

Журнал “Социальные исследования”

Алгоритмы формирования выборки социологического опроса

Фархад Назипович Ильясов^a *

^a Независимый исследователь, кандидат философских наук по специальности «прикладная социология»

О СТАТЬЕ

Прохождение статьи:
Поступила: 21.11.2017
Принята: 14.12.2017
Опубликована онлайн:
28.12.2017

Ключевые слова:
виды выборки; алгоритмы
отбора; репрезентативность

АННОТАЦИЯ

В статье уточняются термины, связанные с понятиями выборки и репрезентативности. Описывается классификация выборок, принятая в социологии. Предлагается и описывается классификация выборок по месту нахождения респондента и по методу контакта с ним.

1. Введение

Проблема выборки в социальных обследованиях имеет очень большое значение, тем не менее, эта тема остается не в полной мере разработанной и проясненной. Особенно это касается не-вероятностных, случайных выборок (англ. random sample). Ред Бейкер с коллегами отмечают: «В отличие от вероятностной выборки, нет единого основания, которое позволило бы адекватно классифицировать и описать все виды не-вероятностных выборок. Не-вероятностные выборки представляет собой набор, коллекцию методов, и трудно, если не невозможно, описать особенности и алгоритмы, применяемые во всех видах не-вероятностных выборок» [Baker et al., 2013: 3].

* E-mail: iliassov.farkhad@yahoo.com

Целью настоящей статьи является уточнение имеющихся классификаций и описаний алгоритмов отбора в социальных исследованиях.

2. Уточнение понятий

Генеральная совокупность - группа, выбранная в качестве объекта исследования. В соответствии с задачами исследования, генеральная совокупность выбирается по двум и более признакам. Однако, даже, если для выделения группы выбраны два «целевых» признака, например: 1. «все население определенной страны», 2. «в возрасте 18+», то выделяются еще и дополнительные репрезентирующие признаки. Это могут быть пол, возраст, образование, социальное положение, национальность, место расположения, детность, брачность, религиозность и т.д.

Генеральные совокупности, формируемые респондентами – это совокупности респондентов, находящихся по своей воле, и в силу сходной детерминации поведения, в определенных местах или осуществляющих определенные онлайн-контакты. Это могут быть:

- торговые центры, зрелищные заведения, точки общепита;
- место работы, учебы;
- лечебные учреждения;
- контакты через базы данных, панели, онлайн-опросных компаний;
- аудитории определенных сайтов.

Выборка, выборочная совокупность – часть генеральной совокупности, отобранная для первичных измерений. Первичные измерения в социологии это измерение свойств первичного объекта измерения – отдельных респондентов, через измерение свойств которых измеряются свойства изучаемой группы (выборки). Конечный объект измерения в отдельном социологическом исследовании – это специфическая группа, выборка или генеральная совокупность.

Репрезентативность выборки – это её свойства точно отражать исследуемые характеристики генеральной совокупности. Строго говоря, репрезентативность – это не единое, общее свойство всей выборочной совокупности, а это характер распределения ответов на тот или иной конкретный вопрос анкеты. То есть репрезентативным являются распределение ответов на определенный вопрос.

Характер рассеяния разных свойств респондентов в генеральной совокупности может быть различным, оттого в одном опросе уровень репрезентативности ответов может быть различным по разным вопросам. Порой используют не совсем корректное выражение «репрезентативный опрос», на самом деле имея в виду репрезентативность выборки, либо репрезентативность результатов опроса, тогда как сам опрос – это процесс сбора эмпирической информации.

Иногда репрезентативной называют выборку, которая по социально-демографическим характеристикам опрошенных воспроизводит генеральную совокупность, при этом репрезентативность ответов на содержательные вопросы может оставаться неизвестной. Порой под репрезентативной понимают выборку, которая с точки зрения теории математической статистики, «должна быть» репрезентативной. Однако в данном случае речь идет только о гипотезе, которая, как правило, сохраняет статус непроверяемой.

Репрезентативность выборки может определяться только эмпирически:

- сравнением социально-демографических параметров выборки и генеральной совокупности;
- сравнением с результатами повторного опроса;
- измерением репрезентативности результатов массового опроса методом анализа повторных подвыборок, подробнее см.: [Ильясов, 2011].

