УДК 303.5

АНТОН ПІГІДА,
аспірант кафедри методології та методів соціологічних досліджень факультету соціології Київського національного університету імені Тараса Шевченка

Побудова оптимальної кластерної вибірки з урахуванням дизайн-ефекту

Анотація

Обсяг ресурсів, виділених на реалізацію вибірки будь-якого дослідження, є досить обмеженим. Тому дослідник зацікавлений у тому, щоб найкращим чином застосувати наявні ресурси й отримати вибірку з найменшою похибкою. У випадку простої випадкової вибірки виконання такого завдання є тривіальним — найкращою буде вибірка найбільшого обсягу. Але у практиці соціологічних досліджень працювати з простою випадковою вибіркою зазвичай не доводиться. Натомість використовують більш складні методи добору респондентів. Отже, цю статтю присвячено питанню побудови оптимальної кластерної вибірки з урахуванням дизайн-ефекту.

Ключові слова: *вибірка, кластерна вибірка, дизайн-ефект, оптимальне розміщення, оптимізація*

Вибірковий метод є основою, на якій базуються соціологічні дослідження. І кожен дослідник прагне побудувати таку вибірку, яка буде якомога точнішою, але при цьому не надто дорогою. Тобто існує завдання максимально ефективно використати ресурси, виділені на реалізацію дослідження, щоб отримати найкращий можливий результат.

Кластерна вибірка — один із найпопулярніших методів формування вибірки, застосовуваних для проведення соціологічних досліджень. Формується вона, як правило, у два етапи. Для початку слід описати генеральну сукупність дослідження як сукупність певних кластерів. У ролі кластерів в Україні можуть виступати населені пункти, виборчі дільниці, поштові

відділення тощо. Далі з цієї сукупності кластерів випадковим чином обирають певну кількість таких кластерів, з якої в подальшому формується остаточна вибірка респондентів. Проте у кластерної вибірки існує одна суттєва вада — вибірка, отримана таким чином, зазвичай менш точна, ніж проста випадкова вибірка такого самого обсягу. Дослідники пов'язують це явище з високою дисперсією середніх, тобто з відмінностями середніх значень для певної ознаки в кожному кластері. Якщо, наприклад, виключити з кластерної вибірки певні кластери, то в результаті отримаємо зсув вибіркової оцінки середнього по всій сукупності. Тобто кластерна вибірка надто чутлива до того, які кластери потраплять до неї. Отже, ми безпосередньо підійшли до розгляду такого явища, як дизайн-ефект.

Дизайн-ефектом є відношення дисперсії оцінки, отриманої за такого дизайну вибірки, щодо дисперсії оцінки, отриманої за умови простого випадкового добору. Запропоновано такий показник було ще Леслі Кішем у 1965 році [Kish, 1965: p. 162]. Тобто такий показник можна інтерпретувати як міру точності, втрачену чи набуту внаслідок застосування поточної вибірки порівняно із застосуванням простої випадкової вибірки.

Для кластерної вибірки дизайн-ефект визначається так (див.: [Kish, 1965: p. 162]):

$$deff_{cl} = 1 + \rho(m - 1), \qquad (1)$$

де

m — обсяг кластера у вибірці;

ρ — коефіцієнт міжкластерної кореляції. У літературі можна також зустріти такий варіант позначення, як *ICC* (Intraclass correlation coefficient).

Обчислення ρ здійснюють за формулою (див.: [Fisher, 1925: p. 178]):

$$\rho = \frac{\sum_{n=1}^{N} (\bar{x}_n - \bar{x})^2}{Ns^2}, \qquad (2)$$

де

N — загальна кількість кластерів;

$\bar{x}_n$ — середнє у кластері;

$\bar{x}$ — середнє у сукупності;

s^2 — дисперсія.

Тобто щоб розрахувати значення *ICC*, необхідно знати значення ознаки для кожного кластера.

Отже, знаючи коефіцієнт міжкластерної кореляції, ми можемо визначити дизайн-ефект від кластеризації.

Оскільки для того, щоб обчислити дизайн-ефект за умови кластерного добору, необхідно знати значення ознаки в кожному кластері (навіть у тих, що не потраплять до вибірки), то зрозуміло, що за результатами самого дослідження дизайн-ефект від кластерного добору з'ясувати неможливо, оскільки відсутня інформація про ті кластери, які не потрапили до вибірки.

