УДК: 303.5

АНТОН ПИГИДА,
аспирант факультета социологии Киевского национального университета имени Тараса Шевченко

О важности корректного округления количества респондентов при построении выборки

Аннотация

В работе рассмотрены проблемы, связанные с необходимостью округлять количество респондентов в стратах при построении выборки. Анализируются отклонения от запланированного объема выборки и смещения в структуре выборочной совокупности, которые могут возникать вследствие применения общепринятых правил округления чисел. На примерах демонстрируется необходимость применения специальных методов случайного округления. Предлагается алгоритм, позволяющий минимизировать погрешности, которые возникают в результате округления, и сохранять запланированный объем выборки. Анализируются преимущества такого случайного округления по сравнению с общепринятыми правилами.

Ключевые слова: *выборка, квота, округление чисел, погрешность, случайное округление*

Проектируя случайную выборку для эмпирического исследования, обычно работают с дробными числами, поскольку объем страты рассчитывается как пропорциональная доля выборки, как правило, являющаяся действительным (дробным) числом. На последнем шаге, при переходе к количеству респондентов становится очевидным, что все числа нужно округлить до натуральных, ведь мы не можем планировать страту, например, из 12,6 респондента. Для этого в основном используют классическое округление до ближайшего натурального числа или до ближайшего наибольшего натурального числа [Turner, 2003; Suhr, 2009; Westfall, 2011; Chaudhuri, 2003]. Однако применение обычных правил дает неудовлетворительный результат. В результате их применения объем выборки может измениться — мы можем получить меньшую выборку (что экономит затраты на исследование), но худшую репрезентативность выборки, или наоборот, большую вы-

борку, что ведет к удорожанию исследования. В случае обоих подходов будет разниться запланированный объем выборки и может возникнуть погрешность как результат смещения структуры выборочной совокупности по отношению к генеральной. В современной литературе этой проблеме не уделяется должное внимание; эту тему не рассматривают, поскольку воспринимают проблему как очевидную или незначимую. Однако на практике встречи с этой проблемой не миновать, а ее преодоление не вполне очевидно.

В этой статье предложен алгоритм вычисления объема страт для случайной выборки с наименьшим отклонением от заданной.

Сформулирую саму проблему. Для этого рассмотрим тривиальный пример: нужно спроектировать пропорциональную выборку объемом 400 респондентов для генеральной совокупности "население старше 18 лет включительно, проживающее в городах с населением свыше 500 тысяч" (табл. 1).

Таблица 1

Выборка для генеральной совокупности

Город	Население города	Население города в возрасте 18+	Доля	Количество объектов	Количество объектов, округленное обычным методом
Киев	2799199	2181161	0,3045	121,8019	122
Харьков	1446500	1110006	0,1550	61,9857	62
Одесса	1009145	765917	0,1069	42,7709	43
Днепропетровск	1004853	750693	0,1048	41,9207	42
Донецк	962049	713514	0,0996	39,8446	40
Запорожье	776535	584886	0,0817	32,6616	33
Львов	732009	559940	0,0782	31,2686	31
Кривой Рог	665080	496859	0,0694	27,7460	28
Всего				**400**	**401**

Как видим, план выборки отличается от ожидаемого: вместо запланированных 400 респондентов после округления получаем 401 респондента. Это только пример с 8 стратами. Но чем больше значений в совокупности неокругленных чисел, тем выше вероятность, что сумма округленных чисел будет отличаться от суммы неокругленных и тем большей будет эта разница.

Разумеется, всегда можно вручную поправить выборку после округления для какого-либо города, чтобы выйти на заданный суммарный объем респондентов, но в случае больших выборок придется довольно много править вручную. Кроме того, при проектировании выборки желательно как можно меньше вмешиваться в выборку ручными правками — во-первых, поскольку нарушается принцип случайности, во-вторых, учитывая банальную возможность наделать еще больше ошибок.

Итак, первое требование — сохранить запланированный объем. Но отклонение от заданной суммы — не самая большая проблема, поджидающая нас при округлении количества респондентов.

Рассмотрим выборку для сел Киевской области с квотами на пол и возраст; поскольку статистика отдельно по каждому селу или району недоступна или вовсе отсутствует, а имеется только для сельского населения области в целом, поэтому квоты в процентном измерении одинаковы (табл. 2).

Таблица 2

Выборка для сел Киевской области с квотами по полу и возрасту

Село	Мужчины						Женщины						Всего
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Новопетровцы	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**
Тарасовка	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**
Новоселки	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**
Красиловка	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**
Счастливое	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**
Вишенки	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**
Рогозив	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**
Масловка	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**
Подгорцы	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**
Всего	**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**

Если применить обычное округление, то получим выборку, представленную в таблице 3.

