УДК 303.5

АНТОН ПИГИДА,
аспирант кафедры методологии и методов социологических исследований факультета социологии Киевского национального университета имени Тараса Шевченко

Построение оптимальной кластерной выборки с учетом дизайн-эффекта

Аннотация

Объем ресурсов, выделяемых на реализацию выборки любого исследования, весьма ограничен. Поэтому исследователь заинтересован в том, чтобы наилучшим образом использовать имеющиеся ресурсы и получить выборку с наименьшей погрешностью. В случае простой случайной выборки выполнение такой задачи тривиально — самой лучшей будет выборка наибольшего объема. Но в практике социологических исследований работать с простой случайной выборкой обычно не приходится. В итоге используют более сложные методы отбора респондентов. Таким образом, данная статья посвящена вопросу построения оптимальной кластерной выборки с учетом дизайн-эффекта.

Ключевые слова: *выборка, кластерная выборка, дизайн-эффект, оптимальное размещение, оптимизация*

Выборочный метод служит основанием, на котором базируются социологические исследования. И каждый исследователь пытается построить такую выборку, которая будет как можно более точной, но при этом не слишком дорогой. То есть существует задача максимально эффективно использовать ресурсы, выделенные на реализацию исследования, чтобы получить наилучший возможный результат.

Кластерная выборка — один из самых популярных методов формирования выборки, применяемых для проведения социологических исследований. Формируется она, как правило, в два этапа. Сначала следует описать генеральную совокупность исследования как совокупность определенных

кластеров. В роли кластеров в Украине могут выступать населенные пункты, избирательные участки, почтовые отделения и т.п. Затем из этой совокупности кластеров случайным образом отбирают определенное количество таких кластеров, из которых в дальнейшем формируется окончательная выборка респондентов. Однако кластерная выборка имеет один существенный недостаток — полученная таким образом выборка, как правило, менее точная, чем простая случайная выборка такого же объема. Исследователи связывают это явление с высокой дисперсией средних, то есть с различиями средних значений для определенного признака в каждом кластере. Если, например, исключить из кластерной выборки определенные кластеры, то в результате получим сдвиг выборочной оценки среднего по всей совокупности. То есть кластерная выборка слишком чувствительна к тому, какие кластеры в нее попадут. Итак, мы непосредственно подошли к рассмотрению такого явления, как дизайн-эффект.

Дизайн-эффектом является отношение дисперсии оценки, полученной при таком дизайне выборки, к дисперсии оценки, полученной при условии простого случайного отбора. Этот показатель был предложен еще Лесли Кишем в 1965 году [Kish, 1965: p. 162]. То есть этот показатель можно интерпретировать как меру точности, утраченную или приобретенную вследствие применения текущей выборки по сравнению с применением простой случайной выборки.

Для кластерной выборки дизайн-эффект определяется следующим образом (см.: [Kish, 1965: p. 162]):

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

где

m — объем кластера в выборке;

ρ — коэффициент межкластерной корреляции. В литературе можно также встретить такой вариант обозначения, как *ICC* (Intraclass correlation coefficient).

Вычисления ρ проводят по формуле (см.: [Fisher, 1925: p. 178]):

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

где

N — общее количество кластеров;

$\bar{x}_n$ — среднее в кластере;

$\bar{x}$ — среднее в совокупности;

s^2 — дисперсия.

То есть чтобы вычислить *ICC* значения, необходимо знать значения признака для каждого кластера.

Зная коэффициент межкластерной корреляции, мы можем определить дизайн-эффект от кластеризации.

Поскольку для того, чтобы вычислить дизайн-эффект при условии кластерного отбора, необходимо знать значения признака в каждом кластере (даже у тех, которые не попадут в выборку), поэтому понятно, что по результатам самого исследования дизайн-эффект от кластерного отбора определить невозможно из-за отсутствия информации о тех кластерах, которые не попали в выборку.