Як видно з формули (1), дизайн-ефект за умови кластерного добору не виникає у двох випадках: або $\rho = 0$, або $m = 1$. Якщо кластер складається

лише з 1 одиниці, то вибірка фактично зводиться до простої випадкової. Дисперсія між кластерами вже не матиме значення. Якщо ж коефіцієнт міжкластерної кореляції становить 0, це свідчить про те, що кластери між собою не відрізняються. Отже, немає значення, яка кількість і які кластери потраплять до вибірки, оскільки кожен з них може репрезентативно представляти генеральну сукупність.

Але зазвичай кластери між собою певним чином відрізняються, і тому коефіцієнт міжкластерної кореляції є більшим за 0. Тому на практиці кластерна вибірка буде тим точнішою, чим більше кластерів вона включатиме (за умови однакового загального обсягу).

Скористаймося результатами виборів до Верховної Ради України 2012 року, щоб оцінити дизайн-ефект залежно від кількості кластерів у вибірці. У ролі кластерів виступатимуть територіальні виборчі дільниці. Для початку необхідно розрахувати коефіцієнт міжкластерної кореляції для кожної партії. Він визначається за формулою (2). Не наводячи поетапно розрахунок цього коефіцієнта, зазначу лише, які дані було використано. Як загальну кількість кластерів використано загальну кількість територіальних виборчих дільниць. Середнє значення ознаки у кластері — це частка голосів за дану партію на цій територіальній виборчій дільниці. Середнє значення у сукупності — загальна частка голосів за дану партію. Результат обчислень наведено в таблиці 1.

Таблиця 1

Коефіцієнт міжкластерної кореляції для кожної партії

Партія	ρ
Комуністична партія	0,075
Свобода	0,137
УДАР	0,042
Батьківщина	0,119
Партія регіонів	0,183

Нехай обсяг нашої вибірки становитиме 1200 респондентів. Застосуймо отриманий коефіцієнт міжкластерної кореляції до формули (1), щоб виявити, як впливає на дизайн-ефект кількість кластерів у вибірці. Середній обсяг кластера візьмемо від 1 до 20, оскільки він лінійно пов'язаний із кількістю кластерів (обсяг вибірки = кількість кластерів × середній обсяг кластера; див. табл. 2).

Очевидно, що чим меншим буде обсяг кластера і чим більше, відповідно, буде цих кластерів у вибірці, тим нижчим буде дизайн-ефект.

Але на практиці, звісно, ми зустрінемося з тим, що вибірка обсягом 1200 респондентів із 60 міст по 20 респондентів у кожному кластері буде значно дешевшою, ніж вибірка обсягом 1200 респондентів із 120 міст по 10 респондентів у кожному. Річ у тім, що кожен новий кластер у вибірці призводить

до істотного подорожчання польових робіт, оскільки транспортні витрати значно перевищують оплату проведення інтерв'юєром додаткових інтерв'ю.

Таблиця 2

Залежність дизайн-ефекту від кількості кластерів у вибірці для кожної партії

Розмір кластера	Кількість кластерів	Комуністична партія	Свобода	УДАР	Батьківщина	Партія реґіонів
1	1200	1	1	1	1	1
2	600	1,07	1,14	1,04	1,12	1,18
3	400	1,15	1,27	1,08	1,24	1,37
4	300	1,22	1,41	1,13	1,36	1,55
5	240	1,30	1,55	1,17	1,48	1,73
6	200	1,37	1,69	1,21	1,60	1,92
7	171	1,45	1,82	1,25	1,72	2,10
8	150	1,52	1,96	1,29	1,84	2,28
9	133	1,60	2,10	1,33	1,96	2,47
10	120	1,67	2,23	1,38	2,08	2,65
11	109	1,75	2,37	1,42	2,19	2,83
12	100	1,82	2,51	1,46	2,31	3,02
13	92	1,90	2,64	1,50	2,43	3,20
14	86	1,97	2,78	1,54	2,55	3,38
15	80	2,04	2,92	1,59	2,67	3,57
16	75	2,12	3,06	1,63	2,79	3,75
17	71	2,19	3,19	1,67	2,91	3,93
18	67	2,27	3,33	1,71	3,03	4,12
19	63	2,34	3,47	1,75	3,15	4,30
20	60	2,42	3,60	1,79	3,27	4,48

Тому за певного фіксованого обсягу ресурсів ми можемо провести дослідження з великою вибіркою, але невеликою кількістю міст у вибірці; а також опитати велику кількість міст, але тоді вибірку доведеться зменшити.