Параметры выборки – определение объема выборки и того, какие параметры генеральной совокупности должна репрезентировать выборка. Это могут быть пол, возраст, доход, профессия, социальное положение, обладание определенными предметами, потребление определенных товаров и т.д.

Проектирование выборки – разработка алгоритма отбора и определение количества опрашиваемых респондентов.

Формирование выборки – реализация алгоритма отбора. Целью формирования выборки часто является создание репрезентативной выборки.

Ошибка выборки. Как отмечал Геннадий Батыгин: «Практически ошибка выборки определяется путем сравнения известных характеристик генеральной совокупности с выборочными средними. ...В качестве контрольных параметров обычно применяются социально-демографические признаки» [Батыгин, 2008: 149]. Важно указать – показатель «статистическая погрешность», описываемый в

терминах доверительного интервала и уровня доверия, является теоретическим понятием математической статистики и никак не характеризует реальную выборочную совокупность. О доверительном интервале и статистической погрешности подробнее см., например: [Антонов, 2013]. Также некорректными следует признать построения, основанные на идее нормального распределения, т.к. нет социологического свойства, относительно которого были бы получены достоверные эмпирические данные, подтверждающие нормальное распределение величин его свойств.

Удобная выборка (англ. convenience sampling) – это виды выборок, в которых отбор осуществляется в генеральных совокупностях, формируемых респондентами.

Размер, объем выборки – количество отбираемых респондентов. Как указывается в известном издании: «Выборочный метод не дает универсального решения относительно необходимого в каждом конкретном случае объема выборки, обязательного соотношения объемов генеральной и выборочной совокупностей» [Рабочая книга..., 1977: 266]. На практике размер выборки определяется эмпирически, исходя из результатов анализа ранее проведенных опросов. Следует указать, что, как известно, так называемый «расчет объема выборки по специальной формуле», основанный на показателе дисперсии величин некоторого свойства (признака), возможен только при известности дисперсии в генеральной совокупности, однако если дисперсия известна, опрос проводить не имеет смысла.

Квотное задание – сформулированное для конкретного интервьюера на основе параметров выборки, задание опросить конкретное число лиц с определенными свойствами, признаками. Это может быть, например, число женщин и число мужчин заданного возраста.

Тип отбора респондентов – при самом общем подходе можно выделить следующие типы отборов:

1. рандомный (называемый также вероятностным или случайным);
2. эмпирический, основанный на опыте предыдущих отборов;
3. стихийный, отбор первого попавшегося респондента, номера телефона и т.д.
4. произвольный, отбор по «эмоциональному основанию».

В первом случае отбор основан на строгом алгоритме с использованием основы выборки, таблицы или генератора случайных чисел. Во втором случае речь

идет об использовании алгоритмов отбора, эмпирически показавших свою возможность обеспечивать репрезентативность выборки. В третьем и четвертом случаях строго алгоритма отбора нет, и вопрос репрезентативности может оставаться открытым.

В настоящем тексте используется термин «рандомная», чтобы не применять многозначное слово «случайная» и не совсем точное «вероятностная» (отбор основан на вероятности, однако производится случайным образом).

Шаг, ступень выборки – элемент алгоритма отбора.

3. Виды выборок

Выражение «виды выборки» используется в разных смыслах, например:

1. техническое задание на отбор респондентов, параметры отбора, квотное задание;
2. метод, алгоритм, критерий, стратегия отбора;
3. этап, шаг отбора.

Соответственно, классификации выборок отражают многозначный характер этого понятия, описывая, как правило, один или два ключевых критерия, подхода, алгоритма.

В литературе выделяются, в частности, следующие виды выборок:

- гнездовая;
- квотная (пропорциональная);
- кластерная;
- удобная (конформная¹);
- маршрутная;
- «уличный отбор»
- районированная (типическая);
- серийная;
- снежного кома²;
- стратифицированная (расслоенная), и др.

¹ Подробнее см.: [Рогозин, 2008].

² Модификацией является «выборка, направляемая респондентом».

Более подробно о классификации видов выборок см., например: [Рабочая книга..., 1977: 258-297; Могильчак, 2015; Alvi, 2016].

Понятия «квотная, кластерная, районированная, стратифицированная» являются в определенной мере сходными и обозначают выделение некоторых непересекающихся групп, страт по отдельным основаниям, определяемым исследователем.