Таблица 3

Выборка для сел Киевской области с квотами по полу и возрасту после обычного округления

Село	Мужчины						Женщины						Всего
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Новопетровцы	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Тарасовка	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Новоселки	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Красиловка	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Счастливое	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Вишенки	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Рогозив	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**
Масловка	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Подгорцы	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Всего	**10**	**10**	**20**	**20**	**30**	**20**	**10**	**10**	**20**	**20**	**30**	**20**	**220**

Как видно из таблицы 3, не только объем выборки увеличился на 20 респондентов, но и полностью изменились квоты (табл. 4).

Таблица 4

Разница между округленной и неокругленной выборками до и после округления

Пол	Возраст	До округления	После округления	Разница
Мужчины	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
Женщины	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

Конечно, полученная после округления выборка не в полной мере соответствует структуре генеральной совокупности. Поэтому ее репрезентативность хуже по сравнению с запланированной выборкой.

Для решения этой проблемы сформулирую алгоритм для округления совокупности чисел с сохранением их суммы.

После округления числа у него “исчезает” или “появляется” часть, являющаяся разностью между начальным числом и округленным числом, то есть остаток округления. Накопление таких остатков и приводит к общей разности между суммой начальной совокупности чисел и результирующей округленной. Этот алгоритм не игнорирует накопления подобных разностей, а после каждого округления числа прибавляет эту разность к следующему, еще не округленному числу. А чтобы избежать возможного систематического сдвига при округлении чисел, расположенных рядом в таблице квот, вводится случайный отбор элемента, который мы будем округлять следующим.

Изображу алгоритм в виде блок-схемы (рис. 1).

Операции прибавления разности из предыдущего округления, вычисление новой разности, округление числа и случайный выбор следующего числа повторяем, пока не будут округлены все элементы.

Поскольку алгоритм содержит случайный отбор, каждое применение алгоритма будет давать несколько отличающийся результат. Один из таких результатов работы этого алгоритма запишем в таблицу 5.

Сравним между собой суммы по столбцам и рядам, чтобы посмотреть, насколько отличается структура выборочной совокупности до и после округления (табл. 6).

Обозначения:

r — переменная, сохраняющая разность между неокругленным и округленным числом;

a — пронумерованная совокупность чисел;

a[i] — *i*-й элемент совокупности *a*;

round() — функция обычного (математического — до ближайшего натурального) округления числа.

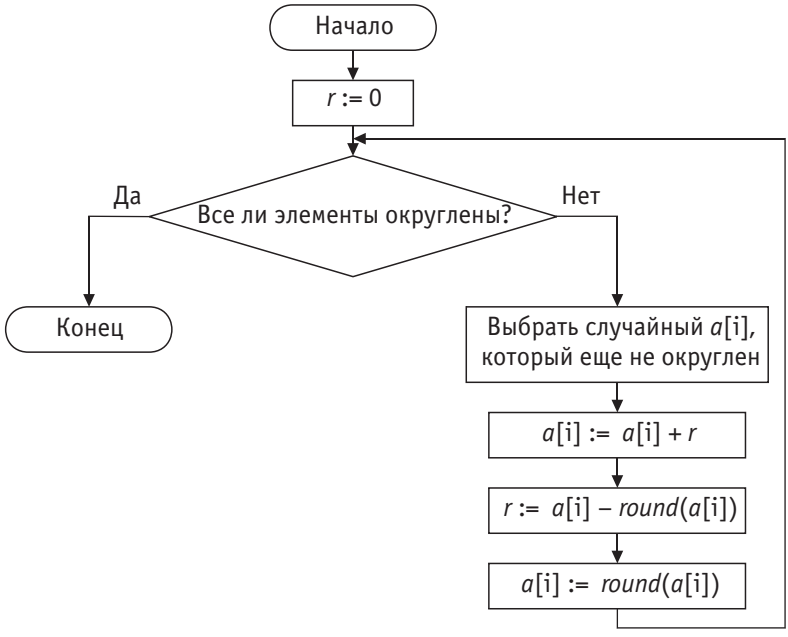

Рис. 1. Алгоритм округления совокупности чисел с сохранением суммы

Таблица 5

Квоты по полу и возрасту после округления специальным алгоритмом

Село	Мужчины						Женщины						Всего
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Новопетровцы	1	1	2	2	3	2	1	1	2	2	3	2	**22**
Тарасовка	0	1	2	2	3	2	0	1	2	2	3	2	**20**
Новоселки	0	1	2	2	3	2	1	1	2	2	3	2	**21**
Красиловка	0	0	2	2	3	1	0	0	2	2	3	2	**17**
Счастливое	1	1	2	2	2	1	1	1	2	2	3	2	**20**
Вишенки	1	0	2	2	3	1	0	1	2	2	3	2	**19**
Рогозив	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**
Масловка	0	0	2	2	3	2	1	0	2	2	3	2	**19**
Подгорцы	1	1	2	1	3	2	1	1	2	2	3	2	**21**
Всего	**6**	**7**	**21**	**19**	**29**	**16**	**6**	**7**	**20**	**19**	**30**	**20**	**200**