Саме тут ми стикаємося із проблемою: яким чином розподілити ресурси на дослідження, щоб отримати найкращий результат? Найбільша вибірка респондентів не означає найнижчої похибки. Якщо провести опитування 1200 респондентів лише у Києві, Львові й Донецьку, це буде значно гірше під кутом репрезентативності, ніж опитати загалом 800 респондентів, але у 10 різних містах України.

Для початку нам необхідно знати, як обчислюється вартість польового етапу дослідження, тобто як впливає на вартість додаткове інтерв'ю для інтерв'юєра та додатковий населений пункт, до якого інтерв'юєрові необхідно дістатися, щоби провести свої інтерв'ю. Іншими словами, слід встановити, як задається функція витрат, що визначає, як витрати ресурсів на дослідження пов'язані з іншими чинниками.

Звісно, кожна дослідницька компанія буде по-своєму обчислювати вартість реалізації конкретної вибірки, і на цю вартість може впливати безліч чинників: відстань населеного пункту до найближчого обласного центру, відстань до залізниці, розташування опитувальних центрів тощо. За бажання їх усі можна врахувати і побудувати досить складну функцію витрат, але в цьому дослідженні вважатимемо, що на вартість реалізації вибірки впливає лише два чинники: транспортні витрати (які є однаковими для всіх кластерів) та оплата за одне інтерв'ю. Тобто на вартість впливатиме кількість респондентів у вибірці та кількість кластерів. Цю функцію витрат можна виразити такою формулою:

$$C = kc_{cl} + nc_r, \qquad (3)$$

де
k — кількість кластерів;
c_{cl} — транспортні витрати на один кластер;
n — обсяг вибірки;
c_r — вартість одного інтерв'ю.

Оскільки кількість кластерів у вибірці визначається як $k = n / m$, де m — обсяг кластера,то можна записати таку формулу:

$$C = (nc_{cl} / m) + nc_r.$$

Розв'яжемо тепер це рівняння для n:

$$n = \frac{C}{(c_{cl} / m) + c_r}. \qquad (4)$$

Отже, якщо знати транспортні витрати на один кластер та вартість одного інтерв'ю і задати загальну суму витрат, то можна порівняти, як це впливатиме на обсяг вибірки.

Нехай, наприклад, вартість одного інтерв'ю становить 32 грошові одиниці, а транспортні витрати — 200. При цьому загальний обсяг ресурсів, виділених на польовий етап дослідження, становить 60000 грошових одиниць. Залежно від розміру кластера ми отримаємо певний обсяг вибірки (табл. 3).

Якби дизайн-ефекту від кластеризації не існувало, то очевидно, що найбільший обсяг вибірки давав би найнижчу похибку. На підставі таблиці 2 ми вже пересвідчилися в тому, що дизайн-ефект зростає зі збільшенням розміру кластера, оскільки дизайн-ефект пов'язаний з обсягом вибірки так (див.: [Kish, 1965: p. 162]):

$$N_{eff} = N/deff. \qquad (5)$$

Отже, якщо ефективний обсяг вибірки дорівнює реальному обсягу, розділеному на дизайн-ефект, то обчислити похибку поточної вибірки можна за формулою:

$$d = 1{,}96\sqrt{\frac{0{,}25}{N/deff}}. \qquad (6)$$

Дані про дизайн-ефект для кожної із партій наведено в таблиці 1. Звідси обчислимо похибку вибірки для кожної партії залежно від обсягу кластера, закладеного у вибірці (табл. 4).

Таблиця 3

Залежність обсягу вибірки від розміру кластера та похибка для простої випадкової вибірки такого самого обсягу

Розмір кластера	Кількість кластерів	Обсяг вибірки	Похибка простої випадкової вибірки
1	259	259	0,061
2	227	455	0,046
3	203	608	0,040
4	183	732	0,036
5	167	833	0,034
6	153	918	0,032
7	142	991	0,031
8	132	1053	0,030
9	123	1107	0,029
10	115	1154	0,029
11	109	1196	0,028
12	103	1233	0,028
13	97	1266	0,028
14	93	1296	0,027
15	88	1324	0,027
16	84	1348	0,027
17	81	1371	0,026
18	77	1392	0,026
19	74	1411	0,026
20	71	1429	0,026