Иногда стратами называют разные подвыборки, формируемые в рамках одного исследования из двух или более сравниваемых генеральных совокупностей, к их числу относятся:

- группы, имеющие существенные социальные различия, например, сельское и городское население;
- группы, различающие по однородности распределения, дисперсии исследуемого показателя, подробнее см., например: [Чуриков, 2007].

Сходными являются также понятия гнездовой и серийной выборки, они обозначают процесс выбора групп (семья, школьный класс, подразделение в организации), подлежащих сплошному опросу.

Не во всех случаях термин выборка обозначает алгоритм отбора респондентов. А в тех случаях, когда речь идет об алгоритмах отбора, указывается, как правило, лишь один или два алгоритма, в соответствии с которыми виду выборки дается наименование. По сути, указание вида выборки, обычно представляет собой лишь маркирование, обозначение некоторого концепта. Основная проблема в понимании вида выборки – это нередкое отсутствие описания полного алгоритма отбора. В виду указанного обстоятельства названия видов выборок в большей мере представляют собой метафоры, в которые исследователи вкладывает некоторое, порой не до конца раскрываемое содержание.

4. Рандомный, случайный отбор

С точки зрения теории статистики, случайный выбор – это выбор одного варианта из нескольких равновероятных. Рандомный, статистически случайный отбор – это алгоритм отбора респондентов, при котором каждая из единиц генеральной совокупности имеет равные шансы, одинаковую вероятность, попасть в выборку. При рандомном отборе есть возможность посчитать теоретическую

вероятность попадания каждого респондента в выборку, потому она называется также вероятностной. Более подробно о рандомных (вероятностных, случайных) выборках см., например: [Чуриков, 2007].

Надо заметить, использование рандомного, вероятностного отбора не гарантирует репрезентативность выборки. Репрезентативность в данном случае – это просто предположение, вытекающее из теории математической статистики, некоторая априорная предпосылка, которая может реализоваться в конкретном отборе, а может и не реализоваться.

Рандомной может считаться, например, выборка, в которой основой выборки является полный список всех единиц генеральной совокупности. Это может быть список избирателей, список адресов, список работников организации и т.д. В соответствии с принятым объемом выборки, отбор может осуществляться через определенный интервал, шаг выборки. Рандомный отбор может быть осуществлен с помощью компьютерной программы, с использованием генератора случайных чисел. В программу вносятся два показателя – размер генеральной совокупности и размер выборки. Эта процедура реализуема и с использованием известных статистических пакетов.

В варианте с территориальной выборкой рандомность обеспечивается случайным выбором каждого элемента на каждом шаге выборки.

Из выборок, которые могут считаться рандомными, довольно распространенными становятся выборки с использованием таких основ выборки как списки стационарных и мобильных, домашних и корпоративных телефонов.

5. Эмпирические выборки

На практике алгоритм строго рандомного отбора реализовать сложно. Потому все чаще используются так называемые «эмпирические» выборки, под которыми понимаются различные совокупности алгоритмов рандомного и «стихийного», «произвольного» отбора, которые, как ожидается исследователем, обеспечивают репрезентативность. Эмпирическими они называются потому, что оптимальный алгоритм отбора в них определяется эмпирически, через апробацию тех или иных алгоритмов и нахождения наиболее точного, оптимального.

Эмпирические выборки также называют не-вероятностными, неслучайными. Однако они называются не-вероятностными, строго говоря, не потому, что не

осуществляется рандомный отбор, а лишь потому, что вероятность попадания каждой единицы отбора в выборку неизвестна.

Иногда эмпирические выборки оценивают как не имеющие теоретического обоснования. Например, Ред Бейкер с коллегами отмечают: «Мы полагаем неприемлемыми для формулирования статистических выводов и предположений, используемые без теоретического обоснования методы сбора данных и сформулированные на их основе оценки» [Baker et al., 2013]. Однако теория лишь обобщает экспериментальные данные, а именно эмпирические данные являются основой позитивной науки. Критерием научности, как известно, является совпадение результатов повторяемых, воспроизводимых эмпирических процедур. Если проводится некая эмпирическая процедура, с хорошо прописанными алгоритмами действий, в итоге которой получается воспроизводимый и повторяемый результат, то этот результат может считаться научным. Например, феномен времени в физике не имеет признанного теоретического обоснования, однако это не делает неприемлемыми эмпирические данные и оценки, полученные на основе измерения времени (длительности).