Таблица 6

Сравнение разных типов округления

Параметры		Выборка			Квадрат разности	
		Оригинальная (1)	Округленная обычным способом (2)	Округленная с помощью специального алгоритма (3)	Между (1) и (2)	Между (1) и (3)
Новопетровцы		20	22	22	4	4
Тарасовка		20	22	20	4	0
Новоселки		20	22	21	4	1
Красиловка		20	22	17	4	9
Счастливое		20	22	20	4	0
Вишенки		20	22	19	4	1
Рогозив		20	22	20	4	0
Стайки		20	22	21	4	1
Масловка		20	22	19	4	1
Подгорцы		20	22	21	4	1
Мужчины	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
Женщины	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
Сумма квадратов разности					112,29	21,17

Как видим, использованный алгоритм значительно уменьшил отклонение структуры выборочной совокупности после округления посредством специального алгоритма по отношению к неокругленной. В качестве меры отклонения использована сумма квадратов разности. По каждому столбцу и ряду вычисляется сумма (отдельно для выборки до и после округления), потом вычисляется разность этих сумм для каждого ряда и столбца и возводится в квадрат. После этого подсчитываем общую сумму квадратов разности по всем рядам и столбцам. Возведение в квадрат позволяет, во-первых, избавиться от одного знака, во-вторых, большая разница нелинейно сильнее меньшей разницы (иначе говоря, разница между 20 и 22 и 20 и 21 не в два, а в четыре раза больше), поскольку сильное отклонение для одной катего-

рии для выборки хуже, чем несколько небольших отклонений для каждой из категорий (так как, например, чтобы исправить сильно заниженную в размере категорию относительно начального плана выборки, придется вводить веса с большим коэффициентом).

Однако возможно добиться еще меньших отклонений по суммам столбцов и рядов, то есть воспроизвести целыми числами нужные нам пропорции с еще большей точностью. Для этого следует немного модифицировать алгоритм (рис. 2).

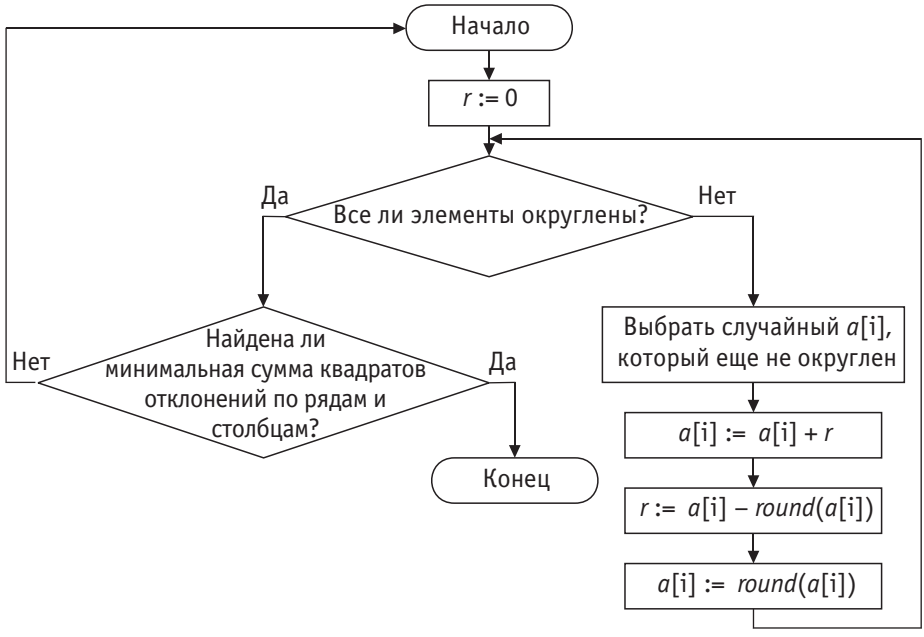

Рис. 2. Модифицированная версия алгоритма

Добавилось одно условие: “Найдена ли минимальная сумма квадратов отклонений по рядам и столбцам?” Понятно, что заранее невозможно сказать, какая сумма квадратов отклонений для определенной совокупности чисел будет минимальной. Нужно применить некоторый критерий для определения достаточной суммы квадратов разности. Он может различаться в зависимости от конкретной реализации алгоритма.