Таблиця 4

Залежність похибки вибірки від обсягу кластера

Розмір кластера	Обсяг вибірки	Комуністична партія	Свобода	УДАР	Батьківщина	Партія регіонів
1	*2*	*3*	*4*	*5*	*6*	*7*
1	259	0,061	0,061	0,061	0,061	0,061
2	455	0,048	0,049	0,047	0,049	0,050
3	608	0,043	0,045	0,041	0,044	0,046
4	732	0,040	0,043	0,038	0,042	0,045
5	833	0,039	0,042	0,037	0,041	0,045
6	918	0,038	0,042	0,036	0,041	0,045
7	991	0,037	0,042	0,035	0,041	0,045
8	1053	0,037	0,042	0,034	0,041	0,046
9	1107	0,037	0,043	0,034	0,041	0,046
10	1154	0,037	0,043	0,034	0,042	0,047

Закінчення табл. 4

1	*2*	*3*	*4*	*5*	*6*	*7*
11	1196	0,037	0,044	0,034	0,042	0,048
12	1233	0,038	0,044	0,034	0,042	0,048
13	1266	0,038	0,045	0,034	0,043	0,049
14	1296	0,038	0,045	0,034	0,043	0,050
15	1324	0,039	0,046	0,034	0,044	0,051
16	1348	0,039	0,047	0,034	0,045	0,052
17	1371	0,039	0,047	0,034	0,045	0,052
18	1392	0,040	0,048	0,034	0,046	0,053
19	1411	0,040	0,049	0,035	0,046	0,054
20	1429	0,040	0,049	0,035	0,047	0,055

Подивімося на таблицю 4 у вигляді графіка, наведеного на рисунку.

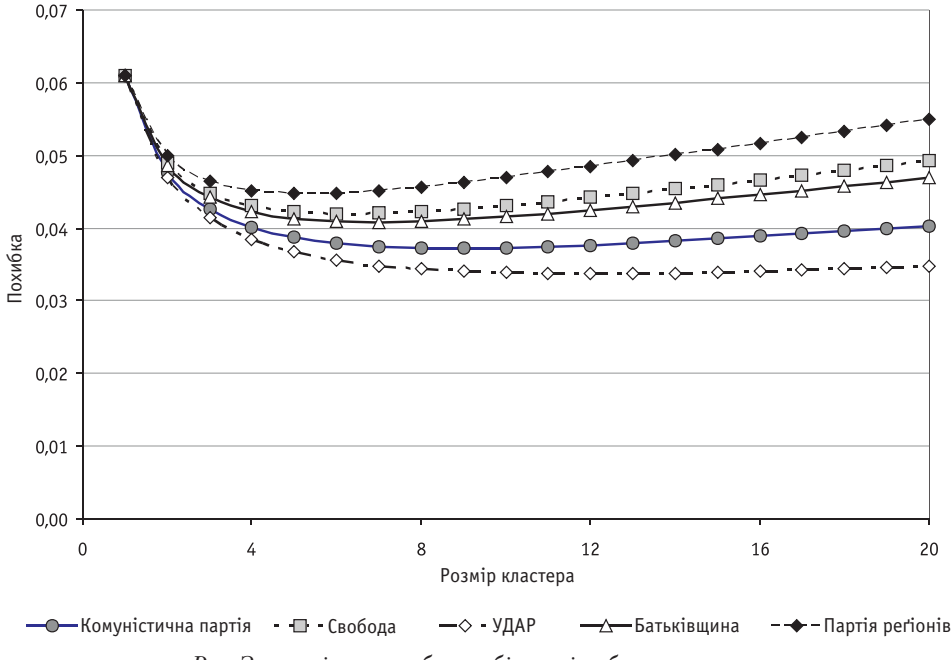

Рис. Залежність похибки вибірки від обсягу кластера

Як бачимо, зв'язок між обсягом вибірки і похибкою не є лінійним, і кожна партія досягає мінімальної похибки за певного обсягу кластера. Причому ця точка оптимуму в кожної партії своя і залежить від коефіцієнта міжкластерної кореляції (див. табл. 5).

Як бачимо, чим нижчим був коефіцієнт міжкластерної кореляції, тим більший розмір кластера є припустимим і, відповідно, тим більший загальний обсяг вибірки. У Партії реґіонів коефіцієнт міжкластерної кореляції найбільший, тому для того, щоб вибірка була якомога репрезентативнішою

для неї, вона має складатися з великої кількості кластерів, що зумовлює скорочення загального обсягу вибірки.