Большинство выборок представляют собой некий алгоритм – последовательность определенных действий, шагов (ступеней), направленных на нахождение и отбор респондентов. Такие выборки называются многоступенчатыми. При этом нередко выборка представляет собой комбинацию различных видов отбора, такая выборка называется комбинированной. Значительная часть эмпирических выборок содержат в себе элементы рандомного и стихийного отбора.

Таким образом, в реальности эмпирическая выборка часто представляет собой алгоритм, состоящий из нескольких неоднородных, последовательных процедур отбора, имеющих рандомный, не строго рандомный и «стихийный» характер. Иными словами, на практике часто используется многоступенчатая, комбинированная выборка, использующая рандомные и стихийные принципы отбора.

Как представляется акцент на количестве ступеней и на комбинации методов отбора, как классифицирующих признаках, в некоторой степени заслоняет собой основную цель выборки – нахождение и отбор респондентов.

Каждая из эмпирических выборок представляет собой алгоритм отбора, состоящий из двух элементов:

1. алгоритм поиска локации респондента, либо выбор способа контакта с ним;
2. алгоритм отбора респондентов по заданным признакам в месте локации, либо через выбранный способ контакта.

Соответственно, можно выделить два основания для классификации видов выборок, алгоритмов отбора:

1. если контакт очный, непосредственный – это отбор по месту расположения, локация респондента;
2. если контакт опосредованный, дистанционный – это отбор по способу контакта с респондентом.

6. Эмпирические виды выборок по месту нахождения респондента

По основанию локации респондентов можно выделить следующие основные эмпирические виды выборок:

1. По месту жительства, см. рис. 1;
2. По месту работы или учебы, см. рис. 2;
3. По локации в городе – на улице, площади;
4. По месту приобретения товаров и услуг.

Объектом исследования в случае выборки по месту жительства могут являться генеральные совокупности – территориальные общности разного уровня: страна, область, город, район. Построение выборки начинается с локации нужного масштаба. Определяется размер и параметры выборки. Алгоритм формирования выборки по месту жительства респондента приведен на рис. 1. На первых пяти ступенях отбора, от региона до наименования улицы, отбор может производиться либо случайным образом, либо отбираются типичные на взгляд исследователя локации, соответствующие целям изучения.

Для каждой из ступеней отбора могут формироваться свои квоты. Они могут быть пропорциональными – отражать социально-демографические и иные

пропорции генеральной совокупности, т.е. воспроизводить структуру генеральной совокупности.

Квоты могут быть «аналитическими», т.е. соответствовать аналитическому плану исследования. Они создаются таким образом, чтобы минимальная, запланированная для анализа группа, была не менее 30-40 респондентов. Например, это может быть группа: женщины, возраст 40-45 лет, доход средний, с определенным видом поведения. Различные случаи выделения квот, страт, см., например: [Могильчак, 2015: 39-53].

Рис. 1. Алгоритм формирования выборки по месту жительства респондентов



До интервьюера эти квоты доводятся в виде квотного задания, в котором указывается, сколько человек с определенными социально-демографическими и иными свойствами он должен опросить.

В случае отбора по месту жительства, см. рис. 1, внутри последней ступени локации (домохозяйства) в начале опроса может производиться рандомный выбор респондента (например, по ближайшему дню рождения) между двумя или более респондентами, соответствующими квотному заданию. После выполнения квотного задания по отдельным позициям, далее отбор производится только в соответствии с квотным заданием.

Рис. 2. Алгоритм формирования выборки по месту работы респондентов



Алгоритм уличного отбора во многом совпадает с алгоритмом отбора по месту жительства, только здесь конечной, опросной точкой является место на улице. Как

правило, уличный опрос проводится на основе квотных заданий. На первом этапе, до выбора квот, респонденту дается задание отбирать для опроса проводящих мимо респондентов через определенный шаг, например, каждого пятого. Этот этап отбора можно считать рандомным. На втором этапе респондент отбирает в соответствии с недобранными квотами, например, только мужчин среднего возраста. Этот этап отбора можно полагать стихийным.