Как видно из формулы (1), дизайн-эффект при условии кластерного отбора не возникает в двух случаях: или $\rho = 0$, или $m = 1$. Если кластер состоит всего из 1 единицы, то выборка фактически сводится к простой случайной. Дисперсия между кластерами уже не будет иметь значения. Если же коэффициент межкластерной корреляции составляет 0, это свидетельствует о том, что кластеры между собой не различаются. Поэтому не имеет значения, какое количество и какие кластеры попадут в выборку, поскольку каждый из них может репрезентативно представлять генеральную совокупность.

Но обычно кластеры между собой определенным образом разнятся, поэтому коэффициент межкластерной корреляции больше 0. Поэтому на практике кластерная выборка будет тем более точной, чем больше кластеров она будет включать (при условии одинакового общего объема).

Используем результаты выборов к Верховную Раду Украины 2012 года, чтобы оценить дизайн-эффект в зависимости от количества кластеров в выборке. В роли кластеров будут выступать территориальные избирательные участки. Сначала необходимо вычислить коэффициент межкластерной корреляции для каждой партии. Он определяется по формуле (2). Не приводя поэтапно расчет этого коэффициента, укажу только, какие данные были использованы. В качестве общего количества кластеров использовано общее количество территориальных избирательных участков. Среднее значение признака в кластере — это доля голосов за данную партию на данном территориальном избирательном участке. Среднее значение в совокупности — общая доля голосов за данную партию. Результаты вычислений приведены в таблице 1.

Таблица 1

Коэффициент межкластерной корреляции для каждой партии

Партия	ρ
Коммунистическая партия	0,075
Свобода	0,137
УДАР	0,042
Батьківщина	0,119
Партия регионов	0,183

Пусть объем нашей выборки будет составлять 1200 респондентов. Применим полученный коэффициент межкластерной корреляции к формуле (1), чтобы выяснить, как влияет на дизайн-эффект количество кластеров в выборке. Средний объем кластера примем от 1 до 20, поскольку он линейно связан с количеством кластеров (объем выборки = количество кластеров × средний объем кластера; см. табл. 2).

Очевидно, что чем меньше будет объем кластера и чем больше, соответственно, будет этих кластеров в выборке, тем ниже будет дизайн-эффект.

Но на практике, разумеется, мы столкнемся с тем, что выборка объемом 1200 респондентов из 60 городов по 20 респондентов в каждом кластере будет значительно дешевле, чем выборка объемом 1200 респондентов из 120 городов по 10 респондентов в каждом. Дело в том, что каждый новый кластер в выборке ведет к существенному удорожанию полевых работ, посколь-

ку транспортные затраты значительно превышают оплату проведения интервьюером дополнительных интервью.

Таблица 2

Зависимость дизайн-эффекта от количества кластеров в выборке для каждой партии

Размер кластера	Количество кластеров	Коммунистическая партия	Свобода	УДАР	Батьківщина	Партия регионов
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

Поэтому при определенном фиксированном объеме ресурсов мы можем провести исследование по большей выборке, но с небольшим количеством городов в выборке; а также опросить большое количество городов, но тогда выборку придется уменьшить.

Именно здесь мы оказываемся перед проблемой: как распределить ресурсы на исследование, чтобы получить наилучший результат? Самая большая выборка респондентов не означает самой низкой погрешности. Если провести опрос 1200 респондентов только в Киеве, Львове и Донецке, это будет значительно хуже в плане репрезентативности, нежели опросить в целом 800 респондентов, но в 10 разных городах Украины.

Для начала нам нужно знать, как вычисляется стоимость полевого этапа исследования, то есть как влияет на стоимость дополнительное интервью для интервьюера и дополнительный населенный пункт, до которого интервьюеру придется добраться, чтобы провести свои интервью. Иными словами, следует установить, как задается функция затрат, определяющая, как затраты ресурсов на исследование связаны с другими факторами.