Таблиця 5

Оптимальна кількість кластерів для кожної партії

Партія	ρ	Розмір кластера	Кількість кластерів	Обсяг вибірки	Похибка
Комуністична партія	0,075	9	123	1107	0,037
Свобода	0,137	6	153	918	0,042
УДАР	0,042	12	103	1233	0,034
Батьківщина	0,119	7	142	991	0,041
Партія регіонів	0,183	5	167	833	0,045

Отже, щоби розрахувати оптимальну кількість кластерів у вибірці, необхідно знати коефіцієнт міжкластерної кореляції та функцію витрат.

Скористаймося даними з нашого прикладу, щоб продемонструвати розрахунок оптимальної кількості кластерів.

Якщо у формулу (6) підставити (1), то побачимо, що повністю формула обчислення похибки вибірки виглядає так:

$$d = 1{,}96\sqrt{0{,}25/n}\sqrt{1+\rho(m-1)}. \qquad (7)$$

Якщо підставити замість n формулу (4), то отримаємо:

$$d = 1{,}96\sqrt{0{,}25\Big/\frac{C}{(c_{cl}/m)+c_r}}\times\sqrt{1+\rho(m-1)}. \qquad (8)$$

Нехай ми оптимізуємо вибірку для досягнення мінімальної похибки для ГО "Свобода". Коефіцієнт міжкластерної кореляції для неї дорівнює 0,137. Вартість одного інтерв'ю становить 32 грошові одиниці, транспортні витрати — 200, загальний обсяг ресурсів — 60000 грошових одиниць.

Підставимо ці значення й отримаємо:

$$d = 1{,}96\sqrt{0{,}25\Big/\frac{60000}{(200/m)+32}}\times\sqrt{1+0{,}137(m-1)}.$$

Тепер необхідно знайти мінімум цієї функції. Для цього знайдемо для неї похідну щодо m:

$$d'(m) = \frac{0{,}00876983m^2 - 0{,}345272}{\sqrt{0{,}137m+0{,}863m^2}\sqrt{32+(200/m)}}.$$

Прирівняємо її до 0:

$$\frac{0{,}00876983m^2 - 0{,}345272}{\sqrt{0{,}137m+0{,}863m^2}\sqrt{32+(200/m)}} = 0.$$

Як розв'язок цього рівняння маємо:

$$m_1 = -6{,}27459,\ m_2 = +6{,}27459.$$

Ми знайшли мінімум розглядуваної функції і тепер знаємо, що найнижчу похибку вибірки отримаємо, якщо розмір кластера дорівнюватиме 6.

Висновки

Найбільший вплив на похибку кластерної вибірки справляють такі чинники: загальний обсяг вибірки, кількість кластерів у вибірці та коефіцієнт міжкластерної кореляції.

За умов обмеженості ресурсів на проведення дослідження від обсягу цих ресурсів та функції витрат залежать загальний обсяг вибірки та кількість кластерів у вибірці. Для створення кластерної вибірки з найнижчою можливою похибкою дослідник має визначити, з якої кількості кластерів має складатися його вибірка та якого обсягу вона має бути, щоби не перевищити межі наявних ресурсів. Спочатку розраховують коефіцієнт міжкластерної кореляції досліджуваної ознаки або ознаки, яку можна використати замість неї. Потім виводять функцію витрат, яка має показати, як пов'язані загальні витрати на дослідження з обсягом вибірки та кількістю кластерів у ній. У кожному випадку може бути своя функція, але загалом вона має показувати ці зв'язки. Далі виводять загальну формулу, яка пов'язує похибку із розміром кластера. Розмір кластера, за якого похибка вибірки буде найнижчою, дорівнюватиме мінімуму розрахованої функції.

Джерела

Черняк О.І. Техніка вибіркових досліджень / Черняк О.І. — К. : МІВВЦ, 2001. — 248 с.

Чурилов Н. Типология и проектирование выборочного социологического исследования (история и современность) / Чурилов Н. — К. : Факт, 2008. — 366 с.

Hansen M.H. Sample Survey Methods and Theory / Hansen M.H., Hurwitz W.N., Madow W.G. — N.Y. : John Wiley and Sons, Inc., 1953. — Vol. 1.

Kish L. Survey sampling / Kish L. — N.Y. : John Wiley & Sons, 1965. — 642 p.