Опыт проведения уличных опросов показывает, что, при правильной организации, они могут обеспечивать высокий уровень репрезентативности.

Аналогичным алгоритму отбора по месту работы, см. рис. 2, является алгоритм отбора по месту учебы.

Выборки по месту приобретения товаров и услуг обычно формируются в рамках исследований потребительского поведения, измерения продвинутости брендов и т.д. Такая выборка может считаться частным случаем выборки по локации в городе. В данном случае речь идет о генеральных совокупностях, формируемых респондентами. Подобные выборки репрезентируют то место скопления потребителей, где проводится опрос. В этих условиях может решаться обратная задача выборки, см., пример: [Ильясов, 2016].

8. Выборки по способу контакта с респондентом

По основанию типа дистанционного контакта с респондентом можно выделить следующие основные виды выборок:

1. рандомный или стихийный отбор респондентов из списка номеров стационарных и мобильных телефонов в домохозяйствах и организациях;
2. само-отбор (англ. self-selection sampling) – участие в опросе респондентов по собственной инициативе, в ответ на предложение принять участие в онлайн опросе, размещенное на определенных сайтах.
3. рандомный отбор респондентов в опросных веб-панелях (англ. web panels), в базах данных лиц, выразивших ранее желание принимать участие в опросах;
4. отбор из основы выборки (базы данных) респондентов для рассылки анкет обычной или электронной почтой.

В случае выборки, создаваемой на основе списка телефонов, она может быть репрезентативной при соблюдении следующих условий:

- генеральная совокупность имеет высокий охват телефонной связью;
- список является полным;
- производится рандомный отбор.

В случае выборки само-отбора, осуществляемой посетителями сайтов, выборка может рассматриваться как репрезентативная только относительно аудитории этих сайтов. Выборка само-отбора имеет признаки стихийной. Однако, она может рассматриваться как эмпирическая, в случае, если результаты онлайн опроса согласуются с результатами другого опроса, определенными как репрезентативные. Подробнее о выборке само-отбора см., например: [Bethlehem, 2008].

Упорядоченным вариантом выборки само-отбора является выборки из взб-панелей, состав которых формируется за счет предшествующего само-отбора. Респонденты сами откликаются на объявления с предложениями включиться в панель, регистрируются на соответствующем сайте, указывает свои соц-дем характеристики, потребительские и иные особенности. Веб-панели создаются организациями, специализированными на формировании этих панелей и проводящих интернет опросы с оплатой ответов респондентам.

Организация, для которой проводится опрос, формулирует техническое задание на опрос, в котором указываются свойства, признаки и квоты (пропорции), по которым должна формироваться выборка. В этом случае выборка может быть репрезентативной относительно самой веб-панели. Репрезентативность же самой веб-панели может быть неизвестной. Она может быть проверена с помощью контрольных вопросов, использованных ранее в опросах, репрезентативность которых является достоверной.

Существуют также репрезентативные веб панели, например, голландская веб панель LISS (www.lissdata.nl) состоит из 5 000 домашних хозяйств, состоящих из 8 000 человек. Панель основана на реальной рандомной (вероятностной) выборке домохозяйств, взятых из реестра населения Статистического управления Нидерландов [Stoop et al., 2012: 17].

Репрезентативность почтового опроса зависит от репрезентативности основы выборки. Популярность этого вида отбора (и опроса) в последнее время уменьшается.