Разумеется, каждая исследовательская компания будет по-своему вычислять стоимость реализации конкретной выборки, и на эту стоимость может влиять множество факторов: расстояние населенного пункта до ближайшего областного центра, расстояние до железной дороги, расположение опросных центров и т.п. При желании все их можно учесть и построить довольно сложную функцию затрат, но в данном исследовании будем считать, что на стоимость реализации выборки влияет только два фактора: транспортные затраты (одинаковые для всех кластеров) и оплата за одно интервью. То есть на стоимость будет влиять количество респондентов в выборке и количество кластеров. Эту функцию затрат можно выразить следующей формулой:

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

где
k — количество кластеров;
c_{cl} — транспортные расходы на один кластер;
n — объем выборки;
c_r — стоимость одного интервью.

Поскольку количество кластеров в выборке определяется как $k = n/m$, где m — объем кластера,то можно записать следующую формулу:

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

Решим теперь это уравнение для n:

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

Итак, если знать транспортные затраты на один кластер и стоимость одного интервью и задать общую сумму затрат, то можно сравнить, как это повлияет на объем выборки.

Пусть, например, стоимость одного интервью — 32 денежные единицы, а транспортные затраты — 200. При этом общий объем ресурсов, выделенных на полевой этап исследования, составляет 60000 денежных единиц. В зависимости от размера кластера мы получим определенный объем выборки (табл. 3).

Если бы дизайн-эффекта от кластеризации не существовало, то очевидно, что наибольший объем выборки давал бы самую низкую погрешность. На основании таблицы 2 мы уже убедились в том, что дизайн-эффект увеличивается по мере увеличения размера кластера, поскольку дизайн-эффект связан с объемом выборки следующим образом (см.: [Kish, 1965: p. 162]):

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

То есть если эффективный объем выборки равен реальному объему, разделенному на дизайн-эффект, то вычислить погрешность текущей выборки можно по формуле:

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

Данные касательно дизайн-эффекта для каждой из партий приведены в таблице 1. Отсюда вычислим погрешность выборки для каждой партии в зависимости от заложенного в выборке объема кластера (см. табл. 4).

Таблица 3

Зависимость объема выборки от размера кластера и погрешность для простой случайной выборки такого же объема

Размер кластера	Количество кластеров	Объем выборки	Погрешность простой случайной выборки
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

Таблица 4

Зависимость погрешности выборки от объема кластера

Размер кластера	Объем выборки	Коммунистическая партия	Свобода	УДАР	Батьківщина	Партия регионов
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

Окончание 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

Посмотрим на таблицу 4 в виде графика, приведенного на рисунке.

