000 — 7,6, за 10000 — 5,5, за 100000 — 4,1. Більш як 10000 ітерацій будуть істотно вповільнювати роботу алгоритму і приведуть лише до незначного покращення.

Отже, розгляньмо результат роботи модифікованого алгоритму (табл. 7, 8).

Таблиця 7

Квоти за статтю та віком після округлення модифікованою версією алгоритму

Село	Чоловіки						Жінки						Усього
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Нові Петрівці	0	1	2	2	2	2	1	1	2	2	3	2	**20**
Тарасівка	0	1	2	2	3	1	1	1	2	2	3	2	**20**
Новосілки	1	1	3	1	3	2	1	0	1	2	3	2	**20**
Красилівка	1	1	2	2	3	1	1	1	2	1	3	2	**20**
Щасливе	1	1	2	2	3	1	1	0	2	2	3	2	**20**
Вишеньки	0	1	2	2	3	2	0	1	2	2	3	2	**20**
Рогозів	0	1	2	2	3	1	1	0	2	2	3	2	**19**
Стайки	1	0	3	2	3	1	0	1	2	2	3	2	**20**
Маслівка	1	1	2	2	3	2	0	1	2	1	3	2	**20**
Підгірці	1	0	2	2	3	2	0	1	2	2	4	2	**21**
Усього	**6**	**8**	**22**	**19**	**29**	**15**	**6**	**7**	**19**	**18**	**31**	**20**	**200**

Як бачимо, отримана після округлення вибірка має в кілька разів меншу суму квадратів різниці за стовпчиками та рядками, ніж у попередньому варіанті, а також у кілька десятків разів меншу, ніж у разі округленої звичайним методом: 5,41 проти 21,17 та 112,29. Отже, отримана вибірка набагато точніше відповідає структурі генеральної сукупності.

Можна поліпшити алгоритм під певні потреби, наприклад, застосувавши замість квадрата різниці зважений квадрат різниці — квадрат різниці, поділений на суму елементів у даному рядку чи стовпчику. Це вможливить досягнення більшої точності там, де вона важливіша — для менших квот.

Отож, можна підсумувати, що під час проектування вибірки при перетворенні кількості респондентів до цілих чисел справді можуть виникати неочевидні на перший погляд проблеми. Для їх розв'язання можна застосовувати наведений у статті алгоритм.

Таблиця 8

Порівняння вибірки до та після округлення

Параметри		Вибірка		Квадрат різниці
		Оригінальна	Після округлення за спеціальним алгоритмом і з мінімізацією СКР	
Нові Петрівці		20	20	0
Тарасівка		20	20	0
Новосілки		20	20	0
Красилівка		20	20	0
Щасливе		20	20	0
Вишеньки		20	20	0
Рогозів		20	19	1
Стайки		20	20	0
Маслівка		20	20	0
Підгірці		20	21	1
Чоловіки	12–15	6,66	6	0,43
	16–19	7,39	8	0,37
	20–29	20,87	22	1,28
	30–39	18,88	19	0,02
	40–54	29,63	29	0,40
	55–65	15,01	15	0,00
Жінки	12–15	6,36	6	0,13
	16–19	6,65	7	0,12
	20–29	19,40	19	0,16
	30–39	18,60	18	0,36
	40–54	30,62	31	0,14
	55–65	19,94	20	0,00
Сума квадратів різниці				5,41

Пропонований підхід до округлення має дві переваги перед звичайним округленням: по-перше, зберігається задана сума респондентів, по-друге, отримана після округлення сукупність респондентів набагато точніше відповідає структурі генеральної сукупності.

Джерела

Chaudhuri S. Optimized Stratified Sampling for Approximate Query Processing [Electronic resource] / Surajit Chaudhuri, Gautam Das, Vivek Narasayya // Journal ACM Transactions on Database Systems (TODS) TODS Homepage archive. — 2007 . — June . — Vol. 32,

Is. 2. — Article No. 9. — Mode of access : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.8286&rep=rep1&type=pdf.

Suhr D. Selecting a Stratified Sample with PROC SURVEYSELECT [Electronic resource] / Diana Suhr // SAS Global Forum 2009, paper 058-2009. — Mode of access : http://support.sas.com/resources/papers/proceedings09/058-2009.pdf.

Turner A.G. Sampling strategies [Electronic resource] / Anthony G. Turner // UNITED NATIONS SECRETARIAT ESA/STAT/AC.93/2, Statistics Division 03. — 2003. — November. — P. 45. — Mode of access : http://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_2.pdf.

Westfall J.A. et al. Post-stratified estimation: within-strata and total sample size recommendations [Electronic resource] / James A. Westfall, Paul L. Patterson, John W. Coulston // Canadian Journal of Forest Research. — 2011. — Vol. 41. — P. 1130–1139. — Mode of access : http://www.fs.fed.us/rm/pubs_other/rmrs_2011_westfall_j001.pdf.