Так, можно сравнивать сумму квадратов разности данной итерации алгоритма с наилучшей на данный момент итерацией. Если, к примеру, в течение 1000 итераций ни одна из сумм квадратов разности не будет меньше наилучшей, то выдать тот результат округления, который был во время наилучшей итерации. Если же нашелся меньший результат, то объявить его наилучшим и провести еще 1000 итераций.

В общем самый лучший результат можно найти, если просто перебирать все возможные комбинации. Но для этого придется сделать $n!$ итераций (где n — количество чисел). Поэтому лучше использовать случайные перестановки. Чем больше итераций делает такой алгоритм, тем сильнее сумма квадратов разности будет приближаться к настоящей минимально возможной сумме квадратов разности для данной совокупности чисел. Это проис-

ходит благодаря тому, что каждый раз мы отбираем случайные элементы для округления из всей совокупности.

Необходимое количество итераций для достижения практически достаточной суммы квадратов разности зависит от количества элементов в совокупности, количества столбцов и рядов. Для данных, использованных в этом примере, увеличение количества итераций в 10 раз приводит в среднем к уменьшению суммы квадратов разности в 1,37 раза. То есть за 100 итераций в среднем будет достигнута сумма квадратов разности 10,4, за 1000 — 7,6, за 10000 — 5,5, за 100000 — 4,1. Более 10000 итераций будут ощутимо замедлять работу алгоритма и приведут лишь к незначительному улучшению.

Итак, рассмотрим результат работы модифицированного алгоритма (табл. 7, 8).

Таблица 7

Квоты по полу и возрасту
после округления по модифицированной версии алгоритма

Село	Мужчины						Женщины						Всего
	12–15	16–19	20–29	30–39	40–54	55–65	12–15	16–19	20–29	30–39	40–54	55–65	
Новопетровцы	0	1	2	2	2	2	1	1	2	2	3	2	**20**
Тарасовка	0	1	2	2	3	1	1	1	2	2	3	2	**20**
Новоселки	1	1	3	1	3	2	1	0	1	2	3	2	**20**
Красиловка	1	1	2	2	3	1	1	1	2	1	3	2	**20**
Счастливое	1	1	2	2	3	1	1	0	2	2	3	2	**20**
Вишенки	0	1	2	2	3	2	0	1	2	2	3	2	**20**
Рогозив	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**
Масловка	1	1	2	2	3	2	0	1	2	1	3	2	**20**
Подгорцы	1	0	2	2	3	2	0	1	2	2	4	2	**21**
Всего	**6**	**8**	**22**	**19**	**29**	**15**	**6**	**7**	**19**	**18**	**31**	**20**	**200**

Как видим, полученная после округления выборка имеет в несколько раз меньшую сумму квадратов разности по столбцам и рядам, чем в предыдущем варианте, а также в несколько десятков раз меньшую, чем в случае округленной обычным методом: 5,41 против 21,17 и 112,29. Следовательно, полученная выборка намного точнее соответствует структуре генеральной совокупности.

Можно улучшить алгоритм под конкретные нужды, например, применив вместо квадрата разности взвешенный квадрат разности — квадрат разности, разделенный на сумму элементов в данном ряду или столбце. Это позволит достичь большей точности там, где она гораздо важнее — для меньших квот.

Подведем итог. Во время проектирования выборки при преобразовании количества респондентов до целых чисел действительно могут возникать неочевидные на первый взгляд проблемы. Для их решения можно применять приведенный в статье алгоритм.

Таблица 8

Сравнение выборки до и после округления

Параметры		Выборка		Квадрат разности
		Оригинальная	После округления посредством специального алгоритма с минимизацией СКР	
Новопетровцы		20	20	0
Тарасовка		20	20	0
Новоселки		20	20	0
Красиловка		20	20	0
Счастливое		20	20	0
Вишенки		20	20	0
Рогозив		20	19	1
Стайки		20	20	0
Масловка		20	20	0
Подгорцы		20	21	1
Мужчины	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
Женщины	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
Сумма квадратов разности				5,41

Предлагаемый подход к округлению имеет два преимущества перед обычным округлением: во-первых, сохраняется заданная сумма респондентов, во-вторых, полученная после округления совокупность респондентов гораздо точнее соответствует структуре генеральной совокупности.

Источники

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.