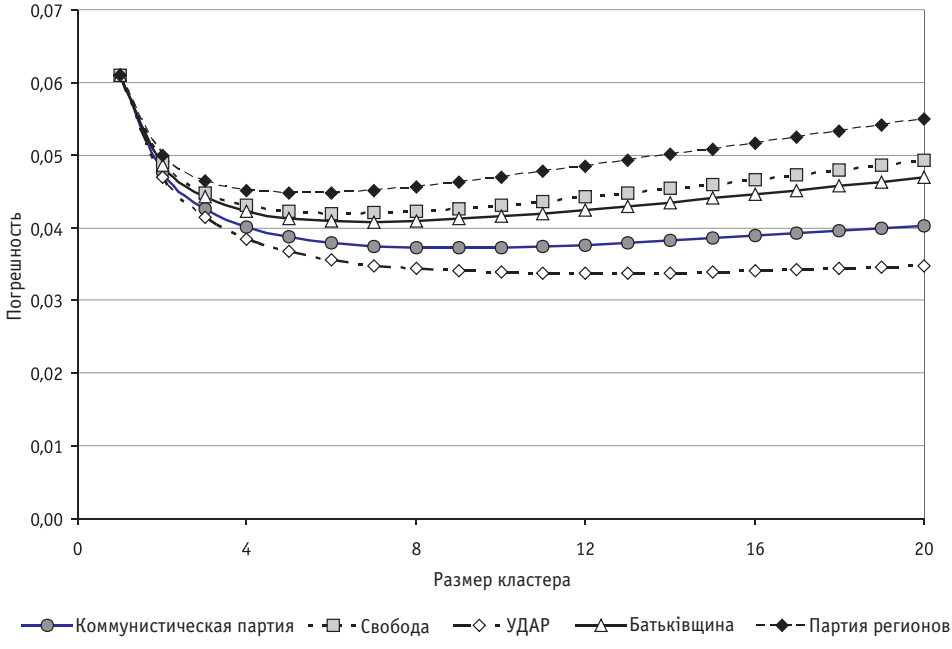

Рис. Зависимость погрешности выборки от объема кластера

Как видим, связь между объемом выборки и погрешностью нелинейная, и каждая партия достигает минимальной погрешности при определенном объеме кластера. Причем эта точка оптимума у каждой партии своя и зависит от коэффициента межкластерной корреляции (см. табл. 5).

Как видим, чем ниже был коэффициент межкластерной корреляции, тем больший размер кластера является допустимым и, соответственно, тем больше будет общий объем выборки. В Партии регионов коэффициент межкластерной корреляции наибольший, поэтому для того, чтобы выборка была как можно более репрезентативной для нее, она должна состоять из

большого количества кластеров, что обусловливает сокращение общего объема выборки.

Таблица 5

Оптимальное количество кластеров для каждой партии

Партия	ρ	Размер кластера	Количество кластеров	Объем выборки	Погрешность
Коммунистическая партия	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
Партия регионов	0,183	5	167	833	0,045

Итак, чтобы рассчитать оптимальное количество кластеров в выборке, необходимо знать коэффициент межкластерной корреляции и функцию затрат.

Воспользуемся данными из нашего примера, чтобы продемонстрировать расчет оптимального количества кластеров.

Если в формулу (6) подставить (1), то увидим, что полностью формула вычисления погрешности выборки выглядит так:

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

Если вместо n подставить формулу (4), то получим:

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

Пусть мы оптимизируем выборку для достижения минимальной погрешности для ГО “Свобода”. Коэффициент межкластерной корреляции для нее равен 0,137. Стоимость одного интервью — 32 денежные единицы, транспортные затраты — 200, общий объем ресурсов — 60000 денежных единиц.

Подставим эти значения и получим:

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

Теперь необходимо найти минимум этой функции. Для этого найдем для нее производную по m:

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

Приравняем ее к 0:

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

В качестве решения этого уравнения получим:

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

Мы нашли минимум рассматриваемой функции и теперь знаем, что самую низкую погрешность выборки получим, если размер кластера будет равен 6.

Выводы

Наибольшее влияние на погрешность кластерной выборки оказывают следующие факторы: общий объем выборки, количество кластеров в выборке и коэффициент межкластерной корреляции.

При условии ограниченности ресурсов на проведение исследования от объема этих ресурсов и функции затрат зависят общий объем выборки и количество кластеров в выборке. Для создания кластерной выборки с самой низкой возможной погрешностью исследователь должен определить, из какого количества кластеров должна состоять его выборка и какого объема она должна быть, чтобы не превышать пределов имеющихся ресурсов. Сначала рассчитывают коэффициент межкластерной корреляции исследуемого признака, или признака, который можно использовать вместо него. Потом выводят функцию затрат, которая должна показать, как связаны общие затраты на исследование с объемом выборки и количеством кластеров в ней. В каждом случае это может быть своя функция, но в целом она должна показывать эти связи. Далее выводят общую формулу, которая связывает погрешность с размером кластера. Размер кластера, при котором погрешность выборки будет самой низкой, будет равняться минимуму рассчитанной функции.

Источники

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.