Литература

- Антонов Г. В. Выборочный метод в социологических исследованиях // Научный диалог. 2013. №11. С. 96-109.
- Батыгин Г. С. Лекции по методологии социологических исследований. М.: РУДН. 2008. - 368 с.
- Ильясов Ф. Н. Репрезентативность результатов опроса в маркетинговом исследовании // Социологические исследования. 2011. № 3. С. 112-116.
- Ильясов Ф. Н. Обратная задача выборки и мотивация на рынке Форекс // Социальные исследования. 2016. №2. С. 49-59.
- Могильчак Елена Львовна Выборочный метод в эмпирическом социологическом исследовании. Екатеринбург: УрФУ. 2015. – 120 с.
- Рабочая книга социолога. М.: Наука. 1977. – 511 с.
- Рогозин Д. М. Конформная выборка в торговых центрах // Социологический журнал. 2008. №1. С. 22-48.
- Чуриков А. Случайные и неслучайные выборки в социологических исследованиях // Социальная реальность. 2007. №4. С. 89-109.
- Alvi, Mohsin Hassan. A Manual for Selecting Sampling Techniques in Research. MPRA Paper No. 70218, posted 25 March 2016. Available online: <https://mpra.ub.uni-muenchen.de/70218/> (Accessed: 2017.12.02)
- Baker R., Brick J. M., Bates N. A., Battaglia M., Couper M. P., Dever J. A., Gile K. J., Tourangeau R. Report of the AAPOR task force on non-probability sampling. June 2013. Available online: http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf (Accessed: 2017.12.02)
- Bethlehem, Jelke. How accurate are self-selection web surveys. The Hague/Heerlen: Statistics Netherlands. 2008. Available online: <https://peilingpraktijken.nl/wp-content/uploads/2014/06/bethlehem04.pdf> (Accessed: 2017.12.02)
- Stoop, Ineke and Harrison, Eric. Classification of Surveys. In: Handbook of Survey Methodology for the Social Sciences. Ed. Gideon L. New York: Springer Science + Business Media. 2012. P. 7-21.

Algorithms for sampling a sociological survey

Farkhad Nazipovich Iliassov *

* - An independent researcher. PhD in «Applied Sociology»

Email: iliassov.farkhad@yahoo.com

Abstract

The article clarifies the terms associated with the concepts of sampling and representativeness. The classification of samplings, adopted in sociology, is described. The classification of samplings according to the location of the respondent and the method of contact with him is proposed and described.

Keywords: types of sample; sampling algorithms, representativeness

References

- Antonov G. V. Vyborochnyi metod v sotsiologicheskikh issledovaniyakh. *Nauchnyi dialog*. 2013. No. 11. P. 96-109.
- Batygin G. S. Lektsii po metodologii sotsiologicheskikh issledovaniy [Lectures on the methodology of sociological research]. Moscow: RUDN. 2008. 368 p.
- Iliassov F. N. Reprezentativnost' rezul'tatov oprosa v marketingovom issledovanii. *Sotsiologicheskie issledovaniya [Sotsiologicheskie Issledovaniia]*. 2011. No. 3. P. 112-116.
- Iliassov F. N. Obratnaya zadacha vyborki i motivatsiya na rynke Foreks. *Sotsial'nye issledovaniya [Journal of Social Research]*. 2016. No.2. P. 49-59.
- Mogil'chak Elena L'vovna Vyborochnyi metod v empiricheskom sotsiologicheskom issledovanii [Sampling method in an empirical sociological study]. Ekaterinburg: UrFU. 2015. 120 p.
- Rabochaya kniga sotsiologa [Working book of a sociologist]. Moscow: Nauka. 1977. 511 p.
- Rogozin D. M. Konformnaya vyborka v torgovykh tsentrakh. *Sotsiologicheskii zhurnal*. 2008. No. 1. P. 22-48.
- Churikov A. Sluchainye i nesluchainye vyborki v sotsiologicheskikh issledovaniyakh. *Sotsial'naya real'nost'*. 2007. No. 4. P. 89-109.
- Alvi, Mohsin Hassan. A Manual for Selecting Sampling Techniques in Research. MPRA Paper No. 70218, posted 25 March 2016. Available online: <https://mpra.ub.uni-muenchen.de/70218/> (Accessed: 2017.12.02)
- Baker R., Brick J. M., Bates N. A., Battaglia M., Couper M. P., Dever J. A., Gile K. J., Tourangeau R. Report of the AAPOR task force on non-probability sampling. June 2013. Available online: http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf (Accessed: 2017.12.02)

Bethlehem, Jelke. How accurate are self-selection web surveys. The Hague /Heerlen: Statistics Netherlands. 2008. Available online: <https://peilingpraktijken.nl/wp-content/uploads/2014/06/bethlehem04.pdf> (Accessed: 2017.12.02)

Stoop, Ineke and Harrison, Eric. Classification of Surveys. In: Handbook of Survey Methodology for the Social Sciences. Ed. Gideon L. New York: Springer Science + Business Media. 2012. P. 7-21.